Bioinformatics and Machine Learning in Prevention, Detection and Treatment of HIV/AIDS

By

Wakaya Brian
17146003

A thesis submitted to the Department of Pharmacy in partial fulfillment of the requirements for the degree of
Bachelor of Pharmacy (Hons.)

Department of Pharmacy
Brac University
July 2021

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I have acknowledged all main sources of help.

Wakyaya Brian
17146003

Approval

The project titled "Bioinformatics and Machine Learning in prevention, detection and treatment of HIV/AIDS" submitted by Wakyaya Brian (17146003) of Spring, 2017 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Pharmacy (Hons) on 18-07-2021.

Examining Committee:

Supervisor:
(Member)

Mohammad Kawsar Sharif Siam
Senior Lecturer, Department of Pharmacy
Brac University

Program Coordinator:
(Member)

Dr. Hasina Yasmin
Professor, Department of Pharmacy
Brac University

Departmental Head:
(Chair)

Dr. Eva Rahman Kabir
Professor, Department of Pharmacy
Brac University

Ethics Statement

This study does not involve any human or animal trial.

Abstract

As the acquired immunodeficiency syndrome (AIDS) pandemic continues to be a major health crisis of global concern, new strategies in the management and treatment of the disease is being explored. This project titled “Bioinformatics and Machine Learning in Prevention, Detection and Treatment of HIV/AIDS” discusses the existing processes and procedures within which computational (Bioinformatics and Machine Learning) techniques and approaches that can be potentially applied in the global fight to end the HIV/AIDS pandemic e.g. homology modeling, virtual screening, Quantity Structural Activity Relationship (QSAR) and molecular docking. It further reviews the bioinformatics and various machine learning techniques such as Support Vector Machine (SVM), Decision Tree Algorithms and Artificial Neural Networks (ANNs) that are incorporated into computational tools (Computer-Aided Drug Design-CADD) to accelerate the process of drug design and development of anti-HIV drugs by reviewing distinguished journals, articles and databases. Attempts were taken to identify gaps within the existing literature.

Keywords: Bioinformatics; Machine learning; Computer Aided Drug Design (CADD); HIV/AIDS

Dedication

I dedicate this work to my family, friends and teachers

Acknowledgement

I am extremely grateful to God Almighty for the precious gift of life, good health and wisdom that HE has given me that has enabled me to successfully accomplish my thesis project as part of the requirements of my bachelor's degree program at Brac University.

I thank my parents, Mr. Wakyaya Francis and Mrs. Mutonyi Allen together with my siblings, Wakyaya Aron, Wakyaya Godwin, Wakyaya Joel, Kakai Suzan and Kakai Gift for the enormous love, care, support and prayers they have always showered me in times of hardships.

With special thanks, I would like to extend my sincere gratitude towards the Department of Pharmacy community starting with our respected Prof. Dr. Eva Rahman Kabir Chairperson, Department of Pharmacy, Brac University; Prof. Dr. Hasina Yasmin, Academic program coordinator Department of Pharmacy, Brac University; Nashrah Mustafa, Teaching Assistant, Department of Pharmacy, Brac University; all the Faculties of the department of Pharmacy and their teaching assistants not forgetting my fellow students, both my seniors and juniors that have played a very big role towards my both professional and personal development. I am forever grateful for everything that you have done for me.

Now I take this chance to give my sincere heart felt appreciation to my thesis supervisor, Mr. Mohammad Kawsar Sharif Siam, Senior Lecturer, Department of Pharmacy, Brac University for the great academic experience that you have made me attain. The criticism, words of encouragement and the guidance you gave me during my thesis have shaped and sharpened my academic career life in terms research to a greater level.

Last but not least, I want to thank my Ugandan friends and fellow students that have always been by my side and given their endless support and comfort.

Table of Contents

List of Tables

List of Figures

List of Acronyms

HIV	Human Immunodeficiency Virus
AIDS	Acquired Immunodeficiency Syndrome
CADD	Computer-Aided Drug Discovery
QSAR	Quantitative Structure Activity Relationship
SVM	Support Vector Machine
ANN	Artificial Neural Networks
NRTIs	Nucleoside reverse transcriptase inhibitors
CXCR4	Chemokine co-receptor type 4
CCR5	Chemokine co-receptor type 5
cART	Combinatorial antiretroviral therapy
BLAST	Basic Local Alignment Search Tool

Glossary

Machine Learning	The process of generating the algorithms based on the previous inputs and past experiences in terms of following instructions to give outputs.
Bioinformatics	The use of computational tools to analyze, organize, comprehend, visualize, and store data on biological macromolecules

Chapter 1

Introduction

Human Immunodeficiency Virus (HIV) refers to an infectious agent which is a causative agent of Acquired Immunodeficiency Syndrome (AIDS) (Worachartcheewan et al., 2018). HIV/AIDS pandemic that started in the early 1980s continues to be a global public health problem (Ghosh et al., 2016). However, it has now become a manageable chronic health disease since the HIV/AIDS patients have the access to treatment and this has increased their lifespan and decreased the mortality rate as compared to the previous situation. The HIV virus comprises of two major variants- HIV-1 and HIV-2. The first variant causes HIV infections in many different parts of the world, whereas the latter is mostly confined to West Africa with its transmission not as fast as the former (Santos et al., 2015). The Human Immunodeficiency Virus (HIV) is commonly spread by bodily fluids such as blood, vaginal fluid, sperm, breast milk, and rectal fluid (Kirchmair et al., 2012).

Currently, the HIV virus has infected an incredibly huge population of over 78 million individuals, with 1.5 million new infections in 2020 leaving almost 38 million people living with the illness. The pandemic has claimed over 35 million people's lives by 2020 with around 0.69 million new deaths in 2020. About 27.4 million people (68% of adults aged between 15 and 49 years and 53% of children below the age of 15) are reported to be living with HIV infection all over the world. These patients are receiving treatment. However, the number of HIV/AIDS infection cases differs from one country to another or one region to another. The African region is the most affected region contributing to about 70% of total world infections (Ghosh et al., 2016). It is seconded by the South or South-East Asia and followed by United States of America (USA) and Europe. Right from the early 1980s when the pandemic started, the HIV/AIDS-related deaths have risen to over 38 million according to

the WHO (World Health Organization) and UNAIDS (The Joint United Nations Programme on HIV/AIDS) (World Health Organization, 2020).

1.1 The Lifecycle of the Human Immunodeficiency Virus (HIV)

The Human Immunodeficiency Virus (HIV) goes through at least seven to nine distinct phases during its lifespan. Binding, fusion, reverse transcription, integration, replication, assembly, and budding are some of the processes that take place in this lifecycle. (Figure 1). In order to develop a proper precise and effectively efficient anti-HIV/AIDS diagnostic tools and therapeutic agents, a detailed understanding of HIV/AIDS infection cycle is necessary. This is because most of the antiviral therapeutic products or agents (molecules) used in combating the virus are specifically intended to target one or more phases of this HIV infection cycle. There are about six types of anti-HIV medicines that have been effective against HIV to date. These include Fusion inhibitors that target the fusion step of the infection cycle, Reverse transcriptase inhibitors such as Nucleoside reverse transcriptase inhibitors (NRTIs) and Non-nucleoside reverse transcriptase inhibitors (NNRTIs) that inhibit the virus's reverse transcription process by targeting reverse transcriptase enzyme, Chemokine co-receptor (CCR5 and CXCR4) antagonists such as maraviroc, Protease inhibitors (PIs), and Integrase inhibitors. These drugs primarily target around 3-4 stages of the HIV lifecycle and are frequently used in combination to offer a better synergistic effect, commonly known as combination antiretroviral treatment (cART). Unfortunately, these numerous combination medications can have a variety of adverse effects and can even lead to HIV-related comorbidities in certain situations (Bala et al., 2018).

1.2 Steps of the life-cycle of the Human Immunodeficiency Virus (HIV)

The steps in the life-cycle (Figure 1) of the human immunodeficiency virus are as follows:

1. Binding and Entry: The HIV virus attaches itself to a CD4 molecule as well as to one sort of co-receptor (either CCR5 or CXCR4). On the cell surface, receptor molecules are abundant, and it is through them that the virus attaches itself to and unites with the cell.
2. Fusion and Penetration: The virus unites with the cell and releases its contents.
3. Reverse Transcription: The reverse transcriptase enzyme converts viral RNA strands into double-stranded DNA by creating a "mirror image" of the RNA strands.
4. Integration: The integrase enzyme is responsible for inserting viral DNA into the cell's own DNA.
5. Replication/Transcription: Each time a cell divides, the viral DNA is read and long chains of proteins are produced by the infected cell.
6. Assembly: Sets of viral proteins chains form and assemble.
7. Budding: The immature virus makes its way out of the cell, dragging a piece of the cell membrane along with it. The protease enzyme begins to process the proteins in the newly formed virus when it is activated.
8. The immature virus is able to break away from the infected cell.
9. Maturation: The protease enzyme completes the process of cleaving HIV protein chains into separate components. These come together to form the viral core, which then produces a new virus that is capable of reproducing itself.

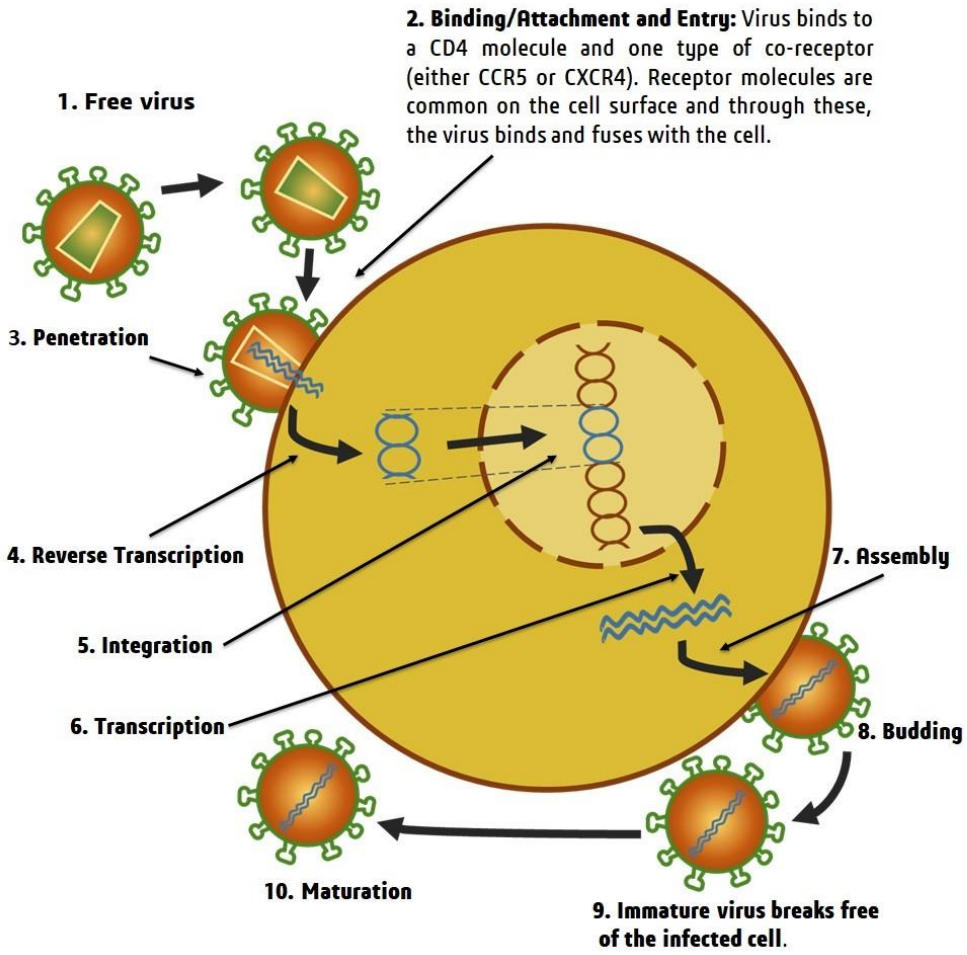

Figure 1: Stages of Human Immunodeficiency Virus (HIV) Lifecycle (Adapted from

(*HIV Drugs and the HIV Lifecycle | The Well Project*, 2019))

