COMPUTER–AIDED APPROACHES TO DISCOVERY OF NOVEL DRUGS
AGAINST THE HUMAN HOOKWORM NECATOR AMERICANUS
(NEMATODA: ANCYLOSTOMATIDAE)

BY

ODAME AGYAPONG
(10204283)

THIS THESIS IS SUBMITTED TO THE UNIVERSITY OF GHANA, LEGON IN
PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF
MPHIL BIOMEDICAL ENGINEERING DEGREE

DEPARTMENT OF BIOMEDICAL ENGINEERING
COLLEGE OF BASIC AND APPLIED SCIENCES
UNIVERSITY OF GHANA

JULY, 2017
DECLARATION

I, ODAME AGYAPONG, do hereby declare that this work, COMPUTER–AIDED APPROACHES TO DISCOVERY OF NOVEL DRUGS AGAINST THE HUMAN HOOKWORM NECATOR AMERICANUS (NEMATODA: ANCYLOSTOMATIDAE), with the exception of the cited references, was written and submitted by me in the University of Ghana from AUGUST 2015 to JULY 2017, under the supervision of Dr. Samuel K. Kwofie and Prof. Michael Wilson.

I further declare that this work has not been submitted to University of Ghana or any other university.

...........................................  ...........................................
Odame Agyapong (10204283)  (Date)
(Student)

...........................................  ...........................................
Samuel K. Kwofie (PhD)  (Date)
(Principal Supervisor)

...........................................  ...........................................
Prof. Michael Wilson  (Date)
(Co-Supervisor)
ABSTRACT

There is a crucial need to develop novel anthelminthic drugs due to the mounting disease burden and increasing evidence of hookworm resistance to drugs such as albendazole and mebendazole, which for decades have been used to treat the infection. Consequently, it is exigent to develop alternative drugs with improved therapeutic efficacy. Natural products due to their unique active ingredients have been shown to possess exceptional structures with chemical diversity that is unmatched by any synthetic libraries. It is imperative to leverage natural products to augment hookworm drug discovery. Therefore, this study aimed to: (i) identify potential novel anthelminthic lead compounds by screening African natural product-derived ligands against beta tubulin of Necator americanus, a known hookworm receptor and (ii) develop support vector machine-based proteochemometric modelling (PCM) for bioactivity profiling of beta tubulins receptors.

The 3D structure of the beta tubulin of hookworm with UniProt entry W2T758, was generated using homology modelling. The model was subjected to molecular dynamics simulations and active site interactions prediction. The first set of ligand libraries comprising 885 natural product compounds obtained from African medicinal plants database (AfroDb) combined with Dichapetalin A, were screened against the receptor. ZINC14760755 and ZINC28462577 compounds were found to be potential leads due to promising binding affinity, active site interactions and pharmacokinetic profiles. Additionally, a second set comprising 2297 compounds derived from Northern African Natural Product Database (NANPDB) were virtually screened. The compound S,5Z,8Z,11Z,13E,17Z-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylcosa-5,8,11,13,17-pentaen-1-one exhibited plausible binding affinity, toxicity and
pharmacokinetic profile. The aforementioned natural compounds are potential leads which can be experimentally characterised for possible pre-clinical trials.

Support vector machine based proteochemometric modelling was also developed to predict the bioactivity relations between beta tubulin variants and small compounds using an interaction dataset retrieved from BindingDB. The model achieved reasonably good performance with a ROC-AUC of 87%, an MCC of 0.75 and a classification error of approximately 4%, although it was trained on a small dataset. The model allows the prediction of the likelihood of interactions between query datasets comprising ligands in SMILES format and protein sequences of beta tubulin targets. In future, larger bioactive datasets of beta tubulins originating from high throughput experiments can be utilised to possibly enhance the performance of the hookworm PCM model.
DEDICATION

I dedicate this thesis to the Almighty God for his protection, knowledge and guidance throughout the research journey.
ACKNOWLEDGEMENTS

