University of Essex
Department of Mathematical Sciences

MA981: DISSERTATION

NIPAH INHIBITOR

GEETHANJALI ERIYIL RAMESH
2211420

Supervisor: IGOR RODIONOV

November 24, 2023
Colchester

PREFACE

This dissertation aims to contribute to the development of a vaccine for the deadly Nipah virus infection. My awareness of this virus began in 2018 when the first case was reported in Kerala, India, my home state. Recognizing the severity of Nipah's spread, I struggled to choose a topic for my dissertation. During a discussion with my professor, Dr. Igor Rodionov, I expressed my interest in the medical field. He suggested considering diseases common in India, prompting me to think of Nipah. We finalized Nipah as the topic.

Finding a dataset for Nipah prediction proved challenging. Despite efforts to connect with hospitals that treated Nipah patients in 2018, I hit a dead end. However, discovering a paper titled "Computational Identification of Inhibitors Using QSAR Approach Against Nipah Virus" gave me hope. After reading this paper, I shifted my focus from Nipah prediction to creating a model for identifying compounds that could aid in developing a cure for the disease.

As someone with no background in biology, microbiology, or chemistry, understanding molecular properties and descriptors was daunting. However, I persevered, immersing myself in online resources and papers to grasp the basics of molecular descriptors, structures, and microbiological terms necessary for the dissertation and report generation. Despite the challenges, I remained determined, recognizing the potential impact of my work on Nipah virus research and the importance of contributing to finding a cure.

Contents

List of Figures

CHAPTER 1

Abstract

The urgent need for effective vaccines against Nipah virus, a highly lethal zoonotic pathogen, has spurred the exploration of innovative strategies for candidate compound identification. This project focuses on leveraging Quantitative Structure-Activity Relationship (QSAR) modeling to discern and prioritize potential compounds with promising vaccine attributes. The methodology involves comprehensive computational analyses of molecular structures, biological activities, and physicochemical properties, amalgamating them into a robust QSAR model. Through this approach, we aim to pinpoint the most potent compound candidates for vaccine creation against Nipah virus. The synthesis of these prioritized compounds will pave the way for subsequent in vitro and in vivo evaluations, ultimately advancing the development of a highly efficacious Nipah virus vaccine. This interdisciplinary endeavor not only showcases the synergy between computational and experimental approaches but also underscores its potential impact on public health by providing a proactive solution to combat the menace of Nipah virus infections.

”’QSAR MODEL architecture formated from the paper "Computational Identification of Inhibitors Using QSAR Approach Against Nipah Virus"[5] 1.1”’.

