

DEVELOPMENT AND VALIDATION OF MOLECULAR DESCRIPTOR BASED ON PHYSICAL AND BIOLOGICAL PREDICTION MODELS

Thesis submitted to
University of Calicut in partial fulfillment of
the requirements for the award of the Degree of

DOCTOR OF PHILOSOPHY IN CHEMISTRY

By

SAJEEV R

Under the Guidance of

Dr. SREEJITH M NAIR

Associate Professor,
Department of Chemistry



**POST GRADUATE AND RESEARCH CENTRE
MALABAR CHRISTIAN COLLEGE
CALICUT
NOVEMBER 2017**

CERTIFICATE

This is to certify that the thesis entitled “**Development and validation of molecular descriptor based on physical and biological prediction models**” is an authentic record of research work carried out by Mr. Sajeev R., under my supervision in partial fulfillment of the requirements for the degree of Doctor of Philosophy, in Chemistry of University of Calicut and further that no part thereof has been presented before for any other degree.

Calicut
November 2017

Dr. Sreejith M. Nair
(Supervising Teacher)

CERTIFICATE

This is to certify that Mr. Sajeev R., Ph. D. student under my guidance has incorporated corrections/suggestions from the adjudicators in the thesis entitled **“Development and validation of molecular descriptor based on physical and biological prediction models”**.

Calicut
December 2018

Dr. Sreejith M. Nair
(Supervising Teacher)

DECLARATION

I, **Sajeev R** hereby declare that thesis, entitled “**Development and validation of molecular descriptor based on physical and biological prediction models**” is a record of original and independent research work carried out by me under the supervision of **Dr. Sreejith M. Nair**, Associate Professor, Department of Chemistry, Malabar Christian College, University of Calicut, Calicut and this work has not been submitted for any degree or diploma to any other University/Institute prior to this date.

Calicut
November 2017

Sajeev R

ACKNOWLEDGEMENTS

First and foremost, I am extremely grateful and indebted to Dr. Sreejith M Nair, Associate Professor, Department of Chemistry, Malabar Christian College, Calicut, for his excellent guidance, invaluable suggestions and unfailing encouragement without which I could not have accomplished this work. I would also like to thank him for taking time to review my dissertation.

I would like to express my sincere thanks to Dr. U C A Jaleel for his invaluable guidance, continued support, infinite patience and immense encouragement during the course of my PhD study. For me this was a journey started in December 2008 from Malabar Christian College Calicut entered into an interdisciplinary world of teaching and subject of informatics (Cheminformatics) lead by Dr. U C A Jaleel and it was my pleasure in continuing the research and I am still holding this passion as far as I can. I would also like to thank the moral support by him in mentoring me in all the difficult situations faced in my life.

I would like to thank Dr. Godwin Samraj Principal, Malabar Christian College, all the faculty members of Chemistry Department, especially former Principal Gladys P Isacc and Pavamani Mary of this college and Head of this Department Dr. Susannah Seth, Prof. Haris, lab staff Mrs. Smitha, for the invaluable support they offered whenever I was in need.

I also thank the Tata CSIR OSDD Fellowship for support and funding for the betterment of my research at Indian Institute of Science, Bangalore during the period 2014-2016. I would like to thank OSDD chief mentor Prof. Sameer K Brahmachari, the Head OSDD Dr. T., S., Balganes, Dr. Bheemrao Ugarkar (scientist) and Dr. U C A Jaleel (P. I). Also my sincere gratitude to our OSDD team members Mr. Jinuraj, Mr. Nufail M, Mr. Andrew Titus Manuel, Mr. Jayan, Mrs. Dhanlakshmi, Mr. Lijo John, Mr. Chandan Kumar, Mrs. Dakshyani, Mr. Yatindra Yadav, Mrs. Swathi Gandhi, Mrs. Aysha, Mrs. Rakhila, Mrs. Athira, Mrs. Jisha, Mr. Jadan Rasnik, Mrs. Sahida, Mr. Sijo Jose and many other friends like Mr. Adarsh,

Mrs. Preetha Anil, Mr. Fahim, Mr. Haris, Tom thomas etc. Without their support my work may not completed within the stipulated time.

Finally, my dream would have been impossible without the incessant co-operation, love and encouragement from my family members.

Sajeev R

*Dedicated to
My Family and Teachers*

CONTENTS

	Page No.
PART I	1-116
PHYSICAL PREDICTIVE MODELS	
Chapter 1 Introduction	1
Chapter 2 Materials and methods	31
Chapter 3 Development of Bayesian models based on organic semiconductors	52
Chapter 4 Predictive models based on Decision Tree analysis	75
Chapter 5 Support Vector Machines-SMO Models	92
Chapter 6 Pattern Search for organic semiconductors	101
References	104
PART II	117-265
Biological Predictive Models	
Chapter 1 Introduction	117
Chapter 2 Materials and methods	140
Chapter 3 Bayesian model against β -lactamase enzyme	152
Chapter 4 Decision tree model against β -lactamase enzyme	163
Chapter 5 Support vector models against β -lactamase enzyme	179
Chapter 6 Docking study against β -lactamase enzyme present in <i>M. tuberculosis</i> and <i>P. aeruginosa</i>	195
Chapter 7 Artificial Neural Network based Self Organizing Maps	213
Chapter 8 Sensitivity of molecular descriptors based virtual screening methods against β -lactamase enzyme	220
References	255
PART III	266-308
Development and Validation of Molecular Descriptor	
Chapter 1 Introduction	266
Chapter 2 Materials and methods	271
Chapter 3 Development and validation of molecular descriptor	272
References	307
Summary	309-312

ABBREVIATIONS

The following abbreviations are used in the thesis for the sake of easiness.

ADMET	Absorption Distribution Metabolism Toxicity
AI	Artificial Intelligence
ANN	Artificial Neural Network
ARFF	Attribute Relation File Format
ChEBI	Chemical Entities of Biological Interest
CSV	Comma Separated Value
D M	Data Mining
DT	Decision tree
Ev	Electron volt
GSK	GlaxoSmithKline
GUI	Graphical User Interface
HOMO	Highest Occupied Molecular Orbital
HTS	High-Throughput Screening
LUMO	Lowest Unoccupied Molecular Orbital
MDR TB	Multi-drug-resistant tuberculosis
ML	Machine learning
MM+	Molecular Mechanics
MTB	Mycobacterium Tuberculosis
N B	Naïve Bayes
OOB	Out of bag error
PDB	Protein Data Bank
PM3	Parameterized Model 3
QSAR	Quantitative Structure Activity Relationship
QSPR	Quantitative Structure Property Relationship
RF	Random Forest
SMARTS	SMiles ARbitrary Target Specification
SMO	Sequential Minimal Optimization

SOM	Self Organizing Maps
SVM	Support Vector Machines
TB	Tuberculosis
TDR TB	Totally Drug-Resistant Tuberculosis
VCC	Virtual Computational Chemistry Laboratory
VS	Virtual Screening
WEKA	Waikato Environment for Knowledge Analysis
XDR TB	Extensively Drug-Resistant Tuberculosis

PREFACE

Molecular descriptors play a pivotal role in a technology driven world where all the calculations are carried out by making use of supercomputers and advance algorithms to carry out better decision making. Molecular descriptors combined with data mining, machine learning and artificial neural networks have led to increased predictive models in various scientific disciplines. In this study molecular descriptor based electronic and anti-bacterial virtual screening predictive models are developed. The machine learning models are based on Bayesian classification, decision tree algorithm and support vector machines that are trained, tested, cross validated and model fitness was checked from various statistical parameters. The electronic predictive machine learning models developed on various algorithms and virtual screening results are summarized in Part I. Also electronic patterns were derived from the maximum common substructure analysis. Materials and methods are briefly described systematically in Part I and Part II.

In the Part II section anti-bacterial machine learning models were developed based on the various classification and decision tree algorithms. Biological methods were developed on the microbes *M. tuberculosis* and *P. aeruginosa* against the target enzyme β -lactamase. Apart from the data mining models, molecular docking and artificial neural network based self organizing maps were also carried out and the results are summarized here. Also the sensitivity of the anti-bacterial Bayesian model, structure based docking study and artificial neural network based high dimensional analysis were carried out among *M. tuberculosis* and *P. aeruginosa* under the selected target. The results are interpreted in this section.

Part III consists of the development and validation of a 2D walk descriptor developed by cross screening of electronic and anti-bacterial screening sets against physical and biological predictive machine learning models. The pattern lone pair pi walk count 8 was postulated and validated on the existing machine learning biological and electronic models. The developed molecular descriptor has the characteristics property both in electronic and anti-bacterial activity.

A detailed reference in serial order is mentioned at the end of each part.

PUBLICATION

1. Sajeev, R.; Athira, R. S.; Nufail, M.; Jinu Raj, K. R.; Rakhila, M.; Nair, S. M.; Abdul Jaleel, U. C.; Manuel, A. T. Computational Predictive Models for Organic Semiconductors. *J. Comput. Electron.* **2013**, *12*, 790–795.

PART 1
PHYSICAL PREDICTIVE MODELS

Sajeev R “Development and validation of molecular descriptor based on physical and biological prediction models” Thesis. Department of Chemistry, Malabar Christian College, Calicut, 2017.

CHAPTER 1

INTRODUCTION

1. Cheminformatics

Cheminformatics is a new discipline that is used to solve chemical problems with computers that a Chemist is facing. The advances made over the past 50 years in the field of Computer Sciences has helped shape the way in which today's chemical research is carried out in comparison to traditional methods.¹ It's a branch of Computational Chemistry that produces useful models that can predict chemical and biological properties of compounds. "Unlike Quantum Chemistry or molecular simulation, which are designed to model physical reality, Chemoinformatics is intended simply to produce useful models that can predict chemical and biological properties of compounds given the two-dimensional (or sometimes three) chemical structure of a molecule".² In earlier days, methods were developed for storing, indexing and retrieval of chemical structural information. Thereafter quantitative structure-activity/property relationship (QSAR/QSPR) models were developed that linked chemical activities with molecular structures and compositions for predicting physical, chemical, biological or environmental data.^{3,4} All of these studies had a common problem, particularly with the representation, manipulation and retrieval of chemical structural information. Later on, within a few decades, a new interdisciplinary field of research emerged where Chemistry, Biology, Computer Science and Mathematics coalesced. However, it was not until the late 1990's that a name was given to this field: Chemoinformatics. It has become clear that Chemoinformatics has applications in any field of Chemistry and related Sciences. Chemoinformatics is also called 'Cheminformatics' or 'Chemical Information Science' in various other definitions.⁵

Definitions of Chemoinformatics.⁶

1. "Cheminformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended

purpose of making better decisions faster in the area of drug lead identification and organization”.

2. “Cheminformatics – A new name for an old problem”.
3. “Chem(o)informatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information”.

There is a major difference between Chemoinformatics and Computational Chemistry. The former uses inductive learning; learning from data, for making predictions on chemical phenomena. While the latter, makes use of a theory for making predictions in a deductive learning process. “Presently, Chemoinformatics has found its most widely-accepted applications in the field of drug design”.^{1,7}

Before the field of Cheminformatics emerged, chemical information was enormous and it was difficult to find the relationship between structure and properties of a compound. Later on, the Chemical Abstracts Service and Swiss-German Chemical companies developed methods for the storage and searching of both structural and textual information in databases. Also the projects like Dendral (Stanford University) and Brandeis, Harvard, Stony Brook and the Technical University of Munich embarked on the developments of Cheminformatics.⁸ Hammett and Taft worked on the QSAR development involving quantification of steric and electronic influences on chemical reactivity while Hansch quantified hydrophobic effects and their influences on a variety of properties, but not effectively on the biological activity of drugs in 1964.⁹ Earlier models were simple linear models, typically built using only a very few features and were valid only for a small series of closely related compounds and later multi-linear regression models were developed.^{2,10} During this period, Free-Wilson analysis introduced the relationship between the presence/absence of certain substructures in a molecule in response to biological activity. Since then, the QSAR with the effects of substituent on the biological response of the molecules described by electronic, steric, and hydrophobic properties were applied to Agro Chemistry, Pharmaceutical Chemistry, and Toxicology. Many other groups like Langridge and Co-workers developed 3D

molecular models that was visualized on the screens of cathode-ray tubes first at Princeton, and then at UC San Francisco. And other visualizations including protein structures were carried out by Marshall from Washington University, USA.¹¹

1.1 Achievements of Cheminformatics

1.1.1 Databases

The most popular achievement of Cheminformatics is that it provides access to chemical information in databases available in the web resources for which the same would be unattainable by working through chemical literatures. For instance the universal chemical database GDB,¹² which presently has 166.4 billion molecules having up to 17 atoms of C, N, O, S and halogens of known compounds, otherwise would have been impossible for obtaining an overview of its chemical activity/property. Furthermore, databases enable the Chemist to communicate in their international language, structure diagrams and reaction equations. Chemical databases/resources are the backbones of the modern *in silico* drug discovery. These databases provide pieces of information that were used to build knowledge based models. Here we have provided a list of some of the major chemical databases in Table 1.

Table 1. Short descriptions of the various chemical databases are mentioned

Name and web link	Short description
PubChem ¹³ <i>pubchem.ncbi.nlm.nih.gov</i>	It's an open database that accepts data submission from various institutions and Govt. agencies. It contains 2,283,533 small molecule samples till date.
PubChem BioAssay ¹⁴ <i>ncbi.nlm.nih.gov/pcassay</i>	Contains screens of over one million records holding 230,000,000 bioactivity outcomes.
ChemSpider ¹⁵ http://www.chemspider.com/	It's a free online chemical database where chemical and physical properties, spectral data and molecular structure can be accessed. It offers nomenclature for over twenty six million unique chemical compounds.
ZINC ¹⁶ <i>zinc.docking.org</i>	It's a free database consisting of 35 million purchasable compounds; it is used for virtual screening, docking etc. Molecules can be searched from ZINC ID, SMILES, catalog, vendor code etc.

Name and web link	Short description
ChEMBL ¹⁷ <i>www.ebi.ac.uk/ChEMBLdb</i>	The database consists of bioactive drug-like small molecules along with their calculated properties like Lipinski Parameters, Molecular Weight, logP, and bioactivity information like Binding Constants, ADMET data etc.
ChemBank ¹⁸ <i>chembank.broadinstitute.org</i>	It is a public database that houses various chemical structures that contains calculated molecular descriptors, human curated bio-active information of small molecule activities and raw experimental results from HTS BioAssays.
DrugBank ¹⁹ <i>drugbank.wishartlab.com</i>	It is a public database containing information on drugs and drug targets. There are 10,507 drug entries and out of which 1,738 are approved small molecule drugs. It is highly accessed by the pharmaceutical companies, Medicinal Chemist's, pharmacists, physicians and researchers.
ChEBI ²⁰ <i>https://www.ebi.ac.uk/chebi/</i>	It is a freely available dictionary of molecular entities focused on "small" chemical substances. And 'molecular entity' refers to any isotopically distinct atom, ion, conformer, radical, molecule etc.

1.1.2 Property Prediction (QSAR/QSPR)

The complex relationships between many biological data of compounds and their structure are predicted either through QSAR/QSPR in a two-step process as shown in Figure 1. "In the first step, a molecular structure is represented by structure descriptors. In the second step, a dataset of structures as represented by their descriptors and their associated properties is submitted to a data analysis and model building method".^{1,21}

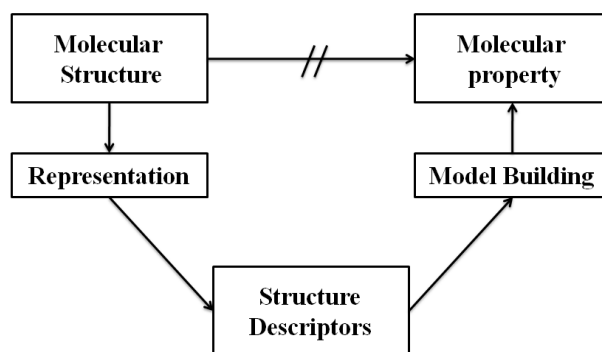


Figure 1. The basic approach for QSAR/QSPR model

1.1.3 Drug Design

By and large Chemoinformatics substantially has contributed towards a large number of applications in the field of drug design.²² Methods have been developed for Lead Discovery, Structure-based,²³ Ligand-based methods,^{24,25} Lead Optimization, Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET)²⁶ properties. This approach has grown so much so that, major drug companies have a Chemoinformatics department, as the newly designed molecules are screened utilizing various Chemoinformatics tools.^{1,27} The importance of computational methods in chemistry has been emphasized in the press release of the Nobel Prize in Chemistry 2013 “Today the computer is just as important a tool for chemists as the test tube.”²⁸ A methodology has been developed for the prediction of properties that cannot directly be calculated by theoretical methods. Thus, many physical, chemical or biological properties have been predicted from the information of the structure of a compound.

Cheminformatics models are built for many properties provided; a database is available with biological response or chemical activity profile. But especially, those of isolated molecules, Chemoinformatics would be a poor choice of methodology for model building. On one hand the calculations of the chemical properties like Dipole moment and Polarizabilities would be better off using Quantum Chemistry. Today, Chemoinformatics becomes a sensible option in explaining a complex biological system that cannot be easily modeled by physics-

based methods. There are also numerous physicochemical properties that are hard to obtain from theoretical chemical methods such as Density Functional Theory or Molecular Dynamics, and hence are often modeled by Chemoinformatics. The properties like aqueous solubility and logP (log of the octanol/water partition coefficient) have direct connection with drug discovery.^{2,29} At present, many problems have been solved, and interesting results were obtained due to the accessibility of chemical information available in various databases. Thus, Chemoinformatics tools and techniques have a great impact on today's chemical research in drug design.

1.2 Learning in Cheminformatics

1.2.1 Inductive learning vs. Deductive learning

Predictions are made through learning and there exists two types of learning systems: deductive and inductive.³⁰ The former makes use of a fundamental theory while the latter learns from a series of observations from which inferences are made to predict new observations. In deductive learning, fundamental theories do exist for Chemistry like Quantum Mechanics, Molecular Mechanics, empirical methods etc.^{28,31} In Quantum Mechanics, the property of a compound depends on its three dimensional structure given by the Schrodinger equation. However, the development of technology in computer application in hardware and software technology has allowed the calculations of many interesting properties of chemical compounds of fairly reasonable size with high accuracy.

In the case of inductive learning a model is built based on a set of observations like the essential features that are in common and that are different. Followed by which, the predictions are made analogically. The examples include inductive and resonance effects in organic chemistry that were not derived from theory but have been introduced to explain the experimental observations. There are enormous amount of data related to chemical, physical and biological properties of a chemical compound that have been determined and made accessible. And one of the major tasks is to derive knowledge from these data by inductive learning. A data can be any observation like the result or numerical value of a physical measurement or a

binary value to determine the action of a biological activity. From this data, information is obtained by putting one data into context with the other data. And finally, knowledge is obtained with some level of abstraction^{6,32} as shown in Figure 2.

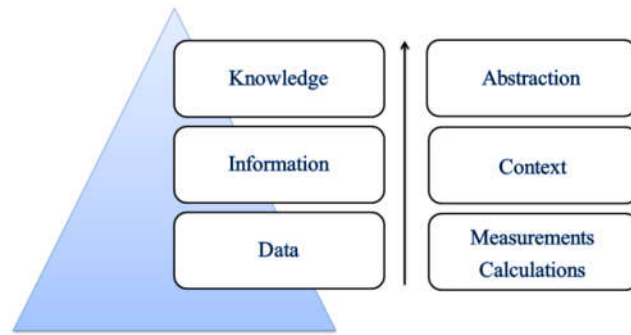


Figure 2. Illustrates the hierarchy from data through information to knowledge

1.3 Data Mining

Over the recent years, data has been increasing exponentially, which has led to the wastage of storage space as well as loss of hidden patterns. The massive data necessitated techniques to process and analyze these records. In the late 90's, the phrase "Knowledge Discovery in Databases" was introduced that later on came to be known as Data mining, as knowledge was the ultimate product.³³ The definition of Data mining includes "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."³⁴ Generally it's a process of extracting interesting hidden information from available chunks of data, which could otherwise have been manually impossible.^{33,35} Though it is meant for knowledge discovery, it encompasses approaches like classification, clustering, detecting anomalies etc.

1.3.1 Scheme of data mining process

Data mining process involves problem identification along with data collection for the betterment of the model analysis through statistical, visualization

tools, correlation analysis etc. A data mining tool needs to be versatile i.e. able to adapt to many different functions or activities, it should be scalable in the sense that models are applied on small to large datasets capable of accurately predicting responses between actions and results that can be automated.³⁶ Data mining models range from simple, parametric equations derived from linear techniques to complex and nonlinear models derived from nonlinear techniques. Data mining is important because it is applicable not only in the field of science but also in the field of marketing, banking and robotics as it uses data from various perspectives as well as summarizes it into useful information.^{37,38} It intersects numerous disciplines like Machine Learning, Statistics, Artificial Intelligence and Database Systems.³⁹ Most of the algorithms developed for data mining are statistically based. So, it is difficult to mention the best algorithm as it is in accordance to the definition of the problem and the structure of data.⁴⁰

1.3.2 Data Treatment

Data mining is essentially finding the useful pattern from bulk data sets in a statistical manner. The standard process involves breaking down of large datasets into two portions. One portion of the data is usually used as the *training set* for the development of the model (no matter what modeling technique is used), while reserving the second portion of the data to be used as the *test set* for testing the model that is built. Whereas in certain applications, a third portion of data is used for validating the model built. By doing so a more accurate model is obtained.^{41,42}

In the data mining process data is the input and knowledge is the output. Here no query is required but “interestingness criteria” as shown in the Figure 3.

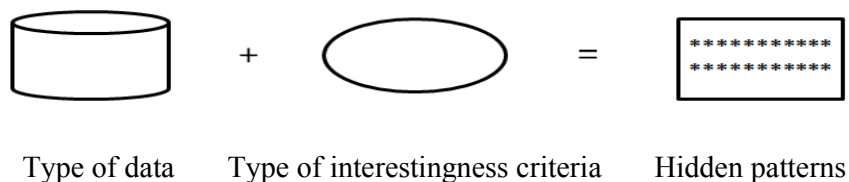


Figure 3. Data mining process for finding hidden patterns

Here we are mentioning the different types of data and interestingness criteria.

Type of data

1. Tabular
 - Relational
 - Multi-dimensional (e.g. Transaction data)
2. Spatial
 - x, y, z coordinates (e.g. Remote sensing data)
3. Tree (e.g. XML data)
4. Graphs (WWW., Bimolecular data)
5. Sequence (e.g. DNA, activity logs)
6. Text etc.

Type of interestingness

1. Frequency
2. Rarity
3. Correlation
4. Consistency
5. Repeating/periodicity
6. “Abnormal” behavior

1.3.3 Data Mining Techniques

Data mining techniques commonly involves four classes namely, clustering, classification, regression and association rule.^{43,44} Among the four, classification is most commonly used.

- (i) Classification, as the name suggests, data are organized into classes by using predetermined class labels. Classification process often employs supervised learning and is mostly used for predictive modeling.⁴⁵ Here the algorithms normally use a training set where all data points are already associated with known class labels. The classification algorithm learns from the training set and builds a model, also called a classifier. The model is then applied to predict the class labels for the unclassified data points in the testing data. Such classification uses mathematical techniques to construct binary decisions, “active” or “inactive” format so that data is split into two different classes according to its attributes as found in binary decision trees, neural networks, linear programming and statistics.
- (ii) Clustering is the task of discovering “similar” groups and structures in the data, without using the known structures. Cluster analysis takes ungrouped data and uses automatic techniques to put this data into groups. Clustering is unsupervised, and does not require a learning set.⁴⁶
- (iii) Regression attempts to find a function which models the data with the least error.
- (iv) Association rule learning searches for relationships between variables. It deals with large scale real time data sets so as to identify things that go together such as those generated each day by retail organizations.

The other techniques used in data mining are: Artificial Neural Networks, Genetic Algorithms, Rule Induction, Nearest Neighbor Method and Memory Based Reasoning, Logistic Regression, Discriminate Analysis and Decision Trees.^{34,47,48}

1.3.4 Data mining tools

Many data mining software products are available; WEKA (from the University of Waikato in New Zealand) is an open source tool with many machine algorithms for performing data mining tasks.^{49,50} This java based product is easily downloadable from www.cs.waikato.ac.nz/ml/weka/. There are many other products from Enterprise Miner by SAS and Intelligent Miner by IBM, CLEMENTINE by

SPSS etc.⁵¹ In our study we used WEKA for developing computational predictive models.

1.3.5 Data mining applications

Data mining methods are largely applied in the *in silico* drug design; here a model is derived that relates to a set of molecular descriptors to biological activity or efficacy or (ADMET) properties. Then this model is used to predict key property values of a new screening set for prioritizing and finding their structure–activity relations (SARs). One of the major application is in the field of cheminformatics spectrum i.e. virtual screening, computational tools are used to search large databases in finding and prioritizing new leads which has high probability on the target structure or in a whole cell screen.^{52,53,54} Other applications include, retail industry in fraud detection, banking firms, credit card, insurance etc.^{36,55} Also DM are frequently used for marketing and campaigning in mass media like TV, radio, newspaper and broadcasts. It is a type of knowledge discovery process.^{47,56}

1.4 Machine learning

Machine learning (ML) is an important branch of artificial intelligence that focuses on the theoretical, algorithmic and applicative aspects of learning from examples and enabling machines to learn. It helps mimic intelligent abilities of humans and enables a machine to learn from experiences, i.e. to build a classifier model and feed it with a set of example objects.^{57,58} A new object can then be assigned to one of a set of classes which are known beforehand. There are various algorithms available in the creation of the classifier model, such as decision tree, neural network, genetic algorithm and support vector machines. Learning in this context is in an inductive manner where knowledge is built from data, using this knowledge new data is predicted. Here also there are two major learning categories, supervised and unsupervised learning. The former tries to couple data against a known response while the later finds regularities and irregularities in data. If the response is discrete a classification algorithm is performed and if the response is continuous regression is used.

In this thesis we implemented supervised learning scheme. The goal of supervised learning is to approximate the function that maps the descriptors, of the examples with a chemical and biological response/activity. Mapping is constructed using training data and can be tested on a validation set or by using cross validation. A validation set is a part of the data set withheld during training of the model. The cross validation approach splits the dataset into n subsets and for each subset a model is built using the remaining portion of the data and tested on the subset. There are also other measures that can be algorithm specific, like out of bag error (OOB) for Random Forest (RF)⁵⁹ and the number of support vectors for Support Vector Machines (SVM).⁶⁰ Different machine learning methods have different ways of deriving these approximations. ML and computational intelligence are well-established in the drug design field for building predictive models for extremely complex biological and pharmacological responses for drug administration.⁶¹

1.4.1 Machine learning process

The ML process usually begins with the selection of a dataset, which is then divided into training set and test set. Where the former set consists of 80% of the data points and trains the system while the latter set about 20% of the data points is used to evaluate ability of the system in predicting the outputs.^{62,63} ML is performed by training and the capability to predict an output is called generalization. That is the system has to learn from the input data rather than just memorizing the input values. Here also the learning strategies are supervised and unsupervised. As discussed previously in supervised learning the machine is also given a set of target outputs and its task is to learn to generate the correct output for a newly given input. The output of the system is compared with the correct output and thus an error is obtained. Supervised learning methods try to minimize this error. Besides classification, supervised learning can be used for modeling or prediction. While in the case of unsupervised learning, clusters are detected within the data. Common tasks of unsupervised learning include data compression, clustering and outlier detection. Among the methods, Kohonen networks Self-Organizing Maps (SOMs) belong to a large group of methods called Artificial Neural Networks. Artificial

neural networks are techniques which process information in a way that is motivated by the functionality of a biological nervous system.⁶⁴ Some of the machine learning methods are given in Table 2.

Table 2. Different types of machine learning methods

Unsupervised	Supervised
Kohonen Networks	Decision Trees
Clustering	Partial Least Squares (PLS)
Principle Component Analysis (PCA)	Multiple Linear Regression (MLR)
	Genetic Algorithms (GA)

1.4.2 Classifiers

In this thesis, we adopted major ML algorithms like Decision Tree (DT), Naïve Bayes and Support Vector Machines, for the model build.

1.4.2.1 Decision Tree analysis

DT represents a supervised approach to classification. A DT is a simple structure consisting of one root, branches, nodes (places where branches are divided) and leaves. And the architecture looks like a flow chart diagram. An ordinary tree has non-terminal nodes where attributes are tested and terminal nodes reflect the decision outcome. Here the data is distributed into sub-groups in an iterative manner until all the data has been grouped accordingly in a desired condition. Finally the items in the sub-groups contain more common features. DT with a range of discrete (symbolic) class labels is called a classification tree, whereas DT with a range of continuous (numeric) value is called a regression tree. Classification and Regressing Tree (CART) is a well-known program, used in the designing of DT. DT make use of the IDE3, C4.5 and CART algorithms.^{40,65}

The tree nodes are represented as circles and branches as connecting nodes. A decision tree is usually drawn from top to bottom from the first node “root” or from left to right and expands as “root-branch-node”. The end node is called

“leaf”.⁶⁶ Such trees main advantage is that it does not take long training process and a lot of modeling time is saved for bigger datasets. The main benefit of decision trees over other classification techniques is that the resulting classification model can be easily interpreted. They not only point out which variables are important in classifying objects/observations, but also indicate that a particular object/observation belongs to a specific class when the built rules are satisfied. Some of the other advantages are they are easily understandable and can be classified into both numerical as well as categorical. But the output is always categorical as there are no prior assumptions about the nature of the data. DT model are capable in handling high dimensional data as well. But they do have some disadvantages, like they can't handle multiple output attributes as they are unstable. Also a slight variation in training data results in different attribute selection which in turn affects the descendent trees and can lead to wrong output prediction.

1.4.2.2 Support Vector Machines (SVM)

SVMs are supervised machine learning algorithms which are applied for binary property/activity prediction. The methods were developed by Vapnik and Cortes in the year 1995 that facilitated compound classification, ranking and regression-based property value prediction.^{67,68} Typically they are used to classify drugs from non drugs or biologically active specific molecules from non specific active molecules. This is achieved by projecting the compound libraries into higher dimensional feature vector space via a kernel function, thereby becoming linearly separable.^{69,70} The two classes of compounds are separable by a hyperplane as shown in Figure 4. In real there are an infinite number of hyperplanes. And the SVM chooses the hyperplane that maximizes the margin between the two classes. Thereby classifier error can be reduced while dealing with an external dataset. The hyperplanes that define such margins are called ‘support hyperplanes’ and the data points that lie on these hyperplanes are the ‘support vectors’. In the case of no separable classes, which are common, the soft-margin hyperplane are applicable, which maximizes the margin while keeping the number of misclassified samples to a minimal.

Apart from the pattern classification SVM, it can also handle nonlinear regression problems in which case they are known as Support Vector Regressors (SVRs). SVM classification uses kernel functions like radial basis (RBF), linear, polynomial and sigmoid. RBF is a local kernel function while the remaining three belongs to global kernels.^{71,72}

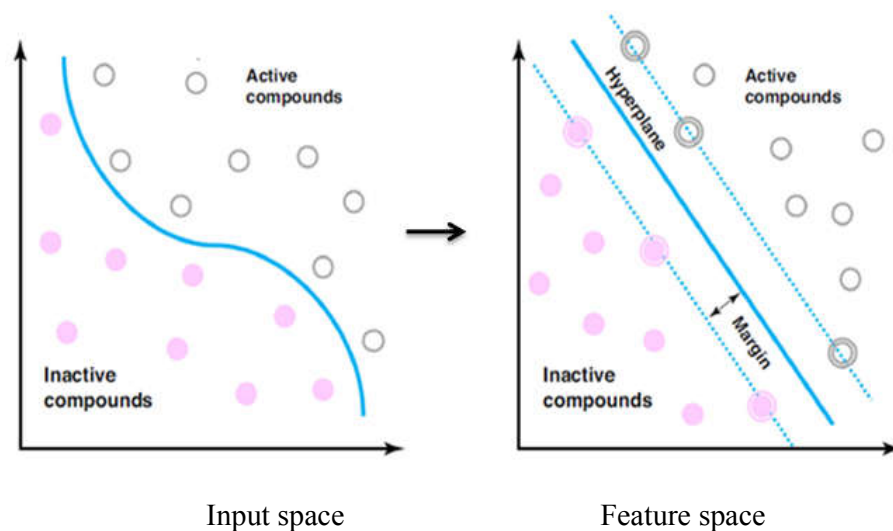


Figure 4. Illustration of the projection of active (empty gray points) and inactive (filled pink points) compounds into high-dimensional feature space that are not linearly separable in a low-dimensional input space. Left panel displays the projected compounds into high-dimensional feature space while in the right panel compounds are separated by the maximum-margin hyperplane and the data points lying on the dotted line are called 'support vectors'. Figure adapted from ref.⁵⁴

The main advantage of SVM model are they can produce nonlinear separating surfaces via nonlinear kernel functions such as Gaussian (RBF) or n^{th} order polynomials which have achieved very good performance of complex problems. Also SVMs are less affected by class imbalance ratio.⁷³ Due to their good generalization performance these models have become very popular with a wide range of applications, including document classification, image classification, bio-informatics, handwritten character recognition etc. One of the major drawbacks of these models is the memory and the computational complexity requirements for handling large datasets. The reason is that the separating surface is obtained by

solving a quadratic programming problem involving an $N \times N$ matrix, where N is the number of items in the dataset.⁷⁴

1.5 Validation

Chemoinformatics models are only useful if they are predictive no matter what classifier is used. Also it is not sufficient simply to fit known data, as a model must be able to generalize the unknown data. After which the model must be validated with an external dataset and not the same dataset that was used to create the classifier itself. One common and effective approach is *cross-validation*.^{75,76} In an n -fold cross-validation, the data is distributed, either randomly or in a stratified way; into n separate folds, with one fold being the initial test set. The remaining folds are the initial training set. Typically, fivefold or 10-fold cross validations are carried out. If relevant, a second fold is used for internal validation. The identities of the folds are then cyclically rearranged in such that every fold is the test fold once and hence each instance is predicted exactly once as shown in Figure 5.



Figure 5. A 10 fold cross validation

Various articles cited on ML algorithms performance shows that there is no best single method for solving all the problems. The relative abilities of methods depend on various factors like the size and distribution in chemical space of the dataset, nature of the chemical problem to be solved and the internal correlation of the descriptor set available. Thus validation of ML project plays an important role and robustness of an *in silico* experiment. The traditional way of training-test split of the dataset is a good validation strategy in an ML process, provided that the two sets share same regions of chemical space.^{2,77}

1.6 ML applications

In recent years, many ML improved algorithm models are in use in various drug design programs. The pharmaceutical companies have invested heavily in pre-clinical discovery pipeline. Previously about 40% of clinical trial failures (in 1980s and 1990s) were due to poor absorption, distribution, metabolism, excretion and toxicity (ADMET) properties. But now a day's most of the drug discovery processes are fully integrated into *in silico* - *in vitro* modeling and high-throughput *in vitro* screening for potency against the target interest. A greater usage of computational intelligence and machine learning methods are adopted. ADMET models with QSAR/QSPR models have proven to be very successful. Since 2010 the pre-clinical failures due to the ADMET was reduced from 40% to 10–14%.^{61,78}

1.6.1 Statistical performance matrices

Typically a classifier is evaluated by using a confusion matrix as illustrated in Figure 6. The column represents the Predicted Class, while the rows are the Actual Class. The confusion matrix in Table 3 describes the following parameters- **TN** is the number of negative examples correctly classified (True Negatives), **FP** is the number of negative examples incorrectly classified as positive (False Positives), **FN** is the number of positive examples incorrectly classified as negative (False Negatives) and **TP** is the number of positive examples correctly classified (True Positives).^{79,80,81,82} From the confusion matrix classifier accuracy can be determined. This provides an honest estimate of the true error rate, i.e. an indicator of how good the classifier is or the probability of it classifying new cases correctly.⁸³

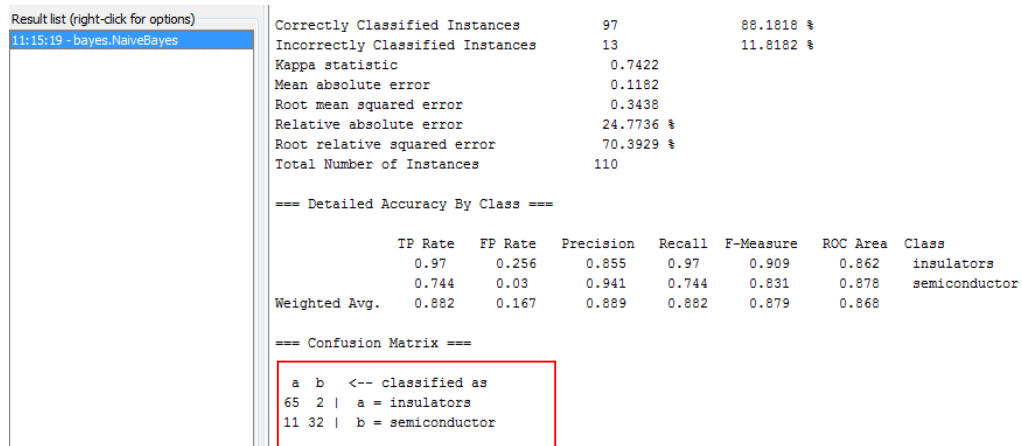


Figure 6. The confusion matrix obtained from the data mining software WEKA

Table 3. Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Various statistical binary classification performance measures^{63, 83, 84, 85, 86,87} used in the study is mentioned in Table 4.

Table 4. Various statistical evaluation parameters used as model robustness evaluators.

Statistical Parameter	Equation	Description
True Positive Rate (TPR)	$(TP)/(TP + FN)$	Proportion of actual positives which are predicted positives.
False Positive Rate (FPR)	$(FP)/(FP + TN)$	Ratio of predicted false actives to actual number of inactive.
Precision	$(TP)/(TP + FP)$	Proportion of predicted positive cases that are correctly real positives
Recall or Sensitivity	$(TP)/(TP + FN)$	Proportion of real positive cases that are correctly predicted positive.
Receiver Operating Characteristic (ROC) Curve	A graphical plot of TPR vs. FPR (for a binary classification system)	ROC plot is defined by FPR and TPR on X and Y axes respectively.
Accuracy	$(TP + TN)/(TP + FP + TN + FN)$	It indicates proximity of measurement of results to the true value and overall effectiveness of the classifier.
Balanced Classification Rate (BCR):	$(0.5 * (\text{Sensitivity} + \text{Specificity}))$	It gives a balanced accuracy for unbalanced datasets.
Matthews Correlation Coefficient (MCC)	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	MCC judges the performance of unbalanced datasets and its values ranges from -1 to +1.
F-measure	$(1 + \beta^2) * \text{recall} * \text{Precision} / (\beta^2 * \text{recall} + \text{precision})$	It is a popular evaluation metric for an imbalance problem. Coefficient β is usually set to 1.
Kappa	$P(A) - P(E) / 1 - P(E)$ The observed agreement among the predicted and the observed class is given by P(A) and the expected agreement by chance is given as P(E).	It includes measures of class accuracy within an overall measurement of classifier accuracy. Kappa value range from -1 to 1. Kappa value 1 means perfect agreement, -1 means perfect disagreement and 0 means that agreement is equal to chance.

1.7 Computational Chemistry

A mathematical description of Chemistry is now what is known as Theoretical Chemistry. This mathematical method which is sufficiently well

developed and that can be automated in a computer is called Computational Chemistry. As computing power has increased over the past 30 years, the field of Computational Chemistry has taken full advantage of this extra power in today's digital age.⁸⁸ More sophisticated algorithms, shorter time steps and more accurate molecular modeling have contributed to making Computational Chemistry a more accurate and reliable method.

The most commonly used constructs in Chemistry are models and approximations. A model is a description that is used to describe and predict some scientific results. So they are very useful in understanding a chemical phenomena without undergoing the complex mathematical calculations derived from a rigorous theory. A model does not describe the molecule as a whole but can describe some features like the strength and chemical bonding patterns as in the Lewis dot structure model. However, none of the Quantum Mechanics equations are used in applying this technique. QM describes electron behavior in a mathematical way and has never gone wrong.⁸⁹

QM equations have never been solved exactly other than the one electron system i.e. hydrogen atom and thus the entire field of Computational Chemistry are built around approximate solutions.⁹⁰ QM description of the hydrogen atom matches the observed spectrum as accurately as any experiment has ever done. One of the main advantages of Computational Chemistry is the modeling of a molecular system, this helps in ruling out 90% of possible compounds as being unsuitable for their intended use. This useful information will be beneficial to Synthesis Chemist. Also, there are some molecular properties and molecular bonding patterns that can be easily obtained from computational method rather from an experimental method. Thus, many experimental Chemists are performing computational studies of the compounds that have been examined in the laboratory.

Some of the major goals of Computational Chemistry are to create efficient mathematical approximations and computer programs so as to apply these programs to concrete physico-chemical systems and to calculate various molecular properties of interest.⁹¹ The properties (of a three dimensional structure) include interaction

energies, electronic charge distributions, dipole moments, vibrational (frequencies and intensities) UV-Visible, ESR, NMR spectra, molecular absolute and relative energies.⁹²

1.7.1 Molecular Mechanics

The fundamental methods of Computational Chemistry are Molecular Mechanics (MM), *Ab Initio*, Semiempirical, and Density Functional Methods (DFT).^{93,94} Molecular Mechanics – could also be Force Field Methods that is based on a molecular model consisting of collection of atoms (balls) and bonds (springs) held together.^{95,96} From this model the static properties of a molecule or a group of molecules, such as structure, energy or electrostatics are calculated. And for the dynamics properties like the time evolution of a molecular system, molecular dynamics is used. It is called as a “force field” because it is the force on an atom that is calculated from the change in energy between its current position and its changed position a small distance away. This is recognized as the negative derivative and the potential energy E with respect to the coordinate's r_{ij} as given in equation (1).

$$F = - \partial E / \partial r_i \quad (1)$$

MM methods are often used because of their fast algorithm and they are even applied on large molecules with many atoms like a steroid (e.g. cholesterol). The geometry optimization can be performed in a few seconds on a workstation.

This type of calculations does not deal with electrons, thus the input of the total charge and the spin of a molecule is not mandatory. The mathematical formulation of a typical molecular mechanics force field which is also called the Potential Energy Function (PEF), as shown in equation (2).^{6,97}

$$V_{PEF} = \sum V_{bonded} + \sum V_{non-bonded}$$
$$V_{PEF} = \sum V_{bonds} + \sum V_{angles} + \sum V_{torsions} + \sum V_{electrostatic} + \sum V_{van\ der\ Waals} \quad (2)$$

The PEF is the summation of many individual contributions in bonded properties like bonds, angles, torsions and non bonded properties like Van der Waals and electrostatic. V , responsible for intermolecular and intramolecular interactions.

Here are some of the force fields for small molecules:

1. Molecular Mechanics (MM2/MM3/MM4, TINKER)
2. Universal Force Field (UFF)
3. Assisted Model Building with Energy Refinement (AMBER)
4. Chemistry at HARvard Macromolecular Mechanics (CHARMM)
5. Groningen Molecular Simulation (GROMOS)
6. Optimized Potentials for Liquid Simulations (OPLS)

1.7.2 Semiempirical methods

Semiempirical calculations are based on Hartree–Fock (HF) calculation involving a Hamiltonian and a wave function.⁹⁵ Usually, certain pieces of information are approximated or are completely omitted and the omitted part of the calculation is parameterized. Parameterization is performed by fitting the results to experimental data or *Ab Initio* calculations. Unlike MM where electrons do not take part in the calculation, here only a minimal basis set is used and the core electrons are excluded from the calculation. The advantage of such calculations is that they are much faster than *ab initio* type calculations. And a disadvantage is that the results can be erratic and only few properties are predicted reliably. It is because the molecule that is being computed shares similar molecules in the database that is used to parameterize the method. This can ultimately end up with good results or else vice-versa. For example, the carbon atoms in cyclopropane and cubane do not share a normal bond angle as in the case of most of the other organic molecules. Thus, these molecules may not be predicted well unless they are included in the parameterization.

Semiempirical methods are often used to calculate geometry, energy (usually the heat of formation), dipole moments, heats of reaction and ionization potentials in the parameterization set. The most widely used methods in semiempirical are Austin Model 1 (AM1) and Parameterization method 3 (PM3). Both use nearly the same equations with an improved set of parameters for PM3. The availability of algorithms for calculating solvation effects makes these methods more popular. The PM3 methods are widely used for organic systems. It is more accurate than AM1 for hydrogen bond angles, but AM1 is more accurate for hydrogen bond energies.⁹⁸ AM1 generally predicts the heats of formation (ΔH_f) more accurately than other methods. Over all PM3 predicts energies and bond lengths more accurately than AM1. Depending on the nature of the system and information desired, either AM1 or PM3 will often give the most accurate results obtainable for organic molecules with semiempirical methods. The other force fields are HUCKEL, Complete Neglect Of Differential Overlap (CNDO), Modified Intermediate Neglect Of Differential Overlap MINDO (inorganic molecules), Zerner's INDO method etc.^{99,100}

1.7.3 *Ab initio* Methods

The term *Ab Initio* means “from the beginning”, is a computational calculation that are derived directly from Quantum Chemistry principles without inclusion of any experimental data.⁹³ The most common type of *ab initio* calculation is Hartree-Fock (HF), which is based on the central field approximation. According to this approximation the Coulombic electron-electron repulsion is taken into account as an average effect of the repulsion, but not the explicit repulsion interaction. Because of this approximation, the energies from HF (1 Hartree = 27.2116 eV) calculations are always greater than the exact energy and tend to a limiting value called the Hartree-Fock limit as the basis set is improved. One of the advantages of this method is that it breaks the many electron Schrödinger equation into many simpler one electron equations. Each one electron equation is solved to yield a single electron wave function, called an orbital and its energy is called an orbital energy. The orbital describes the behavior of an electron in the net field of all the other electrons.

The second approximation in HF calculations involves defining the wave function by means of mathematical function which is exactly known for one electron system. The wave function is formed from linear combinations of atomic orbital's or, stated more correctly, from linear combinations of basis functions usually abbreviated as STO-3G or 6-311++g** (Slater Type Orbitals) that describes the shape of an orbital in an atom. Although semiempirical calculations uses a predefined basis set. But for *ab initio* calculations a basis set must be specified which is the major challenge in determining the accuracy of results. In general, *ab initio* calculations can yield qualitative results provided that all the approximations are made sufficiently for smaller molecules. The main drawback of this method is that it is expensive, uses enormous computational space, memory and CPU time.

1.7.4 Density Functional Theory

Density Functional Theory (DFT) ¹⁰¹ is a Quantum Mechanical theory that is used to determine the energy of a molecule on determination of the *electron density* instead of wave function. DFT is a general purpose computational method, and can be applied to most systems involving metals. Among the two popular methods, B3LYP is often used for reaction calculations, while MPW1K is used for the determination of kinetic problems. The DFT method B3LYP/6-31G (d) is often considered as a standard model for various chemical applications.¹⁰² The main advantage of DFT methods is increased computational accuracy with a reduced amount of computing time. Even though DFT is an advanced computational technique in our study we employed Cheminformatics based method and less priority was given to DFT methods.

1.8 Physical Models

The prediction of physicochemical parameters like logP, pKa, logD, aqueous solubility are both related to drug discovery and environmental studies involving surfactants, wetting agents etc.¹⁰³ The modeling of these properties is best facilitated by obtaining large, structurally diverse, high-quality datasets. The aggregation and curation of such datasets can be very exacting in terms of extraction of the data from the literature and virtual databases for the model built. Creation of a successful

predictive model is a tedious and time consuming process. This process includes data acquisition and preparation, generation of molecular descriptors, application of appropriate machine learning classifiers, evaluation of statistical measurements and assessment of model applicability. One of the most difficult steps is the collection of high quality data (experimental) which is manually extracted from scientific literature and other experimental procedure and the model built upon these data results in a better predictive model. There are various online chemical databases like DrugBank,^{19,104} ChemSpider,¹⁵ ChemExper¹⁰⁵ and PubChem where the molecules are stored. And modeling facilities are provided by other online tools like VCCLAB,¹⁰⁶ ChemBench,¹⁰⁷ QSAR DataBank,^{108,109} and finally a predictive model is built by putting together the online resources and tools. Some of the computational models like melting point, boiling point, water solubility, IC50 etc are available in the OCHEM website.¹¹⁰ Hence, the whole process of training and modeling using a ML method of choice is tedious and iterative. These computational models could significantly reduce the amount of experimental measurements taken for the compound synthesis and related properties. And this helps in screening of a large number of compounds, especially against a particular physicochemical property which might have not been synthesized yet.¹¹¹

Many studies, reported that the life cycle of predictive models ends up with only a publication. More than 50 models alone were published for logP and lipophilicity etc and practically they are of no use, if not taken forward.^{112, 113} Very rarely these models end up becoming a software tool or as virtual online servers. But nowadays the advancement of informatics, computer power, operating system and internet has reduced this problem. For instance the various biological web servers are provided in the website.¹¹⁴ Although attempts to reproduce published models are not always successful. The reason could be many factors like the unavailability of the initial dataset and variation of molecular descriptors calculated (depending upon the software). But models built on memory-based approaches like Bayesian models, Support Vector Machines and neural networks could be reproduced correctly provided with initial dataset.

1.9 Organic Semiconductors

Organic semiconductors are compounds mostly made up of carbon and hydrogen atoms that share some properties with inorganic semiconductors, like silicon or germanium. Apart from carbon and hydrogen atoms they also contain few heteroatoms such as sulfur, oxygen, and nitrogen.^{115,116} The conducting property is typically associated with absorption and emission of light in the visible spectral range and degree of conductivity that is sufficient for the operation of classical semiconductor devices such as light-emitting diodes (LEDs), solar cells, and field-effect-transistors (FETs).^{117, 118} In the view of their conductivity, electro conductive polymers are called as organic “metals”.¹¹⁹ They are easily modifiable, soluble, more mechanically resistant, easily molded into different shapes and thicknesses which makes them different from traditional conductors.

The semiconductor properties exhibited by these molecules strongly differ from “semiconducting” nature between inorganic and organic materials. This can be explained in terms of the valence and conduction bands where charge carriers (i.e. electrons and holes) “travel” through these bands resulting in current conduction in inorganic semiconductors. The same for organic semiconductor are usually named HOMO (Highest Occupied Molecular Orbital) and LUMO (Lowest Unoccupied Molecular Orbital) since the structure is basically related to the molecular properties.¹²⁰ Depending on the molecular structure, the separation between HOMO and LUMO bands are 2, 5 eV or more, which is significantly larger than inorganic semiconductors such as 1.1 for silicon (Si), 0.67 for germanium (Ge), and 1.4 for gallium arsenide (GaAs).

The discovery of organic semiconductors in the late 1970s proved that their high conductivities were obtained from the π -conjugated polymers. That is, hydrocarbon chains with alternating single and double bonds. This property of connected carbon chain atoms, either as closed benzene rings (oligomers) or expanded chains (usually referred to as “backbone”) in polymers are responsible for electric current conduction. Here charge conduction is due to π -bond formation by p-orbitals of carbon atoms of trigonal hybridization between adjacent carbon atoms in

a conjugated system. The σ bond being more stable than the π bond, the least energetic electron excitation of conjugated molecules observed is π - π^* type transition as shown in Figure 7, where the energy gap is usually between 1.5 and 3 eV in the UV-Visible region.¹²¹ The energy gap may be controlled by the degree of conjugation of the individual systems in the macromolecule, which opens various possibilities for the modification of optoelectronic properties of organic semiconductors.^{116,122,123} Some examples of organic semiconductor prototypes are fullerene, poly (p-phenylene vinylene) (PPV), polyfluorene (PFO), pentacene, poly(3-alkylthiophene) (P3AT) etc.

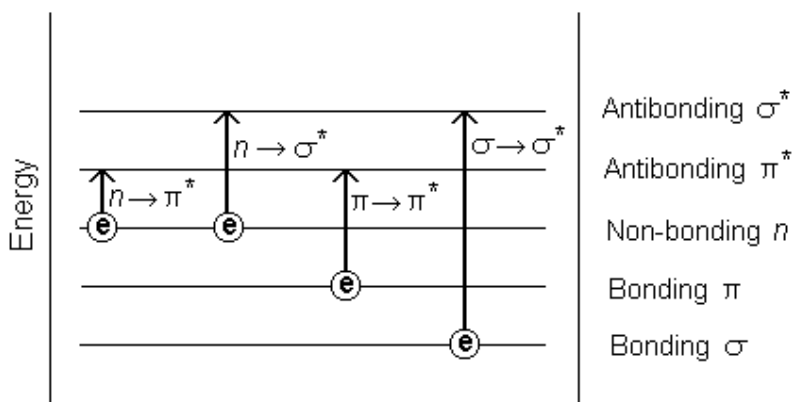


Figure 7. The possible four types of electronic transitions of π , σ , and n electrons are displayed. They are: $\sigma \rightarrow \sigma^*$ Transitions (high energy), $n \rightarrow \sigma^*$ Transitions, $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ Transitions (here less energy is required but for transitions unsaturated group is needed in the molecule to provide the π electrons.). Electronic transitions figure is taken from ref.¹²²

1.9.1 Types of Organic Semiconductors

Based on their charge carriers, organic semiconductors are of two types: p-type (holes as major charge carriers) and n-type (electrons as major charge carriers). Usually charge transfer is facilitated by π -conjugation in oligomers or polymers, where the $\pi - \pi$ stacking direction ideally is along the direction of current flow. This is attained by assembling of semiconductor molecules in a certain orientation upon either vapour or solution deposition. It is also important that the semiconductor thin

film has large, densely packed and well-interconnected grains. The arrangement of organic semiconductor pentacene oriented close to normal to the dielectric surface with the size order of at least a few micrometers is shown in Figure 8a. In Figure 8b the π -conjugated plane of poly (3-hexylthiophene) edge-on orientation on the surface is displayed.^{124,125}

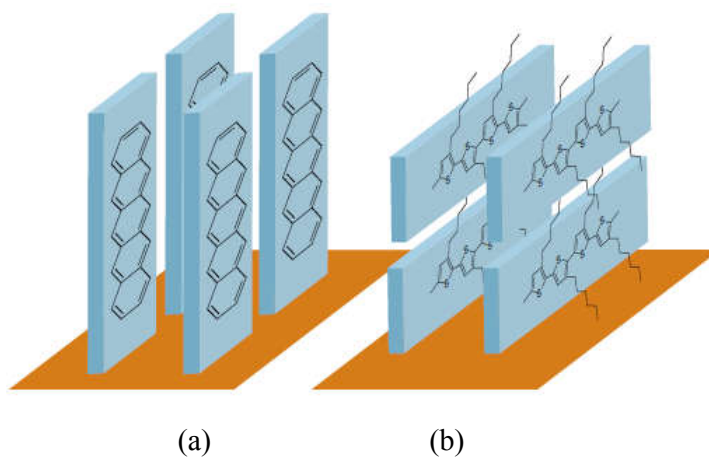


Figure 8. Molecular orientation of organic semiconductors (a) Pentacene (b) poly (3-hexylthiophene) (P3HT). The figure was adapted from ref.¹²⁵

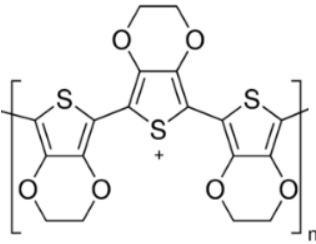
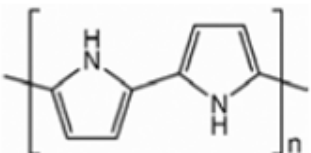
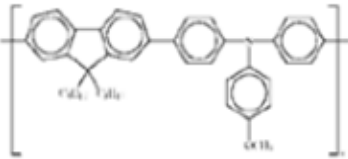
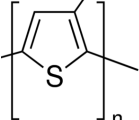
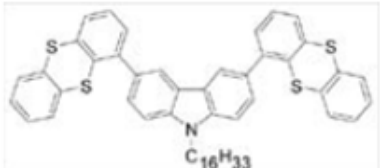
The other types of organic semiconductors include; (i) *Amorphous molecular films* where organic molecules are deposited as thin amorphous film through evaporation or spin-coating that is mostly employed for device applications such as LEDs, xerography etc. (ii) *Molecular crystals* such as naphthalene or anthracenes are held together by van der Waals interactions in the form of a crystal structure. These molecular crystals have high charge mobilities in comparison to those in noncrystalline organic materials and are widely used in transistor applications. (iii) *Polymer films* are arranged as a chain of covalently coupled molecular repeat units that are processed by different deposition techniques like simple spin-coating, ink-jet deposition, or industrial reel-to-reel coating.¹¹⁵

1.9.2 Applications of Organic Semiconductors

One of the main advantages of these organic materials is their simplicity of technological implementation in designing new semiconductor products having

substantial economical effects. Organic (opto) electronic materials like anthracene¹²⁶ have received considerable attention due to their applications in thin-film-transistors, light-emitting diodes, solar cells, sensors, photorefractive devices and various displays like computer displays, palmtops, mobile phones, iPads, iPhones, TV sets etc.¹²⁷ The other applications are mentioned in the following Table 5.¹¹⁹

Table 5. Examples of some organic semiconductors and its applications are mentioned

Organic Semiconductors	Application
<p data-bbox="418 646 829 682">Poly(3,4-ethylenedioxythiophene)</p> 	<p data-bbox="1105 772 1187 808">OLED</p>
<p data-bbox="548 932 695 968">Polypyrrole</p> 	<p data-bbox="1101 1014 1192 1050">Battery</p>
<p data-bbox="321 1129 924 1165">Poly(p-methoxytriphenylamine-9,9-octylfluorene)</p> 	<p data-bbox="1084 1220 1208 1255">Transistor</p>
<p data-bbox="483 1346 764 1381">Poly(3-octylthiophene)</p> <p data-bbox="570 1381 760 1417">$\text{CH}_2(\text{CH}_2)_6\text{CH}_3$</p> 	<p data-bbox="1089 1430 1208 1465">Solar cell</p>
<p data-bbox="358 1543 886 1579">3,6-bis(thianthrene)-N-hexadecyl carbazole</p>  <p data-bbox="613 1724 683 1751">$\text{C}_{16}\text{H}_{33}$</p>	<p data-bbox="1084 1633 1208 1669">Biosensor</p>

1.10 Scope of present investigation

Literature survey shows that for the last few decades, computational models and Cheminformatics have enhanced Chemistry from the traditional way of approach in solving chemical problems. In the informatics era we have identified suitable descriptors for building predictive models to predict the unknown property values based on the known values from the training dataset. We concentrated on methods for supervised learning, particularly algorithms involving Random Forest, Support Vector Machine and Naïve Bayes classifiers. The chemical databases provided with millions of molecules with basic molecular properties made us to take this data in order to transform the data into information, then information into knowledge so as to take an informed decision (better decision).

Virtual screening, the analogy of the High-Throughput Screening are widely applied for making better decisions faster in the area of drug lead identification and optimization. The survey revealed that such virtual screening can be adopted in the Material Chemistry by making use of the data mining and machine learning techniques. In the current study many physical predictive models were developed that can be used to screen large datasets to predict the physical property like semiconductority.

CHAPTER 2

MATERIALS AND METHODS

In this chapter a concise review of the softwares, online servers, statistical tools used and the methods adopted for the organic semiconductor predictive model build are presented. Detailed descriptions are provided in suitable contexts.

2. Packages

2.1 e-dragon

We used the descriptor generator software DRAGON that is provided with a variety of molecular descriptors derived from different molecular representations. This was developed by Milano Chemometrics and QSAR Research Group in 1994. It allows the calculation of 1,664 molecular descriptors that works on both operating systems Windows and Linux. The different versions available are stand-alone mode, background mode (dragon X for Linux). It is a user friendly software that performs descriptor calculations within a few steps that involves the loading of the molecular files, selection of the descriptors, calculation of the descriptors and saving the output of the calculated descriptor.¹²⁸ Molecular descriptor calculations are performed for 3D optimized structures with both hydrogen and hydrogen depleted molecules. And it doesn't perform geometry optimization, nor transform topological structures into the corresponding 3D geometrical structures. And for SMILES notations hydrogen atoms are automatically added.

In our study the electronic version E-DRAGON supported by Virtual Computational Chemistry Laboratory (VCC-LAB) was used for the molecular descriptor calculations.¹⁰⁶ In the background, Java applet is employed in order to calculate the molecular descriptors directly within an HTML page. One of the main limitations that the server faces is that, the maximum number of molecules per batch is 149 with a maximum number of 150 atoms per molecule. E-DRAGON can load three molecular file formats (SYBYL Mol2, MDL sdf and SMILES).

The software provides more than 1,600 molecular descriptors that are divided into 20 logical blocks as shown in Figure 9. The calculations involve simplest atom type, functional group and fragment counts to several topological and geometrical descriptors. It can calculate molecular properties such as H-donors, H-acceptors, number of rotatable bonds, logP, molar refractivity, topological surface area (TPSA), Lipinski's alert "the rule of 5" and some drug-like indices.

Figure 9. Home page of the e-dragon software from VCC lab web server

The 20 different logical blocks depicted in Figure 9 is summarized as:

1. Block 1 is Constitutional descriptors. It is the most simple and commonly used 0D descriptors. It includes number of atoms, bonds, rings, rotatable bonds and specific atom types.
2. Blocks 2 – 10 are topological and topographic descriptors. Both are based on the graph representation of a molecule. The numerical values obtained for

topological descriptors are obtained from the molecular topology by the application of algebraic operators to matrices representing molecular graphs. While Topographic descriptor uses the geometric distances between atoms instead of the topological distances.

3. The blocks 11 – 16 descriptors are derived from 3D structure of a molecule and the block 20 include literature models like Moriguchi logP, Ghose-Crippen logP and Lipinski rule-of-five.
4. The block 17 includes enumerative descriptors that include 154 functional groups counts. They are based on the counting of chemical functional groups. Some of the examples include number of total primary C (sp³), number of total secondary C (sp³), number of total tertiary C (sp³), number of ketones, number of esters etc. And Ghose-Crippen atom-centred fragments belongs to 18th block that are based on the counting of 120 atom-centered fragments.
5. Block 19 includes fourteen charge descriptors like q_pmax (maximum positive charge), q_nmax (maximum negative charge), Q_{pos} (total positive charge), Q_{neg} (total negative charge), TE1 (topological electronic descriptor), TE2 (topological electronic descriptor bond restricted) etc. Here the charges are estimated by quantum molecular method.

2.2 MacroModel

MacroModel¹²⁹ is one of the applications of Schrodinger software and Maestro serves as the graphical user interface for MacroModel. It is a force-field-based molecular modeling program that aids in understanding the chemical structure, energetics, and dynamics. Though MacroModel energy calculation is a classical based molecular mechanics, its implementations are slightly different from the authentic force fields in various ways. From the original force field name, they are distinguished by adding a “*” to the end of the force field name like **MM2***, **MM3***, **AMBER***, **OPLS***, **OPLS_2001**, **OPLS_2005**, **AMBER94**, **MMFF**. Numerous minimization methods are available as shown in Figure 10 that enables

geometry optimizations for a broad selection of structural classes. It also allows various methods for conformational searching for systems ranging from small molecules to macromolecules like proteins. Other features like molecular dynamics simulations, free energy perturbation simulations, pure and mixed methods for ensemble sampling are well performed by MacroModel. MacroModel runs calculations as independent tasks and consequently does not tie up Maestro during lengthy computations.

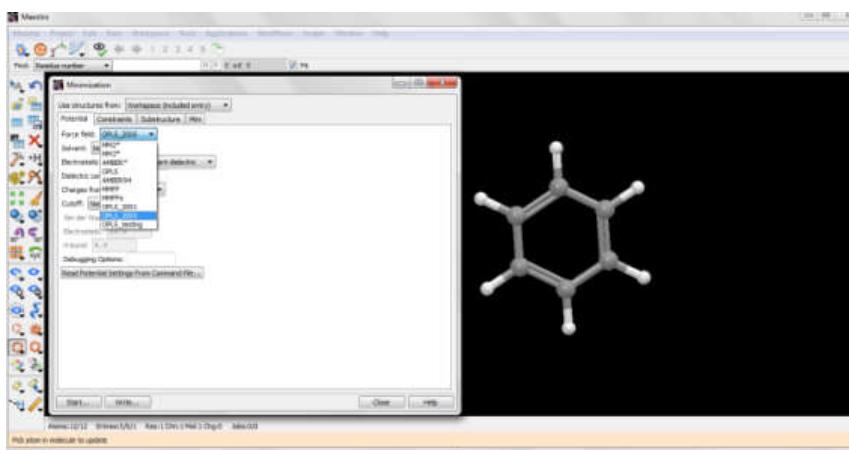


Figure 10. GUI of MacroModel from Schrodinger suite with various geometry optimization methods is displayed.

2.3 HyperChem

HyperChem^{130,131} provides a molecular modeling environment where molecules are drawn, edited and is visualized in both 2D and 3D. It can also perform quantum chemical calculations, molecular mechanics and dynamic simulations. Figure 11 displays the GUI of the molecular modeling environment.

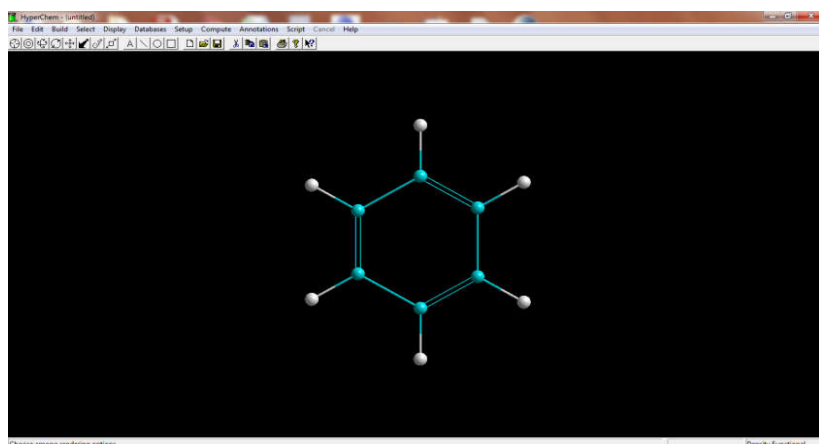


Figure 11. HyperChem workspace where a benzene molecule is modeled in ball and stick.

2.3.1 Structure input and manipulation included in HyperChem

1. Select, rotate, translate, and resize molecular structures.
2. Molecules can be displayed in various forms like ball and stick, fused CPK spheres, sticks, ball and cylinder or tubes.
3. Specific bond lengths, bond angles, torsion angles or the bonding geometry about a selected atom can be measured.
4. Peptides and nucleic acids can be built from amino acid and nucleotide residue libraries.
5. Large molecules can be generated as well as can be mutated.
6. A periodic box of pre-equilibrated water molecules for aqueous solvation studies can be performed i.e. periodic boundary conditions with/without solvent systems.
7. Import structures from standard file formats: Brookhaven PDB, ChemDraw CHM, MOPAC Z-matrix, MDL MOL, ISIS Sketch and Tripos MOL2 files.

2.3.2 Types of calculations performed by HyperChem software

1. Single point energy calculations determine properties for a given fixed geometry can be calculated.
2. Geometry optimization calculations can be performed as the software is provided with five minimization algorithms (Steepest Descent, Fletcher-Reeves, Polak-Ribiere, Eigenvector following and Conjugate Directions).
3. Vibrational spectrum can be generated and each vibrational motion of the molecule can be animated.
4. Molecular dynamics simulations, Langevin dynamics simulations, Metropolis Monte Carlo simulations and simulated annealing can be performed.
5. Computational methods like **Density Functional Theory (DFT)**, *ab initio* Quantum Mechanics, Semiempirical Quantum Mechanics and Molecular Mechanics are included.

2.4 Open Babel

The Open Babel graphical user interface (GUI)¹³² is a program that is used to convert molecules from one file format to another and is available as a cross-platform on Windows, Linux and MacOSX. Although the GUI presents many options, the basic operation is straightforward; the type of the input file is selected from the dropdown list and through “*Convert*” button desired output is selected. As an example Benzene in sdf file format is converted to mol file as shown in Figure 12.

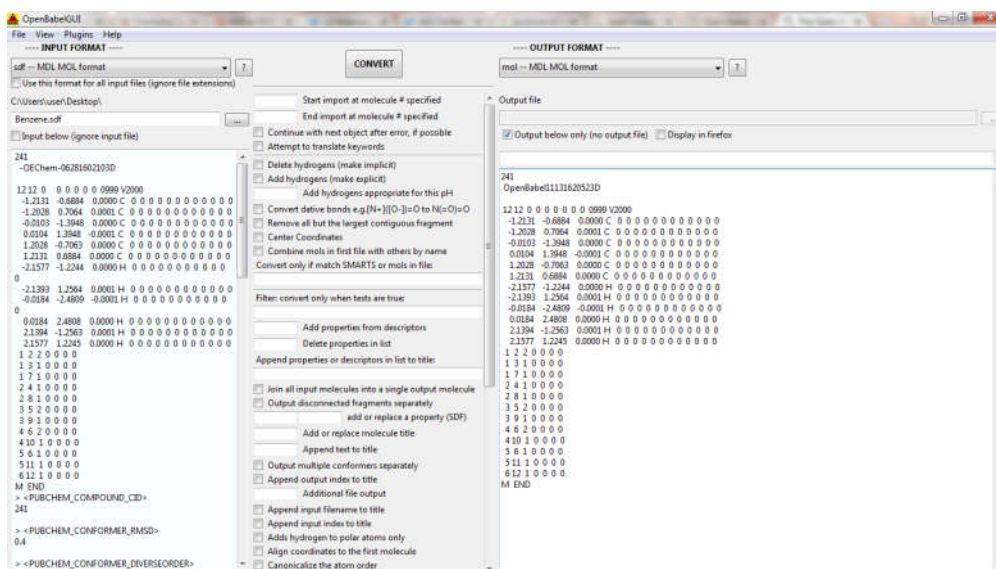


Figure 12. Open Babel GUI of version 2.3.2. Here a benzene molecule in sdf format is converted into benzene in mol and the message window below the button gives contents and the number of molecules of the output file is displayed. If, the output format allows multiple molecules by default, all the molecules in the input are converted.

2.5 WEKA

The Waikato Environment for Knowledge Analysis (WEKA)^{50,133,134} is an open source software that performs many machine learning and data mining algorithms. It is a Java based program that facilitates the availability of data mining tools regardless of the computer platform and it's freely available from the internet. The input data in specific file format is fed into a variety of learning schemes known as “classifiers” as they induce a rule set or decision tree that models the data. The data is pre-processed, model is developed and performance of the classifier is analyzed. Here the algorithms can either be applied directly to a dataset or called from a Java code (currently not used). The data mining algorithms for regression, classification, clustering, association rule mining and attribute selection are included in the WEKA workbench. Since its inception in 1992, WEKA is recognized as a landmark system in data mining and machine learning methods. It is widely accepted within academia and business circles. It has become one of the most used

open software tool for data mining research. In our study we used WEKA 3.6.2 for the ML model construct.

2.5.1 WEKA User Interfaces

WEKA is provided with several graphical user interfaces (GUI). The main graphical user interface the “Explorer” is shown in Figure 13 (a) WEKA GUI chooser; the other applications include Experimenter, Knowledge Flow and Simple CLI. The panel-based interface corresponds to different data mining tasks like the first panel called “Preprocess” panel. It enables to load and transform data using WEKA’s data preprocessing tools, called “filters” as shown in Figure 13 (b). The input data is loaded from various sources including files, URLs and databases provided with supported file formats Attribute-Relation File Format (ARFF) format, Comma-separated values (CSV), a Library for Support Vector Machines (LibSVM’s format) and C4.5’s format.



Figure 13 (a). WEKA GUI chooser is displayed with different applications involving “explorer”, “experimenter”, “knowledge flow” and “simple CLI” where various data mining tasks are performed.

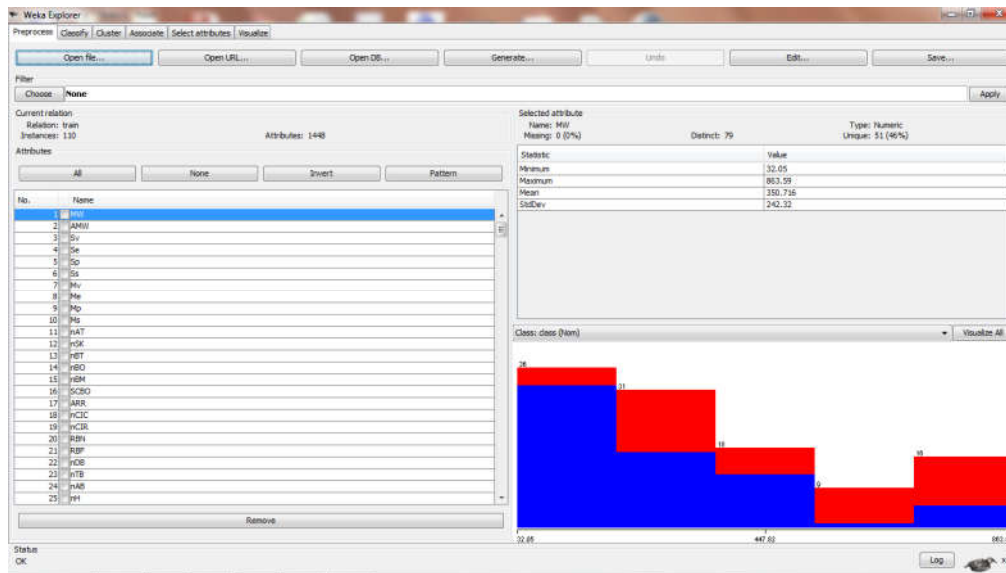


Figure 13 (b). WEKA “preprocess” panel

The second panel “Classify” in the Explorer gives access to various classification and regression algorithms for the model generation. The predictive performance is estimated by running a cross-validation for the pre processed dataset on the selected learning algorithm. The panel also provides access to graphical representations of models, e.g. decision trees along with other visualization modes like scatter plots, ROC curves, threshold curves etc. Buffer results and models can be saved. Models saved can be reloaded for various other tasks in this panel.

Along with classification algorithms, WEKA also supports clustering and association rule mining via the third and fourth panel in the explorer application. The “Cluster” panel allows running a clustering algorithm on the data loaded in the Preprocess panel. The other panels include associate and select attributes. The last panel in the Explorer, called “Visualize”, provides a color-coded scatter plot matrix, along with the option of drilling down by selecting individual plots in this matrix and selecting portions of the data to visualize. Since the Explorer employs a batch-based data processing: training data is loaded into memory in its entirety and then processed. This may not be suitable for large datasets. However, WEKA does have an incremental nature of an algorithm for the model building process called

“Knowledge Flow”. Following which datasets performs a series of tasks as nodes that can be loaded and each individual instance is preprocessed before feeding them into appropriate incremental learning algorithms. Once the nodes are interconnected and configured, it can be saved for later re-use. Later it can be evaluated and visualized.

The third main graphical user interface in WEKA is the “Experimenter” (see Figure 13) which is sparsely used in our study. Compared to WEKA’s other user interfaces, the Experimenter is perhaps used less frequently by data mining practitioners.

2.6 Canvas: Maximum Common Substructure (MCS)

In our study Cheminformatics package Canvas of version 1.3 was used for the Maximum Common Substructure.^{135,136} This package provided with a range of applications for structural and data analysis, including fingerprints, molecular calculations, similarity searching, substructure searching, clustering, building regression and other predictive models based on properties/fingerprints of the structure based classification models.

All the job status is colour-coded and finished jobs are incorporated manually (by clicking incorporate) into the spreadsheet. The Canvas GUI also provides chemical structure storage and organization, data analysis and visualization and access to various other applications as shown in Figure 14. Canvas is supported by different file formats;

1. Maestro, compressed or uncompressed (.mae, .mae.gz, .maegz)
2. SD file, compressed or uncompressed (.sd, .sdf, .sd.gz, .sdf.gz)
3. CSV file with SMILES strings and properties (.csv)
4. SMILES file with SMILES strings and optional titles (.smi)
5. Canvas fingerprint file (.fp)

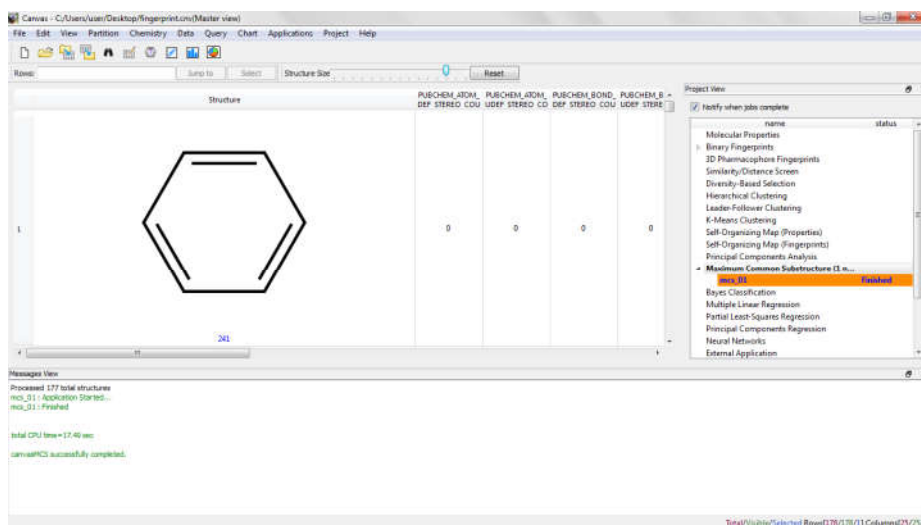


Figure 14. A Canvas project panel loaded with a benzene molecule is displayed and the highlighted application is “Maximum Common Substructure”.

The following are the some of the applications performed under Canvas.

1. Molecular Properties

A wide range of molecular properties and descriptors based on the 2D structure can be calculated. There are four classes of properties- Physicochemical Descriptors (Molecular weight, Polar surface area, Rotatable bonds, AlogP, Hydrogen bond acceptors, Hydrogen bond donors etc.), Topological Descriptors, LigFilter Descriptors and QikProp Descriptors.

2. Fingerprints

Canvas can calculate a variety of fingerprints from the 2D structure, (hashed fingerprints and structural keys) and 3-point or 4-point pharmacophore models.

3. Similarity, Dissimilarity and Clustering

Molecular similarity or dissimilarity is measured by a distance between molecules in the space of a property set. The similarity or distance is defined by a metric. Similarity or distances are used for clustering and screening of structures, or selection of diverse structures.