Chapter 2

Machine Learning

2.1 Background

Machine learning, a language/ technique named by Arthur Samuel in 1959, refers to the process of generating the algorithms based on the previous inputs and past experiences in terms of following instructions to give outputs (Sharma & Sharma, 2018). Machine learning techniques are cutting-edge technologies that are used to develop algorithms that improve recognition of patterns, classification, regression, and prediction by generating specialized models from current data. It is worth noting that machine learning and data mining are connected and the important area of research in both fields is pattern recognition. The most prevalent algorithms, such as classification, are often used to identify active and inactive substances, whereas regression techniques, in addition to prediction, are employed in the training and testing of continuous data. Then, using ensemble techniques such as bagging and boosting, one can make accurate and quick judgments. Cross-validation, on the other hand, is crucial in attaining the desired outcome. A wide range of machine learning approaches have been employed in drug discovery and development studies, including but not limited to QSAR (quantitative structure activity relationship) studies, virtual screening, and computational ADMET (Adsorption, Distribution, Metabolism, Excretion, and Toxicity) investigations. *In-silico*-QSARs normally get differentiated by the inclusion of a large number of chemical structural descriptors in conjunction with several linear and nonlinear optimization methodologies, as well as a strong emphasis on the validation of the models developed. Machine learning approaches are also employed in the identification of the best-suited models among the computational-based cheminformatics softwares like Modeller,

chem. sketch, DRAGON, MOE, VMD, and AUTODOCK, which are frequently used in the identification of targets and the discovery of hits (Dubey, 2018).

2.2 How it Works

Machine learning is a steadily and aggressively evolving field from laboratory, practical and clinical application in the medical sector. This is mostly due to the ever-increasing amounts of sociological, epidemiological, clinical, genomic or other forms of data available for scientists and researchers to derive from, which is ostensibly time-consuming and exhausting, with seemingly limitless human mistakes in the final conclusions. As a result, different machine learning algorithms are projected to improve the speed and accuracy with which physicians and other healthcare practitioners predict and as well make decisions, resulting in lower costs, more time saved, and better drug treatments and patient health (Shapshak et al., 2019). Generally, the machine learning methods function on basis of the viral genome sequence inputs which in turn after their performance give the outputs as the susceptibility or the virus's resistance to a particular medication. The general approach is that the algorithm examines the data set used in the training for both input and output and then performs the necessary calculations with the help of other algorithms like statistical and classification algorithms. Using these data, the learning and construction of a computational model is performed, which may then be utilized to determine the required output. Therefore, if the model is updated to include a new viral genotype, the projected resistance value derived by the model may be interpreted as drug susceptibility, drug resistance, or something in between for the drug under consideration. The several strategies that were used to construct the computational model can be divided into three categories: statistical learning approaches, classification methods, and molecular structure-based methods (Shapshak et al., 2019).

2.3 Methods of Machine Learning

Machine Learning is majorly classified into three broad categories in addition to other different machine learning algorithms. These include supervised machine learning, unsupervised machine learning and semi supervised machine learning. These are discussed in the following sections.

2.3.1 Supervised Machine Learning

A common type of machine learning approach is supervised learning, which is used to develop a predictor or model using samples with known class labels (referred to as training data) in order to predict the class of a new sample using the predictor or model (Noorbakhsh et al., 2019). This approach involves predicting a goal result variable (dependent variable) from a set of predictors (independent variables). The method works by making decisions basing on the given inputs and their desired outputs are obtained. In this method, particular set of inputs will give corresponding output to detect errors and the model is trained until it reaches the appropriate degree of precision and accuracy given the training data. Regression Analysis, Decision Tree, Random Forest, K-Nearest Neighbour, and Regression Analysis or Logistic Regression are examples of supervised learning algorithms (Figure 2) (Sharma & Sharma, 2018).

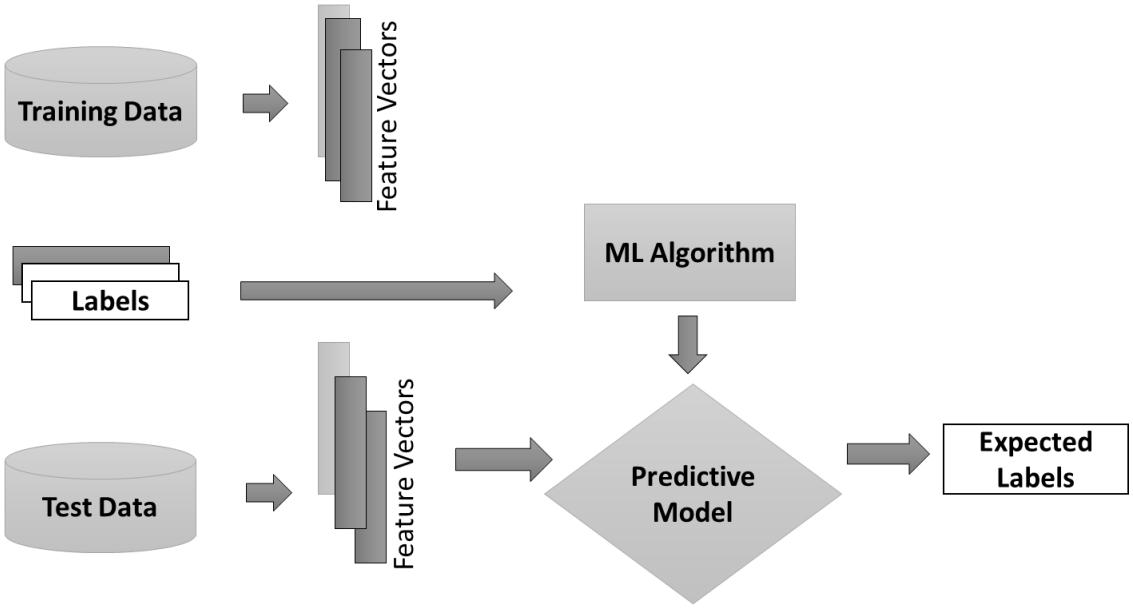

Figure 2: Supervised Machine Learning Model (Adapted from (Sharma & Sharma, 2018))

Supervised Machine Learning can be classified in the following ways.

2.3.1.1 Decision Tree Algorithm

The decision tree (DT) algorithm is a type of supervised machine learning method that is used for classification and regression problems in data mining. Each internal node represents a feature (or characteristic), each branch represents a decision rule, and each leaf node is the conclusion or result of the decision tree. In a decision tree, the root node is the leaf node, which is the node that is located at the very top of the tree. Therefore, decision tree look like flowcharts. According to this technique, a given set of data or population under investigation can be separated into many homogeneous sets. The most significant attributes/independent variables are frequently used to separate as many groups as possible in order to maximize the number of distinct groupings. It learns to segment the data based on the value of the attribute and splits the tree using a recursive partitioning strategy. It is a powerful tool. For example, when a decision tree is being constructed, it begins with a specific data set and gradually breaks it down into smaller and smaller subgroups (Shapshak et al., 2019). This flowchart-like structure of the decision tree is an essential tool in decision making. Decision trees are

easy to comprehend and evaluate because their visual representation is similar to that of a flowchart diagram, which closely reflects human level thinking. These are trees that classify instances by ordering them according to the values of their feature attributes. Each node in a decision tree represents a characteristic of an instance that needs to be classified, and each branch represents a possible value that the node could take on. Starting with the root node, instances are classified and ordered based on the values of their feature attributes (Muhammad & Yan, 2015). The Decision Tree Algorithm, a white box machine learning algorithm, as opposed to black box machine learning methods such as neural networks, is a sort of machine learning system that shares internal decision-making logic. Due to the fact that this information is not available in black box algorithms, the decision tree method's training period is significantly shorter and faster than that of the neural network approach. The quantity of records and characteristics in the provided data are what determine the temporal complexity of decision trees, as is the amount of records and characteristics. Decision trees are non-parametric or distribution-free approaches that do not rely on probability distribution assumptions, such as those used in statistical analysis. Decision trees are capable of dealing with large amounts of data while maintaining excellent accuracy.

2.3.1.2 Artificial Neural Network (ANN) Algorithm

Artificial Neural Network (ANN) is a machine learning technology which performs valuable tasks quickly by mimicking the human brain. It's a mathematical model that uses repeated test and error trials to discover the relationship between experimental elements (input) and response (output) values. (Metwally & Hathout, 2015). It can learn and identify nonlinear functions and patterns, and can hopefully fit them into a broad range of applications in complicated systems, implying that this approach may be used to build a platform on which various models can be molded and created (Du et al., 2017). The Artificial Neural Network algorithm was created to simulate brain structure and function, and it was an inspiration of

human information processing through the interplay of hundreds of millions and even billions of neurons linked to each other. Because neurons have a certain topology, they create neural networks when they are linked to one another. It is referred to as a feed forward network in an ANN. Furthermore, Multiplayer Perceptrons (MLP), Radial Basis Function (RBF) networks, and Kohonen's Self Organizing Maps (Kohonen's SOM) are all examples of feed forward networks in artificial neural networks (ANNs). It is mostly used in chemical categorization, QSAR investigations, primary virtual screening of compounds, identification of probable therapeutic target locations, and pattern identification of structural and functional characteristics of proteins (Dubey, 2018).

2.3.1.3 Support Vector Machine

The Support Vector Machine (SVM) basically is another type of supervised machine learning algorithm that is commonly employed in case of classification, regression, and outlier identification. It's a kernel-based method for binary data categorization and regression that Vapnik and co-workers initially proposed in 1995 (Du et al., 2017). This type of Supervised Machine Learning is important and highly preferred due to its effective high dimensional space, and its capacity to employ a subset of training points (called support vectors) in the decision function, resulting in memory efficiency, as well as its adaptability in that it holds several kernel functions that may be provided for the decision function (Muhammad & Yan, 2015). This approach (SVM), is one of the most promising machine learning approaches for constructing predictive QSAR models using chemical descriptors at various scales for drug discovery and development, as well as dealing with high-dimensional information (Lin et al., 2020). Thus is not, however, a probabilistic approach, and it is unable to analyze the accuracy of its predictions, in addition to processing big databases (Varnek & Baskin, 2012).

2.3.1.4 Bayesian Classifier

Bayesian Classifier is a statistical technique based on the Bayes rule for conditional probability, which categorizes compounds based on the equal and independent contributions of their characteristics (Du et al., 2017). Naive Bayes is a probabilistic model and classification approach that is based on Bayes' theorem and assumes predictor independence (Lo et al., 2018). According to a prior probability distribution (priors) indicating the relative proportions of labels in training sets, this method predicts and approximates the likelihood that a particular item of data would be correctly assigned to a specific label for a given label. If many labels are presented, the probability associated with each label is conditionally independent of the probability associated with the other labels (Lo et al., 2018). In its most basic form, a Naive Bayes classifier claims that the presence of one feature in a class is unrelated to the presence of any other feature in the class. For example, an apple is a red fruit with a spherical shape and a diameter of approximately 3 inches that is approximately 3 inches in diameter. No matter how dependent each of these traits is on the others, or how dependent they are on the presence of other characteristics, a naive Bayes classifier would consider each attribute to contribute to the likelihood that this fruit is an apple on its own.