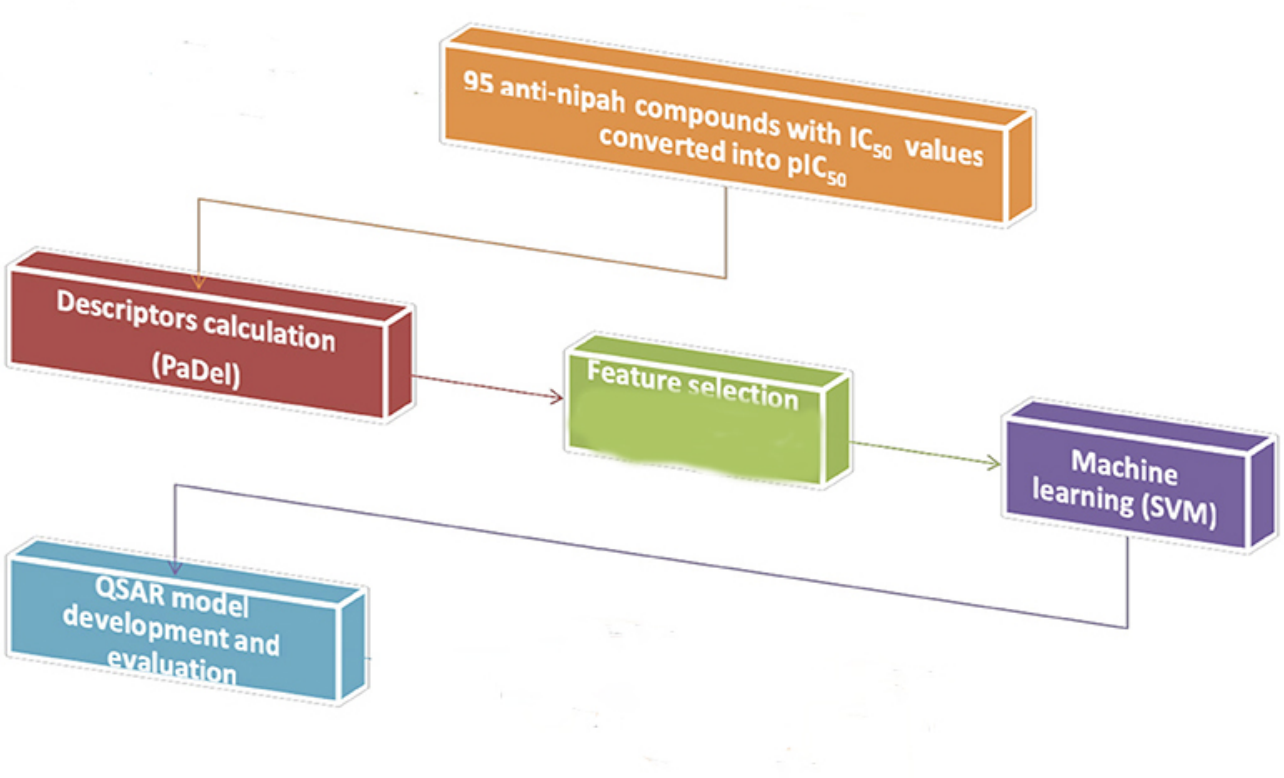


Figure 1.1: Overall architecture for the development of QSAR MODEL

CHAPTER 2

Introduction

2.1 Aim:

The aim of the dissertation is to find the most effective inhibitor for the vaccine development for the dangerous virus NIPAH using QSAR model.

2.2 Nipah:

NiV is a member of the family Paramyxoviridae, genus Henipavirus. It is a zoonotic virus, meaning that it initially spreads between animals and people. The animal host reservoir for NiV is the fruit bat (genus Pteropus), also known as the flying fox. Nipah virus is also known to cause illness in pigs and people. Infection with NiV is associated with encephalitis (swelling of the brain) and can cause mild to severe illness and even death.(Center for disease control and prevention [[1]) Nipah virus is classified as Biosafety Level-4 based on its high pathogenicity in humans and lack of available vaccines and therapeutics.

The structure of nipah virus is given in figure 2.1

Like other henipaviruses, the Nipah virus genome is a single (non-segmented) negative-sense, single-stranded RNA of over 18 kb, which is substantially longer than that of other paramyxoviruses. The enveloped virus particles are variable in shape, and can be filamentous or spherical; they contain a helical nucleocapsid.[2] Six structural proteins are generated: N (nucleocapsid), P (phosphoprotein), M (matrix), F (fusion), G (glycoprotein) and L (RNA

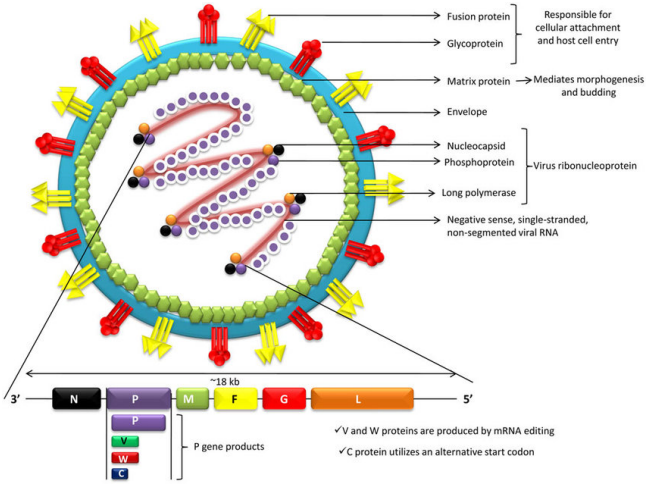


Figure 2.1: Structure of henipah virus

polymerase). The P open reading frame also encodes three nonstructural proteins, C, V and W.[6]

The linear arrangement of genes on the RNA genome is as follows: 3′-N-P-M-f-G-L-5′. This sequence is crucial for the functioning of the virus life cycle and the synthesis of viral proteins. The order from 3′ to 5′ reflects the sequence in which these genes are encountered during transcription and replication processes. Each gene serves a specific purpose:

- The N (Nucleoprotein) gene is essential for maintaining the integrity of the viral RNA.
- The P (Phosphoprotein) gene is involved in RNA synthesis, acting as a co-factor for the viral polymerase.
- The M (Matrix protein) gene is responsible for virus assembly.
- The F (Fusion protein) gene is involved in the fusion of the viral envelope with the host cell membrane, facilitating virus entry into the host cell.
- The G (Adhesion Glycoprotein) gene is responsible for viral attachment to host cells and is crucial for the initial stages of infection.
- The Large protein 'L' serves as the viral RNA-dependent RNA polymerase and is responsible for the replication and transcription of the viral genome.

Furthermore, the P gene also generates three non-structural proteins, V, W, and C, through mRNA editing mechanisms and alternative open reading frames. Protein V frequently

suppresses host antiviral responses by disrupting the host's type 1 interferon signaling pathway. Protein W acts as a regulator of viral transcription and replication. Protein C is engaged in regulating viral RNA synthesis and replication, as well as aiding in immune evasion.

The ability to produse multiple protiens from a single gene highlights the adaptability and complexity of paramoxyviruses in their interactions with the host's immune system and their ability to evade host defenses.