4. Clustering Structures by Similarity or Distance

Canvas provides three methods for clustering structures: hierarchical clustering, leader follower clustering and k-means clustering. Leader-follower and k-means clustering are used on a much larger data set than hierarchical clustering.

2.7 Self-Organizing Maps

Kohonen self-organizing maps is an artificial neural network that is used to project high dimensional data based on molecular property/fingerprint that is visualized in a 2D plane. Self-organizing maps is mentioned in detail in the Part II section of thesis.

2.8 Maximum Common Substructure

The program searches maximum common substructure for a given dataset and the calculations are performed from the Maximum Common Substructure dialog box from the Applications menu. MCS does not encompass all the structures in the set but a specified minimum to maximum number of a substructure must match. And the results include the number of substructures in each group where the substructures are shown in SMARTS format.

2.9 Machine learning classifiers

The classifiers used in this study were also part of the WEKA environment successfully used in different fields as a Machine Learning engine. Four different machine learning classifiers have been tested as possible candidates to predict the organic semiconductivity property. The chosen classifiers cover a range of popular modes of classification: Decision Trees (J48 and Random Forest), Naïve Bayes and Support Vector Machines have been selected for the study. The four classifiers were implemented with their default parameter values.

2.10 Sampling methods

A well balanced dataset is important for building good prediction models. Else it would be a challenge for conventional learning algorithms for building good

predictive models.¹³⁷ Observations from various articles have shown that the natural distribution is often not the best distribution for learning a classifier.^{79,138} To overcome the imbalanced problem, various sampling strategies have been used like under-sampling where some data from the majority class is eliminated and over-sampling where data are duplicated or generated artificially. Both methodologies have merits and demerits. One of the advantages of over-sampling technique is that there is no information loss from the original training set as both majority and minority classes are kept. However, oversampling technique increases the size of the training dataset (over fitting) that leads to a longer training time. And if the time taken to resample is not considered, under-sampling performs better than over-sampling in terms of time and memory complexity.¹³⁹

Sampling methods can be performed randomly. Random over-sampling where data points from the minority class are chosen randomly then, they are duplicated and added to the new training set. An alternative to random over-sampling is random under-sampling, here the number of samples in the majority class is reduced randomly from the original dataset to balance the class distribution.¹⁴⁰ Studies show that random under-sampling yields better minority prediction than random over-sampling. Also, the other studies say that random under sampling method can potentially remove certain important samples that may affect the accuracy of the model performance and random oversampling can lead to over fitting.

Mining from imbalanced datasets is a great problem in perspective of algorithmic and performance of the classifier. Significant loss of performance is mainly due to the skewed class distribution, given by the imbalance ratio (*IR*), defined as the ratio of the number of instances in the majority class to the number of examples in the minority class.^{141,142} Not choosing the right distribution of class while developing a model will automatically introduce bias towards the uninterested class prediction. Classifier evaluation based on the imbalanced datasets will not be an appropriate measure for justifying the accuracy of a predictive model.

2.11 Virtual screening

The process of virtual high-throughput screen involves reducing large chemical libraries by means of theoretical techniques into smaller sets of promising

lead compounds for experimental Chemists to follow up on. Recently the size of the chemical space is estimated to be $> 10^{60}$ molecules that make any kind of rational search challenging. The context of the calculations and chemical space differ between inorganic materials, organic materials and organic pharmaceutical chemistry. Their high-throughput screens are displayed in Figure 15 as they share a common underlying philosophy in these entire areas as mentioned below.¹⁴³

2.11.1 Four philosophies of High-Throughput Virtual Screening

1. Timescale is important.
2. Automated techniques are required; high-throughput approaches require automation.
3. Data-driven discovery; the data and trends in the data both are important.
4. Computational Funnel; allows only promising molecules that are passed through various computational filters as shown in Figure 16.

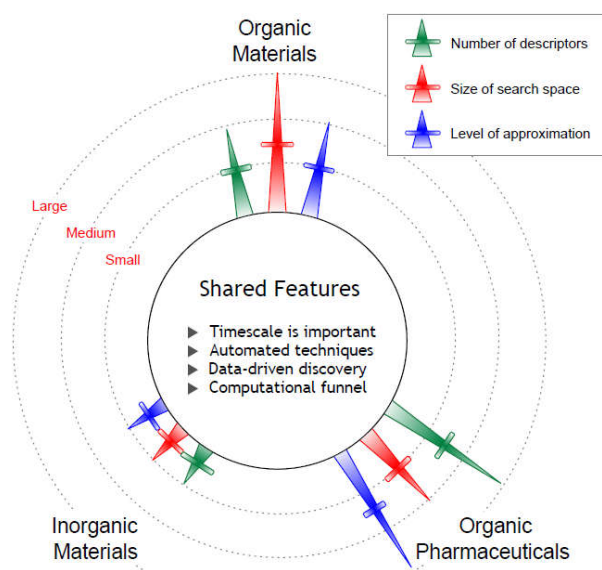


Figure 15. Organic material chemical space involving organic materials, inorganic materials and organic pharmaceuticals is displayed. The number of descriptors, size of search space and level of approximation is least for inorganic materials while organic materials possess a large chemical space. And for organic pharmaceutical both number of descriptors and level of approximation are high with a lowered chemical space. The image is taken from ref.¹⁴³

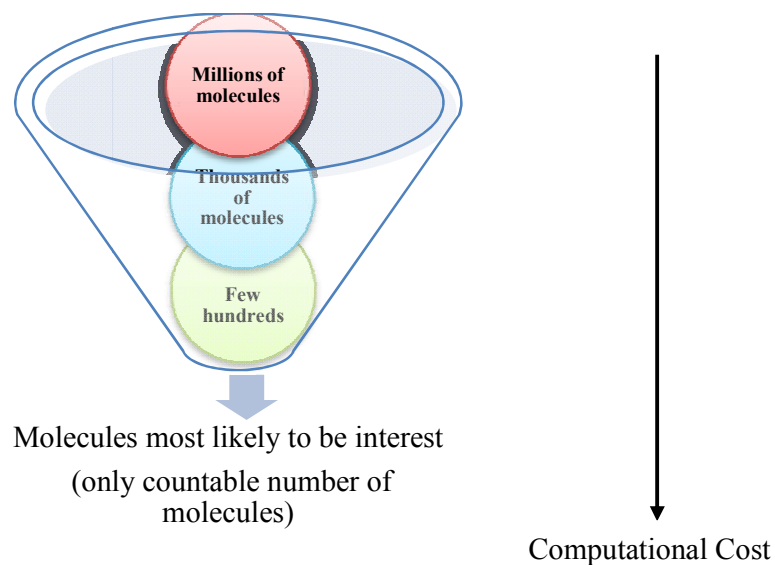
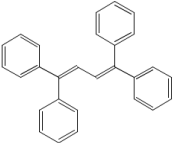
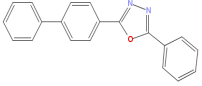
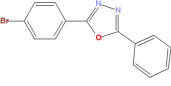
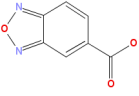
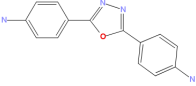
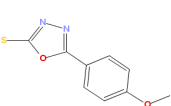
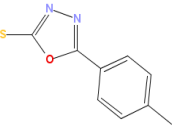
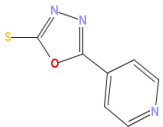
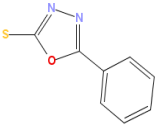
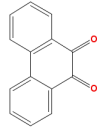
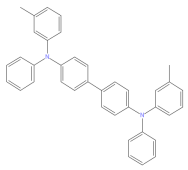
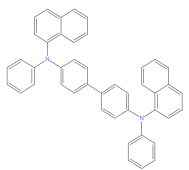
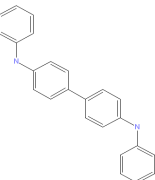
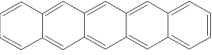
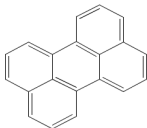
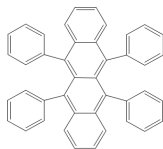
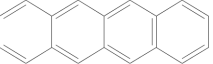
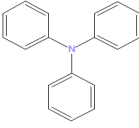
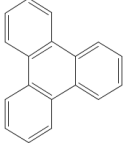
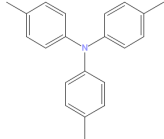
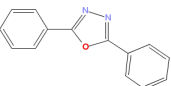
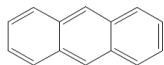
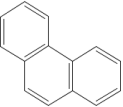
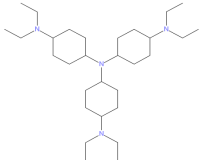


Figure 16. A computational funnel scheme eliminates many molecules that are of less interest and allows identifying the top performing molecules in a virtual library.

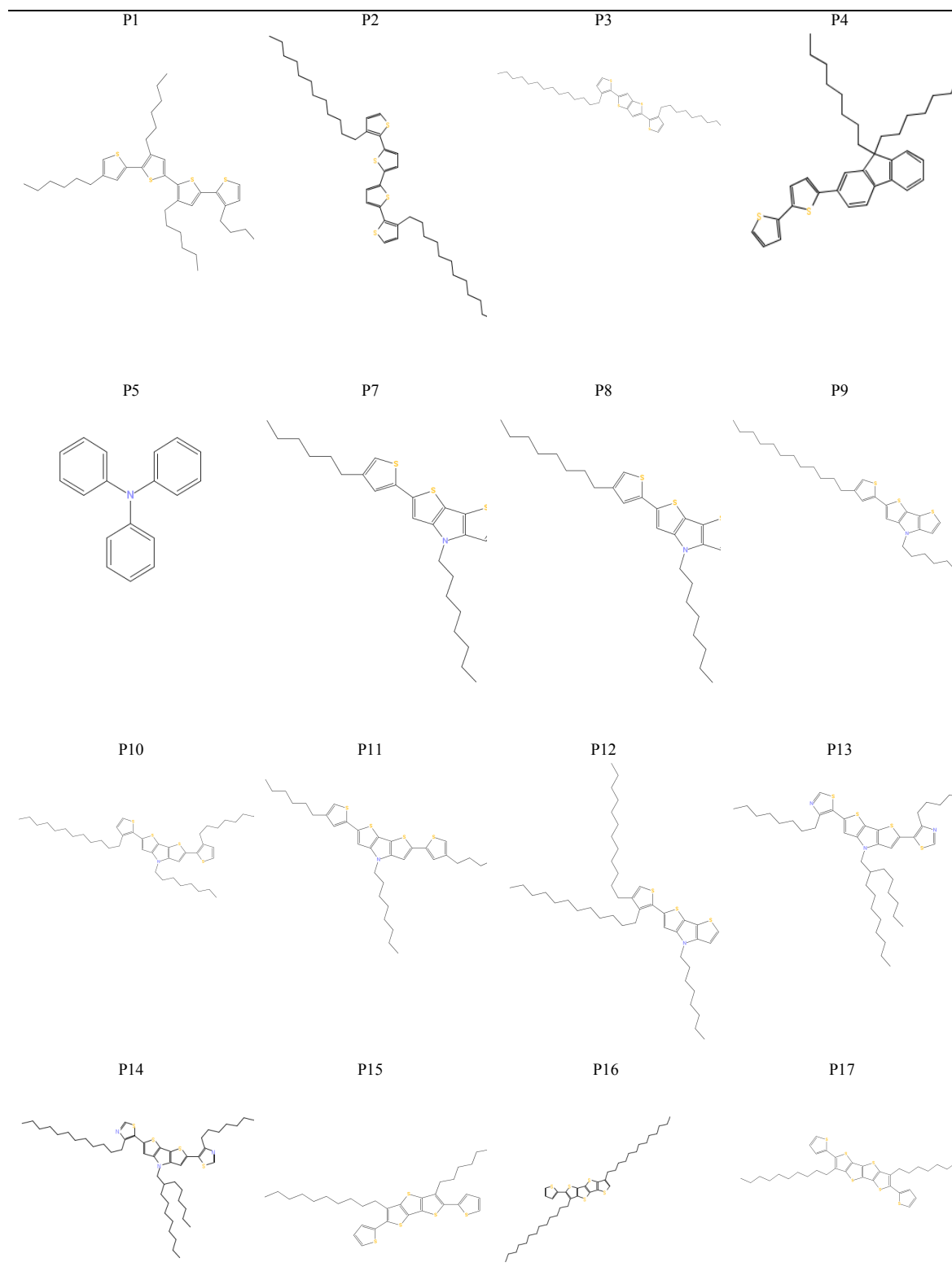
Virtual screening has been widely employed by the pharmaceutical industry¹⁴⁴ and VS can be directly applied in the organic materials science for the prediction of semiconducting nature of the organic molecules. Under this philosophy, high-throughput screening of organic materials has a long way to go. In the field of drug discovery, usually known chemical space is explored or at least starts from a known drug rather than the creation of new chemical libraries de novo. This is due to multiple reasons like large capital investment or the cost of late-stage failures or due to the complexities involved in understanding the structure-biological activity relationships than the structure-property relationship. Unlike the structure-activity relationships in drug discovery, the relation between molecular structures, electronic structure and device properties in organic electronic materials are more straightforward. But the CPU cost of computational methods in organic materials is quite high and it is too important to have a high hit-ratio. Thus custom libraries were generated for the predictive models and virtual screening. In our study custom-generated libraries are reported, in the areas on organic semiconductors and insulators (considered as non semiconductors) as mentioned from Tables 6-8. The small molecule custom library was created from various literatures, chemical database and web resources.^{145,146, 147}

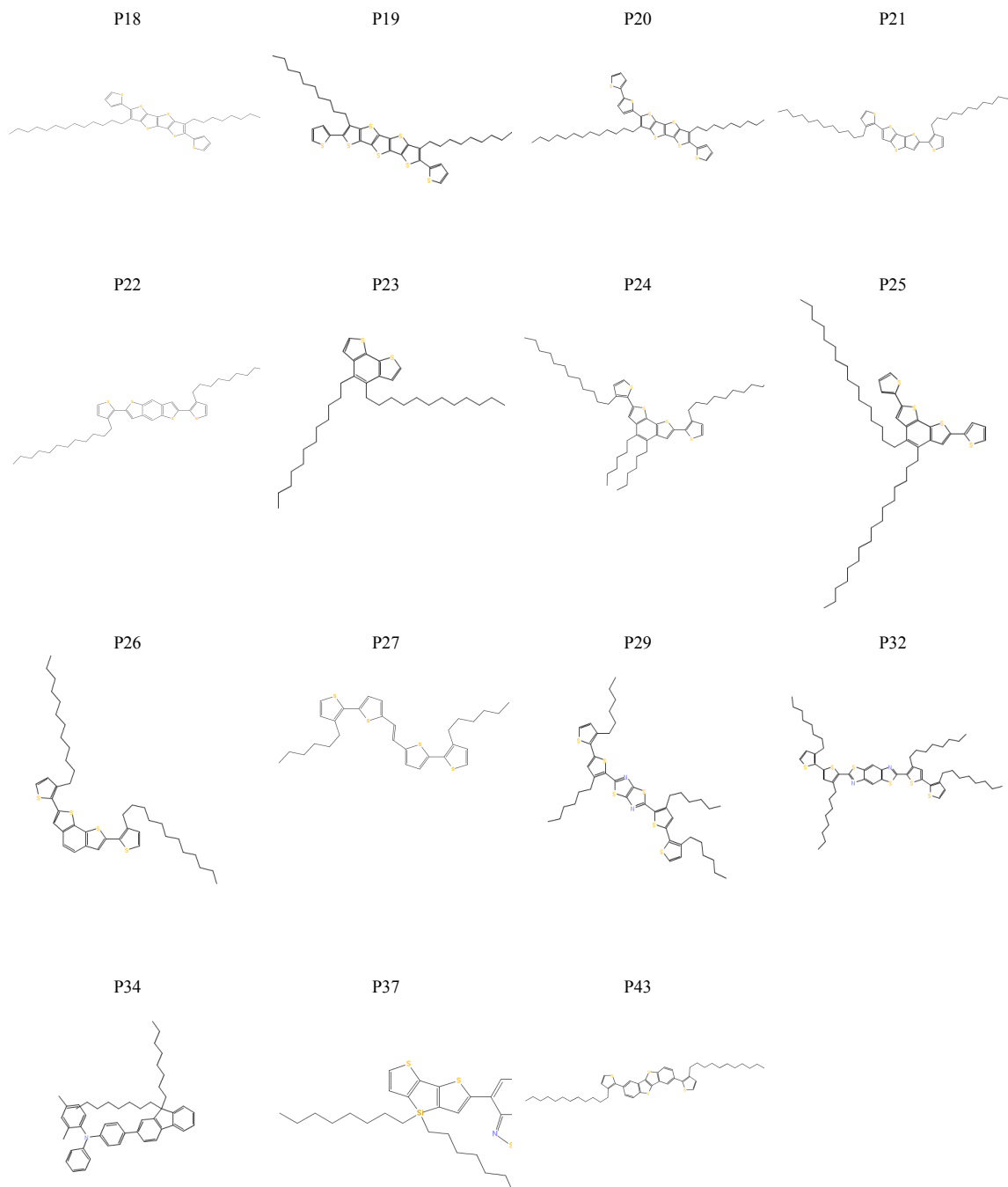
Table 6. List of Organic Semiconductors used in this study

1,1,4,4-tetraphenyl-1,3-butadiene 	2-(4-Biphenyl)-5-phenyl-1,3,4-oxadiazole, (PBD) 	2-(4-Bromophenyl)-5-phenyl-1,3,4-oxadiazole 	2,1,3-Benzoxadiazole-5-carboxylic acid 	2,5-Bis (4-aminophenyl) 1,3,4-oxadiazole 
5-(4-Methoxyphenyl)-1,3,4-oxadiazole-2-thiol 	5-(4-Methylphenyl)-1,3,4-oxadiazole-2-thiol 	5-(4-Pyridyl)-1,3,4-oxadiazole-2-thiol 	5-Phenyl-1,3,4-oxadiazole-2-thiol 	9,10 phenanthrenequinone 
N,N'-Bis(3-methylphenyl)-N,N'-diphenylbenzidine 	N,N'-Di-[(1-naphthalenyl)-N,N'-diphenyl]-1,1'-biphenyl-4,4'-diamine, (NPD) 	N,N'-Diphenylbenzidine 	Pentacene 	Perylene 
Rubrene 	Tetracene 	Triphenylamine 	Triphenylene 	Tri-p-tolylamine 
2,5-Diphenyl-1,3,4-oxadiazole 	Anthracene 	Phenanthrene 	tris-4-diethylaminophenylamine 	

It is a custom made library consisting of small organic semiconductors that were retrieved from online literature and were selected based on the HOMO-LUMO band gap energy.

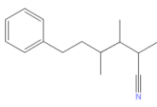
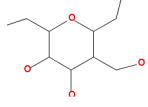
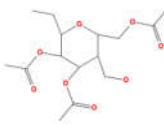
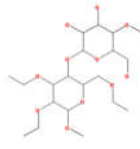
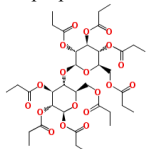
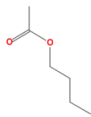

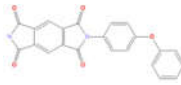

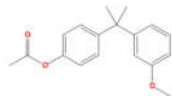
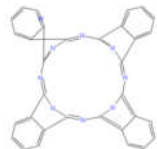
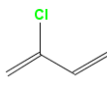
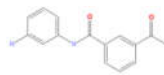
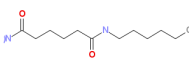
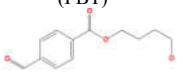
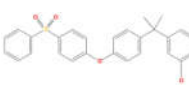
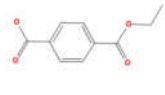
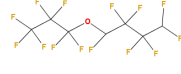

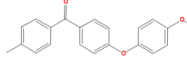
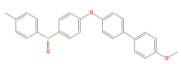

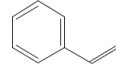
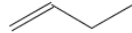
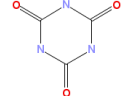
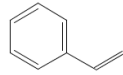
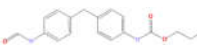
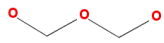
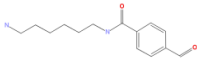
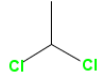
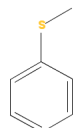
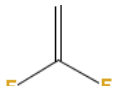
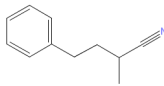


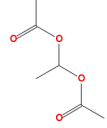
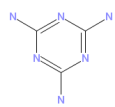
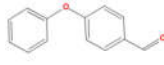
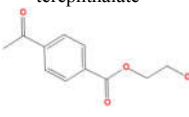
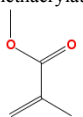
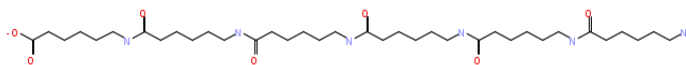
Table 7. List of n-type and p-type Organic Semiconductors used in this study





It is a custom made library consisting of n-type and p-type organic semiconductors that were retrieved from literature.

Table 8. List of non-semiconductors in this study

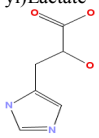
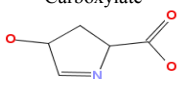
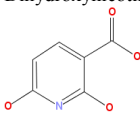
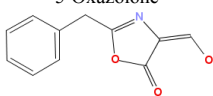
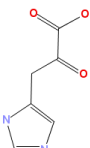
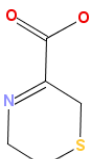
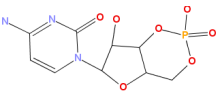
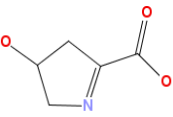
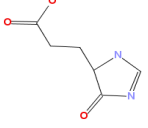
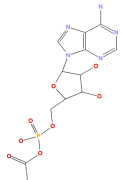
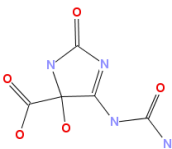
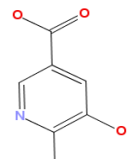
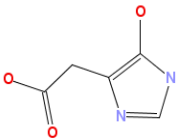
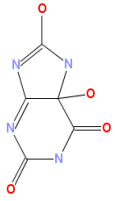
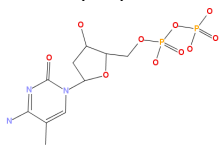
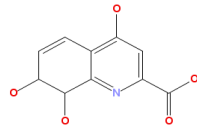
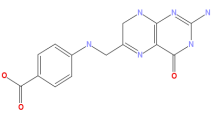
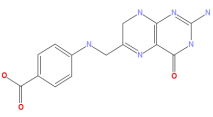

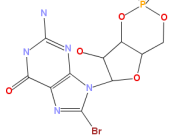
Acrylonitrile Butadiene Styrene (ABS)	Cellulose acetate butyrate	Cellulose acetate	Cellulose ethyl	Cellulose propionate
				
Ethylene-vinyl acetate	Fluoro Ethylene Propylene (FEP)	Kapton	Kynar	Lexan
				
Mica	Neoprene	Nomex	Nylon 66	Polybutylene terephthalate (PBT)
				
Polyethersulfone (PES)	Poly-Ethylene Terphthalate Glycol (PETG)	PerFluoroAlkoxy (PFA)	Poly(p-phenylene oxide)	Poly(ether ether ketone) (PEEK)
				
Polyarylsulfone	Polypropylene	Polyaryletherketone	Polybutylene	Polyisocyanurate
				
Polystyrene	Polyurethane	Poly(oxymethylene)	Polyphthalamide	polyphenylene sulfide
				
Polyphenylene Sulfide	Poly vinylidene Fluoride (PVDF)	Styrene Acrylonitrile (SAN)	Tetrafluoroethylene	Poly(vinyl chloride) (PVC)
				
Delrin	Melamine	polyetherketoneetherketone (PEKEKK)	Poly(ethylene terephthalate	Polymethyl methacrylate
				
		Nylon 6		
				

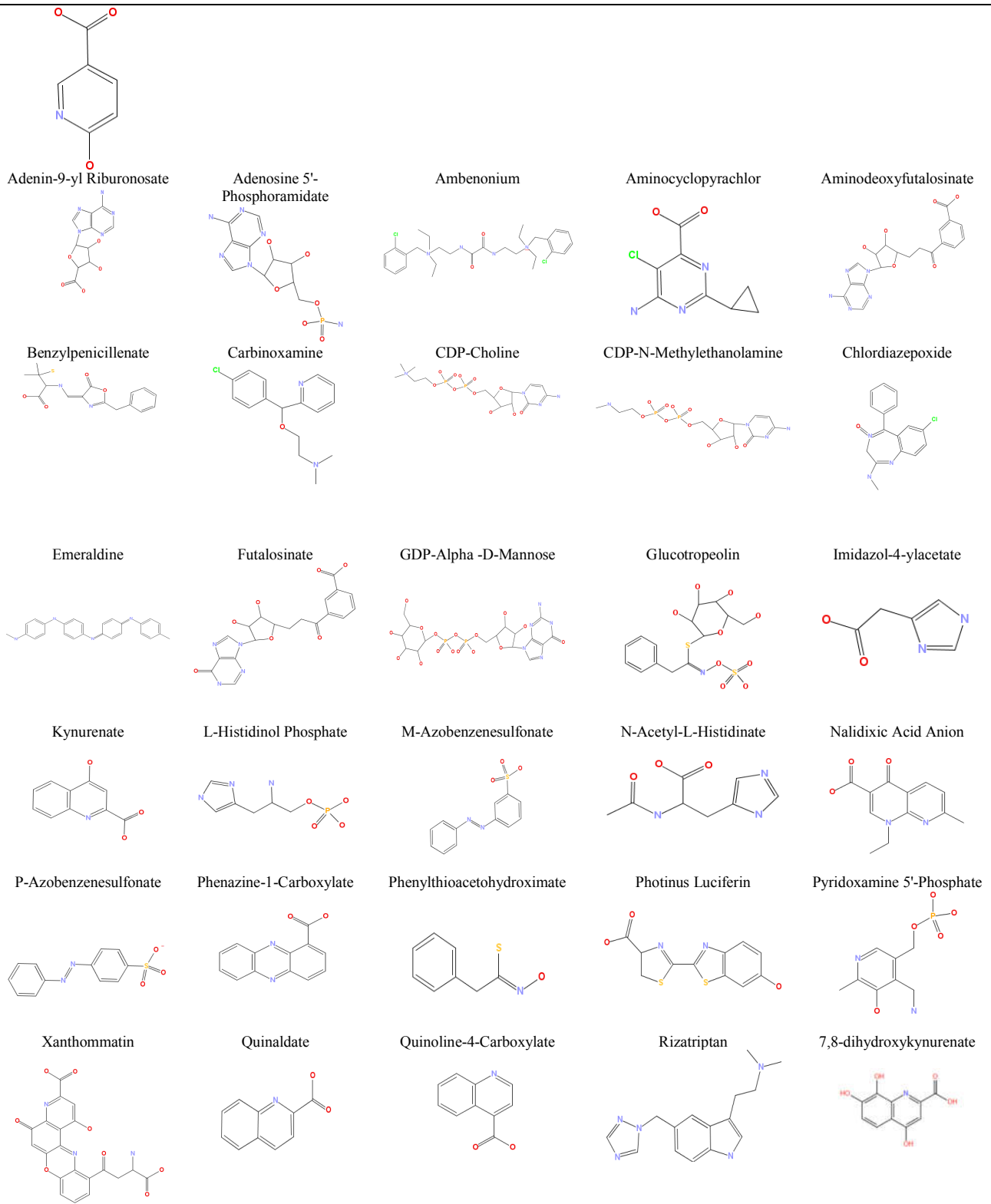
It is a custom made library consisting small molecules of insulators that were treated as non semiconductors were retrieved from various online web resources.

2.11.2 Virtual Screening-Screening Set

In the field of Chemistry, high-throughput analysis has emerged to address the issues associated with larger, more complex datasets. Such analysis has improved the relationships between structure and property and the outcome is data-driven models that are used for the future experiment based on desired properties. And their application will greatly enhance the understanding of the basic physical and chemical principles. For this study, custom generated library selected for the VS as mentioned in the Table 9.

Table 9. List of molecules as screening set used in this study consisting of Schiff Base and Azo compound

(S)-3-(Imidazol-5-yl)Lactate 	1-Pyrroline-3-Hydroxy-5-Carboxylate 	2,6-Dihydroxynicotinate 	2-Benzyl-4-Oxidomethylene-5-Oxazolone 	3-(Imidazol-5-yl) Pyruvate 
3,4-Dehydrothiomorpholine-3-Carboxylate 	3',5'-Cyclic CMP 	4-Hydroxy-1-Pyrroline-2-Carboxylate 	3-(4-oxo-4,5-dihydro-1H-imidazol-5-yl)propanoic acid 	5'-Acylphosphoadenosine 
5-Hydroxy-2-Oxo-4-Ureido-2,5-Dihydro-1H-Imidazole-5-Carboxylate 	5-Hydroxy-6-Methylpyridine-3-Carboxylate 	5-Hydroxyimidazole-4-Acetate 	5-Hydroxyisouric Acid Anion 	5-Methyldeoxycytidine 5'-Diphosphate 
6-Hydroxynicotinate 	7,8-Dihydro-7,8-Dihydroxykynurenate 	7,8-Dihydropteroate 	S-Adenosyl-4-Methylthio-2-Oxobutanoate 	8-Bromo-3',5'-Cyclic GMP 



The screening set was prepared from the random selection of Schiff Bases and azo compound from the online repository Chemical Entities of Biological Interest (ChEBI) database.

CHAPTER 3

DEVELOPMENT OF BAYESIAN MODELS BASED ON ORGANIC SEMICONDUCTORS

3 Bayesian Classifier

Classification is a common task that is practiced by many machine learning algorithms in order to discover knowledge from prior datasets. These datasets provide information on the trends of each class which is used to predict the class for new instances.¹⁴⁸

In this chapter we developed predictive models based on Bayesian (Bayes theorem) classification algorithm where it deals with conditional probability. The concept of *conditional probability* is that an event is a probability obtained with the additional information that some other event has already occurred. The algorithm uses the equation (3) for finding the probability of $P(B|A)$ that denote the conditional probability of event B occurring, given that event A has already occurred. The following formula was provided for finding the probability $P(B|A)$.^{54,149,150}

$$P\left(\frac{B}{A}\right) = \frac{P(A \text{ and } B)}{P(A)} \quad (3)$$

In Bayesian algorithm the posterior probability is calculated for all classes, and the class with the highest probability will be the instance's label. It is based on the Bayesian theorem and it is particularly suited when the dimensionality of the inputs are high. Generally Naive Bayes uses the method of maximum likelihood and

assumptions. It is a simple probabilistic classifier that often performs better in many complex real world situations.

3.1 Materials and methods

The softwares and online web servers used for the model built are specified in Chapter 2.

3.2 Experimental Studies

3.2.1 Dataset preparation

In our study we prepared custom-generated libraries from various literature and available resources as mentioned in the introduction. There were two classes of datasets one for organic semiconductors and the other for non semiconductors were used for the model generation as mentioned in the materials and methods. The library consists of different types of small molecules involving n-type and p-type organic semiconductors like pentacene, tetracene along with thermosetting and thermoplastics etc. For non semiconductor organic molecules we chose insulators like ABS (acrylonitrile butadiene styrene), Nomex, Kapton, PVC (poly vinyl chloride)^{151,152} for training the dataset in the ML process. The entire dataset of organic semiconductors and non semiconductors were built and modeled using HyperChem and later they were saved in .mol format. The geometry optimization was carried out with the Schrodinger suite (MacroModel) and the energy parameters for semiconductors and non semiconductors are mentioned in the tables 11-12. Later, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energy was calculated on the basis of semiempirical level of theory. As an example the acene molecule pentacene being one of the most widely used organic semiconductors is due to its molecular structure which can be used for various applications in the electronics era. Once the library of organic semiconductors and non semiconductors were prepared in the respective file format, molecular descriptors and electronic descriptors were generated from the electronic version of the E-Dragon software of Virtual Computational Chemistry Laboratory (VCC). Table 10 lists the 20 block descriptors calculated by the E-Dragon software.

Table 10. 20 block descriptors generated by the e-dragon software

Blocks	Descriptor list	Calculated descriptors
1	Constitutional descriptors	48
2	Topological descriptors	119
3	Walk and path counts	47
4	Connectivity indices	33
5	Information indices	47
6	2D autocorrelations	96
7	Edge adjacency indices	107
8	Burden eigen value descriptors	64
9	Topological charge indices	21
10	Eigen value-based indices	44
11	Randic molecular profiles	41
12	Geometrical descriptors	74
13	RDF descriptors	150
14	3D-morse descriptors	160
15	WHIM descriptors	99
16	GETAWAY descriptors	197
17	Functional group counts	154
18	Atom-centered fragments	120
19	Charge descriptors	14
20	Molecular properties	29

The energy gap between HOMO and LUMO was calculated using PM3 a semiempirical method and every time there was an increment of 3.6 higher than reference value 1-4.9 eV. Finally, the dataset with a reduction of 3.6 ± 0.1 in band gap was used as organic semiconductors. The geometry optimization was also carried out for organic semiconductors and non semiconductors are mentioned in the Tables 11-12.

Table 11. Organic Semiconductors MacroModel minimization energy values

S.No	Organic Semiconductors	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	van der Waals Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005
1	1,1,4,4-tetraphenyl-1,3-butadiene	163.635162	7.846752	10.767722	60.749863	0.104101	75.703934	8.462793	0.044705
2	2-(4-Biphenyl)-5-phenyl-1,3,4-oxadiazole, (PBD)	124.734329	7.697057	10.093349	0.08368	0	72.440063	34.420181	0.019482
3	2-(4-Bromophenyl)-5-phenyl-1,3,4-oxadiazole	82.348328	3.366554	3.465445	2.092	0	41.038498	32.38583	0.042635
4	2,1,3-Benzoxadiazole-5-carboxylic acid	21.904514	2.848986	12.376802	18.89076	0	22.387562	-34.599598	0.007417
5	2,5-Bis(4-aminophenyl)-1,3,4-oxadiazole	21.986347	3.224056	20.869329	2.092	0	42.781307	-46.980347	0.038328
6	2,5-Diphenyl-1,3,4-oxadiazole	91.694038	3.301597	3.441925	2.092	0	42.340755	40.517757	0.04738
7	5-(4-Methoxyphenyl)-1,3,4-oxadiazole-2-thiol	21.81354	2.461823	12.158251	1.047467	0	26.846714	-20.700714	0.008517
8	5-(4-Methylphenyl)-1,3,4-oxadiazole	15.822765	1.775871	4.026211	1.046	0	20.241783	-11.2671	0.005956
9	5-(4-Pyridyl)-1,3,4-oxadiazole-2-thiol	-184.193787	0.871293	4.628162	1.046	0	24.538191	-215.277435	0.002367
10	5-Phenyl-1,3,4-oxadiazole-2-thiol	24.679617	1.720329	3.684409	1.046	0	20.050566	-1.821685	0.011653
11	9,10 phenanthrenequinone	204.815613	8.016693	5.15384	1.359874	0.000003	61.557442	128.727753	0.034984
12	Anthracene	55.835468	3.668622	0.106178	0.000002	0	54.923141	-2.862471	0.011745
13	N,N'-Bis(3-methylphenyl)-N,N'-diphenylbenzidine	57.863194	8.409162	33.909069	40.336643	0.008738	107.300926	-132.101349	0.038406
14	N,N'-Di-[(1-naphthalenyl)-N,N'-diphenyl]-1,1'-biphenyl-4,4'-diamine, (NPD)	129.654358	15.180117	37.88435	37.147724	0.094064	153.806824	-114.458717	0.039264
15	N,N'-Diphenylbenzidine	40.57243	9.130772	53.003235	-2.00832	0	94.792137	-114.34539	0.04556
16	P1	52.446095	1.736226	29.547152	51.10532	0.290853	-46.456558	16.223103	0.046285
17	P2	54.921268	1.80632	26.502316	36.119034	0.073973	-25.619923	16.039553	0.041843
18	P3	123.976524	17.361467	109.433571	37.072796	0.135849	-62.04047	22.013311	0.046832
19	P4	160.839218	6.826996	95.322723	25.397528	0.286552	26.323788	6.681628	0.044207
20	P5	699.895691	158.648575	374.094543	0	0	202.72197	-35.569389	0.00555
21	P7	241.107529	12.299754	214.818924	23.473501	0.149666	-18.44709	8.812778	0.047726

S.No	Organic Semiconductors	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	van der Waals Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005
22	P8	243.839752	12.435467	215.568939	23.314901	0.146446	-18.658949	11.032952	0.03867
23	P9	252.686234	12.712566	217.892807	26.886288	0.151333	-19.285093	14.32833	0.047271
24	P10	264.575317	15.734925	228.885895	41.130424	0.211975	-50.390789	29.002886	0.041328
25	P11	254.702286	13.899329	221.633621	30.497887	0.157178	-23.16044	11.67471	0.047617
26	P12	251.712006	16.941082	226.686676	25.650227	0.316549	-40.578556	22.696032	0.046273
27	P13	256.049622	16.385752	224.701279	42.62328	0.257951	-67.551346	39.632698	0.043649
28	P14	273.619965	17.354778	229.380249	39.490501	0.315915	-60.34771	47.426235	0.049336
29	P15	210.542191	14.803642	167.980865	33.217884	0.051846	-34.875221	29.363167	0.042515
30	P16	260.665131	14.443419	227.310837	17.203791	0.060342	-30.540651	32.187397	0.048034
31	P17	268.093628	14.964754	228.359787	33.823784	0.067124	-38.354931	29.233101	0.047217
32	P18	283.157074	15.492399	232.211121	40.492882	0.06983	-38.783115	33.673958	0.046939
33	P19	325.298218	15.453766	288.939301	35.493946	0.043822	-43.697994	29.065384	0.0488
34	P20	284.613312	15.730103	234.082016	41.52317	0.242249	-41.566299	34.602081	0.044692
35	P21	204.624039	13.773409	172.664871	33.868771	0.054661	-34.685623	18.947943	0.047988
36	P22	67.137108	2.987127	44.264572	36.766247	0.116079	-29.503082	12.506168	0.04416
37	P23	150.454208	9.865905	93.537315	3.362418	0.009249	15.546271	28.133053	0.048435
38	P24	175.36879	14.417026	112.170738	29.71328	0.097765	-13.060825	32.030815	0.048938
39	P25	179.936584	11.798286	103.504539	19.896652	0.160528	8.867426	35.709145	0.048282
40	P26	148.606659	7.849204	99.794289	34.048164	0.030259	-10.34376	17.228506	0.044255
41	P27	55.650166	2.017098	23.522291	60.602688	0.826962	-34.206104	2.887227	0.038232
42	P29	169.952606	13.929951	83.355728	64.019119	0.82875	-48.990883	56.809944	0.04837
43	P32	89.190338	4.261045	46.127502	64.093811	0.806228	-43.851688	17.753439	0.049324
44	P34	151.582474	12.009183	107.870918	33.06266	0.178809	87.13446	-88.673553	0.042765

S.No	Organic Semiconductors	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	van der Waals Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005
45	P37	222.407791	4.022821	151.790909	24.433558	0.012258	-8.43254	50.580784	0.035437
46	P43	126.145836	10.559002	85.298088	37.483475	0.056059	-6.2416	-1.009186	0.049176
47	Pentacene	83.756203	6.365469	0.157719	0.000014	0.000001	90.511803	-13.278805	0.038772
48	Perylene	204.88591	14.418784	81.540161	-4.01664	0	93.612167	19.331429	0.025164
49	Phenanthrene	75.995804	6.109811	3.734605	0	0	60.31736	5.834027	0.019276
50	Ruberene	390.201263	31.885578	26.763197	106.09552	1.617807	184.645081	39.194073	0.031624
51	Tetracene	69.835953	5.014995	0.131981	0.00004	0.000003	72.719818	-8.030879	0.026517
52	Tri-p-tolylamine	674.942932	159.079651	376.173767	0.000006	0	202.53717	-62.84763	0.025576
53	Triphenylamine	699.895691	158.660934	374.086456	0	0	202.717773	-35.569454	0.020109
54	Triphenylene	140.820557	14.557887	11.55773	0	0	96.434235	18.270702	0.0222
55	Tris-4-diethylaminophenylamine	-116.569824	12.165013	29.590994	-19.548944	0.658188	92.599182	-232.034256	0.047825

Table 12. Non Semiconductors MacroModel minimization energy values

S.No.	Non Semiconductors	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	van der Waals Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005
1	ABS	78.693855	5.303289	13.260797	21.24612	0.001895	29.910728	8.971024	0.042114
2	Cellulose acetate	64.81897	7.170308	64.303688	91.689133	0.009812	14.113544	-112.46751	0.047248
3	Cellulose acetate butyrate	109.25599	4.7099	17.143412	110.57761	0	27.841766	-51.016705	0.038505
4	Cellulose ethyl	219.26242	9.24289	39.016415	134.0665	0	20.125961	16.810659	0.034752
5	Cellulose propionate	-350.6275	14.077498	82.71608	47.807175	0.048986	5.121709	-500.39896	0.04077
6	Delrin	-49.214214	2.78678	9.915618	-9.587226	0.012794	14.915639	-67.25782	0.045236
7	Ethylene-vinyl acetate	5.777717	1.02615	11.30118	13.731949	0	6.297262	-26.578825	0.041276
8	Kapton	250.28003	10.146706	185.92845	78.37928	0.022584	58.871433	-83.068413	0.035287
9	Kynar	-8.740515	0.001606	5.102122	0	0	-0.078661	-13.765581	0.004196
10	Lexan	47.95475	8.010608	26.066835	20.491785	0.1466	63.938072	-70.69915	0.035657
11	Melamine	-952.65027	1.118488	36.27998	0.000109	0.000062	45.075882	-1035.1248	0.043429
12	Mica	648.20709	11.623855	558.13226	2.770439	0.060262	111.77464	-36.154392	0.044549
13	Neoprene	-8.031582	0.166023	2.375197	-3.112896	0	3.348075	-10.80798	0.003764
14	Nomex	90.128036	4.702282	24.597372	65.275261	0.327077	46.403057	-51.177013	0.048615
15	Nylon 6	-238.95413	3.7054	32.014614	35.935429	0.029981	16.60483	-327.24439	0.047167
16	Nylon 66	-153.51817	1.106508	5.290629	0.393815	0.000174	5.615934	-165.92523	0.049797
17	PBT	45.342262	3.069195	3.202442	6.080599	0.000699	30.785944	2.203381	0.048176

S.No.	Non Semiconductors	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	van der Waals Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005
18	PEKEKK	135.69547	5.313781	61.779613	0	0	47.960289	20.641777	0.002485
19	PES(POLYETHERSULFONE)	113.74303	8.921412	14.514061	10.075331	0.099286	83.227959	-3.095021	0.040379
20	PETG	-2.039053	3.528212	4.555061	18.894514	0	33.885757	-62.902599	0.011586
21	PFA	1.352293	1.950069	5.94276	-67.702087	0	16.037195	45.124359	0.04877
22	Poly(ether ether ketone)	257.84787	17.790466	116.58485	0.001469	0	98.022644	25.448448	0.043988
23	Poly(ethylene terephthalate	49.734444	4.019032	8.341609	20.833223	0	36.427906	-19.887327	0.034506
24	Poly(p-phenylene oxide)	20.973429	1.857157	8.664197	0.001456	0	26.111982	-15.661363	0.017507
25	Polyaryletherketone	37.739326	19.125925	32.613247	42.228477	0.131514	152.98503	-209.34486	0.047438
26	Polybutylene	13.037167	0.047337	0.366792	5.535151	0	0.116742	6.971144	0.016935
27	Polyisocyanurate	-416.93433	1.923593	8.405294	0	0	55.771141	-483.03436	0.004785
28	Polymethyl Methacrylate	-19.948334	1.344495	1.996333	-1.832485	0	13.543864	-35.000542	0.009535
29	Polypropylene	9.883329	0.037665	0.459319	0.038085	0	-0.274849	9.623109	0.000571
30	Polystyrene	27.143047	1.476049	2.40964	1.979061	0.000001	22.64933	-1.371033	0.013538
31	polyurethane	24.994802	7.204564	68.114464	16.762411	0	60.569561	-127.6562	0.048778
32	POM (poly(oxyethylene))	-26.47002	0.21336	2.128538	-14.027442	0	2.87358	-17.658056	0.021902
33	PPA polyphthalamide	54.423328	2.414203	9.567377	31.890387	0.351959	24.799288	-14.599883	0.049931
34	PPS(polyphenylene sulfide)	18.215984	1.123308	2.430233	0.000701	0	19.20261	-4.540866	0.000884
35	PVC	1.118785	0.00146	0.579802	0	0	-0.21514	0.752663	0.000135
36	PVDC	-3.762079	0.008291	0.038595	0.006999	0	-0.421063	-3.394902	0.0005

S.No.	Non Semiconductors	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	van der Waals Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005
37	PVDF	-8.740515	0.001603	5.102114	0	0	-0.078671	-13.765561	0.000033
38	SAN	17.82847	1.314577	2.195089	-2.454637	0.001668	16.733192	0.038579	0.041307
39	Tetrafluoroethylene	24.246799	0.010032	9.947694	0	0	0.148026	14.141047	0.000042
40	FEP	26.124664	5.22923	14.742977	-30.073477	0	22.069603	14.156329	0.009722
41	Polyarylsulfone	106.89739	5.909462	13.999976	18.606318	0.021362	77.288246	-8.927973	0.03954

The different energy parameters were calculated for the non semiconductors that were used in the custom made library. And the listed compounds are geometrically optimized from MacroModel package from Schrodinger suite.

3.2.2 Model Generation

For the generation of machine learning models the datasets involving organic semiconductors and non semiconductors were taken. The number of data points corresponding to semiconductors and non semiconductors are given in Table 13.

Table 13. Number of data points corresponding to semiconductors and non semiconductors.

S.No.	Semiconductors	No. of molecules
1	Organic Semiconductors	55
2	Non semiconductors	41
	Total	96

Each dataset was uploaded into the e-dragon software and calculated the 20 block descriptors and the output file was downloaded in the text format. After descriptor generation, the datasets were converted from comma separated value (CSV) to attribute relation file format (ARFF) in the ML package WEKA. The ordered sets of instances were preprocessed, randomized, filtered and split into training set and test set. The former dataset corresponds to the majority class (80%), and the latter to the minority class (20%). The training set was imported to the WEKA in the preprocessor panel. The histogram indicates the descriptor distribution by two different colors as the individual class as “*organic semiconductors*” and “*non-semiconductors*”. A total of 1664 descriptors calculated from the e-dragon server were reduced to about 1448 descriptors by choosing “*Remove Useless*” method as shown in Figure 17.^{153,154}

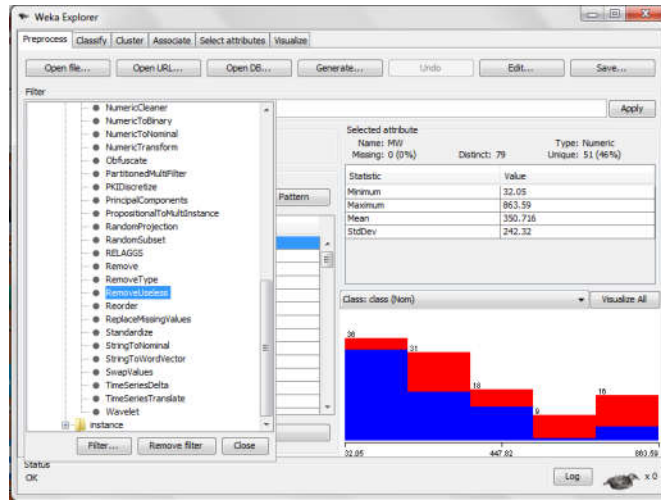


Figure 17. Selection of Remove Useless method from the WEKA explorer.

Bayesian model was developed by importing the training set and test set into the WEKA environment for the default model generation (Model 1). Here the cost sensitive analysis and sampling method is not performed and the model was generated as described in section 3.2.2. The dataset with the imbalanced data points of non semiconductors as in our case were solved through the oversampling method i.e. oversampling of the minority class for the generation of Model 2 (Oversampled model). The data points of training and test set corresponding to the models are given in the following Table 14.

Table 14. Number of data points corresponding to training set and test set.

	Model 1	Model 2
1. Training set	77	110
2. Test set	19	27
3. Attributes	1448	1448

Among the various WEKA classifiers we preferred the Naïve Bayes because of its simplicity and less computational complexity. The performance of the classifier was evaluated by a 10 fold cross validation where the dataset was randomized and split into 10 folds of equal size. Out of 10 fold, one fold was used for testing and the remaining was used for training the classifier in an iterative

manner as mentioned in the materials and methods section. After these processes, two models were generated; the one corresponding to Model 1 (Default Model) was trained upon 77 molecules with 19 molecules as test samples. All the test set samples were re-evaluated upon the training set by ten by ten stratified cross validation. Similar model was generated for the oversampled data points of non semiconductors containing a total of 137 molecules. Here, the training set of 110 molecules was tested with test set of 27 molecules. And finally the two computational predictive Bayesian models for Model 1 and Model 2 are displayed in Figures 18-19.

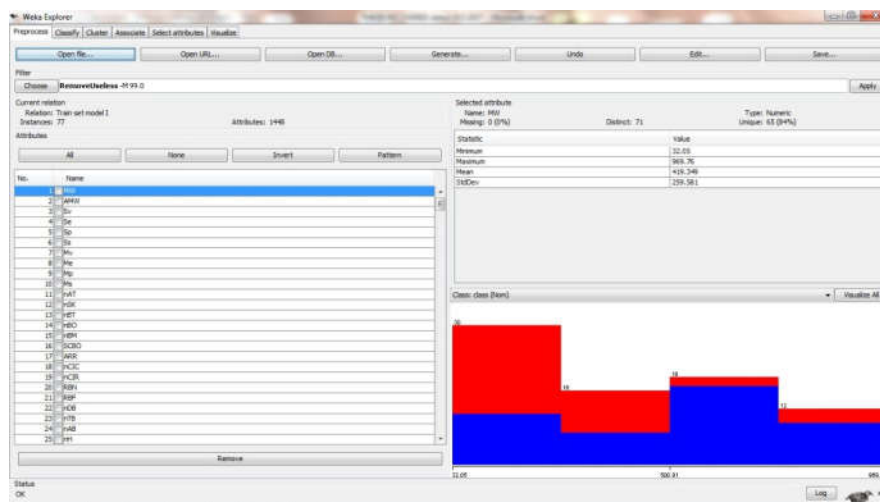


Figure 18 (a). Bayesian Model 1 generation process from preprocess panel.

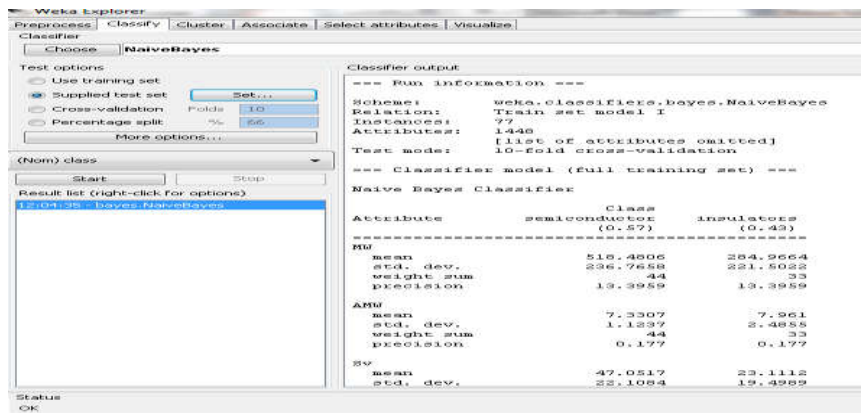


Figure 18 (b). Cross validation and Bayesian Model 1 (Default) generation from WEKA preprocess panel.

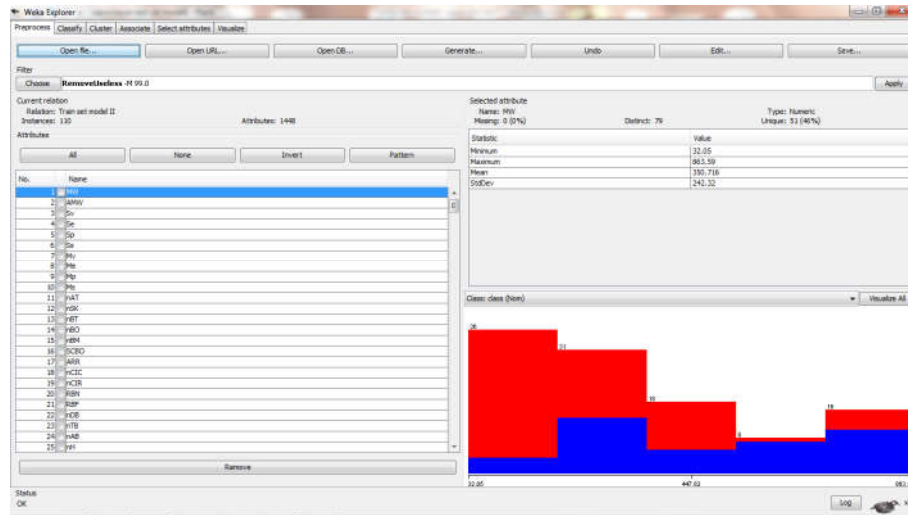


Figure 19 (a). Model generation for Model 2 (Oversampled) from WEKA preprocess panel.

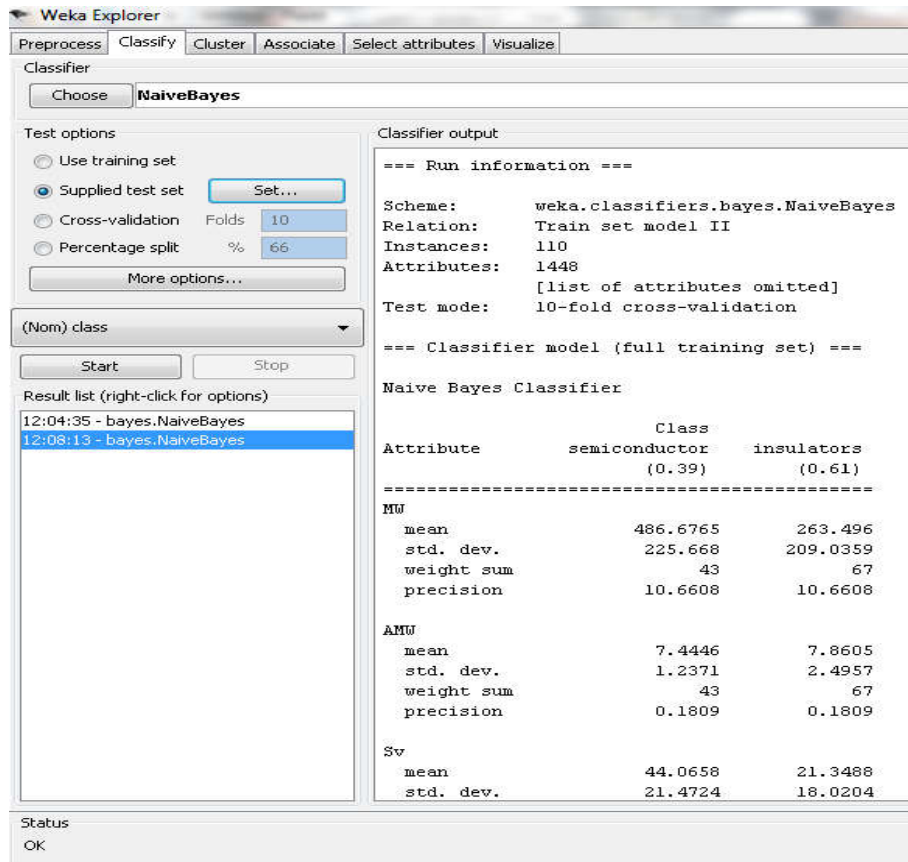


Figure 19 (b). Cross validation and Bayesian classifier Model 2 (Oversampled) generation.

3.2.3 Results and discussion

3.2.3.1 Model robustness

The performance of Bayesian algorithm was analyzed using a confusion matrix of two class problem as shown in Table 15. And the matrix is analyzed and read as given in Table 16.

Table 15. Confusion matrix for Model 1 and Model 2.

Model 1	Model 2
<pre>a b <-- classified as 8 3 a = semiconductor 0 8 b = insulators</pre>	<pre>a b <-- classified as 14 1 a = insulators 2 10 b = semiconductor</pre>

Table 16. The 2 x 2 confusion matrix displayed for Model 1 and Model 2.

	Prediction as Active	Prediction as Inactive
Active	TP	FN
Inactive	FP	TN

The above confusion matrix is read as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

The confusion matrix was analyzed in terms of True Positive (TP), True Negative (TN), False Negative (FN) and False Positive (FP). TP and TN denote the number of real positives and real negatives that are classified correctly, while FN and FP denotes the number of misclassified positive and negative examples.¹⁵⁵ One of the important points during development of the classifiers is that, percentage of false negatives is more important than percentage of false positives for compound selection. To attain this, one could minimize the number of false negatives at the expense of increasing the false positive. Increasing misclassification for false

negatives would lead to increase in both false positives and true positives. The percentage of false positives was kept in check by setting an upper limit on FP rate. In this case, the limit of FP rate was set to a maximum of 20% and cases where standard classifiers producing this result, cost-sensitivity analysis were not used and only default classifiers were used.¹⁵⁶ Table 17 shows all the classification results against the 2 x 2 confusion matrix of Model 1 and Model 2. The fineness of the classifying algorithm was determined by the evaluation measures like Accuracy, Precision, Recall, Kappa, Sensitivity, Specificity, ROC and F-measure^{83,157} as tabulated in Table 18.

Table 17. The evaluation measures TP, FP, FN, TN, TP rate and FP rate generated by the Models 1 and 2

Bayesian Classifier	TP	TN	FP	FN	TP rate %	FP rate %
Model 1	8	8	0	3	72.7	0
Model 2	10	14	1	2	83.3	6.7

Table 18. The evaluation measures Precision, Recall, *F*-measure, ROC, Accuracy and Kappa generated by the Models 1 and 2

Bayesian Classifier	Precision	Recall	Specificity	BAC	F-measure	ROC	Accuracy	Kappa	MCC
Model 1	100	72.7	100	86.35	84.2	87.5	84.2105	0.6919	0.7272
Model 2	90.9	83.3	93.33	88.31	87	91.4	88.889	0.7731	0.7753

The overall robustness of a ML classifier can be judged by the statistical measure accuracy of the generated models. Since we have performed sampling technique in our datasets, the parameter accuracy alone may not turn out to be the real estimator of the Bayesian model fineness. So, another performance balanced accuracy (BAC) was also accounted that equally weights the errors within each class. BAC gives a more precise evaluation of the overall model effectiveness as shown in Table 18. The Receiver Operator Characteristics (ROC) was also used to compare and evaluate the performance of each model (Model 1 and Model 2) for

their efficiency and robustness. The reliability of the model was checked for the Kappa value which was greater than 0.7. Among the two models, the classifier Model 1 produced higher precision and specificity than compared to Model 2 as shown in Table 18. But showcased higher values for Recall, BAC, F-measure with an accuracy > 85 % and with a good Kappa score of 0.77. The graph also displays that the FN for model 1 is higher than model 2 but for TP, TN and FP model 2 has higher number of data points. In the case of TP rate model 2 has performed better in comparison to model 1. While FP rate for Model 1 being 0 and for Model 2 it had increased to 6.7% but the FP rate was under the threshold value 20. From Table 18, it is clear that Model 2 had performed significantly better than Model 1 in terms of accuracy, kappa and MCC. Due to oversampling the accuracy of the Model 2 had been improved. Figure 20 is the comparative graph between the various evaluation measures of both classification models-Model 1 and Model 2. Later the two Bayesian models were selected for the further screening of the Schiff base compounds selected from ChEBI database.

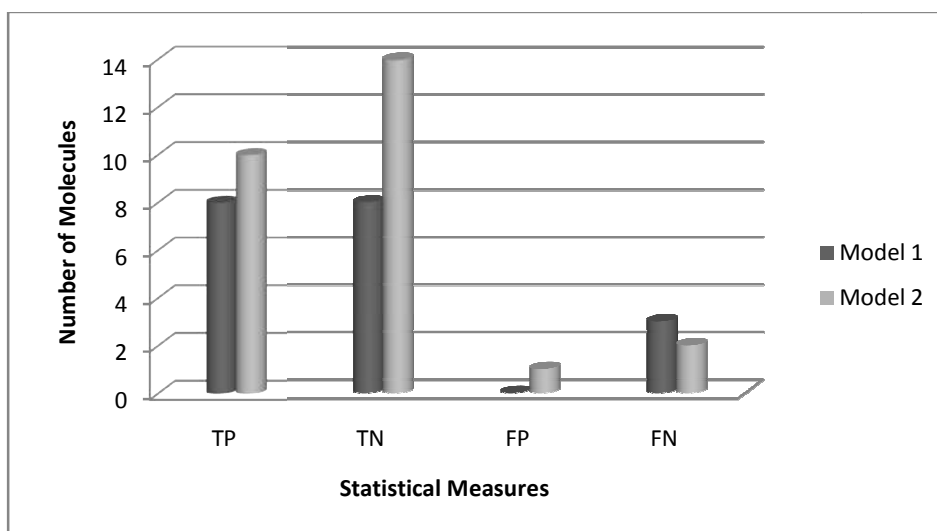


Figure 20 (a). Number of molecules based on confusion matrix

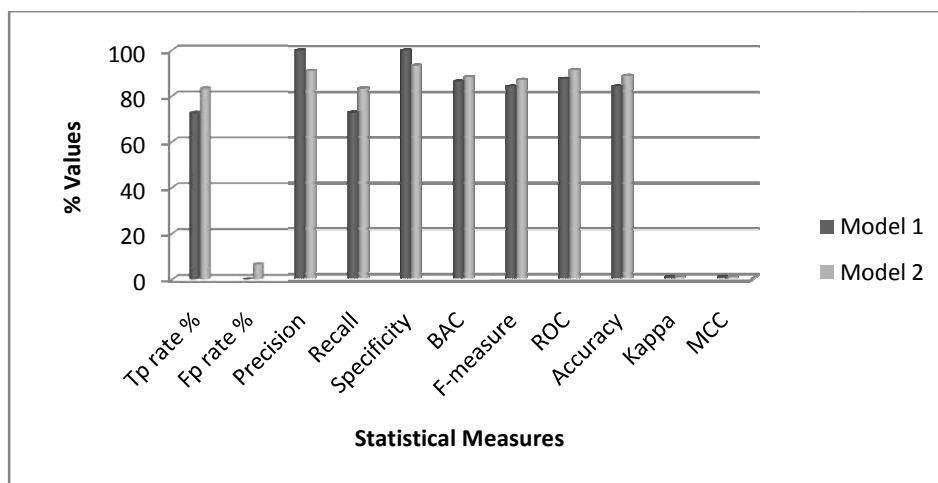


Figure 20 (b). Comparative graph between the various statistical measures of both classification models-Model 1 and Model 2 are presented

3.2.3.2 Virtual Screening and Validation of the model

As mentioned previously virtual screening has been one of the mainstays in the identification of hits in a general drug discovery programme. The cost and the time spent in running high-throughput screens are enormous. Computational virtual screening being a cheaper method could further benefit provided with faster processors, parallel computing and smarter and faster algorithms in prioritizing compound selection. The Schiff base and azo compounds from ChEBI database were selected as screening set for the virtual screening. They were prepared in the similar way as the test set. The geometry optimized screening set energy parameters are presented in Table 19.

Table 19. Geometrically optimized screening set used for virtual screening under the study

S.No.	Screening molecules	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	van der Waals Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005
1	Xanthommatin	-75.955864	14.071092	36.692795	13.742313	0.222804	105.19458	-245.87944	0.048958
2	Rizatriptan	93.851936	2.821414	16.055029	13.000259	0.085479	16.056135	45.833618	0.043825
3	Quinoline-4-Carboxylate	-164.40123	4.077016	11.301263	17.612545	2.378026	48.770332	-248.54041	0.0269
4	Quinaldate	-94.37809	2.934723	6.799718	14.292544	0	45.777847	-164.18292	0.025467
5	Pyridoxamine 5'-Phosphate	-317.56381	2.76606	7.968674	19.46488	0.012511	29.654688	-377.43063	0.047737
6	Photinus Luciferin	-58.87611	3.382584	26.33634	13.026973	0.053052	18.93848	-120.61354	0.023457
7	Phenylthioacetohydroximate	50.522602	1.963237	3.927998	2.895944	0.003488	25.994274	15.737659	0.042059

S.No.	Screening molecules	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	van der Waals Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005
8	Phenazine-1-Carboxylate	108.49406	9.412841	18.136692	18.891111	0.000446	78.098373	-16.045397	0.043114
9	P-Azobenzene-sulfonate	-0.380463	4.534633	14.393931	22.079672	0	49.232201	-90.620903	0.043554
10	Nalidixic Acid Anion	-95.628731	7.354933	11.313037	-25.006132	0.001133	61.428719	-150.72043	0.044079
11	N-Acetyl-L-Histidinate	-133.24266	1.218527	24.053104	31.035831	0.151833	-5.819225	-183.88272	0.046441
12	M-Azobenzene-sulfonate	-31.386225	4.245025	14.415183	22.079885	0	48.878067	-121.00439	0.014585
13	L-Histidinol Phosphate	-307.81546	1.95545	18.5483	3.658328	0.00317	4.744571	-336.72525	0.042675
14	Kynurenate	-131.52396	3.117492	7.061327	14.292544	0	48.552723	-204.54805	0.001087
15	Imidazol-4-ylacetate	21.142469	0.943419	17.28746	11.667429	0.051917	-2.03835	-6.769405	0.024859
16	Glucotropeolin	27.970018	7.465842	36.025681	128.75002	0.049634	36.565819	-180.88698	0.034134
17	GDP-Alpha -D-Mannose	-160.81276	8.398344	50.297684	262.71274	0.021056	43.251011	-525.49359	0.041441
18	Futalosinate	82.931145	11.347034	64.406624	162.78923	0.191959	75.892418	-231.69612	0.042292
19	Emeraldine	73.437103	11.869424	164.05283	-8.368	0	106.23554	-200.35269	0.035073
20	Chlordiazepoxide	-148.01863	7.525411	29.909945	42.850872	1.860856	68.691422	-298.85715	0.042293
21	CDP-N-Methylethanolamine	-336.96542	8.534631	40.16143	94.023506	0.136834	15.606735	-495.42856	0.034316
22	Carbinoxamine	27.763	3.853155	5.997577	31.217896	0.022144	45.832558	-59.160332	0.047216
23	Benzylpenicillenate	80.03849	4.645626	68.965965	-2.784073	0.524035	22.495008	-13.808078	0.04705
24	Aminodeoxyfutalosinate	28.648985	7.896406	36.131264	173.89816	0.201913	60.026554	-249.50531	0.041625
25	Aminocyclopyrachlor	-434.1019	1.350355	16.908974	31.927582	1.783398	30.585552	-516.65778	0.026117
26	Ambenonium	514.7652	31.94396	95.947647	35.228359	0.463823	104.23074	246.95068	0.036841
27	Adenosine 5'-Phosphoramidate	-365.53668	6.136134	21.226974	116.84438	5.902314	31.040613	-546.68707	0.04829
28	Adenin-9-yl Riburonosate	25.347046	5.417504	39.313217	115.2339	1.125029	34.259872	-170.00247	0.042481
29	8-Bromo-3',5'-Cyclic GMP	-363.25967	4.214622	29.439766	107.50671	0.095296	22.39378	-526.90985	0.036892
30	7,8-dihydroxycynurenate	-127.16043	3.974569	11.369832	14.292544	0	53.063019	-209.8604	0.014542
31	7,8-Dihydropteroate	-27.536703	9.60216	25.76516	-8.545625	0.163586	84.185265	-138.70725	0.048476
32	7,8-Dihydro-7,8-Dihydroxycynurenate	-161.76057	3.61037	15.008894	45.561325	0.092607	51.834091	-277.86786	0.02731
33	6-Hydroxynicotinate	-218.59815	1.208356	4.846621	7.715296	0	27.315163	-259.68359	0.035037
34	5-Methyldeoxycytidine 5'-Diphosphate	-334.02719	5.028625	31.087051	83.08445	0.01546	26.528927	-479.7717	0.046374
35	5-Hydroxyisouric Acid Anion	-255.46913	4.4411	89.016068	-34.927589	1.542048	41.473518	-357.01428	0.017515
36	5-Hydroxyimidazole-4-Acetate	-34.799904	1.357585	18.685061	13.406301	0.10873	4.401547	-72.759125	0.034256
37	5-Hydroxy-6-Methylpyridine-3-Carboxylate	-164.75116	1.376217	4.797054	7.715308	0.000002	28.831627	-207.47138	0.013032
38	5'-Acylphosphoadenosine	-125.94065	4.865969	32.906952	150.34139	0.182993	22.13699	-336.37494	0.039965
39	4-Hydroxy-1-Pyrroline-2-Carboxylate	44.738552	1.111397	32.673573	30.068428	0.001262	2.080107	-21.196213	0.030725
40	3',5'-Cyclic CMP	-419.53803	7.565116	21.982203	44.507683	0.247033	49.845345	-543.68543	0.044871
41	3,4-Dehydrothiomorpholine-3-Carboxylate	120.03505	4.175147	21.57925	5.332715	0.134817	14.540237	74.272881	0.033543
42	3-(Imidazol-5-yl) Pyruvate	70.117424	2.210711	18.90831	11.097281	0.242632	5.288191	32.370296	0.038493
43	(S)-3-(Imidazol-5-yl)Lactate	-31.947437	2.341128	25.804852	40.453739	0.021316	15.272335	-115.84081	0.03603
44	2-Benzyl-4-Oxidomethylene-5-Oxazolone	113.78626	1.439101	52.632488	2.93389	0.005588	15.427914	41.347282	0.04888
45	2,6-Dihydroxynicotinate	-205.77779	2.739514	16.237123	7.715296	0	30.629944	-263.09967	0.042158
46	5-Hydroxy-2-Oxo-4-Ureido-2,5-Dihydro-1H-Imidazole-5-Carboxylate	-297.46094	2.557754	74.659363	-3.932234	0.967547	17.604137	-389.31751	0.044869
47	1-Pyrroline-3-Hydroxy-5-Carboxylate	-9.106635	1.211928	31.95422	30.414993	0.033465	5.074837	-77.796074	0.035408
48	3-(4-oxo-4,5-dihydro-1H-imidazol-5-yl)propanoic	-60.420261	1.322617	49.351543	2.419444	1.27291	0.669084	-115.45586	0.034041
49	CDP-Choline	-41.9876	1.134424	35.9830	17.6549	0.01865	9.9076	-56.09872	0.02785
50	S-Adenosyl-4-Methylthio-2-Oxobutanoate	-54.8769	1.76832	46.74530	57.93246	0.06438	17.5973	-88.26938	0.03169

The different energy parameters were calculated for the screening set that was used for virtual screening includes Schiff base compounds and azo compound. All the molecules were geometrically optimized from MacroModel package from Schrodinger suite.

Before performing the virtual screening, the screening set was prepared by calculating molecular descriptors from the e-Dragon server. The calculated 1664 descriptors were later reduced to 1448 by applying the option “*remove useless*” in the WEKA environment. Finally the screening set was preprocessed as screening set.arff format. Then screening set was screened against the Bayesian Model 1 and Model 2 in the data mining package after performing cross validation, re-evaluating of test set upon the training set for the model built. The virtual screening process was obtained from the “output prediction” as shown in Tables 20-21.

Table 20. Semiconductor prediction output from computational Bayesian Model 1.

Inst.	Actual	Predicted	Error	Probability	Distribution
1	semiconductor	insulator	+	0	*1
2	semiconductor	insulator	+	0	*1
3	insulator	insulator		0	*1
4	insulator	insulator		0	*1
5	insulator	insulator		0	*1
6	insulator	insulator		0	*1
7	insulator	insulator		0	*1
8	insulator	insulator		0	*1
9	insulator	insulator		0	*1
10	insulator	insulator		0	*1
11	insulator	insulator		0	*1
12	insulator	insulator		0	*1
13	insulator	insulator		0	*1
14	insulator	insulator		0	*1
15	insulator	insulator		0	*1
16	insulator	insulator		0	*1
17	insulator	insulator		0	*1
18	insulator	insulator		0	*1
19	insulator	insulator		0	*1
20	insulator	semiconductor	+	*1	0
21	insulator	insulator		0	*1
22	insulator	insulator		0	*1
23	insulator	insulator		0	*1

Inst.	Actual	Predicted	Error	Probability	Distribution
24	insulator	insulator		0	*1
25	insulator	insulator		0	*1
26	insulator	insulator		0	*1
27	insulator	insulator		0	*1
28	insulator	insulator		0	*1
29	insulator	insulator		0	*1
30	insulator	insulator		0	*1
31	insulator	insulator		0	*1
32	insulator	insulator		0	*1
33	insulator	insulator		0	*1
34	insulator	insulator		0	*1
35	insulator	insulator		0	*1
36	insulator	insulator		0	*1
37	insulator	insulator		0	*1
38	insulator	insulator		0	*1
39	insulator	insulator		0	*1
40	insulator	insulator		0	*1
41	insulator	insulator		0	*1
42	insulator	insulator		0	*1
43	insulator	insulator		0	*1
44	insulator	insulator		0	*1
45	insulator	insulator		0	*1
46	insulator	insulator		0	*1
47	insulator	insulator		0	*1
48	insulator	insulator		0	*1
49	insulator	insulator		0	*1
50	insulator	insulator		0	*1

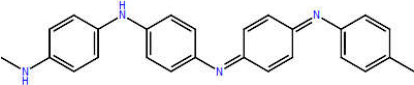
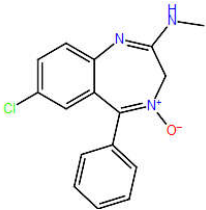
Table 21. Semiconductor prediction output from computational Bayesian Model 2.

Inst.	Actual	Predicted	Error	Probability	Distribution
1	semiconductor	insulator	+	0	*1
2	semiconductor	insulator	+	0	*1
3	insulator	insulator		0	*1
4	insulator	insulator		0	*1
5	insulator	insulator		0	*1
6	insulator	insulator		0	*1
7	insulator	insulator		0	*1
8	insulator	insulator		0	*1
9	insulator	insulator		0	*1
10	insulator	insulator		0	*1
11	insulator	insulator		0	*1
12	insulator	insulator		0	*1
13	insulator	insulator		0	*1
14	insulator	insulator		0	*1
15	insulator	insulator		0	*1
16	insulator	insulator		0	*1

Inst.	Actual	Predicted	Error	Probability	Distribution
17	insulator	insulator		0	*1
18	insulator	insulator		0	*1
19	insulator	insulator		0	*1
20	insulator	semiconductor	+	*1	0
21	insulator	semiconductor	+	*1	0
22	insulator	insulator		0	*1
23	insulator	insulator		0	*1
24	insulator	insulator		0	*1
25	insulator	insulator		0	*1
26	insulator	insulator		0	*1
27	insulator	insulator		0	*1
28	insulator	insulator		0	*1
29	insulator	insulator		0	*1
30	insulator	insulator		0	*1
31	insulator	insulator		0	*1
32	insulator	insulator		0	*1
33	insulator	insulator		0	*1
34	insulator	insulator		0	*1
35	insulator	insulator		0	*1
36	insulator	insulator		0	*1
37	insulator	insulator		0	*1
38	insulator	insulator		0	*1
39	insulator	insulator		0	*1
40	insulator	insulator		0	*1
41	insulator	insulator		0	*1
42	insulator	insulator		0	*1
43	insulator	insulator		0	*1
44	insulator	insulator		0	*1
45	insulator	insulator		0	*1
46	insulator	insulator		0	*1
47	insulator	insulator		0	*1
48	insulator	insulator		0	*1
49	insulator	insulator		0	*1
50	insulator	insulator		0	*1

Tables 20 and 21 represents the output buffer result of virtual screening process for Model 1 and Model 2, where it displays the predicted class i.e. the molecules that are computationally predicted to be organic semiconductors. The screened molecules are displayed in the following Table 22 against each Model.

Table 22. Screened results against Bayesian models (1 and 2)

S.No.	Screened Molecules	N B Models	Band Gap (eV)
1	<p>Emeraldine</p> 	1	3.556
2	<p>Chlordiazepoxide</p> 	1 and 2	4.425

Two Schiff base molecules were prioritized as computationally Bayesian semiconductor actives among the virtual screening library. Out of these, Chlordiazepoxide is active in Model 1 and Model 2 while Emeraldine was active in Model 1 only.

3.3 Conclusion

Bayesian model is widely applied in the pharmaceutical industries in the area of *in silico* drug design due to its simplicity and less computational cost and time. And this approach holds great promise in developing computational models for the development efforts for next-generation materials, such as organic semiconductors. In this work, the virtual screening approach is illustrated as an alternative method for predicting the semiconducting nature of small organic molecules among the class of compounds involving Schiff base. The results presented here demonstrate a powerful approach for exploring the semiconductor nature in the organic semiconductor space. The prospects of predictive material modeling rapidly

accelerate materials discovery, analysis and optimization in the area of small organic semiconductors. And the study of modeling and predicting will reveal structure–property trends and the underlying patterns in an easier way.

The chapter has briefly touched upon the development and virtual screening of computational Bayesian classification models for organic semiconductors. Two models, one Default model (Model 1) and Oversampled model (Model 2) was generated. The purposes of the models were to screen many number of molecules in finding and understanding the semiconductivity that are computationally active. For the model built, custom library was prepared based on the band gap energy involving small organic semiconductors, monomers, n-type and p-type polymers and insulators as non-semiconductors. The models were trained with thousands of molecular descriptors calculated from e-Dragon server (as training set and test set) against class variable semiconductor and non semiconductor in the data mining package WEKA. The predictive models developed are two class problem that were statistically analyzed from various evaluation measures. The two computational models developed are of great importance, since both models were able to screen the relevant molecules even though Model 2 was slightly better in terms of the performance. As a part of the study emeraldine and chlordiazepoxide are reported to be computational organic semiconductors.

CHAPTER 4

PREDICTIVE MODELS BASED ON DECISION TREE ANALYSIS

4. Decision Tree Model (DTM)

This chapter aims to address small data applications of Decision Trees (DTs) models for classification tasks specifically, considering algorithms like Random Forests (RFs) and J48 for the prediction of organic semiconductors from a compound library. As implied by its name, binary DT algorithm follows as a tree-like structure, the parent node is split into two subsets namely child nodes by calculating the best feature split determined by a chosen split criterion. Then the two resulting child nodes become the new parent nodes and are subsequently divided further into two child nodes until all the observations are classified. After the classification, each group member represents more common features as a homogenous set can be realized and evaluated. In our case leaves represent the outcome class labels, i.e. organic semiconductor or non-semiconductor while the branches correspond to conjunctions of input features that resulted in those outcomes.