I would like to first extend my deepest and sincere gratitude to my supervisors whose useful guidance has brought me this far. I would not have been able to finish this research without their expertise, insightful comments and considerable guidance. I would like to specially thank Dr Samuel Kojo Kwofie, my main supervisor, for patiently guiding me to develop my background in computational drug discovery and machine learning. His insightful comments, enthusiasm, corrections of thesis writing, suggestions, guidance and scholarly knowledge were greatly useful. I also want to take this opportunity to thank my co-supervisor, Professor Michael Wilson, for the financial support given to me throughout this research. I am deeply grateful for all the support he gave me including correction of the thesis and the willingness to always support me wherever and whenever needed. I wish to also express special thanks to all the lecturers at the Department of Biomedical Engineering, University of Ghana for their useful criticisms during my thesis defence. Special thanks to Dr Whelton Miller from the University of Pennsylvania, USA, for guidance on molecular dynamics simulation of the beta tubulin of *Necator americanus*. Also, I am grateful to Dr. Christian Parry from the Howard University, USA and Dr Michael Bosu from Waikato Institute of Technology, New Zealand, for their useful suggestions. Last but not the least, I would finally like to thank my family and friends for all the encouragement, support and love for the completion and success of this research.
# TABLE OF CONTENTS

DECLARATION ........................................................................................................................................... i

ABSTRACT ................................................................................................................................................... ii

DEDICATION ................................................................................................................................................ iv

ACKNOWLEDGEMENTS ........................................................................................................................... v

TABLE OF CONTENTS ............................................................................................................................. vi

LIST OF FIGURES ...................................................................................................................................... x

LIST OF TABLES ......................................................................................................................................... xii

LIST OF ABBREVIATIONS ........................................................................................................................ xiv

CHAPTER 1 .................................................................................................................................................... 1

INTRODUCTION ........................................................................................................................................... 1

1.1 Background ........................................................................................................................................... 1

1.2 Problem Statement, Rational and Overall Goal of the Study ............................................................... 3

1.2.1 Problem statement ........................................................................................................................... 3

1.2.2 Rationale of the study ....................................................................................................................... 4

1.2.3 Overall goal of the study .................................................................................................................. 6

1.2.4 Expected outcome (Contribution to knowledge) ............................................................................. 7

CHAPTER 2 ................................................................................................................................................... 8

LITERATURE REVIEW ............................................................................................................................. 8

2.1 The Hookworm, Necator americanus ....................................................................................................... 8
2.1.1 Life cycle .......................................................................................................................... 8

2.1.2 Geographical distribution of nematode (hookworm) infections ......................... 10

2.2 Hookworm Drug Targets ............................................................................................... 11

2.2.1 Beta-tubulin .................................................................................................................. 11

2.2.2 Other potential targets................................................................................................. 12

2.3 Existing Treatment Methods and their Molecular Targets ........................................ 13

2.4 Natural Products (NP) and their Utility as Anthelmintic Therapeutics: ............. 15

2.4.2 Other naturally derived compounds effective against hookworm .................. 17

2.5 Computer-Aided Drug Design (CADD) .................................................................... 17

2.5.1 Economic significance and time factor of CADD .................................................. 18

2.5.3 Structure based drug design ...................................................................................... 20

2.5.4 Ligand based drug design ......................................................................................... 23

2.5.6 Proteochemometric modelling (PCM) ..................................................................... 26

2.6 Recent Efforts in Hookworm Drug Discovery ............................................................. 32

CHAPTER 3 .............................................................................................................................. 34

METHODS ............................................................................................................................... 34