This genetic flexibility is an important factor in the pathogenesity and virulence of paramoxyviruses.[7]

2.2.1 Symptoms:

Early symptoms of nipah includes fever, cough, headache, sore throat, difficulty breathing, vomiting. Severe conditions of nipah are drowsiness, disorientation, seizures, coma, brain swelling [enciphalitis] Typically the symptoms start to occur after 4 to 14 days after the exposure to the virus.The symptoms start with a fever, cough and may have respiratory illness. The next stage of nipah lead to encephalitis[brain swelling] which can rapidly change to coma within 24 to 48 hours. Death may occurs in 40-75 percent of people. Long-term side effects in survivors of Nipah virus infection have been noted, including persistent convulsions and personality changes.[[2]]

2.2.2 Diagnosis:

There are two tests available for testing nipah, during early stages of nipah we use real time polymerase chain reaction [RT-PCR] from throat and nasal swabs, cerebrospinal fluid, urine, blood.while in the course of illness or after recovery we use enzyme-linked immunosorbent assay (ELISA) in-order to find the antibodies present in the individuals. Early detection of Niv infection could be difficult because of the non-specific early symptoms of the illness. However if we could detect and diagnose Niv infection at the earliest will help to increase the chances of survival of the infected person and to prevent transmission to other people, and to manage outbreak response efforts.[[3]]

2.2.3 Treatment:

As of now no licensed treatments available for Nipah virus (NiV) infection. Treatment is limited to supportive care, including rest, hydration, and treatment of symptoms as they occur. However, there are immunotherapeutic treatments, such as monoclonal antibody therapies, currently undergoing development and evaluation for treating NiV infections. One such monoclonal antibody, m102.4, has completed phase 1 clinical trials and has been administered on a compassionate use basis. Additionally, the antiviral medication remdesivir has shown efficacy in nonhuman primates when administered as post-exposure prophylaxis and may complement immunotherapeutic treatments. Ribavirin, another drug, was used to treat a limited number of patients during the initial Malaysian NiV outbreak, but its effectiveness in humans remains uncertain.[[4]]

2.3 QSAR Model:

QSAR : Quantitative Structure-Activity relationship models are regressions used in chemical and biological science. In QSAR model the predictors consists of physicochemical properties or theoretical molecular descriptors of chemicals and the response variable will be the biological activity of the chemicals. A QSAR model first summarise a supposed relationship between the chemical structures and biological activity in a dataset and then the models tries to predict the activities of new chemicals with the same properties. A biological activity can be expressed quantitatively as the concentration of a substance required give a certain biological response. In QSAR model the physicochemical properties or the structures are expressed as numbers which help us to find a mathematical relationship between the predictor and response.[[5]]

The mathematical form : Activity = f(physicochemical properties and structural properties)+error

2.4 Inhibitor

An enzyme inhibitor is a molecule that bimds to an enzyme and blocks its activity. An enzyme facilitates a specific chemical reaction by binding the substrates to its active state,

a specialized area on the enzyme that accelerates the most difficult step of the reaction. An enzyme inhibitor stops ("inhibits") this process, either by binding to the enzyme's active site (thus preventing the substrate itself from binding) or by binding to another site on the enzyme such that the enzyme's catalysis of the reaction is blocked. Enzyme inhibitors are found in nature and also produced artificially in the laboratory. Inhibitors are mainly used for metabolic regulations,in making drugs and are used as pesticides.Here in this dissertation we need to find the best inhibitors that will lead to a vaccine creation against nipah.

CHAPTER 3

Literature Review

3.1 History Of Nipah

NiV outbreaks have so far been reported in three countries: Malaysia, Bangladesh and India. The initial outbreak of Nipah occurred in Malaysia from 1998 to 1999, where a total of 265 confirmed cases were reported, resulting in 105 deaths. The virus initially affected pigs, which exhibited respiratory illness and encephalitis. Subsequently, the virus transmitted to humans. Initially, Malaysian health authorities attributed the infections to "Japanese Encephalitis," leading to a delay in effectively deploying measures to prevent its spread. This misunderstanding stemmed from serum tests conducted on 28 infected individuals in the area, which tested positive for JE-specific immunoglobulin M, as confirmed by the WHO. Nipah, as deadly as Ebola virus disease, targets the brain system instead of blood vessels. Local virologists from the Faculty of Medicine, University of Malaya, identified this as a new infectious agent.[8]

In late February 1999, Nipah Virus (NiV) spread to Singapore after infected pigs were imported from Malaysia. This resulted in 11 abattoir workers contracting the disease, with one fatality [28]. Prompt measures, such as culling infected pigs and implementing a ban on pig imports from Malaysia, were taken to contain the infection. These actions proved effective, and by May 1999, the spread of the virus was successfully contained.

The first recorded outbreak of Nipah Virus (NiV) occurred in the district of Meherpur, Bangladesh, in April 2001, with 13 diagnosed cases. Subsequently, numerous outbreaks of

NiV were reported annually across various parts of Bangladesh, including Naogoan, Rajbari, Faridpur, Tangail, Thakurgaon, Kushtia, Pabna, Natore, Manikganj, Gaibandha, Rangpur, Nilphamari, Madaripur, Gopalganj, Lalmohirhat, Dinajpur, Comilla, Joypurhat, Rajshahi, Jhenaidah, Mymensingh, Ponchoghor, and Magura, spanning from April 2001 to February 2015. Some districts experienced recurrent outbreaks. Transmission of the virus was primarily through the consumption of NiV-contaminated raw date palm sap. However, inadequate surveillance and medical facilities contributed to elevated mortality rates. During this period, approximately 261 confirmed cases were reported, resulting in 199 deaths, with a mortality rate of 76.2

The first outbreak of Nipah Virus (NiV) in India occurred between January and February 2001 in the district of Siliguri, a prominent commercial city in West Bengal. Given Siliguri's proximity to Bangladesh and initial challenges in identifying the causative agent through laboratory investigations, patient samples were retrospectively tested for the NiV virus. This severe outbreak resulted in 45 deaths out of a total of 66 confirmed patients, yielding a mortality rate of 68 percent. Information regarding the true index patient was not available. However, the spread primarily occurred within healthcare settings, with no reports indicating the involvement of animals.