4.1 DT design in this study

The DT design in the present study was based on the random forest algorithm implemented in WEKA package. In the training process, dataset is recursively divided based on a split criteria until the dataset is split into two homogenous set. The split optimization criterion used in this DT model is the Gini's Diversity Index (GDI), which is a measure of node impurity. The node is considered pure when it contains only observations of one class (either organic semiconductor or non-semiconductor) the GDI of a pure node is equal to 0.¹⁵⁸

4.1.1 Random Forest

The DT design in the present study was based on the random forest algorithm implemented in WEKA package. RF is an ensemble learning method used for classification and regression developed by Breiman. The collection of decision trees is constructed by combined methods from Breiman's bagging sampling approach, and the random selection of features, introduced by Ho, Amit and Geman. Each decision tree in the ensemble is constructed using a sample with replacement from the training data known as bagging.

Statistically, the sample is likely to have about 64% of instances referred to as in-bag instances and the remaining instances 36% as out-of-bag instances.¹⁵⁹ The unlabeled instance of the class label is determined by each tree via majority voting where each classifier casts one vote for its predicted class label, and then the class label with the majority votes is used to classify the instance. The voting is so performed because Random Forest Model (RFM) with a single tree is considerably weak as it is trained on a subset of the dataset and the combination of all trees makes the classifier stronger. In the RFM, random selection of candidate variables ensures a low correlation between trees and thus over-training is prevented and pruning is not performed. Thus, the reduced correlation among trees in the forests helps to achieve improved performance in prediction.¹⁶⁰ The expected error rate of a classifier in new samples is usually estimated by cross-validation method such as leave-one-out or K-fold cross-validation. General cross-validation procedures are unnecessary to predict the classification performance of a given RFM. A cross-validation is already built-in, as each tree in the forest has its own training (bootstrap) and test (OOB) data.¹⁶¹

4.1.1.1 Materials and Methods

The software's and online web servers used for the model build are specified in Chapter 2. The dataset (training set and test set) mentioned for the Bayesian model built are used here also.

4.1.1.2 Experimental Studies

For the model construct classification experiments were done using WEKA version 3.6. We started with an increased heap-size of 4 GB to handle out-of-memory exceptions for large datasets. The training and test set that were used to build Bayesian model were used to build the DT models with class nominal organic semiconductors and non-semiconductors. Table 23 includes various capabilities of class and attributes included in random forest construction while Table 24 displays the various parameters set used for the random forest tree construction.

Table 23. Information on the type of class and attributes for developing Random Forest.

Class	Nominal class, Binary class, Missing class values
Attributes	Nominal attributes, Empty nominal attributes, Unary attributes, Numeric attributes, Date attributes, Missing values, Binary attributes.

Table 24. The parameters set for the construction of Random Forest.

Debug	If set to true, classifier may output additional info to the console.
MaxDepth	The maximum depth of the trees, 0 for unlimited.
NumFeatures	The number of attributes to be used in random selection (see RandomTree).
NumTrees	The number of trees to be generated.
Seed	The random number seed to be used.

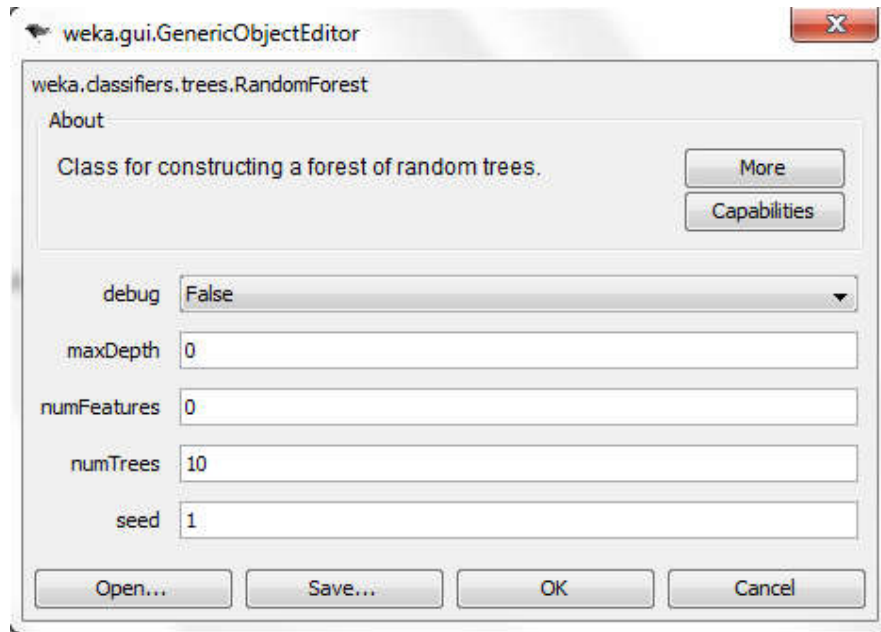


Figure 21. Weka Generic Object Editor.

The models were built from the Weka Generic Object Editor `weka.classifiers.trees.RandomForest` as shown in Figure 21 by providing training set and test set as mentioned above. The models were built after preprocessing and 10 fold cross-validation of the training set. A total of 10 trees were used to construct a Random forest, considering 11 random features. While training, the Out of bag error was found to be 0.1169. After the model construction, the test set was re-evaluated upon the training set to understand the model accuracy and performance. Here also two RF models were constructed, one corresponding to the default dataset (RF Model 1) and the other, the oversampled dataset (RF Model 2). The default Model 1 was trained upon 77 data points with 19 data points test set while Model 2 was trained upon 110 data points with 27 data points as test sample set. All the results of the predictive models (Model 1 and Model 2) were re-evaluated upon the independent test set and various statistical performance matrices have been tabulated from Tables 25-26.

Table 25. The evaluation measures TP, FP, FN, TN, TP rate and FP rate generated by Models 1 and 2

Classifier	TP	TN	FP	FN	TP rate %	FP rate %
RF Model 1	11	7	1	0	100	12.5
RF Model 2	12	15	0	0	100	0

Table 26. The evaluation measures precision, recall, F-measure, ROC, accuracy and kappa generated by Models 1 and 2

Classifier	Precision	Recall	Specificity	BAC	F-measure	ROC	Accuracy	Kappa	MCC
RF Model 1	91.7	100	87.5	93.75	95.7	90.3	94.7368	0.8902	0.8955
RF Model 2	100	100	100	100	100	100	100	1	1

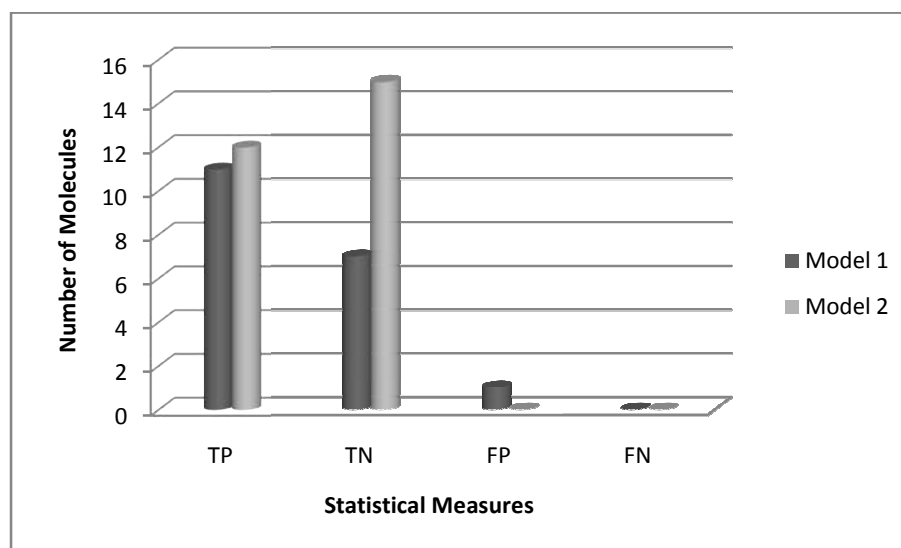


Figure 22 (a). Number of molecules based on confusion matrix

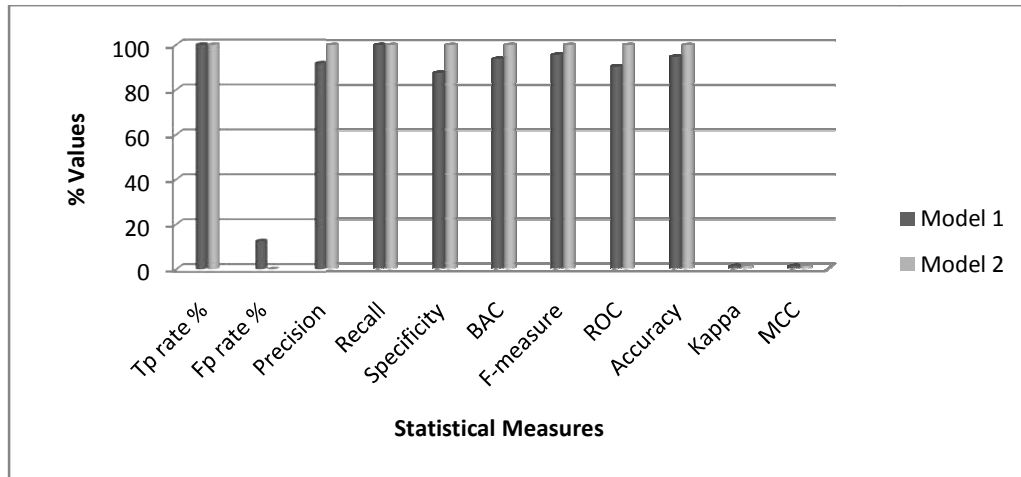


Figure 22 (b). Comparative graph between the various statistical measures of both classification models - Model 1 and Model 2 are presented

4.1.1.3 Results and discussion

The robustness of the RF models (Model 1 and Model 2) was analyzed from the confusion matrix as displayed in Table 27. The performance of the models were selected based on various statistical performance measures like TP Rate, FP Rate, accuracy, precision, recall, F-measure, kappa statistics, ROC area, MCC and BAC. The values of these parameters indicate that Model 2 has outperformed in comparison to Model 1.

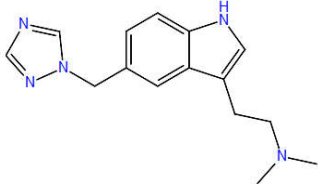
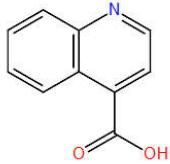
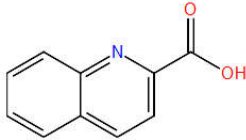
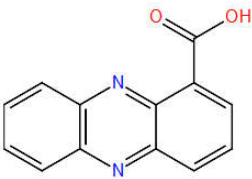
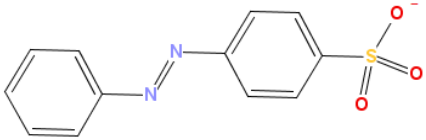
Table 27. Random forest confusion matrix of default and oversampled dataset

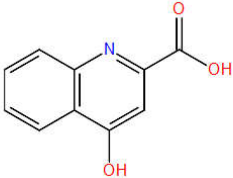
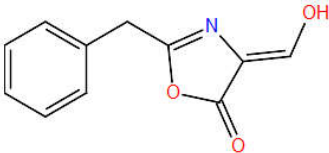
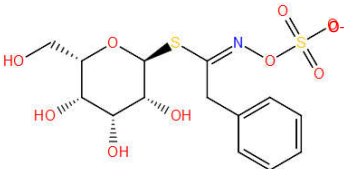
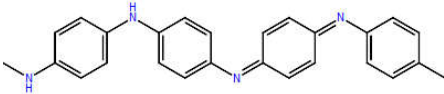
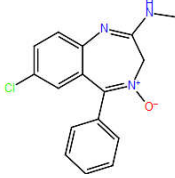
R F Model 1		R F Model 2	
a	b	a	b
←-- classified as		←-- classified as	
11	0	15	0
a = semiconductor		a = insulators	
1	7	0	12
b = insulators		b = semiconductor	

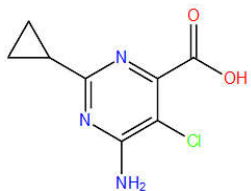
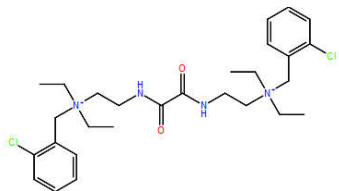
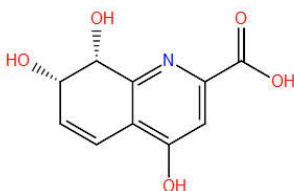
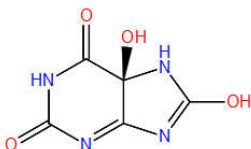
Tables 25-26 summarize the performance measures of the predictive models for Model 1 and Model 2. From our study the FP rate percentage of Model 1 (12.5%) was reduced to 0 in Model 2. The overall robustness of random forest models (R F Model 1 and R F Model 2) classifier was judged from the statistical measure accuracy. Here also the balanced accuracy was accounted for each model from which the model effectiveness was studied as show in Table 26. A Receiver

Operating Characteristic (ROC) curve is a graphical plot of TPR vs. FPR for a binary classification system. ROC space is defined by False Positive Rate and True Positive Rate on X and Y axes respectively. The Area under Curve (AUC) value reported by a ROC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The Receiver Operator Characteristics (ROC) was also compared and evaluated the performance of each model (R F Model 1 and R F Model 2) for their efficiency and robustness. The reliability of the model was also checked from the Kappa value which was found to be 1. The classifier Model 2 showcased greater values in terms of statistical parameters, but it was interesting to note that all the parameters performed well. An extensive analysis of binary classification statistical parameters showed that overall best result was provided by oversampled model. Further, the cross-validation of best models against each of the other dataset also revealed a high predictive accuracy in the same range, substantiating the robustness of the model. During the model development process, a maximum limit of 20% False Positives were allowed or else misclassification cost had to be applied (not used in the current study). Since, both models attained less than the prescribed limit for false positive, cost sensitive analysis was not performed. We have employed a systematic and comprehensive approach to build a supervised classification based predictive models for small organic semiconductor agents from various resources. In contrast with the conventional HOMO-LUMO based methodologies, these models are target-agnostic as they are based on predictive models. Owing to the class imbalance in datasets, introducing sampling techniques led to enhanced sensitivity, specificity and accuracy in the generated models. Figure 22 is the comparative graph between all the statistical measures of both classification RF models-Model 1 and Model 2. Later the screening set that was developed in ARFF was used here also (screening set used against the Bayesian model screen). The WEKA compatibility issue was solved simultaneously followed by virtual screening against RF Model 1 and RF Model 2. The molecules screened against the two models (Model 1 and Model 2) are tabulated in Table 28.

Table 28. Screened results against Random Forest models (1 and 2)

S.No.	Screened Molecules	Models	Band Gap (eV)
1	Rizatriptan 	1,2	4.807
2	Quinoline-4-Carboxylate 	1,2	4.641
3	Quinaldate 	1,2	4.811
4	Phenazine-1-Carboxylate 	1,2	3.799
5	P-Azobenzenesulfonate 	1	4.847

S.No.	Screened Molecules	Models	Band Gap (eV)
6	<p>Kynurenate</p> 	1,2	4.94
7	<p>2-Benzyl-4-Oxidomethylene-5-Oxazolone</p> 	1	5.131
8	<p>Glucotropeolin</p> 	2	3.546
9	<p>Emeraldine</p> 	2	3.556
10	<p>Chlordiazepoxide</p> 	2	4.425

S.No.	Screened Molecules	Models	Band Gap (eV)
11	Aminocyclopyrachlor 	2	4.618
12	Amabenonium 	2	4.621
13	7,8-Dihydro-7,8-Dihydroxykynureate 	2	5.012
14	5-Hydroxyisouric Acid Anion 	2	5.227

A total of fourteen molecules were prioritized as computationally Random Forest semiconductor actives among the virtual screening library. Out of these, thirteen molecules are Schiff base and one azo compound respectively. The two molecules 2-Benzyl-4-Oxidomethylene-5-Oxazolone and P-Azobenzenesulfonate is active only in Model 1 while the rest are active in both Random Forest Models 1 and 2.

4.2 J48 Model

J48 is a slightly modified C4.5 algorithm in WEKA that generates a classification-decision tree for the given dataset by recursive partitioning of data using Depth-first strategy.^{66,162} C4.5 algorithm is an extension of the IDE3 (Iterative

Dichotomiser 3) algorithm, developed by Quinlan Ross (1993). The algorithm creates a small tree and uses divide and conquer approach. Decision trees are grown as proposed by Hunt and his co-workers.¹⁶³ It uses the gain ratio impurity method to evaluate the splitting attribute. Here pruning of the decision tree takes place by replacing the internal node with a leaf node, thereby reducing the error rate.¹⁶⁴ The main advantage of tree pruning is that it reduces misclassification errors, due to noise or too-much detail in the training dataset. Unlike the IDE3, both continuous and categorical attributes in building the decision tree are accepted in C4.5 algorithm. In order to achieve the best splitting attribute the data is sorted at every node of the tree as in the case of IDE3 algorithm. The main disadvantage of this method is the tree depth that matches to the run-time complexity of the algorithm i.e. the tree depth is linked to tree-size and thereby to the number of examples. So, C4.5 algorithm runs slow for noisy and large datasets.

4.2.1 Materials and Methods

The software and online web servers used for the J48 model build are specified in Chapter 2. The dataset (training set and test set) used to build Bayesian model are used here also for the J48 model build.

4.2.2 Experimental Studies

The J48 trees were constructed on the WEKA “explorer” platform. To handle out-of-memory exceptions for the datasets under study, the heap-size was configured to 4 GB. The training set and test that were used to build Bayesian model were used for J48 tree analysis with class nominal semiconductors as actives and non-semiconductors as inactive.

4.2.3 J48 Model Generation

The models were built from the WEKA Generic Object Editor - `weka.classifiers.trees.J48` by providing training set and test set as mentioned in Figure 23. The dataset was randomized, reordered and split into 10 folds of equal size and cross validation was performed. For the model built Generic Object Editor “unpruned” was set to *true* as shown in Figure 23.

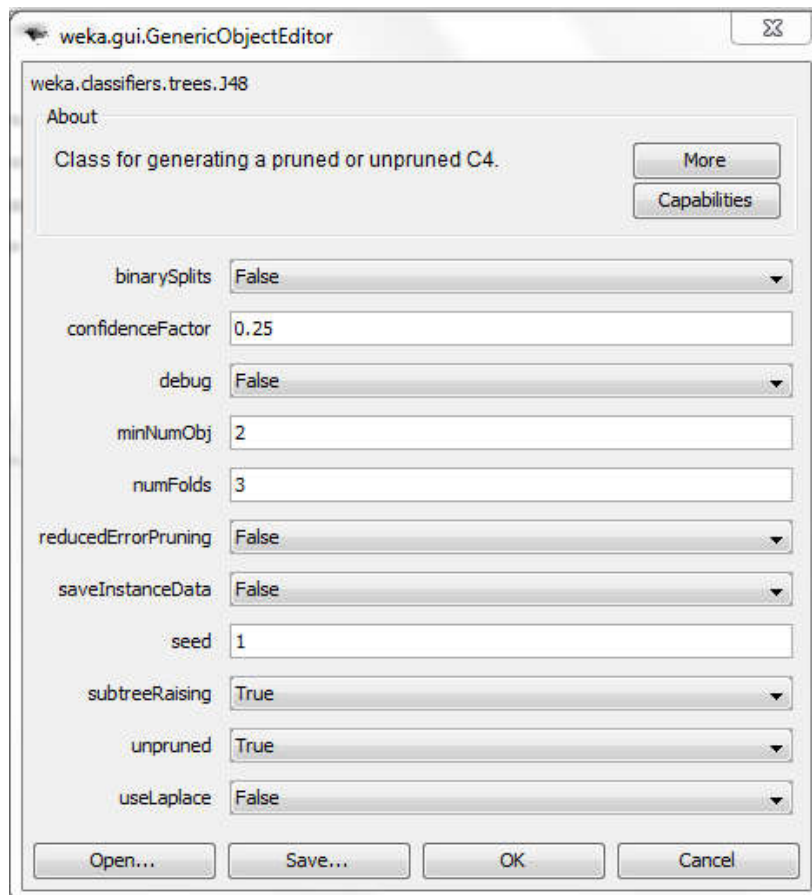


Figure 23. WEKA Generic Object Editor for J48 classifier.

The models were built after preprocessing and 10 fold stratified cross-validation of the training set that ended up with a tree size of seven consisting four leaves. After the construction of J48 Model 1 the test set was re-evaluated upon the training set to understand the model accuracy and performance. Model 2 construction also followed the same method. Here the unpruned tree ended up with a tree size of five consisting of three leaves. J48 tree is visualized from the WEKA console as shown in Figure 24 (a-b). Finally the two models, one corresponding to the default dataset (J48 Model 1) and the oversampled dataset was developed (J48 Model 2). The default J48 Model 1 was constructed using 77 data points and 19 data points as training and test set respectively. While J48 Model 2 was trained upon 110 data points with 27 data points as test sample set.

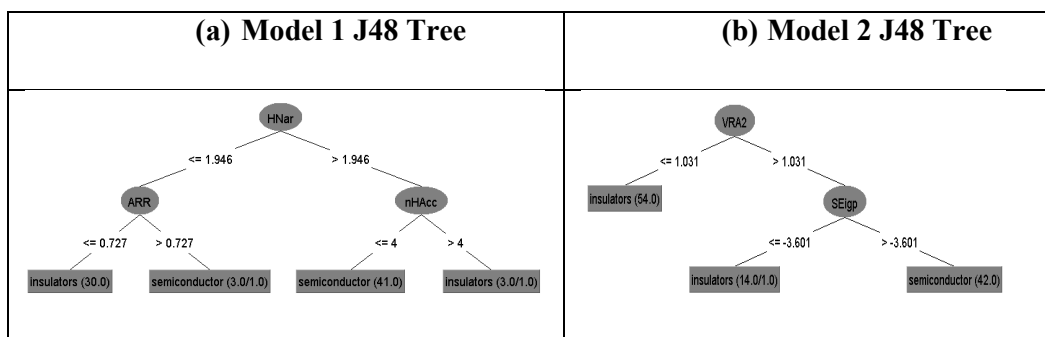


Figure 24. (a) J48 tree developed from the WEKA console for Model 1. (b) J48 tree developed from the WEKA console for Model 2.

All the results of predictive models (J48 Model 1 and J48 Model 2) were re-evaluated upon the independent test set and various statistical performance matrices.

4.2.3.1 Results and discussion

The robustness of the models (J48 Model 1 and J48 Model 2) was analyzed from confusion matrix as shown in Table 29. The performance of the models were selected based on various statistical performance measures like TP Rate, FP Rate, accuracy, precision, recall, F-measure, kappa statistics, ROC area, MCC, BCR, etc. The values of these parameters except FP rate indicate that Model 2 has outperformed in comparison to Model 1. Tables 30-31 summarize the performance measures of the predictive models for Model 1 and Model 2.

Table 29. Confusion matrix for J48 tree Model 1 and Model 2.

J48 Model 1		J48 Model 2	
a	b	a	b
←-- classified as		←-- classified as	
10	1	15	0
a = semiconductor		a = insulators	
1	7	0	12
b = insulators		b = semiconductor	

Table 30. The evaluation measures TP, FP, FN, TN, TP rate and FP rate generated by J48 Models 1 and 2

Classifier	TP	TN	FP	FN	TP rate %	FP rate %
J48 Model 1	10	7	1	1	90.9	12.5
J48 Model 2	12	15	0	0	100	0

Table 31. The evaluation measures Precision, Recall, F-measure, ROC, Accuracy and Kappa generated by J48 Models 1 and 2.

Classifier	Precision	Recall	Specificity	BAC	F-measure	ROC	Accuracy	Kappa	MCC
J48 Model 1	90.9	90.9	87.5	89.2	90.9	88.1	89.4737	0.7841	0.784
J48 Model 2	100	100	100	100	100	100	100	1	1

From our study the FP rate percentage of J48 Model 1 (12.5%) was reduced to 0 in J48 Model 2. While the TP rate for J48 Model 1 was 90.9 which became 100 in J48 Model 2. The reliability of the model was also checked from the evaluation measure accuracy and kappa value which was found to be 1. All the other statistical parameters were also high in comparison to J48 Model 1. The parameter BAC ($\frac{1}{2}$ (Sensitivity + Specificity)) is the mean of specificity and sensitivity introduces a balance among the classification rate of the two classes as displayed in the Table 31. The quality of a binary classification is also measured in terms of Matthews Correlation Coefficient (MCC) which is found to be at the highest is regarded as the balanced measure of a classifier. Figure 25 is the comparative graph between all the performance matrices of both classification J48 models - Model 1 and Model 2. Later the screening set that was used in Bayesian and Random forest model screen was also used here. Here also we encountered the compatibility issue between the training and test set. This was solved simultaneously after which virtual screening against the J48 Model 1 and J48 Model 2 was performed. The molecules screened against the two models (J48 Model 1 and J48 Model 2) are tabulated in Table 32.

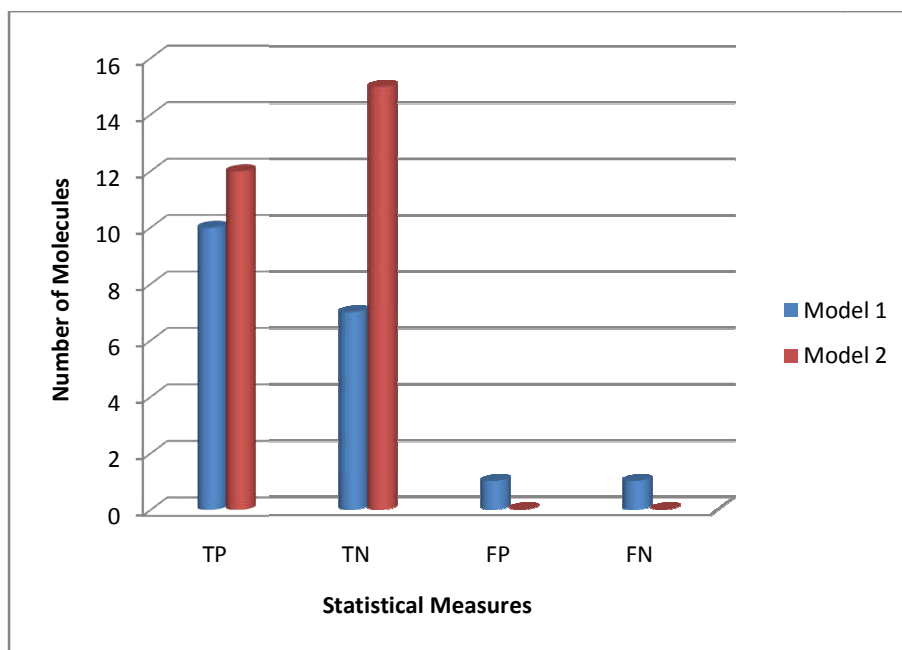


Figure 25 (a). Number of molecules based on confusion matrix

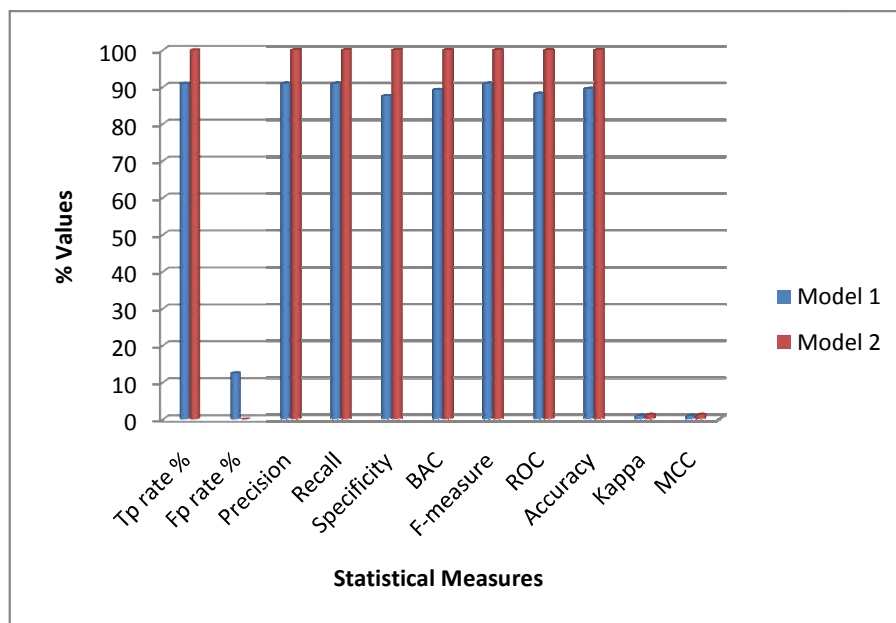
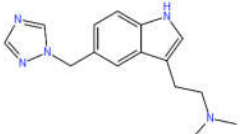

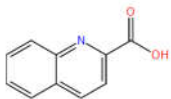
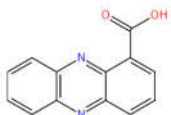
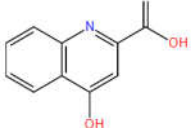
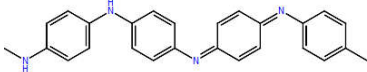
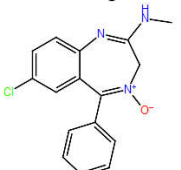


Figure 25 (b). Comparative graph between the various statistical measures of both classification models – J48 Model 1 and J48 Model 2 are presented

4.2.3.2 Screened Results

Table 32. Screened outcome against J48 Model 1 and J48 Model 2

S.No.	Screened Molecules	J48 Models	Band Gap (eV)
1	Rizatriptan 	1 and 2	4.807
2	Quinoline-4-Carboxylate 	1 and 2	4.641
3	Quinaldate 	1 and 2	4.811
4	Phenazine-1-Carboxylate 	1 and 2	3.799
5	Kynurenate 	1	4.940
6	Emeraldine 	1 and 2	3.556
7	Chlordiazepoxide 	2	4.425

A total of seven molecules were prioritized as computationally J48 organic semiconductor actives among the virtual screening library. Out of these, Kynurenate molecule is active in J48 Model 1 and Chlordiazepoxide molecule is active in J48 Model 2 and the rest are active in both J48 Models 1 and 2.

4.3 Conclusion

We explored the potential of DT algorithms Random Forest (RF) and J48 classification models, in the context of small dataset, for outcome prediction in semiconductivity. The ultimate aim was in virtual screening process through decision tree models and compound classification of semiconductor nature. The success rate of DT algorithms in the e-commerce, fraud detection, banking commerce etc is very high in solving high dimensional dataset. The algorithms too are applied in various biological models in high-throughput virtual screening.

In this chapter we developed molecular descriptor based random forest and J48 decision tree models. Further they were used for the virtual screening process against the screening set developed from the database ChEBI. A total of four predictive DT models were generated (RF model 1, RF model 2, J48 model 1 and J48 model 2). The robustness of each models were evaluated from various statistical parameters. The evaluation measures of RF model 1 were higher when compared with J48 model 1 except for the statistical parameter specificity. While the oversampled RF and J48 models outperformed well and the evaluation measures were at the highest. All the predictive models were used for the virtual screening process for finding new computationally active organic semiconductors. As a result we report a total of fourteen computationally predicted organic semiconductors, among them seven molecules were predicted by J48 models and rest of the molecules by random forest models. The developed computational models will be of great importance as they could be utilized for screening various molecules from a broader and larger library.

CHAPTER 5

SUPPORT VECTOR MACHINES-SMO MODELS

5. Sequential Minimal Optimization

SVMs are supervised machine learning algorithms developed by Vapnik and Cortes in the year 1995 for binary property/activity prediction. Typically they facilitate classification task. The classification is achieved by projecting the compound libraries into a high dimensional feature, vector space via a kernel function. The compound classification as in our case as semiconductors and non semiconductors are made linearly separable through a hyperplane an imaginary margin between two classes of compounds as shown in Figure 26. There exist many hyperplanes, but the hyperplane that maximizes the margin between the two classes are chosen by the SVM algorithm and the dataset is separated.^{58 165}

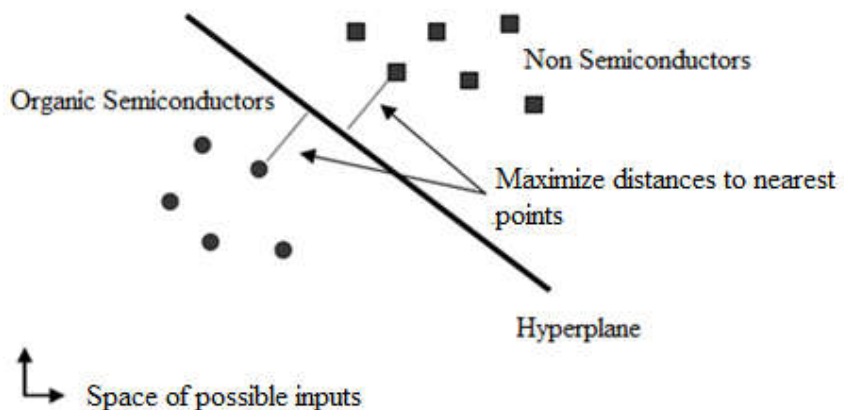


Figure 26. Semiconductors and non semiconductors separated by a hyperplane. Figure adapted from ref.¹⁶⁸

In this chapter, we adopted the algorithm Sequential Minimal Optimization (SMO)^{74,166} for training support vector machines. The training of SVM is based on very large quadratic programming (QP) solutions. SMO breaks the QP problems into a series of small QP problems which is solved analytically. SMO algorithm can

handle large training set easily, since the amount of memory required for the algorithm is linear in the training set size. SMO performs better and fast for linear SVMs and for sparse datasets. Also in real-world sparse datasets, SMO is found to be 1000 times faster than the standard chunking SVM algorithm. The SMO algorithm is conceptually simple, easy to implement, generally faster, and has better-scaling properties for difficult SVM problems than the standard SVM training algorithm. Instead of the numerical quadratic programming as an inner loop used in SVM learning algorithms, SMO uses an analytic QP step for the classification task.¹⁶⁷

5.1 Experimental studies

5.1.1 Materials and Methods

The software's and online web servers used for SMO based predictive models are specified in Chapter 2. For the model development the dataset; training set, test set and screening set are described in Chapter 3.

5.1.2 Generation of SMO Models

For the model construct we implemented John Platt's sequential minimal optimization algorithm for training a support vector classifier in the ML package WEKA version 3.6. John Platt's SMO algorithm globally replaces both missing values and transforms nominal attributes into binary ones. Also by default all attributes are normalized in this algorithm.

During model build, the heap size of the computer was set to 4 GB in order to handle out-of-memory exceptions for the datasets. The dataset was prepared in a systematic manner, by calculating 1448 molecular descriptors (from e-dragon), randomized, divided into two sets; 80% (training set) and 20% (test set), conversion from CSV to ARFF and finally loaded into the WEKA environment. The training and test set were labeled with class nominal organic semiconductors as actives and non-semiconductors as inactive.

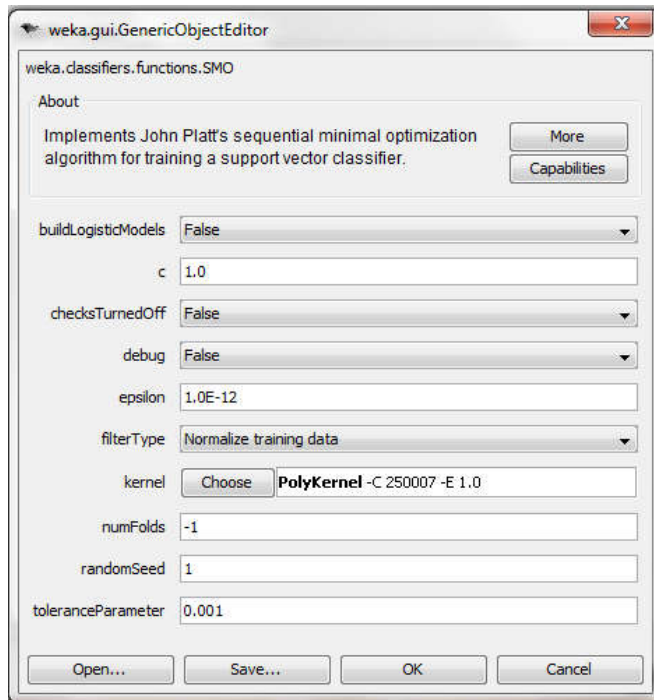


Figure 27. Weka Generic Object Editor for SMO algorithm.

The training set was loaded into the WEKA panel and selected the WEKA-Classifiers-Functions-SMO algorithm for training the dataset as shown in Figure 27. Thereafter test set was re-evaluated upon the model through stratified 10 fold cross validation. The algorithm was set with the default parameters as shown in Figure 27 above with build logistic model made “true” from the WEKA generic object editor. The parameters set for the SMO model built as shown in Table 33 displays the various features included in the algorithm.

Table 33. Parameters set for the SMO model built.

Built Logistic Models	Whether to fit logistic models to the outputs (for proper probability estimates).
C	The complexity parameter C.
Checks Turned Off	Turns time-consuming checks off - use with caution.
Debug	If set to true, classifier may output additional info to the console.
Epsilon	The epsilon for round-off error (shouldn't be changed).
Filter Type	Determines how/if the data will be transformed.
Kernel	Polynomial kernel, $K(x, y) = \langle x, y \rangle^p$ or $K(x, y) = (\langle x, y \rangle + 1)^p$
Num Folds	The number of folds for cross-validation used to generate training data for logistic models (-1 means use training data).
Random Seed	Random number seed for the cross-validation.
Tolerance Parameter	The tolerance parameter (shouldn't be changed).

Here two predictive SMO models were developed; (i) Default Model 1 trained upon 77 molecules with 19 molecules as test samples with class labels as semiconductor and non semiconductors. All the test set samples were re-evaluated upon the training set by ten by ten stratified cross validation. (ii) Model 2 was generated for the oversampled data points of non semiconductors containing a total of 137 molecules. All the results of predictive models (SMO Model 1 and SMO Model 2) upon the test set with its various performance matrices have been tabulated in the results and discussion section.

5.1.3 Results and discussion

5.1.3.1 Robustness of the models

The robustness of the models (SMO Model 1 and SMO Model 2) was analyzed from confusion matrix as shown in Table 34. The performance of the models were analyzed based on various statistical evaluation measures like TP Rate,

FP Rate, accuracy, precision, recall, F-measure, kappa statistics, ROC area, MCC and BCR. The values of these parameters except FP rate indicate that SMO Model 2 has outperformed better than SMO Model 1. Tables 35-36 summarize the performance measures of the predictive models for SMO Model 1 and SMO Model 2.

Table 34. Confusion matrix for SMO Model 1 and Model 2

Model 1				Model 2			
a	b	←-- classified as		a	b	←-- classified as	
8	3	a = semiconductor		15	0	a = insulators	
1	7	b = insulators		0	12	b = semiconductor	

Table 35. The evaluation measures TP, FP, FN, TN, TP rate and FP rate generated by the SMO Models 1 and 2

Classifier	TP	TN	FP	FN	TP rate %	FP rate %
SMO Model 1	8	7	1	3	72.7	12.5
SMO Model 2	12	15	0	0	100	0

Table 36. The evaluation measures Precision, Recall, F-measure, ROC, Accuracy and Kappa generated by SMO Models 1 and 2

Classifier	Precision	Recall	Specificity	BAC	F-measure	ROC	Accuracy	Kappa	MCC
SMO Model 1	88.9	72.7	87.5	80.1	80	95.5	78.9474	0.5824	0.5955
SMO Model 2	100	100	100	100	100	100	100	1	1

From our study the FP rate percentage of SMO Model 1 (12.5%) was reduced to 0 as in the case of SMO Model 2. The reliability of the model was also checked from the Kappa value which was found to be 1 in the case of SMO Model 2. The classifier Model 2 was found to have performed better on comparison with SMO Model 1. An extensive analysis of the binary classification of statistical

parameters showed that overall best result was provided by the oversampled SMO Model 2. Figure 28 is the comparative graph between all the performance matrices of both classification SMO Models 1 and 2.

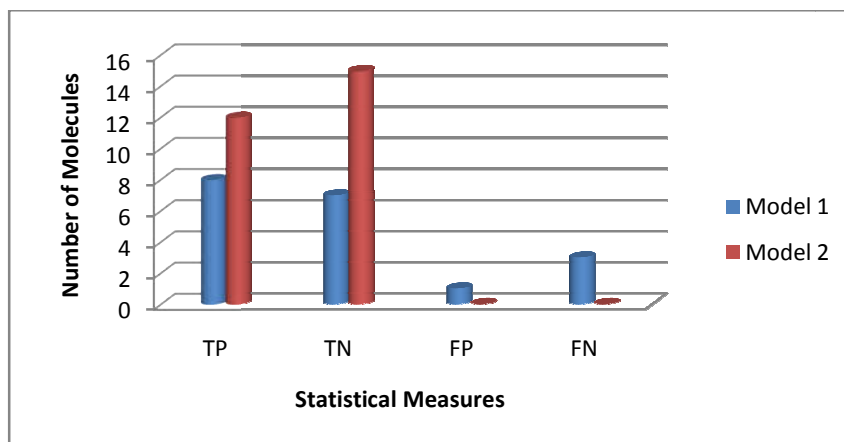


Figure 28 (a). Number of molecules based on confusion matrix

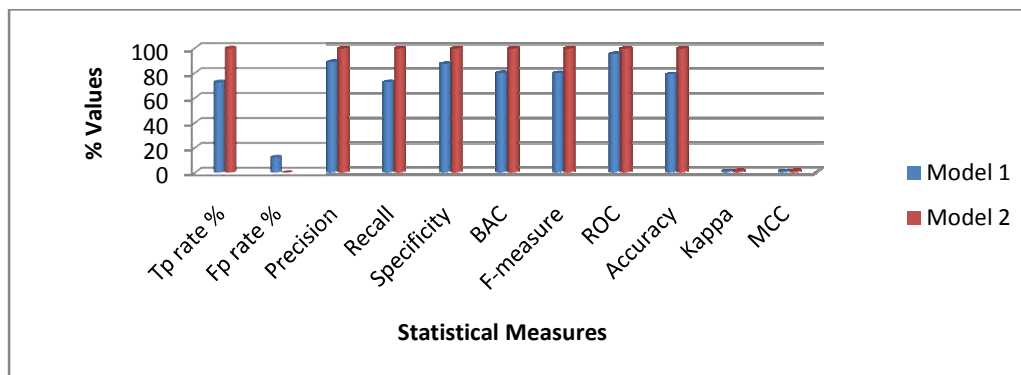


Figure 28 (b). Comparative graph between the various statistical measures of SMO classification models: SMO Model 1 and SMO Model 2 are presented

A Receiver Operating Characteristic (ROC) curve is a graphical plot of TPR vs. FPR for a binary classification system as shown in the Figure 29 (a-b). ROC space is defined by FPR and TPR on X and Y axes respectively. The Area under Curve (AUC) value reported by a ROC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. As per the concern of ROC area SMO algorithms are having different

values i.e. a higher ROC area for SMO Model 2 is better in the aspect of model prediction. It examines the performance of the predictive semiconductor models as an additional way besides the confusion matrix. It examines in respect to FP rate and TP rate as displayed in Figure 29.

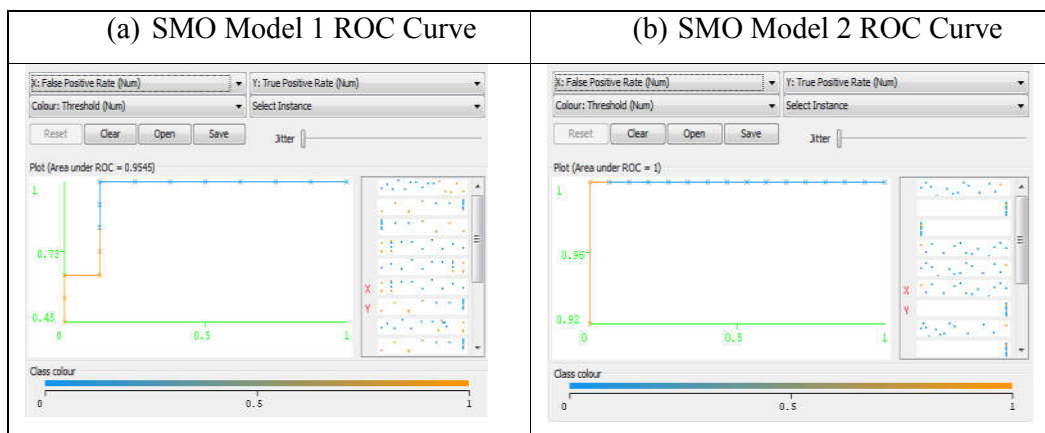


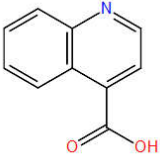
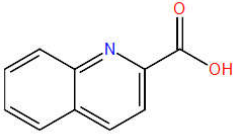
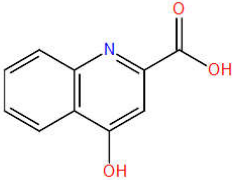
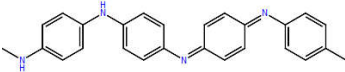
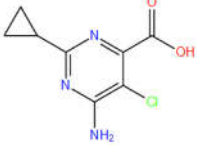
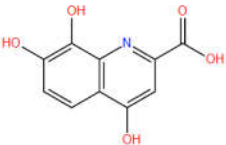
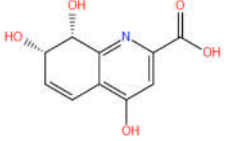
Figure 29. Selected ROC: (a) ROC curve of SMO Model 1 (b) ROC curve of SMO Model 2

Since, both models attained values less than the upper limit for false positive, cost sensitive analysis was not performed. A systematic and comprehensive approach was employed to build supervised classification based SMO predictive models for small organic semiconductors. Owing to the class imbalance in datasets, introducing sampling techniques led to enhanced sensitivity, specificity and accuracy in the models generated.

5.1.3.2 Virtual Screening

The screening set was prepared from the ChEBI database as mentioned in Chapter 3. All the molecules were prepared in a systematic way starting from molecular modeling, geometry optimization, molecular descriptor calculation (1448 from E-Dragon server), conversion from CSV to ARFF and were checked for the compatibility issue. All the molecules were screened against the two models (SMO Model 1 and SMO Model 2). The screened molecules are displayed in Table 37. As a result, SMO predictive models produced a total of seven molecules that are computationally organic semiconductors.

Table 37. Screened molecules from SMO Model 2

S.No.	Screened Molecules	SMO Models	Band Gap (eV)
1	<p>Quinoline-4-Carboxylate</p> 	Model 2	4.641
2	<p>Quinaldate</p> 	Model 2	4.811
3	<p>Kynureate</p> 	Model 2	4.940
4	<p>Emeraldine</p> 	Model 2	3.556
5	<p>Aminocyclopyrachlor</p> 	Model 2	4.618
6	<p>7,8-dihydroxykynureate</p> 	Model 2	4.315
7	<p>7,8-Dihydro-7,8-Dihydroxykynureate</p> 	Model 2	5.012

The prioritized seven molecules were computationally SMO Model 2 semiconductor actives. While no molecules were prioritized from SMO Model 1.

5.1.4 Conclusion

SMO prediction engine has a good generalization performance on various real world problems like handwritten character recognition, face detection, pedestrian detection, text categorization and other biological fields like metabolic pathways identification.¹⁶⁸ The success of the predictive approach has emphasized to extend the algorithmic prediction in the case of organic semiconductivity. In this study SMO based computational semiconductor models were developed. The developed SMO models (SMO Model 1 and SMO Model 2) are descriptor based classification models. Molecules from chemical database (ChEBI) were screened against the models and some molecules were prioritized as computationally active semiconductors. The developed models were based on function (polynomial kernel) where its attributes corresponds to the molecular and electronic descriptors. Two models were developed; default model (SMO Model 1) and oversampled model (SMO Model 2) and were analyzed based on the various statistical parameters. SMO Model 2 outperformed with high accuracy. As a part of the study seven molecules are reported to be computationally active organic semiconductors.

CHAPTER 6

PATTERN SEARCH FOR ORGANIC SEMICONDUCTORS

6. Maximum Common Substructure

Maximum Common Substructure (MCS) is a chemical similarity searching program looking for the largest substructure that appears in two or more chemical structures. The algorithm makes use of the graph theory as representation of molecular structures and extracts maximum common substructure in the form of connected or disconnected graph.¹⁶⁹ MCS has a wide application in filtering and prioritizing large datasets of molecules and often used as a search tool for finding patterns/substructure of structurally related drugs, which are likely to be an important fragment of their biological activities.¹⁷⁰ In our study the relevance of a similarity measure was exclusively based on MCS, implemented using CANVAS cheminformatics suite software. The program searches for the largest common part (substructures) among the set of molecules of interest.

6.1 Materials and Methods

The software used for MCS was CANVAS, a comprehensive cheminformatics suite from Schrodinger package. The software details are furnished in Chapter 2 materials and methods.


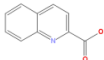
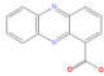
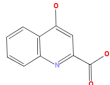
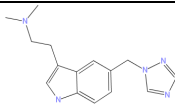
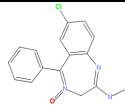
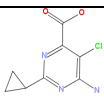
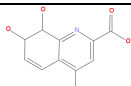
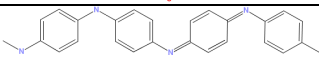
6.2 Experimental Studies

The work started by importing input molecules that are virtually screened on the various classifiers like Naïve Bayes, Random forest, J48 and SMO into the canvas software. For the pattern search, Model 2 was selected from the above mentioned WEKA learning schemes as displayed in Table 38. And the molecules shown in Table 39 are the molecules which were predicted to be computationally semiconductor active.

Table 38. Computationally predicted semiconducting molecules by SMO, Random Forest, and J48

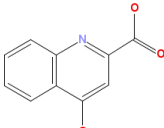
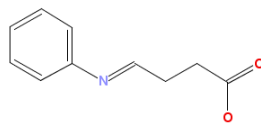
SMO	Random Forest	J48
7,8-Dihydro-7,8-Dihydroxykynurenate	Chlordiazepoxide	Chlordiazepoxide
Aminocyclopyrachlor	Emeraldine	Emeraldine
Chlordiazepoxide	Kynurenic acid	phenazine-1-carboxylic acid
Emeraldine	Quinaldic acid	Quinaldic acid
Kynurenic acid	Quindine-4-carboxylic acid	Quindine-4-carboxylic acid
Quinaldic acid	Rizatriptan	Rizatriptan
Quindine-4-carboxylic acid		

Table 39. Totally predicted screened molecules by Naïve Bayes, Random Forest, SMO and J48 by Model 2

Screened molecules	Molecular Structure
Quindine-4-carboxylic acid	
Quinaldic acid	
Phenazine-1-carboxylic acid	
Kynurenic acid	
Rizatriptan	
Chlordiazepoxide	
Aminocyclopyrachlor	
7,8-Dihydro-7,8-Dihydroxykynurenate	
Emeraldine	

All the molecules were imported in sdf format in the canvas panel. The task of finding the maximum common substructure (MCS) from the above molecules was run from the *Maximum Common Substructure* dialog box from the applications menu in the Canvas software. The search doesn't encompass all the structures in the set but a minimum number of structures were specified that were matched for the substructure and a maximum number that must match. The criteria were set in such a manner i.e. if the minimum is greater than the number of structures in the set, all structures must match. Or else, a series of substructures (MCS groups) were found for each number of structures from the minimum to the maximum. The application was run and as a result SMiles ARbitrary Target Specification (SMARTS)¹⁷¹ strings for the substructures, the number of substructures in each group, and the membership of the groups was visualized in the canvas panel. The matching was performed from the definitions provided in the Atom/bond typing list in the application menu. We have selected the scheme "Atoms distinguished by atomic number and bond order, aromaticity." And the patterns were visualized from the cluster in the canvas panel. The molecules clustered for the common substructure led to the identification of two new patterns (SMARTS) for computationally active semiconductors as shown in Table 40.

Table 40. Common Substructures clustered from SMO, Random Forest and J48

SMO, Random Forest	J48
 <chem>Oc1cc(C(=O)O)nc(c12)cccc2</chem>	 <chem>c1cccc1ncccC(=O)O</chem>

6.3 Conclusion

The MCS has shown to be effective in the study and pattern identification of chemical compounds. The maximum common substructure (MCS) approach provides a more promising and flexible alternative for predicting patterns for semi-conductive compounds. And from the study we developed two patterns configuring the semiconducting nature which was filtered through various ML algorithms.

REFERENCES

- ¹ Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21*.
- ² Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *WIREs Comput. Mol. Sci.* **2014**, *4*, 468–481.
- ³ Nantasenamat, C.; Isarankura-na-ayudhya, C.; Naenna, T. Review Article: A PRACTICAL OVERVIEW OF QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP. *EXCLI J.* **2009**, *8*, 74–88.
- ⁴ Begam, B. F.; Kumar, J. S. Computer Assisted QSAR/QSPR Approaches - A Review. *Indian J. Sci. Technol.* **2016**, *9*.
- ⁵ Medina-Franco, J. L. Advances in Computational Approaches for Drug Discovery Based on Natural Products. *Rev. Latinoam. Quim.* **2013**, *41*, 95–110.
- ⁶ Gasteiger, Johann, T. E. *Cheminformatics: A Textbook*; Wiley-VCH Verlag GmbH & Co. KGaA, 2003.
- ⁷ Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21*.
- ⁸ Chen, W. L. Chemoinformatics: Past, Present, and Future. *J. Chem. Inf. Model.* **2006**, *46*, 2230–2255.
- ⁹ Selassie, C.; Verma, R. P. *History of Quantitative Structure–Activity Relationships*, 7th ed.; Rotella, D. J. A. and D. P., Ed.; John Wiley & Sons, 2010; Vol. 1.
- ¹⁰ Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *WIREs Comput. Mol. Sci.* **2014**, *4*, 468–481.
- ¹¹ Bhalerao, S. A.; Verma, D. R.; Teli, N. C. Chemoinformatics: The Application of Informatics Methods to Solve Chemical Problems. *Res. J. Pharm. , Biol. Chem. Sci.* **2013**, *4*, 475–499.
- ¹² Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.
- ¹³ Han, L.; Wang, Y.; Bryant, S. H. A Survey of across-Target Bioactivity Results of Small Molecules in PubChem. *Bioinformatics* **2009**, *25*, 2251–2255.

-
- ¹⁴ Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955–D963.
- ¹⁵ Pence, H. E.; Williams, A. Chemspider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87*, 1123–1124.
- ¹⁶ Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.
- ¹⁷ Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, 1100–1107.
- ¹⁸ Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S.; Brudz, S.; Sullivan, J. P.; Muhlich, J.; Serrano, M.; Ferraiolo, P.; Tolliday, N. J.; Schreiber, S. L.; Clemons, P. A. ChemBank: A Small-Molecule Screening and Cheminformatics Resource Database. *Nucleic Acids Res.* **2008**, *36*.
- ¹⁹ David S. Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang and Jennifer Woolsey David S. Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, J. W. DrugBank: A Comprehensive Resource for *in silico* Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672.
- ²⁰ Degtyarenko, K.; De matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; Mcnaught, A.; Alcántara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: A Database and Ontology for Chemical Entities of Biological Interest. *Nucleic Acids Res.* **2007**, *36*, 344–350.
- ²¹ Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21*.
- ²² Firdaus Begam, B.; Satheesh Kumar, J. A Study on Cheminformatics and Its Applications on Modern Drug Discovery. *Procedia Eng.* **2012**, *38*, 1264–1275.
- ²³ Anderson, A. C. The Process of Structure-Based Drug Design. *Chem. Biol.* **2003**, *10*, 787–797.
- ²⁴ Lee, C. H.; Huang, H. C.; Juan, H. F. Reviewing Ligand-Based Rational Drug Design: The Search for an ATP Synthase Inhibitor. *Int. J. Mol. Sci.* **2011**, *12*, 5304–5318.

-
- 25 Aparoy, P.; Reddy, K. K.; Reddanna, P. Structure and Ligand Based Drug Design Strategies in the Development of Novel 5-LOX Inhibitors. *Curr Med Chem* **2012**, *19*, 3763–3778.
- 26 van de Waterbeemd, H.; Gifford, E. ADMET *in silico* Modelling: Towards Prediction Paradise? *Nat. Rev. Drug Discov.* **2003**, *2*, 192–204.
- 27 Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21*.
- 28 Ghosh, S. K. Nobel Prize in Chemistry 2013: Chemistry in Cyberspace. *Curr. Sci.* **2013**, *105*, 1455–1456.
- 29 Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *WIREs Comput. Mol. Sci.* **2014**, *4*, 468–481.
- 30 Engel, T. Basic Overview of Chemoinformatics. *J. Chem. Inf. Model.* **2006**, *46*, 2267–2277.
- 31 Ghosh, S. K. Nobel Prize in Chemistry 2013: Chemistry in Cyberspace. *Curr. Sci.* **2013**, *105*, 1455–1456.
- 32 Gasteiger, Johann, T. E. *Cheminformatics: A Textbook*; Wiley-VCH Verlag GmbH & Co. KGaA, 2003.
- 33 Maksood, F. Z.; Achuthan, G.; Lecturer, S. Analysis of Data Mining Techniques and Its Applications. *Int. J. Comput. Appl.* **2016**, *140*, 6–14.
- 34 Fawagreh, K.; Gaber, M. M.; Elyan, E. Random Forests: From Early Developments to Recent Advancements. *Syst. Sci. Control Eng.* **2014**, *2*, 602–609.
- 35 Maksood, F. Z.; Achuthan, G.; Lecturer, S. Analysis of Data Mining Techniques and Its Applications. *Int. J. Comput. Appl.* **2016**, *140*, 6–14.
- 36 David L. Olson, D. D. *Advanced Data Mining Techniques*; David L. Olson, D. D., Ed.; Springer-Verlag Berlin Heidelberg, 2008.
- 37 Festus Ayetiran, E.; Barnabas Adeyemo, A. A Data Mining-Based Response Model for Target Selection in Direct Marketing. *Int. J. Inf. Technol. Comput. Sci.* **2012**, *4*, 9–18.
- 38 Elsalamony, H. A. Bank Direct Marketing Analysis of Data Mining Techniques. *Int. J. Comput. Appl.* **2014**, *85*, 975–8887.
- 39 Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In *Int Conf on Knowledge Discovery and Data Mining*; 1996; pp 82–88.

-
- 40 Özsoy, S.; Gümüş, G.; Khalilov, S. C4.5 Versus Other Decision Trees: A Review. *Comput. Eng. Appl.* **2015**, *4*, 2252–4274.
- 41 Jamal, S.; Scaria, V. Cheminformatic Models Based on Machine Learning for Pyruvate Kinase Inhibitors of *Leishmania Mexicana*. *BMC Bioinformatics* **2013**, *14*, 329.
- 42 Gaba, S.; Jamal, S.; Open Source Drug Discovery Consortium; Scaria, V. Cheminformatics Models for Inhibitors of *Schistosoma Mansoni* Thioredoxin Glutathione Reductase. *ScientificWorldJournal*. **2014**, *2014*, 9.
- 43 Zaveri, Samiksha H, N. J. A Comparative Study of Data Analysis Techniques in the Domain of Medicative Care for Disease Predication. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*, 564–566.
- 44 Anita Devi, J. K. A Survey on Data Mining and Its Current Research Directions. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*, 23–27.
- 45 Pappa, G. L.; Freitas, A. *Automating the Design of Data Mining Algorithms*; Springer-Verlag Berlin Heidelberg 2010, 2010.
- 46 Barai, A. K. Performance Based Association Rule-Mining Technique Using Genetic Algorithm. *J. Ind. Intell. Inf.* **2015**, *3*, 114–118.
- 47 Karim, M.; Rahman, R. M. Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. *J. Softw. Eng. Appl.* **2013**, *6*, 196–206.
- 48 Fawagreh, K.; Gaber, M. M.; Elyan, E. Random Forests: From Early Developments to Recent Advancements. *Syst. Sci. Control Eng.* **2014**, *2*, 602–609.
- 49 Ian H.Witten, E.; Frank. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Elsevier, 2005.
- 50 Hall M, Eibe F, Holmes G, Pfahringer B, et al. The WEKA Data Mining Software: An Update. *SIGKDD Explor.* **2009**, *11*, 10–18.
- 51 Al Ghoson, A. M. Decision Tree Induction & Clustering Techniques in SAS Enterprise Miner, SPSS Clementine, and IBM Intelligent Miner-a Comparative Analysis. *Int. J. Manag. Inf. Syst.* **2010**, *14*, 57.
- 52 Stachová, M.; Sobíšek, L. Data Mining Classification Methods Applied in Drug Design. *Int. J. Medical, Heal. Biomed. Bioeng. Pharm. Eng.* **2012**, *6*, 122–125.
- 53 Bajorath, J. Improving Data Mining Strategies for Drug Design. *Futur. Med. Chem.* **2014**, *6*, 255–257.

-
- 54 Lavecchia, A. Machine-Learning Approaches in Drug Discovery: Methods and Applications. *Drug Discov. Today* **2015**, *20*, 318–331.
- 55 David L. Olson, D. D. *Advanced Data Mining Techniques*; David L. Olson, D. D., Ed.; Springer-Verlag Berlin Heidelberg, 2008.
- 56 Karim, M.; Rahman, R. M. Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. *J. Softw. Eng. Appl.* **2013**, *6*, 196–206.
- 57 Melville, J. L.; Burke, E. K.; Hirst, J. D. Machine Learning in Virtual Screening. *Comb. Chem. High Throughput Screen.* **2009**, *12*, 332–343.
- 58 Helgee, E. A. Improving Drug Discovery Decision Making Using Machine Learning and Graph Theory in QSAR Modeling, Gothenburg, 2010.
- 59 Mitchell, M. W. Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters. *Open J. Stat.* **2011**, *1*, 205–211.
- 60 Ivanciuc, O. Applications of Support Vector Machines in Chemistry. In *Reviews in computational chemistry*; Kenny B. Lipkowitz, T. R. C., Ed.; Wiley-VCH, John Wiley & Sons, 2007; Vol. 23, pp 291–400.
- 61 Hecht, D. Applications of Machine Learning and Computational Intelligence to Drug Discovery and Development. *Drug Dev. Res.* **2011**, *72*, 53–65.
- 62 Jamal, S.; Periwai, V.; Scaria, V. Predictive Modeling of Anti-Malarial Molecules Inhibiting Apicoplast Formation. *BMC Bioinformatics* **2013**, *14*, 55.
- 63 Periwai, V.; Rajappan, J. K.; Jaleel, A. U.; Scaria, V. Predictive Models for Anti-Tubercular Molecules Using Machine Learning on High-Throughput Biological Screening Datasets. *BMC Res. Notes* **2011**, *4*, 504.
- 64 Zgurovsky, M. Z.; Zaychenko, Y. P. *The Fundamentals of Computational Intelligence: System Approach*; Springer International Publishing Switzerland, 2017; Vol. 652.
- 65 Özsoy, S.; Gümüş, G.; Khalilov, S. C4.5 Versus Other Decision Trees: A Review. *Comput. Eng. Appl.* **2015**, *4*, 2252–4274.
- 66 Zhao, Y.; Zhang, Y. Comparison of Decision Tree Methods for Finding Active Objects. *Adv. Sp. Res.* **2008**, *41*, 1955–1959.
- 67 Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
- 68 Wen, L.; Li, Q.; Li, W.; Cai, Q.; Cai, Y.-M. A QSAR Study Based on SVM for the Compound of Hydroxyl Benzoic Esters. *Bioinorg. Chem. Appl.* **2017**, *2017*, 1–10.

-
- 69 Lu, S.-X. L. S.-X.; Meng, J. M. J.; Cao, G.-E. C. G.-E. Support Vector Machine Based on a New Reduced Samples Method. In *Proceedings of the Ninth International Conference on Machine Learning and Cybernetics*; 2010; Vol. 3, pp 11–14.
- 70 Lucieer, A. Visualization of Hyperplanes for SVM Classification. *Int. Geosci. Remote Sens. Symp.* **2007**, 2034–2035.
- 71 Pahwa, S.; Sinwar, D. Comparison Of Various Kernels Of Support Vector Machine. *Int. J. Res. Appl. Sci. Eng. Technol.* **2015**, 3, 532–536.
- 72 Zanaty, E. A. Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in Data Classification. *Egypt. Informatics J.* **2012**, 13, 177–183.
- 73 Prati, R. C.; Batista, G.; Silva, D. F. Class Imbalance Revisited: A New Experimental Setup to Assess the Performance of Treatment Methods. *Knowl. Inf. Syst.* **2015**, 45, 247–270.
- 74 Yang, C. Y.; Su, K. H.; Jan, G. E. An Elaboration of Sequential Minimal Optimization for Support Vector Regression. In *2014 IEEE International Conference on System Science and Engineering (ICSSE)*; 2014; pp 88–93.
- 75 Krstajic, D.; Buturovic, L. J.; Leahy, D. E.; Thomas, S. Cross-Validation Pitfalls When Selecting and Assessing Regression and Classification Models. *J. Cheminform.* **2014**, 6, 1–15.
- 76 McDonagh, J. L.; Nath, N.; De Ferrari, L.; Van Mourik, T.; Mitchell, J. B. O. Uniting Cheminformatics and Chemical Theory to Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Inf. Model.* **2014**, 54, 844–856.
- 77 Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *WIREs Comput. Mol. Sci.* **2014**, 4, 468–481.
- 78 Hecht, D. Applications of Machine Learning and Computational Intelligence to Drug Discovery and Development. *Drug Dev. Res.* **2011**, 72, 53–65.
- 79 Chawla, N. V. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*; Oded Z. Maimon, L. R., Ed.; Springer Science & Business Media, 2005; pp 853–867.
- 80 Schierz, A. C. Virtual Screening of BioAssay Data. *J. Cheminform.* **2009**, 1, 1–12.
- 81 Klepsch, F.; Vasanathan, P.; Ecker, G. F. Ligand and Structure-Based Classification Models for Prediction of P-Glycoprotein Inhibitors. *J. Chem. Inf. Model.* **2014**, 54, 218–229.

-
- 82 Hajjo, R.; Grulke, C. M.; Golbraikh, A.; Setola, V.; Huang, X. P.; Roth, B. L.; Tropsha, A. Development, Validation, and Use of Quantitative Structure-Activity Relationship Models of 5-Hydroxytryptamine (2B) Receptor Ligands to Identify Novel Receptor Binders and Putative Valvulopathic Compounds among Common Drugs. *J. Med. Chem.* **2010**, *53*, 7573–7586.
- 83 Correa, M.; Bielza, C.; Pamies-Teixeira, J. Comparison of Bayesian Networks and Artificial Neural Networks for Quality Detection in a Machining Process. *Expert Syst. Appl.* **2009**, *36*, 7270–7279.
- 84 Periwai, V.; Kishtapuram, S.; Scaria, V. Computational Models for in-Vitro Anti-Tubercular Activity of Molecules Based on High-Throughput Chemical Biology Screening Datasets. *BMC Pharmacol.* **2012**, *12*, 1.
- 85 Periwai, V.; Rajappan, J. K.; Jaleel, A. U.; Scaria, V. Predictive Models for Anti-Tubercular Molecules Using Machine Learning on High-Throughput Biological Screening Datasets. *BMC Res. Notes* **2011**, *4*, 504.
- 86 Seal, A.; Passi, A.; Jaleel, U. A.; Wild, D. J. In-Silico Predictive Mutagenicity Model Generation Using Supervised Learning Approaches. *J. Cheminform.* **2012**, *4*, 10.
- 87 Correa, M.; Bielza, C.; Pamies-Teixeira, J. Comparison of Bayesian Networks and Artificial Neural Networks for Quality Detection in a Machining Process. *Expert Syst. Appl.* **2009**, *36*, 7270–7279.
- 88 Breslow, R.; et. al. *Beyond the Molecular Frontier*; 2003.
- 89 Young, D. C. *COMPUTATIONAL CHEMISTRY: A Practical Guide for Applying Techniques to Real-World Problems*; 2001; Vol. 9.
- 90 Gupta, M. C. *Atomic And Molecular Spectroscopy*; New Age International, 2007.
- 91 Khilari, Sunil, D. S. K. Representing of Problem , Solution and Implementation Spaces with Interrelated Attributes for Developing Knowledge Management Base in Computational Chemistry Area. *Int. J. New Innov. Eng. Technol.* **2016**, *4*, 1–5.
- 92 Dorsett, H.; White, A. Overview of Molecular Modelling and *Ab initio* Molecular Orbital Methods SuiTable for Use with Energetic Materials, 2000, 45.
- 93 Errol Lewars. Introduction to the Theory and Applications of Molecular and Quantum Mechanics. In *Computational Chemistry*; Kluwer Academic Publishers, 2003; p 664.
- 94 Johansson M.P., Kaila V.R.I., S. D. *Ab initio*, Density Functional Theory, and Semi-Empirical Calculations. In *Biomolecular Simulations. Methods in*

-
- Molecular Biology (Methods and Protocols)*; Monticelli L., S. E., Ed.; Springer, 2013; Vol. 924.
- 95 Rogers, D. W. *Computational Chemistry Using the PC*, 3rd ed.; John Wiley & Sons, 2003.
- 96 Errol G. Lewars. *Molecular Mechanics*. In *Computational Chemistry*; Springer, 2004.
- 97 Johann Gasteiger, T. E. *Chemoinformatics: A Textbook*; John Wiley & Sons, 2006.
- 98 Olasunkanmi, L. O.; Ige, J.; Ogunlusi, G. O. Theoretical Study of the Molecular Geometries, Electronic and Thermodynamic Properties of Chlorinated Dipyrindo-(3,2-a:2,3-c)-Phenazine. *J. Chem.* **2013**, *2013*, 1–7.
- 99 El-shamy, O. A. A. Semiempirical Theoretical Studies of 1,3-Benzodioxole Derivatives as Corrosion Inhibitors. *Int. J. Corros.* **2017**, *2017*, 10.
- 100 Bruna-Larenas, T.; Gómez-Jeria, J. S. A DFT and Semiempirical Model-Based Study of Opioid Receptor Affinity and Selectivity in a Group of Molecules with a Morphine Structural Core. *Int. J. Med. Chem.* **2012**, *2012*, 1–16.
- 101 Shallangwa, G. A.; Uzairu, A.; Ajibola, V. O.; Abba, H. MNDO and DFT Computational Study on the Mechanism of the Oxidation of 1,2-Diphenylhydrazine by Iodine. *ISRN Phys. Chem.* **2014**, *2014*, 8.
- 102 Bouzzine, S. M.; Salgado-Morán, G.; Hamidi, M.; Bouachrine, M.; Pacheco, A. G.; Glossman-Mitnik, D. DFT Study of Polythiophene Energy Band Gap and Substitution Effects. *J. Chem.* **2015**, *2015*.
- 103 Tetko, I. V.; Lowe, D.; Williams, A. J. The Development of Models to Predict Melting and Pyrolysis Point Data Associated with Several Hundred Thousand Compounds Mined from PATENTS. *J. Cheminform.* **2016**, *8*, 1–18.
- 104 David S. Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang and Jennifer Woolsey David S. Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, J. W. DrugBank: A Comprehensive Resource for *in silico* Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672.
- 105 <https://www.chemexper.com/>
- 106 Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual Computational Chemistry

-
- Laboratory - Design and Description. *J. Comput. Aided. Mol. Des.* **2005**, *19*, 453–463.
- 107 Walker, T.; Grulke, C. M.; Pozefsky, D.; Tropsha, A. Chembench: A Cheminformatics Workbench. *Bioinformatics* **2010**, *26*, 3000–3001.
- 108 Ruusmann, V.; Sild, S.; Maran, U. QSAR DataBank - An Approach for the Digital Organization and Archiving of QSAR Model Information. *J. Cheminform.* **2014**, *6*, 1–17.
- 109 Ruusmann, V.; Sild, S.; Maran, U. QSAR DataBank Repository: Open and Linked Qualitative and Quantitative Structure–activity Relationship Models. *J. Cheminform.* **2015**, *7*, 32.
- 110 Novotarskyi, S.; Sushko, I.; Körner, R.; Kumar, A.; Rupp, M.; Prokopenko, V.; Tetko, I. OCHEM - on-Line CHEmical Database & Modeling Environment. In *Journal of Cheminformatics*; 2010; Vol. 2, p P5.
- 111 Vorberg, S.; Tetko, I. V. Modeling the Biodegradability of Chemical Compounds Using the Online Chemical Modeling Environment (OCHEM). *Mol. Inform.* **2014**, *33*, 73–85.
- 112 Tetko, I. V. 5 . 27 Rule-Based Systems to Predict Lipophilicity. In *omprehensive Medicinal Chemistry II*; I V Tetko, D J Livingstone, C., Ed.; Elsevier Ltd, 2007; pp 649–668.
- 113 Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. *In silico* Approaches to Prediction of Aqueous and DMSO Solubility of Drug-Like Compounds: Trends, Problems and Solutions. *Curr. Med. Chem.* **2006**, *13*, 223–241.
- 114 <http://crdd.osdd.net/raghava//old/webservices.html>.
- 115 Anna Köhler, H. B. *The Electronic Structure of Organic Semiconductors*, 1st ed.; Wiley-VCH Verlag GmbH & Co. KGaA, 2015.
- 116 Brütting, W. Organic Semiconductors. In *Physics of Organic Semiconductors*; Wolfgang Brütting, Ed.; John Wiley & Sons, 2006; p 554.
- 117 Jacob, M. Organic Semiconductors: Past, Present and Future. *Electronics* **2014**, *3*, 594–597.
- 118 Kowalsky, W.; Becker, E.; Benstem, T.; Johannes, H. H.; Metzdorf, D.; Neuner, H.; Schöbel, J. Organic Semiconductors: Fundamentals and Applications. In *Advances in Solid State Physics 40*; Kramer B., Ed.; Springer, Berlin, Heidelberg, 2007; Vol. 40, pp 795–808.
- 119 Agnieszka ŚWIST, J. S. Organic Semiconductors – Materials of the Future? *Chemik* **2012**, *66*, 289–296.

-
- 120 Koželj, M.; Cvikl, B.; Korošak, D. Application of Organic Semiconductors for the Detection of Ionizing Radiations. In *International Conference Nuclear Energy for New Europe 2006*; 2006; pp 1–10.
- 121 <https://teaching.shu.ac.uk/hwb/chemistry/tutorials/molspec/uvvisab1.htm>
- 122 Anna Köhler, H. B. *The Electronic Structure of Organic Semiconductors*, 1st ed.; Wiley-VCH Verlag GmbH & Co. KGaA, 2015.
- 123 Kaloni, T. P.; Schreckenbach, G.; Freund, M. S. Band Gap Modulation in Polythiophene and Polypyrrole-Based Systems. *Sci. Rep.* **2016**, *6*, 36554.
- 124 *Organic Electronics*; 2007; Vol. 2.
- 125 Bao, P. Z. Self-Assembly in Organic Thin Film Transistors for Flexible Electronic Devices <http://www.sigmaaldrich.com/technical-documents/articles/material-matters/self-assembly-in-organic.html>.
- 126 Ostroverkhova, O. Organic Optoelectronic Materials: Mechanisms and Applications. *Chem. Rev.* **2016**, *116*, 13279–13412.
- 127 Sulaiman, K.; Ahmad, Z.; Fakir, M. S.; Abd Wahab, F.; Mah Abdullah, S.; Abdul Rahman, Z. Organic Semiconductors: Applications in Solar Photovoltaic and Sensor Devices. *Mater. Sci. Forum* **2013**, *737*, 126–132.
- 128 Mauri, a; Consonni, V.; Pavan, M.; Todeschini, R. Dragon Software: An Easy Approach to Molecular Descriptor Calculations. *Match Commun. Math. Comput. Chem.* **2006**, *56*, 237–248.
- 129 MacroModel, version 9.7, Schrödinger, LLC, New York, NY, 2009.
- 130 <http://www.hyper.com/> student version
- 131 HyperChem(TM) Professional 7.51, Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA
- 132 O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 1–14.
- 133 Witten, I. H.; Frank, E. WEKA Machine Learning Algorithms in Java. In *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*; 2000; pp 265–320.
- 134 Hall M, Eibe F, Holmes G, Pfahringer B, et al. The WEKA Data Mining Software: An Update. *SIGKDD Explor.* **2009**, *11*, 10–18.
- 135 Canvas, version 1.5, Schrödinger, LLC, New York, NY, 2012.

- 136 Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and Comparison of 2D Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods. *J. Mol. Graph. Model.* **2010**, *29*, 157–170.
- 137 Rahman, M. M.; Davis, D. N. Addressing the Class Imbalance Problem in Medical Datasets. *Int. J. Mach. Learn. Comput.* **2013**, *3*, 224–228.
- 138 Chawla, N. V. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*; Oded Z. Maimon, L. R., Ed.; Springer Science & Business Media, 2005; pp 853–867.
- 139 Rahman, M. M.; Davis, D. N. Cluster Based Under-Sampling for Unbalanced Cardiovascular Data. In *Proceedings of the World Congress on Engineering 2013*; 2013; Vol. III, pp 1–6.
- 140 Elrahman, S. M. A.; Abraham, A. A Review of Class Imbalance Problem. *Netw. Innov. Comput.* **2013**, *1*, 332–340.
- 141 López, V.; Fernández, A.; García, S.; Palade, V.; Herrera, F. An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. *Inf. Sci. (Ny)*. **2013**, *250*, 113–141.
- 142 V. García, J.S. Sánchez, R. A. M. I. On the Effectiveness of Preprocessing Methods When Dealing with Different Levels of Class Imbalance. *Knowledge-Based Syst.* **2012**, *25*, 13–21.
- 143 Edward O. Pyzer-Knapp, Changwon Suh, R. G.; Omez-Bombarelli, J. A.-I.; Al', A.; Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **2015**, *1*, 1–25.
- 144 Subramaniam, S.; Mehrotra, M.; Gupta, D. Virtual High Throughput Screening (vHTS) - A Perspective. *Bioinformation* **2008**, *3*, 14–17.
- 145 Organic Semiconductors for Advanced Electronics <http://www.sigmaaldrich.com/chemistry/chemical-synthesis/learning-center/chemfiles/chemfile-2001-2003.html#2003>.
- 146 Facchetti, A. π -Conjugated Polymers for Organic Electronics and Photovoltaic Cell Applications. *Chem. Mater.* **2011**, *23*, 733–758.
- 147 <http://www.gopolymers.com/plastic-types/abs-plastic.html>.
- 148 Mohd Sobran, N. M.; Arfah, A.; Ibrahim, Z. Classification of Imbalanced Dataset Using Conventional Naïve Bayes Classifier. In *International Conference on Artificial Intelligence in Computer Science and ICT (AICS 2013)*; 2013; pp 35–42.
- 149 Townsend, J.; Glen, R.; Mussa, H. Note on Naive-Bayes Based on Binary Descriptors in Cheminformatics. *J. Chem. Inf. Model.* **2012**, *52*, 2494–2500.