Bayesian classifier models are a collection of highly quick and simple classification methods that are frequently acceptable for very large datasets. They wind up being highly effective as a quick-and-dirty baseline for a classification issue since they are so fast and have so few adjustable parameters. That said, they have several advantages. For example, this Bayesian Classifier model is uncomplicated to create and is particularly successful when dealing with large data sets. Because of its simplicity, it is widely regarded as outperforming even the most sophisticated classification algorithms.

2.3.1.5 Logistic Regression Analysis

Logistic Regression is a technique for predicting given outcomes. It's a classification algorithm, that is used to estimate discrete values (like 0/1, yes/no, and true/false) basing on a given set of independent variables(s). In simple terms, it fits data to a logit function to estimate the likelihood of an event occurring, thus, being called logit regression. Its output values are between 0 and 1 since it forecasts probability (as expected).

Logistic regression is an important classification technique. Linear classifiers are members of the same family as polynomial and linear regression, and it is included in this group. Logistic regression is a straightforward and quick approach of data analysis, and the results are straightforward to comprehend. Despite the fact that it is largely a binary classification strategy, it can also be utilized to handle problems involving several classes. It does, however, have a lower accuracy than these other machine learning techniques, at least to a certain degree.

2.3.1.6 K-Nearest (kNN) Neighbour:

The k-nearest neighbor technique is among the most basic, flexible, and straightforward method of all machine learning algorithms, and it's commonly employed in conjunction with other feature space, classification, and regression algorithms based on example learning (Lin et al., 2020). It's a nonparametric approach for classifying objects based on the feature space's nearest training samples (Du et al., 2017). In kNN, the data containing labeled and unlabeled nodes is represented in a high-dimensional feature space, and the labels from the closest nodes are transferred to the query using a majority-voting rule, where the parameter k indicates the number of closest neighbors who are participating in the voting process (Lo et al., 2018). However, this method or algorithm is computationally costly in general. Higher range variables might cause bias if variables are not standardized. It is extremely advised that

data be normalized on the same scale in order to achieve better results. The normalization range is typically between 0 and 1, with the exception of rare circumstances. KNN is more useless when dealing with large amounts of dimensional data. In such cases, it is necessary to lower the size of the component in order to improve performance. Additionally, addressing the issue of missing numbers will contribute in the refinement of our conclusions.

2.3.1.7 Random Forest Algorithm

Leo Breiman proposed Random Forest (RF) Algorithm as a machine learning method in the early 2000s. Random Forest algorithm is a supervised machine learning method that uses ensemble learning to build numerous decision trees based on training data and a majority-voting system similar to kNN to generate classification or regression predictions for fresh inputs (Lo et al., 2018). It is essentially a collection of randomly generated separate decision trees that work together to enhance classification or regression accuracy, and it employs the bootstrap sampling approach that is an upgraded form of bagging. Each tree is distinct from the others due to the randomization introduced in the RF method in two ways: one in the sample dataset used to develop the tree and the other in the subset of characteristics used for node splitting (Shapshak et al., 2019). It is applied in building classification models, for example the SARS-CoV-2 activity prediction classification models (Gaudêncio & Pereira, 2020).

2.3.2 Unsupervised Machine Learning

Unsupervised learning refers to a group of techniques that are used to infer underlying organization and connections between data without the usage of relevant labels. There is no objective or result variable to forecast or estimate in this method (Noorbakhsh et al., 2019). It is the problem of inferring a function to portray hidden structure from unlabeled data using machine learning (Dash et al., 2021). In this type of machine learning, the data is used

without distinctions between input or output variables (labels). The objective of unsupervised learning is to evaluate the data distribution, reduce data dimensionality, and discover hidden patterns in the data without making distinctions between input and output variables (labels). It's primarily utilized to divide data or set of information or people into distinct categories simply known as clustering (Varnek & Baskin, 2012).

2.3.2.1 Clustering

Cluster analysis is a technique that divides samples into different groups, with samples from the same group being more similar than samples from other groups. There are numerous clustering methods available, each with its own set of advantages and disadvantages. K-means, hierarchical clustering, and self-organizing maps are the most widely used clustering algorithms in biomedical research (Noorbakhsh et al., 2019).

2.3.3 Semi Supervised Machine Learning

Semi Supervised Machine Learning (SSML), is a kind of supervised and unsupervised learning technique. It is not supervised learning since it does not rely on a set of labeled training data, yet it is not unsupervised learning because it is based on the reward that the learning algorithm must optimize. The computer is trained to make particular judgments in semi-supervised learning, and the outputs are defined only for some cases, such as labelled and not unlabeled outputs (Varnek & Baskin, 2012). The learning strategy is to figure out the best actions to perform in various scenarios in order to maximize the reward, i.e., the machine or computer is exposed to an environment where it is constantly training itself via trial and error. This computer learns from its mistakes and strives to capture as much information as possible in order to make appropriate judgments or decision. The Semi Supervised Machine Learning technique has been effectively utilized to improving antiretroviral treatment and finding the optimum way to treating sepsis in human immunodeficiency virus positive

patients. Unlike supervised learning, which provides one-time predictions, the Semi Supervised Machine Learning algorithm's choice has an impact on both the patient's future health and treatment options (Noorbakhsh et al., 2019). Among these are the Inductive learning system (which learns categorization from training examples and utilizes induced rules for classifying new instances), and the Transductive machine learning system (which was designed to predict the objects in a specified test set). It is possible that the model's performance will increase significantly as a result of the specificity of the test set being taken into consideration throughout the learning process (Varnek & Baskin, 2012).

Chapter 3

Bioinformatics

Bioinformatics first named by Paulien Hogeweg, a Dutch system-biologist in 1970 (Aamer Mehmood, 2014). However, the origins of bioinformatics can be stretched back to the late 1950s and since Watson and Crick's revelation of the DNA structure (Levin et al., 2018). Bioinformatics is defined as the use of computational tools to analyze, organize, comprehend, visualize, store and retrieve data on biological macromolecules (Diniz & Canduri, 2017). Many of the difficulties in biology have now become challenges in computers as a result of the rise in data quantities. If faster and accurate findings are to be obtained, a strategy that can simplify the process by which computers can manage vast amounts of data and guide in comprehending the complex dynamics seen in nature is required (Yamamoto & Nakao, 2001). Bioinformatics generally aims at arranging and organizing these macromolecular information or data in such a format whereby scientists can access the desired information easily and also be able to contribute new entries that they possibly generate. This helps in managing large volumes of data to be analyzed which is the ultimate goal. And to do so, bioinformatics as a discipline enables the development of different bioinformatics tools and resources which and accelerate the process of analyzing such enormous amounts of data (Yamamoto & Nakao, 2001). Once these data are analyzed, they must be interpreted to decipher the biological information in details and this can be done by comparing the globally analyzed data with the previously studied data to discover both similarities and differences in such data so as to make informed decisions on how to use the obtained results. As a result, Bioinformatics is a branch of science in which several disciplines such as biology, computing, and information technology collaborate to organize, store, analyze, and manage large amounts of biological data, with contributions from

genetics, molecular biology, and biotechnology, among other fields (Soria-Guerra et al., 2015).

3.1 Bioinformatics approaches

Bioinformatics techniques are broadly divided into different categories, that is, genomics, proteomics and metabolomics among others.

3.1.1 Genomics

The study of how genes and genetic information are arranged within the genome and how this arrangement influences their function is known as genomics. That is, mapping, sequencing, and characterizing genomes to learn about the structure and function of a live organism's complete genome (Solanke & Tribhuvan, 2017).

3.1.2 Proteomics

Proteomics is the systematic isolation and characterization of proteins in biological systems at a high throughput. Disease processes appear at the protein level, and most medicines operate there as well (Prakash & Devangi, 2010). Proteomics is an attempt to characterize the biological state of cells and extracellular biological materials, as well as qualitative and quantitative changes in protein composition, under various situations in order to better understand biological processes. It is a stage of functional genomics in which the genome is described at the protein level (Solanke & Tribhuvan, 2017).

3.1.3 Metabolomics

Metabolomics is the study of metabolite levels in an organism and how they vary over time as a result of stimuli in a systematic and complete manner. Antibiotics, pigments, carbohydrates, fatty acids, and amino acids are examples of metabolites, which are

intermediates and products of metabolism. They are usually less than 1 kDa (Solanke & Tribhuvan, 2017).

3.2 Bioinformatics Tools

Some of the important bioinformatics tools have been discussed in the following sections.

3.2.1 BLAST Alignment Tool

The BLAST algorithm (Basic Local Alignment Search Tool) is a local alignment technique developed from the Smith-Waterman algorithm that displays the highest alignment score of two sequences (Diniz & Canduri, 2017). It is a comparison technique for amino acid sequences of various proteins or nucleotide sequences of nucleic acids. A BLAST search compares a query sequence to a database of sequences, allowing researchers to find library sequences that have a high degree of similarity to the query sequence (Eric et al., 2014).

3.2.2 FASTA Alignment Tools

FASTA was created in 1995 as a result of a FASTP enhancement. It's a similar method to BLAST for quickly aligning protein and DNA sequences (Issac & Raghava, 2005). BLAST, on the other hand, is quicker than FASTA because it simply looks for the most significant patterns in the sequences. For nucleic acid and protein sequences, the sensitivity (or accuracy) of BLAST and FASTA differs (Eric et al., 2014).

3.2.3 Dot Matrix Tools

DOT matrix is an old and very useful technique used to detect similar regions between two sequences. In addition, it is also possible to detect insertions and deletions (indels) between sequences. However, it is manual and subjective in nature thus posing as the major limitation of this technique (Issac & Raghava, 2005).

Chapter 4

Applications

4.1 *In silico* anti-HIV Drug Design and Discovery

Drug discovery refers to the process of developing novel drug compounds to treat various illnesses. It is essentially a procedure aimed at discovering a chemical that may be used to cure and treat a certain ailment or disease condition. Before the era of bioinformatics and computational biology, scientists used the traditional process of drug discovery to design and formulate new drug molecules with the basis of chemistry, pharmacology and clinical sciences through conventional experimentation procedures. This traditional drug discovery and development method is a time-consuming and resource-intensive procedure that necessitates partnerships from a wide variety of experts in disciplines such as medicinal chemistry, drug metabolism, animal pharmacology, and clinical research (Gao, 2016). The process normally involves Identification of candidates, chemical synthesis, characterization, screening, validation, optimization, and tests for therapeutic effectiveness and toxicity (Luxminarayan et al., 2019). This old method, on the other hand, is not only time-consuming, but also costly and inconvenient (Aamer Mehmood, 2014). For example, developing a single innovative medication molecule from discovery to commercialization for patient use takes around 10-15 years, and the total cost of the process is estimated to be between $800 million and $2 billion, with an average of $1 billion (Hassan et al., 2017). That is, the Tufts Center for the Study of Medication Development projected in 2014 that the cost of developing and commercializing a drug had risen by approximately 150 percent in the previous decade. The price tag has now been projected to reach approximately $2.6 billion dollars (Leelananda & Lindert, 2016). Due to these factors, together with the ever increasing demand for new drug products, scientists were forced to develop a shorter and quicker drug discovery process with

minimum room for error and wastage and in this way, they opted for the computational approaches of drug design that employs techniques like machine learning and bioinformatics and its tools to be applied in drug design and development process. These approaches have proved to be more effective as they are precise, time saving and economically efficient as compared to the traditional process (Nair, 2007). This new computational (Bioinformatics and Machine Learning) process can also be termed as Computer-Aided Drug Design (CADD). It is essentially the application of computer approaches to reduce the time spent searching for therapeutic compounds (Aamer Mehmood, 2014). Unlike the new computational methods, the traditional drug discovery process (Figure 3) conventionally involves a number of steps, starting from Target identification to Pre-clinical and clinical development. That is, Target identification (also known as Target selection), Target validation, lead discovery and optimization, medicinal chemistry and pre-clinical and clinical development (Luxminarayan et al., 2019). The first stage, target selection, is identifying a biological target inside the body and creating a model to simulate a real disease state that may be used as a site of therapeutic intervention. This enables the identification of a lead chemical or molecule with drug-like characteristics. The therapeutic molecule can then proceed to the preclinical phase of animal pharmacology and toxicology investigations when the lead chemical has been optimized with regard to the target protein of interest. Finally, if the medication passes all of the previous tests, it moves on to many rounds of human clinical trials, when it is given to human volunteers in at least three phases (Gao, 2016).