The second NiV outbreak was reported in the district of Nadia, West Bengal, in 2007. All five NiV-positive patients succumbed to the infection within 10 days, resulting in a 100 percent fatality rate.

The third outbreak in India was in Kerala in 2018 May where 23 NiV positive patients were identified with a case-fatality rate of 91 percent.The outbreak commenced on May 2, 2018, in Kozhikode, with a 27-year-old man who was admitted to the hospital presenting with fever and myalgia. As his condition worsened with the onset of high-grade fever, vomiting, and altered sensorium, he was transferred to another hospital but unfortunately succumbed to death. Regrettably, his blood sample was not collected for testing the presence of Nipah Virus (NiV). The spread of the virus was determined to be exclusively nosocomial, with 22 cases contracting the NiV infection from the index patient. Among the 23 cases, only 2 individuals survived while 21 succumbed to the infection (comprising 18 confirmed NiV cases), resulting in the highest mortality rate recorded at 91 percent. The outbreak was declared contained after May 30, 2018.[9] After this kerala has shown occurrences in 2019, 2021 and 2023. In 2019 only one case was reported. The 2021 outbreak began with the death of a 12-year-old boy.The

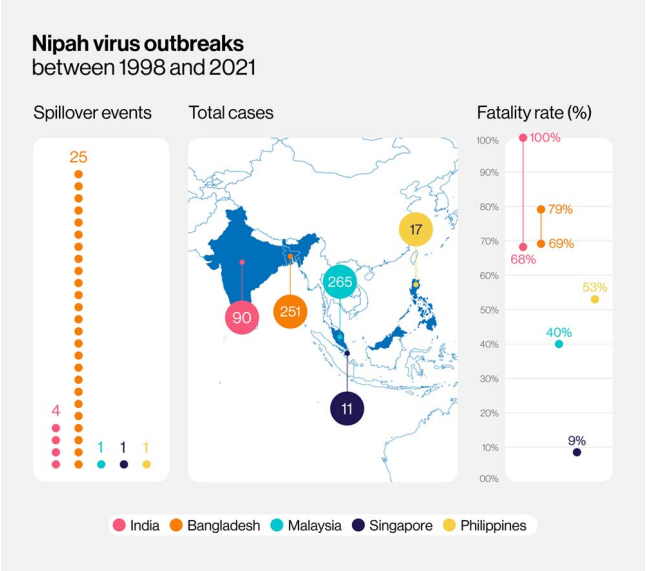


Figure 3.1: outbreaks of Nipah[10]

outbreak was localized in the village and claimed one life.And in 2023 again six cases were reported and two deaths happened.Here i am including a figure that represents the number of cases in different countries and the mortality rates 3.1.

In the quest to uncover the origin of the virus, a team led by Lam Kai Sit from the University of Malaya was formed. During their research, they discovered that the newly identified virus, Nipah, closely resembled a virus called Hendra. Both viruses belong to the paramyxoviridae virus family. The Hendra virus emerged in Australia from 1994 to 1995, resulting in the deaths of a dozen horses and two people. Australian researchers had previously demonstrated that the Hendra virus was widespread among three species of flying foxes, which primarily feed on fruit and nectar.

This discovery led the team to hypothesize that a similar phenomenon might be occurring with Nipah virus. Consequently, they conducted a study focusing on the bat colony near the farms where pigs were infected with Nipah. After collecting 1000 samples, they detected the virus in the urine of one animal, a species known as the flying fox. Additionally, they found live virus in a piece of fruit that had been partially consumed by a bat, suggesting that the virus was also present in the animal's saliva.

These findings confirmed that flying foxes serve as the reservoir for Nipah virus. The research team concluded that the pigs likely became infected when they ingested bat urine or saliva, possibly by scavenging half-eaten fruit dropped from trees. Given that Malaysian pig farms often have fruit trees, this scenario was deemed plausible.[11]

3.2 vaccine and therapeutics for nipah

There is no treatment or vaccine for nipah virus as we said this in section 2.2.3. But there are numerous studies are happening inorder to get a vaccine developed.

Ribavirin was the initial drug treatment administered during the NiV outbreak in 1998 due to its broad-spectrum antiviral activity against both DNA and RNA viruses. Reports indicated that it reduced mortality by 36 percent in treated patients, with no apparent side effects observed. The antimalarial drug chloroquine showed promising results in vitro, either alone or in combination with ribavirin. However, studies in various in vivo NiV infection models, including hamsters, ferrets, and AGMs, demonstrated that chloroquine was ineffective. In a hamster model, ribavirin only delayed death from NiV infection, suggesting that the reduced mortality observed in humans during the 1998 outbreak might have been due to improved patient management and empirical treatment rather than solely the use of ribavirin.