-
- 150 Klon, A. E.; Lowrie, J. F.; Diller, D. J. Improved Naive Bayesian Modeling of Numerical Data for Absorption, Distribution, Metabolism and Excretion (ADME) Property Prediction. *J. Chem. Inf. Model.* **2006**, *46*, 1945–1956.
- 151 Charles A. Harper, E. M. P. *Plastics Materials and Processes- A Concise Encyclopedia- Book Review*; John Wiley & Sons, 2003; Vol. 20.
- 152 https://www.chem.uwec.edu/chem405_s01/malenirf/project.html
- 153 Bhuvaneswari, E.; Sarma Dhulipala, V. R. The Study and Analysis of Classification Algorithm for Animal Kingdom Dataset. *Inf. Eng.* **2013**, *2*, 6–13.
- 154 Gibert, K.; Izquierdo, J.; Holmes, G.; Athanasiadis, I.; Comas, J.; Sánchez-Marrè, M. On the Role of Pre and Post-Processing in Environmental Data Mining. In *Proc. iEMSs 4th Biennial Meeting - Int. Congress on Environmental Modelling and Software: Integrating Sciences and Information Technology for Environmental Assessment and Decision Making, iEMSs 2008*; 2008; Vol. 3, pp 1937–1958.
- 155 Sokolova, M.; Lapalme, G. A Systematic Analysis of Performance Measures for Classification Tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437.
- 156 Jamal, S.; Arora, S.; Scaria, V. Computational Analysis and Predictive Cheminformatics Modeling of Small Molecule Inhibitors of Epigenetic Modifiers. *PLoS One* **2016**, *11*, 1–13.
- 157 Correa, M.; Bielza, C.; Pamies-Teixeira, J. Comparison of Bayesian Networks and Artificial Neural Networks for Quality Detection in a Machining Process. *Expert Syst. Appl.* **2009**, *36*, 7270–7279.
- 158 Shaikhina, T.; Lowe, D.; Daga, S.; Briggs, D.; Higgins, R.; Khovanova, N. Decision Tree and Random Forest Models for Outcome Prediction in Antibody Incompatible Kidney Transplantation. *Biomed. Signal Process. Control* **2015**, 1–7.
- 159 Khaled Fawagreh, Mohamed Medhat Gaber, E. E. Intelligent Data Engineering and Automated Learning-IDEAL 2014 15th International Conference Salamanca, Spain, September 10-12, 2014 Proceedings. In *Diversified Random Forests Using Random Subspaces*; 2014; Vol. 8669.
- 160 Kamath RS, K. R. Modelling Fetal Morphologic Patterns through Cardiotocography Data: A Random Forest Based Approach. *Res. J. Pharm. Biol. Chem. Sci.* **2016**, *7*, 2449–2455.
- 161 Touw, W. G.; Bayjanov, J. R.; Overmars, L.; Backus, L.; Boekhorst, J.; Wels, M.; Sacha van Hijum, A. F. T. Data Mining in the Life Science Swith Random Forest: A Walk in the Park or Lost in the Jungle? *Brief. Bioinform.* **2013**, *14*, 315–326.

-
- 162 Zhao, Y.; Zhang, Y. Comparison of Decision Tree Methods for Finding Active Objects. *Adv. Sp. Res.* **2008**, *41*, 1955–1959.
- 163 Latha, G.; Naik, B.; Kiranmai, R. Data Balancing Using SVM as Pre-Processor. In *Proc. of Int. Conf. on Advances in Communication, Network, and Computing, CNC*; Elsevier, 2014.
- 164 Chauhan, H.; Chauhan, A. Evaluating Performance of Decision Tree Algorithms. *Int. J. Sci. Res. Publ.* **2014**, *4*, 4–5.
- 165 Helgee, E. A. Improving Drug Discovery Decision Making Using Machine Learning and Graph Theory in QSAR Modeling, 2010.
- 166 Yang, C. Y.; Su, K. H.; Jan, G. E. An Elaboration of Sequential Minimal Optimization for Support Vector Regression. In *2014 IEEE International Conference on System Science and Engineering (ICSSE)*; 2014; pp 88–93.
- 167 Platt, J. C. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Adv. kernel methods* **1998**, 185–208.
- 168 Chen, L.; Chu, C.; Feng, K. Predicting the Types of Metabolic Pathway of Compounds Using Molecular Fragments and Sequential Minimal Optimization. *Comb. Chem. High Throughput Screen.* **2016**, *19*, 136–143.
- 169 A. Golbamaki, A.M. Franchi, G. G. The Maximum Common Substructure (MCS) Search as a New Tool for SAR and QSAR. In *Advances in QSAR Modeling, Challenges and Advances in Computational Chemistry and Physics*; Roy, K., Ed.; Springer International Publishing, 2017; pp 149–165.
- 170 Cao, Y.; Jiang, T.; Girke, T. A Maximum Common Substructure-Based Algorithm for Searching and Predicting Drug-like Compounds. *Bioinformatics* **2008**, *24*, 366–374.
- 171 <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>

PART II
BIOLOGICAL PREDICTIVE MODELS

Sajeev R “Development and validation of molecular descriptor based on physical and biological prediction models” Thesis. Department of Chemistry, Malabar Christian College, Calicut, 2017

CHAPTER 1

INTRODUCTION

1.1 Biological databases

In recent years, a variety of public-domain bioactivity resources have been developed like PubChem BioAssay and ChemBank.¹ They are large archival databases that provide access to millions of deposited screening results, typically from high-throughput screening (HTS) experiments. And several bio-active information was extracted from literature that largely focuses on thematic areas like BindingDB.² They chiefly contain binding information of potential three-dimensional structures of proteins targets extracted from various publications. Also, Protein Data Bank (PDB) contains many thousands of binding affinity information for protein–ligand complexes and BRENDA³ provides binding constants for enzymes while DrugBank,⁴ provide detailed annotation about the properties and mechanistic action of approved drugs.⁵ Such enormous amount of biological information existing in the online resources has made it possible to generate various computational predictive models to predict ADME/toxicity properties in the biological models.⁶

1.2 QSAR/ QSPR

In early days chemical structures were related to biological or chemical activity using mathematical models (QSAR).⁷ The constructed model describes the relationship once the activity of ligands is determined. Structure quantification is not a trivial problem since it cannot be represented by a mere value. Instead, molecular descriptors can be computed from the structure and used to quantify it. Thereby a relationship can be described through a computational model by using structural descriptors as independent variables and activity as a dependent variable.⁸ Always it is very important that the calculated descriptors are related to the biological or chemical activity for which the model is built or if a descriptor is not related to activity, it shouldn't be incorporated in the modeling process so as to avoid the wrong prediction. Then the built QSAR model is used for prediction of biological activity of novel molecules. A QSAR model can also screen potentially active

molecules from a database. One of the main advantages of a QSAR model is it can incorporate a wide range of different variables, be it physical, chemical or biological, it can also be utilized in industries apart from drug design such as toxicology, food chemistry, and other fields. A typical QSAR study would involve Hammett's constants, partition coefficients, molar refractivity and many other descriptors. Examples of some 3D-QSAR algorithms are the Comparative Molecular Field Analysis (CoMFA), the Comparative Molecular Similarity Indices Analysis (CoMSIA), Comparative Molecular Moment Analysis (CoMMA) and GRID.⁹

1.2.1 QSAR Approach to Drug Design

QSAR models are widely used in the field of drug design in comparison to the traditional way of approach. Drug identification is performed largely through random experimentation which is very less effective and the mechanism of action of a successful drug would typically remain obscured. The idea behind QSAR approaches is to use the known responses/activity of simple compounds to predict the responses of unknown compounds. And only predicted compounds found to have desired properties would then be tested. Very often the input for QSAR applications consists of various physico-chemical properties like solubility, pKa, pKd, partition coefficients, surface areas (polar and non-polar), topological indices, atom connectivities, and intramolecular energies. And the output of such models is a decision concerning biological activity such as binding affinity, inhibition, absorption, bioavailability and toxicity.¹⁰

1.3 Drug Design

The use of medicines and drugs dates back to as early as 3100 BC. In the early days, drug discovery has been a trial-and-error process for the majority of the time. Conventionally, the drug development process is very time-consuming and laborious as well as a blind screening approach. And the disadvantages of such method have led to the concept of "Rational drug design" in the 1960's.¹¹

As chemistry advanced, compounds were extracted and purified the active compounds known to have medicinal properties and deduce the structures of these active compounds. The science of drug design progressed further with advances in molecular biology and biochemistry, which elaborated the concepts of genes and

ligand-receptor relationships. With the introduction of data integration and knowledge management solutions with the help of computer power, informatics in drug discovery has cut the development cost of traditional drug discovery by almost a third. Now the development time is reduced from 10-16 years to 6-8 years thus make drug discovery more cost-effective.¹²

Drug discovery is very complex and a time-consuming process that requires an interdisciplinary effort to design effective and commercially feasible drugs. The main purpose of drug design is to find a molecular structure that can fit a specific pocket on a protein target both geometrically and chemically. After passing the *in vitro-in vivo* and human clinical trials, the drug is subjected to approval by regulatory authorities following which it is then available to patients via market. The conventional drug design methods are a long design cycle and high cost that involves random screening of chemicals found in nature or synthesized in laboratories. Modern techniques like structure-based and ligand-based drug approaches have speeded up the drug discovery process in an efficient manner.^{13,14} Significant improvement has been made during the past few years in major areas concerned with drug design and discovery. A comparison involving salient features of traditional approaches of drug discovery with computational approaches is given in Table 1.

Table 1. A comparison of conventional and modern drug discovery approaches.

Parameter	Traditional approach	Modern approach
Procedure	Trial and error method	logical
Screening type	Blind screening	Specific and target oriented
Execution of steps involved	Sequential	Parallel
Drug development cost	Very high	About one-third
Drug development duration	10-16 years	6-8 years
Interdisciplines of drug development	Strictly separate	Coordinated
Transparency of drug development process	Less	More
Management of drug development process	Not easy	Easy
Redundancy	Exists	Can be reduced
Communication between disciplines	More Complicated	Less Complicated

A drug is any chemical that affects the body and its processes but legally “Under the US law, a drug is any substance (other than a food or device), which is used in the diagnosis, cure, relief, treatment or prevention of disease, or intended to affect the structure or function of the body”. And some of the important features of an “ideal” drug must be (i) safe and effective (ii) should be well absorbed orally and bioavailable (iii) metabolically stable and with a long half-life (iv) nontoxic with minimal or no side effects (v) should have selective distribution to target tissues.^{12,15}

The development of any potential drug begins with years of scientific study to determine the etiology of a disease, for which pharmaceutical intervention is possible. The result is the determination of specific receptors (targets) and to find one or more compounds (lead) which interacts with that target to alter their bioactivity by some means. From this point onward, medicinal chemistry plays an important role in refinement and testing in an iterative manner until a drug is developed that undergoes clinical trials. The techniques used to refine drugs are ligand-based design (little knowledge of protein/target structure is available), combinatorial and structure-based design (three dimensional structure of protein is available).¹⁶ After the successful clinical phase, the drug is subjected to approval by regulatory authorities and then marketed. The modern-day drug discovery pipeline is shown in Figure 1.

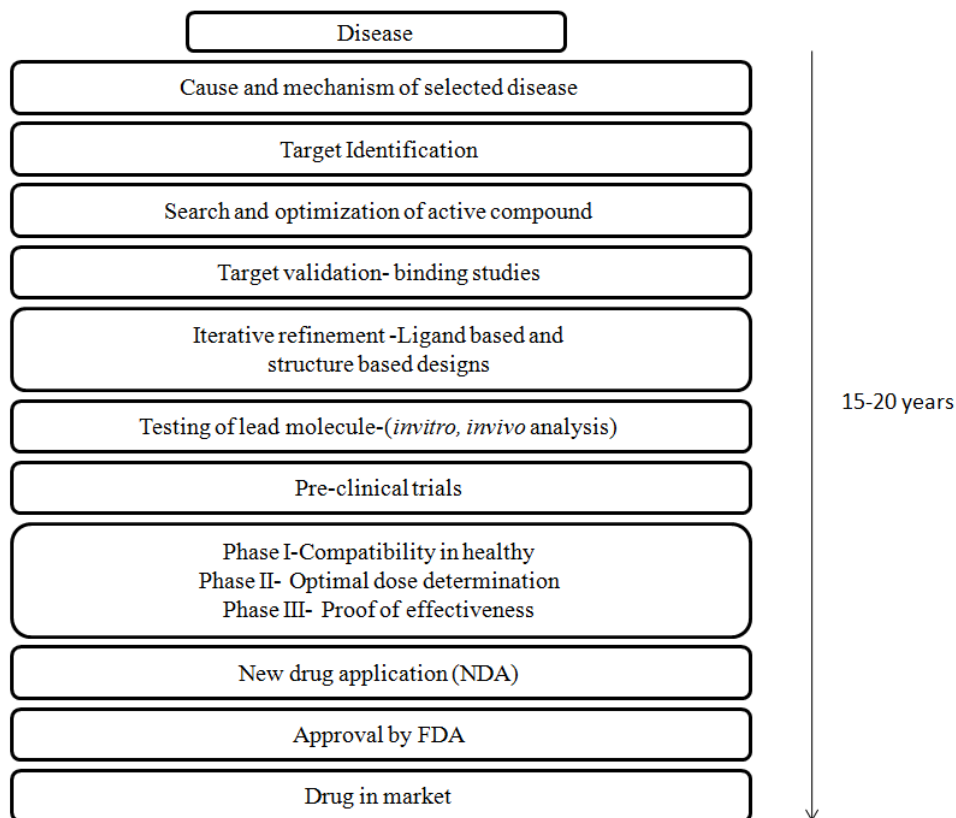


Figure 1. Different phases of drug discovery process

1.3.1 Drug design process

1.3.1.1 Selection of Disease

Choosing a disease basically depends on where there is a need for new drugs. However, pharmaceutical companies do consider both economic factors as well as medical ones because of a huge money investment made towards the research and development of a new drug. Therefore, companies ensure that they get a good financial return on their investment. As a result, most of the research projects tend to focus on diseases (e.g., cancer, cardiovascular diseases, depression, diabetes, flu, migraine, obesity) that are important in the “first world” countries. Less is carried out on the tropical diseases of the developing world which only affects a small subset of the population.

1.3.1.2 Selection of Target

Once a therapeutic area has been identified then the next step is to find out a molecular target for e.g. receptor, enzyme, or nucleic acid. An understanding of biomacromolecules type in a particular disease state is very important as it allows the medicinal research team to design agonists or antagonists (against a particular receptor) or inhibitors (against a particular enzyme).

1.3.1.3 Target specificity and selectivity between species

In modern medicinal chemistry research target specificity and selectivity is a crucial factor in the drug design process. The more selective a drug is for its target, less likely that they will interact with different targets and undesirable side effects can be reduced. Therefore the best targets are selected which are unique to microorganism that are not present in human being. For example the antibiotic Penicillin targets an enzyme that is responsible for the bacterial cell wall synthesis. For the same, mammalian cell does not possess cell wall and the target enzyme is absent.

1.3.1.4 Identify a Bioassay

Choosing the right bioassay or test system is very important in the drug design program. Since very large molecules are screened the test should be simple, quick and relevant. In the early stages, human testing is not possible so the tests are carried out *in vitro* (i.e. on isolated cells, tissues, enzymes or receptors) or *in vivo* (on animals). In general, *in vitro* tests are preferred over *in vivo* test because they are cheaper, easier, faster and requires a relatively small amount of compounds for testing. And it doesn't involve live animals instead specific tissues, cells or enzymes are used.

1. 3.1.4 Lead Identification

A lead¹⁷ is a compound that shows some level of pharmacological activity. The bio-activity may be less with undesirable side effects but it provides a good start for the drug design and development process. This means that structural

modification is required and further molecular optimization is carried out. For the development of an orally active compound the lead compound has to account for the “rule of three” criteria as suggested below. Also Lipinski's Rule of Five or Veber's parameters are obeyed by most orally active drugs.¹⁸

Rule of three criteria:

1. Molecular weight less than 300.
2. No more than 3 hydrogen bond donors.
3. No more than 3 hydrogen bond acceptors.
4. $c\text{LogP} = 3$.
5. No more than 3 rotatable bonds.
6. A polar surface area = 60 \AA^2 .

1.3.1.5 Drug discovery

1. Identify structure-activity relationship (SARs).
2. Identify the pharmacophore.
3. Improve target interactions (pharmacodynamics).
4. Improve pharmacokinetics properties.

1.3.1.6 Drug development

1. Drug patent procedure.
2. IND - Investigational New Drug allows the drug to be studied in a human.
3. Preclinical trials are carried out involving drug metabolism, toxicology, formulation and stability test, pharmacology studies.
4. Clinical trials (Phase I, Phase II, Phase III).

5. NDA - New Drug Application – here application is made for permission to market the new agent provided Phase III results meet expectations.
6. Food & Drug Administration (FDA) is an agency in USA that oversees the drug evaluation process and grants approval for marketing of new drug products.
7. Design, a manufacturing process involves chemical and process development.
8. Registration and marketing.

Many of these stages run parallel to and are dependent on each other. But still the whole process of discovery, design and development of a new drug takes more than 12-15 years with a cost estimate in the region of \$1 billion.¹⁹

1.3.2 Virtual Screening (VS)

Nowadays, there are multiple computational methodologies used in cheminformatics and bioinformatics tools for the study of biological systems and drug discovery. VS is a computational technique widely used in the area of drug design. Its main objective is to search for specific information in compounds or molecular libraries with similar structural properties that can acceptably interact with a therapeutic target to understand the drug target interactions and drug likeliness of a molecule.²⁰ VS methods can be classified into structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS) approaches depending on the availability of the protein structure and the ligand information. If the 3D structure of the target is known, a molecular docking is performed that sample ligand poses with respect to the target binding site, but where the information on the receptor is scant, LBVS methods are commonly used by performing similar compounds by 2D/3D similarity or pharmacophore searches.²¹ The process of drug development aims towards the identification of compounds with pharmacological interest to assist in the treatment of diseases and ultimately to improve the quality of life. The compounds used in the process are usually small organic molecules (ligands) which interact with specific macromolecules (receptors/targets). Usually

compounds are compiled into a large collection based on certain physicochemical property or particular protocol termed as libraries. And the desired activity of the compound libraries are achieved by a high-throughput screening (HTS). Since large compound libraries often contain millions of drug like molecules, their robotic testing is very expensive. It also depends on various experimental factors like compounds selected for screening should be highly stable, high solubility in the testing media etc. This reduces the usability of HTS. An alternative to experimental HTS is high-throughput virtual screening (HTVS or VS).²²

VS is a knowledge-driven approach that is based on the amount and type of information available about the system under inspection. The quality and the amount of information regarding the system under inspection is a critical factor in the *in-silico* drug design program see Table 2. Even though HTS has identified several structurally diverse compounds compared to VS, in some cases virtual hits were found to be better than the experimental compounds. Moreover, not only real molecules but purely theoretical construct virtual molecules are screened *in silico*. This feature allows entire study significantly in a cost effective and efficient manner.

Table 2. Classification of virtual screening methods based on the information available on target and ligands.

	Ligand information is known	Ligand information is unknown
Availability of 3D protein structure (X-ray Crystallographic, NMR or close homologue)	Structure-Based Virtual Screening (SBVS): Molecular Docking studies	De novo Structure-Based Virtual Screening
Protein Structure Unknown	Ligand-Based Virtual Screening (LBVS): 1. Similarity search (Fewer ligand molecules) 2. Pharmacophore model (several ligand molecules)	Virtual Screening not possible

Molecular docking involves a complex optimization task of finding the most favorable 3D binding conformation of the ligand to the receptor molecule. The 3D models of biomolecules are obtained from NMR spectroscopy, X-ray crystallography or homology modeling. Currently, public repository for three dimensional protein structure Protein Data Bank (PDB) contains more than 70,000 experimentally solved 3D structures of proteins that are used for various VS process. 3D protein structure is obtained from NMR spectroscopy that produces several real conformations for the receptor while X-ray crystallography offers one single state of the crystallized protein. Even though all these methods have been successfully applied in structure-based VS, X-ray Crystallography remains the most powerful source of structural data.

Molecular docking is not only limited to protein-ligand scenarios but also with other macromolecule like RNA, DNA and protein-protein interaction. Usually they are referred as ligand-docking software as commercial license (GLIDE, GOLD), academic users and freely available software like DOCK or Auto-Dock Vina. Each docking softwares, shares a different algorithm in generation and scoring of the various ligands poses. There are also online web servers available like Supercomputing Facility for Bioinformatics & Computational Biology (<http://www.scfbio-iitd.res.in/>), docking server (<https://www.dockingserver.com/web/>) etc, from which docking can be performed by providing the target and ligand information. Structural analysis of biomolecules like protein-ligand complexes is visualized from packages like PyMOL, VMD, Swiss-PDB Viewer and Chimera. Major online database like ZINC a non-commercial database with more than 22 million compounds allows the virtual screening to be simple, provided with the target structure.

Docking is computationally intensive and not suitable to carry out very large VS experiments. By contrast, LBVS methods are computationally inexpensive and easy to use as it aims to find compounds in a database that matches best to a given query. In LBVS new hits can be identified even with one or more compounds with a specific activity. Conceptually, it is based on the similarity property principle (SPP)

as described by Johnson and Maggiora in 1990 which states that “similar molecules should have similar biological properties (activity)”.²³ However; a small structural modification of the active compounds can either increase or decrease the bioactivity. And the difference between active and inactive compound can be distinguished by a small chemical difference.

LBVS approaches include similarity search and compound classification techniques including pharmacophore searching, shape comparison, and machine learning.²⁴ Similarity search is performed by making use of the molecular fingerprints derived from 2D molecular graphs or 3D conformations. They are compared in a pair-wise manner against a database (molecular fragments) using a similarity metric like Tanimoto coefficient and ranked in accordance with the order of decreasing molecular similarity to the reference molecule. Tanimoto is a measure of similarity, that quantifies the compound similarity by determining the overlap between the fingerprint strings or feature set which is calculated from the equation (1).²⁵

$$N_{ab}/(N_a + N_b - N_{ab}) \quad (1)$$

Where N_a and N_b are the number of features/bits set in the fingerprint of compounds a and b , N_{ab} is the number of features/bits set present in both fingerprints of a and b . From this ranking, molecules are selected.

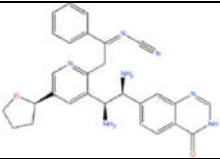
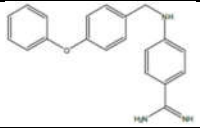
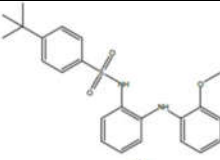
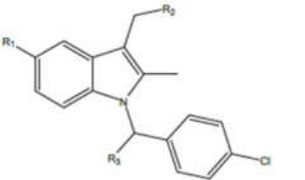
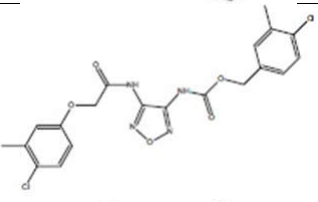
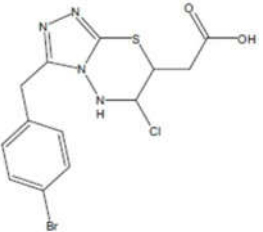
LBVS also includes pharmacophore search which is “a molecular framework that carries the essential features responsible for a drug’s biological activity” as defined by Paul Ehrlich during the late 1800s.²⁶ A pharmacophore model is constructed from a set of known ligands. For an active compound the features like H-donor, H-acceptor, anionic, cationic and steric factors constitute a pharmacophore model. While a 3D pharmacophore reflects the positioning of key amino acid residues present in the pocket of a target protein. These models can be built from data on target protein structure by studying the binding site and possible ligand-protein interactions. Use of pharmacophore models are widely adopted in finding specific inhibitors for G-protein coupled receptors, enzymes and ion channels.^{8,27} Other approaches include compound classification techniques like classification

methods, such as clustering and partitioning (for which many different algorithms exist). Recently ML approaches, such as support vector machines (SVM), decision trees (DT), k-nearest neighbors (k-NN), naive bayesian methods and artificial neural networks (ANN), are popular and widely used as LBVS. The purpose of all these algorithms is to predict compound class labels either active or inactive through predictive models derived from training sets and test sets. The first application of ML in drug discovery was sub-structural analysis (SSA), as described by Cramer et al. ML is used as a computational tool for the automated analysis of very large biological screening dataset. It is an attractive area of research in computer science and related science, with the increasing availability and accessibility of big data collections that promote the development of novel tools for data mining methods. The broad spectrum of ML algorithms helps in decision making process in a better way in the drug design process. Thus, LBVS methods have an increasingly important role at the beginning of the drug discovery projects, especially where little 3D information is available for the receptor.

1.3.3 Applications of VS

In recent years, the usage of VS against potential therapeutic agents is observed to be far more than the previous years for example: inhibitors of HIV-1 reverse transcriptase, SENP2 inhibitors, kinases proteins agonists, inhibitors with vasodilator activity. The following Table 3, lists some of the active compounds identified through various computational methods.²⁰

Table 3. Some active compounds identified by various computational methods.

Molecules	Use and Methods
	Antiretroviral agents (VS)
	Thrombin inhibitor (Combinatorial docking)
	Kv1.5 channel blocker (Fragment based)
	Aldose reductase inhibitor (LBVS)
	SNEP2 inhibitors (VS)
	Antimycobacterial agents (VS)

Software based computational virtual screening in drug discovery overcome the limitations of traditional HTS methods. It takes advantage of fast algorithms and computer power to filter chemical space, select and prioritize potential drug candidates successfully. The filtering step ensures that rejection of compounds from the compound library is minimal that do not meet specific drug-like criteria. Thus VS is a useful and promising tool for *in silico* drug discovery program.^{21, 28}

1.4 Tuberculosis

Tuberculosis (TB) is a communicable disease caused by the bacteria *Mycobacterium tuberculosis*.²⁹ It typically affects the lungs (pulmonary TB) but can affect other sites of the body, such as brain, kidneys, or spine (extra pulmonary TB). An airborne disease, that spreads among people who are sick with pulmonary TB expel bacteria, for example by coughing. Despite the availability of TB treatment, the threat the disease represents is painful because 10 million people are infected and an average of two million people die each year. However, the probability of developing TB is much higher among people infected with the immune compromised disease like HIV. Also with the increased prevalence of multi-drug resistance (MDR) and extensively drug-resistant (XDR) TB mortality rates can approach 100% of those infected with XDR-TB. According to global TB report published in 2016 there were an estimated 10.4 million new TB cases worldwide: of which 5.9 million were men, 3.5 million were women and 1.0 million were among children. And people living with immunocompromised disease accounted for 1.2 million of all new TB cases. There were also 1.4 million TB deaths in the year 2016.³⁰ According to these reports, the number of TB deaths is unacceptably high, but it is curable when timely diagnosed with correct treatment.

Effective TB drug treatments were developed in the early 1940s. The most effective first-line anti-TB drug, rifampicin/rifampin (RIF), was available in the 1960s. Currently, the treatment for drug-susceptible TB is a six-month regimen of four first-line drugs (FLD): isoniazid (INH), rifampicin (RIF), ethambutol (EMB) and pyrazinamide (PZA). However, FLD often fails to cure TB for many reasons. Relapse and spreading of the disease is the main cause for the emergence of drug-resistant bacteria. This includes multidrug-resistant TB (MDR-TB), which is resistant to at least isoniazid (INH) and rifampicin (RIF). And it is of great concern to go with more toxic and expensive second-line drugs. The second line drugs are Fluoroquinolones-Ofloxacin (OFX), levofloxacin (LEV), moxifloxacin (MOX) and ciprofloxacin (CIP) and injectable antituberculosis drugs- Kanamycin (KAN), amikacin (AMK) and capreomycin (CAP). Extensively drug resistant-TB (XDR-

TB) that is resistant to either isoniazid or rifampicin, any fluoroquinolone and at least one of the three second-line anti-tuberculosis injectable drugs.³¹ First and second line drugs and its target enzymes are presented in Tables 4-5.

Table 4. First line drugs and its respective enzyme information.

Drug	Target/enzyme
Isoniazid	catalase/oxidase, enoyl reductase, alkyl hydroperoxide reductase
Rifampicin	β -subunit of RNA polymerase
Pyrazinamide	PZase
Streptomycin	S12 ribosomal protein
Ethambutol	arabinosyl transferase

Table 5. Second line drugs and its respective enzyme information.

Drug	Target/enzyme
Fluoroquinolones	DNA gyrase
Kanamycin/Amikacin	16S rRNA
Capreomycin	rRNA methyltransferase

Sputum smear microscopy remains the most widely used technique for diagnosing TB. It is a century old method where bacteria are observed in sputum sample under a microscope. TB is completely curable through DOTS or Directly Observed Treatment Short course.^{32,33} It is an internationally recommended strategy for TB control that has been recognized most effective means of eliminating TB from a population. DOTS comprise five major elements as mentioned below.

1. Government commitment to sustained TB control activities.
2. Patients reporting to health services as case detection by sputum smears microscopy.

3. Standardized treatment regimen of six to eight months for at least all confirmed sputum smear positive cases, with directly observed treatment (DOT) for at least the initial two months.
4. A regular, uninterrupted supply of all essential anti-TB drugs.
5. A standardized recording and reporting system that allows assessment of treatment results for each patient and of the TB control programme overall.

1.5 Target – β -lactamase

In our study we selected the enzyme β -lactamase present in the bacteria *M. tuberculosis* and *P. aeruginosa*. Productions of these enzymes are the most common and important mechanism of resistance to β -lactam antibiotics.^{34,35,36} They inactivate β -antibiotics efficiently by hydrolyzing the amide group of the β -lactam ring. β -lactamases are chromosomal enzyme encoded by the gene *blaC* the only gene encoding a β -lactamase in *M. tuberculosis*.^{37,38} Resistance to the β -lactam antibiotics primarily occurs through the horizontal transfer of β -lactamase genes contained on plasmids.³⁹ There are other mechanisms by which the bacterium forms resistance to β -lactam antibiotics.^{40,41}

1. Changes in the active site of Penicillin Binding Proteins (PBPs) can lower the affinity for β -lactam antibiotics.
2. Decreased expression of outer membrane proteins (OMPs) makes it difficult for β -lactams to access PBPs present in inner plasma membrane bacterial cell walls.
3. Efflux pumps are intrinsic resistance phenotype as their main function is to export substrates from periplasm to the surrounding environment.

1.5.1 Bacterial cell wall: mechanism of action of β -lactam antibiotics

Gram positive bacteria cell wall is made up of glycopeptide called peptidoglycan (murein). The backbone of peptidoglycan is made of alternating units of N-acetylglucosamine (NAG) and N-acetylmuramic acid (NAM).⁴² Each NAM unit has a tetrapeptide side chain attached to it as shown in Figure 2. And the cross-linking of two D-alanine–D-alanine NAM pentapeptides is catalyzed by Penicillin Binding Protein (PBPs), which act as transpeptidase. This cross-linking of adjacent glycan strands is responsible for rigidity of the bacterial cell wall.



Figure 2. Cross-linking of two D-alanine–D-alanine NAM pentapeptides. Peptidoglycan model is taken from ref.⁴³

The β -lactam ring of the antibiotic is sterically identical to d-alanine-d-alanine. In the bacterial cell the d-d transpeptidase enzyme mistakenly binds to the β -lactam antibiotic instead of its natural substrate. These enzymes are also known as Penicillin Binding Protein (PBP) since the antibiotic containing β -lactam ring binds to d,d transpeptidase.⁴⁴ The binding results in acylation of the enzyme leading to the production of an inactive penicilloyl-enzyme.⁴⁵ As a result, further cross-linkages between the layers of peptidoglycan halts which weakens the cell wall and ultimately the cell undergoes osmotic instability and lyses.⁴⁶ Bacteria have multiple PBPs where each has a distinct role for e.g., bacterium *E. coli* has seven PBPs. Study says that

the sensitivity of individual PBP is known to vary with individual β -lactam drug but at clinical doses most β -lactam drugs bind to more than one PBP.

1.5.2 Classification of β -lactamase enzyme

β -lactamase enzymes have been categorized into two classification schemes; i) Ambler classes A through D, based on amino acid sequence homology, ii) Bush-Jacoby-Medeiros groups 1 through 4, based on substrate and inhibitor profile. The Ambler classification portrays class A, C, and D as serine β -lactamases while class B are metallo- β -lactamases (MBLs) possess either a single Zn^{2+} ions or a pair of Zn^{2+} ions coordinated to His/Cys/Asp residues in the active site. In our study we used the Ambler classification scheme.^{47,48, 49}

1.5.2.1 Class A Serine β -lactamases

Generally class A serine β -lactamase are susceptible to most of the commercially available β -lactamase inhibitors like Clavulanate, Tazobactam and less in Sulbactam. “TEM” was the first plasmid-mediated β -lactamase that was identified in *E. coli* in 1963. It was named after the patient Temoniera from whom it was isolated. Another common β -lactamase SHV was named from the term “sulfhydryl reagent variable” is primarily found in *K. pneumonia*. In the 3D protein structure Ser 70 in the active site residue corresponds to the mechanistic action against β -lactam antibiotics.⁵⁰

1.5.2.2 Class B Metallo- β -lactamases

MBLs are Zn dependant β -lactamases that demonstrate a mechanism (hydrolytic) different from that of other class of serine β -lactamases of classes A, C, and D. Hydrolytic profile of MBLs makes these enzymes resistant to penicillins, cephalosporins, carbapenems, along with β -lactamase inhibitors with less likely against the antibiotic aztreonam. The *bla*_{MBL} genes responsible for their production are located on the chromosome, plasmid, and integrons. In contrast to serine β -lactamase mechanism, they use the hydroxyl group from a water molecule which is coordinated by Zn^{2+} to hydrolyze the amide bond of a β -lactam antibiotic. Based on their Zn^{2+} dependency, MBLs are classified into three categories (i) whether they are

fully active with either one or two ions (ii) require two ions (iii) employ one ion and are inhibited by binding of an additional ion.⁵¹

1.5.2.3 Class C Serine cephalosporinases

Class C are AmpC β -lactamases which are encoded by *bla* genes either located on the bacterial chromosome or plasmid-borne AmpC enzymes. They are typically resistant to β -lactam- β -lactamase inhibitor combinations, penicillin, and cephalosporins. Class C enzymes hydrolytic mechanism is based on the nucleophilic residue Ser 64 present in the active site where Tyr 150 behaves as a general base thereby increasing the nucleophilicity of serine residue for acylation.⁵²

1.5.2.4 Class D Serine oxacillinases

Class D β -lactamases are “oxacillinases” due to their ability in hydrolyzing oxacillin at a rate of at least 50% of that of benzylpenicillin in comparison to the slower hydrolysis rate in class A and C. OXA enzymes confer resistant to β -lactamase inhibitors with some exceptions; e.g., OXA-2 and OXA-32 are inhibited by tazobactam but not sulbactam and clavulanate, but OXA-53 is inhibited by clavulanate. In the mechanism of inhibition Lys 70 serve as the general base by activating nucleophilic residue Ser 67 for both acylation and deacylation step.⁵³

1.5.2.5 Hydrolytic Mechanism in Class A β -Lactamase

The binding of β -lactam antibiotics on serine β -lactamases is much like PBPs, where they use strategically positioned water molecules to hydrolyze the acylated β -lactam.⁴¹ During this process the β -lactamase is regenerated that is used to inactivate additional β -lactam molecules. This enzymatic reaction for a penicillin β -lactam substrate and a class A serine β -lactamase enzyme is represented in the Figure 3.

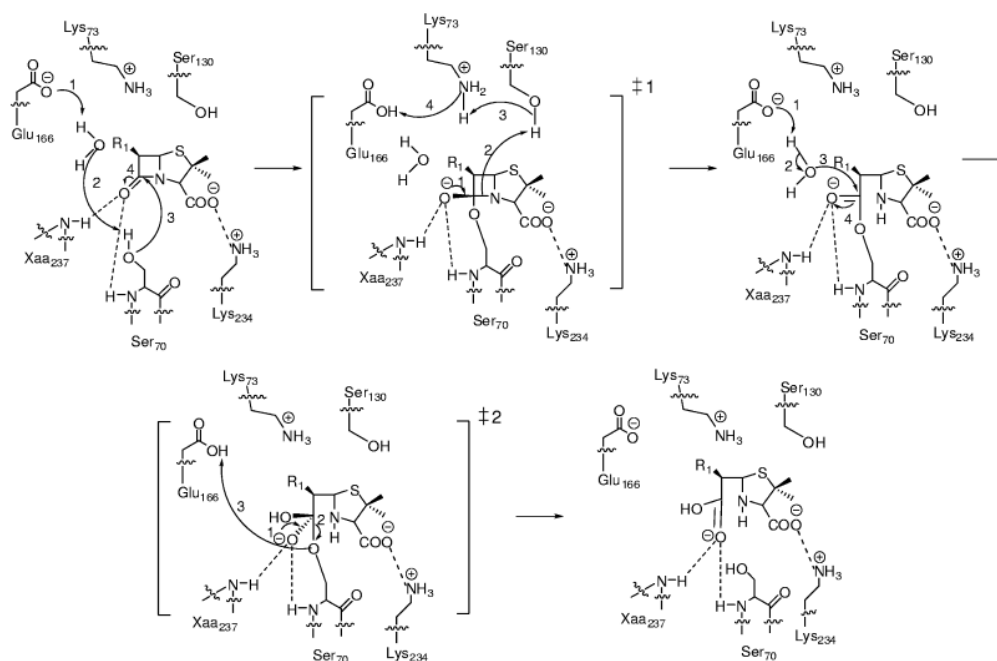


Figure 3. Illustrates the hydrolytic mechanism in the active site of Serine β -lactamase enzyme. Here we represent the reaction scheme of a typical class A β -lactamase. In the binding of penicillin β -lactam substrate with serine β -lactamase, the residue Glu 166 (also in some reference Lys 73 acts as the general base to activate Ser 70) participates in activating a water molecule for both acylation, deacylation and nucleophilic residue Ser 70 is activated. Dashed lines presented in the scheme represent the hydrogen bonds. After activation of the hydroxyl group, the nucleophile Ser 70 in the active site attacks on the carbonyl group of the β -lactam antibiotic, leading to a high-energy acylation intermediate. Then C-N bond is cleaved followed by protonation of the β -lactam nitrogen. This results in the formation of the covalent acyl-enzyme, which adopts a lower energy state. Then the catalytic water molecule attacks in the formation of a high-energy deacylation intermediate. Subsequently, hydrolysis of the bond between the β -lactam carbonyl and the oxygen of Ser 70 takes place and deacylation regenerates active β -lactamase and inactive β -lactam fragment. Figure of the reaction mechanism is adapted from ref.⁴¹

1.5.2.6 β -Lactamase Inhibitors in Clinical Practice

β -lactamase inhibitors can be classified as reversible and irreversible. Most of them contain a β -lactam ring which is a four-membered cyclic amide consisting of three carbon atoms and one nitrogen atom. Since the nitrogen atom is attached to the β -carbon relative to the carbonyl (C=O) in the four membered ring it is named

as β -lactam. And molecules possessing this structure are called β -lactam antibiotics. In the case of reversible inhibitors, they bind to the active site of the enzyme with high affinity but are poorly hydrolyzed as they act as poor substrates. But in the case of irreversible inhibitors they are effective and inactivate the enzyme completely. Such molecules are termed as “suicide inhibitors” as they initially bind to the enzyme’s active site and get hydrolyzed to a form, which in turn inactivates the whole enzyme.⁵⁴

The three common β -lactamase inhibitors currently in clinical use are Clavulanate, Sulbactam and Tazobactam^{55,56} as shown in Figure 4. The presence of a leaving group at position C-1 of the 5-membered ring makes all the three inhibitors different from penicillins. A better leaving group facilitates secondary ring opening and modification of β -lactamase enzyme. Clavulanate possess an enol ether oxygen at this position, while sulbactam and tazobactam have sulfones. Thus sulbactam is relatively less efficient than clavulanate due to the poor leaving group present in it.

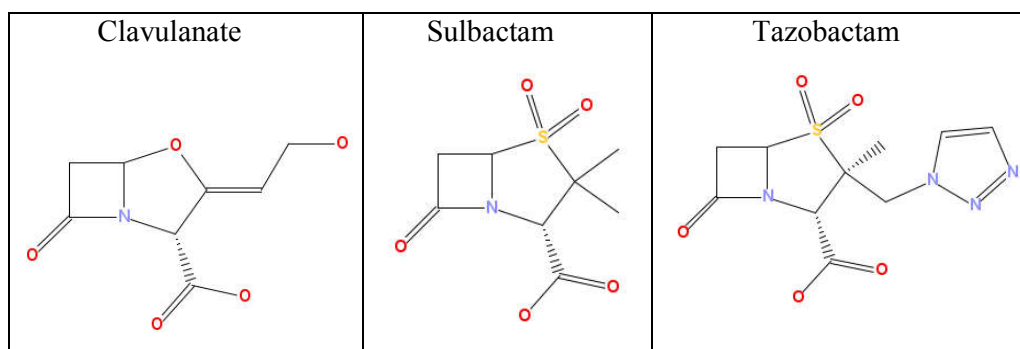


Figure 4. β -lactamase inhibitors Clavulanate, Sulbactam and Tazobactam.

Among the three, Clavulanate was the first β -lactamase inhibitor introduced into clinical practice that was isolated from the species *Streptomyces clavuligerus* in 1970s, more than four decades ago.⁵⁷ The other inhibitors sulbactam and tazobactam are penicillinate sulfones developed in the late 1980.⁵⁸ All inhibitors are structurally similar to penicillin with a common β -lactam moiety. They are found to be effective against class A β -lactamase including TEM-1, TEM-2 and SHV-1 but less effective against class B, C and D β -lactamases. And they don’t have any inhibitory activity against PBPs, however there are exceptions as in the following cases; (i) Sulbactam

is active against *Bacteroides spp.*, *Acinetobacter spp.*, and *N. gonorrhoeae* (ii) Clavulanate is active against *Haemophilus influenzae*, *Acinetobacter spp.*, and *N. gonorrhoeae* (iii) Tazobactam is active against *Borrelia burgdorferi*.⁴¹

The consumption of inhibitors alone is relatively weak to bring out the desired antibacterial effects; they are always combined with β -lactam antibiotics for clinical use. Currently, there exist five β -lactam- β -lactamase inhibitor combinations: amoxicillin-clavulanate, ticarcillin-clavulanate, ampicillin-sulbactam, cefoperazone-sulbactam and piperacillin-tazobactam.^{59, 60} And the other types of β -lactam antibiotics are mentioned in Table 6.

Table 6. Types of β -lactam antibiotics

Penicillins	Benzylpenicillin, Benzathine (Narrow spectrum) Amoxicillin, Ampicillin, Epicillin (Broad spectrum)
Cephalosporins	Cefalotin, Cefaloridine, Cefalexin, Cefroxadine etc. (1 st Generation) Ceforanide, Cefotiam, Cefprozil etc. (2 nd Generation) Ceftriaxone, Cefoperazone, Cefixime, Cefteram etc. (3 rd Generation) Cefpirome, Cefquinome, Cefepime (4 th Generation) Ceftaroline fosamil, Ceftobiprole (5 th Generation)
Cephamicin	Cefotetan, Cefmetazole, Cefoxitin
Carbacephem	Loracarbef
Monobactam	Aztreonam
Carbapenem	Meropenem, Ertapenem, Imipenem

The activity of a β -lactamase inhibitor is evaluated from the turnover number (*tn*) which is defined as the number of inhibitor molecules that are hydrolyzed per unit time before one β -lactamase enzyme molecule is irreversibly inactivated.^{41,61,62} For example, *S. aureus* PC1 requires one clavulanate molecule to inactivate one β -lactamase, while TEM-1 needs 160 clavulanate molecules and for SHV-1 requires 60. Comparative studies shows that sulbactam *tns* are 10,000 and 13,000 for TEM-1 and SHV-1 respectively.⁴¹

1.5.3 Scope of present investigation

From the literature survey it is clear that TB is one of the major diseases that affect third world countries. And there is an urgent need of TB drugs to treat TB. TB has become worse due to the resistance in the current antibiotics. Since drug discovery being a long process, the discovery of a new drug is very difficult that involves a huge sum of money and time. This process can be shortened by making use of the ML algorithms, computational models and data mining softwares.

In our study, we tried to find and prioritize computationally active β -lactamase inhibitors from GSK library of 177 anti-TB molecules. We tried to develop Bayesian, decision tree models like random forest and J48, support vector machines that can predict whether a molecule is β -lactamase active or not. A more detail study was carried out by performing structure based docking methods, ligand based methods and artificial neural network. Molecular docking and ANN was carried out for *M. tuberculosis* (gram positive) and *P. aeruginosa* (gram negative) since both enzyme exhibited same enzymatic mechanism in resistance to the current β -lactam antibiotics.

CHAPTER 2

MATERIALS AND METHODS

2.1 PubChem

PubChem is a public online repository from the National Institutes of Health (NIH's Molecular Libraries Roadmap Initiative which is designed to provide information on biological activities of small compounds). PubChem is integrated with other search engines like NCBI's Entrez that provides various functions involving sub/superstructure, similarity structure, bioactivity data etc. The whole PubChem database is linked to three databases; (i) PubChem Compound, (ii) PubChem Substance and (iii) PubChem BioAssay as a part of information retrieval system. Thus, the chemical structure records in PubChem are easily linked to its biological property information in PubMed and to NCBI's Protein 3D Structure Resource.

(i) PubChem Compound

PubChem Compound is a searchable database of unique chemical structures with validated chemical depiction information and computed properties. It includes over 5M compounds that are pre-clustered and cross-referenced by identity and similarity groups. The search engine allows searching with a variety of chemical synonyms simply by providing molecular name searches like Tylenol, Benzene etc.

(ii) PubChem Substance (deposited structures)

PubChem Substance is a searchable chemical database containing descriptions of chemical compounds from a variety of sources that are linked to PubMed and protein 3D databases. It also includes biological screening results available in PubChem BioAssay. PubChem substance includes over 8 million records and the substances with known content are linked to PubChem Compound. The database allows searching for molecule synonyms like all substances with 'deoxythymidine' as a name fragment, or substances that contain 3'-Azido-3'-

deoxythymidine or to search biology linked searches like substances with tested, active or inactive bioassays etc.

(iii) **PubChem BioAssay**

PubChem BioAssay (<http://pubchem.ncbi.nlm.nih.gov>) is a searchable public repository database containing bioactivity screens of small molecules generated through various high-throughput screening experiments, medicinal chemistry studies, drug discovery programs and chemical biology research. It is hosted by the National Center for Biotechnology Information (NCBI) under NIH. The web-based bioassay provides access to each bioassay that includes descriptions of screening procedural conditions, bioassay test results and readouts. In addition to this, the information content is linked to several other databases like Protein, Gene, BioSystems, PubMed, Online Mendelian Inheritance in Man (OMIM) and protein 3D structure associated with bioassay targets.⁶³

The BioAssay database has been growing substantially since 2004. Currently, the PubChem BioAssay contains over one million records holding 230,000,000 bioactivity results deposited by various organizations around the world. All data in the database are freely accessible and can be downloaded in various structure data format. The PubChem BioAssay statistics⁶⁴ for the time period of 2004–2013 and 2014–2016 are given in Table 7.

Table 7. PubChem BioAssay statistics

	Chemical assays	
	2004–2013	2014–2016
Assay records (AID)	737 994	480 616
Substance samples (SID)	2 755 032	1 396 693
Chemical structures (CID)	1 956 998	986 237
Bioactivity outcomes	222 198 148	8 764 075
Data points	1 403 289 248	100 451 032
Species	2730	1895
Protein targets	7450	6972
Protein targets (human)	3378	3495
Gene targets	-	-
Gene targets (human)	-	-
Gene targets (phenotype)	-	-

2.1.1 Dataset Preparation

For the model generation two PubChem bioassay datasets corresponding to the protein target β -lactamase were selected; (i) AID 434987⁶⁵ (ii) AID 2184⁶⁶ as shown in Figure 5 (a-b).



Figure 5. Bioassay datasets from PubChem database (a) AID 434987 corresponds to *M. tuberculosis* (b) AID 2184 corresponds to *P. aeruginosa*

AID 434987 is a confirmatory bioassay that belongs to the assay project **"Summary of assays used to identify novel compounds that sensitize mycobacterium tuberculosis to beta-lactam antibiotics"** from the Southern Research's Specialized Biocontainment Screening Center (SRBSC). The assay was provided by Dr. William Bishai, Johns Hopkins University Tuberculosis Research Center which was deposited in 15.06. 2010.

AID 2184 is a primary screening bioassay that belongs to assay project **"Epi-absorbance-based counter screen assay for common VIM-2 and IMP-1 inhibitors: biochemical high throughput screening assay to identify inhibitors of TEM-1 serine-beta-lactamase"** from the Scripps Research Institute Molecular Screening Center (SRIMSC). The assay was provided by Peter Hodder, TSRI that was deposited in 08.12.2009. The details of the selected bioassays are furnished in Table 8.

Table 8. The details of the bio-active information

	AID 434987	AID 2184
Bioassay type	Confirmatory	Primary
Protein target	β -lactamase <i>M. tuberculosis</i>	β -lactamase <i>P. aeruginosa</i>
Gene	bla	bla
Number of actives	372	97
Number of inactive	819	100
Number of inconclusive	11	0

2.2 PowerMV

PowerMV is a molecular descriptor calculator.⁶⁷ The software provides various tasks like viewing of compound structure files from SDF, calculation of basic biologically relevant chemical properties and searching against biologically annotated chemical structure databases. The calculation maximum limit remains to be 50k compounds.

The operating environment can compute descriptors useful for judging if a compound is drug-like (Reactive Group Present, Blood-Brain Penetration, Molecular Weight, logP, Number of Hydrogen Bond Donors, Number of Hydrogen Bond Acceptors and Polar Surface Area). Also similarity searching can be performed against annotated databases. PowerMV loaded with molecules are displayed in Figure 6.

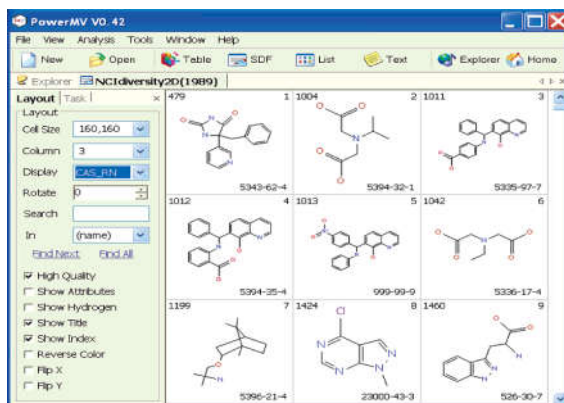


Figure 6. Screenshot of PowerMV descriptor generator software

2.2.1 Table Generation from Compound Data

A total of six molecular descriptor sets are shown in Figure 7, among them four are bit string and two are continuous. In the case of bit string descriptors, each bit is set to either '1' or '0' in accordance to the presence or absence of a particular feature. When a certain feature is presented the value is designated as '1' and '0' when it is not. We implemented both binary descriptors and continuous descriptors, pharmacophore based fingerprint, weighted burden number and properties. For continuous descriptors, Euclidean distance was used to measure similarity. We used weighted burden number; it is a connectivity matrix containing property electronegativity, gasteiger partial charge or atomic lipophilicity, XLogP on the diagonal of the matrix. And the off-diagonal elements were weighted by one of the following values: 2.5, 5.0, 7.5 or 10.0 as a default parameter in the software in generation of 24 numerical descriptors. Finally, we computed eight descriptors; molecular weight, H-bond donors, H-bond acceptors, number of rotatable bonds, XlogP, PSA, blood-brain indicator and bad group indicator for judging the drug-like nature of a molecule.⁶⁸

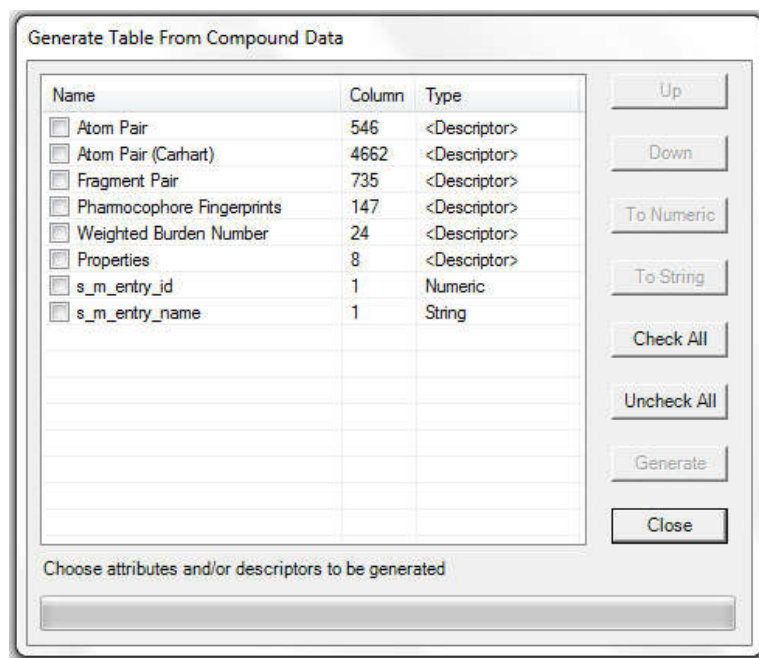


Figure 7. Generation of Biological Descriptors from table

2.3 WEKA

Biological predictive models were developed from the data mining package WEKA 3.6.2. This version was selected as it is supported by many standard data mining tasks like data preprocessing, classification, clustering, regression, visualization, and feature selection. It poses no restriction for data size and the data is modeled, visualized and evaluated statistically.⁶⁹ The original dataset contains 179 attributes but there are 180 attributes. And the reason is the addition of class attribute, which evaluates the agreement either biologically active or inactive.

All the process were carried out in WEKA “explorer” panel, the loaded dataset is displayed as histogram as shown in Figure 8. The graph is represented in red and blue color which is located at the right bottom of user interface where blue color indicates class “inactive” and red color indicates the class of “actives”. The graphs of all the attributes are also visualized in Figure 8.⁷⁰

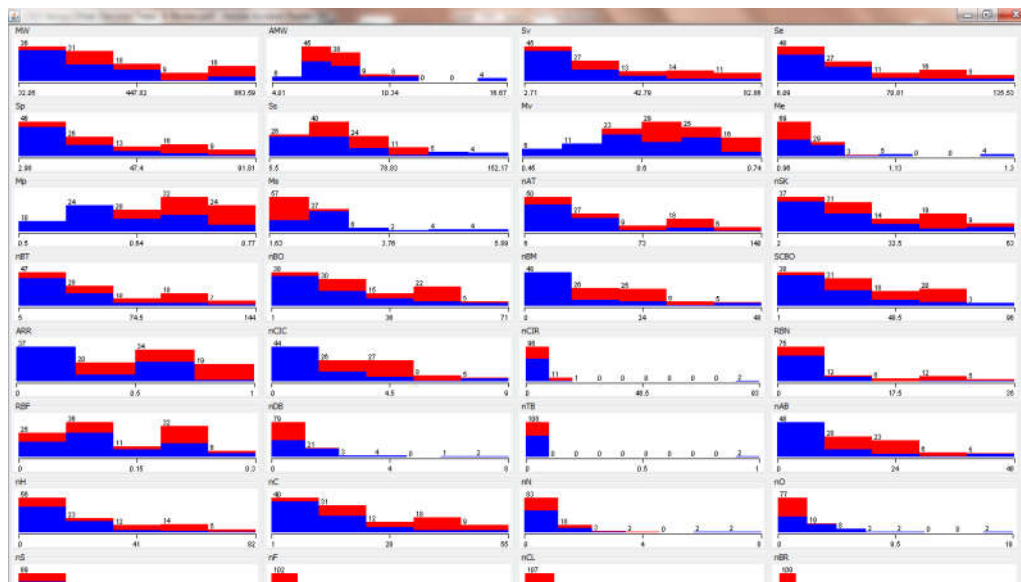


Figure 8. Demonstration of all attributes in terms of actives and inactive by WEKA software

The algorithms selected for the classification task are Naïve Bayes, Random Forest, J48 and SMO. The biological models were built, analyzed and evaluated based on their statistical parameters.

2.4 Schrodinger Suite

Schrödinger is a chemical simulation software package used in pharmaceutical, biotechnology and materials science research. Its products range from general molecular modeling programs like MacroModel to a complete suite of drug design software involving both structure and ligand based methods. Some of the applications include MacroModel (for energy minimization), sitemap (active site prediction of an enzyme or receptor), protein preparation wizard, Glide (ligand docking), QikProp (ADME predictions of drug candidates) etc.

In our study, we implemented Glide (grid-based ligand docking with energetic) for docking study. Glide is a package from Schrodinger suite that offers a full range of speed vs. accuracy options for performing a HTVS (high-throughput virtual screening) from SP (standard precision) to XP (extra precision) mode. Furthermore false positives were eliminated by extensive sampling and advanced scoring functions which will be described in the following chapter.

2.4.1 Canvas-Cheminformatics

Canvas is a cheminformatics package that provides a range of applications for structural and data analysis, including fingerprints, similarity searching, substructure searching, selection by diversity, clustering, building regression and classification models. Canvas graphical interface is project-oriented and provides chemical structure storage and organization, data analysis and visualization, and access to various other applications. The interface also provides links to Maestro that allows to easily transfer structures and data between the two applications.

The main window of Canvas has a menu bar, toolbars, spreadsheet area, project, messages view panel, and a status area. In the status area the number of rows and columns can be visualized while the spreadsheet contains the 2D structures including structure name and its properties.

2.5 Ligplot Packages

LigPlot+ is a graphical front-end to the LIGPLOT and DIMPLOT programs where ligand protein interactions are visualized from multiple 2D plots.^{71,72} The schematic 2D diagrams of protein-ligand interactions are generated provided with a given ligand in a PDB file. The diagrams portray the different interactions like hydrogen bond, hydrophobic interactions between the ligand-protein main chain and side chain elements. The system facilitates various tasks like binding of a series of small molecules to the same protein target and vice versa i.e. similarities and differences between related proteins binding on the same/similar ligand, or the same/similar ligand binding to different proteins can be highlighted. And the most general case where both protein and ligand change in analyzing ligand-protein interactions. The plot displayed in Figure 9 shows the binding of a same molecule (GSK 1365028A) to two different targets with PDB entry 2GDN and 2WKH.

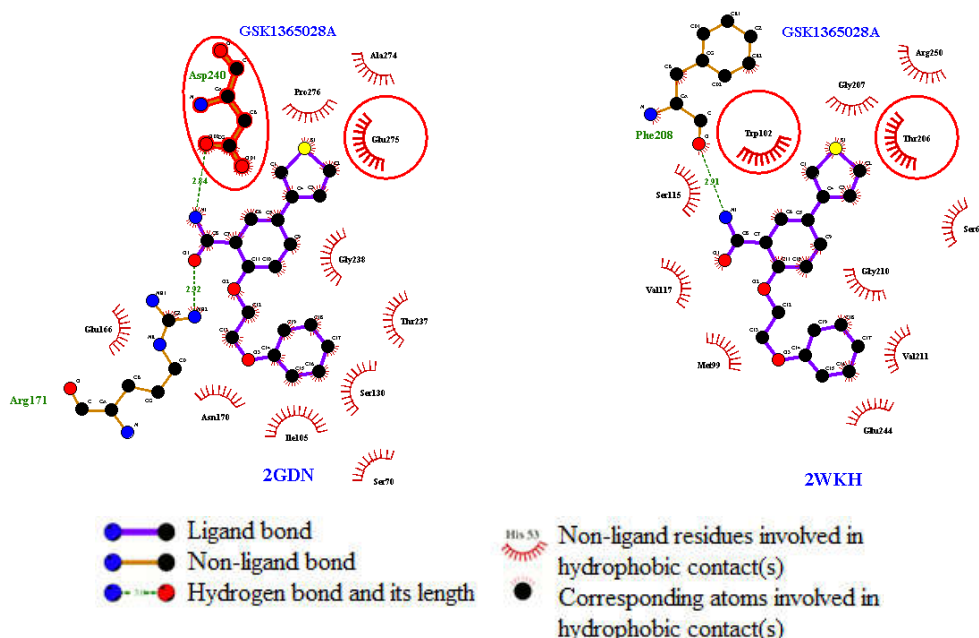


Figure 9. Ligplot interaction diagram of a protein ligand complex. Here ligand-protein interaction 2D diagrams of GSK1365028A bounded to the active site of β -lactamases 2GDN (*M. tuberculosis*) and 2WKH (*P. aeruginosa*) are displayed. The diagrams in each plot portray the hydrogen bond interactions (green dotted lines) and hydrophobic interactions (spoked semi circled arcs). The red circles and ellipses identify the equivalent residues and the side chains residues are engaged in hydrophobic interactions are highlighted with a thicker red line.

2.6 Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a mathematical construct that tries to simulate the structure and functionalities of biological neural networks.^{73,74} The basic element of every ANN is an artificial neuron, that is, a simple mathematical model (function). ANN model is based on three rules; multiplication, summation and activation.⁷⁵ An ANN is developed in a manner that the every input value is multiplied with individual weight then sum function that sums all weighted inputs and bias and at the exit sum of previously weighted inputs and bias pass through the transfer function as shown in Figure 10.⁷⁶

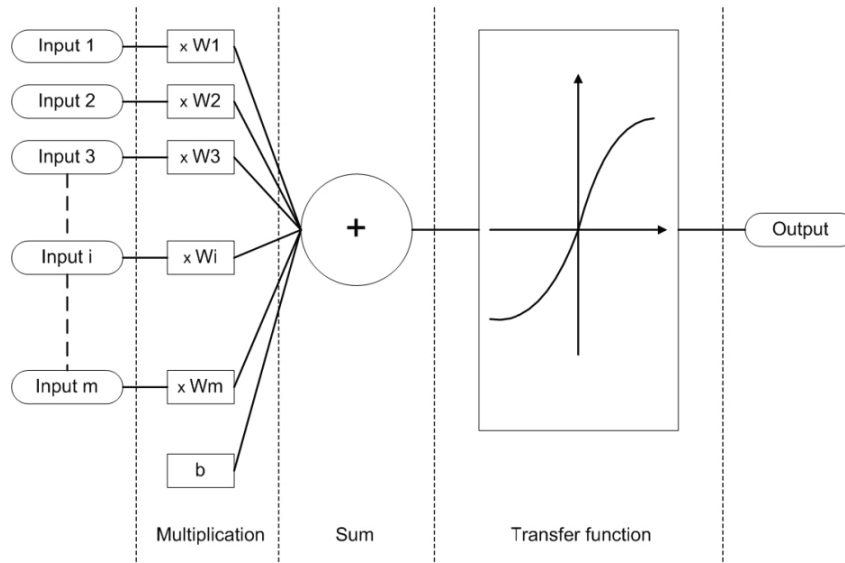


Figure 10. Working principle of an Artificial Neuron. ANN model is adapted from ref.⁷⁶

Mathematical description of an artificial neuron model is given in equation (2) below:

$$y(k) = F(\sum_{i=0}^m w_i(k) \cdot x_i(k) + b) \quad (2)$$

Where, $x_i(k)$ is input value in discrete instant k where i goes from 0 to m , $w_i(k)$ is weight value in discrete instant k where i goes from 0 to m , b is bias, F is a transfer function and $y(k)$ is output value in discrete instant k .

As seen in the equation (2) the unknown variable transfer function can be any mathematical function that defines the properties of the artificial neuron. The function is chosen on the basis of a problem that the artificial neuron (artificial neural network) needs to solve. In most cases either *step function*, *linear function* or *non-linear (Sigmoid) function* is chosen.⁷⁷ Step function is a binary function which has only two possible output values 0 and 1. That is the output value results in one value provided that it meets specific threshold and vice-versa for the value zero as described in the equation (3).

$$y = 1 \text{ if } w_i x_i \geq \text{threshold} \text{ and } 0 \text{ if } w_i x_i < \text{threshold} \quad (3)$$

This type of transfer function used in artificial neuron is called *perceptron* that is mainly used for solving classification problems and as such it can be most commonly found in the last layer of an ANN.⁷⁸

2.6.1 Self-Organizing Map (SOM)

SOM is an ANN that is related to feed-forward networks⁷⁹ where the information flow from input to output is only in one direction with no back-loops. In the feed-forward networks, there are no limitations on the number of layers, type of transfer function used in individual artificial neuron or number of connections between the individual artificial neurons.

In our study we used SOM based 2D maps that were generated from comprehensive cheminformatics computing environment (Canvas) a Schrödinger suite. In comparison to other ANNs, SOM is different in a manner that it uses a neighborhood function to preserve the topological properties of the input space. They use unsupervised learning paradigm to produce a lower-dimensional, discrete representation of the input space of the training samples, called a map that makes them especially useful for visualizing lower-dimensional views of higher-dimensional data. Here the higher dimensional data is mapped into a 2D arrangement of neurons in a hexagonal or rectangular grid.⁸⁰ The main advantages of such networks are it can detect regularities and correlations in their input and adapt their future responses to that input accordingly.

ANN has three major learning paradigms; supervised learning, unsupervised learning and reinforcement learning. In our study we used the unsupervised learning algorithm in the analysis of the biologically active compounds from the inactive. They are employed by any given type of artificial neural network architecture and each learning paradigm has many training algorithms. SOM has a broader application in solving of problems like classification, clustering, regression analysis, pattern recognition, decision making etc. The other areas include chemistry, genetics, radar systems, automotive industry, space industry, astronomy, banking, fraud detection and gaming.^{76,81}

2.7 Sampling Methods

The various sampling methods involving oversampling and undersampling are discussed in Chapter 2 of Part 1 section. In addition to the sampling technique as proposed by Chawla a new approach called Synthetic Minority Oversampling Technique (SMOTE) was used for the development of biological models.⁸² Here the minority class was oversampled by creating “synthetic” data points rather than with duplicated real data. As a part of the SMOTE algorithm the synthetic data points are generated by taking the difference between the feature vector (sample) under consideration and its nearest neighbor. Then the difference is multiplied by a random number between 0 and 1 and the output value is added to the feature vector under consideration. This approach effectively makes the minority class to become more general. When comparing SMOTE technique and oversampling technique with replication, SMOTE results in larger decision region that contain nearby minority class points while the oversampling technique the classification decision for minority class is small and more specific. And SMOTE had performed better than some of the other algorithms like Ripper's loss ratio (not studied here) and Naive Bayes.⁸³ This algorithm was implemented in WEKA software that was used for the development of SMOTE biological models.

2.8 Selection of Screening Set

For structure based and ligand based virtual screening we used the 177 anti-TB molecules from the pharmaceutical company GlaxoSmithKline (GSK) library.

The 177 molecules were the result of anti-mycobacterial phenotypic screening campaign against *M. bovis* BCG (Bacillus Calmette-Guerin), a non-virulent strain with hit confirmation in *M. tuberculosis*. The 177 molecules are confirmed compounds active against MTB strain H37Rv. The cell based screen has many advantages as they provide lead structures for further optimization within the drug development program and they were exploited as tools to identify new targets. Additionally, the whole cell screen fulfills one important criteria and that is the permeability issues which is a very troublesome in TB treatment owing to their thick nature of the mycobacterium cell wall.^{84,85}

2.9 BLAST Sequence Comparison

The amino acid sequence alignment was performed from the online SIM program.⁸⁶ The algorithm searches for the user-defined number of best non-intersecting alignments between two protein sequences or within a sequence. And the pairwise alignments were viewed from the standalone graphical viewer program LALNVIEW software.⁸⁷

CHAPTER 3

BAYESIAN MODEL AGAINST β -LACTAMASE ENZYME

3.1 Introduction

Naive Bayes classifiers are frequently used to predict molecular and pharmacological properties through cheminformatics tools and softwares.⁸⁸ A practical application of the Bayesian algorithms is in the field of VS, but they are also used in various other fields like toxicity prediction of a compound,⁸⁹ phospholipidosis mechanism, bioactivity classification for drug-like molecules etc. It is in principle, possible to use Bayesian classifiers for regression, but they are generally used as a classification. And regression is rarely seen in cheminformatics.

The Bayesian classification is carried out by estimating the probabilities of class membership and the output is from a test instance to the class with the highest estimated probability. The essence of the Bayesian algorithm lies in the mathematical approach in explaining the change in hypothesis in light of new evidence. That is the theory combines new data with their existing knowledge or expertise for the probability prediction.⁹⁰

3.2 Materials and Methods

The software's and tools used for the model build are specified in Chapter 2. For the development of ML models the datasets are provided in Chapter 2 materials and methods.

3.3 Experimental Studies

3.3.1 Dataset Preparation

PubChem BioAssay is a searchable database containing bioactivity screens of chemical substances from a variety of sources. For the current study we selected two AID datasets 434987 and 2184. The PubChem bioassay AID 434987 dataset details the inhibition of serine β -lactamase in *M. tuberculosis*. While the latter

details the inhibition of serine β -lactamase in *P. aeruginosa*. PubChem bioassays are provided with the descriptions of screening procedural conditions and readouts. The bio-active information for the AID datasets is tabulated in Table 9. The contradicting test results from bioassays, such as inconclusive bioactivity outcome were excluded from our analysis.

Table 9. Bio-active information for *M. tuberculosis* and *P. aeruginosa* bioassay datasets

Bioassay	Actives	Inactive	Inconclusive
AID 434987	372	819	11
AID 2184	97	100	0

The β -lactamase Bayesian models were developed for the bacteria *M. tuberculosis* and *P. aeruginosa*. The experiment started by searching bioassay corresponding to the selected microbes. The bioassays AID 434987 and AID 2184 was downloaded in sdf from the PubChem site (<https://pubchem.ncbi.nlm.nih.gov/bioassay/434987> and <https://pubchem.ncbi.nlm.nih.gov/bioassay/2184>). The downloaded datasets both corresponding to *M. tuberculosis* and *P. aeruginosa* was imported into descriptor generator software PowerMV for the calculation of biological descriptors. 179 molecular descriptors were calculated from pharmacophore fingerprint (147), weighted burden number (24) and properties (8). After the calculation of biological descriptors, the datasets were preprocessed, randomized, filtered and split into training set (80%) and test set (20%). The datasets were converted from comma separated value (CSV) to attribute relation file format (ARFF) in the data mining software WEKA. The training set was imported into WEKA in the preprocessor panel. The histogram indicates the descriptor distribution by two different colors as the biological class as “active” and “inactive” as shown in Figure 11.

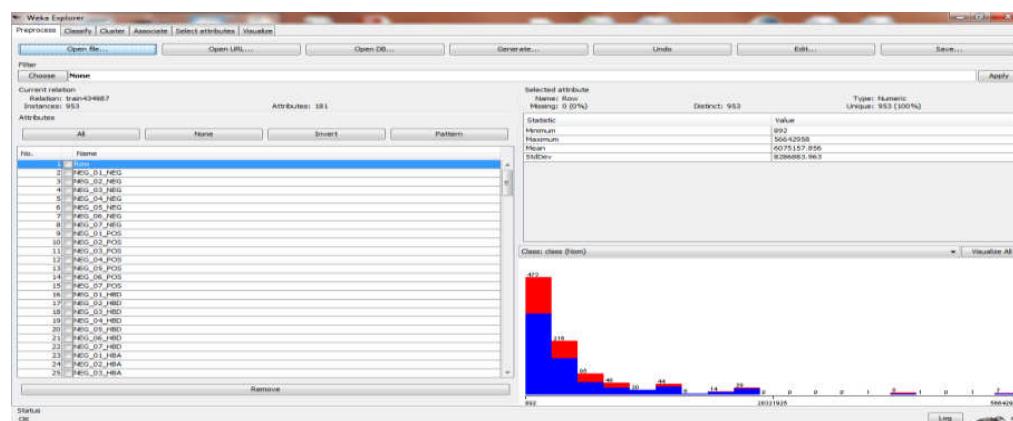


Figure 11. WEKA preprocess panel showing the biological attributes and class. Histogram indicates the descriptor distribution in two different colors as the biological class as “active” and “inactive”.

Bayesian model was developed by importing the training set and test set into the WEKA environment for the default model generation (Model 1) for AID 434987 and AID 2184. The ML models were developed through a 10 fold cross validation as mentioned in the previous chapters. The dataset with the minority class in AID 434987 has only 372 actives in comparison to 819 inactive. So performed the sampling techniques Oversampling and SMOTE for the minority class.⁹¹ Generated Model 2 and Model 3; the former corresponds to the oversampling of “active” class while the latter corresponds to the dataset with synthetic data points. For the AID 2184 default model (Model 4) was generated without performing the sampling technique. The data points of training and test set for Model 1, Model 2, Model 3 and Model 4 are given in Table 10. All the test set samples were re-evaluated upon the training set by ten by ten stratified cross validation. And finally four computational predictive Bayesian models were generated- Model 1, Model 2, Model 3 and Model 4 from WEKA software as shown in Figures 12-15.

Table 10. Number of data points of training and test set used against various ML models

		Training Set	Test Set
1.	Model 1	953	238
2.	Model 2 (TB Oversampled)	1251	312
3.	Model 3 (TB SMOTE)	1251	312
4.	Model 4 (Pseudomonas)	158	39

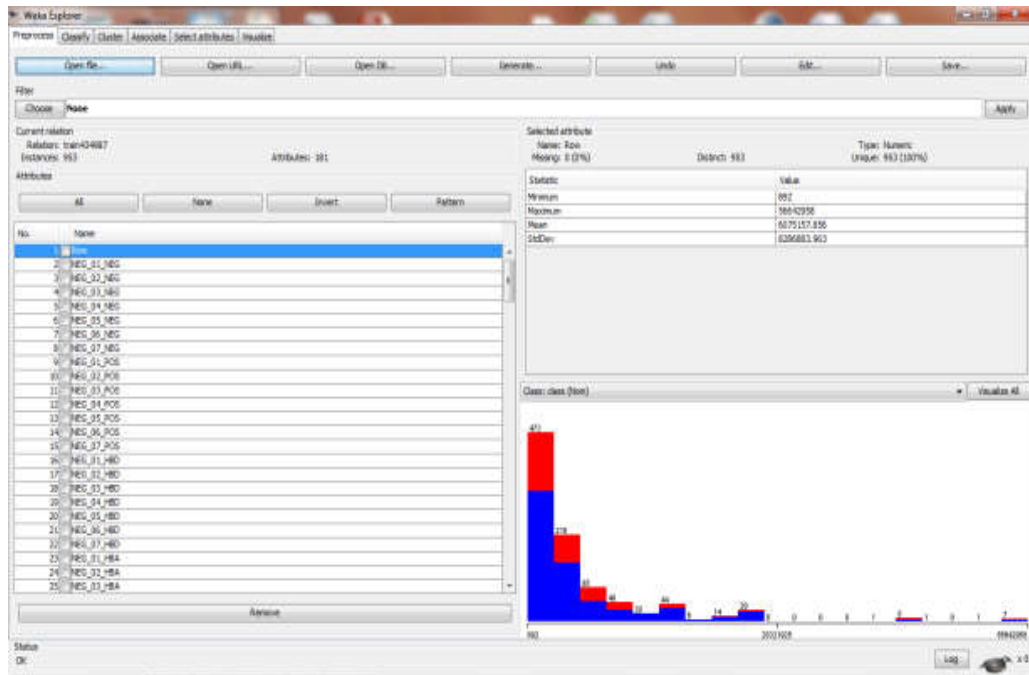


Figure 12. Generation of Bayesian Default Model 1 is displayed.

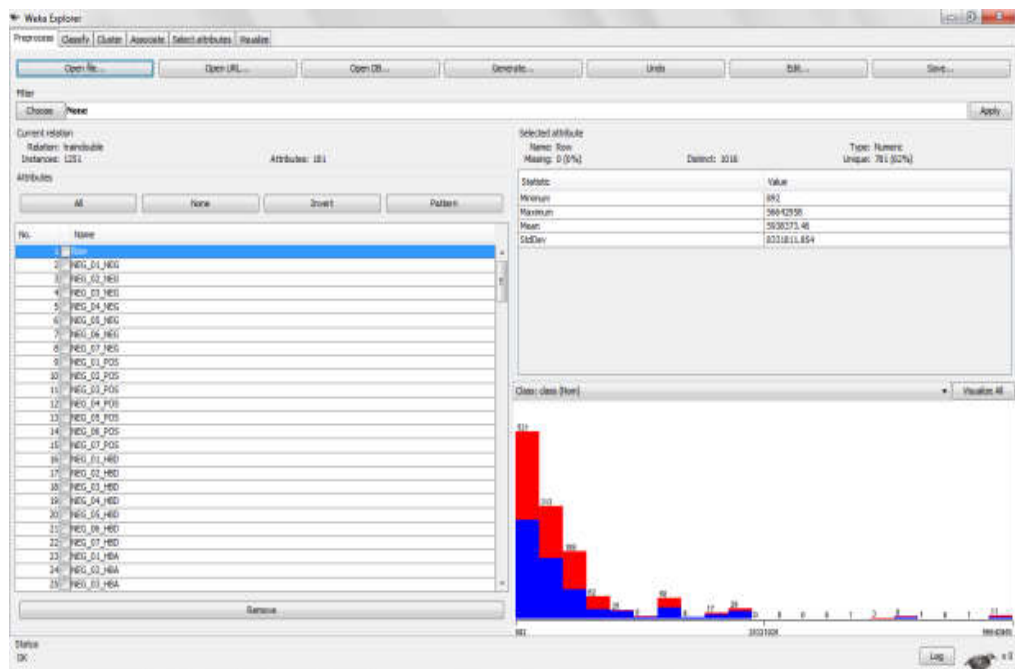


Figure 13. Generation of Bayesian Oversampled TB Model 2 panel is displayed.