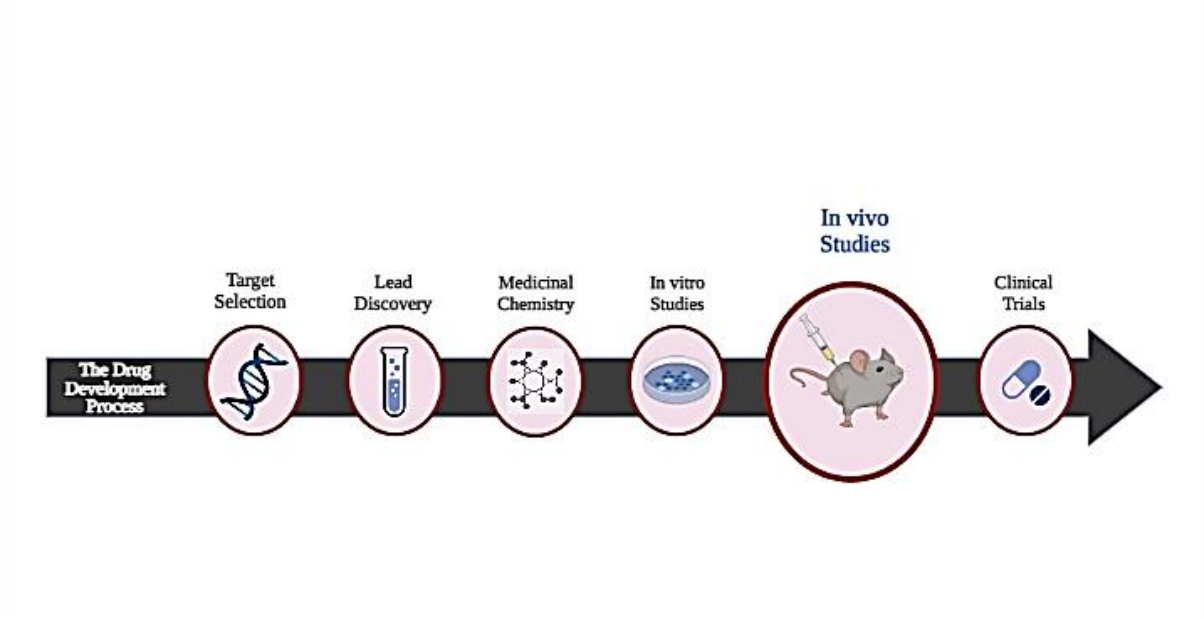

Figure 3: Stages of Traditional Drug Discovery and Development Process (Adapted from (Gao, 2016))

4.1.1 Computer-Aided Drug Design (CADD)

To efficiently and effectively expedite any drug discovery and development process of a novel therapeutic drug molecule process from the first step to the very last, rational drug design and development methods and tools have to be employed together with structural biology so as to register the desired success (Hassan et al., 2017). The latest advancements in the technology of drug design and or discovery has registered much success with results being significant reduction in costs and time taken for the entire process of drug research and development to be executed and such new drug put to market for patient use. Therefore, unlike the traditional drug discovery process characterized with poor drug development efficiency and high failure rate with respect to investments, the novel viable approaches to drug design, discovery and development has changed the game significantly (Xiang et al., 2012). These new approaches are basically computational methods that utilize the current advancements in biochemistry and structural biology, technologies (i.e. Machine Learning) and or bioinformatics (genomics and proteomics) to shorten and ease the whole drug design and discovery process (Xiang et al., 2012). The term Computer Aided Drug Design (CADD)

is a general term for these contemporary or computational drug design techniques. These computational methods include molecular modeling, chemoinformatics, bioinformatics, artificial intelligence, machine learning, and other approaches that are involved in *In-silico* drug design. For the reason that it is a computational strategy that generates a product while also documenting the design process, computer-aided drug design (CADD) (Figure 4) is a particularly successful approach for lowering the cost of drug development. The use of CADD can expedite or accelerate the manufacturing process by transforming complex and detailed diagrams of a product's materials, processes, tolerances, and dimensions (either two-dimensional or three-dimensional diagrams) into specific conventions for the product in question (either two-dimensional or three-dimensional diagrams), and such diagrams can be rotated to be viewed from an alternate perspective (Hassan et al., 2017). Most aspects and steps of the drug discovery and development process are covered by these novel computational methods, tools, and strategies. They have had a significant influence on rational drug design in general (Zheng et al., 2013). The bioinformatics tools, their applications, and various databases are all used in CADD techniques (Hassan Baig et al., 2016). Bioinformatics is a developing discipline that provides functionally predictive information mined from databases and experimental datasets utilizing a variety of computer-based techniques. It is increasingly being used in drug development procedures particularly computer-aided drug design (CADD) and development processes (Prakash & Devangi, 2010). Computer-aided drug design provides a foundation for developing new drugs. Because it allows for the simulation of interactions between active compounds and their targets, which may include receptors, enzymes, and transporters and physiologically-based pharmacokinetic simulations among others, it is possible to develop highly specific and efficacious molecules which are more potent and with good pharmacokinetics based on knowledge of the target structure, functional qualities, and action processes (Hung & Chen, 2014).

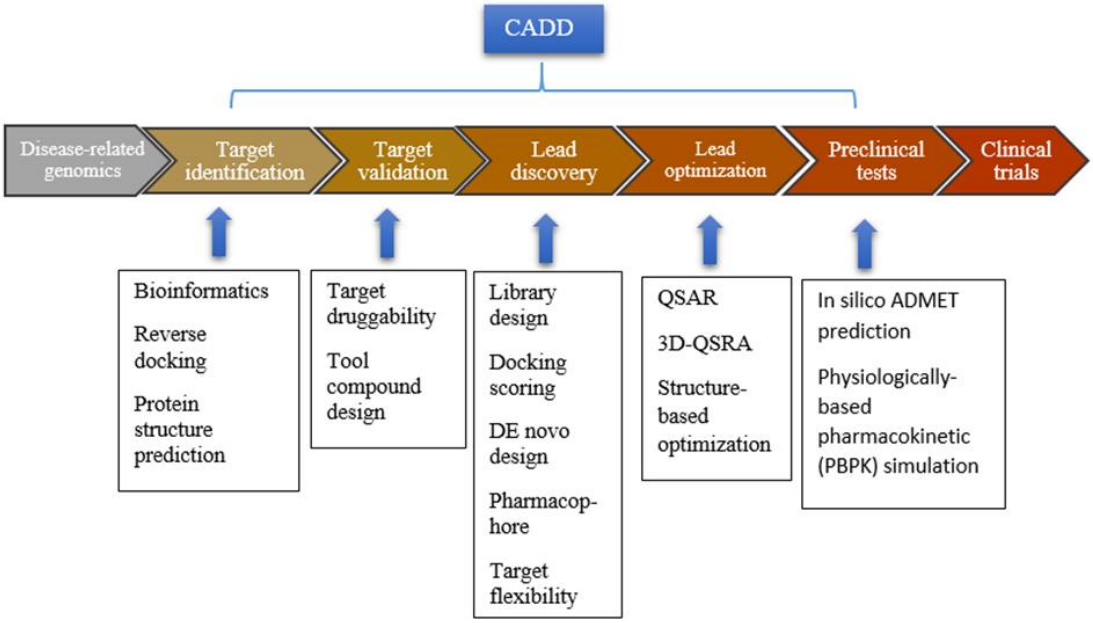

Figure 4: In silico Computer-Aided Drug Design (Adapted from (Hung & Chen, 2014))

4.1.2 Classification of Computer-Aided Drug Design (CADD)

These modern drug design strategies (i.e. CADD) mainly consist of two major approaches or methods. That is, LBDD (ligand-based drug design) and SBDD (structure-based drug design) methods to develop new drug candidates (Shapshak et al., 2019). The use or selection of either of these two approaches depends on whether the information about the biological target is available or not. For example, the ligand-based drug design (LBDD) strategies are mostly used in situations where there is no information (i.e. 3-D structure information) about the biological target accessible, and once knowledge about the biological target structure becomes available, structure-based drug design (SBDD) methodologies can be applied with the major goal of figuring out how the ligand interacts to the target molecule's biological active site using the particular interactions between ligands and proteins (Lima et al., 2016). Therefore, in the hunt for novel compounds active against a specific target, SBDD uses the 3D structure information of the drug target. However, when a target's experimental 3D crystal structure is unavailable, a theoretical or hypothetical 3D structure of the protein can be

constructed using homology modeling to assist SBDD methods (Njogu et al., 2016). These Structure-based drug discovery (SBDD) approaches have greatly impacted the drug companies and pharmaceutical industry generally in terms of new drug development of both infectious diseases like HIV/AIDS and non-infectious diseases. For example, many licensed HIV protease inhibitors, such as Saquinavir and Amprenavir, were created utilizing structure-based molecular docking in the early 1990s to target HIV infections (Leelananda & Lindert, 2016).

For instance, ligand-based drug design, as opposed to structure-based drug design, is concerned with the knowledge of known molecules (ligands) that interact with the target macromolecule of interest rather than experimental or even theoretical 3D structural information about the target protein macromolecule of interest. It is feasible to develop a pharmacophore model using these well-known compounds, which illustrates the minimal structural features (pharmacophores) that a molecule must possess in order to bind to the target. A new molecular entity that interacts with the target may then be created using the pharmacophore model developed previously. QSARs (quantitative structure–activity relationships), which are derived from a connection between the computed properties of compounds and their experimentally determined biological activity, can also be used to predict the activity of novel analogs (Lin et al., 2020). Methods for ligand-based drug design (LBDD), such as scaffold hopping, 3D-quantitative structure activity relationship (3D-QSAR), 2D similarity-based searching, and pharmacophore investigations, are helpful for hit enrichment together with activity prediction based on previously discovered inhibitor (ligands) information. Because of breakthroughs in crystallography and successful applications of homology modeling, structure-based virtual screens (SBVS) have also shown to be effective tools for rapidly discovering bioactive hits in early-stage discovery efforts (Lu et al., 2018). Although, new approaches, such as sequence-based approaches that use

bioinformatics methods to analyze and compare multiple sequences, have been developed to identify potential targets from scratch and conduct lead discovery in situations where both target structure and ligand related information are inefficient or unavailable, these approaches are still in their early stages (Ou-Yang et al., 2012).