Developing a vaccine is primarily the domain of experts in microbiology, immunology, virology, molecular biology, biochemistry, or pharmacology. However, as a data scientist, I can contribute by assisting them in identifying which compounds with specific properties may be more suitable for vaccine development. This can be achieved by creating computational models to analyze data and predict potential candidates for further testing. Despite limited information on computational models for compound selection in Nipah vaccine development, I found one paper that serves as inspiration for this dissertation.

CHAPTER 4

Data Collection

To create a computational method for selecting inhibitor compounds for Nipah vaccine development, I've relied on a single reference paper. To build the dataset for this dissertation, I've collected a PDF containing the structural details of the compounds mentioned in that paper.The PDF containing the structural details of the compounds used in the reference paper is available in the supplementary Material of the paper . This document lists the compounds along with their structure details, providing the necessary information for your research Subsequently, I utilized PaDEL software to extract the molecular descriptors necessary for constructing the model.

PaDEL-Descriptor was developed in Java and comprises two main components: a library component and an interface component. The library component facilitates seamless integration into quantitative structure-activity relationship (QSAR) software, enabling descriptor calculations. Meanwhile, the interface component allows standalone usage of the software. The software designed for computing molecular descriptors and fingerprints. Presently, it offers calculations for 797 descriptors, encompassing 663 1D and 2D descriptors, along with 134 3D descriptors. Additionally, the software provides 10 types of fingerprints. These descriptors and fingerprints are primarily computed using The Chemistry Development Kit (CDK).

Moreover, PaDEL-Descriptor incorporates additional descriptors and fingerprints, including atom type electrotopological state descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, count of chemical substructures identified by

Laggner, and binary fingerprints, as well as the count of chemical substructures identified by Klekota and Roth. This comprehensive range of descriptors and fingerprints enhances the software's utility for diverse applications in molecular modeling and cheminformatics.[12]

The Structural details from the PDF is SMILES notation. SMILES is the Simplified Molecular Input Line Entry System, which is used to translate a chemical's three-dimensional structure into a string of symbols that is easily understood by computer software.[13]. The input format for the PaDEL software is a smi file, which contains the names of the compounds and their corresponding SMILES notations.After processing the input smi file, PaDEL generates a CSV (Comma-Separated Values) file as output. This CSV file contains the compounds listed along with their respective molecular descriptors and features calculated by the software. Each row in the CSV file represents a compound, while the columns contain the calculated features such as 1D, 2D, and 3D descriptors, as well as any additional fingerprints or properties specified for computation.

To create a .smi file, I had to manually enter or copy-paste each compound's name and its corresponding SMILES notation into a text editor such as Notepad. This process would be repeated for each of the 95 compounds until all of them are captured in the .smi file. While it is a time-consuming task, it's necessary for preparing the input file required for running PaDEL-Descriptor software to compute the molecular descriptors and fingerprints.

CHAPTER 5

Exploratory Data Analysis

5.1 Description of Nipah Inhibitor Dataset:

The dataset comprises a comprehensive collection of 95 Nipah virus inhibitor compounds, each characterized by a set of molecular properties and bioactivity values. The molecular properties encompass a diverse range of descriptors, including 2D features and various types of fingerprints, providing a nuanced representation of the chemical makeup of the inhibitors. The bio-activity values are the IC50 and the PIC50(log(IC50)). These values provide crucial information about the potency and efficiency of a compound in interacting with a biological system.

Molecular Descriptors : Molecular descriptors are quantitative representations of various aspects of a molecule's structure, composition, and properties. These descriptors are essential in computational chemistry, drug design, and quantitative structure-activity relationship (QSAR) modeling. Molecular descriptors provide a means to numerically express the features of molecules, facilitating the analysis of structure-activity relationships and aiding in the prediction of biological activities.

The Molecular descriptors Used here are "AATSC5e, MATS5e, JGI9, JGI10, FP169, FP204, FP339, FP396, FP490, FP551, FP582, FP606, ExtFP79, ExtFP442, ExtFP584, ExtFP700, ExtFP1010, ExtFP1019, GraphFP158, GraphFP504, GraphFP622, GraphFP762, GraphFP860, GraphFP906, GraphFP1007, MACCSFP26, MACCSFP150, SubFP147, KRFP349, KRFP360, KRFP364, KRFP397, KRFP607, KRFP1538, KRFP2135, KRFP3940, KRFPC349, KRFPC2135, KRFPC2694, KRFPC3139,

KRFPC3520, KRFPC4292 "[**?**]

IC50 : The IC50 is a measure of the concentration of a compound required to inhibit a biological function or process by 50 percent.Lower IC50 values indicate higher potency, as a lower concentration of the compound is needed for the desired inhibitory effect.