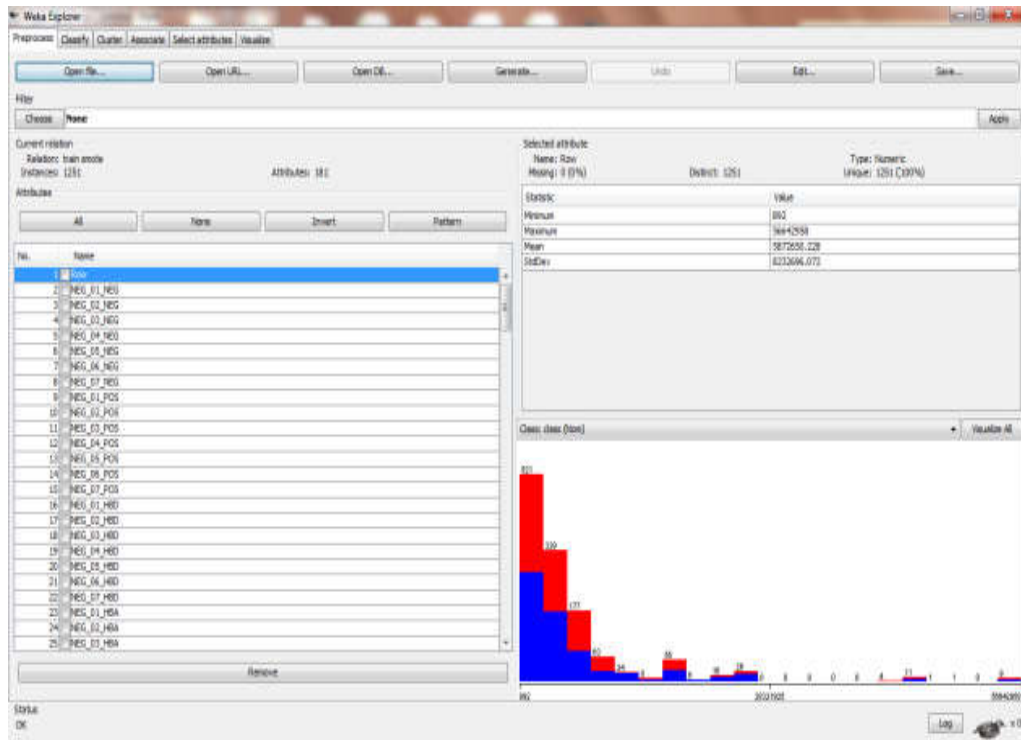


Figure 14. Generation of the Bayesian Model 3 panel is displayed.

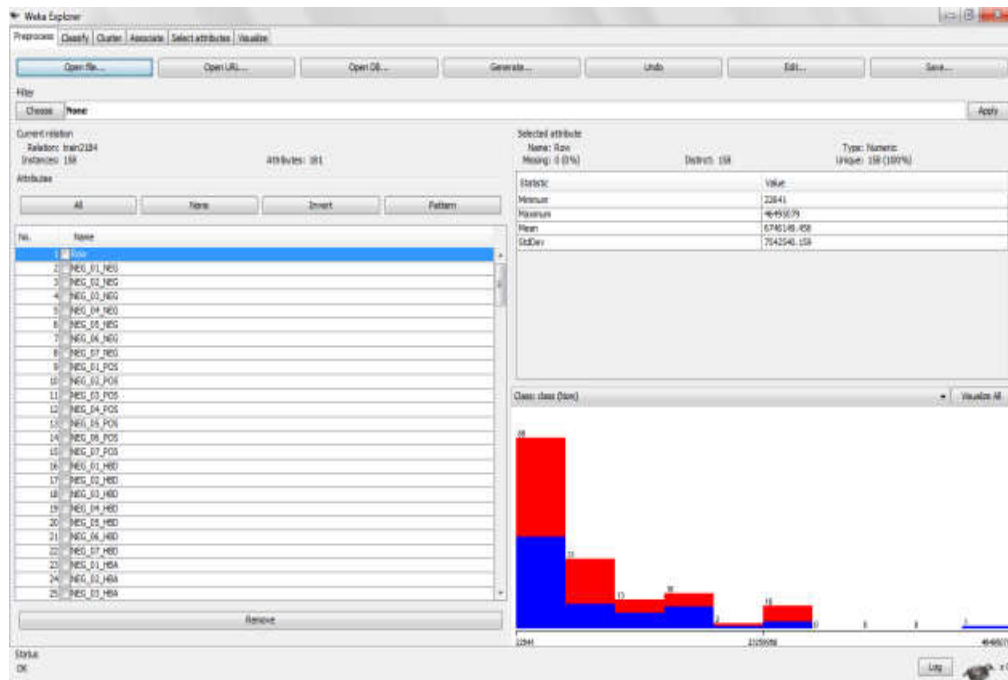


Figure 15. Generation of the Bayesian Model 4 panel is shown.

3.3.2 Results and Discussion

The performance of Bayesian algorithm was analyzed using a confusion matrix of the two class problem based on the test set re-evaluated on the training set as shown in Table 11 against Model 1, Model 2, Model 3 and Model 4.

Table 11. Confusion matrix for Bayesian model; Model 1, Model 2, Model 3 and Model 4

Model 1 <pre>a b <-- classified as 93 70 a = inactive 40 35 b = active</pre>	Model 2 <pre>a b <-- classified as 78 79 a = inactive 52 103 b = active</pre>
Model 3 <pre>a b <-- classified as 87 75 a = inactive 41 109 b = active</pre>	Model 4 <pre>a b <-- classified as 16 10 a = inactive 4 9 b = active</pre>

Table 12. Confusion matrix for biological model is given below

	Prediction as Active	Prediction as Inactive
Active	TP	FN
Inactive	FP	TN

The confusion matrix shown in Table 12 was analyzed as true positive (TP), true negative (TN), false negative (FN) and false positive (FP). TP and TN denote the number of real positives and real negatives that are classified correctly, while FN and FP denotes the number of misclassified positive and negative examples. The β -lactamase Bayesian models were developed giving more emphasis on the percentage of false negatives than percentage of false positives for compound selection. To attain this, one could minimize the number of false negatives at the expense of increasing the false positive. As mentioned previously the percentage of false positives can easily be kept in check by setting an upper limit on FP rate. Here also, the limit of FP rate was set to a maximum of 20% and cases where standard

classifiers producing this result, cost-sensitivity analysis was not used and only default classifiers were used. The fineness of the classifying algorithm was determined from true positive rate, false positive rate, accuracy, precision, recall, F-measure, ROC, kappa, MCC etc. Various statistical parameters were calculated upon the test set against all the Bayesian models (Model 1, Model 2, Model 3 and Model 4) are given in Table 13.

Table 13. The evaluation measures TP, TN, FP, FN, Recall, Precision, F-measure, ROC, Accuracy and Kappa generated by Model 1, Model 2, Model 3 and Model 4.

Statistical Measures	Model 1 (TB Default)	Model 2 (TB Oversampled)	Model 3 (TB SMOTE)	Model 4 (Pseudomonas)
TP	35	103	109	9
TN	93	78	87	16
FP	70	79	75	10
FN	40	52	41	4
TP rate %	46.7	66.5	72.7	69.2
Fp rate %	42.9	50.3	46.3	38.5
Precision	33.3	56.6	59.2	47.4
Recall	46.7	66.5	72.7	69.2
Specificity	57.05521	49.6815	53.7037	61.1538
BAC	51.8776	58.0907	63.2018	65.1769
F-measure	38.9	61.1	65.3	56.3
ROC	53.3	61.9	67.4	74
Accuracy	53.7815	58.0138	62.8205	64.1026
Kappa	0.0336	0.1612	0.2615	0.2759
MCC	0.03483	0.2197	0.2678	0.2901

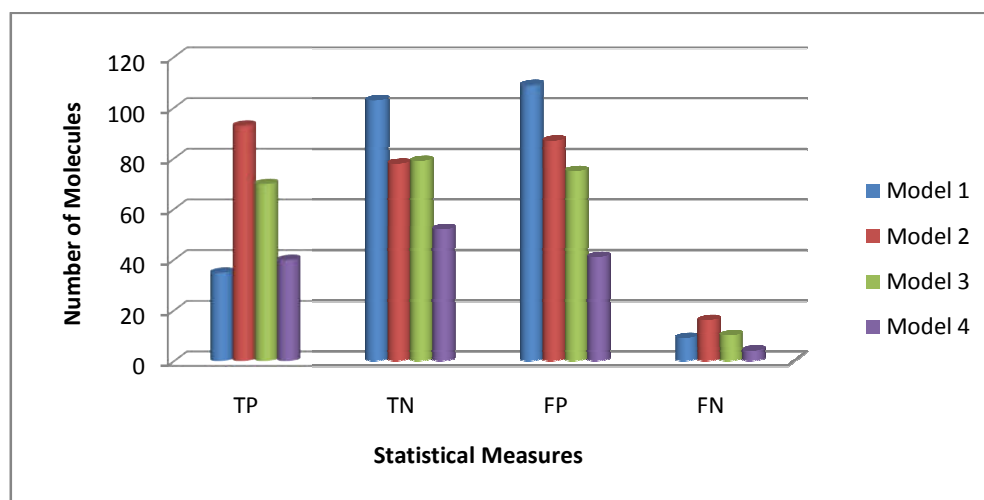


Figure 16 (a). Number of molecules based on confusion matrix

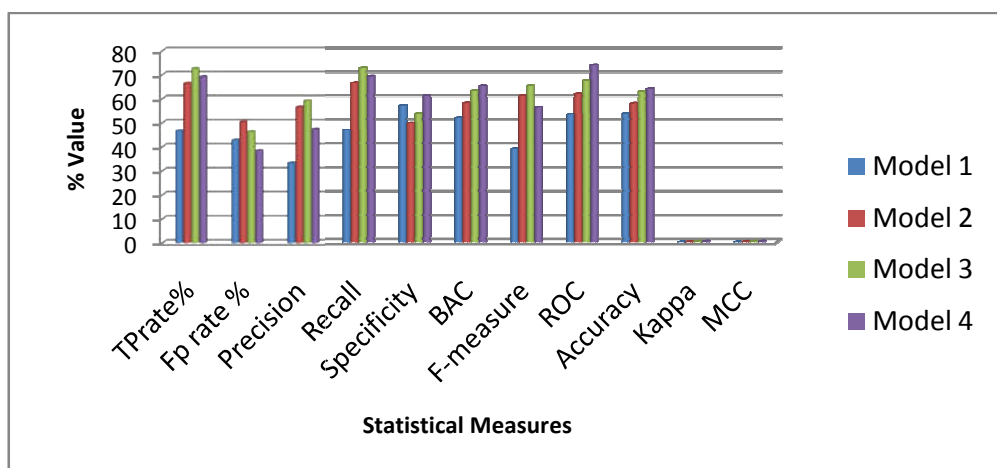


Figure 16 (b). Comparative graph between the various statistical measures of classification models-Model 1, Model 2, Model 3 and Model 4 are presented.

3.3.2.1 Robustness of the Model

Various statistical binary classification performance measures were used to evaluate the performance of the β -lactamase anti-bacterial ML models. The models accuracy was found to be the highest around 64% for the pseudomonas (Model 4) and least for the TB Model 1. The same trends were maintained for the other measures like kappa, MCC, ROC and BAC. For the biased datasets the oversampled

and SMOTE models were found to be very positive as it enhanced the model accuracy which is 62.8% while the oversampled model ended with 58%. The overall robustness of the Bayesian anti-bacterial models (Model 1, Model 2, Model 3 and Model 4) was judged from the statistical parameter accuracy. Here also the balanced accuracy was evaluated for each model was studied as shown in Table 13. The BAC was highest for Model 4 (65%) among all four models. But when compared among TB predictive models Model 3 (63.2%) performed well in comparison to Model 2 (58%) and Model 1 (51%). We checked the TP rate and FP rate. All the models couldn't achieve the threshold FP rate (20%). The models displayed a higher TP rate in the case of oversampled TB (66.5%) and SMOTE models (72%). The TP rate was least for the TB default model (46.7%). Further the default TB and the oversampled and SMOTE models were compared for the reason to check the robustness of the anti-bacterial model when sampling methods were applied. The results were very interesting as statistical parameters TP rate, precision, recall, BAC, f-measure, ROC and accuracy has increased from TB default model through TB oversampled model to TB SMOTE model. Figure 16 is the comparative graph between the various evaluation measures of the classification models-Model 1, Model 2, Model 3 and Model 4. From our study we understand that Model 3 (best among TB predictive models) and Model 4 (best among all anti-bacterial ML models) are efficient models. Later GSK 177 anti-TB molecules were screened against all the β -lactamase anti-bacterial ML models.

3.3.2.2 Virtual Screening and Validation of the Model

As mentioned previously virtual screening has been one of the mainstays in the identification of hits in a general drug discovery program. The cost and time spent in running high-throughput screens are enormous. Computational virtual screening being a cheaper method could further benefit provided with faster processors, parallel computing, smarter and faster algorithms in prioritizing compound selection. The GSK 177 anti-TB molecules were selected for the virtual screening. The screening set was prepared in the similar way as the training sets and test set. All the molecules in the screening set were geometry optimized. The energy

minimized 177 GSK molecules were imported into PowerMV software and calculated 179 molecular descriptors; Pharmacophore fingerprint (147), weighted burden number (24) and properties (8). Then the screening set was preprocessed by adding a new class to the existing column for the prediction of active or inactive. The screening set was converted from comma separated value (CSV) to attribute relation file format (ARFF) from the data mining software WEKA. Each Bayesian models (Model 1, Model 2, Model 3 and Model 4) developed were imported into WEKA panel individually and the parameters were set to output prediction and virtual screening was performed. As a result some of the GSK 177 molecules were screened against β -lactamase Bayesian model for *M. tuberculosis* and *P. aeruginosa* is displayed in Table 14.

Table 14. No. of computationally active β -lactamase inhibitors against each models

Models	No. of Molecules Screened
Model 1 (TB)	54
Model 2 (TB Oversampled)	102
Model 3 (TB SMOTE)	108
Model 4 (Pseudomonas)	34

3.4 Conclusion

In this chapter, we have discussed briefly on biological predictive models and virtual screening techniques based on Naive Bayesian Classifier (NBC). Bayesian models are commonly used in the area of *in silico* drug design program due to its simplicity and less computational cost and time. In the present work, Bayesian models were developed for the target β -lactamase enzyme present in *M. tuberculosis* and *P. aeruginosa*. The PubChem database bioassays AID 434987 and AID 2184 were used as the starting material for the model build. Computational models were developed using biological descriptors from PowerMV software that included various descriptors including pharmacophore fingerprint, weighted burden and drug like properties. The 179 biological descriptors enriched the dataset for developing computational models that were used to screen 177 anti-TB molecules from GSK library. Before model development the biased dataset (AID 434987) was

treated with two forms of sampling technique: (i) Oversampling, (ii) SMOTE. Models were built using WEKA software by using ML classifier Naïve Bayes. As a result four models were developed against the target β -lactamase enzyme present in *M. tuberculosis* and *P. aeruginosa*. The robustness of the models was studied in terms of various statistical parameters accuracy, kappa, ROC, sensitivity, specificity etc. As an outcome, many of the molecules from 177 GSK were reported to be computationally β -lactamase inhibitors.

CHAPTER 4

DECISION TREE MODEL AGAINST β -LACTAMASE ENZYME

4.1 Decision Tree Models

In this study, independent evaluation of the DT based models (Random Forest and J48) was attempted by using two HTS assays, PubChem AID 434987 and AID 2184, which were aimed at identifying β -lactamase inhibitors in *M. tuberculosis* and *P. aeruginosa*. As discussed in the previous chapter a binary DT algorithm is a tree-like structure, where the parent node is split into child nodes by calculating the best feature split determined by a chosen split criterion. Subsequently child nodes are further divided until all the observations are classified. The classification ends up with a group where each group member represents more common features as a homogenous set can be realized and evaluated.^{92,93} In our case leaves represent the biological class labels, i.e. “active” or “inactive” while the branches correspond to conjunctions of input features that resulted in those outcomes.

4.1.1 Random Forest

As discussed in the previous chapter RF is a classification algorithm based on an ensemble or forest of decision trees. The main advantage of RF is large number of independent trees allows RF to benefit from the *wisdom of crowds* effect.^{94,95} Here the trees are built using training multiple biological features for each of a training set of molecules. The algorithmic architecture is briefly discussed in the previous chapter section part 1 and the decision is based on the biological attributes in differentiating β -lactamase actives from inactive.

RF has proven to be a very successful method in cheminformatics and bioinformatics. These include QSAR, mutagenicity, phospholipidosis, hERG blockade and skin sensitization, postdock scoring functions and predicting protein–

ligand binding affinity, genetic epidemiology (response is categorical either diseased/healthy).⁹⁶

4.1.1.1 Materials and Methods

The softwares and online web servers used for the model build are specified in Chapter 2. The datasets used in this experiment were downloaded from PubChem bioAssay (AID 434987 and AID 2184).

4.1.1.2 Experimental procedure

The DT algorithm carried out in the present study is RF from the data mining package WEKA. For the model construct classification experiments we increased the heap-size to 4 GB to handle out-of-memory exceptions for large datasets. The training and test sets that were used to build β -lactamase Bayesian model were used for DT analysis with class nominal “active” and “inactive”.

RF model was developed by importing the training set and test set into the WEKA environment for the default β -lactamase RF model generation (RF Model 1 for AID 434987 and RF Model 4 for AID 2184). The dataset with the minority class in AID 434987 has only 372 actives in comparison to 819 inactive. So performed the sampling techniques Oversampling and SMOTE for the minority class. Generated RF Model 2 and RF Model 3; the former corresponds to the oversampling of “active” class while the latter corresponds to the dataset with synthetic data points. For the AID 2184 default model (RF Model 4) was generated without performing sampling technique. The data points of training and test set for RF Model 1, RF Model 2, RF Model 3 and RF Model 4 are given in Table 15.

Table 15. Data points of training set and test set used in the study

	Training set	Test set	Attributes
1. RF Model 1 (TB Default)	953	238	179
2. RF Model 2 (TB Oversampled)	1251	312	179
3. RF Model 3 (TB SMOTE)	1251	312	179
4. RF Model 4 (Pseudomonas)	158	39	179

The training set for all the models were evaluated by a 10 fold cross validation. After these processes, four models were generated, one corresponded to *P. aeruginosa* while the remaining three to *M. tuberculosis*. The models were built from the WEKA Generic Object Editor-weka.classifiers.trees.RandomForest as shown in Figure 17 by providing training set and test set as mentioned below.

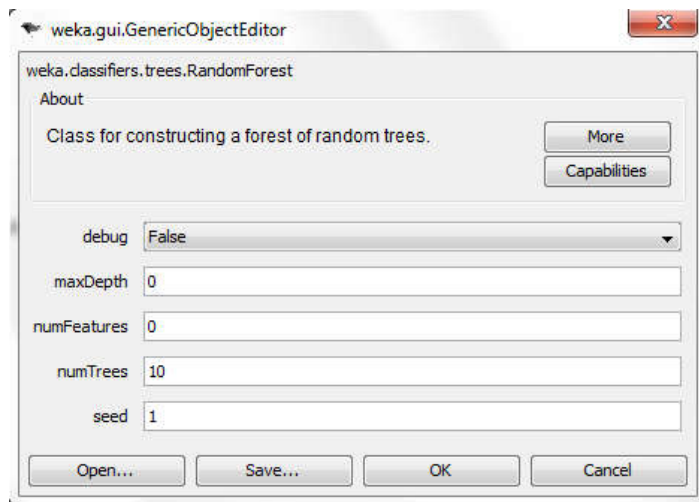


Figure 17. WEKA Generic Object Editor displays the parameters set for the model generation

All the test set samples were re-evaluated upon the training set by ten by ten stratified cross validation. While constructing the RF classifier model the total number of trees that were used to construct a RF considering random features with out-of-bag error is tabulated in Table 16. And finally four computational β -lactamase RF models were generated, RF Model 1 (TB Default), RF Model 2 (TB Oversampled), RF Model 3 (TB SMOTE) and RF Model 4 (Pseudomonas).

Table 16. RF model constructed with the following number of trees with Out-of-Bag error

Models	Total no. of trees	No. of random features	Out-of-Bag error
RF Model 1 (TB Default)	10	08	0.3578
RF Model 2 (TB Oversampled)	10	08	0.2502
RF Model 3 (TB SMOTE)	10	08	0.3837
RF Model 4 (Pseudomonas)	10	08	0.3797

4.1.1.3 Results and Discussion

The performance of Bayesian algorithm was analyzed using a confusion matrix of two class problem based on the test set re-evaluated on the training set is shown in Table 17 against RF Model 1, RF Model 2, RF Model 3 and RF Model 4.

Table 17. Confusion matrix generated against anti-bacterial RF ML models

RF Model 1			RF Model 2		
a	b	<-- classified as	a	b	<-- classified as
143	20	a = inactive	134	23	a = inactive
59	16	b = active	24	131	b = active
RF Model 3			RF Model 4		
a	b	<-- classified as	a	b	<-- classified as
137	25	a = inactive	18	8	a = inactive
58	92	b = active	3	10	b = active

All the results of RF predictive models (RF Model 1, RF Model 2, RF Model 3 and RF Model 4) were re-evaluated upon the independent test set and various statistical performance matrices have been tabulated in Table 18.

Table 18. Statistical parameters for the RF Model 1, RF Model 2, RF Model 3 and RF Model 4

Statistical Parameters	RF Model 1	RF Model 2	RF Model 3	RF Model 4
TP	16	131	92	10
TN	143	134	137	18
FP	20	23	25	8
FN	59	24	58	3
TP rate %	21.3	84.5	61.3	76.9
Fp rate %	12.3	14.6	15.4	30.8
Precision	44.4	85.1	78.6	55.6
Recall	21.3	84.5	61.3	76.9
Specificity	87.73	85.3503	84.5679	69.2307
BAC	54.515	84.3503	72.9333	73.0653
F-measure	28.8	84.8	68.9	64.5
ROC	59	91.5	80.9	82.5
Accuracy	66.8067	84.9359	73.3974	71.7949
Kappa	0.1054	0.6987	0.4628	0.4211
MCC	0.1251	0.6987	0.4737	0.4364

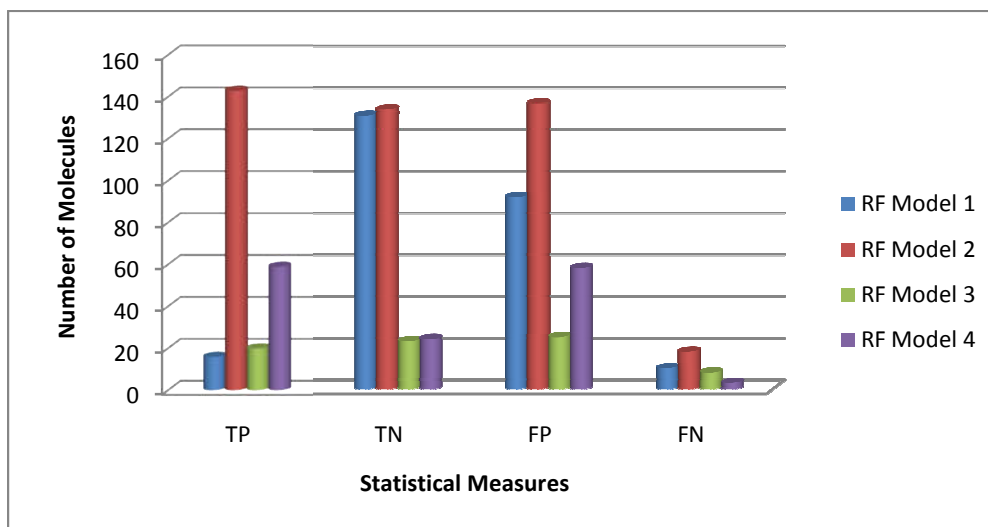


Figure 18 (a). Number of molecules based on confusion matrix

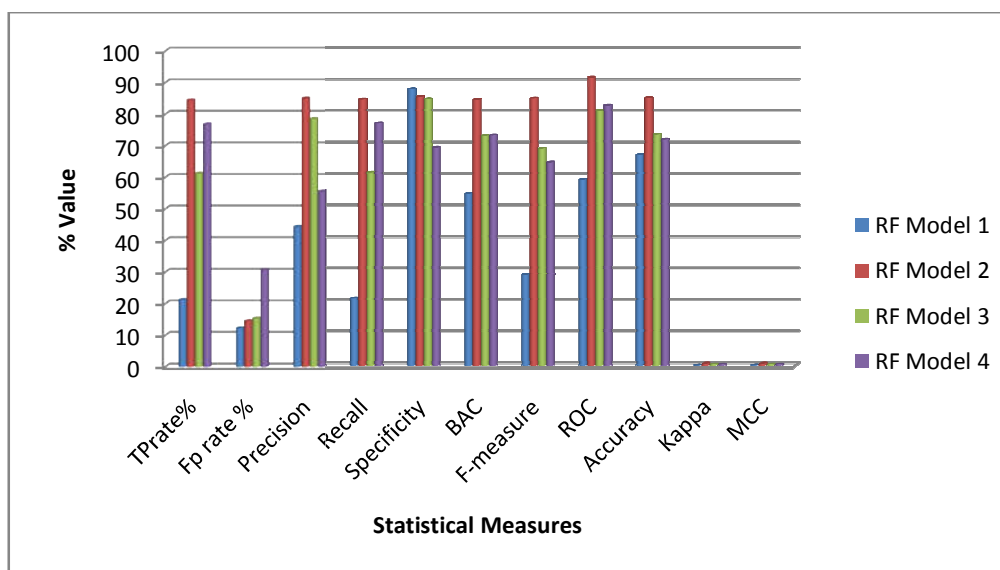


Figure 18 (b). Comparative graph between the various statistical measures of classification models-RF Model 1, RF Model 2, RF Model 3 and RF Model 4 are presented.

The Confusion Matrix was analyzed from TP, TN, FN and FP. The β -lactamase RF models were developed giving more emphasis on the percentage of FN than percentage of FP for compound selection as mentioned in the anti-bacterial Bayesian models. As mentioned previously the percentage of FP can easily be kept

in check by setting an upper limit on FP rate to 20% without usage of the cost-sensitivity analysis and models were built on default parameters. The fineness of the classifying algorithm was determined from TP rate, FP rate, accuracy, precision, recall, F-measure, ROC, kappa, MCC etc. All the statistical parameters were calculated upon the test set against all the RF models (RF Model 1, RF Model 2, RF Model 3 and RF Model 4) are given in Table 18.

4.1.1.4 Robustness of the models

Various statistical binary classification performance measures were used to evaluate the performance of the RF ML models. The models accuracy were found to be the highest around 84% for the TB oversampled (RF Model 2) and least for the default TB RF Model 1. For the biased datasets the oversampled and SMOTE models were found to be very positive as it enhanced the model accuracy which is 84% for RF Model 2 and 73% for RF Model 3. The overall robustness of the random forest anti-bacterial models (RF Model 1, RF Model 2, RF Model 3 and RF Model 4) was judged from the statistical parameter accuracy. Since the dataset for TB was biased balanced accuracy was evaluated for each model as show in Table 18. The BAC was highest for RF Model 2 (84.3%) and least for RF Model 1 (54.5%) among TB models. Also the RF Model 2 has showcased greater performance for the parameters like kappa, MCC, ROC, f-measure, BAC, recall and precision. And the only parameter that produced lesser value was against specificity. Among the TB models, RF Model 2 had outperformed in comparison to the TB default and SMOTE models. For the biased datasets the oversampled RF Model 2 found to be very positive as it enhanced the model accuracy. We checked the TP rate and FP rate. The FP rate was found to be gradually increasing for the TB models (default-oversampled-SMOTE) but the values were under threshold FP rate (20%). For the pseudomonas model, FP rate was higher but TP rate was higher in comparison to TB default and SMOTE models. The oversampled model had outperformed in the DT model which shows the significance of the doubling of the data points in the active region in developing better ML anti-bacterial models. Figure 18 displays comparative graph between the various evaluation measures of

the DT models-RF Model 1, RF Model 2, RF Model 3 and RF Model 4. From our study we understand that RF Model 2 has performed better in respect to all the random forest anti-bacterial ML Models. Later the screening set GSK 177 anti-TB molecules (as prepared in the prescribed procedure) were screened against all the β -lactamase anti-bacterial RF ML models. As a result we could prioritize many GSK anti-TB molecules as computationally active β -lactamase inhibitors against the two targets. The numbers of the screened results are mentioned in Table 19.

Table 19. No. of computationally active β -lactamase inhibitors against RF Models

Models	No. of Molecules Screened
RF Model 1 (TB Default)	17
RF Model 2 (TB Oversampled)	26
RF Model 3 (TB SMOTE)	29
RF Model 4 (Pseudomonas)	55

4.2 J48 Decision Tree Algorithm

The J48 algorithm had been briefly discussed in Part I section. In the current study β -lactamase J48 predictive models were developed and the algorithmic architecture was the modified C4.5 algorithm implemented in the WEKA 3.6 software. Here also the classification decision tree for the given dataset is carried out by recursive partitioning of data using Depth-first strategy.^{97,98,99} Here biological models were developed that are based on the drug related descriptors having a class variable as “active” and “inactive”. The class signifies the molecules activity against β -lactamase present in *M. tuberculosis* and *P. aeruginosa*.

4.2.1 Materials and Methods

The softwares and online web servers used for the J48 model build were specified in Chapter 2. The datasets used in this experiment were downloaded from PubChem bioassay (AID 434987 and AID 2184). The screening set was prepared from the GSK library consisting of 177 anti-TB molecules.

4.2.2 Experimental Procedure

The J48 models against β -lactamase target were constructed from WEKA “explorer” platform. The heap-size was configured to 4 GB in order to handle out-of-memory exceptions for large datasets. The training set and test that were used to build anti-bacterial Bayesian model were used for J48 tree analysis with class nominal β -lactamase actives as “active” and β -lactamase inactive as “inactive”. The parameters set for the J48 model generation is given in Figure 19.

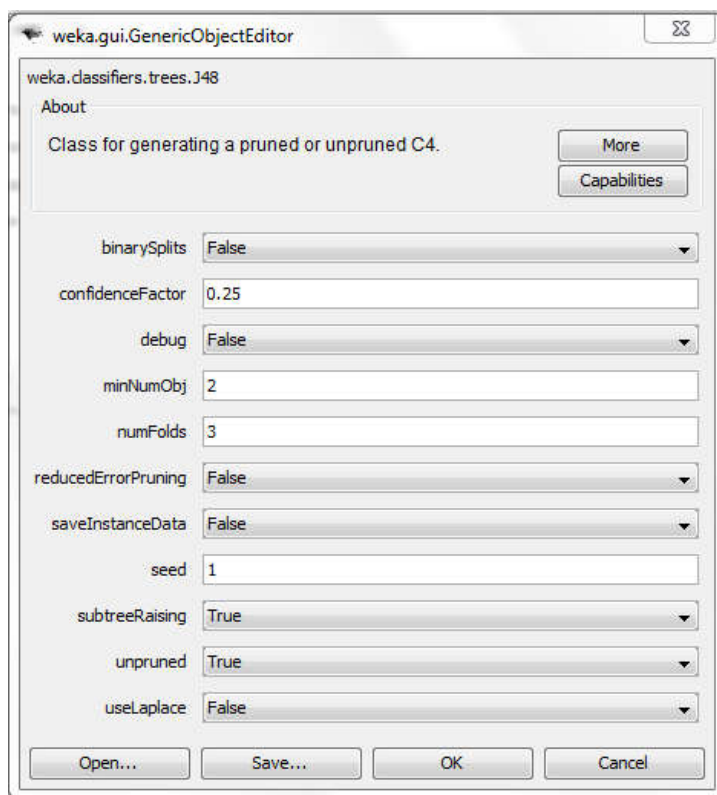


Figure 19. WEKA Generic Object Editor for J48 classifier

4.2.2.1 J48 Model Generation

The models were built from the WEKA Generic Object Editor - weka.classifiers.trees.J48 as shown in Figure 19. The J48 ML Models were constructed against the dataset AID 434987 and AID 2184 in a similar way as developed against anti-bacterial RF models. The dataset was randomized, reordered and split into 10 folds of equal size, one fold was used for testing and rest of them

were used for training the classifier in an iterative manner. For the model build the parameter “unpruned” was set to “*True*” from the Generic Object Editor. The J48 trees were built after preprocessing and 10 fold stratified cross-validation of the training set.

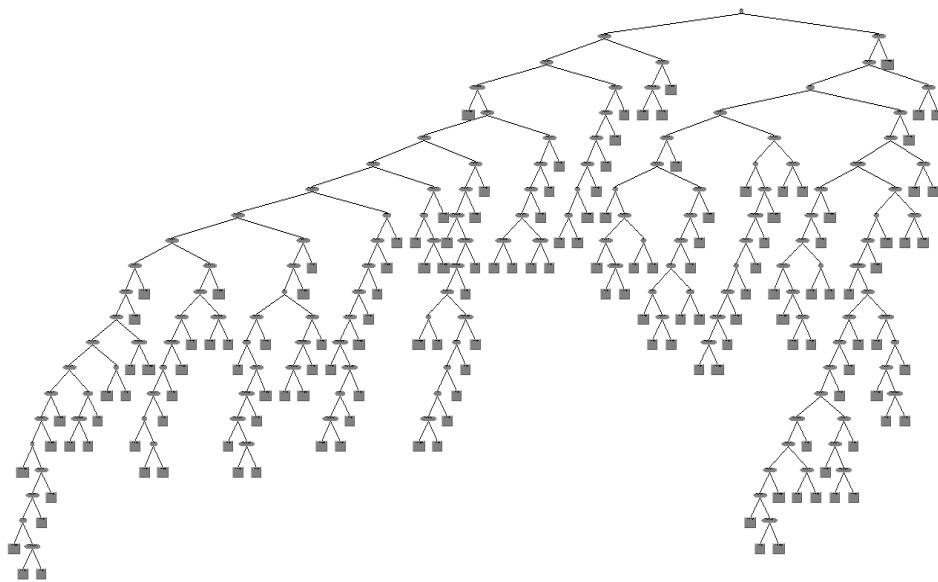
J48 model was developed by importing the training set and test set into the WEKA environment for the default β -lactamase J48 ML model generation (J48 Model 1) for AID 434987. For the AID 434987 dataset oversampling and SMOTE sampling methods were carried out for the minority class (actives). And generated J48 Model 2 and J48 Model 3 former corresponds to the oversampling of “active” class while the latter corresponds to the dataset with synthetic data points. For the AID 2184 default model (J48 Model 4) was generated without performing sampling technique, here also the “unpruned” parameter was set to “*True*” for the model generation. Each J48 models developed ended up with a tree size consisting of numbers of leaves charted below in Table 20.

Table 20. Displays the number of trees and corresponding leaves generated by each J48 models

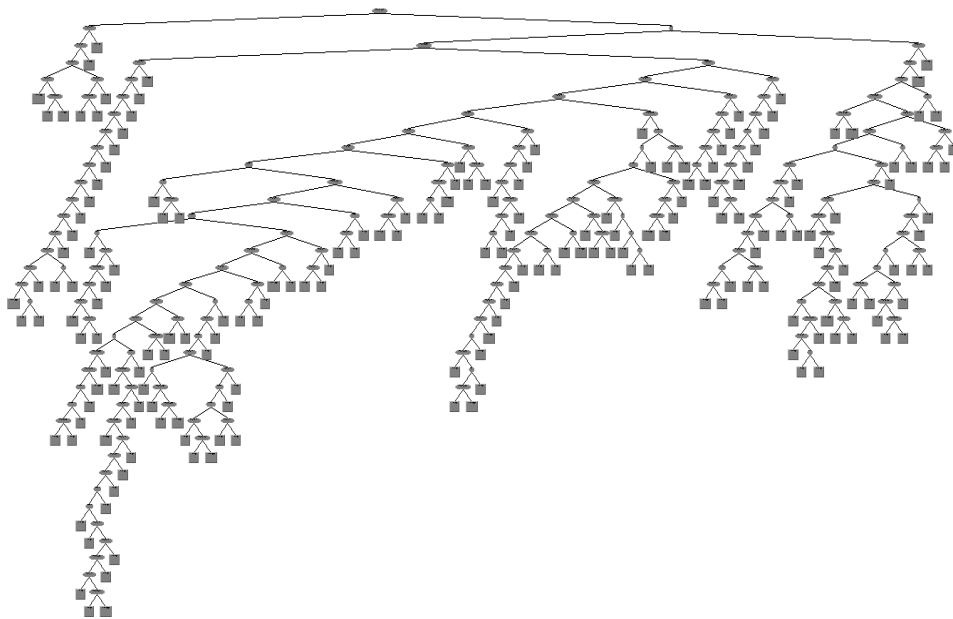
Models	Tree size	No. of leaves
J48 Model 1 (TB Default)	149	297
J48 Model 2 (TB Oversampled)	196	391
J48 Model 3 (TB SMOTE)	159	317
J48 Model 4 (Pseudomonas)	26	51

All the J48 ML Models were constructed. All the test set was re-evaluated upon the training set and performance of the models and accuracy were studied from various statistical parameters. The data points of training and test set for J48 Model 1, J48 Model 2, J48 Model 3 and J48 Model 4 are given in Table 15 (Section RF). The J48 tree developed was visualized from the WEKA console as shown in Figure 20 (a-d).

(a) Model 1 J48 Tree



(b) Model 2 J48 Tree



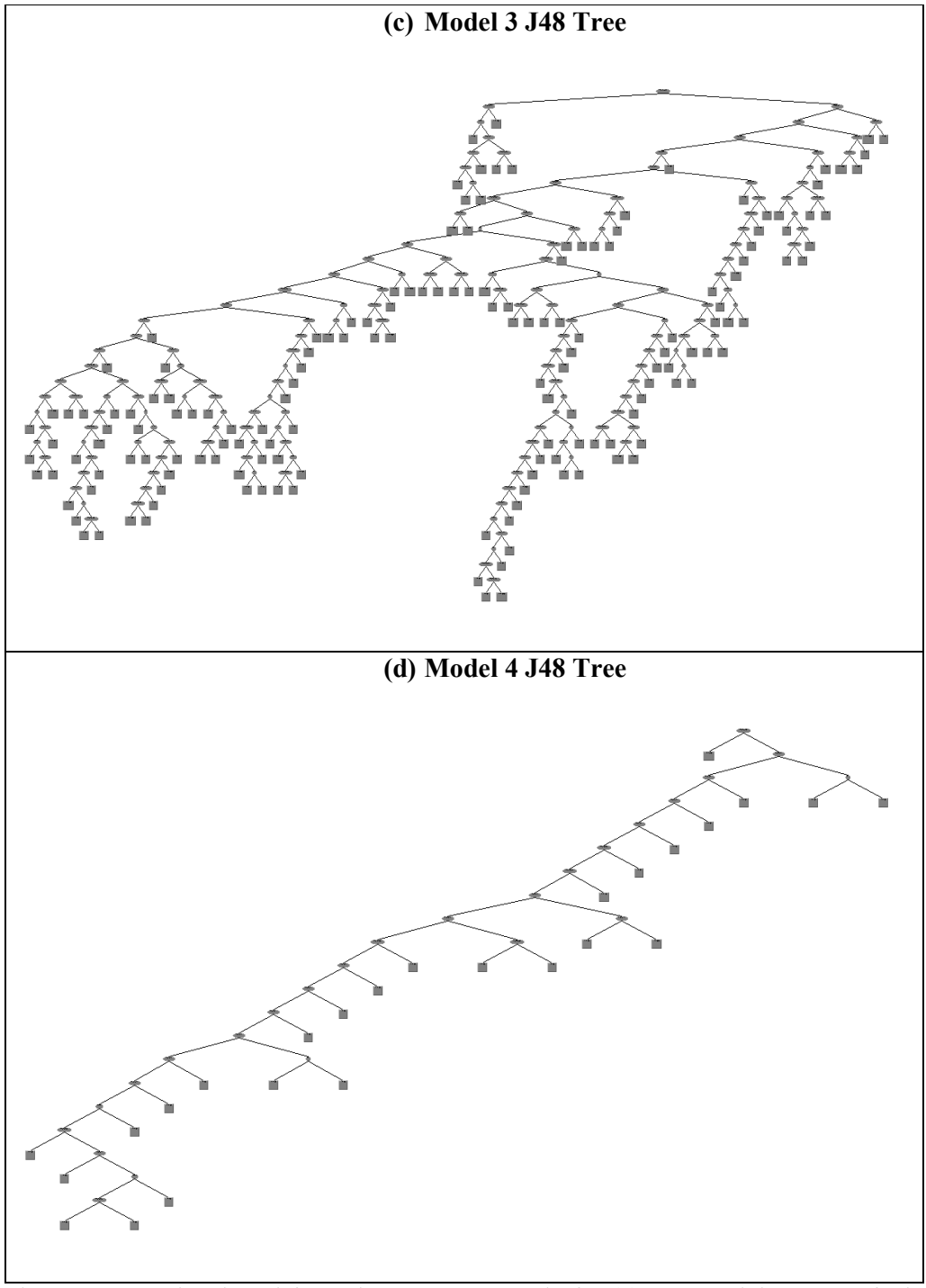


Figure 20. DT developed from the WEKA console for (a) J48 Model 1 TB Default, (b) J48 Model 2 TB Oversampled, (c) J48 Model 3 TB SMOTE, (d) J48 Model 4 Pseudomonas

4.2.2.2 Results and Discussion

All the predictive anti-bacterial J48 models were re-evaluated upon the independent test set and the confusion matrix was generated as shown in Table 21.

Table 21. Confusion matrix generated against anti-bacterial J48 ML models

<p>J48 Model 1 TB Default</p> <pre> a b <-- classified as 112 51 a = inactive 49 26 b = active </pre>	<p>J48 Model 2 TB Oversampled</p> <pre> a b <-- classified as 111 46 a = inactive 40 115 b = active </pre>
<p>J48 Model 3 TB SMOTE</p> <pre> a b <-- classified as 113 49 a = inactive 59 91 b = active </pre>	<p>J48 Model 4 Pseudomonas</p> <pre> a b <-- classified as 20 6 a = inactive 4 9 b = active </pre>

4.2.2.3 Robustness of the Models

All the results of J48 anti-bacterial predictive models (J48 Model 1, J48 Model 2, J48 Model 3 and J48 Model 4) was re-evaluated upon the independent test set and various statistical performance matrices are tabulated in Table 22.

Table 22. The number of TP, TN, FP, FN, TP rate, FP rate and the evaluation measures Precision, Recall, F-measure, ROC, Accuracy and Kappa generated by the J48 models

Statistical Parameters	J48 Model 1	J48 Model 2	J48 Model 3	J48 Model 4
TP	27	115	91	8
TN	109	111	113	20
FP	54	46	49	6
FN	48	40	59	5
TP rate %	36	74.2	60.7	61.5
Fp rate %	33.1	29.3	30.2	23.1
Precision	33.3	71.4	65	57.1
Recall	36	74.2	60.7	61.5
Specificity	66.8711	70.7006	69.753	76.923
BAC	51.4355	72.4503	65.2265	69.2115
F-measure	34.6	72.8	62.8	59.3
ROC	51.8	74.4	67.3	67
Accuracy	57.1429	72.4359	65.3846	71.7949
Kappa	0.0281	0.4488	0.305	0.3774
MCC	0.23903	0.44916	0.3055	0.37796

The model robustness was analyzed from the confusion matrix. The β -lactamase J48 anti-bacterial models were developed giving more emphasis on the percentage of false negatives than percentage of false positives for compound selection as mentioned in the anti-bacterial Bayesian models. As mentioned previously the percentage of FP rate was kept with an upper limit 20%. Here we did not apply cost-sensitivity analysis and models were built on default parameters. The fineness of the classifying algorithm was determined from number of TP, FP, FN, TN, TP rate, FP rate, accuracy, precision, recall, F-measure, ROC, kappa, MCC etc. All the statistical parameters were calculated upon the test set against all the J48 anti-bacterial models (J48 Model 1, J48 Model 2, J48 Model 3 and J48 Model 4) are given in Table 22.

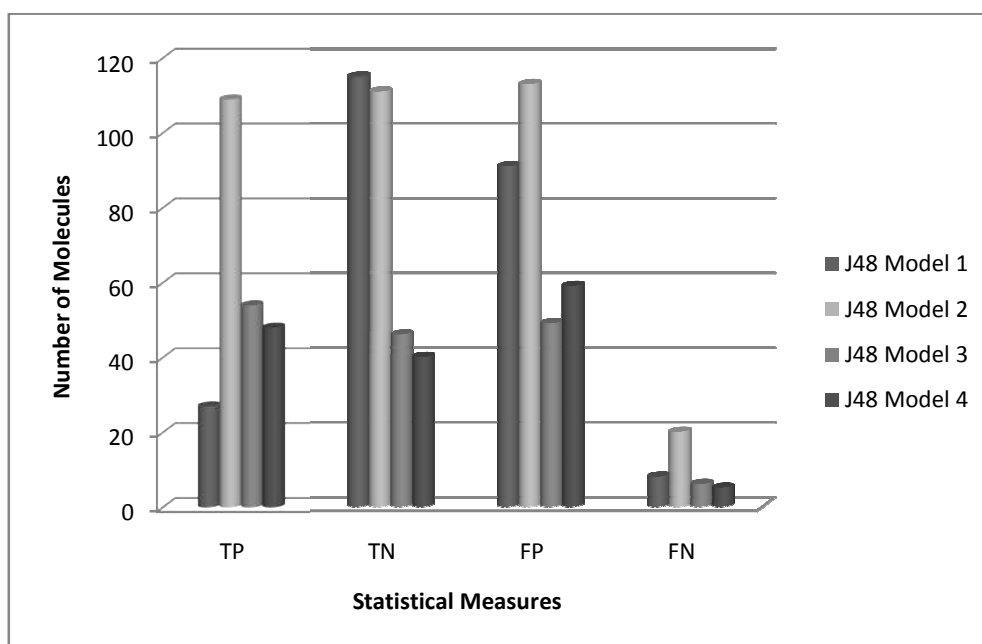


Figure 21 (a). Number of molecules based on confusion matrix

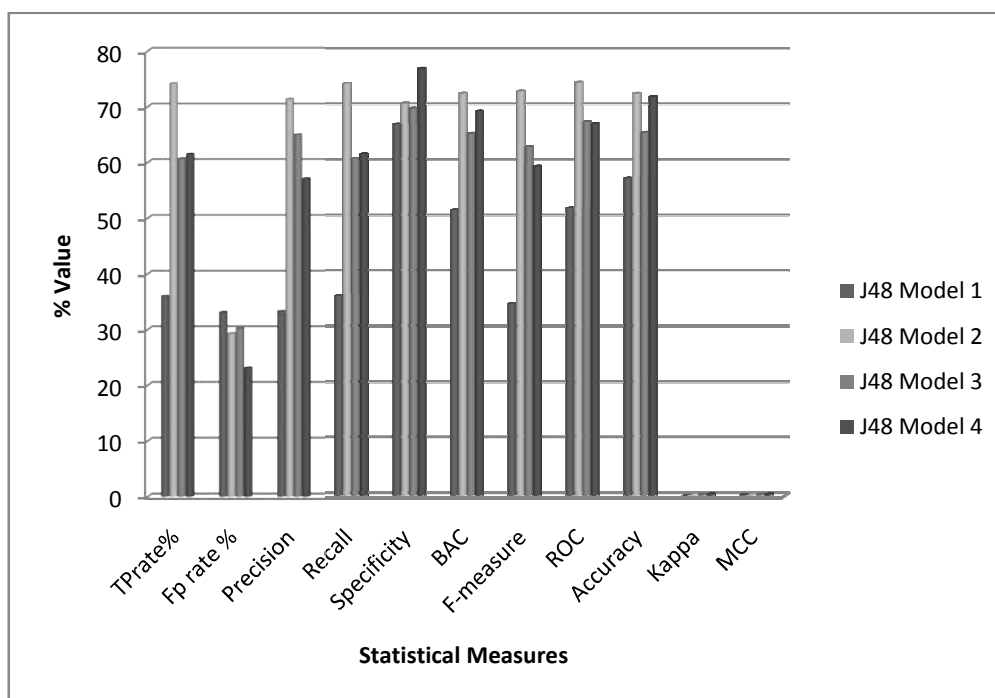


Figure 21 (b). Comparative graph between the various statistical measures of classification models-J48 Model 1, J48 Model 2, J48 Model 3 and J48 Model 4 are presented

Various statistical binary classification performance measures were used to evaluate the performance of the J48 anti-bacterial ML models. The model accuracy was found to be the highest around 72% for the TB oversampled (J48 Model 2) and least for the default TB J48 Model 1. Also the J48 ML Model 2 showcased better performance for the parameters like kappa, MCC, ROC, f-measure, BAC, recall and precision among all the ML models. The parameter specificity was higher for the Pseudomonas ML model in comparison to rest of the models. In comparison to TB default and SMOTE models TB oversampled J48 model outperformed. The FP rate of all the TB models was slightly higher, the values were in the range 30-33% with respect to the threshold FP rate. But for the Pseudomonas model the FP rate was better as it was close to the FP rate limit. The TP rate was higher for all the models except for the TB default model and TB oversampled model resulted with the highest. Here too the oversampled J48 ML model is better which shows the significance of the doubling of the data points in the active region in developing

better ML anti-bacterial models. From our study we understand that J48 Model 2 has performed better in respect to all the J48 anti-bacterial ML Models. Figure 21 displays comparative graph between the various evaluation measures against J48 ML models-J48 Model 1, J48 Model 2, J48 Model 3 and J48 Model 4. Later the screening set GSK 177 anti-TB molecules (as prepared in the prescribed procedure mentioned in the previous chapters) were screened against all the β -lactamase anti-bacterial J48 ML models. As a result we prioritized many GSK anti-TB molecules as computationally active β -lactamase inhibitors against the two targets computationally that is based on J48 algorithmic architecture. The numbers of the screened results are mentioned in Table 23.

Table 23. No. of computationally active β -lactamase inhibitors against J48 ML models

Models	No. of molecules virtually screened
J48 Model 1 (TB Default)	45
J48 Model 2 (TB Oversampled)	51
J48 Model 3 (TB SMOTE)	42
J48 Model 4 (Pseudomonas)	44

4.3 Conclusion

In this chapter, we discussed briefly in prioritizing computationally predicted β -lactamase inhibitors by two decision tree algorithms-Random forest and J48. The binary DT models were commonly used to discriminate compound bioactivities by using their chemical descriptors in the area of *in silico* drug design program. They are widely used in bioinformatics and cheminformatics due to its simple architecture. DT model is simple and produces readable and interpretable rules that provide insight into problematic domains. In the present work, DT models were developed against the target β -lactamase enzyme present in *M. tuberculosis* and *P. aeruginosa*. Here eight ML anti-bacterial models were developed based on the biological descriptor set generated from PowerMV software. RF and J48 DT models were developed, for the biased dataset sampling techniques (oversampling and SMOTE) were applied and carried out virtual screening against all the eight models.

The fitness of each model was studied in terms of various statistical parameters accuracy, kappa, ROC, sensitivity, specificity, precision, recall etc. The TB oversampled models (RF Model 2 and J48 Model 2) performed well against all the anti-bacterial DT models. Also FP rate was achieved low in the case of RF DT models and vice-versa for the J48 models. The developed DT models could screen many of the molecules from 177 GSK and it is reported as computational β -lactamase inhibitors. Our results suggested that the designed anti-bacterial DT models can be used as a virtual screening technique in prioritizing β -lactamase inhibitors as well as a complement to traditional approaches for hit selection.

CHAPTER 5

SUPPORT VECTOR MODELS AGAINST β -LACTAMASE ENZYME

5.1 Sequential Minimal Optimization (SMO)

Support Vector Machines (SVM) is a powerful classification and regression tool that is becoming increasingly applied in various ML models.¹⁰⁰ We used SMO for estimating the activity of β -lactamase enzyme inhibitors from the GSK library. In this study, SMO models were generated by using two HTS assays, PubChem AID 434987 and AID 2184, which were aimed at identifying β -lactamase inhibitors in *M. tuberculosis* and *P. aeruginosa*. As discussed in the previous chapter, SVM maps the data into a high-dimensional space, using a kernel function that is typically nonlinear. The SVM seeks to find an optimal separation between two classes, such that each in their entirety lies on opposite sides of a separating hyperplane. This is achieved by maximizing the margin between the closest points, known as support vectors, and the hyperplane. SVM can be adapted to either multiclass classification or to regression.¹⁰¹ SVMs are used in bioactivity prediction like drug repurposing,^{60,61} kinase inhibition,²⁵ estrogen receptor agonists²³ and opioid activity. Also the algorithm is used to predict toxicity-related properties like hERG blockade, mutagenic toxicity, toxicity classification and phospholipidosis. Applications in physicochemical property prediction include solubility, pKa,²⁹ logP and melting point.⁹⁴

5.2 Materials and Methods

The software's, online web servers and sampling techniques used for the model build are specified in Chapter 2. The bioassay datasets (AID 434987 and AID 2184) used in this experiment was downloaded from PubChem database and biological descriptors were generated from PowerMV software.

5.3 Experimental Procedures

For the model construct we implemented John Platt's sequential minimal optimization algorithm for training a support vector classifier from the ML package WEKA version 3.6. John Platt's SMO algorithm globally replaces both missing values and transforms nominal attributes into binary ones. Also by default all attributes are normalized in this algorithm for the ML SMO model build.

Before model development heap-size of WEKA was increased to 4 GB in order to handle out-of-memory exceptions for the AID bioassay datasets (434987 and 2184). The dataset was prepared in a systematic manner, by calculating 179 molecular descriptors from PowerMV, randomized, divided into two sets; 80% (training set) and 20% (test set), converted from CSV to ARFF, and finally loaded into the WEKA environment. The training set and test set were labeled with class nominal active and inactive respectively. The parameters set for the SMO model generation is given in Figure 22.

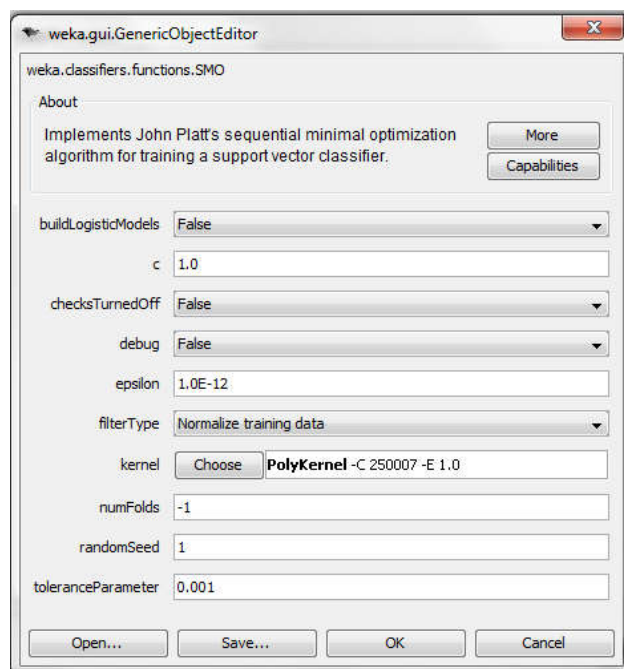


Figure 22. WEKA Generic Object Editor for SMO classifier

The training set was loaded into the WEKA panel and selected the WEKA-Classifiers-Functions-SMO algorithm for training the dataset, thereafter test set was re-evaluated upon the model through stratified 10 fold cross validation. The algorithm was set with the default parameters as shown in Figure 22 with build logistic model made “true”. The features included in the algorithm in the SMO model built is described in section Part I (Chapter 5).

SMO predictive models were developed by importing the training set and test set into the WEKA environment for the default β -lactamase SMO model generation SMO Model 1 for AID 434987 and SMO Model 4 for AID 2184. The dataset with the minority class in AID 434987 has only 372 active classes in comparison to 819 inactive classes. So performed the sampling techniques Oversampling and SMOTE for the minority class. Generated SMO Model 2 and SMO Model 3; the former corresponds to the oversampling of “active” class while the latter corresponds to the dataset with synthetic data points. For the AID 2184 default model (SMO Model 4) was generated without performing sampling technique. The data points of training and test set for SMO Model 1, SMO Model 2, SMO Model 3 and SMO Model 4 are given in Table 24.

Table 24. The number of data points for training set and test set in respect to SMO ML models

		Training set	Test set
1.	SMO Model 1 (TB Default)	953	238
2.	SMO Model 2 (TB Oversampled)	1251	312
3.	SMO Model 3 (TB SMOTE)	1251	312
4.	SMO Model 4 (Pseudomonas)	158	39

The training set for all the models were evaluated by a 10 fold cross validation. After these processes four models were generated corresponding to *M. tuberculosis* and *P. aeruginosa*. The models were built from the WEKA Generic Object Editor-weka.classifiers.functions.SMO by providing training set and test set as mentioned above. All the test set samples were re-evaluated upon the training set by ten by ten stratified cross validation.

Table 25. Confusion matrix generated for SMO ML models

SMO Model 1	SMO Model 2
<pre> a b <-- classified as 136 27 a = inactive 65 10 b = active </pre>	<pre> a b <-- classified as 111 46 a = inactive 59 96 b = active </pre>
SMO Model 3	SMO Model 4
<pre> a b <-- classified as 104 58 a = inactive 61 89 b = active </pre>	<pre> a b <-- classified as 18 8 a = inactive 4 9 b = active </pre>

The SMO ML models robustness was analyzed from the confusion matrix as shown in Table 25. The β -lactamase SMO anti-bacterial ML models were developed giving more emphasis on the percentage of FN than percentage of FP for compound selection as mentioned previously in the anti-bacterial Bayesian models. The false positive rate for the SMO ML models was set to an upper limit 20%. For the model generation only the parameter “build logistic models” was set “true” and rest of the parameter kept in the default state. The model fineness was checked from the number of TP, FP, FN, TN, TP rate, FP rate, accuracy, precision, recall, F-measure, ROC, kappa, MCC etc. All the statistical parameters were calculated upon the test set against all the SMO anti-bacterial ML models (SMO Model 1, SMO Model 2, SMO Model 3 and SMO Model 4) are given in Table 26.

Table 26. The number of TP, TN, FP, FN, TP rate, FP rate and the evaluation measures Precision, Recall, F-measure, ROC, Accuracy and Kappa generated by the SMO models

Statistical Parameters	SMO Model 1	SMO Model 2	SMO Model 3	SMO Model 4
TP	10	96	89	9
TN	136	111	104	18
FP	27	46	58	8
FN	65	59	61	4
TP rate %	13.3	61.9	59.3	69.2
Fp rate %	16.6	29.3	35.8	30.8
Precision	27	67.6	60.5	52.9
Recall	13.3	61.9	59.3	69.2
Specificity	83.435	70.7006	64.19753	69.2307
BAC	48.867	66.303	61.7487	69.2153
F-measure	17.9	64.6	59.9	60
ROC	51	71.5	69.7	69.2
Accuracy	61.34	66.3462	61.859	69.2308
Kappa	-0.4141	0.3265	0.2355	0.3571
MCC	-0.04142	0.3276	0.2356	0.3656

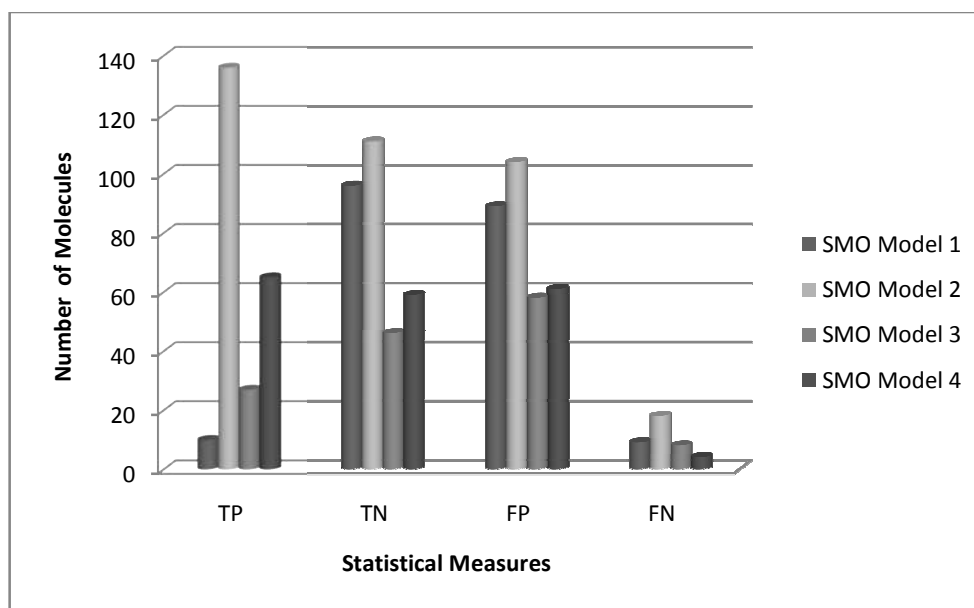


Figure 23 (a). Number of molecules based on confusion matrix

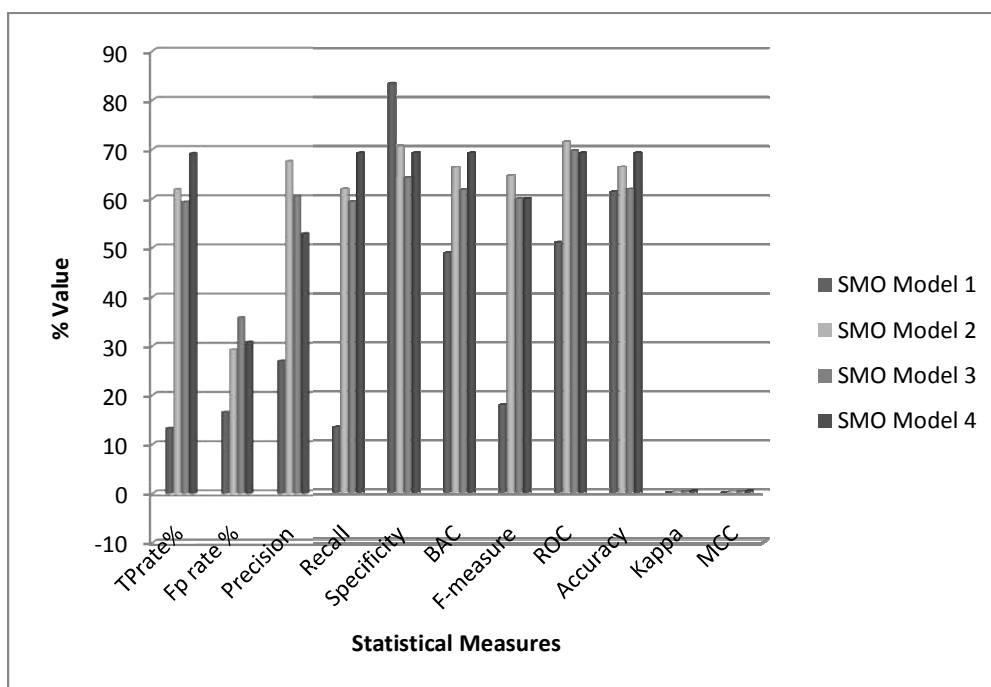


Figure 23 (b). Comparative graph between the various statistical measures of classification models-SMO Model 1, SMO Model 2, SMO Model 3 and SMO Model 4 are presented.

5.3.1 Robustness of the models

Various statistical binary classification performance measures were used to evaluate the performance of the β -lactamase SMO anti-bacterial ML models. The model accuracy was found to be the highest 69.23% for the Pseudomonas SMO model 4 and least for the default TB SMO Model 1. For the biased datasets the oversampled and SMOTE models were found to be very positive as it enhanced the model accuracy which is 66.3% for SMO Model 2 and 61.8% for SMO Model 3. The overall robustness of the SMO anti-bacterial models (SMO Model 1, SMO Model 2, SMO Model 3 and SMO Model 4) was judged from the statistical parameter accuracy. Since the dataset for TB was biased BAC was studied for each model as shown in Table 26. The BAC was highest for SMO Model 2 (66.3%) and least for SMO Model 1 (48.8%) among TB models. Among the TB ML models, SMO ML Model 2 showcased better performance for the parameters like kappa, MCC, ROC, f-measure, BAC, recall and precision among all the ML models. But the Pseudomonas model performed better against the parameters kappa, MCC, BAC

and recall. The FP rate was low and under the threshold value for the TB default Model 1 and high for TB SMOTE model. For the same, Pseudomonas model was 30%. Figure 23 displays comparative graph between the various evaluation measures of the SMO ML models-SMO Model 1, SMO Model 2, SMO Model 3 and SMO Model 4. Later the screening set GSK 177 anti-TB molecules (as prepared in the prescribed procedure mentioned in the previous chapters) were screened against all the β -lactamase anti-bacterial based SMO ML models. As a result we prioritized many GSK anti-TB molecules as computationally active β -lactamase inhibitors against the two targets that are based on SMO algorithm. The numbers of the screened molecules are mentioned in the Table 27. And the complete list of all the molecules that were predicted to be computationally active β -lactamase inhibitors against Bayesian, random forest, J48 and SMO anti-bacterial ML models are tabulated in Table 28.

Table 27. No. of computationally active β -lactamase inhibitors against SMO ML models

ML Models	No. of molecules virtually screened
SMO Model 1	9
SMO Model 2	56
SMO Model 3	28
SMO Model 4	48

S.No.	Molecules	NAÏVE BAYES				RANDOM FOREST				SMO				J48			
		TB D*	TB O**	TB S#	P##	TB D	TB O	TB S	P	TB D	TB O	TB S	P	TB D	TB O	TB S	P
20	GSK1829732A			Active				Active				Active					
21	GSK463114A		Active	Active				Active					Active	Active	Active		
22	GSK1783710A													Active		Active	
23	SB-204804-A		Active	Active	Active			Active				Active					
24	GSK1180781A			Active									Active				
25	GSK1220329A				Active							Active					Active
26	GSK1829819A		Active	Active	Active			Active		Active				Active	Active		
27	GSK1121877A		Active														
28	GW339742X													Active	Active	Active	
29	GSK1955236A		Active	Active	Active	Active		Active									
30	GSK810016A				Active				Active	Active		Active		Active		Active	
31	GW356807A	Active	Active	Active				Active						Active			
32	GSK731389A	Active	Active	Active			Active							Active			
33	GW876411A		Active	Active								Active				Active	
34	GSK921190A	Active	Active	Active				Active		Active	Active		Active	Active		Active	
35	GW859039X	Active	Active	Active									Active	Active	Active		
36	GSK1519001A		Active						Active	Active							
37	GSK2059310A							Active									
38	SB-435634	Active	Active	Active					Active	Active	Active	Active					Active
39	GSK1744926A																
40	GSK1750922A												Active				
41	GSK957094A	Active	Active	Active	Active		Active	Active	Active	Active				Active			
42	GSK1829733A			Active				Active				Active					

S.No.	Molecules	NAÏVE BAYES				RANDOM FOREST				SMO				J48			
		TB D*	TB O**	TB S#	P##	TB D	TB O	TB S	P	TB D	TB O	TB S	P	TB D	TB O	TB S	P
43	GSK695914A		Active	Active	Active				Active								
44	GRI35487X	Active	Active	Active													Active
45	GSK735816A		Active	Active				Active								Active	
46	BRL-7940SA		Active	Active							Active		Active				
47	GSK124576A		Active	Active													
48	BRL-8088SA		Active	Active	Active	Active	Active	Active				Active	Active	Active	Active		
49	BRL-10988SA					Active								Active			
50	CCI7967	Active	Active	Active										Active			
51	GSK254610A				Active				Active				Active				Active
52	GSK810037A																
53	GSK426032A													Active	Active		Active
54	GR223839X		Active	Active				Active								Active	
55	GSK735826A	Active	Active	Active				Active	Active		Active		Active	Active		Active	
56	GSK861337A															Active	Active
57	GSK829969A														Active		
58	GSK124945A			Active			Active							Active			
59	GSK1859936A																Active
60	GSK1742694A							Active									
61	GSK498315A		Active	Active	Active				Active		Active		Active	Active	Active	Active	
62	GSK1996236A				Active				Active		Active	Active	Active				Active
63	GW360240X				Active		Active						Active				
64	SB-650816				Active		Active		Active				Active	Active		Active	Active
65	GSK1829674A			Active					Active				Active	Active			

S.No.	Molecules	NAÏVE BAYES				RANDOM FOREST				SMO				J48			
		TB D*	TB O**	TB S#	P##	TB D	TB O	TB S	P	TB D	TB O	TB S	P	TB D	TB O	TB S	P
66	GSK762874A		Active	Active										Active			
67	BRL-10143SA		Active								Active		Active	Active			
68	GSK353071A	Active	Active	Active	Active			Active					Active			Active	Active
69	GW857165X		Active	Active							Active						
70	GSK1829671A			Active					Active				Active				
71	GSK847913A			Active					Active					Active	Active	Active	
72	GSK994258A																
73	GSK353496A	Active	Active	Active												Active	Active
74	GSK1829728A			Active					Active				Active	Active		Active	
75	GSK547481A																
76	GSK1365028A	Active	Active	Active							Active	Active		Active			Active
77	GSK920684A	Active	Active	Active							Active				Active		
78	BRL-51091AM		Active	Active									Active				
79	GSK2200157A		Active								Active				Active		
80	GSK2043267A			Active													
81	SB-516933													Active		Active	
82	GSK385518A	Active	Active	Active	Active	Active		Active	Active		Active	Active	Active			Active	
83	GSK798463A	Active	Active	Active							Active	Active	Active	Active		Active	
84	GSK1072678A	Active	Active	Active				Active			Active	Active		Active	Active		Active
85	BRL-8903SA		Active	Active							Active	Active	Active	Active			
86	GSK1829727A			Active					Active				Active	Active		Active	
87	GSK237561A		Active	Active					Active		Active						Active
88	GSK262906A	Active	Active					Active						Active	Active	Active	

S.No.	Molecules	NAÏVE BAYES				RANDOM FOREST				SMO				J48			
		TB D*	TB O**	TB S#	P##	TB D	TB O	TB S	P	TB D	TB O	TB S	P	TB D	TB O	TB S	P
89	GRI53167X		Active	Active							Active	Active		Active			Active
90	GSK2032710A								Active					Active			
91	GSK1985270A															Active	
92	GSK1589671A													Active			
93	GSK381407A		Active	Active	Active	Active			Active		Active	Active					
94	SB-712970														Active	Active	Active
95	GW713556X					Active								Active		Active	Active
96	GSK1826247A			Active												Active	
97	SB-811137-V								Active								
98	GSK754716A	Active	Active	Active	Active			Active			Active	Active	Active	Active		Active	
99	GSK705278A	Active	Active	Active				Active			Active	Active		Active			
100	GSK1941290A																
101	GSK1826825A			Active			Active		Active								Active
102	GSK1589673A													Active			
103	GSK847920A		Active	Active												Active	
104	GSK275628A	Active	Active	Active													
105	GSK636544A	Active	Active	Active													
106	GSK1434490A								Active						Active	Active	
107	GSK1731114A														Active	Active	Active
108	GSK345724A			Active	Active		Active		Active				Active				Active
109	BRL-51093AM		Active	Active							Active	Active	Active	Active			
110	GSK937733A			Active										Active			
111	GSK1402290A		Active	Active	Active	Active		Active	Active						Active		

S.No.	Molecules	NAÏVE BAYES				RANDOM FOREST				SMO				J48			
		TB D*	TB O**	TB S#	P##	TB D	TB O	TB S	P	TB D	TB O	TB S	P	TB D	TB O	TB S	P
112	GSK1588120A		Active	Active	Active								Active			Active	
113	GSK130506A		Active	Active							Active	Active					Active
114	GSK1925843A		Active	Active													Active
115	GSK991960A								Active								Active
116	GSK445886A	Active	Active	Active			Active				Active					Active	
117	GSK1302651A		Active	Active	Active			Active	Active				Active				
118	GSK937213A						Active								Active		
119	GSK270670A			Active									Active				
120	GI247341A																Active
121	GSK1829660A	Active	Active	Active	Active	Active			Active				Active	Active	Active		
122	GSK547543A																
123	GI103688B					Active	Active								Active		
124	GSK1729177A																
125	GSK1758774A																
126	GSK1650514A	Active	Active	Active							Active	Active			Active	Active	Active
127	GSK1812410A	Active	Active					Active								Active	
128	GSK831784A		Active						Active	Active	Active						
129	GSK1857145A		Active	Active							Active						
130	GSK848336A						Active	Active									
131	GSK1051703A	Active	Active	Active			Active			Active	Active		Active	Active			
132	GSK437009A	Active	Active	Active				Active		Active	Active	Active					
133	GSK276001A	Active	Active						Active								
134	GSK1826089A			Active												Active	

S.No.	Molecules	NAÏVE BAYES				RANDOM FOREST				SMO				J48			
		TB D*	TB O**	TB S#	P##	TB D	TB O	TB S	P	TB D	TB O	TB S	P	TB D	TB O	TB S	P
135	GSK479031A	Active	Active	Active	Active			Active	Active					Active			
136	GSK1055950A																
137	GSK1829816A	Active	Active	Active	Active							Active	Active	Active			
138	GSK1905227A	Active	Active	Active	Active	Active	Active					Active		Active	Active	Active	
139	SB-811796-V								Active								
140	GSK1863309A	Active	Active	Active										Active			Active
141	GSK316438A		Active	Active							Active			Active			Active
142	SB-706404	Active	Active				Active										
143	GSK547511A								Active								
144	GSK1691553A	Active	Active	Active							Active	Active					Active
145	GSK146660A				Active				Active		Active	Active					Active
146	GSK468214A						Active		Active						Active		
147	GSK1733953A	Active	Active	Active							Active						Active
148	GSK690382A	Active	Active	Active							Active						
149	SB-552112																
150	GSK2157753A	Active	Active	Active													
151	GSK1832831A	Active	Active					Active			Active						
152	GSK347301A		Active	Active							Active				Active		Active
153	GSK1385423A								Active				Active				
154	GSK1372568A	Active	Active	Active		Active	Active				Active	Active		Active	Active		Active
155	GSK1788487A		Active	Active							Active			Active			
156	GSK920703A	Active	Active	Active								Active					
157	GSK1107112A	Active	Active	Active		Active	Active	Active									Active

S.No.	Molecules	NAÏVE BAYES				RANDOM FOREST				SMO				J48			
		TB D*	TB O**	TB S [#]	P ^{##}	TB D	TB O	TB S	P	TB D	TB O	TB S	P	TB D	TB O	TB S	P
158	GSK1598164A		Active	Active					Active								
159	GSK921295A	Active	Active	Active				Active			Active	Active		Active	Active		Active
160	GW664700A		Active								Active	Active			Active		
161	GW369335X		Active	Active	Active				Active				Active				
162	GSK889423A					Active									Active		
163	GSK1668869A			Active							Active	Active					
164	GSK547487A																
165	GSK892651A		Active	Active			Active		Active						Active		
166	GSK275984A	Active	Active					Active			Active				Active		
167	GSK352635A																Active
168	SB-746177	Active	Active	Active	Active	Active		Active	Active				Active		Active		
169	GSK1518999A									Active	Active						
170	GSK1570606A	Active	Active	Active		Active	Active	Active			Active						
171	GV187303X	Active	Active	Active							Active		Active				
172	SB-354364	Active	Active								Active	Active			Active		
173	GSK1635139A	Active	Active	Active		Active				Active	Active	Active		Active			
174	GSK1310678A		Active				Active		Active				Active			Active	Active
175	GSK1611550A	Active	Active	Active				Active			Active				Active		
176	GSK1174628A				Active				Active				Active				Active
177	GSK133167A				Active				Active				Active		Active		

*Default, **Oversampled, #SMOTE, ##Pseudomonas

5.4 Conclusion

In this chapter, we discussed briefly the development of predictive models based on SMO. SMO algorithms are typically used in *in silico* drug design mainly to classify biologically active specific molecules from non-specific active molecules or to classify drugs from non drugs. The SMO models have wide applications in the field of bioinformatics and cheminformatics due to their good generalization and are less affected by the class imbalance ratio. In our study we developed SMO ML models against biased and non biased dataset against anti-bacterial activity. And we checked the algorithmic adaptation of the SMO models through sampling methods and virtual screening of anti-TB molecules from GSK library. Here we prioritized computationally predicted β -lactamase inhibitors from four SMO ML anti-bacterial models. From our study we understood that the SMO models do handle the imbalanced dataset and is a very good tool used for the computational virtual screening in the area of drug designing.

CHAPTER 6

DOCKING STUDY AGAINST β -LACTAMASE ENZYME PRESENT IN *M. TUBERCULOSIS* AND *P. AERUGINOSA*

6.1 Molecular Docking

Structure based virtual screening has become a crucial component of many drug discovery programs, from hit identification to lead optimization. One key methodology known as ‘docking program’ is used to place computer-generated representations of a small molecule to protein binding sites (via a user-define active site of an enzyme or from an online web servers like Computed Atlas of Surface Topography-CASTp.¹⁰²) in a variety of positions, conformations and orientations.¹⁰³ For performing a docking study, knowledge of the three dimensional structures or model of the target is a must.¹⁰⁴ Main goal of a docking study is to identify the energetically most favorable pose (also referred as ‘pose prediction’) of a ligand. The pose is evaluated by a scoring function which is based on the complementarily to the target in terms of shape and electrostatics. A good score for a ligand molecule indicates that it is potentially a good binder. The process is repeated and ranked subsequently by their scores for the molecules in the library that can be used later for any biological investigation for the compounds that are predicted to be active.^{105,106} In the past few decades molecular docking has been widely used in rational drug discovery as displayed in Figure 24. Here the evolution of the publications where molecular docking has been used extensively starting from 1948 to 2015 as obtained from PubMed database with keyword “docking”.²⁰

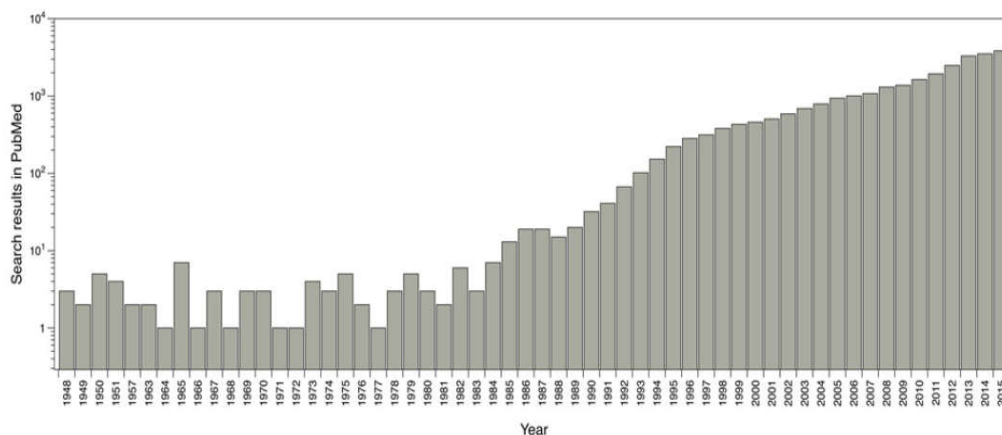


Figure 24. Reported publications number where ‘docking’ as search criteria in the PubMed database. Figure adapted from ref.²⁰

In molecular docking study energy scoring function is an important parameter that can rapidly and accurately describe the interaction between protein and ligand.¹⁰⁷ There are several reviews on scoring functions available in the literature. Some of the most important applications of scoring functions in molecular docking are it is used to rank ligand orientations/conformations by evaluating the binding tightness of each ligand protein complexes. An ideal scoring function would rank the experimentally determined ligand protein binding mode most highly. In spite of significant success, accuracy and prediction of protein–ligand interactions is still a challenge in structure based virtual screening.