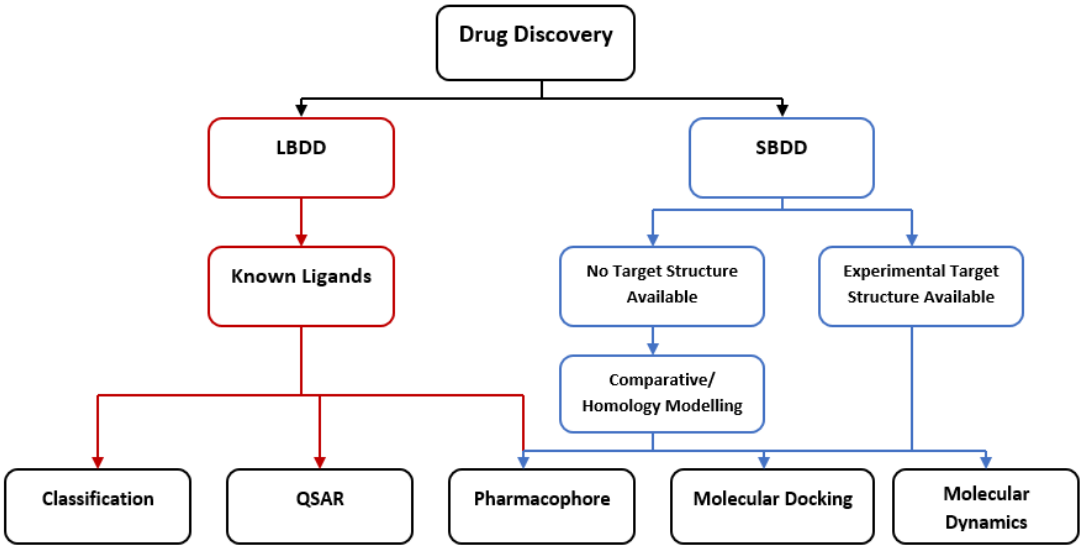

Figure 5: Bioinformatics and machine learning approaches employed in drug design and discovery i.e. ligand-based drug design (LBDD) and structure-based drug design (SBDD) together with other relevant tools.

4.2 *In silico* Computer-Aided Drug Design (CADD) Steps and Tools

As a result of the continuous accumulation of biological macromolecule and small molecule knowledge, the application of computational drug discovery has been greatly widened and extensively deployed to almost every stage in the drug research and development process (Figure 6). Among a variety of activities or stages involved in drug design, including target selection and validation, lead discovery and optimization, and preclinical and clinical testing, computational or in-silico tools have been applied in stretch from target identification to preclinical testing (Ou-Yang et al., 2012).

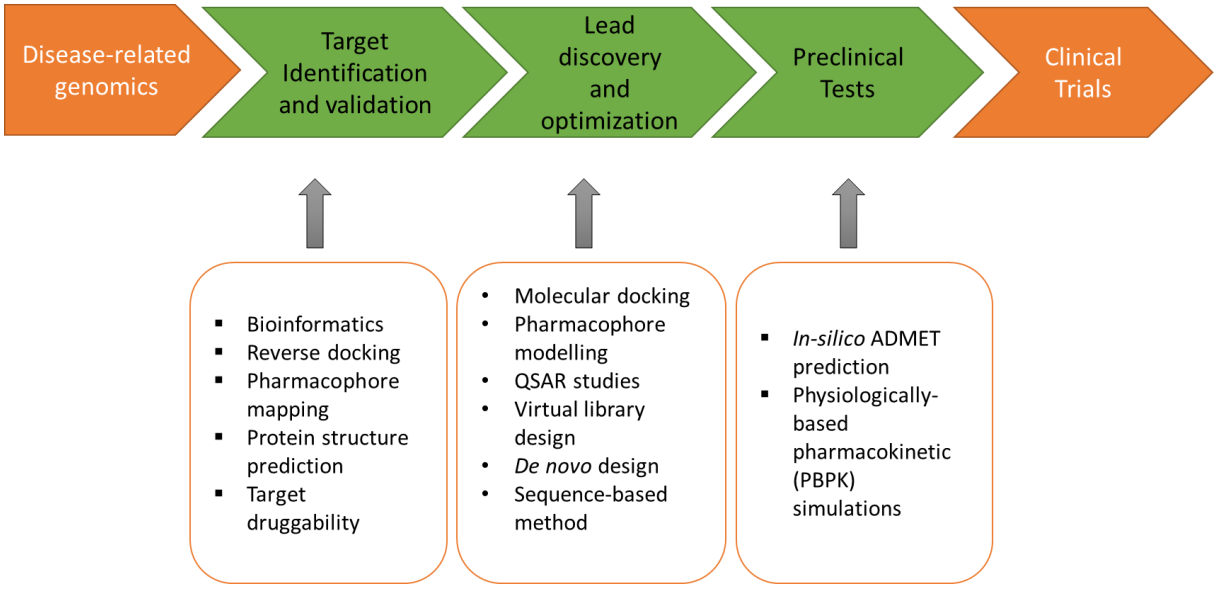

Figure 6: Computational drug discovery approaches that have been applied in various stages of the drug discovery and development pipeline.

4.2.1 Target Identification and Validation

Drug targets also referred to as drug receptors are basically biological sites onto which the ligand (drug) binds. These are tiny essential biomolecules that allow a drug agent or molecule to have desired effects on metabolic or signaling pathways linked to the illness under investigation without interfering with the cell's normal function (Aamer Mehmood, 2014). They can be classified into different kinds of targets such as enzymes, several kinds of receptors, ion channels, transporters, and others (Caldwell, 2015). Therefore, an effective the identification and confirmation of feasible targets (disease's biological origins and possible intervention targets) is a seemingly critical and crucial first stage in the drug design, discovery, and development process (Master et al., 2010). It all starts with determining the function and involvement of a potential therapeutic target (gene/nucleic acid/protein) in the disease. Determination of the molecular mechanisms associated with these target molecules is also important in predicting ideal druggable (efficacious and safe) targets with druggability that meets clinical and commercial requirements (Luxminarayan et al., 2019). Understanding

how a gene works, for example, is critical when selecting a gene as a target, as demonstrated by the bioinformatics sequencing of the human genome, which has supplied researchers with hundreds of potential novel drug targets (Aamer Mehmood, 2014).

Conventionally, this Target Identification and Validation process takes much time and requires a lot of resources, however, with the new CADD techniques, it has been simplified. The Bioinformatics (genomics, proteomics and metabolomics) approaches incorporated in the CADD techniques have exponentially improved the target identification and validation process (Caldwell, 2015). Computational tools assist in *in silico* predicting of genes, nucleic acids and or proteins of small molecules as targets. Protein information to assist in target selection can be obtained by using a variety of *in silico* approaches like reverse docking, pharmacophore mapping, protein structure predictions and among others (Caldwell, 2015). That is, methods like Homology modeling and molecular docking or reverse docking are some of the mostly used bioinformatics and machine learning tools (Figure 5) in the target identification stage (Zheng et al., 2013). For example, reverse docking computational methodology uses docking and scoring technologies to screen a particular ligand or drug against protein-target databases to identify and establish a single or multiple potential targets or to find several targets that could better represent the disease or be responsible for off-target toxicity and side effects (Caldwell, 2015). Support Vector Machines (SVM), another in-silico machine learning approach, has been investigated as a novel way for predicting druggable proteins from amino acid sequences (Master et al., 2010). For example, according to one particular research paper, a support vector machine (SVM) classifier was created utilizing several available genomic data sets to classify proteins into drug targets and non-drug targets for the treatment of breast, pancreatic, and ovarian cancers, among other diseases. In order to research HIV druggable protein targets, such comparable models can be used in a similar fashion to those used in other studies (Vamathevan et al., 2019). These Bioinformatics and

Machine Learning tools can be used to analyze the target structure for possible active sites for binding, check for their likeness by docking the molecules with the target to generate the candidate molecules and rank them according to their binding affinities. They can also optimize the molecule to improve their binding characteristic (Hassan et al., 2017).

4.2.1.1 Homology Modeling

Homology modeling is the process of predicting a protein's three dimensional structure based on the structure of another homologous protein whose structure has already been established. Homology modeling is nothing more than a search for pharmacological analogs based on similarities (Rahman et al., 2012). It's also known as comparative modeling, and it's a computer approach for creating a 3-D model for an unknown structure (target) from the amino acid sequence using one or more homologous proteins linked to known structures (templates) (Lima et al., 2016). Understanding the function, dynamics, and interaction of proteins, as well as functional prediction and identification of therapeutic targets, is aided by determining the 3-D structure (Diniz & Canduri, 2017). This method is based on the idea that two sequences that are identical or evolutionarily related have comparable three-dimensional structures. The creation of a more trustworthy model will be aided by a better sequence identity between the target and template structures (Singh, 2020). Homology modeling is a dependable method for creating a 3D model of a protein from its amino acid sequence if the template structure is known and the sequences have more than 30% similarity. This can be done in mainly four stages starting with identifying and selecting qualified structural template(s), then aligning the target and template sequences with the help of certain alignment tools like FASTA and BLAST. It is then followed by the third step of predicting the secondary structure and build model based on the template structure and finally refining the model by evaluating its quality to ensure that the desired 3D target structural model is obtained (Gao, 2016). The whole process is an In-silico process and can be executed with

help of different modelling tools (Table 2) such as Modeller, Prime and SWISS-MODEL. For example, a study shows that homology modelling technique was used to construct a homology model for CCR5 receptor or target structure using the available 3-D structural information of bovine rhodopsin and CXCR4 receptors as templates (Gu et al., 2014).

4.2.1.2 Reverse Docking

To make *in-silico* target identification, computational methods take use of the rising amount of large-scale human genomes and proteomics data sets (Dezso & Ceccarelli, 2020). One of the popular approaches for in-silico drug target prediction or identification is the reverse docking. The essential idea of reverse docking is that each docking score is generated when the query molecule or tiny molecular ligand is individually docked against a structural grid database containing a large number of protein targets. The protein targets are then ranked by docking energy, with a higher rank (protein with higher docking energy) suggesting a better likelihood of the protein being a target. The basic principle of reverse docking is dependent on the determination of the docking energy or interaction energy between the small-molecule ligand (query molecule under study) and a potential protein target which reflects the binding strength of the ligand with the target (Huang et al., 2018). It is structural-based approach that generates a connection hypothetical relationship within protein targets determining possible potential targets by assessing the binding affinity of the ligand or compound that bind to these targets individually. Therefore, a given ligand is docked to different targets from a set of protein target database to obtain the potential drug target. Using different computational and machine learning tools and softwares like AutoDock, TarFisDock and idTarget, reverse docking can be implemented in predicting and finding drug targets. This approach can also be

used to determine off-targets that are sometimes responsible for side effects of drugs (Agamah et al., 2020).

4.2.1.3 Pharmacophore Mapping

A pharmacophore refers to the spatial arrangement of functional properties that permits molecules or ligands (drugs) to bind to target proteins in a certain manner. For example, hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), hydrophobic groups (H), positively charged center (P), negatively charged center (N) and aromatic rings(Huang et al., 2018). It is the configuration or arrangement of steric and electrical characteristics required to initiate or inhibit a biological reaction (Hassan Baig et al., 2016). The principle is that important functional pharmacophores have a major role in the binding of some drug molecules to their protein targets. As a result, matching these essential pharmacophores can be utilized to find novel small-molecule therapeutic targets (Huang et al., 2018).

Target validation involves proper understanding of the cell type in which the target is present together with the target biology role in cell signaling and the target participation in the metabolic pathways for healthy and diseased individuals. Therefore, because of this complexity, Bioinformatics comprising of many computational methods, is employed to support target validation by providing a global integration of all available data. In addition to determining common principles and providing functionally predictive information that has been mined (Caldwell, 2015).

4.2.2 Primary HIV Viral Targets

The key targets of HIV virus include the following, reverse transcriptase, protease, integrase, CCR5 and CXCR4.