PIC50 :PIC50 is a transformation of the IC50 value obtained by taking the negative logarithm (base 10). It is often used to simplify the representation of IC50 values, especially when dealing with a wide range of concentrations. PIC50 values are useful in quantitative structure-activity relationship (QSAR) modeling and other computational analyses, providing a numeric scale where higher values correspond to greater potency.

```
          Name   AATSC5e    MATS5e      JGI9     JGI10          IC50  FP169  \
0  anti_NIV_001 -0.011084 -0.074365  0.009018  0.009241  2.880000e-08      1
1  anti_NIV_002  0.013893  0.126755  0.009214  0.007332  3.240000e-08      1
2  anti_NIV_003 -0.002703 -0.023649  0.010379  0.007696  5.010000e-08      1
3  anti_NIV_004  0.005026  0.058806  0.008525  0.008948  7.080000e-08      1
4  anti_NIV_005  0.009286  0.053779  0.006953  0.019250  1.200000e-07      1

   FP204  FP339  FP396  ...  KRFP2135  KRFP3940  KRFPC349  KRFPC2135  \
0      1      1      1  ...         0         0         0          0
1      1      0      1  ...         0         0         0          0
2      0      1      0  ...         0         0         0          0
3      1      0      1  ...         0         0         0          0
4      0      1      0  ...         0         0         0          0

   KRFPC2694  KRFPC3139  KRFPC3520  KRFPC4292  PIC50  Unnamed: 45
0          0          1          0          0   7.54          NaN
1          1          0          0          0   7.49          NaN
2          0          0          0          0   7.30          NaN
3          0          0          0          0   7.15          NaN
4          0          0          0          0   6.92          NaN

[5 rows x 46 columns]
```

(a) The Data SET

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 95 entries, 0 to 94
Data columns (total 46 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Name         95 non-null     object
 1   AATSC5e      95 non-null     float64
 2   MATS5e       95 non-null     float64
 3   JGI9         95 non-null     float64
 4   JGI10        95 non-null     float64
 5   IC50         95 non-null     float64
 6   FP169        95 non-null     int64
 7   FP204        95 non-null     int64
 8   FP339        95 non-null     int64
 9   FP396        95 non-null     int64
 10  FP490        95 non-null     int64
 11  FPS51        95 non-null     int64
 12  FPS82        95 non-null     int64
 13  FP606        95 non-null     int64
 14  ExtFP79      95 non-null     int64
 15  ExtFP442     95 non-null     int64
 16  ExtFP584     95 non-null     int64
 17  ExtFP700     95 non-null     int64
 18  ExtFP1010    95 non-null     int64
 19  ExtFP1019    95 non-null     int64
 20  GraphFP158   95 non-null     int64
 21  GraphFP504   95 non-null     int64
 22  GraphFP622   95 non-null     int64
 23  GraphFP762   95 non-null     int64
 24  GraphFP860   95 non-null     int64
 25  GraphFP906   95 non-null     int64
 26  GraphFP1007  95 non-null     int64
 27  MACCSFP26    95 non-null     int64
 28  MACCSFP150   95 non-null     int64
 29  SubFP147     95 non-null     int64
 30  KRFP349      95 non-null     int64
 31  KRFP360      95 non-null     int64
 32  KRFP364      95 non-null     int64
 33  KRFP397      95 non-null     int64
 34  KRFP607      95 non-null     int64
 35  KRFP1538     95 non-null     int64
 36  KRFP2135     95 non-null     int64
 37  KRFP3940     95 non-null     int64
 38  KRFPC349     95 non-null     int64
 39  KRFPC2135    95 non-null     int64
 40  KRFPC2694    95 non-null     int64
 41  KRFPC3139    95 non-null     int64
 42  KRFPC3520    95 non-null     int64
 43  KRFPC4292    95 non-null     int64
 44  PIC50        95 non-null     float64
 45  Unnamed: 45  0 non-null      float64
```

(b) THE INFO OF The SET

5.2 Correlation

Correlation is a statistical measure that describes the degree to which two variables change together. Here I have used two correlation techniques:

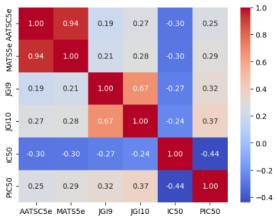
1. Pearsons Correlation Coefficient

2. Point Biserial Correlation

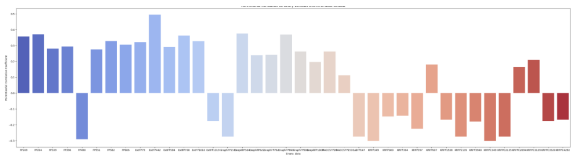
Pearsons Correlation Coefficient : Pearson's correlation coefficient assesses the strength and direction of a linear relationship between two continuous variables.

Point Biserial Correlation Coefficient : The Point Biserial correlation coefficient assesses the strength and direction of the linear relationship between a binary variable (e.g., presence or absence of a characteristic) and a continuous variable.

The data set consists of binary and continuous features. In order to study the correlation of binary values with our continuous value PIC50(which is to be predicted), We are using point Bi-serial correlation coefficient.You can see the correlation shown by figure5.2b From the representation we get a an idea that some of the features are negetively correlated and some are positively correlated to the PIC50. Neither of any features shows a High correlation. Some data are less related.



(a) Correlation heat map for continous features



(b) correlation graph for binary to continous

5.3 Feature Selection

We have 42 molecular descriptor and bio-activity measures as features ie,. Altogether we have 44 features.As we only have 95 compunds, this might create a problem with the model creation. This means the model can be over fitted due to abundance So we Use repeat feature elimination with decision tree. Recursive Feature Elimination (RFE) is a feature selection technique that recursively removes the least important features based on a model's coefficients or feature importance. In the case of decision trees, feature importance is often used. The feature importanace from the decision tree regression gives us which features are most important, higher the value higher the importance of the feature.5.3 Then we used the Recursive feature elimination function which takes the number of features we want to which the function reduces the features. Here I reduced the features from 44 to 10