Currently scoring functions are categorized into three following groups according to how they are derived: force-field-based (FF), empirical and knowledge-based functions.^{108,109} In our study we applied the empirical scoring function that describes hydrogen-bonding, ionic and polar interactions, as well as desolvation and entropic effects in the formation of the ligand-receptor complex. Here the binding affinity of a ligand complex is estimated based on a set of weighted energy terms as given by the equation (4).

$$\Delta G = \sum_i W_i \cdot \Delta G_i \quad (4)$$

Where ΔG_i is the various energy terms like van der Waals (VDW) energy, electrostatics, hydrogen bond, desolvation, entropy, hydrophobicity, etc. while the

coefficient W_i are ascertained by fitting the binding affinity data of a training set of protein–ligand complexes with known three-dimensional structures.

The other scoring function Force-field-based scoring functions estimate the binding energy from physical atomic interactions,⁵¹ including van der Waals (VDW) interactions, electrostatic interactions and bond stretching, angle bending and torsional forces. This type of scoring function is usually derived from both experimental data and ab initio quantum mechanical calculations according to equations of classical mechanics. Despite its lucid physical meaning, the major problem is solvent in the ligand binding which is not considered. In comparison to this scoring function, empirical scoring functions are much faster in binding score calculations due to their simple energy terms. The third one, knowledge-based scoring functions are statistical-potential based scoring functions employed on the energy potentials that are derived from the structural information embedded in experimentally determined atomic structures.¹¹⁰ In the current study we employed only the empirical based scoring functions in determining the binding affinity of protein-ligand complex.

Table 29. Examples of scoring functions implemented in widely used docking softwares

Force-Field-Based	Empirical	Knowledge-Based
DOCK, AutoDock, GoldScore, SYBYL G-Score, Molegro Virtual Docker, ICM, SYBYL D-Score, LigandFit	AutoDock, GlideScore, ChemScore, X Score, F Score, Fresno, SCORE, LUDI, SFCscore, HYDE, LigScore	SMoG, DrugScore, PMF Score, MotifScore, RF Score, PoseScore

Over the years, many small molecule docking methods have been developed and reviewed. They have been applied not only to protein-ligand scenarios but also in RNA-ligand docking, DNA-ligand docking and to the modulation of protein-protein interactions. They are commonly referred to as ligand-docking software. Several widely used docking implementations are summarized in Table 29. Although most of them require a commercial license, academic users and

researchers can experiment using freely available software like DOCK or AutoDock Vina. Each docking program has its own algorithms for the generation and scoring of the ligand poses, and a list of docking programs with VS experiments is mentioned in Table 30. While most of the software available performs poorly in generating near native poses for highly flexible ligands, Glide software overcomes this issue by running a more accurate search along the conformational space (random search followed by structure refinement), which results in an overall slow speed.^{21,111,112}

Table 30. A list of widely used docking programs with examples related to VS

Docking Programs	SBVS
AutoDock and AutoDock Vina	Cdc25 phosphatase inhibitors, Glutamate Transporter 1 (GLT 1), Cyclodextrin-based receptors, D-Ala:D-Ala ligase inhibitors.
DOCK	SARS-CoV 3C-like proteinase inhibitors, Hepatitis C virus helicase inhibitors, Cyclooxygenase (COX-2) inhibitors.
Glide	HIV-1 reverse transcriptase inhibitors and HIV-1: CD4-gp 120 binding inhibitors, Liver X receptor modulators.
GOLD	Non-peptide β -secretase inhibitors, Serotonin 5-HT(7)R antagonists, Sarco/endoplasmic reticulum calcium ATPase inhibitors.

6.2 Materials and Methods

The software's and tools used for the docking study are specified in Chapter 2. Glide from Schrödinger suite was used to study the ligand-protein interaction. The protein target β -lactamase was selected from protein data bank (PDB). And structure based virtual screening was performed with 177 anti-TB molecules from GSK library. The protein-ligand interactions were analyzed from Ligplot analysis and the dock pose was visualized from Pymol software.

6.3 Experimental Studies

The focus of the study was to identify the binding affinity of GSK molecules on the target β -lactamases of *M. tuberculosis* and *P. aeruginosa*. The idea was to

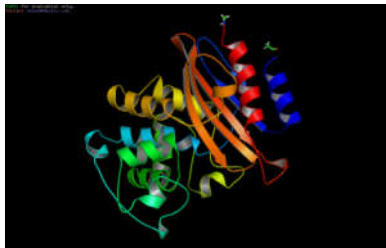
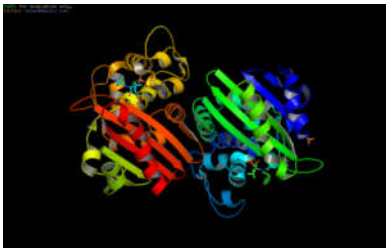
predict computationally active β -lactamase inhibitors from GSK library in the perspective of docking study of two bacterial species. Selected the protein β -lactamase (PDB id: 2GDN)¹¹³ for *M. tuberculosis* and (PDB id: 2WKH)¹¹⁴ for *P. aeruginosa* from PDB. The protein structure 2GDN was the structural characterization of *M. tuberculosis* that was understood based on the deletion of the *blaC* gene the only gene encoding a β -lactamase in the organism.¹¹³ For the same there are also other protein structures like 4QB8, 3CG5, 4QHC that are deposited in the PDB database. The protein structure 2WKH was selected in order to study the sensitivity of the computational docking based analysis of two strains of proteins having similar mechanistic action. Shishido H et al. emphasized that *P. aeruginosa* were the major pathogenic infection in chronic lung diseases in the post-tuberculosis patient. Another report from Ho-Kee Yum et al. discussed on the recurrent *P. aeruginosa* infection that are very common in lung diseases like bronchiectasis and chronic obstructive lung disease (COPD) due to TB. They also show drug resistance to the antibiotics. We have selected proteins from the X-ray crystallography structure because it remains the most powerful source of structural data among other methods like NMR spectroscopy and homology modeling. NMR spectroscopy produces several real conformations for the receptor while x-ray crystallography offers one single state of the crystallized protein. The active site of the two proteins was comparable for the reason that of sharing a common serine residue in the pocket of the two proteins. The key residues in the active site Ser 70 and Ser 67 for *M. tuberculosis* and *P. aeruginosa* plays an important role in the mechanistic action of β -lactamase activity against the currently available β -lactam antibiotics.^{113,114} The other active site residues for *M. tuberculosis* corresponds to Glu 166, Lys 73, Ser 130, Lys 234, Gly 132, Thr 235 and Thr 237 while Lys 70, Ser 115, Lys 205 and Gly 207 belongs to active site residues in *P. aeruginosa*.^{115,116,117,118,119}

6.3.1 Molecular Docking

DBVS was used to understand the binding affinity of anti-TB molecules against the serine β -lactamases of *M. tuberculosis* and *P. aeruginosa*. The selected proteins β -lactamase (PDB ID: 2GDN)¹¹⁸ for *M. tuberculosis* and (PDB ID:

2WKH)¹¹⁹ for *P. aeruginosa* was downloaded from the PDB in the .pdb format. The details of the targets β -lactamase present in *M. tuberculosis* and *P. aeruginosa* is summarized in Table 31.

Table 31. Details of the selected protein targets PDB 2GDN and 2WKH

PDB id	2GDN	2WKH
Organism	<i>M. tuberculosis</i> 	<i>P. aeruginosa</i> 
Experimental Method	X-ray diffraction	X-ray diffraction
Resolution	1.72 Å	1.79 Å
Classification	hydrolase	hydrolase
Chain	A	A, B
Amino Acid Sequence Length	267	248
Deposited Authors	Wang, F., Cassidy, C., Sacchettini, J.C., TB Structural Genomics Consortium (TBSGC)	Vercheval, L., Bauvois, C., Kerff, F., Sauvage, E., Guet, R., Charlier, P., Galleni, M.

The Schrodinger suite Grid based Ligand Docking with Energetics (GLIDE) was used for the docking study.¹²⁰ The crystal structure of the proteins 2GDN and 2WKH were imported into the protein preparation wizard panel and carried out the optimization procedure independently. The protein for *M. tuberculosis* was a monomer and homodimer (Chain A and B) in the case of *P. aeruginosa*. For the latter energy optimization was carried out for the chain A. The β -lactamase enzyme for *P. aeruginosa* was a co-crystallized protein with ampicillin substrate bounded in the active site. The protein was refined by hydrogen bond assignment and water molecules with less than 3A° hydrogen bonds to non-waters were removed. The grid was generated from the “receptor grid generation menu”. It is an automated process and the active site residues that were obtained from the literature^{50,115,116,117} were

given during the receptor grid generation. The active site residues Ser 70 and Ser 67 was given for *M. tuberculosis* and *P. aeruginosa* respectively. The screening set was prepared; all the 177 GSK molecules were geometry optimized from the MacroModel package. The docking was carried out independently against *M. tuberculosis* and *P. aeruginosa* in the standard precision (SP) mode. The binding affinity of the ligand-receptor complex was analyzed based on the docking score (glide score) which is the most important component in the structure-based drug discovery (SBDD). Glide uses an empirical scoring function in calculating the binding affinity of the complex which is faster as compared to the other force field scoring function. The β -lactamase inhibitor clavulanic acid (reacts with enzyme quickly to form hydrolytically stable and inactive enzyme in comparison with sulbactam and tazobactam) was set as a threshold docking score against the two targets. The glide score for the β -lactamase inhibitor clavulanic acid against the proteins 2GDN (*M. tuberculosis*) and 2WKH (*P. aeruginosa*) was -5.471 and -5.384. The molecules that passed the threshold score were treated as actives and the rest as inactive. The outcome, number of screened actives and inactive based on the docking score against the selected microbe is shown in Table 32. Except for the two GSK molecules GSK1121877A and GW356807A that failed to dock in the active site of β -lactamase enzyme in *M. tuberculosis* were considered to be inactive. The number of actives predicted by the DBVS method showed high target specificity against *M. tuberculosis* than *P. aeruginosa*. Table 33 summarizes the docking scores, various geometry optimization energy values and screened result (glide actives and glide inactive) for GSK 177 anti-TB molecules against the targets under study.

Table 32. Number of actives and inactives based on the docking scores

	<i>M. tuberculosis</i>	<i>P. aeruginosa</i>
Glide Actives	62	9
Glide Inactive	115	168

Table 33. Summary of the binding efficiencies of GSK 177 ligands onto β -lactamase enzyme present in *M. tuberculosis* and *P. aeruginosa* in reference to Clavulanic acid.

S. No.	GSK Molecules	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	Van der Waal Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005	TB Dock Score	Pseudomonas Dock Score	TB Glide	P Glide
1	BRL-10143SA	-182.022	7.09151	37.4288	-5.45662	3.84854	62.0109	-286.946	0.04636	-4.925459	-4.904949	x	x
2	BRL-10988SA	-161.318	7.63457	42.9544	15.3442	2.07546	64.6859	-294.012	0.044252	-4.710024	-5.640783	x	✓
3	BRL-51091AM	-170.378	6.1609	37.6982	5.46168	1.26255	53.2572	-274.218	0.041447	-5.744285	-4.820541	✓	x
4	BRL-51093AM	-191.292	6.42565	37.1027	-22.9031	3.99318	61.2907	-277.201	0.038584	-5.825566	-5.021055	✓	x
5	BRL-7940SA	-161.687	11.7644	47.035	-6.19724	3.96718	76.9089	-295.165	0.04112	-5.552213	-4.887263	✓	x
6	BRL-8088SA	-256.191	8.42653	28.3595	-15.578	3.8947	78.4422	-359.735	0.041232	-5.591093	-4.341675	✓	x
7	BRL-8903SA	-185.577	6.85393	37.4118	-7.60897	3.95231	64.0768	-290.263	0.03516	-4.83667	-4.832077	x	x
8	CCI7967	-487.357	3.77537	28.0276	-40.728	1.77567	79.5881	-559.796	0.046378	-5.562162	-4.662066	✓	x
9	Clavulanic acid	-65.4284	2.97103	61.8421	10.1185	5.79664	3.68911	-149.846	0.046257	-5.471	-5.384	✓	✓
10	GI103688B	-374.358	5.3179	51.3587	18.0159	0.544284	30.4763	-480.071	0.044781	-6.082088	-6.249893	✓	✓
11	GI247341A	-20.2591	5.99359	19.1687	34.9051	0.867764	50.7134	-131.908	0.04672	-4.021379	-4.471773	x	x
12	GR135486X	-194.353	2.94792	30.5843	0	0	53.5339	-281.419	0.000617	-5.078666	-4.339886	x	x
13	GR135487X	-171.462	2.51784	14.7274	16.6635	5.43037	56.5808	-267.381	0.036037	-5.343352	-5.023262	x	x
14	GR153167X	-55.8453	3.32483	8.91265	35.5703	2.12027	21.4927	-127.266	0.037852	-4.599758	-3.938989	x	x
15	GR223839X	162.231	8.91974	68.6961	10.7274	0.048712	32.0001	41.8385	0.046825	-5.847569	-3.762507	✓	x
16	GSK1051703A	-217.295	4.91056	24.0119	49.8316	1.64295	81.2436	-378.936	0.028475	-6.236361	-4.367053	✓	x
17	GSK1055950A	31.9489	7.4968	28.485	84.5583	3.34255	52.6417	-144.575	0.033299	-4.201657	-4.070941	x	x
18	GSK1072678A	-224.961	3.13598	14.1156	17.4802	0.406603	52.0519	-312.152	0.040944	-4.690078	-4.158446	x	x
19	GSK1107112A	10.4457	3.88297	51.4966	45.5454	0.7244	22.3608	-113.564	0.033079	-4.378685	-4.724175	x	x
20	GSK1121877A	-207.93	10.4163	30.7429	131.781	3.27558	138.292	-522.438	0.048818	-	-5.647076	x	✓
21	GSK1174628A	-117.577	9.62456	30.8108	21.1813	0.054644	48.331	-227.579	0.029202	-5.340502	-6.136228	x	✓
22	GSK1180781A	10.8261	3.98804	24.1783	52.2825	0.973429	49.3032	-119.899	0.042412	-4.577114	-3.80396	x	x
23	GSK1220329A	-105.028	6.16562	61.7862	3.20697	0.147564	10.4857	-186.82	0.044376	-5.197891	-4.540029	x	x

S. No.	GSK Molecules	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	Van der Waal Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005	TB Dock Score	Pseudomonas Dock Score	TB Glide	P Glide
24	GSK124576A	-182.269	4.67551	13.3866	1.62101	0.040629	69.6706	-271.664	0.038908	-4.822552	-4.148671	x	x
25	GSK124945A	185.341	3.7059	15.9063	1.77823	0.003108	37.3091	126.639	0.036286	-5.095692	-3.855036	x	x
26	GSK1302651A	58.3007	6.09865	16.4265	-3.44327	0.855404	70.6505	-32.2871	0.035663	-4.824759	-4.540315	x	x
27	GSK130506A	10.7537	3.98314	13.9726	13.7454	0.626606	47.8937	-69.4678	0.04937	-4.794501	-4.09419	x	x
28	GSK1310678A	180.831	5.24569	21.1996	25.7325	0.467586	29.3702	98.8158	0.04816	-7.263893	-5.126418	✓	x
29	GSK1329419A	-191.417	4.596	21.253	18.8256	0.020982	58.1547	-294.267	0.032399	-5.868955	-5.119634	✓	x
30	GSK133167A	103.219	10.1049	91.8853	134.897	0.954551	60.7918	-195.415	0.03514	-5.860265	-4.564013	✓	x
31	GSK1365028A	-25.5923	6.76062	25.7264	15.0445	0.624086	60.4705	-134.218	0.040503	-5.795231	-5.524003	✓	✓
32	GSK1372568A	-202.233	1.95211	15.3286	13.6305	0.000911	30.5233	-263.668	0.047812	-4.603222	-3.799022	x	x
33	GSK1385423A	92.6166	13.2299	34.0158	87.4923	1.83736	110.867	-154.826	0.041345	-5.091986	-3.948361	x	x
34	GSK1402290A	145.282	5.68304	45.6568	1.10031	0.007682	64.2154	28.6186	0.044327	-5.752629	-3.949569	✓	x
35	GSK1434490A	203.723	9.01372	64.4595	74.5273	0.27401	55.3546	0.093512	0.046434	-5.618112	-4.58761	✓	x
36	GSK146660A	105.091	6.6782	26.9975	-6.02521	0.118563	70.6835	6.63855	0.042778	-4.098286	-4.206953	x	x
37	GSK1518999A	-337.319	6.69738	31.8932	78.1503	1.41394	96.9423	-552.416	0.046807	-5.73743	-4.454925	✓	x
38	GSK1519001A	-359.398	6.59355	31.4917	75.2126	1.91377	97.0103	-571.62	0.04519	-4.948158	-4.636751	x	x
39	GSK153890A	-871.942	4.94662	34.8951	31.0511	0.371366	92.734	-1035.94	0.04259	-5.622324	-4.872096	✓	x
40	GSK1570606A	-175.399	3.14575	13.8399	16.7318	0.078225	48.4923	-257.687	0.038687	-5.366569	-4.483416	x	x
41	GSK1588120A	59.4805	3.85641	23.1102	93.7617	0.006079	63.0889	-124.343	0.04066	-4.311554	-5.43586	x	✓
42	GSK1589671A	-3.37862	3.27745	26.6399	55.6908	2.68584	27.3535	-119.026	0.046552	-5.512963	-4.685194	✓	x
43	GSK1589673A	10.9575	5.1734	33.0953	57.4248	1.36416	56.1323	-142.232	0.039686	-5.48464	-3.14357	✓	x
44	GSK1598164A	-202.427	4.73551	23.8369	56.7106	1.01688	74.1373	-362.865	0.031689	-4.904349	-4.202812	x	x
45	GSK1611550A	-216.7	4.40889	25.8257	5.52781	0.011563	56.9801	-309.454	0.032725	-5.482702	-4.057204	✓	x
46	GSK1635139A	39.6657	5.67671	38.5888	59.3114	0.237048	64.0944	-128.243	0.039784	-5.115107	-1.880659	x	x
47	GSK163574A	-610.704	5.95325	27.7168	35.5267	0.674005	87.3704	-767.946	0.048185	-4.045683	-4.176861	x	x
48	GSK1650514A	-23.4792	2.68842	18.7996	10.5764	0.052475	24.7053	-80.3014	0.045307	-4.216401	-4.150087	x	x
49	GSK1668869A	40.7	2.9164	8.24988	-2.42986	0.115862	26.2055	5.64229	0.034043	-4.524381	-3.916103	x	x
50	GSK1691553A	-200.927	3.77431	28.4471	47.2426	0.339256	45.5913	-326.322	0.047425	-5.228742	-4.130722	x	x

S. No.	GSK Molecules	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	Van der Waal Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005	TB Dock Score	Pseudomonas Dock Score	TB Glide	P Glide
51	GSK1729177A	31.9498	7.49687	28.4514	84.5405	3.34546	52.6714	-144.556	0.044411	-4.280701	-3.808163	x	x
52	GSK1731114A	307.247	11.7597	126.842	152.692	0.85347	35.0198	-19.9204	0.034286	-5.099303	-4.578677	x	x
53	GSK1733953A	108.973	7.61222	30.3576	41.975	0.149863	80.7333	-51.8552	0.04722	-4.743176	-4.51376	x	x
54	GSK1742694A	-86.652	2.03215	15.0904	5.22974	0.176065	19.5788	-128.759	0.041193	-5.296827	-4.148205	x	x
55	GSK1744926A	128.542	4.99215	27.7927	99.7582	0.347906	27.5152	-31.8637	0.043035	-5.750897	-4.079467	✓	x
56	GSK1750922A	52.0423	6.76833	27.8411	113.044	3.44239	72.5813	-171.635	0.041431	-4.049035	-3.689326	x	x
57	GSK1758774A	72.7394	5.65154	21.6839	119.421	1.05614	48.0432	-123.117	0.036928	-5.676211	-4.38173	✓	x
58	GSK1759150A	50.0223	3.9824	15.6102	12.5711	0.043429	47.2574	-29.4422	0.038956	-5.434897	-4.386236	x	x
59	GSK1783710A	27.303	8.12797	46.4532	70.5304	1.0923	53.5106	-152.411	0.032902	-4.648597	-3.663739	x	x
60	GSK1788487A	186.473	5.23981	20.449	62.6292	0.576983	17.8776	79.7004	0.039547	-4.521581	-4.529761	x	x
61	GSK1812410A	200.441	13.0806	23.4208	11.6556	0.099786	86.6539	65.5299	0.033172	-5.027874	-4.387707	x	x
62	GSK1826089A	-202.925	3.98577	209.417	-0.913377	0.344606	72.5625	-488.321	0.043645	-4.901519	-4.950106	x	x
63	GSK1826247A	-170.495	5.05987	31.4074	20.787	0.328406	68.1587	-296.236	0.047407	-5.262256	-4.092032	x	x
64	GSK1826825A	215.104	13.9569	16.2178	49.1365	0.138027	93.8342	41.8208	0.045137	-5.737437	-4.581582	✓	x
65	GSK1829660A	10.737	6.94213	22.2723	39.2091	1.16676	53.3386	-112.192	0.038795	-6.177868	-4.11545	✓	x
66	GSK1829671A	17.5196	7.44773	54.6419	37.8486	0.433833	35.9585	-118.811	0.037684	-4.675169	-4.352627	x	x
67	GSK1829674A	-46.7782	4.18426	8.81411	32.2914	0.472233	35.7217	-128.262	0.039212	-4.488235	-4.414894	x	x
68	GSK1829676A	-13.5252	5.61924	8.99257	34.2562	0.409479	62.4866	-125.289	0.046897	-4.856796	-4.463195	x	x
69	GSK1829727A	-73.5336	3.31542	10.4041	27.5757	0.497021	20.3498	-135.676	0.047105	-5.095832	-4.65595	x	x
70	GSK1829728A	-95.0475	3.29292	10.1716	27.7747	0.487073	16.2136	-152.987	0.04786	-5.011102	-4.307195	x	x
71	GSK1829729A	-48.7501	3.06424	12.6171	29.916	0.390608	14.6688	-109.407	0.040971	-6.212781	-4.821215	✓	x
72	GSK1829732A	2.22315	7.39785	54.8004	39.2623	0.441447	33.6291	-133.308	0.030589	-7.523095	-4.489654	✓	x
73	GSK1829733A	-58.8144	5.233	22.164	29.8952	0.447968	37.7445	-154.299	0.040775	-4.34916	-4.374612	x	x
74	GSK1829736A	-61.7978	4.19444	9.20312	32.3806	0.455483	34.0953	-142.127	0.046361	-6.236021	-4.381309	✓	x
75	GSK1829816A	-56.8496	3.25623	8.58264	31.9006	0.628243	19.9317	-121.149	0.038065	-5.02456	-5.217832	x	x
76	GSK1829819A	-84.5779	3.28836	8.2524	22.9791	0.556161	19.5568	-139.211	0.048649	-6.185953	-5.316149	✓	x
77	GSK1829820A	-56.969	3.09863	10.7573	25.7821	0.653836	17.2331	-114.494	0.048144	-5.352699	-5.143	x	x

S. No.	GSK Molecules	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	Van der Waal Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005	TB Dock Score	Pseudomonas Dock Score	TB Glide	P Glide
78	GSK1832831A	66.9012	9.09719	20.7868	58.5002	6.78373	100.428	-128.695	0.045217	-6.127531	-4.1216	✓	×
79	GSK1857145A	72.6705	6.3572	26.26	25.3939	0.967171	41.9685	-28.2762	0.045008	-4.167561	-4.108283	×	×
80	GSK1859936A	-369.851	9.5977	73.4642	20.122	0.141699	83.872	-557.049	0.042593	-4.71878	-4.184577	×	×
81	GSK1863309A	-19.4133	3.44094	16.9975	22.1807	0.479053	36.9178	-99.4293	0.048503	-5.087439	-3.487572	×	×
82	GSK1905227A	-24.5148	3.14805	10.915	51.0101	0.031137	39.958	-129.577	0.038332	-4.360943	-4.053539	×	×
83	GSK1925843A	-446.264	4.6594	26.8211	4.12069	0.201798	67.4842	-549.551	0.035721	-6.159658	-3.594705	✓	×
84	GSK1941290A	-59.7268	8.31848	19.2494	94.3949	0.298711	58.8159	-240.804	0.034526	-4.793047	-3.236873	×	×
85	GSK1955236A	-28.6743	4.54756	14.6787	4.90187	2.24202	69.479	-124.523	0.031768	-5.089971	-4.730871	×	×
86	GSK1985270A	113.235	7.2609	13.0701	46.2563	0.001766	63.685	-17.0393	0.048233	-3.41155	-4.362997	×	×
87	GSK1996236A	12.2389	6.41115	28.7152	-22.6789	0.046303	73.5342	-73.7891	0.047758	-4.038049	-4.674773	×	×
88	GSK2032710A	180.663	7.63849	28.5719	85.0523	0.224136	46.8121	12.3639	0.031993	-5.295409	-3.755368	×	×
89	GSK2043267A	96.1767	9.47139	46.3598	16.8261	0.058231	92.1729	-68.7116	0.048344	-5.922338	-4.800879	✓	×
90	GSK2059310A	-603.317	4.48205	50.3642	7.86045	8.94723	71.6876	-746.659	0.041978	-5.550714	-3.926131	✓	×
91	GSK2111534A	-38.5434	4.09025	6.77176	29.172	0.4232	36.1618	-115.162	0.044871	-5.818013	-5.102715	✓	×
92	GSK2157753A	-67.7002	1.77752	18.1958	9.71258	0.002021	28.0915	-125.48	0.044766	-4.900896	-4.082925	×	×
93	GSK2200150A	171.257	8.98644	30.4493	73.2636	0.045373	61.3089	-2.79723	0.049977	-4.363393	-3.942104	×	×
94	GSK2200157A	75.4518	4.7127	22.7489	44.9196	0.039439	12.3092	-9.27814	0.032081	-3.878424	-4.217975	×	×
95	GSK2200160A	72.2344	4.64268	25.745	33.1523	0.039555	12.1801	-3.52516	0.034686	-5.354315	-4.396907	×	×
96	GSK237561A	-496.707	4.76935	33.0671	25.6269	0.517939	63.393	-624.082	0.04917	-6.086947	-4.06445	✓	×
97	GSK254610A	26.2805	6.88927	17.5903	38.1997	0.01253	77.171	-113.582	0.049157	-5.134372	-4.859504	×	×
98	GSK262906A	23.179	10.5354	25.5066	-2.75804	0.320534	66.863	-77.2886	0.048782	-4.736116	-3.096196	×	×
99	GSK270670A	-138.958	3.96848	39.016	17.2918	2.99033	29.7935	-232.019	0.03303	-4.936422	-4.437894	×	×
100	GSK275628A	-417.126	5.73646	26.9176	14.2622	1.83321	82.3651	-548.241	0.047663	-4.580708	-4.334243	×	×
101	GSK275984A	-582.593	6.46441	14.011	71.784	4.38677	113.799	-793.038	0.01821	-3.803786	-3.874174	×	×
102	GSK276001A	-578.073	6.17544	14.2937	72.673	4.36577	112.628	-788.21	0.024735	-3.91484	-4.37488	×	×
103	GSK316438A	-71.0676	1.06936	19.1027	3.00411	0	0.39898	-94.6427	0.002942	-4.036971	-3.991625	×	×
104	GSK345724A	69.5158	8.15305	24.313	1.50151	1.20637	64.5997	-30.2578	0.049937	-5.190288	-4.572399	×	×

S. No.	GSK Molecules	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	Van der Waal Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005	TB Dock Score	Pseudomonas Dock Score	TB Glide	P Glide
105	GSK347301A	33.4626	3.70695	18.9645	20.0881	0.362086	48.1099	-57.7689	0.049842	-3.5351	-4.00121	x	x
106	GSK352635A	76.3214	2.58602	7.24298	5.35964	0	20.2049	40.9279	0.000649	-6.428837	-4.549626	✓	x
107	GSK353069A	-426.57	2.20559	20.1079	-9.20227	0.000754	59.9441	-499.626	0.044917	-5.894691	-4.897256	✓	x
108	GSK353071A	-443.289	2.21626	16.279	-8.09356	2.05466	58.9788	-514.724	0.042904	-5.081144	-3.963855	x	x
109	GSK353496A	-466.613	0.832227	30.4028	-9.20264	0.001001	37.5213	-526.167	0.026943	-5.123067	-4.586514	x	x
110	GSK358607A	-152.404	6.68457	43.5534	8.38196	0.001263	87.6451	-298.67	0.039921	-5.487358	-4.304648	✓	x
111	GSK381407A	-319.45	11.2913	33.8162	12.5436	0	126.999	-504.1	0.027348	-4.497637	-5.0295	x	x
112	GSK385518A	-19.3478	6.61063	59.946	42.2813	0.001552	51.568	-179.755	0.049007	-5.607946	-5.884787	✓	✓
113	GSK426032A	54.5108	5.23786	36.291	27.6396	0.369312	61.8708	-76.8978	0.048142	-4.45128	-4.294382	x	x
114	GSK437009A	-650.924	7.09087	25.8762	30.7528	0.053757	99.175	-813.872	0.035337	-4.263662	-4.548821	x	x
115	GSK445886A	-334.812	2.25108	30.9716	4.52709	0	52.3228	-424.884	0.008386	-4.744651	-4.379147	x	x
116	GSK463114A	-185.664	5.09648	22.1563	4.44775	0.011149	72.7177	-290.093	0.039971	-5.66555	-4.185975	✓	x
117	GSK468214A	-227.547	1.91598	29.4308	17.1979	0.840829	22.4275	-299.36	0.048998	-5.309476	-4.1851	x	x
118	GSK479031A	-74.0444	2.96495	12.2096	15.6638	0.051081	44.5313	-149.465	0.046341	-6.133645	-5.138022	✓	x
119	GSK498315A	6.15853	7.36419	52.7504	33.5751	0.42631	35.4114	-123.369	0.049824	-6.915556	-5.027683	✓	x
120	GSK547481A	-389.46	7.18275	45.6188	121.397	4.74788	73.4333	-641.839	0.036505	-5.166088	-4.787793	x	x
121	GSK547487A	-647.659	6.27769	36.513	74.6083	5.80508	83.7606	-854.624	0.047681	-4.331938	-4.92963	x	x
122	GSK547511A	-419.002	8.90326	42.3127	118.727	5.90562	90.0448	-684.895	0.048589	-5.527249	-3.718299	✓	x
123	GSK547543A	-522.996	10.3194	46.3052	115.541	5.06032	106.355	-806.577	0.049938	-6.459044	-5.196833	✓	x
124	GSK636544A	88.4174	3.41407	43.3593	55.5501	0.466593	19.1007	-33.4734	0.041286	-4.138962	-4.103778	x	x
125	GSK690382A	-259.34	2.75402	30.7002	23.0357	0.539425	47.3849	-363.754	0.032378	-5.355261	-4.176003	x	x
126	GSK695914A	74.0455	9.91832	22.5367	23.7171	0.591968	100.799	-83.5176	0.0436	-5.703629	-4.722775	✓	x
127	GSK705278A	136.167	6.17983	50.2997	14.3784	0.158958	53.2495	11.9005	0.043501	-5.460412	-5.270821	x	x
128	GSK731389A	149.578	6.06605	42.2458	84.1391	0.018322	29.7327	-12.6239	0.032341	-5.393793	-4.248966	x	x
129	GSK735816A	-263.126	2.69356	29.7368	12.5436	0	44.9435	-353.043	0.008909	-5.660585	-4.433318	✓	x
130	GSK735826A	-231.183	5.34523	68.0381	27.1704	0.534171	39.068	-371.339	0.028996	-5.742838	-4.502916	✓	x
131	GSK749336A	-127.985	6.13617	33.1036	9.09699	0.02453	83.148	-259.494	0.042031	-5.669826	-4.385861	✓	x

S. No.	GSK Molecules	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	Van der Waal Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005	TB Dock Score	Pseudomonas Dock Score	TB Glide	P Glide
132	GSK754716A	18.4816	6.38264	61.2356	11.1438	0.029604	41.743	-102.053	0.025292	-4.54943	-3.820543	x	x
133	GSK762874A	-13.4381	4.99888	15.1916	24.2938	0.735436	44.9615	-103.619	0.047678	-5.720321	-4.422411	✓	x
134	GSK798463A	42.6077	6.81002	58.7838	33.9854	0.540751	41.9813	-99.4936	0.04581	-5.907448	-4.749784	✓	x
135	GSK810016A	353.678	17.1683	163.553	54.6202	2.06757	19.8738	96.3953	0.031642	-5.442965	-4.220134	x	x
136	GSK810037A	325.403	14.6451	165.637	24.4704	0.0067	-16.4386	137.083	0.03985	-5.856571	-4.95457	✓	x
137	GSK829969A	-70.793	3.33571	23.0506	3.53363	0.078459	22.779	-123.57	0.033533	-4.509008	-4.292939	x	x
138	GSK831784A	-515.262	4.95959	24.0841	-10.9542	0.159844	94.5719	-628.083	0.038246	-5.07314	-4.126366	x	x
139	GSK847913A	-484.416	4.99239	10.2706	13.4158	0.007637	80.6874	-593.79	0.048777	-3.589416	-3.987311	x	x
140	GSK847920A	25.2308	3.90388	6.96087	19.7392	0.138118	33.1019	-38.6131	0.038001	-5.48285	-4.012116	✓	x
141	GSK848336A	-562.755	4.48841	23.4226	47.4911	0.003287	67.6952	-705.856	0.040942	-4.451997	-3.855008	x	x
142	GSK861337A	255.387	10.2091	63.3804	125.387	0.677695	57.3899	-1.65617	0.034137	-5.158176	-4.789905	x	x
143	GSK888636A	-21.4633	5.35337	23.3075	4.48008	2.09875	77.2223	-133.925	0.041142	-4.98263	-3.982144	x	x
144	GSK889423A	38.0992	5.10084	19.2258	12.1211	0.028841	39.0951	-37.4725	0.048309	-4.78812	-4.405249	x	x
145	GSK892651A	-42.9018	3.05381	10.4185	25.9848	0.674735	17.7694	-100.803	0.013964	-5.940762	-5.307386	✓	x
146	GSK920684A	-207.81	3.69633	20.7427	4.52911	0.000074	57.7328	-294.511	0.046952	-5.685314	-4.195351	✓	x
147	GSK920703A	-213.85	2.81221	14.4876	15.6259	0.069625	47.7274	-294.572	0.031353	-5.548655	-4.108738	✓	x
148	GSK921190A	-168.037	4.71046	17.5675	25.4969	0.969921	56.0701	-272.852	0.037555	-5.400508	-3.606963	x	x
149	GSK921295A	-34.078	4.01765	13.2676	14.0339	0.464272	47.8889	-113.75	0.049845	-5.536	-4.329229	✓	x
150	GSK937213A	98.2221	4.30592	22.0474	43.5455	0.182207	28.9541	-0.81304	0.030265	-5.053564	-3.35988	x	x
151	GSK937733A	54.2155	7.77845	70.5486	60.3678	1.3564	49.1	-134.936	0.044206	-5.889491	-4.765693	✓	x
152	GSK957094A	-33.9572	4.05878	6.95686	29.3152	0.434666	36.9178	-111.641	0.033495	-6.06936	-5.287924	✓	x
153	GSK991960A	-73.9297	2.61722	19.3121	1.07966	4.22242	18.2001	-119.361	0.034876	-4.077586	-3.450489	x	x
154	GSK994258A	98.8408	6.59274	34.2651	36.2152	0.052163	45.2884	-23.5727	0.049772	-4.397974	-4.151907	x	x
155	GV187303X	-18.0373	4.49943	17.9585	28.0935	4.97186	73.3783	-146.939	0.03963	-5.416388	-4.37542	x	x
156	GW339742X	-374.456	6.58984	25.6893	13.1339	1.41935	74.49	-495.778	0.047022	-6.116786	-5.019073	✓	x
157	GW356807A	-16.5881	7.54477	95.7136	74.161	5.7716	79.8264	-279.605	0.039188	-	-5.134416	x	x
158	GW360240X	-683.064	7.75714	34.1323	25.6609	6.50935	93.9022	-851.026	0.045176	-5.401248	-4.9346	x	x

S. No.	GSK Molecules	Potential Energy-OPLS-2005	Stretch Energy-OPLS-2005	Bend Energy-OPLS-2005	Torsional Energy-OPLS-2005	Improper Torsional Energy-OPLS-2005	Van der Waal Energy-OPLS-2005	Electrostatic Energy-OPLS-2005	RMS Derivative-OPLS-2005	TB Dock Score	Pseudomonas Dock Score	TB Glide	P Glide
159	GW369335X	-304.731	17.2366	113.1	0	0	106.814	-541.881	0.008824	-6.700112	-5.424153	✓	✓
160	GW623128X	3890.56	1433.02	1569.57	125.622	0.47213	894.591	-132.71	0.043011	-4.091625	-3.969966	×	×
161	GW664700A	79.1277	4.13309	28.3894	78.9805	0.371852	32.458	-65.2052	0.047798	-5.147898	-4.457291	×	×
162	GW713556X	164.47	9.02295	64.4633	104.752	0.358321	49.0909	-63.2175	0.043851	-3.74818	-4.749814	×	×
163	GW857165X	-128.253	5.43852	24.0206	8.93293	0.017772	74.9006	-241.563	0.036192	-4.842953	-4.580092	×	×
164	GW859039X	-27.9347	4.99053	16.8616	20.2826	0.008577	36.8392	-106.917	0.049086	-5.002763	-4.030854	×	×
165	GW861072X	192.65	9.14733	66.568	105.319	0.360875	49.433	-38.1777	0.048715	-4.036657	-3.829027	×	×
166	GW876411A	-615.2	5.68945	26.9236	36.4315	0.787782	87.8366	-772.869	0.04637	-4.095902	-4.303594	×	×
167	SB-204804-A	113.63	3.75838	20.9912	39.7789	0.177425	24.9771	23.9468	0.044686	-5.032691	-4.410498	×	×
168	SB-354364	279.216	9.86875	146.829	87.0636	0.003839	52.6923	-17.242	0.042979	-4.492187	-4.215178	×	×
169	SB-435634	-19.8289	5.77867	55.2263	24.7815	5.00992	46.4654	-157.091	0.042931	-5.886289	-4.782097	✓	×
170	SB-516933	208.494	10.5819	64.9888	105.928	0.363908	72.092	-45.46	0.0471	-3.849518	-4.726249	×	×
171	SB-552112	205.147	9.19876	25.1914	112.654	0.878422	98.1166	-40.8924	0.041889	-5.688828	-4.59511	✓	×
172	SB-650816	294.482	10.7407	29.5337	83.9042	3.25132	94.3501	72.7015	0.038372	-6.236126	-4.773735	✓	×
173	SB-706404	-71.5536	4.47074	15.7594	1.15775	0.392778	28.9608	-122.295	0.027937	-5.165198	-5.378961	×	×
174	SB-712970	281.712	9.90606	76.6494	136.289	0.694666	45.133	13.0401	0.041508	-4.067237	-5.446238	×	✓
175	SB-746177	-286.077	4.50018	11.8643	54.2228	0.624094	69.8824	-427.17	0.043952	-5.506079	-4.31357	✓	×
176	SB-811137-V	143.003	6.61081	16.1338	100.321	3.92422	45.582	-29.569	0.036726	-5.305719	-4.258882	×	×
177	SB-811796-V	209.052	6.01252	15.5844	138.852	0.428166	44.331	3.84322	0.041776	-4.812373	-4.422011	×	×
178	SB-829405	134.469	5.44469	30.3724	107.664	0.018864	50.0345	-59.0651	0.045824	-3.141303	-4.582211	×	×

Note: Active molecule is represented by the symbol “✓” and inactive by “×”. Two GSK molecules GSK1121877A and GW356807A showed no binding with β -lactamase enzyme present in *M. tuberculosis* and were considered as inactive for the current study.

6.3.2 Results and discussion

As shown in Table 33 maximum docking scores were obtained for ligands against PDB 2GDN than PDB 2WKH. From this it is evident that the binding affinity is more towards the TB β -lactamase enzyme than Pseudomonas. After structure based virtual screening, the best four compounds with respective binding affinities to the active site of β -lactamase enzyme present in *M. tuberculosis* and *P. aeruginosa* were identified from the ligand screening set GSK 177 anti-TB molecules. The two best compounds identified as TB β -lactamase inhibitors, each with the docking score -7.5230 and -7.2638, respectively. Similarly the other two best compounds each with docking score -6.2498 and -6.1362 were identified as computational active Pseudomonas β -lactamase inhibitors. The docking results of TB β -lactamase enzyme ended with a total of 62 actives in which 17 GSK molecule scored above -6 and 43 molecules scored above -5.47. For the same in Pseudomonas β -lactamase enzyme, seven molecules scored above -5.38. An in-depth analysis of the protein ligand interaction was studied from the Schrodinger suite. The top four molecules schematic presentations of binding modes of the target protein structure with the compounds, i.e. the substrates (β -lactamase from TB and Pseudomonas) are shown in Figure 25 (a-d).

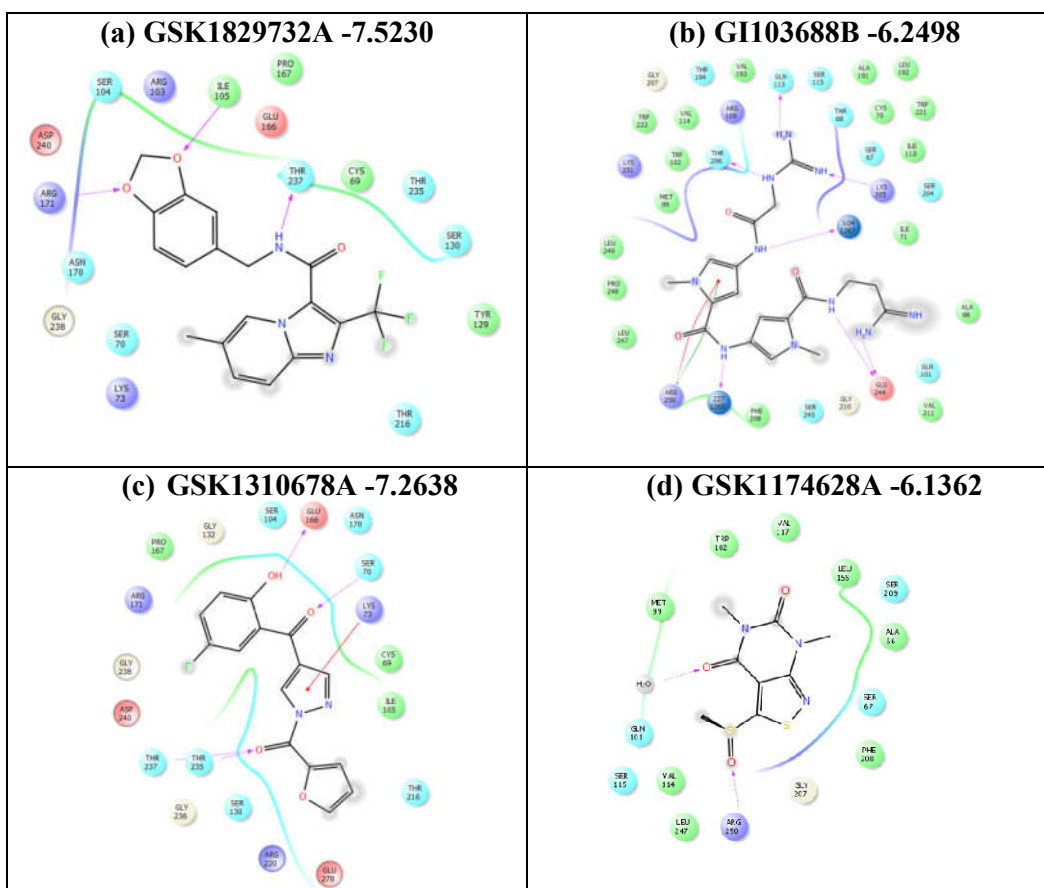


Figure 25 (a-d) β -lactamase actives 2D interaction diagrams with highest docking score against *M. tuberculosis* and *P. aeruginosa* in reference to the threshold docking score of Clavulanic acid against the targets under study. Highlighted colors give information on the interaction types and nature of residues.

Molecular docking of GSK1829732A showed that four hydrogen bonds were formed within the active site of target protein. The polar residues Ser 70, Asn 170, Ser 104, Thr 237 and Thr 216 constituted the binding scheme with the ligand molecule. Out of these, Thr 237 has a hydrogen bond with the ligand. And all the other residues like negative charged residues (Asp 240 and Glu 166), positive

charged residues (Lys 73, Arg 171, Arg 103) and the hydrophobic residues (Ile 105, Cys 69, Tyr 129) are in the vicinity of the active site region in TB β -lactamase enzyme. The interaction of the molecule GSK1310678A displayed four hydrogen bonds with the side chain residues Ser 70, Glu 166, Thr 235 and Thr 237. Here a π -cation interaction is also displayed with Lys 73 residue. The polar residues, positive charged, negative charged and hydrophobic are present in the active site region in TB β -lactamase enzyme. The docking of GI103688B has five hydrogen bonds formed within the active site of Pseudomonas β -lactamase enzyme. The polar residues Ser 67, Ser 115, Ser 245, Gln 101 and Ser 204 constitute the binding scheme with the ligand molecule. Also, Thr 206, Gln 113, Lys 205, Glu 244 has a hydrogen bond with the ligand. And all the other residues like positive charged residues (Arg 109, Lys 205 and Lys 251), negative charged residues (Glu 244) and many hydrophobic residues are in the vicinity of the active site region. Here a π -cation and π - π stacking interaction is also displayed with Arg 250 residue. GSK1174628A makes two hydrogen bond interaction with Arg 250 and a water molecule. The polar residues (Ser 67, Ser 209, Gln 101, Ser 115), positive charged residue Arg 250, hydrophobic residues (Met 99, Trp 102, Val 117, Leu 155, Ala 66, Phe 208, Leu 247 and Val 114) with no negative charged residues available are displayed. The structure of the β -lactamase enzyme complexed with the GSK molecules (Figure 25) revealed that hydrogen bonds and hydrophobic interaction is the main force in the active site. Consequently, our candidate compounds screened from SBVS method have the potential to be considered as new computationally active β -lactamase inhibitors. A structure based approach and calculation of binding affinity was used to investigate the ligand-protein interaction between the compounds and the protein in detail. In addition to this the experimental procedures involving whole-cell screening assays fulfill one of the major criteria like permeability issue which is a serious problem in MTB owing to thick mycobacterial cell wall. Thus, it would be interesting to investigate the prioritized molecules as potential candidates against anti-bacterial drug discovery.

6.4 Conclusion

Molecular docking was carried out against the target β -lactamase enzyme present in *M. tuberculosis* and *P. aeruginosa* and binding affinity of 177 GSK anti-TB molecules were studied. As the screening set under the study are experimentally active against *M. tuberculosis* is a whole cell screen. The selection of the targets were from two kinds of bacterial species for the reason of sharing common pocket residue Ser 70 (*M. tuberculosis*) and Ser 67 (*P. aeruginosa*) since they play an important role in the mechanistic action of β -lactamase activity against the currently available β -lactam antibiotics. From our study we predicted that 62 anti-TB molecules were computationally active against the target selected under study and for the same nine anti-TB molecules showed a high affinity to the enzyme in *P. aeruginosa* which is a gram negative bacterium. The activity profile was not only based on the glide docking score but we put a threshold condition that the activity is determined based on the docking score of the β -lactamase inhibitor clavulanic acid which was -5.471 for *M. tuberculosis* and -5.384 for *P. aeruginosa*. Our finding proves that some GSK anti-TB molecules are computational β -lactamase inhibitors and are active against both the gram positive and gram negative bacteria. As shown in our study, molecular docking has been able to identify promising compounds that might represent future solutions in designing new combination of treatment for drugs in TB.

CHAPTER 7

ARTIFICIAL NEURAL NETWORK BASED SELF ORGANIZING MAPS

7.1 Introduction to Artificial Neural Networks

An artificial neural network (ANN), often simply called a neural network is a mathematical model used for pattern recognition and machine learning. As discussed earlier the network's structural design consist of connected neurons with input layer, a hidden layer or layers, and an output layer. Here each connection between neurons carries a weight that are varied during the training phase as the network learns to connect input and output data, before being tested on unknown dataset. While ANN is inspired by the design and functionalities of a biological neuron that is a basic building block of biological neural networks that includes the brain, spinal cord and peripheral ganglia. The similarities in design and functionalities of biological neuron and an artificial neuron can be seen in Figure 26.

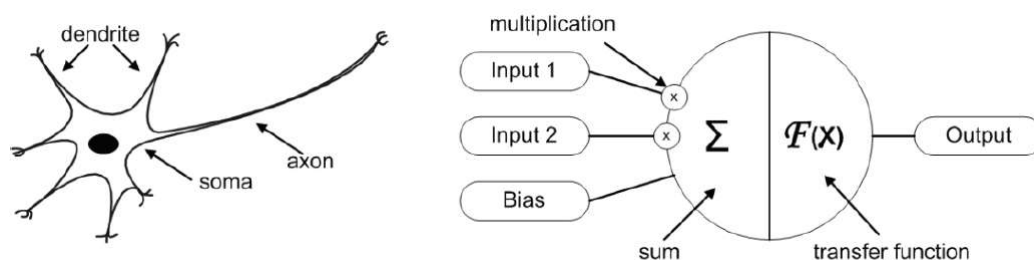


Figure 26. A similarity between a biological neuron and an artificial neuron is displayed. On the left side it is a biological neuron with its soma, dendrites and axon and while on the right side an artificial neuron with its inputs, weights, transfer function, bias and outputs are shown. Model depicted in the figure is taken from ref.⁷⁶

ANN has wide applications in QSAR and drug design where it is used to correlate physicochemical parameters of compounds with chemical or biological activities and predicts the activities of new compounds. ANN-based QSAR models are also used as prediction methods in the virtual screening process. ANN in drug

discovery as reported by Haider on a predictive PK/PD model for an oral hypoglycemic agent (repaglinide) using ANN are few of its application.^{73,121}

In our study we used the Self-Organizing Maps (SOM), which is called Kohonen networks, have been used with a great deal of success in numerous applications.¹²² It was developed by Tuevo Kohonen, a professor emeritus of the Academy of Finland in the year 1982. And it was aptly named because no supervision is required.¹²³ Kohonen ANN is a ‘self-organizing’ system that is capable to solve the unsupervised problems like clustering than the supervised problems and can be used for exploratory tasks. The algorithm automatically adapts itself in a manner that the similar input data are associated with the topological close neurons in the ANN. This means that the neurons that are physically located close to each other will react similarly to similar inputs and vice-versa for the neurons that are far apart in the layout of the ANN structure.¹²⁴ This often assumes a colored visualization of the SOM where feature vectors, from an input space usually a high dimensional data is projected into an output space with a low dimensional 2D map. The main aim of SOM is to find the hidden patterns in high dimensional data. In this chapter we try to identify β -lactamase inhibitors of computationally active molecules by projecting the high dimensional data where the molecules were screened and visualized in the low dimensional 2D map.

7.2 Materials and Methods

The software’s and tools used for SOM analysis are specified in Chapter 2. SOM analysis was performed from cheminformatics software Canvas (Schrodinger suite). The bioassays AID 434987, AID 2184 and 179 biological descriptors calculated from PowerMV was selected for the SOM analysis. ANN based virtual screening was performed against the screening set from GSK library consisting of 177 anti-TB molecules.

7.3 Experimental Studies

ANN method Kohonen Self-Organizing Map was used to identify the β -lactamase activity on the basis of 179 biological molecular descriptors in the

multidimensional 2D map. Kohonen's Self-Organizing Map is one of the most popular neural network models among the existing neural networks algorithm.^{125,126} A powerful suite of the cheminformatics, comprehensive cheminformatics computing environment (Canvas) was used in the virtual screening against the targets β -lactamase in *M. tuberculosis* and *P. aeruginosa*. The screening sets for *M. tuberculosis* and *P. aeruginosa* include the 179 descriptors calculated from PowerMV software. The *M. tuberculosis* screening set 1 contained 1368 molecules from the PubChem bioassay AID 434987 including actives, inactive and 177 molecules from the GSK library. Similarly, *P. aeruginosa* screening set 2 contained 374 molecules in total of which 197 molecules are from AID 2184 and the rest from 177 molecules from GSK library. The inconclusive molecules were excluded from the dataset preparation.

Kohonen self-organizing maps were built from the Applications menu in the canvas software. First we selected "Build New Map", and then chose the 179 biological descriptors calculated from PowerMV software. In the Lattice options section, the dimensions of the lattice were set to 15x15 for the screening set 1 (*M. tuberculosis*) and 8x8 for screening set 2 (*P. aeruginosa*). Both the screening sets were projected through SOM separately, were visualized in the 2D map consisting of a single layer of rectangular grid of nodes (neurons). A network of 15x15 (225 neurons) based on molecular descriptors is displayed in Figure 27 (a-b) against the *M. tuberculosis* screening dataset. The same procedure was repeated with 8x8 (64 neurons) lattice dimension for the *P. aeruginosa* screening dataset.

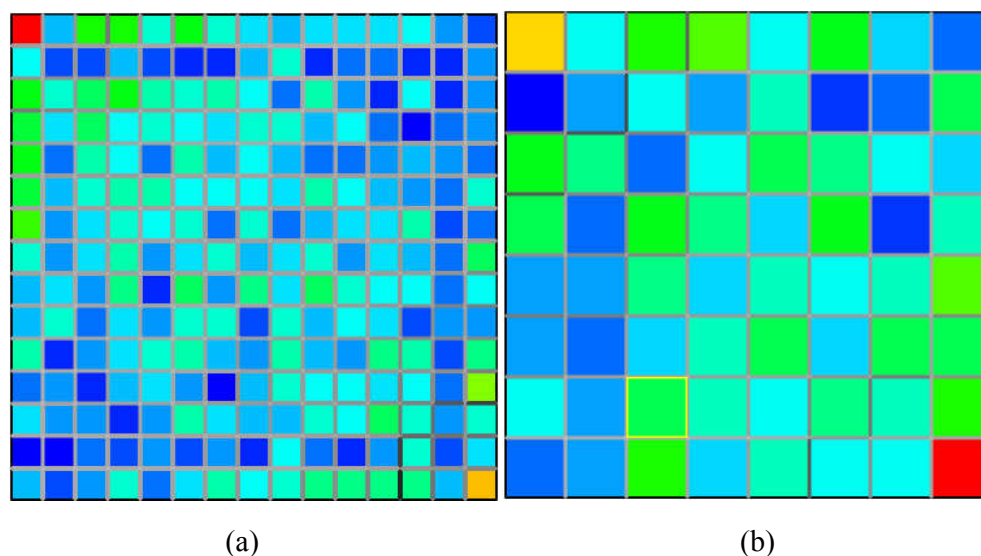


Figure 27. SOM distribution of *M. tuberculosis* and *P. aeruginosa* dataset in the multidimensional 2D map (a) left panel displays the population distribution of the 1368 compounds projected through 225 neurons against *M. tuberculosis* screening set (b) The right panel showing the population distribution of 374 compounds projected through 64 neurons against *P. aeruginosa* is displayed.

7.3.1 Results and Discussion

In this work, we presented and evaluated a method for the identification and prioritization of β -lactamase inhibitors on the basis of Self-Organizing Maps (SOMs). In the SOM analysis results are displayed as a color coded low dimensional 2D map showing the distribution of the compounds among the neurons. The colored cells in the 2D map represent the population; dark blue for empty cells or fewer numbers of molecules and red for cells containing the maximum. The shading of cell borders indicates the distance between adjacent cells; darker borders indicate larger distance and lighter shades for shorter distance in the multidimensional chemical space. Each neuron (cell) was analyzed based on the molecular population distribution in the 2D map as shown in Figure 28. The most important advantage of such method is the ability to interactively analyze the results. By clicking the mouse on each neuron cells, all the compounds are displayed in separate tab which makes easy interpretation of the results.

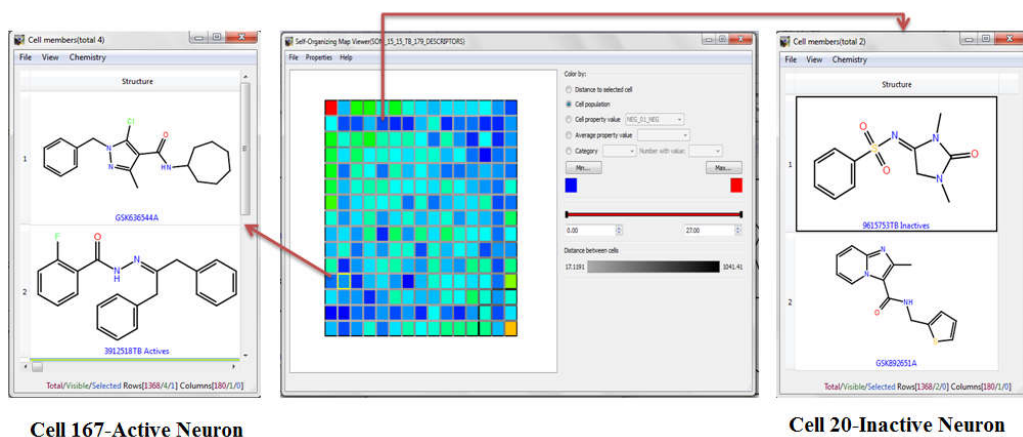


Figure 28. Population distribution of SOM in TB dataset is displayed. SOM 2D map visualized for TB screening set, where the active and inactive neuron is highlighted from their respective cell.

For both screening sets the neurons inclusive of GSK molecule with bioassay actives were treated as active and vice-versa for the inactive ones. Though the study was focused in screening out the molecules in the active neuron, there were mixed neurons (GSK molecules mixed with actives and inactive) and singleton neurons involving only GSK molecules or any other class (actives or inactives from bioassay). The main advantage of such classification was the screening protocol was not only limited to a two class problem involving only actives and inactive neuron but was further extended to four class problem. The number of actives, inactive and the molecules present in the active neurons against SOM TB and Pseudomonas models is shown in Tables 34-36. This signifies the screening ability of the SOM method in classifying actives from inactive in a much finer manner.

Table 34. Number of actives and inactive from the SOM analysis against *M. tuberculosis* and *P. aeruginosa*.

SOM	<i>M. tuberculosis</i>	<i>P. aeruginosa</i>
Actives	6	26
Inactive	29	8

Table 35. TB Actives from 15x15 SOM network (each cell is an active neuron)

Cell Neuron	Screened Molecules
Cell 11	GW360240X, GW876411A, GSK153890A, GSK2157753A
Cell 167	GSK636544A, GSK276001A

Table 36. Pseudomonas actives from 8x8 SOM network (each cell is an active neuron)

Cell Neuron	Screened Molecules
Cell 7	GSK1750922A, GW857165X, GSK1985270A
Cell 8	GSK1832831A
Cell 11	GSK2043267A, GSK1589671A, GSK1598164A
Cell 12	GSK831784A, GSK1744926A
Cell 21	SB-712970, GSK130506A, GSK1731114A, GSK1996236A, GSK146660A
Cell 24	GSK275984A, GW623128X, GSK994258A
Cell 32	GSK754716A, GSK2200157A, GSK1588120A
Cell 42	GSK1650514A
Cell 51	GW360240X, GSK735816A, GSK921190A, GSK1072678A, GSK163574A

Tables 35-36 displays the molecules in the active neuron were retrieved from the multidimensional 2D maps against TB and Pseudomonas screening sets. Each neuron was visualized and the screened molecules were sorted out manually. This was performed in such a manner that the cell (active neuron) consisted at least a minimum of one active molecule from the respective datasets (AID 434987 and AID 2184) were considered as SOM-active and vice-versa for inactive molecules. The same procedure was followed for all the other neurons incorporated with the screening sets. As mentioned above, four class classifications had ruled out the mixed and singleton neuron from the screening set that helped in picking out the actives from inactives in a much better way. ANN based SOM analysis also involves various other parameters by which cell are colored like “distance to selected cell”, “cell property value”, “average property value” and “category” as mentioned in

Figure 28. But in our study we employed the cell population for performing VS and prioritization of anti-TB molecules against the targets under study.

7.4 Conclusion

The present study illustrates the application of Kohonen self-organizing maps (SOMs), in predicting the β -lactamase activity of anti-TB molecules from GSK library from a multidimensional 2D map. Two multidimensional 2D maps were constructed. One map is related to the *M. tuberculosis* (AID 434987) bioassay dataset, involving 1368 molecules that were projected into 225 neurons. While the second map was constructed against *P. aeruginosa* (AID 2184) bioassay dataset consisting of 374 compounds projected through 64 neurons. The virtual screening was analyzed in such a way that each screening set (screening set 1 and screening set 2) involving the bio-active information added to the GSK molecules was allowed to cluster in the SOM. The 2D maps for both species were analyzed based on the population distribution of each neuron to determine activity of a two class problem. The GSK molecules reclining to the active neuron (in AID 434987) were considered to be active and vice-versa for the inactive neuron. The same procedure was followed for AID 2184. From our study we understood that instead of the two class problem involving only actives and inactive neurons, the neurons ended with four class problem involving active, inactive, mixed and singleton. Only active neurons were considered for the screening purpose as a result. We prioritized a total of 32 GSK molecules that were found to be exhibiting β -lactamase activity against *M. tuberculosis* and *P. aeruginosa*. As shown in our study, SOM signifies the screening ability of biologically active molecules predicting actives from inactives in a much better manner due to its algorithmic architecture.

CHAPTER 8

SENSITIVITY OF MOLECULAR DESCRIPTORS BASED VIRTUAL SCREENING METHODS AGAINST β -LACTAMASE ENZYME

8.1 Introduction

Sensitivity is used to understand the behavior of the system being modeled, to verify if the model is doing what it is intended to do, to evaluate the applicability of the model, and to determine the stability of a model.¹²⁷ By definition, sensitivity in data mining (also called in Psychology) is the proportion of real positive cases that are correctly predicted positive and it relates to the test's ability to identify positive results. This measures the coverage of the real positive cases by the +P (predicted positive) rule. Its desirable feature is that it reflects how many of the relevant cases the +P rule picks up. It identifies all real positive cases. In this context it is referred to as TP rate. It is the proportion of actual positives which are predicted positives. Sensitivity can be explained in terms of the mathematical equation (5) as;

$$(TP)/(TP + FN) \quad (5)$$

Where, TP is true positive and FN is the false negative.

Sensitivity analysis has been applied in various fields including complex engineering systems, economics, physics, social sciences, medical decision making, risk assessment and many others. In this study we checked the sensitivity of the computational tools Docking, Bayesian and Artificial Neural Network to understand the pocket dissimilarity against similar strains of β -lactamase enzymes present in *M. tuberculosis* and *P. aeruginosa*. Here the TP and FN are discussed below in the experimental section.

8.2 Materials and Methods

The software's and tools used for the sensitivity are specified in Chapter 2. SIM alignment tool and lanview software was used to analyze protein sequence similarity. The selected proteins are β -lactamase enzyme present in *M. tuberculosis* (PDB id 2GDN) and *P. aeruginosa* (PDB id 2WKH). The bioassays AID 434987 and AID 2184 were selected from PubChem database. The other tools include computational models; docking study (glide suite from Schrodinger software), Bayesian classifier (WEKA data mining software) and ANN SOM software (Canvas suite from Schrodinger software).

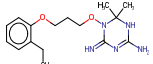
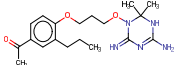
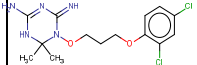
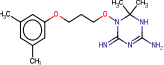
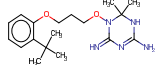
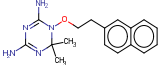
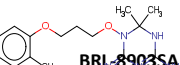

8.3 Experimental Studies









8.3.1 Sensitivity of the computational models

Sensitivity analysis was carried out from the results obtained from the anti-bacterial computational methods Bayesian statistics, molecular docking and ANN based SOM analysis. The activity profile of the 177 GSK anti-TB molecules against the three diverse computational models is displayed in Table 37. Since our study focused on the sensitivity of the computational tools in distinguishing the pockets and which among them are sensitive in pocket similarity/dissimilarity against the anti-bacterial enzyme β -lactamase. For that reason the actives from the mentioned computational methods were used against the two strains of microorganism sharing a common bacterial enzyme β -lactamase.






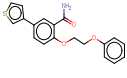


The experiment started from the anti-bacterial β -lactamase actives obtained from the data mining Bayesian classifier and ANN-SOM maps. The Bayesian actives from *M. tuberculosis* and *P. aeruginosa* were compared to check the possibility of common molecules and non common molecules. In a similar way common molecules and non common molecules were sorted from ANN-SOM method and molecular docking along with the docking score as mentioned in Table 37.


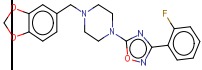
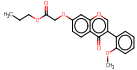





Table 37. Virtual screening results (actives and inactive) from Bayesian (Oversampled ML model), ANN-SOM analysis and molecular docking against β -lactamase enzyme present in *M. tuberculosis* and *P. aeruginosa*.




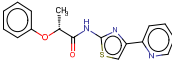
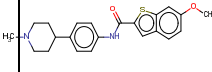



S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
1	 BRL-10143SA	-4.925459	-4.904949	x	✓	-	x	x	x
2	 BRL-10988SA	-4.710024	-5.640783	x	x	x	✓	x	-
3	 BRL-51091AM	-5.744285	-4.820541	✓	✓	-	x	x	-
4	 BRL-51093AM	-5.825566	-5.021055	✓	✓	-	x	x	x
5	 BRL-7940SA	-5.552213	-4.887263	✓	✓	-	x	x	x
6	 BRL-8088SA	-5.591093	-4.341675	✓	✓	-	x	✓	-
7	 BRL-8903SA	-4.83667	-4.832077	x	✓	-	x	x	-
8	 CCI7967	-5.562162	-4.662066	✓	✓	x	x	x	-

S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
9	 G1103688B	-6.082088	-6.249893	✓	✗	-	✓	✗	-
10	 G1247341A	-4.021379	-4.471773	✗	✗	✗	✗	✗	-
11	 GR135486X	-5.078666	-4.339886	✗	✓	-	✗	✗	-
12	 GR135487X	-5.343352	-5.023262	✗	✓	-	✗	✗	-
13	 GR153167X	-4.599758	-3.938989	✗	✓	-	✗	✗	✗
14	 GR223839X	-5.847569	-3.762507	✓	✓	-	✗	✗	✗
15	 GSK1051703A	-6.236361	-4.367053	✓	✓	-	✗	✗	-
16	 GSK1055950A	-4.201657	-4.070941	✗	✗	-	✗	✗	-









S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
17	 GSK1072678A	-4.690078	-4.158446	x	✓	-	x	x	✓
18	 GSK1107112A	-4.378685	-4.724175	x	✓	-	x	x	x
19	 GSK1121877A	-	-5.647076	x	✓	x	✓	x	-
20	 GSK1174628A	-5.340502	-6.136228	x	x	-	✓	✓	-
21	 GSK1180781A	-4.577114	-3.80396	x	x	-	x	x	-
22	 GSK1220329A	-5.197891	-4.540029	x	x	-	x	✓	-
23	 GSK124576A	-4.822552	-4.148671	x	✓	-	x	x	-
24	 GSK124945A	-5.095692	-3.855036	x	x	-	x	x	-

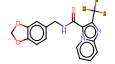
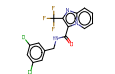
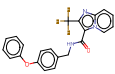
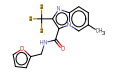
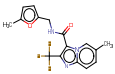
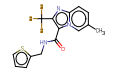
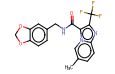
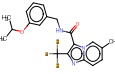
S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
25	 GSK1302651A	-4.824759	-4.540315	x	✓	-	x	✓	-
26	 GSK130506A	-4.794501	-4.09419	x	✓	-	x	x	✓
27	 GSK1310678A	-7.263893	-5.126418	✓	✓	-	x	x	-
28	 GSK1329419A	-5.868955	-5.119634	✓	✓	-	x	x	-
29	 GSK133167A	-5.860265	-4.564013	✓	x	x	x	✓	-
30	 GSK1365028A	-5.795231	-5.524003	✓	✓	-	✓	x	-
31	 GSK1372568A	-4.603222	-3.799022	x	✓	-	x	x	-
32	 GSK1385423A	-5.091986	-3.948361	x	x	-	x	x	-


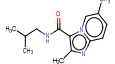
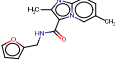
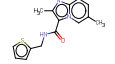



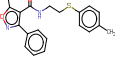
S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
33	 GSK1402290A	-5.752629	-3.949569	✓	✓	-	✗	✓	-
34	 GSK1434490A	-5.618112	-4.58761	✓	✗	✗	✗	✗	-
35	 GSK146660A	-4.098286	-4.206953	✗	✗	-	✗	✓	✓
36	 GSK1518999A	-5.73743	-4.454925	✓	✗	✗	✗	✗	-
37	 GSK1519001A	-4.948158	-4.636751	✗	✓	-	✗	✗	-
38	 GSK153890A	-5.622324	-4.872096	✓	✓	✓	✗	✗	-
39	 GSK1570606A	-5.366569	-4.483416	✗	✓	-	✗	✗	-
40	 GSK1588120A	-4.311554	-5.43586	✗	✓	-	✓	✓	✓


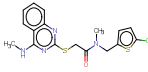



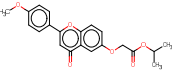


S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
41	 GSK1589671A	-5.512963	-4.685194	✓	✗	-	✗	✗	✓
42	 GSK1589673A	-5.48464	-3.14357	✓	✗	-	✗	✗	-
43	 GSK1598164A	-4.904349	-4.202812	✗	✓	-	✗	✗	✓
44	 GSK1611550A	-5.482702	-4.057204	✓	✓	-	✗	✗	-
45	 GSK1635139A	-5.115107	-1.880659	✗	✓	-	✗	✗	-
46	 GSK163574A	-4.045683	-4.176861	✗	✓	-	✗	✓	✓
47	 GSK1650514A	-4.216401	-4.150087	✗	✓	-	✗	✗	✓
48	 GSK1668869A	-4.524381	-3.916103	✗	✗	-	✗	✗	-



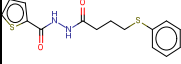
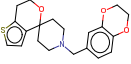
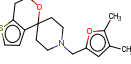
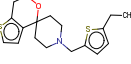

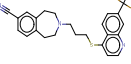
S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
49	 GSK1691553A	-5.228742	-4.130722	x	✓	x	x	x	x
50	 GSK1729177A	-4.280701	-3.808163	x	x	-	x	x	-
51	 GSK1731114A	-5.099303	-4.578677	x	x	-	x	x	✓
52	 GSK1733953A	-4.743176	-4.51376	x	✓	-	x	x	-
53	 GSK1742694A	-5.296827	-4.148205	x	x	-	x	x	-
54	 GSK1744926A	-5.750897	-4.079467	✓	x	-	x	x	✓
55	 GSK1750922A	-4.049035	-3.689326	x	x	-	x	x	✓
56	 GSK1758774A	-5.676211	-4.38173	✓	x	-	x	x	-

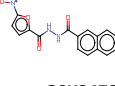
S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
57	 GSK1759150A	-5.434897	-4.386236	x	x	x	x	x	-
58	 GSK1783710A	-4.648597	-3.663739	x	x	x	x	x	-
59	 GSK1788487A	-4.521581	-4.529761	x	✓	x	x	x	-
60	 GSK1812410A	-5.027874	-4.387707	x	✓	-	x	x	-
61	 GSK1826089A	-4.901519	-4.950106	x	x	-	x	x	-
62	 GSK1826247A	-5.262256	-4.092032	x	x	-	x	x	-
63	 GSK1826825A	-5.737437	-4.581582	✓	x	-	x	x	-
64	 GSK1829660A	-6.177868	-4.11545	✓	✓	-	x	✓	-







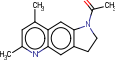
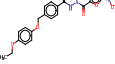
S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
65	 GSK1829671A	-4.675169	-4.352627	x	x	-	x	x	-
66	 GSK1829674A	-4.488235	-4.414894	x	x	-	x	x	-
67	 GSK1829676A	-4.856796	-4.463195	x	x	-	x	x	-
68	 GSK1829727A	-5.095832	-4.65595	x	x	-	x	x	-
69	 GSK1829728A	-5.011102	-4.307195	x	x	-	x	x	-
70	 GSK1829729A	-6.212781	-4.821215	✓	x	-	x	x	-
71	 GSK1829732A	-7.523095	-4.489654	✓	x	-	x	x	-
72	 GSK1829733A	-4.34916	-4.374612	x	x	x	x	x	-



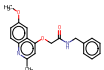
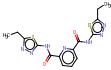

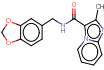
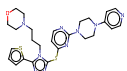

S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
73	 GSK1829736A	-6.236021	-4.381309	✓	✗	-	✗	✗	-
74	 GSK1829816A	-5.02456	-5.217832	✗	✓	-	✗	✓	-
75	 GSK1829819A	-6.185953	-5.316149	✓	✓	-	✗	✓	-
76	 GSK1829820A	-5.352699	-5.143	✗	✓	-	✗	✗	-
77	 GSK1832831A	-6.127531	-4.1216	✓	✓	-	✗	✗	✓
78	 GSK1857145A	-4.167561	-4.108283	✗	✓	-	✗	✗	-
79	 GSK1859936A	-4.71878	-4.184577	✗	✗	-	✗	✗	-
80	 GSK1863309A	-5.087439	-3.487572	✗	✓	-	✗	✗	-

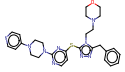
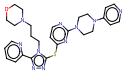

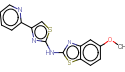
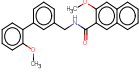


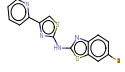
S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
81	 GSK1905227A	-4.360943	-4.053539	x	✓	-	x	✓	-
82	 GSK1925843A	-6.159658	-3.594705	✓	✓	-	x	x	-
83	 GSK1941290A	-4.793047	-3.236873	x	x	x	x	x	-
84	 GSK1955236A	-5.089971	-4.730871	x	✓	-	x	✓	-
85	 GSK1985270A	-3.41155	-4.362997	x	x	-	x	x	✓
86	 GSK1996236A	-4.038049	-4.674773	x	x	-	x	✓	✓
87	 GSK2032710A	-5.295409	-3.755368	x	x	-	x	x	-
88	 GSK2043267A	-5.922338	-4.800879	✓	x	-	x	x	✓

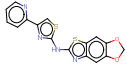

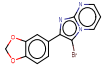
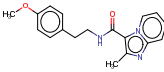
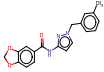
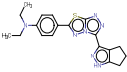

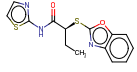
S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
89	 GSK2059310A	-5.550714	-3.926131	✓	✗	-	✗	✗	-
90	 GSK2111534A	-5.818013	-5.102715	✓	✓	-	✗	✓	-
91	 GSK2157753A	-4.900896	-4.082925	✗	✓	✓	✗	✗	-
92	 GSK2200150A	-4.363393	-3.942104	✗	✓	✗	✗	✗	-
93	 GSK2200157A	-3.878424	-4.217975	✗	✓	-	✗	✗	✓
94	 GSK2200160A	-5.354315	-4.396907	✗	✗	-	✗	✗	-
95	 GSK237561A	-6.086947	-4.06445	✓	✓	-	✗	✗	-
96	 GSK254610A	-5.134372	-4.859504	✗	✗	-	✗	✓	-

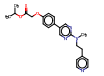

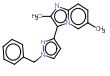
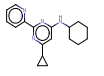
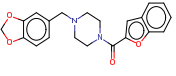

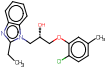
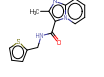
S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
97	 GSK262906A	-4.736116	-3.096196	x	✓	-	x	x	-
98	 GSK270670A	-4.936422	-4.437894	x	x	x	x	x	-
99	 GSK275628A	-4.580708	-4.334243	x	✓	x	x	x	-
100	 GSK275984A	-3.803786	-3.874174	x	✓	-	x	x	✓
101	 GSK276001A	-3.91484	-4.37488	x	✓	✓	x	x	-
102	 GSK316438A	-4.036971	-3.991625	x	✓	x	x	x	-
103	 GSK345724A	-5.190288	-4.572399	x	x	-	x	✓	-
104	 GSK347301A	-3.5351	-4.00121	x	✓	-	x	x	-

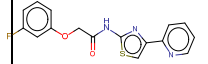
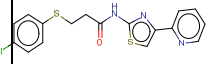

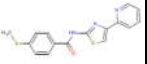




S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
105	 GSK352635A	-6.428837	-4.549626	✓	✗	-	✗	✗	-
106	 GSK353069A	-5.894691	-4.897256	✓	✓	-	✗	✓	-
107	 GSK353071A	-5.081144	-3.963855	✗	✓	-	✗	✓	-
108	 GSK353496A	-5.123067	-4.586514	✗	✓	-	✗	✗	-
109	 GSK358607A	-5.487358	-4.304648	✓	✓	-	✗	✗	-
110	 GSK381407A	-4.497637	-5.0295	✗	✓	✗	✗	✓	-
111	 GSK385518A	-5.607946	-5.884787	✓	✓	-	✓	✓	-
112	 GSK426032A	-4.45128	-4.294382	✗	✗	-	✗	✗	-

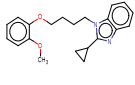

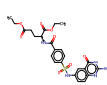


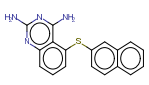


S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
113	 GSK437009A	-4.263662	-4.548821	x	✓	-	x	x	-
114	 GSK445886A	-4.744651	-4.379147	x	✓	-	x	x	-
115	 GSK463114A	-5.66555	-4.185975	✓	✓	-	x	x	-
116	 GSK468214A	-5.309476	-4.1851	x	x	-	x	x	-
117	 GSK479031A	-6.133645	-5.138022	✓	✓	-	x	✓	-
118	 GSK498315A	-6.915556	-5.027683	✓	✓	-	x	✓	-
119	 GSK547481A	-5.166088	-4.787793	x	x	-	x	x	-
120	 GSK547487A	-4.331938	-4.92963	x	x	-	x	x	-

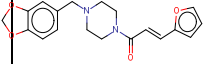
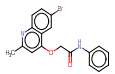
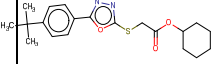
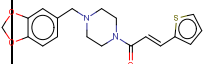



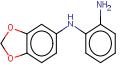
S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
121	 GSK547511A	-5.527249	-3.718299	✓	✗	-	✗	✗	-
122	 GSK547543A	-6.459044	-5.196833	✓	✗	-	✗	✗	-
123	 GSK636544A	-4.138962	-4.103778	✗	✓	✓	✗	✗	-
124	 GSK690382A	-5.355261	-4.176003	✗	✓	-	✗	✗	-
125	 GSK695914A	-5.703629	-4.722775	✓	✓	✗	✗	✓	-
126	 GSK705278A	-5.460412	-5.270821	✗	✓	-	✗	✗	-
127	 GSK731389A	-5.393793	-4.248966	✗	✓	-	✗	✗	-
128	 GSK735816A	-5.660585	-4.433318	✓	✓	-	✗	✗	✓

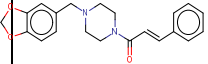

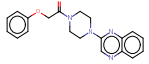
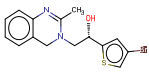
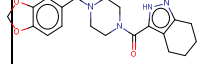



S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
129	 GSK735826A	-5.742838	-4.502916	✓	✓	x	x	x	-
130	 GSK749336A	-5.669826	-4.385861	✓	✓	x	x	x	-
131	 GSK754716A	-4.54943	-3.820543	x	✓	-	x	✓	✓
132	 GSK762874A	-5.720321	-4.422411	✓	✓	-	x	x	-
133	 GSK798463A	-5.907448	-4.749784	✓	✓	-	x	x	-
134	 GSK810016A	-5.442965	-4.220134	x	x	-	x	✓	-
135	 GSK810037A	-5.856571	-4.95457	✓	x	-	x	x	x
136	 GSK829969A	-4.509008	-4.292939	x	x	-	x	x	-

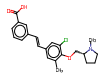
S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
137	 GSK831784A	-5.07314	-4.126366	x	✓	-	x	x	✓
138	 GSK847913A	-3.589416	-3.987311	x	x	x	x	x	-
139	 GSK847920A	-5.48285	-4.012116	✓	✓	-	x	x	-
140	 GSK848336A	-4.451997	-3.855008	x	x	-	x	x	-
141	 GSK861337A	-5.158176	-4.789905	x	x	x	x	x	-
142	 GSK888636A	-4.98263	-3.982144	x	✓	-	x	x	-
143	 GSK889423A	-4.78812	-4.405249	x	x	-	x	x	-
144	 GSK892651A	-5.940762	-5.307386	✓	✓	x	x	x	-

S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
145	 GSK920684A	-5.685314	-4.195351	✓	✓	-	x	x	-
146	 GSK920703A	-5.548655	-4.108738	✓	✓	x	x	x	-
147	 GSK921190A	-5.400508	-3.606963	x	✓	-	x	x	✓
148	 GSK921295A	-5.536	-4.329229	✓	✓	-	x	x	-
149	 GSK937213A	-5.053564	-3.35988	x	x	-	x	x	-
150	 GSK937733A	-5.889491	-4.765693	✓	x	-	x	x	-
151	 GSK957094A	-6.06936	-5.287924	✓	✓	-	x	✓	-
152	 GSK991960A	-4.077586	-3.450489	x	x	-	x	x	-

S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
153	 GSK994258A	-4.397974	-4.151907	x	x	-	x	x	✓
154	 GV187303X	-5.416388	-4.37542	x	✓	-	x	x	-
155	 GW339742X	-6.116786	-5.019073	✓	x	-	x	x	-
156	 GW356807A	-	-5.134416	x	✓	-	x	x	-
157	 GW360240X	-5.401248	-4.9346	x	x	✓	x	✓	✓
158	 GW369335X	-6.700112	-5.424153	✓	✓	x	✓	✓	-
159	 GW623128X	-4.091625	-3.969966	x	✓	-	x	✓	✓
160	 GW664700A	-5.147898	-4.457291	x	✓	-	x	x	-

S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
161	 GW713556X	-3.74818	-4.749814	x	x	-	x	x	-
162	 GW857165X	-4.842953	-4.580092	x	✓	-	x	x	✓
163	 GW859039X	-5.002763	-4.030854	x	✓	-	x	x	-
164	 GW861072X	-4.036657	-3.829027	x	x	-	x	x	-
165	 GW876411A	-4.095902	-4.303594	x	✓	✓	x	x	-
166	 SB-204804-A	-5.032691	-4.410498	x	✓	-	x	✓	-
167	 SB-354364	-4.492187	-4.215178	x	✓	x	x	x	-
168	 SB-435634	-5.886289	-4.782097	✓	✓	-	x	x	-

S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
169	 SB-516933	-3.849518	-4.726249	x	x	x	x	x	-
170	 SB-552112	-5.688828	-4.59511	✓	x	-	x	x	-
171	 SB-650816	-6.236126	-4.773735	✓	x	-	x	✓	-
172	 SB-706404	-5.165198	-5.378961	x	✓	x	x	x	-
173	 SB-712970	-4.067237	-5.446238	x	x	-	✓	x	✓
174	 SB-746177	-5.506079	-4.31357	✓	✓	-	x	✓	-
175	 SB-811137-V	-5.305719	-4.258882	x	x	-	x	x	-
176	 SB-811796-V	-4.812373	-4.422011	x	x	-	x	x	-

S.No.	GSK	TB DOCK SCORE	PSEUDOMONAS DOCK SCORE	TBGLIDE	TBBAYES	TBSOM	PGLIDE	PBAYES	PSOM
177	 SB-829405	-3.141303	-4.582211	✘	✓	-	✘	✘	-

Note: The symbol “✓” and “✘” corresponds to active and inactive against β -lactamase enzymes. And the “-” corresponds to mixed class/Single class (in SOM) and no docking. Pseudomonas- Glide, Bayes, SOM (PGLIDE, PBAYES, PSOM).