4.2.2.1 Reverse transcriptase

HIV as described before is a retrovirus, therefore, reverse transcriptase (RT) is its key enzyme in terms of its survival. The enzyme Reverse transcriptase reverse transcribes the viral RNA into a provirus. It also plays a multifunctional role as it is an essential component for HIV to complete the replication cycle. As a result, this serves as an important target for the development of anti-HIV medicines, such as reverse transcriptase inhibitors. For example, there are several crystal structures, mutants and RNA/DNA hybrid structures of HIV-1 complexes that can help in developing HIV-1 reverse transcriptase inhibitors (Gu et al., 2014).

4.2.2.2 Protease

One of the three main enzymes involved for the production of viral proteins is protease (PR). That is, following translation, it cleaves the viral polyprotein to liberate functionally mature proteins. As a result, it has been deduced as an essential target for the development of anti-HIV medicines. For example, *Table 1;* shows a list of Protease inhibitors that have been developed by computer-aided drug design tools target the protease enzyme as a way of preventing the HIV infection from escalating.

4.2.2.3 Integrase

Integrase (IN) is the viral enzyme that catalyzes the reaction process that involves the integration of viral DNA into the host genome that results into the permanent infection if the human cell (DNA) by the virus. For this reason, integrase enzyme is as well considered for targeting of the HIV virus in case of developing HIV drugs. For example, Table 1; in 2007, the FDA authorized the first integrase inhibitor (raltegravir) for the clinical treatment of AIDS (Table 1) (Gu et al., 2014).

4.2.2.4 CXCR4 and CCR5

CXCR4 (chemokine receptor type 4) and CCR5 (chemokine receptor type 5) are co-receptors that participate in the HIV viral binding, fusion and entry in to the host cell. These major co-receptors are found on CD4+ T-cells and Macrophages respectively. Since these two co-receptors are essential for viral entry into the host cell, they can be potential targets for inhibition of viral entry.

4.2.2.5 Human α-Glucosidase

Human-Glucosidase is an enzyme that cleaves the HIV glycoprotein gp160 and forms a non-covalent complex with the two glycoproteins gp41 and gp120. This complex allows HIV to bind and sub-sequently enter the CD4+ lymphocytes (host cell). Therefore, inhibition of Human α-Glucosidase helps in blocking the entry of the virus into the host cell (Kirchmair et al., 2012).

Table 1: Examples of anti-HIV drugs discovered by CADD:

No.	Drug	Drug Target	Disease	Approved Year	Reference
1	Saquinavir	HIV-1 and HIV-2 proteases	AIDS	1995	(Hassan Baig et al., 2016)
2	Indinavir	HIV protease	AIDS	1996	(Hassan Baig et al., 2016)
3	Ritonavir	HIV protease	AIDS	1996	(Hassan Baig et al., 2016)
4	Raltegravir	Integrase	AIDS	2007	(Kirchmair et al., 2012)

5	Nelfinavir	HIV protease	AIDS	2007	(Hassan Baig et al., 2016)
6	Rilpivirine	Reverse transcriptase	AIDS	2011	(Santos et al., 2015)
7	Dolutegravir	Integrase	AIDS	2013	(Santos et al., 2015)

4.2.3 Lead Discovery and Optimization

Among all of the phases of the drug development process, lead identification and optimization are critical. Drug development efforts could not be started or continued without lead molecules (Xiang et al., 2012). Leads are traditionally discovered in a variety of methods. For example, they can be discovered through a pure serendipitous process, systematic or random screening, and or chemically modifying known active compounds. The use of rational approaches involving in-silico computational techniques and tools that begin with a validated biological target, exploit the structural specificity of that target, and end with the identification of a drug candidate that optimally interacts with it and stimulates the desired physiological or biological action, on the other hand, is becoming increasingly popular (Prakash & Devangi, 2010). Except the rational *in silico* computational lead discovery strategies, these traditional methods have proved to take a long time and cost a lot of money. However, with the new *in silico* drug discovery and development process together with its new computational techniques and tools employed in almost all the drug discovery and development steps with lead discovery and optimization step inclusive, the game has greatly changed. This Drug discovery step begins with data collection of probable compounds that could be small molecules or peptides as this information is of great use (Usha et al., 2018). Bioinformatics and machine learning or *In-silico* methods are expected to play a key role in utilizing structural information and functionally available data to better

understand specific molecular recognition events of the target macromolecule with candidate hits, ultimately leading to the development of improved leads for the target macromolecule (Afshan Shaikh et al., 2012).

4.2.3.1 Structure-Based Drug Discovery (SBDD) Approach

The structure-based approach is often used when information of the three dimensional structure of the target protein macromolecule is always known (Njogu et al., 2016). The crystal structure of the therapeutic target, especially when cocrystallized with a known ligand, provides the majority of the 3D structural information. This information is then utilized to guide and facilitate the discovery of new ligands, either by directing the design of novel compounds or by identifying new ligands through virtual compound library screening (Njogu et al., 2016). Also the 3D structure of the target protein macromolecules already determined experimentally under X-ray crystallography or nuclear magnetic resonance (NMR) techniques, can be retrieved from storage biological databases like Protein Data Bank (PDB) (Yu & Jr, 2017). In the PDB, for instance, there are over 240 HIV-1 RT crystal structures and mutants complexed with over 80 small chemical ligands (Kirchmair et al., 2012). In the lack of structural knowledge on the target (i.e. the target's 3D structure), homology modeling can be used to identify the target protein's structure (Hassan Baig et al., 2016).Beginning with the atomic coordinates, potential ligands can be analyzed by molecular automatic docking calculations in order to find the best fit with the corresponding binding site on the target protein macromolecule. These techniques can also predict the affinity of the ligand. Also structure-based pharmacophore building, which involves 3D mapping of the physicochemical features enables a ligand to bind with a defined protein pocket (Maga et al., 2013). In the discovery of HIV protease inhibitors (ritonavir, indinavir, saquinavir), for example, different structure-based design and discovery approaches were heavily used. These drugs, which were introduced to market in the mid-1990s and played a critical role in reversing a rapid increase

in AIDS mortality and morbidity in the United States, were developed using a variety of structure-based design and discovery approaches (Drie, 2007).

4.2.3.1.1 Virtual Screening:

Virtual screening of chemical libraries, also known as virtual high-throughput screening (vHTS), is one of the most popular applications of CADD (Sliwoski et al., 2014). Virtual screening entails docking of molecules and screening of a chemical database against a pharmacological target, then rating the results based on the binding free energy of the compounds with the target (Singh, 2020.). Virtual screening, a scoring and ranking process that estimates the binding affinity of a given ligand to its specific target is knowledge based. That is, depending on the available information on either the ligand or target in terms of their structure, ligand or target based virtual screening is employed (Ekins et al., 2007). For instance, in target based virtual screening, a target's 3-D structure is compared to libraries of potentially active small molecules or ligands, with the computer docking each ligand into the active site of the target and scoring its geometric and electrostatic fit and ranking them accordingly to obtain the probable most bioactive ligand (Cobb, 2007). Principally, ligand based virtual screening is based on the molecular or compound similarity property theory. This theory states that two or more similar molecules or compounds are expected to show similar properties and in this way there is a higher chance that they bind to same target proteins (Ekins et al., 2007). Virtual screening (vHTS) has been classified in many categories and can come in different forms, for, example, Chemical similarity searches using fingerprints or topology, compound selection using QSAR models, pharmacophore mapping, and or virtual docking of compounds into targets of interest, also known as structure-based docking (Sliwoski et al., 2014). Pharmacophore virtual screening approaches among others

have been employed in the various therapeutic target hit identifications such as HIV integrase and CCR5 antagonists (Ekins et al., 2007).

4.2.3.1.2 Molecular Docking

Molecular docking is one of the virtual screening methods that are highly applied in the structural based drug discovery to screen large libraries of chemical compounds against a biological targets. Virtual screening is an advanced computational form of High throughput screening(Afshan Shaikh et al., 2012). Computer software is used to simulate ligand-target binding and predict binding conformations and molecular interactions between ligands and target macromolecules (Njogu et al., 2016). Molecular docking is used to investigate how a particular ligand interacts with a protein or to explore a library of chemicals for possible binders to a target protein. The whole docking procedure may be broken down into two steps: first, docking algorithms are used to position a ligand in an active site, and then a scoring function is used to measure the strength of the binding posture (Gao, 2016). There are different computer based softwares listed in Table 2*;* that can be applied to execute this process automatically, for example, AutoDock, DOCK, GOLD and Surflex-Doc with the application of large compound libraries like DrugBank and ZINC, where several ligand molecules can be obtained (Sliwoski et al., 2014).

4.2.3.2 For Ligand-Based Drug Discovery (LBDD) approach:

The approaches are discussed in the following sections.

4.2.3.2.1 Pharmacophore Modeling

A pharmacophore is defined by IUPAC as the three-dimensional arrangement of steric and electronic characteristics required to activate or inhibit a biological reaction. Chemical

characteristics such as hydrogen bond donors or hydrogen bond acceptors, aromatic rings, hydrophobic groups, and positive and negative ionizable moieties are examples of three-dimensional chemical features (Kirchmair et al., 2012). Basically, this method depends on the principle of molecular similarity or chemical structure similarity stating that a group or set of similar ligand molecules or compounds may have similar bioactivity i.e. would bind to similar targets with relatively similar binding affinity (Agamah et al., 2020). Therefore, the aim in this approach is to predict and sort out common pharmacophores (specific chemical bonds, HBA, HBD, cations, anions, and hydrophobic groups) within ligands that are treated as ligand–target interaction active sites and these can be used to generate pharmacophore models (Hung & Chen, 2014). For example, in one research study, pharmacophore models of HIV protease were examined using a test set of HIV protease inhibitors as well as inhibitors known to be active on other proteases, together with inactive compounds, in order to determine which model was the most selective based on the inhibitors' selectivity. The findings suggest that the use of ensembles of pharmacophore models, as well as ensemble voting, may be beneficial in improving the signal-to-noise ratio of a dataset. As a result, it is possible to dramatically boost the retrieval rate of known active drugs while maintaining or improving the selectivity of the pharmacophore-based strategy (Koutsoukas et al., 2011).

4.2.3.2.2 Quantitative Structure Activity Relationship (QSAR)

This technique is basically a quantitative study of the interactions between small organic molecules and biological macromolecules. It comprises a relationship between a molecule's predicted characteristics (e.g., absorption, distribution, and metabolism of tiny organic compounds in live organisms) and their biological activity as established empirically (Lin et al., 2020). The QSAR model is constructed utilizing structural factors to predict biological qualities, as well as statistical and analytical techniques to determine the connection between compound structures and biological behaviors (Singh, 2020). CoMFA (Comparative

Molecular Field Analysis) and CoMSIA (Comparative Molecular Shape Indices Analysis) are two of the most widely used 3D QSAR approaches. As a result of both of these procedures, statistical models of the molecule are generated and shown as color-coded contours around the molecule, indicating places where electrostatic characteristics and spatial arrangements are favorable or unfavorable for bioactivity (Kirchmair et al., 2012). Quantitative Structure Activity Relationship studies mostly help to sort and remove compounds or ligands possessing undesirable pharmacokinetic or projected toxic properties from large compounds libraries so that only druggable compounds are optimized for further experimentation (Santos et al., 2015).