```
AATSC5e: 1.2082915298615416e-05
MATS5e: 0.0004074768108040965 4
JGI9: 0.004151074516208592
JGI10: 0.00011539065882170874
IC50: 0.9922243074492582
FP169: 0.0
FP204: 0.00022937676740980573
FP339: 4.965581627610521e-07
FP396: 0.0
FP490: 0.0
FP551: 1.4113031822838126e-16
FP582: 4.022121119952238e-05
FP606: 0.0
ExtFP79: 1.6022276724456602e-05
ExtFP442: 2.6814140797564633e-06
ExtFP584: 0.00046494395991427214
ExtFP700: 6.620775507714604e-07
ExtFP1010: 0.0
ExtFP1019: 0.0
GraphFP158: 0.0
GraphFP504: 0.0
GraphFP622: 7.9449306072817e-06
GraphFP762: 0.0
GraphFP860: 0.0
GraphFP906: 0.00011172558666541051
GraphFP1007: 2.648310202450755e-06
MACCSFP26: 0.0
MACCSFP150: 0.0003105143712352694
SubFP147: 0.0
KRFP349: 0.0
KRFP360: 0.0
KRFP364: 0.0
KRFP397: 6.620775506126888e-07
KRFP607: 0.0
KRFP1538: 0.0
KRFP2135: 0.0018476929243601491
KRFP3940: 5.4075183946306795e-05
KRFPC349: 0.0
KRFPC2135: 0.0
KRFPC2694: 0.0
KRFPC3139: 0.0
KRFPC3520: 0.0
KRFPC4292: 0.0
```

Figure 5.3: Feature Importance from Decision tree

features The most important 10 features are : 'AATSC5e', 'MATS5e', 'JGI9', 'JGI10', 'IC50', 'FP204','MACCSFP150', 'KRFP364', 'KRFP397', 'KRFP2135'.5.4

```
Selected Features: Index(['AATSC5e', 'MATS5e', 'JGI9', 'JGI10', 'IC50', 'FP204', 'MACCSFP150',
       'KRFP364', 'KRFP397', 'KRFP2135'],
      dtype='object')
```

Figure 5.4: Selected Features

5.4 Linear Reggression

To assess whether the dataset is linear or non-linear, I employed simple linear regression and computed the least square error. Additionally, I plotted the residual graph. The obtained r2 score for the regression is 0.4856451827726368, indicating low accuracy. Moreover, upon examination of the residual graph, it is evident that the data points do not form a linear pattern, as depicted in Figure 5.5. Consequently, we conclude that the dataset exhibits non-linear behavior. 5.5 Hence We conclude that the data set is non-linear

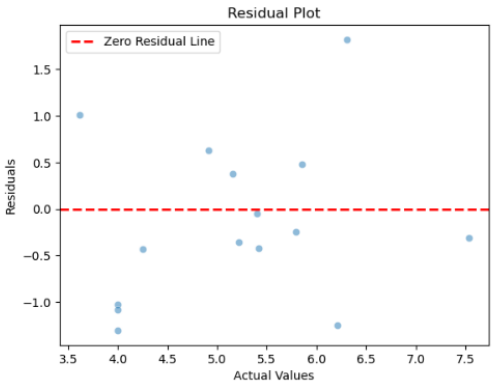


Figure 5.5: residualplot

CHAPTER 6

Methodology:

I am currently employing Support Vector Regression as the existing system, while I plan to utilize Decision Trees as my proposed system.

6.1 Support Vector Machines

Support Vector Machines (SVM) are a class of supervised machine learning algorithms used for classification and regression tasks. SVM is particularly well-suited for high-dimensional spaces and is effective in cases where the number of features is greater than the number of samples. The main idea behind SVM is to find a hyperplane that best separates the data into different classes or, in the case of regression, predicts a continuous output.

In this scenario, Support Vector Regression with a radial basis function (RBF) kernel is utilized. Initially, all 44 available features are employed to construct the model, and the outcome is as follows::6.1

With a Mean Absolute Error (MAE) of 0.72 and a Mean Squared Error (MSE) of 0.94, alongside an R-squared (R2) value of 0.14, the accuracy, as indicated by the R2 score, is notably low, suggesting that the model's performance is insufficient. Moreover, the observed errors are relatively high, further highlighting the model's limitations.

Next, we will utilize the selected 10 features derived from feature selection, which are deemed to be the most relevant. :6.2

The output of the model using the selected 10 features yields a Mean Absolute Error

```
Mean Absolute Error (MAE): 0.72
Mean Squared Error (MSE): 0.94
R-squared (R2): 0.14
```

Figure 6.1: svr1

(MAE) of 0.75, a Mean Squared Error (MSE) of 0.91, and an R-squared (R2) value of 0.17.