The common active molecules and non common molecules were sorted out from the Table 37. The highest number of common actives among the TB and Pseudomonas β -lactamase Bayesian model was twenty four. For the same docking resulted with a lesser number of common actives four and ANN has the least number of common actives as one. Thereafter sensitivity of each method was computed in terms of common molecules (actives) and non common molecules i.e. the molecules commonly found in both *M. tuberculosis* and *P. aeruginosa* in respect to the selected algorithmic methods under study. Consequently the common actives and non-common actives from VS resulted in the similarities and dissimilarities of the pockets for which the sensitivity was studied from the sensitivity equation (5).^{128, 129} The sensitivity used for checking the ML model fineness in the data mining was calculated from the sensitivity equation.

In our study we took this equation and the parameter (TP and FN) were replaced by the number of non common actives (as TP) and number of common actives (FN) and the equation (5) was modified into equation (6) and calculated the sensitivity of all the computational methods as shown in Table 38.

$$\text{Sensitivity} = \frac{\text{Non common molecules}}{\text{Non common molecules} + \text{Common molecules}} \quad (6)$$

Table 38. Sensitivity in terms of pocket dissimilarity of the computational methods against the target β -lactamase enzyme present in *M. tuberculosis* and *P. aeruginosa*.

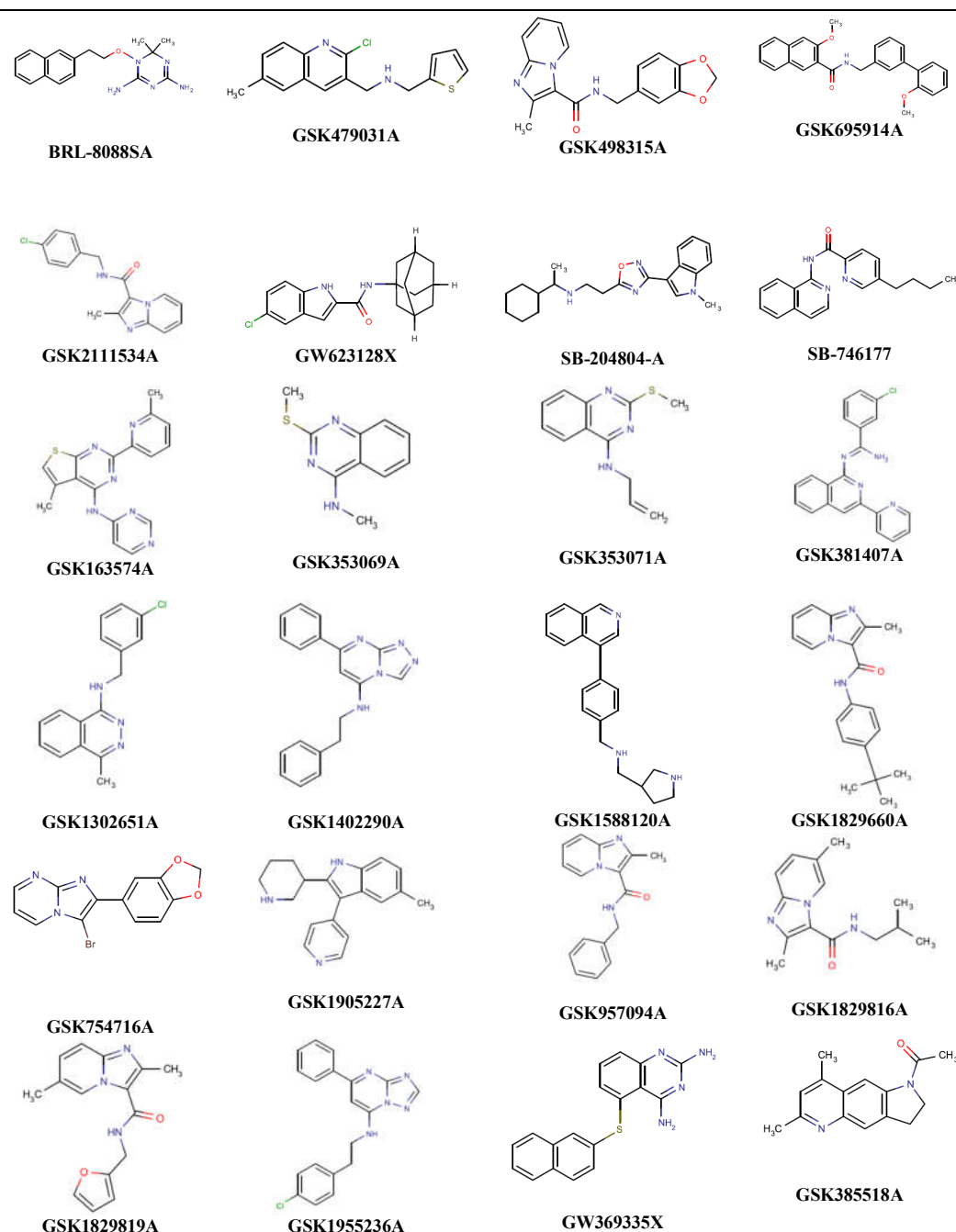
Methods	Actives (<i>M. tuberculosis</i>)	Actives (<i>P. aeruginosa</i>)	Common Active Molecules	Non Common Molecules	Sensitivity
Bayesian	102	34	24	153	0.86
Docking	62	9	4	173	0.97
ANN	6	26	1	176	0.99

From the sensitivity analysis the computational methods under study predicted more of the pocket dissimilarity since the number of non common actives was higher. Even though we received few common molecules as predicted by various computational methods under study. And this parameter common active plays an important role in the sensitivity analysis as it is the one of the deciding factor in the sensitivity equation. So all the common actives from all the computational methods are studied deeply as mentioned in the result and discussion section.

8.3.2 Common actives from Bayesian, Docking and ANN-SOM methods

From the Bayesian model VS against β -lactamase target of *M. tuberculosis* and *P. aeruginosa* we obtained 24 GSK molecules that were common to both as shown in Table 39. The molecules mentioned in Table 39 are β -lactamase actives and common in both *M. tuberculosis* and *P. aeruginosa*. The highlighted molecules GW369335X and GSK385518X are common actives in Bayesian and DBVS methods. The docking results ended with four molecules (GW369335X, GI103688B, GSK1365028A and GSK385518A) that are common actives while Bayesian model resulted 24 and ANN based SOM resulted with one common active (GW360240) among the *M. tuberculosis* and *P. aeruginosa*. Also there were no molecules common in both docking based scoring and ANN-SOM method.

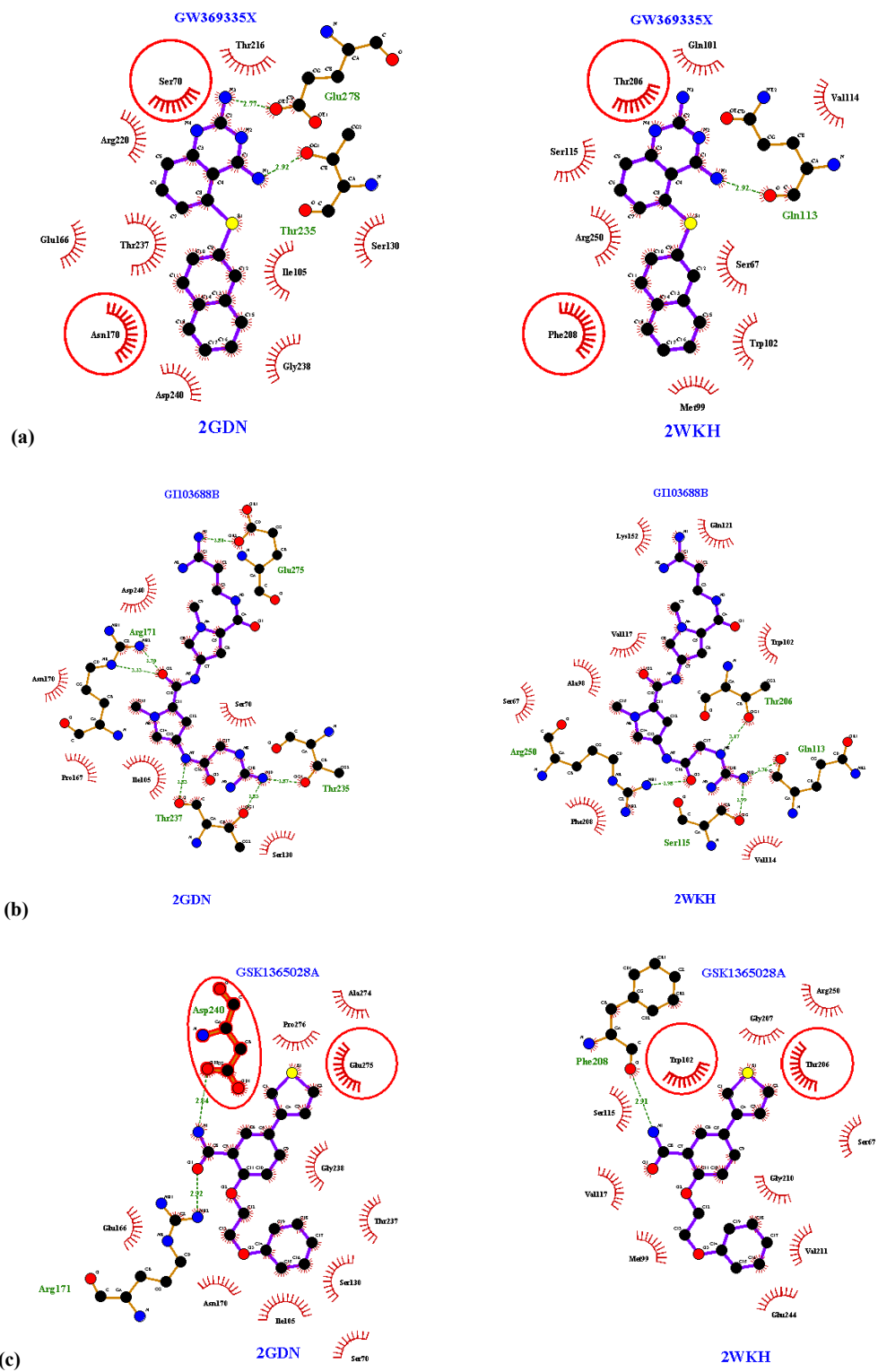
Table 39. Common actives in *M. tuberculosis* and *P. aeruginosa* from the Bayesian models.



Note: The molecules GW369335X and GSK385518A are common actives in molecular docking method.

8.3.2.1 Docking study- Validation of dissimilarity of active site

An in-depth analysis of the pocket dissimilarity was understood from the protein-ligand interactions visualized from ligplot+ program by overlapping the docked molecules of two proteins (TB and Pseudomonas) simultaneously and the active site was visualized 2D. The docking study was carried out against the targets under study (2GDN and 2WKH) where 2WKH is mutated by one residue (K70C-point mutation). For this reason pocket is dissimilar but still we got four molecules (GW369335X, GI103688B, GSK1365028A and GSK385518A) that are common actives in *M. tuberculosis* and *P. aeruginosa*. In depth analysis were carried out with respect to hydrogen bonds and hydrophobic interactions by using ligplot+ program. The ligplot analyses of the four GSK molecules bound to the enzymes 2GDN and 2WKH are shown in Figure 29. Each docked molecule in the active site of 2GDN and 2WKH were overlaid upon each other in the ligplot+ program to visualize the active site in 2D. And the automated generated program superimposed the ligands bound to the active site of 2GDN and 2WKH resulting in the equivalent residues as highlighted in Figure 29 (a-d). The hydrogen bonds, hydrophobic interactions and equivalent residues were visible for all molecules except GI103688B for which the equivalent residues were missing. The glide scores, number of hydrogen bonding interactions, number of hydrogen bonded residues, number of hydrophobic interactions and equivalent residues of the common actives in the pockets of β -lactamase enzyme in *M. tuberculosis* and *P. aeruginosa* are shown in Table 40.



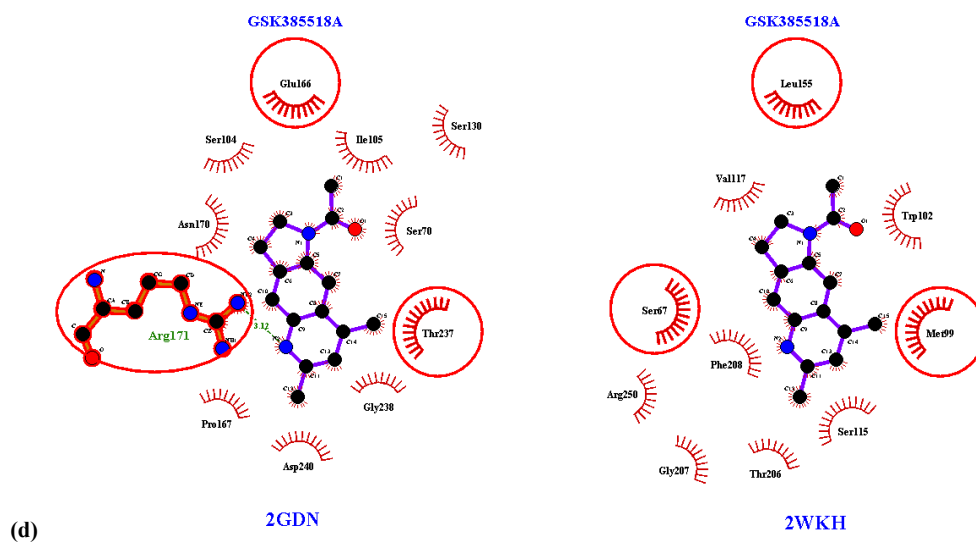
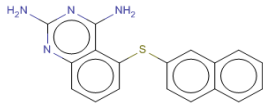
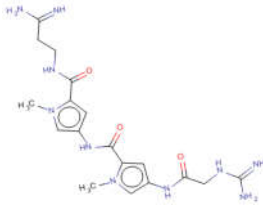
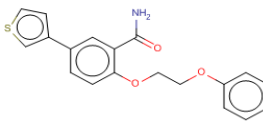
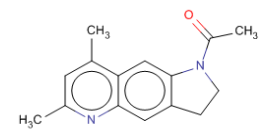


Figure 29 (a-d). Ligand-protein interaction 2D diagrams of the common actives (GW369335X, GI103688B, GSK1365028A, GSK385518A) bounded to the active site of β -lactamases 2GDN (*M. tuberculosis*) and 2WKH (*P. aeruginosa*) is generated from Ligplot+ program. The diagrams in each plot portray the hydrogen bonds interactions (green dotted lines) and hydrophobic interactions (spoked semi circled arcs). The red circles and ellipses identify the equivalent residues and the side chains residues are engaged in hydrophobic interactions are highlighted with a thicker red line.

Table 40. Common actives from the Structure-based Virtual Screen against β -lactamase in *M. tuberculosis* and *P. aeruginosa*.

S.No.	GSK Compounds	Glide score	<i>M. tuberculosis</i>			<i>P. aeruginosa</i>				
			HB *	H-bonded Residues	HYP*	Glide score	HB *	H-bonded Residues	HYP*	Equivalent residues
1	 GW369335X	-6.70	2	Thr235 Glu278	10	-5.42	1	Gln113	9	Ser70, Thr206 Asn170, Phe208
2	 GI103688B	-6.08	6	Arg171 Thr235 Thr237 Glu275	10	-6.24	4	Gln113 Ser115 Thr206 Arg250	12	-
3	 GSK1365028A	-5.79	2	Arg171 Asp240	11	-5.52	1	Phe208	11	Asp240, Trp102 Glu275, Thr206
4	 GSK385518A	-5.60	1	Arg171	11	-5.88	-	-	10	Glu166, Leu155 Arg171, Ser67 Thr237, Met99

HB* = No. of hydrogen bonds, HYP* = No. of hydrophobic interactions

The common actives mentioned in the Table 40 give information about all the molecules except GSK385518A that lack hydrogen bond against the target under study in *P. aeruginosa*. The glide score is highest for GW369335X in TB β -lactamase enzyme and for the same in *P. aeruginosa* is GI103688B. All the molecules do make a good hydrophobic interaction in the active site of both targets.

8.3.2.2 Active site similarity

The structure based sequence alignment was performed for the pocket similarity from SIM programme (<http://web.expasy.org/sim/>) against two strains of β -lactamase enzymes, one from *M. tuberculosis* and the other from *P. aeruginosa*. To comprehend the pocket similarity, class A strains of β -lactamase enzyme from Protein Data Bank (PDB), 2GDN were selected for *M. tuberculosis* and the protein sequence available from the PubChem bioassay 2184 for *P. aeruginosa* were selected in respect of data mining and ANN based SOM method. And for docking study PDB 2GDN and PDB 2WKH (mutated by one residue) of the strain β -lactamase were compared for the pocket dissimilarity check. The amino acid sequences of the bacterial enzymes under study were aligned from SIM program and viewed using Lanlview software as shown in Figure 30. It displays the alignment of FASTA sequence between protein 2GDN and AID 2184 while Figure 31 showcases the amino acid sequence comparison between the selected protein targets 2GDN and 2WKH.

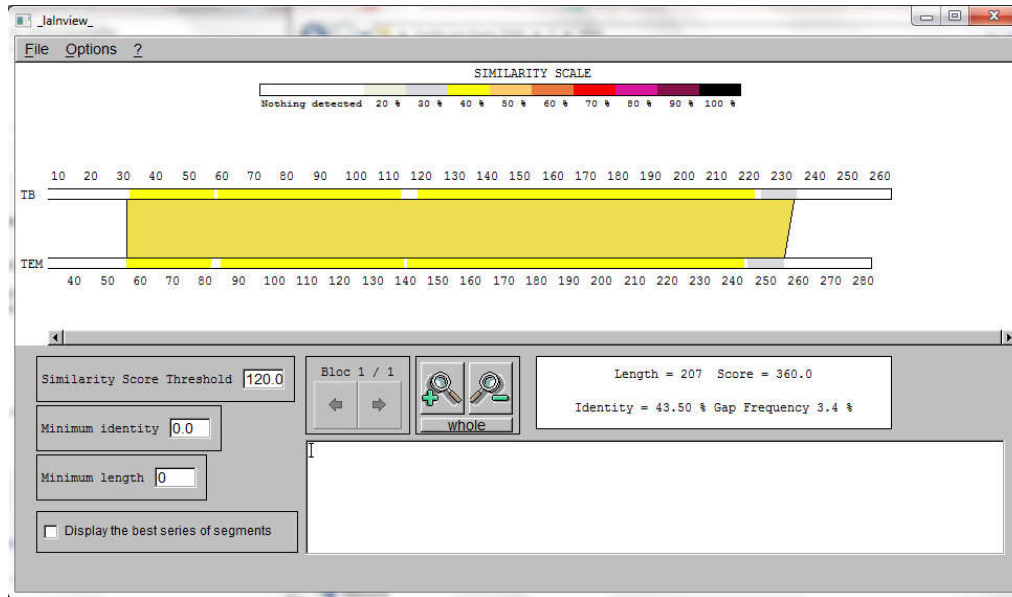


Figure 30. Protein sequence comparison between *M. tuberculosis* (2GDN) and *P. aeruginosa* (FASTA) sequence from AID 2184.

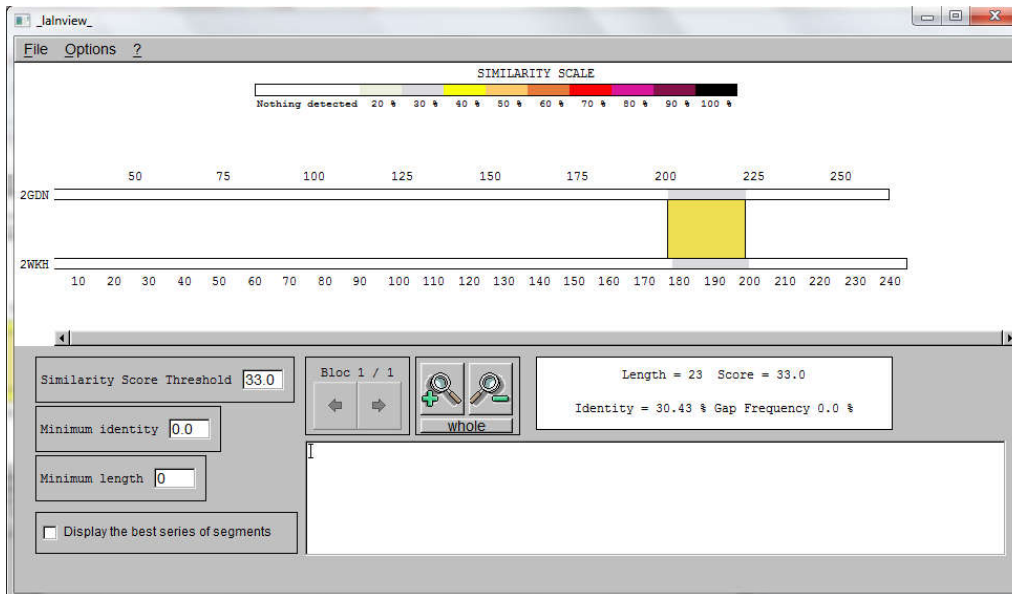


Figure 31. Protein sequence comparison between *M. tuberculosis* (2GDN) and *P. aeruginosa* (2WKH).

As a result, structure based sequence alignment showcased only 43.50% identity between β -lactamase from *M. tuberculosis* and *P. aeruginosa* as shown in Figure 30 while for the same with mutated strain displayed only 30.43% identical

residues (see Figure 31). And thus targets under the study are more towards dissimilarity. At the same time both enzymes possess catalytic residues for *M. tuberculosis*-Ser 70, Lys 73, Ser 130, Lys 234 and Gly 236 and *P. aeruginosa*-Ser 67, Lys 70, Ser 115, Lys 205 and Gly 207. From all the observations involving the number of common actives from different computational methods, comparative study of sequence alignments and the presence of catalytic residue from this we figured out that there is dissimilarity on the proteins structure. When considered, ML models based virtual screening and ANN based SOM analysis both methods were provided with equal input dataset and the dimensionality of the descriptor set remained same. From this study it was clear that both methods resulted with more dissimilarity since from Bayesian based ML screen generated only 24 common active (out of 177) and for the same for ANN based SOM analysis predicted 1. To validate this result we docked the 177 GSK anti-TB molecules into the selected targets (2GDN and 2WKH) of β -lactamase enzymes present in *M. tuberculosis* and *P. aeruginosa*. As mentioned previously the target enzyme 2WKH is a mutated (K70C) strain and Lys70 is an important residue in the cavity of the enzyme. Since the pockets under study are dissimilar and the common actives (which is fewer in number only 4 out of 177) obtained by the comparative study of docking based virtual screening emphasize that the pockets are dissimilar. Thus pocket dissimilarity was analyzed through sensitivity a data mining based model robustness check was introduced in our study and looked out which among the three are more sensitive in predicting pocket dissimilarity. As a result structure based docking study is more sensitive to Bayesian based model screen but less compared to ANN based SOM method. And the superiority of SOM analysis can be explained because of the introduction of high dimensional data and the innate algorithmic architecture of the ANN-SOM method. Finally we concluded that the sensitivity of the computational tool is determined by the active molecules that exist in common as it is one of the factors in the sensitivity equation. Ultimately each virtual screening method makes use of the algorithm in predicting the output and sensitivity was used as a measure to check the accuracies of the algorithmic architecture.

8.4 Conclusion

From our study we understood that sensitivity of the computational methods is essential in differentiating the pockets of structurally dissimilar targets of serine β -lactamase enzyme present in *M. tuberculosis* and *P. aeruginosa*. The hypothesis in our study was to check the sensitivity of the computational tools Docking, Bayesian and Artificial Neural Network in understanding the pocket dissimilarity against similar strains of β -lactamase enzymes present in *M. tuberculosis* and *P. aeruginosa* i.e. which method in perspective of algorithm is better in distinguishing the pocket dissimilarity of targets mentioned in the study. For that reason we performed virtual screening against each method and studied the pocket dissimilarity in terms of common and non common actives from the GSK library consisting of 177 molecules. The Bayesian and ANN-SOM methods are descriptor based analysis while molecular docking is a structure based study. Later performed the virtual screening against the computational models and results were analyzed for pocket dissimilarity in terms of common actives. We also checked the proteins sequence of the targets under study. Both virtual screening and sequence comparison ended with target dissimilarity. Since the targets are dissimilar we took a mutated protein strain for *P. aeruginosa* and non mutated strain for *M. tuberculosis* and performed docking study. And finally sensitivity was checked in terms of the computational common β -lactamase actives derived from three computational methods under study. The pocket dissimilarity is more sensitively proved by the ANN based SOM method in comparison to Bayesian ML model and structure based molecular docking. While in docking study the pocket dissimilarity was checked by mutating one of the active site residues (K70C) present in *P. aeruginosa*. Thus both pockets of the selected targets are non identical and docking is more sensitive with respect to Bayesian models but less compared to ANN based SOM method. From our study, we could also establish the relevance of high dimensional interpretation in understanding the pocket dissimilarity of two similar strains of β -lactamase enzymes which is a hypothesis and result based analysis and hence no reference is provided.

REFERENCES

- ¹ Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S.; Brudz, S.; Sullivan, J. P.; Muhlich, J.; Serrano, M.; Ferraiolo, P.; Tolliday, N. J.; Schreiber, S. L.; Clemons, P. A. ChemBank: A Small-Molecule Screening and Cheminformatics Resource Database. *Nucleic Acids Res.* **2008**, *36*.
- ² Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, 198–201.
- ³ Schomburg, I. BRENDA, Enzyme Data and Metabolic Information. *Nucleic Acids Res.* **2002**, *30*, 47–49.
- ⁴ Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36*, 901–906.
- ⁵ Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, 1100–1107.
- ⁶ Jamal, S.; Scaria, V. Cheminformatic Models Based on Machine Learning for Pyruvate Kinase Inhibitors of *Leishmania Mexicana*. *BMC Bioinformatics* **2013**, *14*, 329.
- ⁷ Shahlai, M. Descriptor Selection Methods in Quantitative Structure-Activity Relationship Studies: A Review Study. *Chem. Rev.* **2013**, *113*, 8093–8103.
- ⁸ Lee, C. H.; Huang, H. C.; Juan, H. F. Reviewing Ligand-Based Rational Drug Design: The Search for an ATP Synthase Inhibitor. *Int. J. Mol. Sci.* **2011**, *12*, 5304–5318.
- ⁹ Durdagi, S.; Mavromoustakos, T.; Papadopoulos, M. G. 3D QSAR CoMFA/CoMSIA, Molecular Docking and Molecular Dynamics Studies of Fullerene-Based HIV-1 PR Inhibitors. *Bioorganic Med. Chem. Lett.* **2008**, *18*, 6283–6289.
- ¹⁰ Hecht, D. Applications of Machine Learning and Computational Intelligence to Drug Discovery and Development. *Drug Dev. Res.* **2011**, *72*, 53–65.
- ¹¹ Reddy, M. R.; Abby, L. P.; Therapeutics, M.; Diego, S. Overview of Rational Drug Design. *Ration. Drug Des. Am. Chem. Soc.* **1999**, *719*, 1–11.
- ¹² Baldi, A. Computational Approaches for Drug Design and Discovery: An Overview. *Syst. Rev. Pharm.* **2010**, *1*, 99.

-
- 13 Aparoy, P.; Reddy, K. K.; Reddanna, P. Structure and Ligand Based Drug Design Strategies in the Development of Novel 5-LOX Inhibitors. *Curr Med Chem* **2012**, *19*, 3763–3778.
- 14 Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Comput. Methods Drug Discov.* **2014**, *66*, 334–395.
- 15 Baldi, A. Computational Approaches for Drug Design and Discovery: An Overview. *Syst. Rev. Pharm.* **2010**, *1*, 99.
- 16 Guido, R. V. C.; Oliva, G.; Andricopulo, A. D. Structure-and Ligand-Based Drug Design Approaches for Neglected Tropical Diseases*. *Pure Appl. Chem* **2012**, *84*, 1857–1866.
- 17 Stevens, E. Lead Discovery. In *Medicinal Chemistry - The Modern Drug Discovery Process*; Pearson Education, Inc, 2014; p 439.
- 18 Davis, B. J.; Roughley, S. D. Fragment-Based Lead Discovery. *Annu. Rep. Med. Chem.* **2017**, *3*, 660–672.
- 19 Hughes, J. P.; Rees, S. S.; Kalindjian, S. B.; Philpott, K. L. Principles of Early Drug Discovery. *Br. J. Pharmacol.* **2011**, *162*, 1239–1249.
- 20 Ramírez, D. Computational Methods Applied to Rational Drug Design. *Open Med. Chem. J.* **2016**, *10*, 7–20.
- 21 Bielska, E.; Lucas, X.; Czerwoniec, A.; Kasprzak, J. M.; Kaminska, K. H.; Bujnicki, J. M. Virtual Screening Strategies in Drug Design - Methods and Applications. *Biotechnologia* **2011**, *92*, 249–264.
- 22 Walker, J. M. *Drug Designing and Discovery Methods and Protocols*; Seetharama D. Satyanarayanajois, Ed.; Springer Science+Business Media, LLC 2011, 2011.
- 23 Monev, V. Introduction to Similarity Searching in Chemistry. *MATCH Commun. Math. Comput. Chem* **2004**, *51*, 7–38.
- 24 Heikamp, K. Application and Development of Computational Methods for Ligand-Based Virtual Screening, 2014.
- 25 Michalsky, E.; Dunkel, M.; Goede, A.; Preissner, R. SuperLigands - a Database of Ligand Structures Derived from the Protein Data Bank. *BMC Bioinformatics* **2005**, *6*, 122.
- 26 Qing, X.; Lee, X. Y.; De Raeymaeker, J.; Tame, J. R.; Zhang, K. Y.; De Maeyer, M.; Voet, A. R. Pharmacophore Modeling: Advances, Limitations, And Current Utility in Drug Discovery. *J. Receptor. Ligand Channel Res.* **2014**, *7*, 81–92.

-
- 27 Lee, C. H.; Huang, H. C.; Juan, H. F. Reviewing Ligand-Based Rational Drug Design: The Search for an ATP Synthase Inhibitor. *Int. J. Mol. Sci.* **2011**, *12*, 5304–5318.
- 28 Bielska, E.; Lucas, X.; Czerwoniec, A.; Kasprzak, J. M.; Kaminska, K. H.; Bujnicki, J. M. Virtual Screening Strategies in Drug Design - Methods and Applications. *Biotechnologia* **2011**, *92*, 249–264.
- 29 Zaman, K. Tuberculosis: A Global Health Problem. *J. Heal. Popul. Nutr.* **2010**, *28*, 111–113.
- 30 WHO. *Global Tuberculosis Report 2016*; 2016.
- 31 Hum Nath Jnawali, S. R. First – and Second – Line Drugs and Drug Resistance. In *Tuberculosis-Current issues in diagnosis and management*; Bassam H. Mahboub and Mayank G. Vats, Ed.; InTech, 2013; pp 163–180.
- 32 Sisay, S.; Mengistu, B.; Erku, W.; Woldeyohannes, D. Directly Observed Treatment Short-Course (DOTS) for Tuberculosis Control Program in Gambella Regional State, Ethiopia: Ten Years Experience. *BMC Res. Notes* **2014**, *7*, 44.
- 33 Deun, A. Van; Rieder, H. L. DOT, S, or DOTS? *International Union Against Tuberc. Lung Dis. Heal.* **2012**, *2*, 3–4.
- 34 Sandanayaka, V. P.; Prashad, A. S. Resistance to β -Lactam Antibiotics: Structure and Mechanism Based Design of β -Lactamase Inhibitors. *Curr. Med. Chem.* **2002**, *9*, 1145–1165.
- 35 Sabath, L. D. Mechanisms of Resistance to Beta-Lactam Antibiotics in Strains of Staphylococcus Aureus. *Ann.Intern.Med.* **1982**, *97*, 339–344.
- 36 Lin, J.; Nishino, K.; , Marilyn C. Roberts , Marcelo Tolmasky, R. I. A. and; Zhang, L. Mechanisms of Antibiotic Resistance. *Front. Microbiol.* **2015**, *6*, 1–3.
- 37 Hugonnet, J.-E.; Blanchard, J. S. Irreversible Inhibition of the Mycobacterium Tuberculosis Beta-Lactamase by Clavulanate. *Biochemistry* **2007**, *46*, 11998–12004.
- 38 Kurz, S. G.; Wolff, K. A.; Hazra, S.; Bethel, C. R.; Hujer, A. M.; Smith, K. M.; Xu, Y.; Tremblay, L. W.; Blanchard, J. S.; Nguyen, L.; Bonomo, R. A. Can Inhibitor-Resistant Substitutions in the Mycobacterium Tuberculosis β -Lactamase BlaC Lead to Clavulanate Resistance?: A Biochemical Rationale for the Use of β -Lactam- β -Lactamase Inhibitor Combinations. *Antimicrob. Agents Chemother.* **2013**, *57*, 6085–6096.

-
- 39 Tasha Smith, Kerstin A. Wolff, and L. N. Molecular Biology of Drug Resistance in Mycobacterium Tuberculosis Tasha. *Curr Top Microbiol Immunol* **2013**, *374*, 53–80.
- 40 Dever, Laura A, T. S. D. Mechanisms of Bacterial Resistance to Antibiotics. *Arch. Intern. Med.* **1991**, *151*, 886–895.
- 41 Drawz, S. M.; Bonomo, R. a. Three Decades of β -Lactamase Inhibitors. *Clin. Microbiol. Rev.* **2010**, *23*, 160–201.
- 42 Sung Joon Kim, James Chang, and M. S. Peptidoglycan Architecture of Gram-Positive Bacteria by Solid- State NMR. *Biochim Biophys Acta* **2015**, *1848*, 350–362.
- 43 Lenka Malinicova, Maria Piknova, Peter Pristas, P. J. Peptidoglycan Hydrolases as Novel Tool for Anti-Enterococcal Therapy. *Curr. Res. Technol. Educ. Top. Appl. Microbiol. Microb. Biotechnol.* **2010**, 463–472.
- 44 Worthington, R. J.; Melander, C. Overcoming Resistance to β -Lactam Antibiotics. *J Org Chem.* **2013**, *78*, 4207–4213.
- 45 Yocum, R. R.; Waxman, D. J.; Rasmussen, J. R.; Strominger, J. L. Mechanism of Penicillin Action: Penicillin and Substrate Bind Covalently to the Same Active Site Serine in Two Bacterial D-Alanine Carboxypeptidases. *Proc. Natl. Acad. Sci. U. S. A.* **1979**, *76*, 2730–2734.
- 46 Kohanski, M. A.; Dwyer, D. J.; Collins, J. J. How Antibiotics Kill Bacteria: From Targets to Networks. *Nat. Rev. Microbiol.* **2010**, *8*, 423–435.
- 47 Bush, K.; Jacoby, G. A. Updated Functional Classification of β -Lactamases. *Antimicrob. Agents Chemother.* **2010**, *54*, 969–976.
- 48 Ghafourian, S.; Sadeghifard, N.; Soheili, S.; Sekawi, Z. Extended Spectrum Beta-Lactamases: Definition, Classification and Epidemiology. *Curr. Issues Mol. Biol.* **2014**, *17*, 11–22.
- 49 Hakime Öztürk, Elif Ozkirimli, A. O. Classification of Beta-Lactamases and Penicillin Binding Proteins Using Ligand-Centric Network Models. *PLoS One* **2015**, *10*, 1–23.
- 50 Hazra, S.; Kurz, S. G.; Wolff, K.; Nguyen, L.; Bonomo, R. A.; Blanchard, J. S. Kinetic and Structural Characterization of the Interaction of 6-Methylidene Penem 2 with the β -Lactamase from Mycobacterium Tuberculosis. **2015**, *54*, 5657–5664.
- 51 Palzkill, T. Metallo- β -Lactamase Structure and Function. *Ann N Y Acad Sci* **2014**, *1277*, 91–104.
- 52 Jacoby, G. A. AmpC β -Lactamases. *Clin. Microbiol. Rev.* **2009**, *22*, 161–182.

-
- 53 Antunes, N.; Fisher, J. Acquired Class D β -Lactamases. *Antibiotics* **2014**, *3*, 398–434.
- 54 Bush, K. β -Lactamase Inhibitors from Laboratory to Clinic. *Clin. Microbiol. Rev.* **1988**, *1*, 109–123.
- 55 Payne, D. J.; Cramp, R.; Winstanley, D. J.; Knowles, D. J. C. Comparative Activities of Clavulanic Acid, Sulbactam, and Tazobactam against Clinically Important β -Lactamases. *Antimicrob. Agents Chemother.* **1994**, *38*, 767–772.
- 56 Higgins, P. G.; Wisplinghoff, H.; Stefanik, D.; Seifert, H. In Vitro Activities of the β -Lactamase Inhibitors Clavulanic Acid, Sulbactam, and Tazobactam Alone or in Combination with β -Lactams against Epidemiologically Characterized Multidrug-Resistant *Acinetobacter Baumannii* Strains. *Antimicrob. Agents Chemother.* **2004**, *48*, 1586–1592.
- 57 Paradkar, A. Clavulanic Acid Production by *Streptomyces Clavuligerus*: Biogenesis, Regulation and Strain Improvement. *J. Antibiot. (Tokyo)*. **2013**, *66*, 411–420.
- 58 Betrosian, A. P.; Douzinas, E. E. Ampicillin-Sulbactam: An Update on the Use of Parenteral and Oral Forms in Bacterial Infections. *Expert Opin. Drug Metab. Toxicol.* **2009**, *5*, 1099–1112.
- 59 Toussaint, K. A. P.; Gallagher, J. C. β -Lactam/ β -Lactamase Inhibitor Combinations: From Then to Now. *Ann. Pharmacother.* **2014**, 1–13.
- 60 Yewale, V.; Dharmapalan, D. Beta Lactam and Beta Lactamase Inhibitor Combinations. *Pediatr. Infect. Dis.* **2010**, *2*, 69–76.
- 61 Farmer, T. H.; Degnan, B. A.; Payne, D. J. Penetration of β -Lactamase Inhibitors into the Periplasm of Gram-Negative Bacteria. *FEMS Microbiol. Lett.* **1999**, *176*, 11–15.
- 62 Frase, H.; Smith, C. A.; Toth, M.; Champion, M. M.; Mobashery, S.; Vakulenko, S. B. Identification of Products of Inhibition of GES-2 β -Lactamase by Tazobactam by X-Ray Crystallography and Spectrometry. *J. Biol. Chem.* **2011**, *286*, 14396–14409.
- 63 Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem BioAssay: 2014 Update. *Nucleic Acids Res.* **2014**, *42*, 1075–1082.
- 64 Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 Update. *Nucleic Acids Res.* **2017**, *45*, D955–D963.

-
- 65 National Center for Biotechnology Information. PubChem BioAssay Database; AID=434987, <https://pubchem.ncbi.nlm.nih.gov/bioassay/434987> (accessed Oct. 29, 2017).
- 66 National Center for Biotechnology Information. PubChem BioAssay Database; AID=2184, <https://pubchem.ncbi.nlm.nih.gov/bioassay/2184> (accessed Oct. 29, 2017).
- 67 Liu, K.; Feng, J.; Young, S. S. PowerMV: A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation. *J. Chem. Inf. Model.* **2005**, *45*, 515–522.
- 68 Syed Asif Hassan, A. H. O. An Improved Machine Learning Approach to Enhance the Predictive Accuracy for Screening Potential Active USP1 / UAF1 Inhibitors. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 144–148.
- 69 Hall M, Eibe F, Holmes G, Pfahringer B, et al. The WEKA Data Mining Software: An Update. *SIGKDD Explor.* **2009**, *11*, 10–18.
- 70 J Uma Mahesh, K.V.Naganjaneyulu, P. Likitha, K. N. S. S. A. Analysis of J48 Algorithm in Classification- Ebola Virus. *Int. J. Emerg. Trends Sci. Technol.* **2014**, *1*, 1289--1292.
- 71 Laskowski, R. A.; Swindells, M. B. LigPlot +: Multiple Ligand À Protein Interaction Diagrams for Drug Discovery. *Chem. Inf. Model.* **2011**, *51*, 2778–2786.
- 72 <https://www.ebi.ac.uk/thornton-srv/software/LigPlus/manual/manual.html>
- 73 Cheng, Feng, V. S. Applications of Artificial Neural Network Modeling in Drug Discovery. *Clin. Exp. Pharmacol.* **2012**, *2*, 2–3.
- 74 Tino, P.; Benuskova, L.; Sperduti, A. Artificial Neural Network Models. In *Handbook of computational Intelligence*; Janusz Kacprzyk, Pedrycz, W., Eds.; Springer-Verlag Berlin Heidelberg, 2015; Vol. 8, pp 455–472.
- 75 Roohi, F. Artificial Neural Network Approach to Clustering. *Int. J. Eng. Sci.* **2013**, *2*, 33–38.
- 76 Andrej Krenker, J. B. and A. K. Introduction to the Artificial Neural Networks. In *Artificial Neural Networks - Methodological Advances and Biomedical Applications*; Prof. Kenji Suzuki, Ed.; InTech, 2011; p 362.
- 77 Mohammad Dorofki, Ahmed H. Elshafie, Othman Jaafar, O. A. K. and S. M. Comparison of Artificial Neural Network Transfer Functions Abilities to Simulate Extreme Runoff Data. In *International Conference on Environment, Energy and Biotechnology (IPCBE)*; 2012; Vol. 33, pp 39–44.

-
- 78 Marini, F. Artificial Neural Networks in Foodstuff Analyses: Trends and Perspectives A Review. *Anal. Chim. Acta* **2009**, *635*, 121–131.
- 79 Zgurovsky, M. Z.; Zaychenko, Y. P. Neural Networks with Feedback and Self-Organization. In *The Fundamentals of Computational Intelligence: System Approach*; M.Z. Zgurovsky and Y.P. Zaychenko, Ed.; Springer International Publishing Switzerland, 2016; Vol. 652, pp 39–79.
- 80 Wankhede, S. B. Analytical Study of Neural Network Techniques: SOM, MLP and Classifier-A Survey. *IOSR J. Comput. Eng. Ver. VII* **2014**, *16*, 86–92.
- 81 Fuzail Misarwala, KausarMukadam, and K. B. Applications of Data Mining in Fraud Detection. *Int. J. Comput. Sci. Eng. (e-IJCSE 2347-269)* **2015**, *3*, 45–53.
- 82 Lusa, L.; Others. SMOTE for High-Dimensional Class-Imbalanced Data. *BMC Bioinformatics* **2013**, *14*, 106.
- 83 Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.
- 84 Ballell, L.; Bates, R. H.; Young, R. J.; Alvarez-Gomez, D.; Alvarez-Ruiz, E.; Barroso, V.; Blanco, D.; Crespo, B.; Escribano, J.; Gonzalez, R.; Lozano, S.; Huss, S.; Santos-Villarejo, A.; Martin-Plaza, J. J.; Mendoza, A.; Rebollo-Lopez, M. J.; Remuinan-Blanco, M.; Lavandera, J. L.; Perez-Herran, E.; Gamo-Benito, F. J.; Garcia-Bustos, J. F.; Barros, D.; Castro, J. P.; Cammack, N. Fueling Open-Source Drug Discovery: 177 Small-Molecule Leads against Tuberculosis. *ChemMedChem* **2013**, *8*, 313–321.
- 85 Martínez-Jiménez, F.; Papadatos, G.; Yang, L.; Wallace, I. M.; Kumar, V.; Pieper, U.; Sali, A.; Brown, J. R.; Overington, J. P.; Marti-Renom, M. a. Target Prediction for an Open Access Set of Compounds Active against Mycobacterium Tuberculosis. *PLoS Comput. Biol.* **2013**, *9*, e1003253.
- 86 Huang, X. A Time-Efficient, Linear-Space Local Similarity Algorithm. *Adv. Appl. Math.* **1991**, *12*, 337–357.
- 87 Duret, L.; Gasteiger, E.; Perrière, G. LALNVIEW: A Graphical Viewer for Pairwise Sequence Alignments. *Comput. Appl. Biosci.* **1996**, *12*, 507–510.
- 88 Kasner, E.; Hunter, C. A.; Ph, D.; Kariko, K.; Ph, D. Ligand Classifier of Adaptively Boosting Ensemble Decision Stumps (LiCABEDS) and Its Application on Modeling Ligand Functionality for 5HT-Subtype GPCR Families. *Chem. Inf. Model.* **2011**, *51*, 521–531.

-
- 89 Su, R.; Li, Y.; Zink, D.; Loo, L.-H. Supervised Prediction of Drug-Induced Nephrotoxicity Based on Interleukin-6 and -8 Expression Levels. *BMC Bioinformatics* **2014**, *15*, S16.
- 90 Zhang, Z. Naive Bayes Classification in R. *Ann. Transl. Med.* **2016**, *4*, 241.
- 91 Wang, Q.; Luo, Z. H.; Huang, J. C.; Feng, Y. H.; Liu, Z. A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM. *Comput. Intell. Neurosci.* **2017**, *2017*, 11.
- 92 Kingsford, C.; Salzberg, S. What Are Decision Trees? *Nat. Biotechnol.* **2008**, *26*, 1011–1013.
- 93 Barros, R. C.; Winck, A. T.; Machado, K. S.; Basgalupp, M. P.; de Carvalho, A. C.; Ruiz, D. D.; de Souza, O. N. Automatic Design of Decision-Tree Induction Algorithms Tailored to Flexible-Receptor Docking Data. *BMC Bioinformatics* **2012**, *13*, 310.
- 94 Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *WIREs Comput. Mol. Sci.* **2014**, *4*, 468–481.
- 95 Marbach, D.; Costello, J. C.; Küffner, R.; Vega, N. M.; Prill, R. J.; Camacho, D. M.; Allison, K. R.; Aderhold, A.; Allison, K. R.; Bonneau, R.; Camacho, D. M.; Chen, Y.; Collins, J. J.; Cordero, F.; Costello, J. C.; Crane, M.; Dondelinger, F.; Drton, M.; Esposito, R.; Foygel, R.; de la Fuente, A.; Gertheiss, J.; Geurts, P.; Greenfield, A.; Grzegorzczak, M.; Haury, A.-C.; Holmes, B.; Hothorn, T.; Husmeier, D.; Huynh-Thu, V. A.; Irrthum, A.; Kellis, M.; Karlebach, G.; Küffner, R.; Lèbre, S.; De Leo, V.; Madar, A.; Mani, S.; Marbach, D.; Mordeliet, F.; Ostrer, H.; Ouyang, Z.; Pandya, R.; Petri, T.; Pinna, A.; Poultney, C. S.; Prill, R. J.; Rezny, S.; Ruskin, H. J.; Saeys, Y.; Shamir, R.; Sîrbu, A.; Song, M.; Soranzo, N.; Statnikov, A.; Stolovitzky, G.; Vega, N.; Vera-Licona, P.; Vert, J.-P.; Visconti, A.; Wang, H.; Wehenkel, L.; Windhager, L.; Zhang, Y.; Zimmer, R.; Kellis, M.; Collins, J. J.; Stolovitzky, G. Wisdom of Crowds for Robust Gene Network Inference. *Nat. Methods* **2012**, *9*, 796–804.
- 96 Boulesteix, A. L.; Janitza, S.; Kruppa, J.; König, I. R. Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 493–507.
- 97 Rezapour, M.; Khavanin Zadeh, M.; Sepehri, M. M. Implementation of Predictive Data Mining Techniques for Identifying Risk Factors of Early AVF Failure in Hemodialysis Patients. *Comput. Math. Methods Med.* **2013**, *2013*, 8.
- 98 Ingolikar, R. A.; Gedam, S. R. Star Classification Using Tree Based Data Mining Techniques. *IOSR J. Comput. Eng.* **2016**, 30–36.

-
- 99 Kapoor, P.; Rani, R. A Survey of Classification Methods Utilizing Decision Trees. *Int. J. Eng. Trends Technol.* **2015**, *22*, 188–194.
- 100 Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug Discovery Using Support Vector Machines. The Case Studies of Drug-Likeness, Agrochemical-Likeness, and Enzyme Inhibition Predictions. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048–2056.
- 101 Chen, P.; Lee, T.; Lee, Y. Multiclass Support Vector Classification via Regression. In *Chen2006MulticlassSV*; 2006.
- 102 Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: Computed Atlas of Surface Topography of Proteins with Structural and Topographical Mapping of Functionally Annotated Residues. *Nucleic Acids Res.* **2006**, *34*, 116–118.
- 103 Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.
- 104 Tovchigrechko, A.; Wells, C. a; Vakser, I. a. Docking of Protein Models. *Protein Sci.* **2002**, *11*, 1888–1896.
- 105 Kroemer, R. T. Structure-Based Drug Design: Docking and Scoring. *Curr. Protein Pept. Sci.* **2007**, *8*, 312–328.
- 106 Meng XY, Zhang HX, Mezei M, C. M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr Comput Aided Drug Des* **2011**, *7*, 146–157.
- 107 Jain, A. N. Scoring Functions for Protein-Ligand Docking. *Curr. Protein Pept. Sci.* **2006**, *7*, 407–420.
- 108 Liu, J.; Wang, R. Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **2015**, *55*, 475–482.
- 109 Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing Scoring Functions for Protein-Ligand Interactions. *J. Med. Chem.* **2004**, *47*, 3032–3047.
- 110 Ferreira, L. G.; Dos Santos, R. N.; Oliva, G.; Andricopulo, A. D. *Molecular Docking and Structure-Based Drug Design Strategies*; 2015; Vol. 20.
- 111 Bielska, E.; Lucas, X.; Czerwoniec, A.; Kasprzak, J. M.; Kaminska, K. H.; Bujnicki, J. M. Virtual Screening Strategies in Drug Design - Methods and Applications. *Biotechnologia* **2011**, *92*, 249–264.
- 112 Fourches, D.; Politi, R.; Tropsha, A. Target-Specific Native/decoy Pose Classifier Improves the Accuracy of Ligand Ranking in the CSAR 2013 Benchmark. *J. Chem. Inf. Model.* **2015**, *55*, 63–71.

-
- 113 Wang, F.; Cassidy, C.; Sacchettini, J. C. Crystal Structure and Activity Studies of the Mycobacterium Tuberculosis β -Lactamase Reveal Its Critical Role in Resistance to β -Lactam Antibiotics. *Antimicrob. Agents Chemother.* **2006**, *50*, 2762–2771.
- 114 Vercheval, L.; Bauvois, C.; di Paolo, A.; Borel, F.; Ferrer, J.-L.; Sauvage, E.; Matagne, A.; Frère, J.-M.; Charlier, P.; Galleni, M.; Kerff, F. Three Factors That Modulate the Activity of Class D β -Lactamases and Interfere with the Post-Translational Carboxylation of Lys70. *Biochem. J.* **2010**, *432*, 495–504.
- 115 Brown, N. G.; Shanker, S.; Venkataram Prasad, B. V.; Palzkill, T. Structural and Biochemical Evidence That a TEM-1 β -Lactamase N170G Active Site Mutant Acts via Substrate-Assisted Catalysis. *J. Biol. Chem.* **2009**, *284*, 33703–33712.
- 116 Maveyraud, L.; Golemi, D.; Kotra, L. P.; Tranier, S.; Vakulenko, S.; Mobashery, S.; Samama, J. P. Insights into Class D β -Lactamases Are Revealed by the Crystal Structure of the OXA10 Enzyme from Pseudomonas Aeruginosa. *Structure* **2000**, *8*, 1289–1298.
- 117 Tao Sun, Michiyoshi Nukaga, Kayoko Mayama, Emory H. Braswell, J. R. K. Comparison of β -Lactamases of Classes A and D: 1.5-A Crystallographic Structure of the Class D OXA-1 Oxacillinase. *Protein Sci.* **2003**, *12*, 82–91.
- 118 <http://www.rcsb.org/pdb/explore.do?structureId=2GDN>
- 119 <http://www.rcsb.org/pdb/explore.do?structureId=2WKH>
- 120 Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
- 121 Cheng, Feng, V. S. Applications of Artificial Neural Network Modeling in Drug Discovery. *Clin. Exp. Pharmacol.* **2012**, *2*, 2–3.
- 122 Schneider, P.; Tanrikulu, Y.; Schneider, G. Self-Organizing Maps in Drug Discovery: Compound Library Design, Scaffold-Hopping, Repurposing. *Curr. Med. Chem.* **2009**, *16*, 258–266.
- 123 Miljković, D. Brief Review of Self-Organizing Maps. In *MIPRO 2017/CTS; 2017*; pp 1252–1257.
- 124 Zupan, J. Introduction to Artificial Neural Network (ANN) Methods: What They Are and How to Use Them. *Acta Chim. Slov.* **1994**, *41*, 327–352.

-
- 125 Santosh B. Katwal, John C. Gore, René Marois, B. P. R. Unsupervised Spatiotemporal Analysis of fMRI Data Using Graph-Based Visualizations of Self-Organizing Maps. *IEEE Trans Biomed Eng* **2013**, *60*, 2472–2483.
- 126 Gandia-Aguilo V, Cibrian R, Soria E, Serrano AJ, Aguilo L, Paredes V, G. J. Use of Self-Organizing Maps for Analyzing the Behavior of Canines Displaced towards Midline under Interceptive Treatment. *Med. Oral Patol. Oral Cir. Bucal* **2017**, *22*, e233–e241.
- 127 Yao, J. T. Sensitivity Analysis for Data Mining. In *22nd International Conference of the North American Fuzzy Information Processing Society, NAFIPS 2003.*; 2003; pp 272–277.
- 128 Baratloo, A.; Hosseini, M.; Negida, A.; Ashal, G. El; El Ashal, G. Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. *Emergency; 2015 Press* **2015**, *3*, 48–49.
- 129 Kim, S. J.; Cho, K. J.; Oh, S. Development of Machine Learning Models for Diagnosis of Glaucoma. *PLoS One* **2017**, *12*, 1–16.

PART III
DEVELOPMENT AND VALIDATION OF
MOLECULAR DESCRIPTOR

Sajeev R, "Development and validation of molecular descriptor based on physical and biological prediction models" Thesis, Department of Chemistry, Malabar Christian College, Calicut, 2017.

CHAPTER 1

INTRODUCTION

1.1 Molecular Descriptors

Molecular descriptor is defined as "the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment".^{1,2} The term "useful" means that the number can give more insight into the interpretation of the molecular properties and/or further it can be used to build predictive models for some interesting property of molecules.

The field of study is interdisciplinary in the sense that a lot of different theories are involved like algebra, graph theory, information theory, computational chemistry, theories of organic reactivity and physical chemistry are usually required, although at different levels.³ The most important concept of the XX century in the development of the scientific knowledge is the concept of "molecular structure". As it has become the main engine for the development of quantum chemistry, physical chemistry, medicinal chemistry, molecular physics, polymer chemistry etc. A molecule is represented in various forms; 3-dimensional Euclidean representation, 2-dimensional representations based on the graph theory, or fingerprint representations. And their representation has important role in the development of molecular descriptors as each representation constitutes a different conceptual model of the molecule from which different chemical information can be retrieved. A molecule considered as a real object contains all the chemical information, but practically only a part of this information is extracted. As discussed above molecular descriptors are numerical that are used to extract small pieces of chemical information from various forms of molecular representations as depicted in Figure 1.

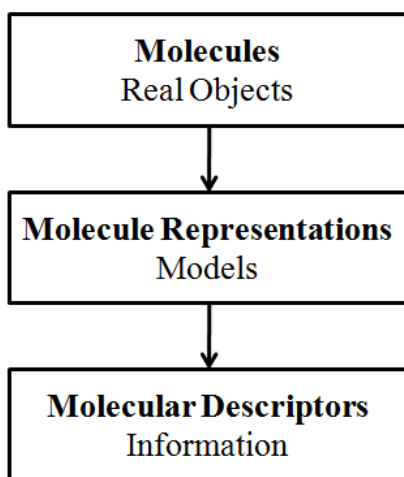


Figure 1. Extraction of chemical information from molecular representation.

A good molecular descriptor should have the following characteristics.⁴

1. It should have a structural interpretation.
2. It should exhibit a good correlation with at least one property.
3. It should be simple.
4. No trivial relation with other molecular descriptors.
5. It should gradually change in its values with gradual changes in the molecular structure.
6. Not including in the definition of experimental properties.
7. Not restricted to a too small class of molecules.
8. Preferably, some discrimination power among isomers.
9. Preferably, allowing reversible decoding (back from the descriptor value to the structure).

1.1.1 Type of molecular descriptors

There are different types of molecular descriptors and the difference is the ways or perspectives a molecule is viewed, taking into account the various features of its chemical structure. They have become one among the most important component used in the various field of molecular modeling, cheminformatics, chemometrics and statistics.

Molecular descriptors are categorized into 0D, 1D, 2D, 3D and 4D.^{4,5,6} The simpler descriptors can be derived from counting some atom types in a molecule. One of the simplest molecular representations is the chemical formula for example benzene molecular formula C_6H_6 contains number of Carbon atoms and Hydrogen atoms to be 6. Here the number of occurrences of the atoms in the molecule is considered. This representation is independent from the kind of molecular structure, and hence, such descriptors obtained from the chemical formula is often comes under the 0D descriptors. Other examples include atomic number, molecular weight etc.

1D descriptor represents a list of substructure of a molecule that includes molecular fragments, functional groups, or substituents of interest etc and thus a complete knowledge of the molecule structure is not necessary. While a 2D representation of a molecule defines the atom connectivity in terms of presence and nature of chemical bonds within the molecule. This approach is usually based on the molecular graph (topology) representation. Here a molecular graph usually consist of a set of vertices V denoting molecular atoms and E is a set of elements representing the binary relationship between pairs of vertices; unordered vertex pairs are called edges, which correspond to bonds between atoms. In this manner two types of molecular graphs are obtained H-depleted molecular graphs (Hydrogens excluded) and H-filled molecular graphs.

The other molecular descriptors are 3D and 4D. 3D molecular descriptors are derived from the spatial (x, y, z) coordinates of the molecule and are usually called as geometrical descriptors. While 4D descriptors are derived from the interaction energies between the molecules imbedded into a grid with some probe molecules like water molecule, methyl group and hydrogen etc. Though the geometrical 3D/4D

descriptors have higher information content than other simpler descriptors, such as counting descriptors or topological descriptors, which often show relevant levels of degeneracy. It is not always that 3D/4D derived descriptors are the “best descriptors”. The “best descriptors” are those whose information content is comparable with the information content of the response for which the model is searched for. In conclusion, it can be stated that the best descriptor(s) valid for all the problems does not exist.⁴

There are many molecular descriptor calculating softwares and web servers available online here are some of them with description.

Table 1. Molecular descriptor generating software and its descriptions.

Software/ Web servers	Descriptions
e-dragon	It is the electronic version of the software DRAGON from VCCLab. It contains more than 1600 molecular descriptors that are divided into 20 logical blocks containing constitutional, simple atom type, functional group descriptors, topological, Lipinski-rule of five, 3D descriptors, enumerative descriptors, charge descriptors etc.
PowerMV	It is a windows based software that can calculate 1000 molecular descriptors involving constitutional, atom pairs, fingerprints, BCUT etc. Other applications include property generation, statistical analysis, molecular visualizing and similarity search.
Chemistry Development Kit (CDK) ⁷	CDK is a java based descriptor calculation tool that calculates constitutional, topological, charge based and geometrical descriptors.
OCHEM	It is a web-based platform widely used for QSAR modeling. The database in OChem is used to develop predictive models, calculate molecular descriptors and many machine learning methods are available.
PreADMET ⁸	It is a web-based descriptor calculator involving drug-likeness prediction, ADME prediction and toxicity prediction.
MODEL ⁹	It is a descriptor generating software that can be used to calculate various topological indices, physical chemistry properties, geometrical molecular descriptors, and quantum chemistry.
PaDEL ¹⁰	It calculates 863 descriptors, (729 1D, 2D descriptors and 134 3D descriptors) and 10 types of fingerprints.

1.1.2 Role of a Molecular Descriptor

A quantitative relationship between structures and properties, biological activities and other experimental properties is established by converting the encoded information present in a molecular structure into useful one or more numbers (molecular descriptors) through a theoretical pathway.¹¹

Molecular descriptors are derived from many theories like graph theory, information theory, quantum theory, organic chemistry etc and are processed by statistics and cheminformatics tools which are later applied into various other field of scientific discipline toxicology, medicinal chemistry, virtual screening etc.¹² To date many thousands of molecular descriptors are generated which is computable by means of various software tools as mentioned in Table 1. Though whole chemical information is not accessible by a molecular descriptor but a small part of the whole system is accounted.¹³ As a result new molecular descriptors are increasing in a large number exhibiting a strong relationship between statistics, data mining, ML methods and cheminformatics. The usage of molecular descriptors has become an important part in the scientific paradigm in understanding the relationship between the molecular structures and physicochemical properties (QSPR) and molecular structure and biological activities (QSAR). For e.g. physicochemical properties like boiling point, melting point, solubility and biological properties like binding affinities, druglikeness, mutagenicity etc.^{14,15,16}

1.2 Scope of Present Investigation

The literature survey made it clear that molecular descriptors have a major role in the modern scientific field and shows how molecular descriptor is directly related to molecular structure, physicochemical property and biological activity. The aim of the study was to connect two entirely different classes of compounds through the era of new knowledge systems. The methodology is virtually screening those molecules which seem to be actives in both biological screen and semiconductor screen. And this study aims to provide a descriptor level and validation for such activity.

CHAPTER 2

MATERIALS AND METHODS

2.1 Methods

For the development and validation of molecular descriptor, we took the molecules that were virtually screened against the semiconductor models (Bayesian, DT, SMO) and β -lactamase (docking, machine learning classifiers, ANN-SOM) methods from *M. tuberculosis* and *P. aeruginosa*. That is the anti-TB molecules from GSK library and Schiff base from ChEBI database. The ML software WEKA was used for the cross screening of Schiff base and GSK molecules against the physical and biological models.

CHAPTER 3

DEVELOPMENT AND VALIDATION OF MOLECULAR DESCRIPTOR

3.1 Development of Molecular Descriptor

For the present study we collected all the developed predictive models against semiconductor and anti-bacterial model from the WEKA software. Here we made use of the twenty four descriptor based predictive models. The eight electronic predictive models that is based on the Bayesian, Decision tree and SOM corresponds to semiconductivity nature while the rest corresponds to the anti-bacterial activity. The input for the development of molecular descriptor is the virtual screened molecules based on the electronic and anti-bacterial models. The virtual screened results obtained from each models (electronic and biological) are mentioned in Tables 2-3. Tables 2-3 shows the number of molecules screened from screening set 1 (Schiff base) and screening set 2 (GSK molecules) against each model and each classifier. A total of fifteen molecules (computationally active semiconductors from both semiconductor machine learning models) and 171 molecules (computationally active GSK molecules which are screened from biological models-Bayesian, random forest, J48, SMO, ANN based SOM and docking based analysis). They were collected for the cross screening process.

Table 2. Number of Schiff base molecules screened against organic semiconductor model

Models	Model 1	Model 2
Naïve Bayes	1	2
Random Forest	7	12
SMO	0	7
J48	6	6

Table 3. Number of GSK molecules screened against anti-bacterial models

Models	Model 1	Model 2 (Oversampled)	Model 3 (SMOTE)	Model 4 (Pseudomonas)
Naïve Bayes	54	102	108	34
Random Forest	17	26	29	55
SMO	1	56	28	48
J48	45	51	42	44

Molecular descriptor was developed by making use of all the electronic models (default and oversampled) and biological models (ML models, docking methods and ANN-SOM method). As a result we got fifteen molecules that are electronic model actives and 171 GSK molecules that are anti-bacterial actives. The screening sets (15 Schiff base and 171 GSK molecules) were cross screened against anti-bacterial models and semiconductor model i.e. 171 anti-TB molecules were screened against all the semiconductor ML models and 15 electronic actives were screened against all the ML biological models (*M. tuberculosis* and *P. aeruginosa*). This task was performed by making use of the descriptor generating softwares. Initially 1664 molecular descriptors were calculated for 171 GSK molecules from the e-dragon web server and were tabulated in csv format. The semiconductor models (default and oversampled models) for each classifier were loaded and screened against the predictive models. Similarly for the Schiff base screening set, 179 biological descriptors were calculated from PowerMV software and screened against the anti-TB models (TB Default, TB Oversampled, TB SMOTE and Pseudomonas default). A brief workflow shown in Figure 2 showcases the various steps carried out to produce the cross screened molecules, these molecules are computationally both semiconducting as well as anti-bacterial. Tables 4-5 describes the total number of molecules screened as a result of cross screening while Tables 6-7 gives the screened molecules list that were computationally predicted both active in electronic and anti-bacterial ML models as a result of cross virtual screening.

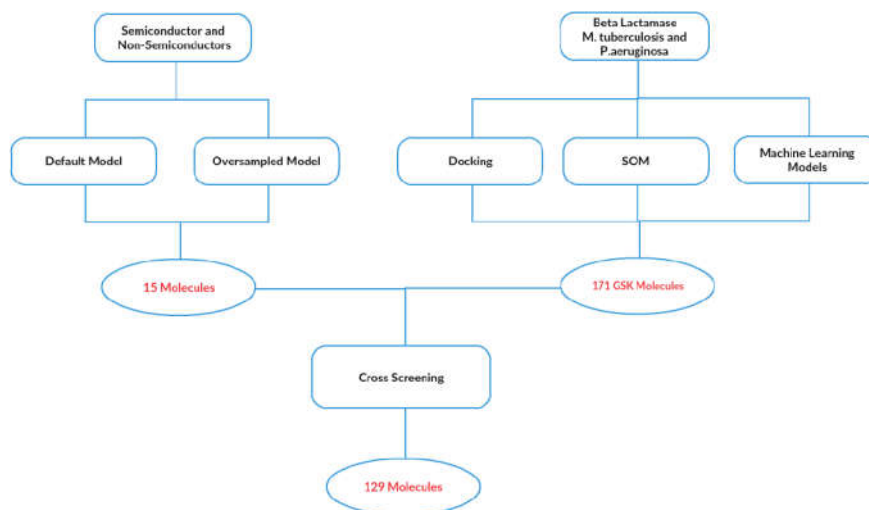


Figure 2. Workflow for the generation of cross screened molecules against electronic and anti-bacterial models

The following Table displays the total number of cross screened results.

Table 4. Number of GSK molecules virtually screened against electronic semiconductor ML models

Classifiers	Model 1	Model 2
Naïve Bayes	40	87
Random Forest	31	52
SMO	7	51
J48	42	62

Table 5. Screening set involving Schiff base molecules screened against ML anti-bacterial biological models

Classifiers	Model 1	Model 2 (Oversampled)	Model 3 (SMOTE)	Model 4 (Pseudomonas)
Naïve Bayes	8	10	6	4
Random Forest	1	5	6	2
SMO	4	7	7	5
J48	2	8	8	6

Table 6. The 15 electronic actives screened against various anti-bacterial models as a part of cross screening.

	Naïve Bayes				Random Forest				SMO				J48			
	TB D*	TB O**	TB S [#]	P ^{##}	TB	TB O	TB S	P	TB D	TB O	TB S	P	TB D	TB O	TB S	P
Schiff base																
Quinoline-4-Carboxylate	active	active	active	active		active	Active	active				active		active		active
Quinaldate	active	active	active			active	Active		active	active	active	active		active	active	active
Phenazine-1-Carboxylate			active	active				active				active		active		
P-Azobenzenesulfonate											active					
Kynurenate15		active	active			active	active			active	active			active	active	Active
Chlordiazepoxide	active	active		active			active									
Aminocyclopyrachlor	active	active	active			active					active			active	active	
7,8-dihydroxykynurenate		active			active					active	active				active	
2-Benzyl-4-Oxidomethylene-5-Oxazolone	active	active							active	active		active	active		active	
7,8-Dihydro-7,8-Dihydroxykynurenate	active	active				active				active	active		active	active	active	Active
Emeraldine	active	active	active				active									
Glucotropeolin														active	active	Active
Rizatriptan	active	active		active			active		active	active		active				Active
Ambenonium																
5-Hydroxyisouric Acid Anion									active	active	active			active	active	

* Default, **Oversampled, [#]SMOTE, ^{##}Pseudomonas

Table 7. GSK 177 anti-TB molecules screened against ML electronic models

S.No.	Molecules	MODEL 2				MODEL 1			
		NB	RF	SMO	J48	N B	RF	SMO	J48
1	GSK1941290A	semiconductor		semiconductor		semiconductor			
2	GSK1826825A								
3	GSK1589673A								
4	GSK847920A	semiconductor	semiconductor	semiconductor	semiconductor				semiconductor
5	GSK275628A	semiconductor		semiconductor			semiconductor		
6	GSK636544A		semiconductor		semiconductor				
7	GSK1434490A								
8	GSK1731114A								
9	GSK345724A								
10	BRL-51093AM								
11	GSK2111534A		semiconductor		semiconductor				
12	GSK937733A								
13	GSK1402290A	semiconductor	semiconductor	semiconductor	semiconductor		semiconductor		semiconductor
14	GSK1588120A	semiconductor	semiconductor		semiconductor	semiconductor	semiconductor		semiconductor
15	GSK130506A								
16	GSK1925843A	semiconductor	semiconductor	semiconductor	semiconductor				
17	GSK991960A								
18	GSK445886A		semiconductor	semiconductor	semiconductor		semiconductor	semiconductor	semiconductor
19	GSK1302651A		semiconductor		semiconductor				semiconductor
20	GSK937213A	semiconductor		semiconductor		semiconductor	semiconductor		
21	GSK270670A								
22	SB-829405	semiconductor				semiconductor			
23	GI247341A								
24	GSK1829660A	semiconductor			semiconductor				
25	GSK547543A	semiconductor		semiconductor		semiconductor	semiconductor		
26	GI103688B								
27	GSK1729177A								
28	GSK1758774A								
29	GSK1650514A	semiconductor		semiconductor	semiconductor	semiconductor			
30	GSK1812410A	semiconductor	semiconductor		semiconductor		semiconductor		semiconductor
31	GSK831784A	semiconductor							
32	GSK1857145A	semiconductor							semiconductor
33	GSK153890A	semiconductor			semiconductor	semiconductor	semiconductor		
34	GSK848336A		semiconductor	semiconductor	semiconductor		semiconductor		semiconductor
35	GSK1051703A								
36	GSK437009A	semiconductor	semiconductor		semiconductor		semiconductor		
37	GSK276001A	semiconductor							
38	GSK1826089A	semiconductor				semiconductor			
39	GSK479031A	semiconductor	semiconductor		semiconductor				semiconductor
40	GSK1055950A								
41	GSK1829816A		semiconductor						
42	GSK1905227A	semiconductor	semiconductor		semiconductor		semiconductor		semiconductor
43	SB-811796-V	semiconductor				semiconductor			
44	GSK888636A	semiconductor							
45	GSK1863309A	semiconductor			semiconductor				semiconductor
46	GSK316438A		semiconductor	semiconductor	semiconductor			semiconductor	

			MODEL 2				MODEL 1		
		NB	R F	SMO	J48	N B	R F	SMO	J48
S.No.	Molecules								
47	SB-706404		semiconductor		semiconductor				
48	GSK547511A	semiconductor		semiconductor		semiconductor	semiconductor		
49	GSK1691553A	semiconductor				semiconductor	semiconductor		
50	GSK146660A								
51	GSK468214A			semiconductor					
52	GSK1733953A								
53	GSK690382A	semiconductor	semiconductor	semiconductor	semiconductor		semiconductor	semiconductor	
54	SB-552112	semiconductor			semiconductor				semiconductor
55	GSK1829736A								
56	GSK2157753A		semiconductor						
57	GSK1832831A	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor			
58	GSK347301A								
59	GSK1385423A								
60	GSK1372568A	semiconductor	semiconductor	semiconductor	semiconductor				semiconductor
61	GSK1788487A	semiconductor				semiconductor			
62	GSK920703A	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor			semiconductor
63	GSK1107112A				semiconductor				
64	GSK1598164A	semiconductor	semiconductor		semiconductor	semiconductor	semiconductor		
65	GSK921295A	semiconductor	semiconductor	semiconductor	semiconductor				semiconductor
66	GSK1829729A								
67	GW664700A	semiconductor		semiconductor			semiconductor		
68	GW369335X	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor		semiconductor
69	GSK889423A	semiconductor			semiconductor				
70	GSK1668869A								
71	GSK547487A	semiconductor		semiconductor		semiconductor			
72	GSK892651A	semiconductor	semiconductor		semiconductor	semiconductor			semiconductor
73	GSK275984A	semiconductor	semiconductor		semiconductor				
74	GSK352635A		semiconductor	semiconductor			semiconductor		
75	SB-746177	semiconductor	semiconductor		semiconductor				semiconductor
76	GSK1518999A	semiconductor							
77	GSK2200160A	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor			semiconductor
78	GSK1570606A								
79	GV187303X	semiconductor	semiconductor		semiconductor				semiconductor
80	SB-354364								semiconductor
81	GSK1635139A	semiconductor			semiconductor	semiconductor			semiconductor
82	GSK1310678A								
83	GSK1611550A								
84	GSK1174628A			semiconductor					
85	GSK133167A	semiconductor		semiconductor		semiconductor			
86	GW623128X	semiconductor	semiconductor		semiconductor	semiconductor			semiconductor
87	GSK2200150A								semiconductor
88	GSK1759150A	semiconductor							semiconductor
89	GSK1829676A								
90	GSK1829732A								
91	GSK463114A	semiconductor							
92	GSK1783710A								
93	SB-204804-A	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor		semiconductor
94	GSK1180781A	semiconductor							

			MODEL 2				MODEL 1		
S.No.	Molecules	NB	R F	SMO	J48	N B	R F	SMO	J48
95	GSK1220329A								
96	GSK1829819A	semiconductor							
97	GSK1121877A	semiconductor		semiconductor		semiconductor	semiconductor		
98	GW339742X								
99	GSK1955236A	semiconductor	semiconductor	semiconductor	semiconductor		semiconductor		semiconductor
100	GSK353069A		semiconductor	semiconductor	semiconductor				semiconductor
101	GSK810016A	semiconductor		semiconductor		semiconductor	semiconductor		
102	GW356807A								
103	GSK731389A			semiconductor	semiconductor				
104	GW876411A	semiconductor	semiconductor		semiconductor	semiconductor	semiconductor		
105	GSK921190A	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor			semiconductor
106	GW859039X			semiconductor					
107	GSK1519001A	semiconductor							
108	GSK2059310A	semiconductor				semiconductor			
109	SB-435634				semiconductor				semiconductor
110	GSK1744926A	semiconductor	semiconductor	semiconductor	semiconductor				semiconductor
111	GSK1829820A	semiconductor	semiconductor		semiconductor	semiconductor			
112	GSK1750922A								
113	GSK957094A		semiconductor		semiconductor				semiconductor
114	GSK1829733A								
115	GSK695914A	semiconductor							semiconductor
116	GR135487X								
117	GSK735816A	semiconductor	semiconductor	semiconductor		semiconductor	semiconductor	semiconductor	
118	BRL-7940SA								
119	GSK124576A	semiconductor			semiconductor	semiconductor			
120	BRL-8088SA								
121	BRL-10988SA								
122	GR135486X						semiconductor		semiconductor
123	CCI7967	semiconductor			semiconductor				
124	GSK254610A	semiconductor				semiconductor			
125	GSK810037A	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor		semiconductor	semiconductor
126	GSK426032A								
127	GR223839X	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor			semiconductor
128	GSK735826A			semiconductor				semiconductor	
129	GSK861337A								
130	GSK829969A	semiconductor		semiconductor	semiconductor				
131	GSK124945A						semiconductor		
132	GSK1859936A								
133	GSK163574A	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor		
134	GSK1742694A	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor			
135	GSK498315A		semiconductor						
136	GSK1996236A								
137	GW360240X	semiconductor							
138	SB-650816	semiconductor							
139	GSK1829674A								
140	GSK762874A	semiconductor							
141	BRL-10143SA								
142	GSK353071A			semiconductor					

S.No.	Molecules	MODEL 2				MODEL 1			
		NB	R F	SMO	J48	N B	R F	SMO	J48
143	GW857165X	semiconductor			semiconductor				
144	GSK749336A	semiconductor							
145	GSK1829671A								
146	GSK847913A								
147	GSK994258A	semiconductor			semiconductor				semiconductor
148	GSK353496A		semiconductor	semiconductor	semiconductor			semiconductor	
149	GSK1829728A								
150	GSK547481A	semiconductor		semiconductor		semiconductor			
151	GSK1365028A								semiconductor
152	GSK920684A			semiconductor					
153	BRL-51091AM								
154	GSK2200157A	semiconductor	semiconductor		semiconductor				semiconductor
155	GW861072X								
156	GSK2043267A								
157	SB-516933								
158	GSK385518A		semiconductor		semiconductor				
159	GSK798463A								
160	GSK1072678A	semiconductor	semiconductor	semiconductor	semiconductor				semiconductor
161	BRL-8903SA								
162	GSK1829727A								
163	GSK237561A	semiconductor					semiconductor		
164	GSK262906A								
165	GR153167X			semiconductor					
166	GSK1329419A	semiconductor							
167	GSK2032710A	semiconductor		semiconductor		semiconductor	semiconductor		
168	GSK1985270A								
169	GSK1589671A	semiconductor							
170	GSK381407A	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor	semiconductor		semiconductor
171	SB-712970								
172	GW713556X								
173	GSK1826247A	semiconductor							
174	SB-811137-V	semiconductor	semiconductor			semiconductor			
175	GSK754716A			semiconductor			semiconductor		semiconductor
176	GSK705278A								
177	GSK358607A	semiconductor	semiconductor	semiconductor	semiconductor				

Among GSK 177 anti-TB molecules, 171 were found to be active against all the biological ML models were cross screened against the electronic ML models. As a result we got 114 electronic actives.