4.2.4 Preclinical and Clinical Trials

In silico ADMET studies or computational pharmacokinetic properties assessment allows for the prediction of absorption, distribution, metabolism, elimination, and toxicity (ADMET) of drug candidates, which is an important computational drug design and discovery process in preclinical trials to estimate the drug molecules' safety and efficacy levels before conducting clinical trials (Lu et al., 2018). This is because, in order to get through clinical trials, a druggable molecule must possess specific ADMET characteristics; otherwise, it is likely to be rejected. To prevent such scenarios, Support Vector Machine (SVM), k-Nearest Neighbour (k-NN), Naive Bayes Classifiers, Decision Trees, and Random Forests (RF) among other statistical and machine learning techniques are used to create the ADMET model prediction tools (Singh, 2020). For example, several machine learning based ADMET prediction tools like ADMET predictor and ADMEWORKS Predictor are used in the preclinical phase to estimate and ascertain the safety and effectiveness of the drug candidate before proceeding to the experimental synthesis and clinical trials phase (Table 2). Many other similar tools have been developed to predict drug-like properties and ADMET properties of drug candidates, including OSIRIS Property Explorer (web-based tool),

ChemSilico (neural net based prediction method), Pre-ADME, PASS Online (Prediction of Activity Spectra for Substances), a Bayesian-based tool, and DREADD (designer receptors exclusively activated by designer drugs) (Aamer Mehmood, 2014).

Table 2: Bioinformatics and Machine Learning tools used in Computer Aided Drug Design (CADD)

CADD Software(Tool)	**Developer**	**Brief description and algorithm used**	**Website**
Docking			
AutoDock	The Scripps Research Institute	Simulated annealing, genetic algorithms (GA)	http://autodock. scripps.edu/
DOCK	University of California	Incremental construction, merged target structure ensemble	https://www.biosolveit. de/products/#FlexX
GOLD	The Cambridge Crystallographic Data Centre	Genetic Algorithms (GA)	https://www. ccdc.cam.ac.u/ solutions/csd- discovery/ components/ gold
Surflex-Doc	Tripos Inc.	Hammerhead's empirical scoring function with morphological similarity	http://www. jainlab.org
Homology modeling			

Modeller	University of California	It performs Homology/comparative modeling using spatial restraints and performs optimization of protein structure models	https://salilab. org/modeller
Prime	Schrödinger, Inc.	Prime performs Comparative modeling using homology modeling and fold recognition	https://www. schrodinger.com/prime
SWISS-MODEL	Swiss Institute of Bioinformatics	Fully automated protein homology modeling server	https://swissmodel. expasy.org/
ADMET prediction			
ADMET predictor	Simulations Plus, Inc.	ANN, SVM, Kernel partial least squares (KPLS), and multiple linear regression (MLR)	https://www.simulations-plus.com/software/ membraneplus/admet-predictor/
ADMEWORKS Predictor	Fujitsu Kyushu Systems	It is a virtual screening system with a simultaneous evaluation of ADMET properties	https://www.fujitsu.com/ global/

Sarchitect	Syngene	Bayesian methods, ANN, SVM, decision trees and forests, and other algorithms are used for model building and prediction	https://www.syngeneintl.com
Hazard Expert Pro	CompuDrug, Ltd	A neural network-based approach is used to model the relationship between human cytotoxicity and atomic descriptors	https://www.compudrug.com/hazardexpertpro

Chapter 5

Future Prospects

5.1 HIV/AIDS Vaccines

Despite the high infectious rate of HIV virus and its prolonged existence of over 40 years among humans, there is not even a single effective HIV-1 vaccine that has ever been clinically approved to date and the hope for protective immunization is still little despite it being the best solution for HIV eradication(Liu et al., 2020). The initial HIV vaccine development process was based on the principle of preventing the infection. The immune system would be trained to respond to HIV by generating antibodies that bind to the virus and prevent it from infecting cells, or by stimulating other immunological responses that kill the virus However, the extraordinary failures of the initial studies involving different HIV vaccine candidates have initiated the perception of changing the course of vaccines development. The current research is therefore, aiming at designing a vaccine that possesses therapeutic properties that aim at curing the infection (Larijani et al., 2019). Currently, the only vaccine candidate among the six completed HIV vaccine trials so far that has ever registered a reasonable amount of success is the RV144 vaccine (Shin, 2016). This vaccine trial conducted in Thailand was a combination of canarypox vector (ALVAC-HIV) and gp120, B/E (AIDSVAX) vaccine that showed nondurable vaccine efficacy of 31.2% (after reducing from 60%) and could not be continued as there was no effect on the viral load (Larijani et al., 2019) (M Barry, 2014). Generally, over time these vaccines have been developed using traditional and experimenting methods that tend to consume a prolonged period of time estimated at 5-15 years at an extremely huge cost. However, with the current development and advancements in the drug and vaccine development technologies, computational (bioinformatics and or immunoinformatics) *in-silico* vaccine design

technologies can accelerate the whole process of vaccine development with a highly reduced cost (Parvizpour et al., 2020). For example, the research that has embarked on finding a therapeutic vaccine for HIV after failure of all other primary vaccine trials, is focusing on a rational design of broadly neutralizing antibodies (bnAbs) inducing vaccines with multi-immunogenic sequence strategies. This exploits reverse vaccinology technologies where templates of known bnAbs are used to design immunogens that help in activating bnAb-producing cells to enable production of mature bnAbs. Reverse vaccinology is a strategy implemented in the designing and construction of vaccines depending the available genomic data using different bioinformatics approaches and tools that can identify and verify compounds/ entities with the potential to induce immune responses (immunogenicity) (Burton, 2019). Another promising field of HIV vaccine development is the Epitope-based vaccine development approach that widely exploits the availability of genomic data and state of the art *in silico* and bioinformatics tools to execute the process. The whole process of this design starts with *in silico* epitope mapping to predict immunogenic regions, followed by designing the immunogenic construct and lastly evaluating its vaccine efficacy or immunogenicity (Parvizpour et al., 2020). For example, there has been a group of researchers that applied this method and designed and artificial epitope-based polypeptide immunogens against HIV-1 (CombiHIVvac). This proposed HIV-1 vaccine design aimed at delivering both T-cell and B-cell epitopes that could stimulate both humoral and cellular mediated immune responses (I. Karpenko et al., 2018).

5.2 Antiviral Drug Resistance Predictions

Antiviral medication resistance can be anticipated and biologically assessed in two ways. Phenotyping and genotyping are the two biological procedures, with genotypic testing being the most frequent and recommended method, in which the sequence of the viral genome is examined for the presence of known treatment resistance mutations (Shapshak et al., 2019).

However, in phenotypic testing, the resistance or susceptibility to drugs is measured with cells infected with the viral strain in vitro which determines whether a mutation of the virus can possibly raise any resistance to a given drug or not. Therefore, genotypic testing quantifies drug susceptibility whilst phenotyping determines the mutational pattern. Unlike genotypic testing, phenotypic testing is slower and more expensive, making it nearly impossible to investigate the mutational resistances that emerge on a regular basis. Genotyping is currently advised for new HIV infections in order to detect resistant mutations, guide medication selection, and guarantee successful treatment of HIV/AIDS patients. As a result, the best approach is to create computer methods that predict resistance from a particular genotype (Shapshak et al., 2019).

Chapter 6

Conclusion

Regardless of the failures, drug design and development procedures have followed a completely new course as a result of constant technological improvements. With a plethora of anti-HIV medications on the market, there is a large public need for another line of therapy that will put an end to the fatal HIV epidemic, and that therapy will be none other than the HIV vaccine. However, in light of the massive losses suffered in HIV vaccine research and testing, it is necessary to change the whole process or at the very least adapt it in light of the knowledge and information obtained from past studies.

References

Aamer Mehmood, M. (2014). Use of Bioinformatics Tools in Different Spheres of Life Sciences. *Journal of Data Mining in Genomics & Proteomics*, *05*(02). https://doi.org/10.4172/2153-0602.1000158

Afshan Shaikh, S., Jain, T., Sandhu, G., Soni, A., & Jayaram, B. (2012). From Drug Target to Leads-Sketching a Physico-Chemical Pathway for Lead Molecule Design In Silico. *Frontiers in Medicinal Chemistry*, 324–360. https://doi.org/10.2174/978160805464011306001 5

Agamah, F. E., Mazandu, G. K., Hassan, R., Bope, C. D., Thomford, N. E., Ghansah, A., & Chimusa, E. R. (2020). Computational/in silico methods in drug target and lead prediction. *Briefings in Bioinformatics*, *21*(5), 1663–1675. https://doi.org/10.1093/bib/bbz103

Burton, D. R. (2019). Advancing an HIV vaccine; advancing vaccinology. *Nature Reviews Immunology*, *19*(2), 77–78. https://doi.org/10.1038/s41577-018-0103-6

Caldwell, G. W. (2015). In silico tools used for compound selection during target-based drug discovery and development. *Expert Opinion on Drug Discovery*, *10*(8), 901–923. https://doi.org/10.1517/17460441.2015.1043885

Cobb, B. Y. K. (2007). Dock This : In Silico Drug Design Feeds Drug Development. *Biomedical Computation Review*, 20–30.

Dash, S. S., Nayak, S. K., & Mishra, D. (2021). A review on machine learning algorithms. *Smart Innovation, Systems and Technologies*, *153*(October), 495–507. https://doi.org/10.1007/978-981-15-6202-0_51

Dezso, Z., & Ceccarelli, M. (2020). Machine learning prediction of oncology drug targets based on protein and network properties. *BMC Bioinformatics*, *21*(1). https://doi.org/10.1186/s12859-020-3442-9

Diniz, W. J. S., & Canduri, F. (2017). Bioinformatics: An overview and its applications. In *Genetics and Molecular Research* (Vol. 16, Issue 1, p. 16019645). https://doi.org/10.4238/gmr16019645

Du, H., Cai, Y., Yang, H., Zhang, H., Xue, Y., Liu, G., Tang, Y., & Li, W. (2017). In Silico Prediction of Chemicals Binding to Aromatase with Machine Learning Methods. *Chemical Research in Toxicology*, *30*(5), 1209–1218. https://doi.org/10.1021/acs.chemrestox.7b00037

Dubey, A. (2018). Machine learning approaches in drug development of HIV/AIDS. *International Journal of Molecular Biology*, *3*(1). https://doi.org/10.15406/ijmboa.2018.03.00044

Ekins, S., Mestres, J., & Testa, B. (2007). In silico pharmacology for drug discovery: Methods for virtual ligand screening and profiling. In *British Journal of Pharmacology* (Vol. 152, Issue 1, pp. 9–20). https://doi.org/10.1038/sj.bjp.0707305

Eric, S. D., Nicholas, T. K. D. D., & Theophilus, K. A. (2014). Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). *Journal of Bioinformatics and Sequence Analysis*, *6*(1), 1–6. https://doi.org/10.5897/ijbc2013.0086

Gao, C. (2016). *Computer-aided drug design approaches in developing anti-cancer inhibitors* (765432156S ed.). https://gupea.ub.gu.se/bitstream/2077/48857/5/gupea_2077_48857_5.pdf

Gaudêncio, S. P., & Pereira, F. (2020). A Computer-Aided Drug Design Approach to Predict Marine Drug-Like Leads for SARS-CoV-2 Main Protease Inhibition. *Marine Drugs*, *18*(12). https://doi.org/10.3390/md18120633

Ghosh, A. K., Osswald, H. L., & Prato, G. (2016). Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS. In *Journal of Medicinal Chemistry* (Vol. 59, Issue 11, pp. 5172–5208).

https://doi.org/10.1021/acs.jmedchem.5b01697

Gu, W. G., Zhang, X., & Yuan, J. F. (2014). Anti-HIV drug development through computational methods. In *AAPS Journal* (Vol. 16, Issue 4, pp. 674–680). https://doi.org/10.1208/s12248-014-9604-9

Hassan Baig, M., Ahmad, K., Roy, S., Mohammad Ashraf, J., Adil, M., Haris Siddiqui, M., Khan, S., Amjad Kamal, M., Provazník, I., & Choi, I. (2016). Send Orders for Reprints to reprints@benthamscience.ae Computer Aided Drug Design: Success and Limitations. *Current Pharmaceutical Design*, *22*, 572–581.