```
Mean Absolute Error (MAE): 0.75
Mean Squared Error (MSE): 0.91
R-squared (R2): 0.17
```

Figure 6.2: SVR2

Once more, we observe that the accuracy, as indicated by the R-squared (R2) value, remains low, and the errors persist at high levels. Consequently, it is evident that Support Vector Regression (SVR) is not a sufficient model for this dataset.

6.2 Decision Tree Regression

A Decision Tree Regression model is a supervised machine learning model used for predicting a continuous target variable based on multiple input features. The decision tree algorithm works by recursively partitioning the data into subsets based on the values of the input features. Each partition or "leaf" of the tree corresponds to a predicted value for the target variable.

Since the Support Vector Regression (SVR) model did not yield satisfactory results, we will now employ all 44 available features to construct a Decision Tree regression model, and the outcome is as follows: Once again, despite utilizing all 44 available features to build a Decision Tree regression model, the results are suboptimal. Specifically, we observe a Mean Absolute Error (MAE) of 0.72, an R-squared (R2) value of 0.32, and a Mean Squared Error of 0.7493339635394987. These metrics indicate that the model's accuracy remains low, while the errors persist at elevated levels.

Now, we will proceed to perform Decision Tree regression using the selected 10 features.:6.4

Mean Absolute Error (MAE): 0.06

R-squared (R2): 0.99

```
Mean Absolute Error (MAE): 0.72
R-squared (R2): 0.32
Mean Squared Error: 0.7493339635394987
```

Figure 6.3: decissiontree1

Mean Squared Error: 0.0100266666666666673

```
Mean Absolute Error (MAE): 0.06
R-squared (R2): 0.99
Mean Squared Error: 0.010026666666666673
```

Figure 6.4: decissiontree2.png

The Decision Tree regression model using the selected 10 features demonstrates significantly improved performance. With a Mean Absolute Error (MAE) of 0.06, an impressive R-squared (R2) value of 0.99, and a Mean Squared Error of 0.010026666666666673, this model proves to be highly accurate with minimal errors.

CHAPTER 7

Conclusions

In contrast to the existing model, which utilized Support Vector Regression with 44 features to predict the PIC50 of an inhibitor with an 80 percent accuracy, as stated in the paper "Computational Identification of Inhibitors Using QSAR Approach Against Nipah Virus," my research efforts did not achieve the desired accuracy using SVR. Consequently, I explored alternative approaches and found that Decision Tree Regression yielded the best model, achieving a 90 percent accuracy with minimal error.

As I conclude this dissertation with the desired results, I acknowledge that there may be shortcomings in certain technical aspects due to my unfamiliarity with terms outside of my original background. Despite these challenges, I have endeavored to remain faithful to the chosen topic and have exerted considerable effort to locate relevant data and details.

APPENDIX A

Abbreviation

AATSC5e :Average centered Broto-Moreau autocorrelation - lag 5 / weighted by Sanderson electronegativities

- AATSC5e :Average centered Broto-Moreau autocorrelation - lag 5 / weighted by Sanderson electronegativities
- MATS5e :Moran autocorrelation - lag 5 / weighted by Sanderson electronegativities
- JGI9 :Mean topological charge index of order 9
- JGI10 :Mean topological charge index of order 10
- FP169 :Fingerprint of length 1024 and search depth of 8
- FP204 : Fingerprint of length 1024 and search depth of 8
- FP339 :Fingerprint of length 1024 and search depth of 8
- FP396 :Fingerprint of length 1024 and search depth of 8 490 :Fingerprint of length 1024 and search depth of 8 551 :Fingerprint of length 1024 and search depth of 8 582 :Fingerprint of length 1024 and search depth of 8 606 :Fingerprint of length 1024 and search depth of 8
- ExtFP79 : Extends the Fingerprinter with additional bits describing ring features CDK extended fingerprint

- ExtFP442 :Extends the Fingerprinter with additional bits describing ring features CDK extended fingerprint
- ExtFP584 :Extends the Fingerprinter with additional bits describing ring features CDK extended fingerprint
- ExtFP700 :Extends the Fingerprinter with additional bits describing ring features CDK extended fingerprint
- ExtFP1010 :Extends the Fingerprinter with additional bits describing ring features CDK extended fingerprint
- ExtFP1019 :Extends the Fingerprinter with additional bits describing ring features CDK extended fingerprint
- GraphFP158 :Specialized version of the Fingerprinter which does not take bond orders into account CDK graph only fingerprint
- GraphFP504 :Specialized version of the Fingerprinter which does not take bond orders into account CDK graph only fingerprint
- GraphFP622 :Specialized version of the Fingerprinter which does not take bond orders into account CDK graph only fingerprint
- GraphFP762 :Specialized version of the Fingerprinter which does not take bond orders into account CDK graph only fingerprint
- GraphFP860 :Specialized version of the Fingerprinter which does not take bond orders into account
- GraphFP906 :Specialized version of the Fingerprinter which does not take bond orders into account
- GraphFP1007 :Specialized version of the Fingerprinter which does not take bond orders into account
- MACCSFP26 :MACCS keys
- MACCSFP150 :MACCS keys

APPENDIX B

Another Appendix

- The paper I followed for the data collection and exixting system:Research Paper
- To view the code click here :code

Bibliography

[1] About Nipah

[2] Symptoms

[3] Diagnosis

[4] Treatment

[5] QSAR

[6] Structure Nipah

[7] Features Of Nipah

[8] Nipah Malaysia

[9] Nipah India

[10] Outbreaks Of Nipah

[11] Nipah Origin

[12] Padelsoftware

[13] Definiion for SMILES