3.2 Results and discussion

The cross screening ended up with a total 129 molecules to be computationally active (molecules filtered through both electronic and biological models) and the rest 98 as inactive molecules. As a result of cross virtual screening the 15 semiconductor actives were also predicted to be biologically active and among 171 anti-TB molecules, 114 filtered through the semiconductor electronic model. The 129 molecules are computationally both semiconducting and anti-bacterial actives. In order to find the key descriptor among the two sets of class we worked on the Pi bond lone pair conjugation (semiconducting nature).

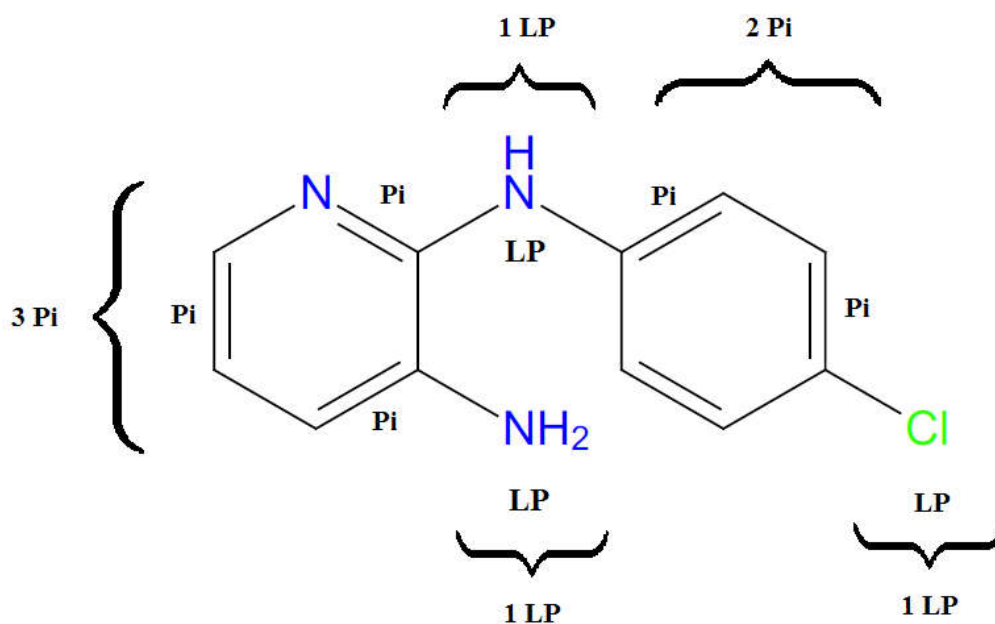


Figure 3. GSK molecule GR135486X lone pair pi walk count 8 is displayed. The new descriptor is read as; **lp3πlp2πlp** from NH₂ through 3 Pi bonds, Lone pair, 2pi and Lone pair having a total count 8.

The patterns were developed based on the lone pair pi bond walk count beginning from the smallest path to the longest directed path. The pattern shown in Figure 3 is a lone pair pi walk with count 8. Table 8-9 shows the longest lone pair pi conjugated walk for active and inactive molecules.

Table 8. Lone pair pi walk count path for 129 active molecules

S.No.	Molecules	Lone pair pi walk	Total No. of Counts
1	CCI7967	lp2πlp1π 3πlp1π	5
2	GR135486X	lp3πlp2πlp lp3πlp3π lp2πlp3πlp 3πlp3πlp	8
3	GR153167X	3πlp2πlp lp2πlp3π	7
4	GR223839X	lp1πlp5π 5πlp1πlp	8
5	GSK124576A	lp4πlp lp5π	6
6	GSK124945A	lp5πlp	7
7	GSK133167A	lp7π	8
8	GSK153890A	3πlp1πlp3π	9
9	GSK163574A	2πlp2πlp3π	9
10	GSK237561A	1πlp3πlp1π 5πlp1π	7
11	GSK254610A	lp5π	6
12	GSK275628A	6π lp5π	6
13	GSK275984A	6π	6
14	GSK276001A	6π 5πlp	6
15	GSK316438A	1πlp4πlp	7
16	GSK352635A	lp1πlp2π 2πlp2π 2πlp1πlp	5
17	GSK353069A	lp5π	6
18	GSK353071A	lp5π 5πlp	6
19	GSK353496A	lp3πlp1π	6
20	GSK358607A	lp4πlp lp3πlp1π	6
21	GSK381407A	9π	9
22	GSK385518A	1πlp5π	7
23	GSK437009A	2πlp3π 6π	6
24	GSK445886A	lp2πlp2πlp	7

S.No.	Molecules	Lone pair pi walk	Total No. of Counts
25	GSK463114A	lp2πlp3π lp5π	6
26	GSK468214A	lp4πlp lp2πlp1π	5
27	GSK479031A	lp2πlp1π 5π	5
28	GSK498315A	lp4π 2πlp3π	6
29	GSK547511A	3πlp2π	7
30	GSK547543A	3πlp2πlp 5πlp3π	9
31	GSK636544A	3πlp5π 2πlp1π	4
32	GSK690382A	4πlp2πlp	8
33	GSK695914A	lp2πlp4π lp6π	7
34	GSK731389A	lp3πlp1π	6
35	GSK735816A	lp3πlp2πlp 4πlp2πlp	8
36	GSK735826A	lp2πlp4πlp	9
37	GSK749336A	lp4πlp	6
38	GSK754716A	lp5π 3πlp4π 3πlp3πlp lp3πlp3π	8
39	GSK762874A	3πlp2π	6
40	GSK810016A	lp3πlp3πlp1π	10
41	GSK810037A	lp1πlp3πlp1π	8
42	GSK829969A	3πlp1π 4πlp 1πlp2πlp	5
43	GSK831784A	lp4πlp 6π 5πlp lp4πlp	6
44	GSK847920A	2πlp3πlp1π	8
45	GSK848336A	6π	6
46	GSK888636A	6πlp1π	8
47	GSK889423A	lp3πlp	5
48	GSK892651A	2πlp3π	6
49	GSK920684A	4πlp1π	6

S.No.	Molecules	Lone pair pi walk	Total No. of Counts
50	GSK920703A	4π1p1π	6
51	GSK921190A	4π1p1π	6
52	GSK921295A	4π1p1π	6
53	GSK937213A	1p7π	8
54	GSK957094A	3π1p2π	6
55	GSK1072678A	4π1p1π	6
56	GSK1107112A	1p3π1p 4π1p 5π	5
57	GSK1121877A	6π 1p5π	6
58	GSK1174628A	1π1p1π1p2π	6
59	GSK1180781A	1p2π1p	4
60	GSK1302651A	1p5π	6
61	GSK1329419A	1p2π1p3π 6π1p	7
62	GSK1365028A	1π1p4π 1π1p3π1p	6
63	GSK1372568A	4π1p1π	6
64	GSK1402290A	6π1p1π	8
65	GSK1518999A	2π1p2π1p3π	9
66	GSK1519001A	2π1p2π1p3π	9
67	GSK1588120A	8π	8
68	GSK1589671A	1p2π1p	4
69	GSK1598164A	1p7π	8
70	GSK1635139A	1p3π1p 5π 4π1p 1π1p3π	5
71	GSK1650514A	1p5π 1p4π1p	6
72	GSK1691553A	3π1p3π	7
73	GSK1742694A	1p2π1p3π	7
74	GSK1744926A	1p4π	5
75	GSK1759150A	1p4π1p	6
76	GSK1788487A	1p4π1p1π	7
77	GSK1812410A	7π	7
78	GSK1826089A	1p4π1p3π	9
79	GSK1826247A	3π1p5π	9
80	GSK1829660A	2π1p3π	6
81	GSK1829816A	3π1p2π	6

S.No.	Molecules	Lone pair pi walk	Total No. of Counts
82	GSK1829819A	$2\pi 1p3\pi$	6
83	GSK1829820A	$2\pi 1p3\pi$	6
84	GSK1832831A	6π	6
85	GSK1857145A	$1p3\pi 1p3\pi$	8
86	GSK1863309A	$2\pi 1p4\pi$	7
87	GSK1905227A	$3\pi 1p4\pi$	8
88	GSK1925843A	$5\pi 1p$	6
89	GSK1955236A	$6\pi 1p1\pi$	8
90	GSK2032710A	$2\pi 1p2\pi 1p$	6
91	GSK2059310A	$1p5\pi$	6
92	GSK2111534A	$3\pi 1p2\pi$	6
93	GSK2157753A	$1\pi 2 1p1\pi$ $3\pi 1p$	4
94	GSK2200150A	$1p3\pi$	4
95	GSK2200157A	$1p2\pi$ $1\pi 1p1\pi$	3
96	GSK2200160A	$1p2\pi$ $2\pi 1p$	3
97	GV187303X	$1p3\pi 1p3\pi$	8
98	GW360240X	$1p5\pi$	6
99	GW369335X	$5\pi 1p4\pi$	10
100	GW623128X	$4\pi 1p$ 5π $1p2\pi 1p1\pi$	5
101	GW664700A	$1p5\pi 1p1\pi$	8
102	GW857165X	$1p4\pi 1p$ $1p5\pi$	6
103	GW859039X	$5\pi 1p$	6
104	GW876411A	$3\pi 1p4\pi$ $6\pi 1p1\pi$	8
105	SB-204804-A	$3\pi 1p2\pi 1p1\pi$	8
106	SB-354364	$1p3\pi 1p$	5
107	SB-435634	$1p2\pi 1p3\pi 1p$	8
108	SB-552112	5π	5
109	SB-650816	5π $4\pi 1p$	5
110	SB-706404	$4\pi 1p$	5
111	SB-746177	$5\pi 1p1\pi$	7
112	SB-811137-V	4π $1p3\pi$	4

S.No.	Molecules	Lone pair pi walk	Total No. of Counts
113	SB-811796-V	4 π 1 π 1 π 3 π	5
114	SB-829405	7 π 6 π 1 π	7
115	Am benonium	1 π 3 π	4
116	5-Hydroxyisouric Acid Anion	1 π 3 π	4
117	Quinoline-4-Carboxylate	6 π	6
118	Quinaldate	6 π	6
119	Phenazine-1-Carboxylate	7 π	7
120	P-Azobenzenesulfonate	7 π	7
121	Kynurenate	5 π 1 π	6
122	Chlordiazepoxide	1 π 3 π 1 π	5
123	Aminocyclopyrachlor	4 π 1 π	5
124	7,8-dihydroxykynurenate	5 π 1 π 1 π 5 π 1 π	7
125	2-Benzyl-4-Oxidomethylene-5-Oxazolone	1 π 1 π 2 π 1 π	5
126	7,8-Dihydro-7,8-Dihydroxykynurenate	4 π	4
127	Emeraldine	1 π 2 π 1 π 8 π	12
128	Glucotropeolin	1 π 1 π 1 π 1 π	4
129	Rizatriptan	1 π 1 π 3 π 4 π 1 π	5

Table 9. Lone pair pi walk count path for 98 inactive molecules

S. No	Molecules	Lone pair pi walk	Total No. of Counts
1	BRL-7940SA	lp3 π	4
3	BRL-8903SA	lp3 π	4
4	BRL-10143SA	lp3 π	4
5	BRL-10988SA	lp3 π	4
6	BRL-51091AM	lp2 π lp	4
7	BRL-51093AM	lp3 π	4
8	GSK426032A	4 π	4
		lp3 π	
		lp2lp1 π	
		lp2 π lp	
		lp3 π	
9	GSK798463A	lp1 π lp1 π	4
		3 π lp	
10	GSK937733A	lp2 π lp	4
		lp3 π	
11	GSK1589673A	lp2 π lp	4
		lp3 π	
12	GSK1731114A	lp3 π	4
13	GSK1758774A	lp3 π	4
14	GSK1985270A	lp3 π	4
15	SB-712970	lp3 π	4
16	Pyridoxamine 5'-Phosphate	lp3 π	4
		3 π lp	
17	CDP-N-Methylethanolamine	3 π lp	4
18	CDP-Choline	3 π lp	4
19	Carbinoxamine	lp3 π	4
20	6-Hydroxynicotinate	3 π lp	4
		lp3 π	
21	5-Methyldeoxycytidine 5'-Diphosphate	lp1 π lp1 π	4
22	5-Hydroxyimidazole-4-Acetate	lp2 π lp	4
23	5-Hydroxy-6-Methylpyridine-3-Carboxylate	4 π	4
24	3',5'-Cyclic CMP	3 π lp	4
		lp3 π	

S. No	Molecules	Lone pair pi walk	Total No. of Counts
25	BRL-8088SA	5 π	5
26	GI103688B	1 π 1 π 1 π 1 π	5
27	GSK347301A	5 π	5
28	GSK861337A	5 π	5
29	GSK994258A	1 ρ 3 π 1 ρ 1 ρ 4 π	5
30	GSK1310678A	1 ρ 4 π	5
31	GW713556X	1 ρ 4 π	5
32	GW861072X	1 ρ 4 π	5
33	SB-516933	5 π	5
34	Photinus Luciferin	1 ρ 4 π 5 π	5
35	Benzylpenicillenate	1 π 1 ρ 2 π 1 ρ	5
36	7,8-Dihydropteroate	1 π 1 ρ 3 π	5
37	2,6-Dihydroxynicotinate	1 ρ 4 π	5
38	GI247341A	1 ρ 5 π	6
39	GSK262906A	1 ρ 5 π	6
40	GSK345724A	1 ρ 5 π	6
41	GSK705278A	2 π 1 ρ 2 π 1 ρ 2 π 1 ρ 3 π	6
42	GSK847913A	6 π	6
43	GSK1055950A	1 ρ 3 π 1 ρ 1 π	6
44	GSK1570606A	4 π 1 ρ 1 π	6
45	GSK1611550A	4 π 1 ρ 1 π	6
46	GSK1668869A	6 π 5 π 1 ρ	6
47	GSK1729177A	1 ρ 3 π 1 ρ 1 π	6
48	GSK1750922A	1 ρ 3 π 1 ρ 1 π	6
49	GSK1829671A	2 π 1 ρ 3 π	6
50	GSK1829674A	2 π 1 ρ 3 π	6
51	GSK1829727A	2 π 1 ρ 3 π	6
52	GSK1829728A	2 π 1 ρ 3 π	6
53	GSK1829729A	2 π 1 ρ 3 π	6
54	GSK1829732A	2 π 1 ρ 3 π	6
55	GSK1829733A	2 π 1 ρ 3 π	6
56	GSK1829736A	2 π 1 ρ 3 π	6

S. No	Molecules	Lone pair pi walk	Total No. of Counts
57	GSK1859936A	5 π lp	6
58	GSK1996236A	lp2 π lp2 π	6
59	GSK2043267A	lp4 π lp 4 π lp1 π	6
60	GW356807A	1 π lp2 π lp1 π	6
61	S-Adenosyl-4-Methylthio-2-Oxobutanoate	lp4 π lp	6
62	GDP-Alpha -D-Mannose	1 π lp3 π lp	6
63	Futalosinate	1 π lp3 π lp	6
64	Aminodeoxyfutalosinate	lp4 π lp	6
65	Adenosine 5'-Phosphoramidate	lp4 π lp	6
66	Adenin-9-yl Riburonosate	lp4 π lp	6
67	8-Bromo-3',5'-Cyclic GMP	lp3 π lp1 π	6
68	5'-Acylphosphoadenosine	lp4 π lp	6
69	GSK130506A	lp4 π lp1 π	7
70	GSK991960A	1 π lp4 π lp	7
71	GSK1220329A	1 π lp2 π lp2 π	7
72	GSK1434490A	lp4 π lp1 π	7
73	GSK1733953A	lp6 π lp5 π lp	7
74	GSK1826825A	3 π lp3 π lp2 π lp3 π	7
75	GSK1829676A	3 π lp3 π	7
76	GSK1941290A	6 π lp	7
77	GW339742X	1 π lp3 π lp1 π	7
78	Nalidixic Acid Anion	4 π lp2 π	7
79	M-Azobenzenesulfonate	7 π	7
80	GR135487X	lp3 π lp3 π 3 π lp3 π lp	8
81	GSK270670A	3 π lp2 π lp1 π	8
82	GSK1783710A	3 π lp2 π lp1 π	8
83	GSK146660A	lp4 π lp3 π 3 π lp4 π lp	9
84	GSK547481A	lp4 π lp3 π	9
85	GSK547487A	5 π lp3 π	9
86	GSK1051703A	lp2 π lp3 π lp1 π	9

S. No	Molecules	Lone pair pi walk	Total No. of Counts
87	GSK1385423A	lp2πlp3πlp1π	9
88	Xanthommatin	7πlp2π 5πlp4π 2πlp7π	10
89	Phenylthioacetohydroximate	3π lp1πlp	3
90	N-Acetyl-L-Histidinate	2πlp	3
91	L-Histidinol Phosphate	1πlp1π	3
92	Imidazol-4-ylacetate	1πlp1π	3
93	3-(Imidazol-5-yl) Pyruvate	lp2π	3
94	(S)-3-(Imidazol-5-yl)Lactate	2πlp	3
95	5-Hydroxy-2-Oxo-4-Ureido-2,5-Dihydro-1H-Imidazole-5-Carboxylate	lp2π	3
96	3-(4-oxo-4,5-dihydro-1H-imidazol-5-yl)propanoic	lp2π	3
97	4-Hydroxy-1-Pyrroline-2-Carboxylate	2π	2
98	3,4-Dehydrothiomorpholine-3-Carboxylate	2π	2

The pie chart displayed in Figure 4 shows the lone pair pi walk patterns in terms of number of actives and number of inactive. Figure 4 also displays the lone pair walk pattern count from the order count 2 to 12 where the number of actives was steadily increasing in respect to inactive. Thus it resembles that the lone pair pi conjugation present in the molecules. At the lone pair pi walk count 8 number of actives is maximum thereafter a decrease and increase due to the presence of lesser number of active.

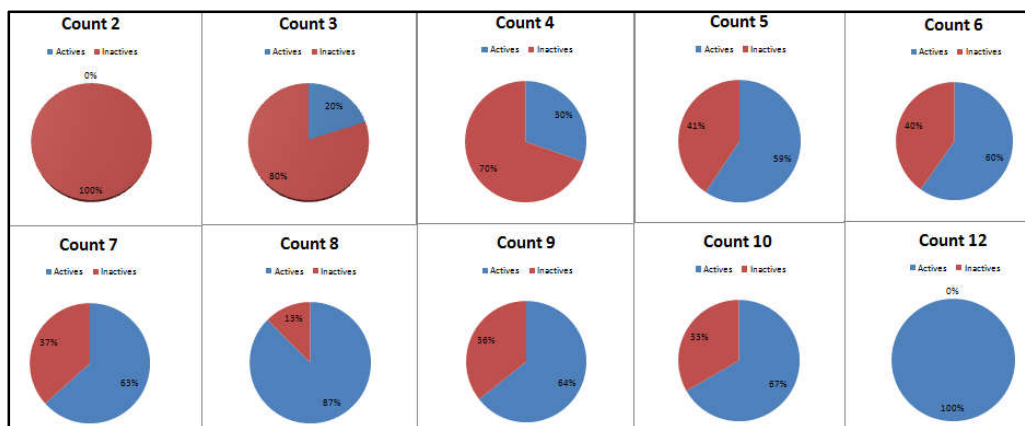


Figure 4. Lone pair pi walk patterns in terms of number of actives and number of inactive

We differentiated between actives and inactive based on the lone pair pi conjugated system with maximum walk count. During this process the 2D walk descriptor was exhibited by both actives and inactive until at walk count 8, we were able to distinguish between active class and inactive class. As a result we could observe the new descriptor namely lone pair pi walk count 8 (see Figure 3). The new descriptor is a graph albeit with a simple vocabulary like atom and bond symbols and all the long possible paths between every atom are taken into account. It is a forward direct 2D graph based descriptor involving longest lone pair and conjugated pi bonds in a molecular representation. It is calculated by counting the number of lone pairs and pi bonds in a molecule (usually starting from a lone pair of electrons/pi electrons). The pattern (**lp3πlp2πlp**) shown in Figure 3 is from the molecule GR135486X which is read in a way from NH₂ through 3 pi bonds, lone pair, 2 pi and lone pair having a total count 8. Further in order to avoid

complications the connectivity of each atom and bond is recorded only once in the forward direction. The patterns in actives and inactives were indistinguishable with lesser walk count as they possessed the lone pair pi conjugation path in common. The walk count 8 is displayed for both active and inactive as mentioned in Table 10. Among the active descriptors the walk count, the descriptors $lp3\pi lp3\pi$ were present in very few molecules GR135486X, GSK754716A, GSK1857145A, GV187303 and $3\pi lp2\pi lp1\pi$ for SB-204804-A.

Table 10. Molecules consisted with lone pair pi walk count 8

Molecules	Actives with lone pair pi walk count 8	Molecules	Inactives with lone pair pi walk count 8
GR135486X	$lp3\pi lp2\pi lp$ $lp3\pi lp3\pi$ $lp2\pi lp3\pi lp$	GR135487X	$lp3\pi lp3\pi$
GR223839X	$lp1\pi lp5\pi$ $5\pi lp1\pi lp$	GSK270670A	$3\pi lp2\pi lp1\pi$
GSK133167A	$lp7\pi$	GSK1783710A	$3\pi lp2\pi lp1\pi$
GSK690382A	$4\pi lp2\pi lp$		
GSK735816A	$lp3\pi lp2\pi lp$		
GSK754716A	$3\pi lp4\pi$ $3\pi lp3\pi lp$ $lp3\pi lp3\pi$		
GSK810037A	$lp1\pi lp3\pi lp1\pi$		
GSK847920A	$2\pi lp3\pi lp1\pi$		
GSK888636A	$6\pi lp1\pi$		
GSK937213A	$lp7\pi$		
GSK1402290A	$6\pi lp1\pi$		
GSK1588120A	8π		
GSK1598164A	$lp7\pi$		
GSK1857145A	$lp3\pi lp3\pi$		
GSK1905227A	$3\pi lp4\pi$		
GSK1955236A	$6\pi lp1\pi$		
GV187303X	$lp3\pi lp3\pi$		
GW664700A	$lp5\pi lp1\pi$		
GW876411A	$3\pi lp4\pi$ $6\pi lp1\pi$		
SB-204804-A	$3\pi lp2\pi lp1\pi$		
SB-435634	$lp2\pi lp3\pi lp$		

In our study we developed a 2D walk descriptor based on lone pair pi bond conjugated system. The descriptor was developed from all of the physical and biological models based data mining models, docking study and ANN-SOM method.

3.3 Validation of molecular descriptor

We developed the molecular descriptor lone pair pi walk count 8 that was calculated manually. The validation was carried out by incorporating the newly developed descriptor into the existing models used against semiconductor and anti-TB activity. For this purpose we considered the default models for semiconductor (Model 1), TB default model and pseudomonas default model against each classifier (Bayesian, random forest, J48 and SMO). The ML model was build against each classifier and the test set was re-evaluated upon the newly built ML model and statistical performance of each model is presented in Tables 11-16. From the data it is clear that most of the statistical performance had improved in accordance to the original models. The higher values may be explained on the basis of the induction of the lone pair pi bond descriptor, since the training sets are improved on addition of the descriptor. The increased statistical parameters enhance the test set ability in screening out the test molecules as depicted from Figures 5-19.

Table 11. TP, TN, FP, FN comparison of Model 1 Default Organic Semiconductor model 1 with added new descriptor (lone pair pi walk count 8 descriptor)

Classifier	TP	TP New	TN	TN New	FP	FP New	FN	FN New
Naive Bayes	8	9	8	8	0	0	3	2
Random Forest	11	11	7	7	1	1	0	0
SMO	8	9	7	7	1	1	3	2
J48	10	11	7	7	1	0	1	1

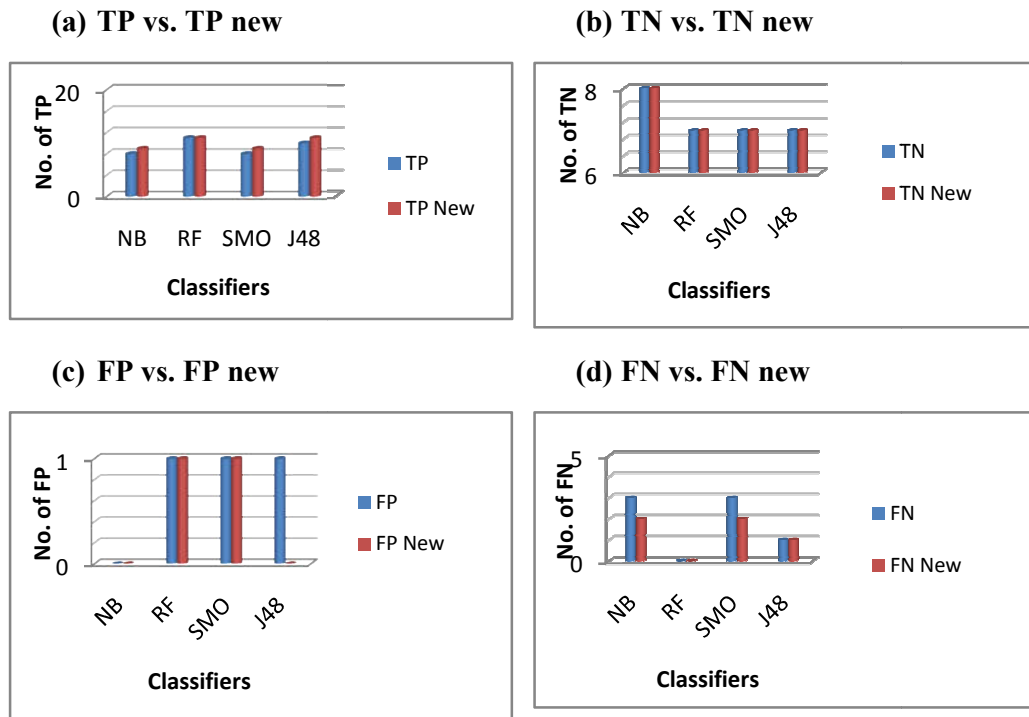


Figure 5. The comparative analysis of the number of TP, TN, FP and FN were studied vs. newly added descriptor. (a) Comparison of TP vs. New TP for each classifier is mentioned (b) comparison of TN vs. New TN for each classifier is mentioned (c) comparison of FP vs. New FP (d) comparison of FN vs. New FN for each classifier is mentioned. We can figure out that the number of TP in the model added with new descriptor has increased slightly for Bayesian, J48 and SMO model but remained same for random forest. Also number of TN remained same for all the classifier and in the new FP showcased a lesser value for J48 model.

Table 12. Statistical parameters of ML classifiers from semiconductor models vs. newly added descriptor

Statistical Parameters	NB	NB New	RF	RF New	SMO	SMO New	J48	J48 New
Tp rate %	72.7	81.81	100	100	72.7	81.81	90.9	91.66
Fp rate %	0	0	12.5	12.5	12.5	12.5	12.5	0
Precision	100	100	91.7	91.7	88.9	90	90.9	100
Recall	72.7	81.81	100	100	72.7	81.81	90.9	91.66
Specificity	100	100	87.5	87.5	87.5	87.5	87.5	100
BAC	86.35	90.9	93.75	93.75	80.1	84.655	89.2	95.83
F-measure	84.2	89.99	95.7	95.7	80	85.7	90.9	95.64
ROC	87.5	91.356	90.3	90.3	95.5	97.1	88.1	92.7
Accuracy	84.2105	89.47	94.7368	94.7368	78.9474	84.21	89.4737	94.73
Kappa	0.69	0.79	0.8902	0.8902	0.5824	0.679	0.7841	0.889
MCC	0.72	0.8	0.8955	0.8955	0.5955	0.6854	0.784	0.8955

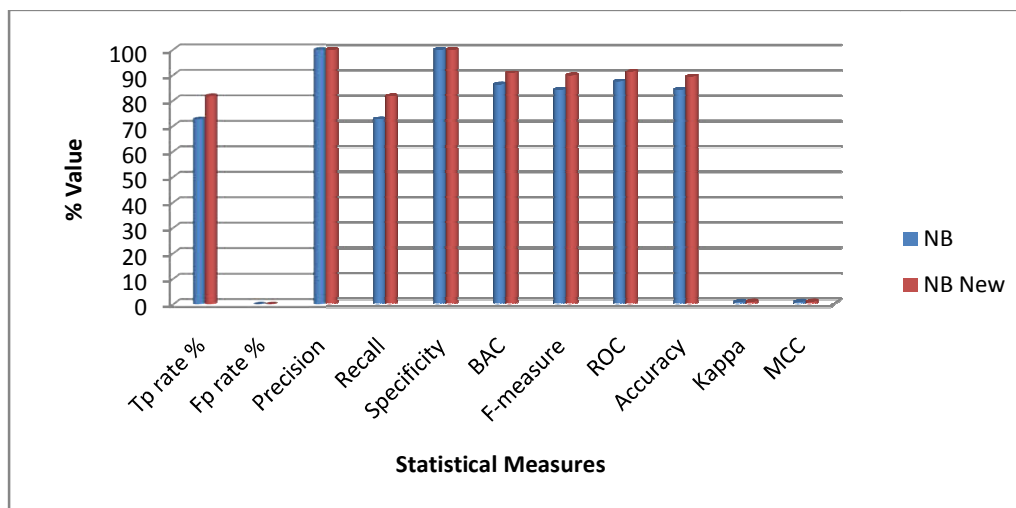


Figure 6. Comparison of Naïve Bayes Default Model 1 semiconductor vs. newly added descriptor Naïve Bayes statistical parameters

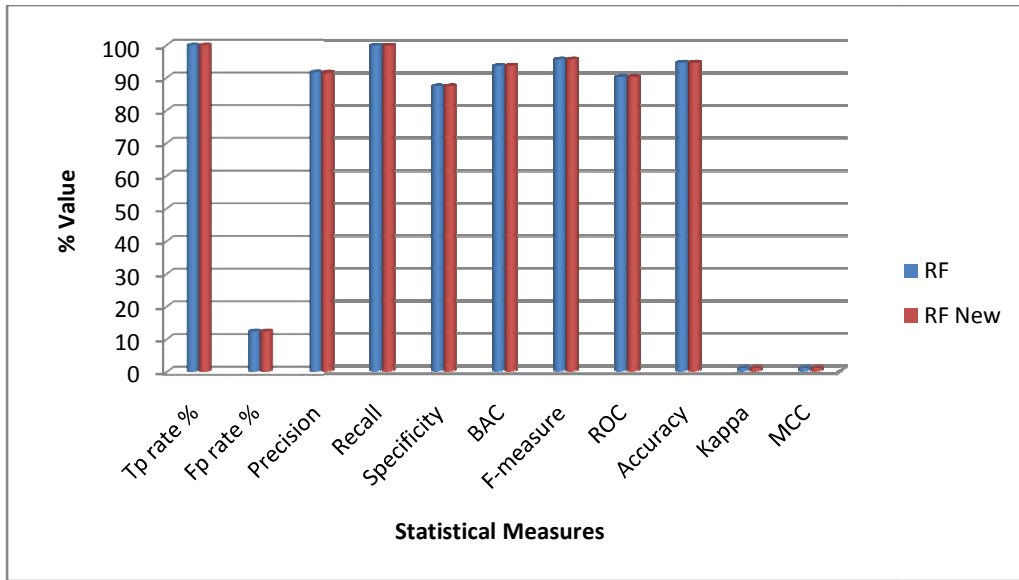


Figure 7. Comparison of Random Forest Default Model 1 semiconductor vs. newly added descriptor Random forest statistical parameters

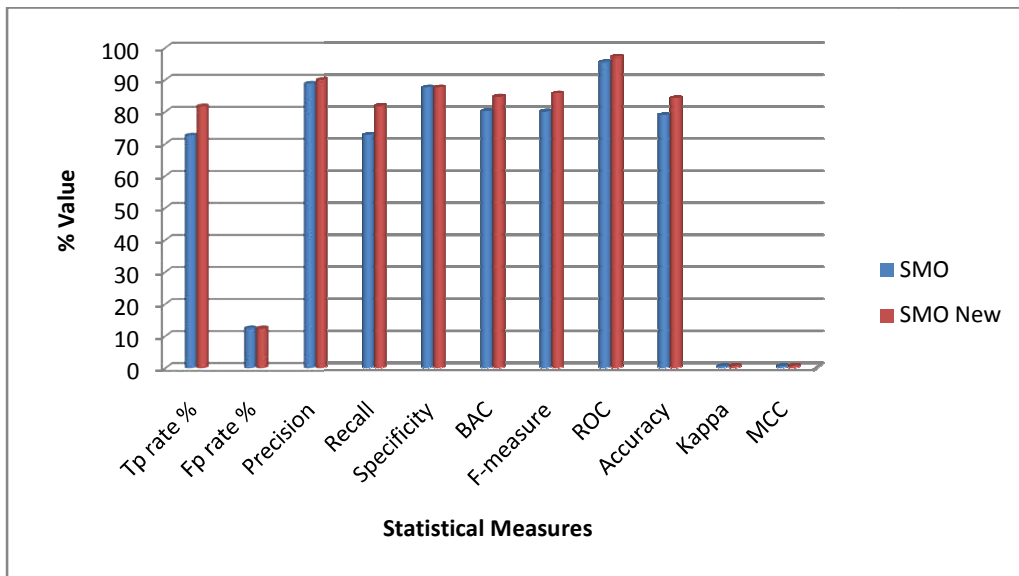


Figure 8. Comparison of SMO Default Model 1 semiconductor vs. newly added descriptor SMO statistical parameters

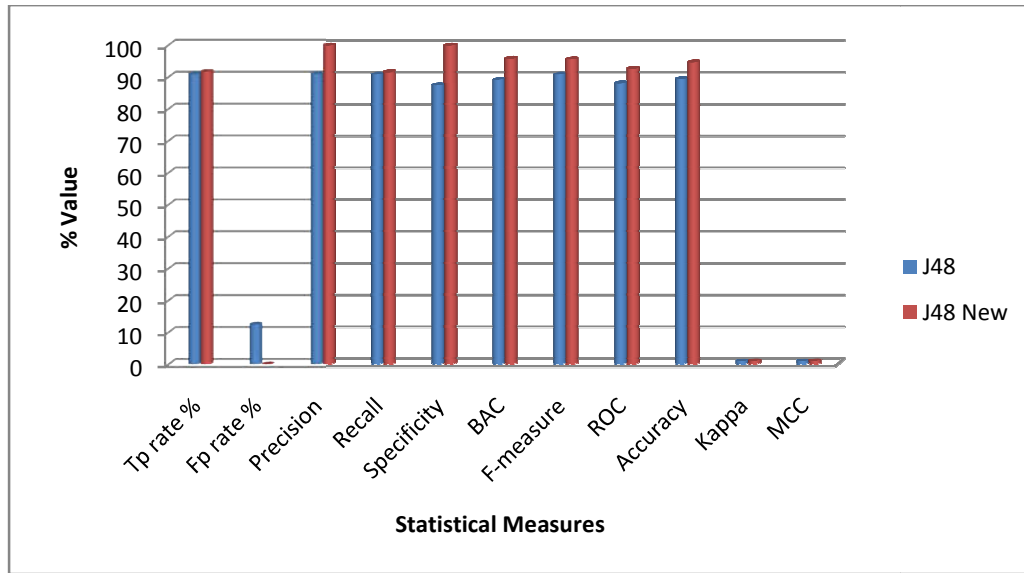


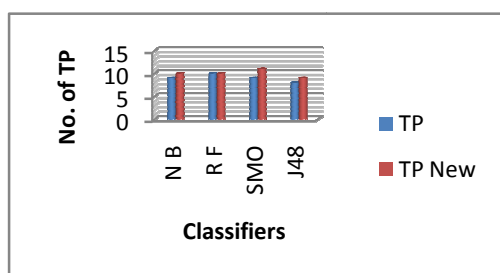
Figure 9. Comparison of J48 Default Model 1 semiconductor vs. newly added descriptor J48 statistical parameters

The Figures 5-9 depict the comparative study of each ML classifier against the newly added descriptor model. In Figure 6, the new descriptor had affected the Bayesian semiconductor model. Since the parameters TP rate, recall, BAC, F-measure, ROC, accuracy, kappa and MCC had elevated that enhances the model robustness. For the same the parameters FP rate, precision, and specificity remained same and thus the newly descriptor added model was better in comparison to the default electronic model. In the case of RF model as shown in Figure 7 the newly added descriptor had not affected the model accuracy and the model fineness remained to be same as semiconductor default model. The newly added SMO model performed better in comparison with SMO default as all the evaluation measures showcased higher accuracy as depicted in Figure 8. The same was observed for the newly added J48 model where the FP rate was reduced to zero and the accuracy to the highest 94.73% as shown in Figure 9. Overall the models RF new and J48 new outperformed against the electronic model with respect to all the other ML models.

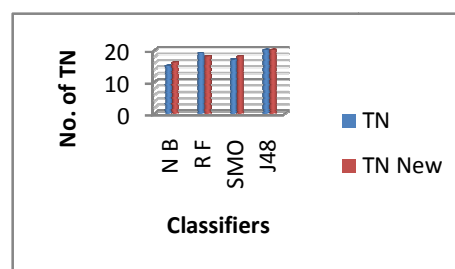
Table 13. TP, TN, FP, FN comparison of Pseudomonas Model vs. Pseudomonas Model added new descriptor (lone pair pi 2d walk descriptor)

Classifier	TP	TP New	TN	TN New	FP	FP New	FN	FN New
N B	9	10	15	16	10	11	3	4
R F	10	10	19	18	8	8	2	3
S M O	9	11	17	18	8	7	4	4
J 4 8	8	9	20	20	6	5	5	5

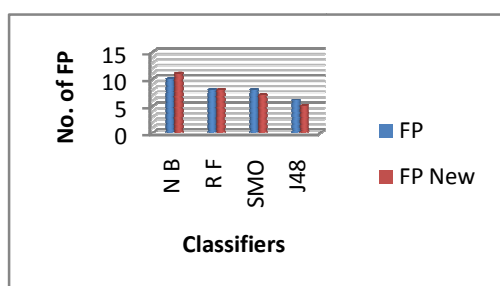
(a) TP vs. TP New



(b) TN vs. TN New



(c) FP vs. FP New



(d) FN vs. FN New

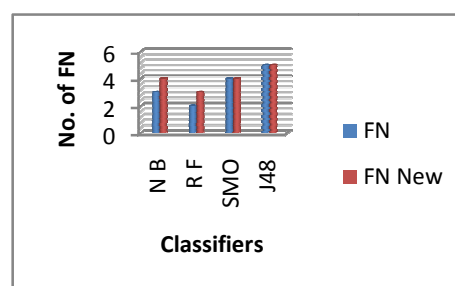


Figure 10. The comparative analysis of the number of TP, TN, FP and FN were studied vs. newly added descriptor model for the *P. aeruginosa* bacterial ML models. (a) Comparison of TP vs. New TP against each ML Model is mentioned (b) comparison of TN vs. New TN against each ML Model is mentioned (c) comparison of FP vs. New FP against each ML Model is mentioned (d) comparison of FN vs. New FN against each ML Model is mentioned. Number of TP for RF model, TN for J48, FP for RF and FN for J48 remained same.

Table 14. Statistical parameters of ML classifiers from Pseudomonas model vs. newly added descriptor Pseudomonas model

Statistical parameters	N B	N B New	R F	R F New	SMO	SMO New	J48	J48 New
TP rate %	69.2	76.92	76.9	83.33	69.2	73.33	61.5	64.28
Fp rate %	38.5	42.307	30.8	42.3	30.8	29.16	23.1	20
Precision	47.4	47.619	55.6	55.55	52.9	61.11	57.1	64.28
Recall	69.2	76.92	76.9	83.33	69.2	73.33	61.5	64.28
Specificity	61.1538	57.69	69.2307	70.37	69.2307	70.83	76.923	80
BAC	65.1769	67.305	73.0653	76.85	69.2153	72.08	69.2115	64.28
F-measure	56.3	58.822	64.5	66.66	60	66.65	59.3	32.13
ROC	74	74.09	82.5	78.82	69.2	73.67	67	70.98
Accuracy	64.1026	64.1	71.7949	74.35	69.2308	71.79	71.7949	74.35
Kappa	0.2759	0.317	0.4211	0.483	0.3571	0.425	0.3774	0.4413
MCC	0.2901	0.3273	0.4364	0.4972	0.3656	0.431	0.37796	0.4428

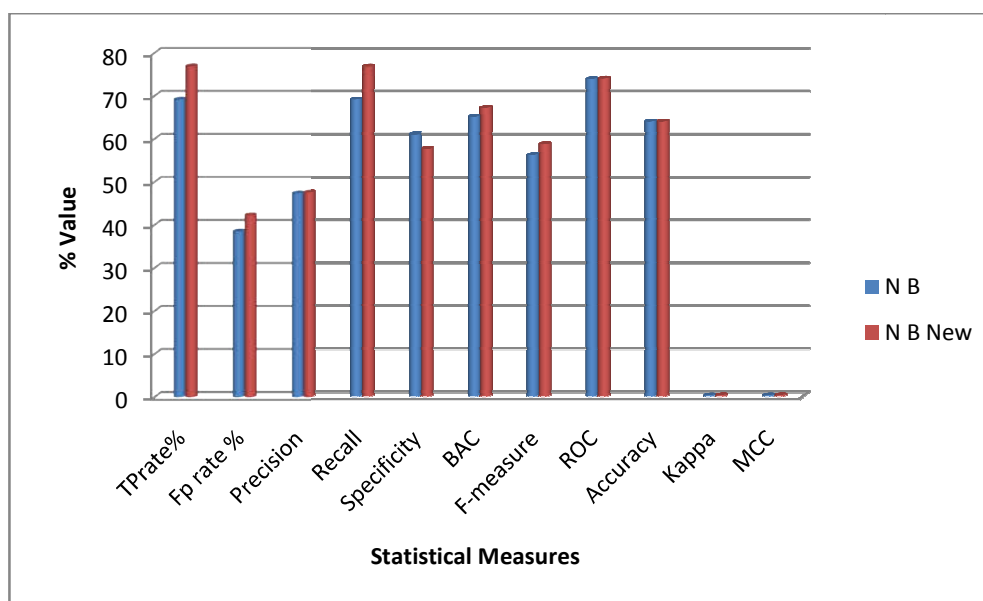


Figure 11. Comparison of N B Pseudomonas Model 1 vs. newly added descriptor N B statistical parameters

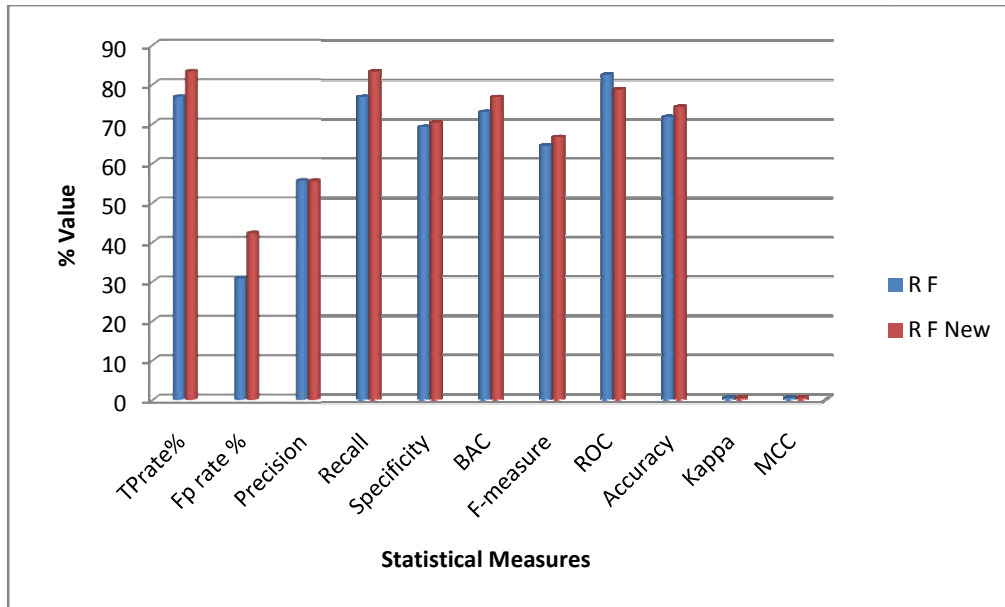


Figure 12. Comparison of R F Pseudomonas Model 1 vs. newly added descriptor R F statistical parameters

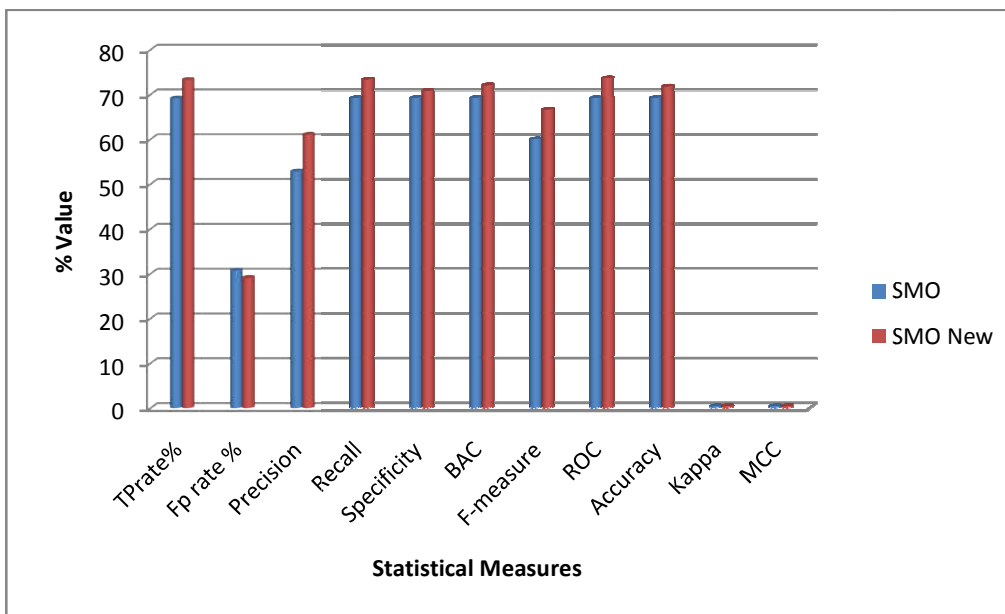


Figure 13. Comparison of SMO Pseudomonas Model 1 vs. newly added descriptor SMO statistical parameters

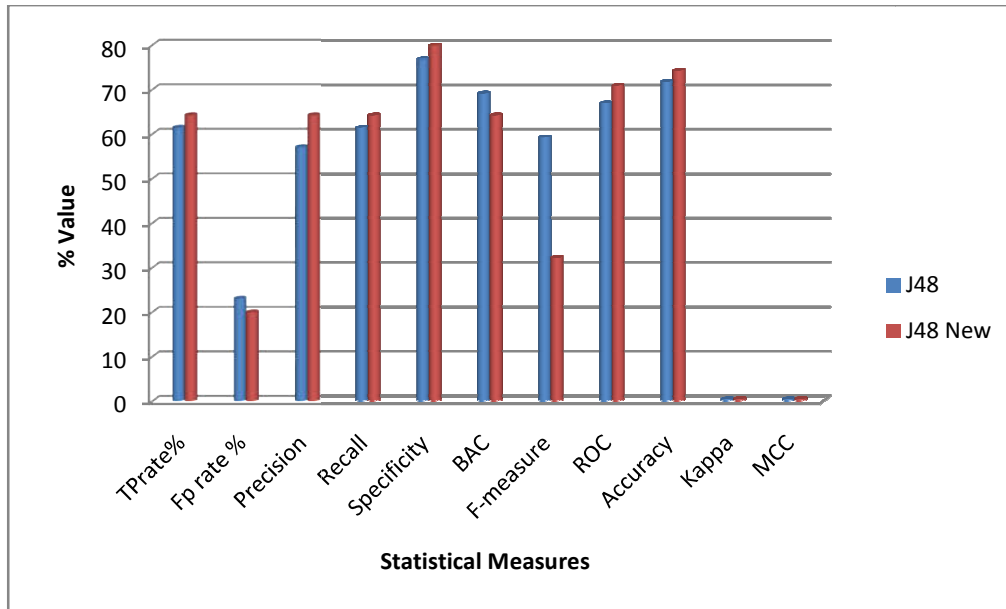


Figure 14. Comparison of J48 Pseudomonas Model 1 vs. newly added descriptor J48 statistical parameters

The Figures 10-14 depict the comparative study of each classifier against the newly added descriptor in the Pseudomonas ML models. In Figure 11 the new descriptor has slightly affected the Pseudomonas model. The parameters TP rate, recall, BAC, F-measure, kappa and MCC had elevated that enhances the Pseudomonas Bayesian model robustness. For the same the parameters FP rate was also increased to 42.3 though it was not under the threshold value. Precision, ROC and accuracy remained same and specificity was reduced to 57.69. Thus the newly descriptor added model has only marginal performance in comparison to the Bayesian default pseudomonas model. In the case of RF model the newly added descriptor had increased various evaluation measures but the model produced a higher FP rate as shown in Figure 12. At the same time accuracy, kappa and MCC was slightly better in the RF new descriptor added ML model. SMO ML model comparison with newly added descriptor SMO model had produced a better model with less FP rate as displayed in Figure 13. In the comparative graph as shown in Figure 14 J48 new model had performed better in comparison to J48 since FP rate is

under the threshold value 20%. Here also the model accuracy was higher for RF new and J48 new Pseudomonas model in respect to other models.

Table 15. TP, TN, FP, FN comparison of TB Model vs. TB Model added new descriptor (lone pair pi walk count descriptor)

Classifiers	TP	TP New	TN	TN New	FP	FP New	FN	FN New
N B	35	40	93	94	70	66	40	38
R F	16	21	143	139	20	20	59	58
S M O	10	7	136	157	27	9	65	65
J 4 8	27	28	109	110	54	51	48	49

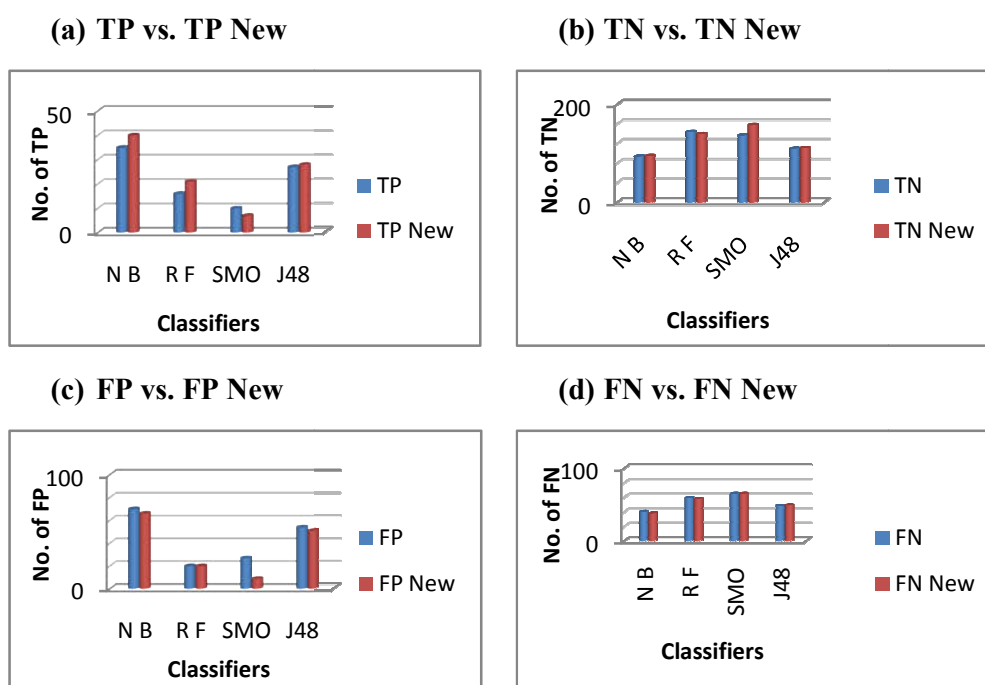


Figure 15. The comparative analysis of the number of TP, TN, FP and FN were studied vs. newly added descriptor model for the *M. tuberculosis* biological ML models. (a) Shows comparison between number of TP vs. New TP (b) shows comparison between number of TN vs. New TN (c) shows comparison between number FP vs. New FP (d) shows comparison between number of FN vs. New FN. The number of TP New has increased for all the classifier except for SMO model. While the number of TN new has reduced for R F model and in the case of FP new SMO model has the lowest. And the number of FN has remained same for FN and FN new model.

Table 16. Statistical parameters of ML classifiers from TB model vs. newly added descriptor TB model

Statistical parameters	N B	N B New	R F	R F New	SMO	SMO New	J48	J48 New
TP rate %	46.7	51.28	21.3	26.582	13.3	9.722	36	36.36
Fp rate %	42.9	41.25	12.3	12.578	16.6	5.42	33.1	31.677
Precision	33.3	37.73	44.4	51.219	27	43.75	33.3	35.44
Recall	46.7	51.28	21.3	26.582	13.3	9.722	36	36.36
Specificity	57.0552	58.75	87.73	87.421	83.435	94.57	66.8711	68.322
BAC	51.8776	55.015	54.515	57.001	48.867	52.146	51.4355	52.341
F-measure	38.9	43.47	28.8	35	17.9	15.908	34.6	35.89
ROC	53.3	53.297	59	62.43	51	67.92	51.8	55.13
Accuracy	53.7815	56.302	66.8067	67.226	61.34	68.907	57.1429	57.983
Kappa	0.0336	0.0975	0.1054	0.171	-0.4141	0.051	0.0281	0.0482
MCC	0.03483	0.0947	0.1251	0.1746	-0.4142	0.0522	0.0239	0.0465

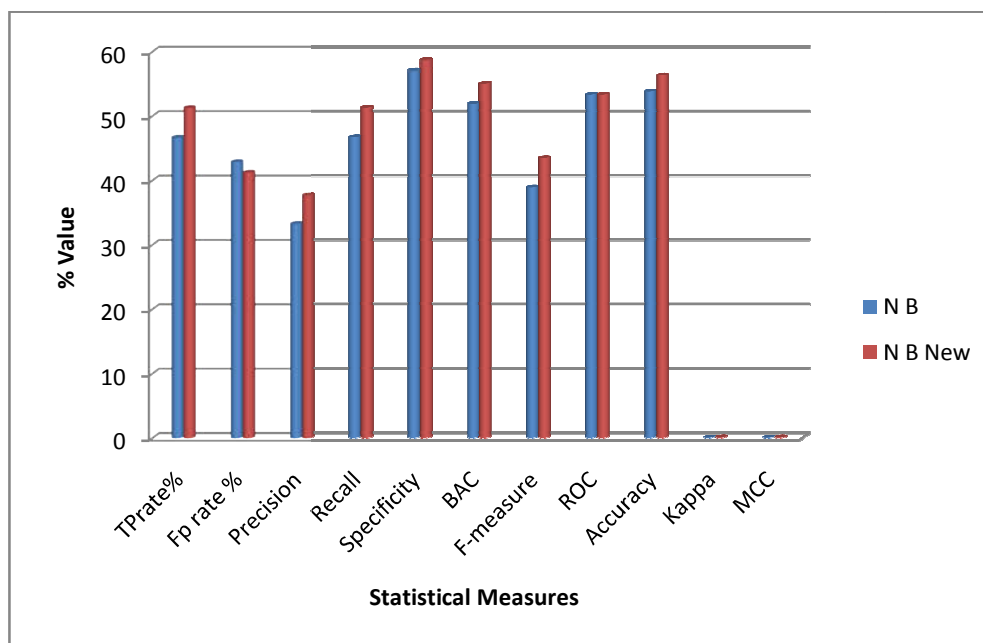


Figure 16. Comparison of N B TB Model 1 vs. newly added descriptor N B statistical parameters

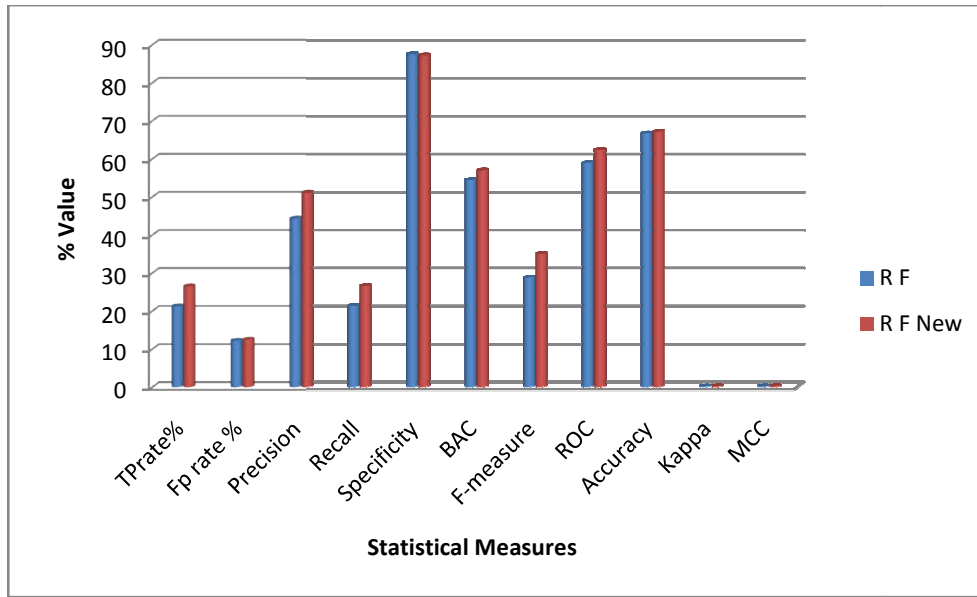


Figure 17. Comparison of R F TB Model 1 vs. newly added descriptor R F statistical parameters

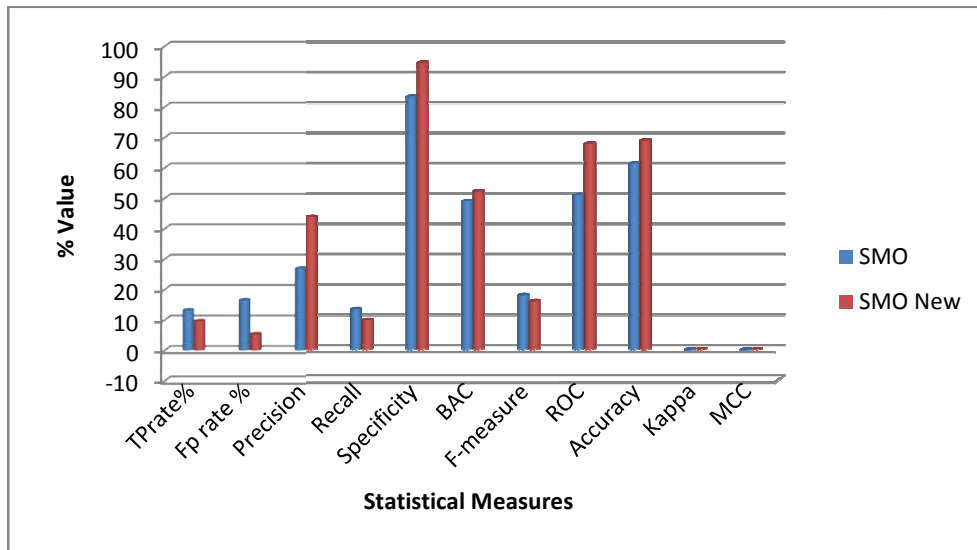


Figure 18. Comparison of SMO TB Model 1 vs. newly added descriptor SMO statistical parameters

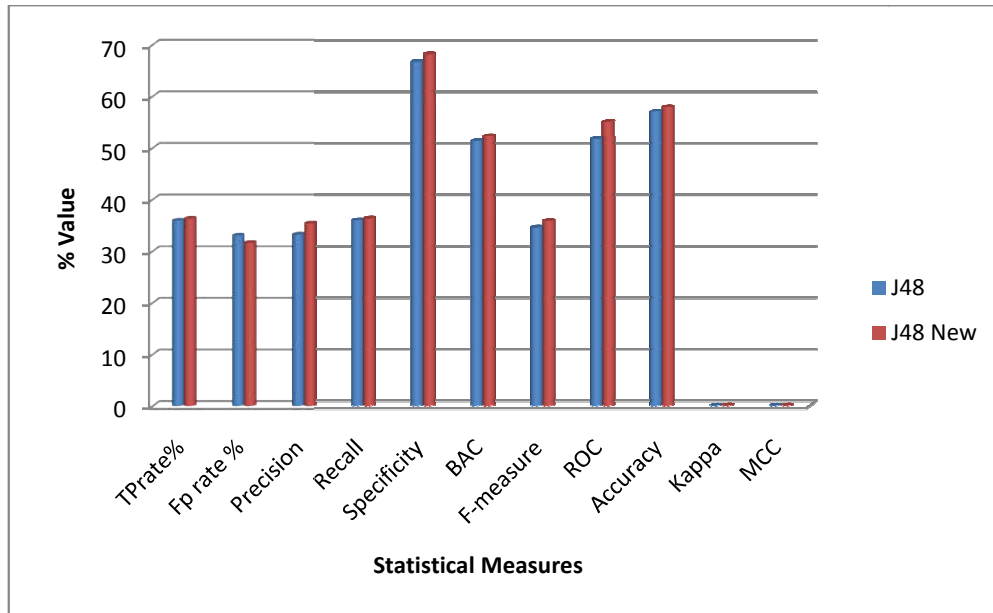


Figure 19. Comparison of J48 TB Model 1 vs. newly added descriptor J48 statistical parameters

The Figures 15-19 depict the comparative study of each ML classifier vs. the newly added descriptor in the TB ML models. As shown in Figure 16 the parameters TP rate, precision, recall, specificity, BAC, F-measure, accuracy, kappa and MCC had elevated that enhanced the TB Bayesian model robustness. For the same the parameters FP rate was lowered to 41.25 though it was not under the threshold value. ROC remained same for the NB new model. Thus the NB new model has only marginal performance in comparison to the Bayesian default TB model. In the case of RF model the RF new model had increased various evaluation measures having a lowered FP rate as shown in Figure 17. At the same time accuracy, kappa and MCC was slightly improved in the RF new descriptor added ML model. SMO ML model comparison with newly added descriptor SMO model had produced a model with high accuracy but the FP rate, recall, F-measure were lowered at the cost of less FP rate as shown in Figure 18. The FP rate of SMO new has lowered from 16.6 to 5.42 which were found to be in accordance to threshold FP rate. Also, kappa and MCC were elevated from negative value to a positive value. J48 new TB model performance was found to be very similar to the J48 TB model as depicted in Figure

19. In comparison to all other TB ML models SMO new and RF new model performed well in respect to accuracy and FP rate.

3.4 Conclusion

By correlating semiconductivity with biological activity we could establish a graph theoretical relationship between two entirely different class systems and could develop a descriptor which we could establish/approves a higher dimensional level of similarities between two different physical and biological processes. The high-throughput virtual screening philosophy has been extensively employed by many pharmaceutical industries and in our study we adopted the virtual screening in organic material science for predicting semiconductor molecules. And HTS in organic materials is a long way. Unlike the complex structure-biological activity relationships in drug discovery, the relation between molecular structures, electronic structure and device properties in organic electronic materials are more straightforward. But the problem is with the CPU cost of computational methods which is quite high and it is too important to have a high hit-ratio in organic materials. So, the idea was to repurpose the semiconductor class of compounds with biological molecules and vice versa. In drug discovery repurposing of drugs is very effective as it reduces cost, time and the various stages including pre-clinical and clinical trials in the drug development. Similarly repurposing of computationally active biological molecules and semiconductor active molecules, the plausibility of *in vitro-in vivo* analysis for bio-active molecules and synthesis part for the semiconducting molecules can be avoided. By cross screening the semiconductor class with anti-bacterial class we could establish some connectivity like the charges or electronic flow in a biological mechanism which has to be confirmed from bigger experiments. The study could provide a new direction that many molecules selected semiconductor are passed through biological screen similarly the biologically active anti-bacterial molecules were passed through semiconductor model. It indicates that more underlying features exist which control both properties anti-bacterial and semiconductor. Hence a new 2D descriptor namely lone pair pi walk count eight was postulated and validated. The walk descriptor developed is based on various virtual

screening computational predictive models against the electronic and anti-bacterial activity. The pattern lone pair pi conjugation was visible for the 129 molecules which were virtually screened through various electronic and anti-bacterial methods. The pattern count was manually calculated starting from 2 to 12 until they were distinguishable from actives to inactives. In this regard a new descriptor was identified with count 8 and was validated on the existing electronic and biological ML models. The ML model robustness was checked from the various evaluation measures. The accuracy for all the classifiers was higher except for the classifier in semiconductor model (Random forest-remained unchanged), Pseudomonas model (Naïve Bayes-remained unchanged). The elevated values may be explained based on the new information gathering and methods in the electronic and biological systems. We believe our results to be encouraging and will provide useful insight in designing new molecules possessing both anti-bacterial and electronic nature. The studies were initiated because of the role of the molecular descriptor in the development of various machine learning predictive models.

REFERENCES

- ¹ Aukasz PaBkowski, Jerzy BBaszczyNski, Andrzej Skrzypczak, Jan BBaszczak, Alicja Nowaczyk, Joanna Wróblewska, Sylwia Kohuszko, Eugenia Gospodarek, Roman SBowiNski, J. K. Prediction of Antifungal Activity of Gemini Imidazolium Compounds. *Biomed Res. Int.* **2015**, *2015*, 10.
- ² Hastings, J.; Magka, D.; Batchelor, C.; Duan, L.; Stevens, R.; Ennis, M.; Steinbeck, C. Structure-Based Classification and Ontology in Chemistry. *J. Cheminform.* **2012**, *4*, 8.
- ³ Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*, 2nd ed.; John Wiley & Sons, 2009; Vol. 2.
- ⁴ R. Todeschini, V. Consonni, P. G. Chemometrics in QSAR. In *Comprehensive Chemometrics*; Brown S, Tauler R, W. R., Ed.; Oxford: Elsevier, 2009; Vol. 4, pp 129–172.
- ⁵ Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold², Molecular Descriptors from 2D Structures for Chemoinformatics and. *Chem. Inf. Model.* **2008**, *48*, 1337–1344.
- ⁶ Ying, J.; Zhang, T.; Tang, M. Metal Oxide Nanomaterial QNAR Models: Available Structural Descriptors and Understanding of Toxicity Mechanisms. *nanomaterials* **2015**, *5*, 1620–1637.
- ⁷ Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.
- ⁸ <https://preadmet.bmdrc.kr/>
- ⁹ Li, Z. R.; Han, L. Y.; Xue, Y.; Yap, C. W.; Li, H.; Jiang, L.; Chen, Y. Z. A RTICLE MODEL — Molecular Descriptor Lab□: A Web-Based Server for Computing Structural and Physicochemical Features of Compounds. *Biotechnol. Bioeng.* **2007**, *97*, 389–396.
- ¹⁰ Yap, C. W. E. I. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2010**, 1466–1474.
- ¹¹ Kunal Roy Supratik Kar Rudra Narayan Das. Chemical Information and Descriptors. In *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*; Elsevier Inc., 2015; pp 47–80.
- ¹² Puzyn, T.; Leszczynski, J.; Cronin, M. T. Recent Advances in QSAR Studies. In *Challenges and Advances in Computational Chemistry and*

Physics Series Volume; Puzyn, T., Leszczynski, J., Cronin, M. T., Eds.; Springer Netherlands, 2010; p 414.

- ¹³ Todeschini, R.; Lasagni, M. NEW MOLECULAR DESCRIPTORS FOR 2D AND 3D STRUCTURES . THEORY. *J Chemom.* **1994**, *8*, 263–272.
- ¹⁴ Brüstle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. Descriptors, Physical Properties, and Drug-Likeness. *J. Med. Chem.* **2002**, *45*, 3345–3355.
- ¹⁵ Paster, I.; Shacham, M.; Brauner, N. Investigation of the Relationships between Molecular Structure , Molecular Descriptors , and Physical Properties. *Ind. Eng. Chem. Res.* **2009**, *48*, 9723–9734.
- ¹⁶ Guha, R.; Willighagen, E. A Survey of Quantitative Descriptions of Molecular Structure. *Curr. Top. Med. Chem.* **2012**, *12*, 1946–1956.

SUMMARY

The present structure-data-information for chemical and biological system has grown exponentially that more computer power, online resources are required to develop modern computational predictive models. And it is so performed that molecular descriptors are decoded from the chemical representation by using various data mining techniques and different machine learning algorithms. Their application is interdisciplinary and can be adapted to medicinal chemistry, Quantitative Structure Activity Relationships (QSAR), Quantitative Structure Property Relationships (QSPR), toxicology, *in silico* drug design, virtual high-throughput screening and even in non scientific fields like banking, fraud detection, face detection etc. Its remarkable applications have given a lot of inspiration for developing electronic semiconductor property and biological anti-bacterial activity as well as for developing a molecular descriptor having the characteristic property of both. It is also proposed that the descriptor would enhance in designing new molecules in the material and anti-bacterial chemical space.

The thesis was carried out into three parts. Part I deals with the development of electronic models that can classify a set of molecules into organic semiconductors and non semiconductors. Part I comprises six chapters. The first chapter consists of an introduction and a critical review of the published work on data mining methods, machine learning models and its various statistical parameters, virtual screening and organic semiconductors and its applications. In the second chapter, materials, methods, softwares and web servers that were used for the various studies are described.

Development of electronic Bayesian organic semiconductor models is described in chapter 3. All the organic semiconductor machine learning models were developed from the experimental data based on electronic band gap energy. Insulators were considered for the non semiconductors dataset. Two machine learning Bayesian models were developed one corresponding to the Default Model while the other belonged to Oversampled Model. And the screening set involving

Schiff base molecules from ChEBI database were virtually screened against the computational models and prioritized three molecules that were predicted to be computationally organic semiconductor actives.

Chapter 4 deals with two kinds of organic semiconductor decision tree algorithms: Random Forest and J48. Here four types of decision tree predictive models were developed. Two default models one corresponding to Random Forest and J48 while the other two belonged to Oversampled Random Forest and Oversampled J48 model. Unpruned was adapted for the J48 model built for better performances. Oversampled model of Random Forest and J48 performed better in comparison to the respective Default Random Forest and J48 models. And all the computational predictive decision tree models were used for the virtual screening and a few of the molecules were prioritized to be computationally organic semiconductor active.

Predictive model generations of electronic SMO models are described in chapter 5. SMO is a type of support vector machines that make use of “hyperplane” in the multidimensional data space to split active compounds from inactive. Such models were developed for predicting semi conductivity nature from the screening set selected from ChEBI database. Four logistics SMO models were built from data mining software WEKA. Two SMO predictive models; Default and Oversampled model were developed and virtually screened against the screening set from the ChEBI library. A total of seven molecules were prioritized as organic semiconductor actives based on SMO models.

Chapter 6 deals with the pattern recognition among the virtually screened computationally active organic semiconductors. The patterns identified in the SMARTS format were generated from the Maximum Common Substructure from the Canvas Schrodinger suite. The reported pattern signifies the semiconductor nature was published in the journal Springer. Part I ends with references.

Part II deals with various machine learning anti-bacterial biological models and structure based and ligand based approaches. This part comprises eight chapters. The first chapter gives information on various biological databases, biological

descriptor generator software, drug discovery pipeline, selection of protein target and various virtual screening methods like different machine learning classifiers, structure based approach, ligand based approach and artificial neural methodology. In chapter 2 material and methods are explained. Chapter 3 describes the various Bayesian machine learning models for *M. tuberculosis* and *P. aeruginosa* against the target β -lactamase enzyme under the study. Sampling methods like Oversampling and SMOTE were performed for the biased dataset. Four machine learning models were developed from the public repository database PubChem enriched with 179 biological descriptors. Further carried out virtual screening of 177 anti-TB molecules from the GSK library which is tabulated in the respective section. The models could prioritize computationally active β -lactamase inhibitors against the microbes under the study. In chapter 4 various biological decision tree models for the microbes *M. tuberculosis* and *P. aeruginosa* are discussed. Her eight decision tree models were developed that comprises four each from Random Forest and J48 models. The models were built based on the 179 biological descriptors. The mentioned models were used to screen 177 anti-TB molecules from the GSK library. All the results are summarized in the respective tables in the results and discussion section. Chapter 5 discusses on the various logistics SMO machine learning anti-bacterial models. Four models built were screened against the GSK 177 anti-TB molecule for prioritizing computationally active β -lactamase inhibitors. In chapter 6 Structure Based Virtual Screening approach is discussed, where GSK 177 anti-TB molecules are docked against the target β -lactamases from *M. tuberculosis* and *P. aeruginosa*. Molecular docking was performed for the targets selected from Protein Data Bank, PDB id: 2GDN for *M. tuberculosis* and PDB id: 2WKH for *P. aeruginosa* from the software Schrodinger suite glide. Protein ligand docking was analyzed based on the threshold docking scores of β -lactamase inhibitor Clavulanic acid. The result docked β -lactamase actives for the selected microbes is summarized on the respective section. Chapter 7 describes virtual screening of GSK 177 anti-TB molecules through Artificial Neural Network based Self Organizing Maps. Self Organizing Maps was generated for the target organism under study, performed virtual screening and prioritized computationally active β -lactamase inhibitors. In

chapter 8 sensitivity of the computational methods Bayesian classifier, molecular docking and self organizing maps were studied for the pocket dissimilarity of the selected targets under study. All the results are summarized in their respective sections. Part II concludes with references.

Part III consists of development and validation of molecular descriptor. Chapter 1 gives a brief introduction of molecular descriptor. Chapter 2 deals with materials and methods. In chapter 3 a molecular descriptor is postulated namely “lone pair pi walk count 8” based on descriptor-based virtual screening computational predictive models against the electronic and anti-bacterial activity. The developed descriptor is validated on the existing electronic and biological machine learning models. Part III ends with list of references.