Hassan, I., Bream, A. S., El-Sayed, A., & Yousef, A. M. (2017). International Journal of Advanced Research in Biological Sciences Assessment of disinfection by-products levels in Aga surface water plant and its distribution system, Dakhlia, Egypt. *Int. J. Adv. Res. Biol. Sci*, *4*(4), 37–43. https://doi.org/10.22192/ijarbs

HIV Drugs and the HIV Lifecycle | The Well Project. (2019). https://www.thewellproject.org/hiv-information/hiv-drugs-and-hiv-lifecycle

Huang, H., Zhang, G., Zhou, Y., Lin, C., Chen, S., Lin, Y., Mai, S., & Huang, Z. (2018). Reverse screening methods to search for the protein targets of chemopreventive compounds. In *Frontiers in Chemistry* (Vol. 6, Issue MAY, p. 138). https://doi.org/10.3389/fchem.2018.00138

Hung, C. L., & Chen, C. C. (2014). Computational approaches for drug discovery. In *Drug Development Research* (Vol. 75, Issue 6, pp. 412–418). https://doi.org/10.1002/ddr.21222

I. Karpenko, L., I. Bazhan, S., M. Eroshkin, A., V. Antonets, D., N. Chikaev, A., & A. Ilyichev, A. (2018). Artificial Epitope-Based Immunogens in HIV-Vaccine Design. *Advances in HIV and AIDS Control*. https://doi.org/10.5772/intechopen.77031

Issac, B., & Raghava, G. P. S. (2005). FASTA Servers for Sequence Similarity Search. *The*

Proteomics Protocols Handbook, *October*, 503–525. https://doi.org/10.1385/1-59259-890-0:503

Kirchmair, J., Distinto, S., Roman Liedl, K., Markt, P., Maria Rollinger, J., Schuster, D., Maria Spitzer, G., & Wolber, G. (2012). Development of Anti-Viral Agents Using Molecular Modeling and Virtual Screening Techniques. *Infectious Disorders - Drug Targets*, *11*(1), 64–93. https://doi.org/10.2174/187152611794407782

Larijani, M. S., Ramezani, A., & Sadat, S. M. (2019). Updated Studies on the Development of HIV Therapeutic Vaccine. *Current HIV Research*, *17*(2), 75–84. https://doi.org/10.2174/1570162x17666190618160608

Leelananda, S. P., & Lindert, S. (2016). Computational methods in drug discovery. In *Beilstein Journal of Organic Chemistry* (Vol. 12, pp. 2694–2718). https://doi.org/10.3762/bjoc.12.267

Levin, C., Dynomant, E., Gonzalez, B. J., Mouchard, L., Landsman, D., Hovig, E., & Vlahovicek, K. (2018). A data-supported history of bioinformatics tools. In *arXiv*.

Lima, A. N., Philot, E. A., Trossini, G. H. G., Scott, L. P. B., Maltarollo, V. G., & Honorio, K. M. (2016). Use of machine learning approaches for novel drug discovery. *Expert Opinion on Drug Discovery*, *11*(3), 225–239. https://doi.org/10.1517/17460441.2016.1146250

Lin, X., Li, X., & Lin, X. (2020). A Review on Applications of Computational Methods. *Molecules*, *25*(6), 1375.

Liu, Y., Cao, W., Sun, M., & Li, T. (2020). Broadly neutralizing antibodies for HIV-1: efficacies, challenges and opportunities. In *Emerging Microbes and Infections* (Vol. 9, Issue 1, pp. 194–206). https://doi.org/10.1080/22221751.2020.1713707

Lo, Y. C., Rensi, S. E., Torng, W., & Altman, R. B. (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, *23*(8), 1538–1546.

https://doi.org/10.1016/j.drudis.2018.05.010

Lu, W., Zhang, R., Jiang, H., Zhang, H., & Luo, C. (2018). Computer-aided drug design in epigenetics. In *Frontiers in Chemistry* (Vol. 6, Issue MAR, p. 57). https://doi.org/10.3389/fchem.2018.00057

Luxminarayan, L., Neha, S., Amit, V., & Khinchi, M. P. (2019). The Stages of Drug Discovery and Development Process. *Asian Journal of Pharmaceutical Research and Development*, *7*(6), 62–67.

M Barry, S. (2014). Trial, Error, and Breakthrough: A Review of HIV Vaccine Development. *Journal of AIDS & Clinical Research*, *05*(11), 359. https://doi.org/10.4172/2155-6113.1000359

Maga, G., Veljkovic, N., Crespan, E., Spadari, S., Prljic, J., Perovic, V., Glisic, S., & Veljkovic, V. (2013). New in silico and conventional in vitro approaches to advance HIV drug discovery and design. *Expert Opinion on Drug Discovery*, *8*(1), 83–92. https://doi.org/10.1517/17460441.2013.741118

Master, A. M., Rodriguez, M. E., Kenney, M. E., Oleinick, N. L., & Sen Gupta, A. (2010). Delivery of the photosensitizer Pc 4 in PEG–PCL micelles for in vitro PDT studies. *Journal of Pharmaceutical Sciences*, *99*(5), 2386–2398. https://doi.org/10.1002/jps

Metwally, A. A., & Hathout, R. M. (2015). Computer-Assisted Drug Formulation Design: Novel Approach in Drug Delivery. *Molecular Pharmaceutics*, *12*(8), 2800–2810. https://doi.org/10.1021/mp500740d

Muhammad, I., & Yan, Z. (2015). Supervised Machine Learning Approaches: a Survey. *ICTACT Journal on Soft Computing*, *05*(03), 946–952. https://doi.org/10.21917/ijsc.2015.0133

Nair, A. (2007). Computational biology & bioinformatics: a gentle overview. *Communications of the Computer Society of India*, *January*, 1–13.

http://sites.google.com/site/printachuth/BINFTutorialV5.0CSI07.pdf

Njogu, P. M., Guantai, E. M., Pavadai, E., & Chibale, K. (2016). Computer-Aided Drug Discovery Approaches against the Tropical Infectious Diseases Malaria, Tuberculosis, Trypanosomiasis, and Leishmaniasis. *ACS Infectious Diseases*, *2*(1), 8–31. https://doi.org/10.1021/acsinfecdis.5b00093

Noorbakhsh, J., Chandok, H., Karuturi, R. K. M., & George, J. (2019). Machine Learning in Biology and Medicine. *Advances in Molecular Pathology*, *2*(1), 143–152. https://doi.org/10.1016/j.yamp.2019.07.010

Ou-Yang, S. S., Lu, J. Y., Kong, X. Q., Liang, Z. J., Luo, C., & Jiang, H. (2012). Computational drug discovery. In *Acta Pharmacologica Sinica* (Vol. 33, Issue 9, pp. 1131–1140). https://doi.org/10.1038/aps.2012.109

Parvizpour, S., Pourseif, M. M., Razmara, J., Rafi, M. A., & Omidi, Y. (2020). Epitope-based vaccine design: a comprehensive overview of bioinformatics approaches. *Drug Discovery Today*, *25*(6), 1034–1042. https://doi.org/10.1016/j.drudis.2020.03.006

Prakash, N., & Devangi, P. (2010). Drug Discovery. *Journal of Antivirals and Antiretrovirals*, *2*(4), 063–068. https://doi.org/10.4172/jaa.1000025

Rahman, M. M., Karim, M. R., Ahsan, M. Q., Khalipha, A. B. R., Chowdhury, M. R., & Saifuzzaman, M. (2012). Use of computer in drug design and drug discovery: A review. *International Journal of Pharmaceutical and Life Sciences*, *1*(2), 1–21. https://doi.org/10.3329/ijpls.v1i2.12955

Santos, L. H., Ferreira, R. S., & Caffarena, E. R. (2015). Computational drug design strategies applied to the modelling of human immunodeficiency virus-1 reverse transcriptase inhibitors. In *Memorias do Instituto Oswaldo Cruz* (Vol. 110, Issue 7, pp. 847–864). https://doi.org/10.1590/0074-02760150239

Shapshak, P., Somboonwit, C., Sinnott, J. T., Menezes, L. J., Kangueane, P., Balaji, S., &

Chiappelli, F. (2019). Global virology III: Virology in the 21st century. In *Global Virology III: Virology in the 21st Century*. https://doi.org/10.1007/978-3-030-29022-1

Sharma, A., & Sharma, A. (2018). Machine Learning: A Review of Techniques of Machine Learning. *Journal of Applied Science and Computations*, *5*(December), 538–541.

Shin, S. Y. (2016). Recent update in HIV vaccine development. *Clinical and Experimental Vaccine Research*, *5*(1), 6. https://doi.org/10.7774/cevr.2016.5.1.6

Singh, D. B. (2020). *Computer-Aided Drug Design*.

Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. In *Pharmacological Reviews* (Vol. 66, Issue 1, pp. 334–395). https://doi.org/10.1124/pr.112.007336

Solanke, A., & Tribhuvan, K. (2017). *Genomics : An Integrative Approach for Molecular Biology Genomics : An Integrative Approach for Molecular Biology*. *October 2015*.

Soria-Guerra, R. E., Nieto-Gomez, R., Govea-Alonso, D. O., & Rosales-Mendoza, S. (2015). An overview of bioinformatics tools for epitope prediction: Implications on vaccine development. In *Journal of Biomedical Informatics* (Vol. 53, pp. 405–414). Academic Press Inc. https://doi.org/10.1016/j.jbi.2014.11.003

Usha, T., Shanmugarajan, D., Goyal, A. K., Kumar, C. S., & Middha, S. K. (2018). Recent Updates on Computer-aided Drug Discovery: Time for a Paradigm Shift. *Current Topics in Medicinal Chemistry*, *17*(30), 3296–3307. https://doi.org/10.2174/1568026618666180101163651

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. In *Nature Reviews Drug Discovery* (Vol. 18, Issue 6, pp. 463–477). https://doi.org/10.1038/s41573-019-0024-5

Varnek, A., & Baskin, I. (2012). Machine learning methods for property prediction in

chemoinformatics: Quo Vadis? *Journal of Chemical Information and Modeling*, *52*(6), 1413–1437. https://doi.org/10.1021/ci200409x

Worachartcheewan, A., Songtawee, N., Siriwong, S., Prachayasittikul, S., Nantasenamat, C., & Prachayasittikul, V. (2018). Rational Design of Colchicine Derivatives as anti-HIV Agents via QSAR and Molecular Docking. *Medicinal Chemistry*, *15*(4), 328–340. https://doi.org/10.2174/1573406414666180924163756

World Health Organization. (2020). *Global HIV & AIDS statistics. Fact sheet*. UNAIDS. https://www.unaids.org/en/resources/fact-sheet

Xiang, M., Cao, Y., Fan, W., Chen, L., & Mo, Y. (2012). Computer-Aided Drug Design: Lead Discovery and Optimization. *Combinatorial Chemistry & High Throughput Screening*, *15*(4), 328–337. https://doi.org/10.2174/138620712799361825

Yamamoto, M., & Nakao, M. (2001). Bioinformatics and physiology--measurement, analysis, and interpretation of biological data. In *Journal of the Physiological Society of Japan* (Vol. 63, Issue 1). https://doi.org/10.1055/s-0038-1638103

Yu, W., & Jr, A. D. M. (2017). Chapter 5 Computer-Aided Drug Design Methods. *Antibiotics: Methods and Protocols*, *1520*, 85–106. https://doi.org/10.1007/978-1-4939-6634-9

Zheng, M., Liu, X., Xu, Y., Li, H., Luo, C., & Jiang, H. (2013). Computational methods for drug design and discovery: Focus on China. *Trends in Pharmacological Sciences*, *34*(10), 549–559. https://doi.org/10.1016/j.tips.2013.08.004