3.1 Template Identification and Homology Modelling of Proteins ............................... 34

3.2 Molecular Dynamic Simulations of Modelled Protein .............................................. 37

3.3 Prediction and Analysis of Binding Site ...................................................................... 39

3.4 Protein Preparation ......................................................................................................... 39
4.4 Virtual Screening Analysis results ................................................................. 60
4.5 Interaction Profile using LIGPLOT .................................................................. 65
4.6 ADME Prediction and Pharmacokinetic Properties ......................................... 73
4.7 Toxicity Prediction Analysis .......................................................................... 77
4.8 Scaffold Analysis ............................................................................................ 78
4.9 Proteochemometric Modelling ...................................................................... 81
  4.9.1 Exploratory principal component analysis (PCA) of compounds and target
datasets ............................................................................................................. 83
  4.9.2 Model development .................................................................................. 85
    4.9.2.1 Model validation ................................................................................ 85
CHAPTER 5 ........................................................................................................... 88
CONCLUSION AND RECOMMENDATION ......................................................... 88
REFERENCES ...................................................................................................... 91
APPENDICES ...................................................................................................... 103
LIST OF FIGURES

Figure 2.1. Life cycle of Hookworm .......................................................... 9
Figure 2.2. Global distribution of the human hookworm infection ................. 10
Figure 2.3. The drug design pipeline ................................................................ 20
Figure 2.4. Structure based drug design .......................................................... 23
Figure 2.5. Ligand based drug design ............................................................... 24
Figure 2.6. Proteochemometric modelling ....................................................... 27
Figure 3.1. Workflow of protein modelling to scaffold analysis ..................... 35
Figure 3.2. Workflow for PCM modelling of beta tubulin bioactivity profiling .... 45
Figure 4.1. A pairwise sequence alignment between the beta tubulin sequence of *N. americanus* and D chain of the crystal structure with PDB ID, 5c8y ................................................................. 54
Figure 4.2. Predicted binding site from I-TASSER and rendered in PYMOL .......... 55
Figure 4.3. 3D model of the beta tubulin of *N. americanus* ............................. 55
Figure 4.4. Ramachandran plot of beta tubulin model from *N. americanus* ........ 57
Figure 4.5. Errat plot ...................................................................................... 58
Figure 4.6. RMSD plot of the Molecular dynamic simulation using GROMACS .... 59
Figure 4.7. Predicted colchicine binding site of beta tubulin from


Necator americanus .................................................................................................................. 60

Figure 4.8. Docking pose of ZINC14760755- beta-tubulin receptor complex .................. 63

Figure 4.9. Docking pose of Dichapetalin A and albendazole beta tubulin receptor complex .............................................................................................................................. 64

Figure 4.10. Docking pose of S,5Z,8Z,11Z,13E,17Z-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylcosa-5,8,11,13,17-pentaen-1-one from NANPDB ............................................................................................................................................. 65

Figure 4.11. Interaction profile of ZINC14760755 and ZINC28462577 .......................... 68

Figure 4.12. Interaction profile of Dichapetalin A and albendazole .................................. 69

Figure 4.13. Interaction profile of S,5Z,8Z,11Z,13E,17Z-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylcosa-5,8,11,13,17-pentaen-1-one from NANPDB as predicted by LIGPLOT ........................................................................................................................................... 70

Figure 4.14. A bar plot of scaffold counts versus the ring systems of 201 compounds from list A compared to list B and list C ................................................................................................................. 80

Figure 4.15. Distribution of response variable (class) in the dataset ................................. 82

Figure 4.16. Chemical and biological space (compound–target interaction space) of beta tubulin-inhibitor dataset ......................................................................................................................... 84

Figure 4.17. Area under Receiver Operating Curve (AUC) .................................................. 87
LIST OF TABLES

Table 1.1. Existing therapeutic agents and their mechanism of action.......................... 14

Table 3.1. Complete dataset used for PCM ................................................................. 47

Table 4.1. Results of the molecular docking scores of the top 20 ligands from
AfroDB library plus Dichapetalin A and albendazole........................................... 61

Table 4.2. Results of the molecular docking scores of the top 20 ligands
from the Northern African Natural Product Database......................................... 62

Table 4.3. Results of number of hydrogen-bonds/hydrophobic-bonds and contact
residues of top ten ligands from AfroDB, and that of Dichapetalin A
and albendazole. ........................................................................................................... 71

Table 4.4. Results of number of hydrogen-bonds/hydrophobic-bonds and contact
residues of top ten ligands from the NANPDB. ......................................................... 72

Table 4.5. Results of ADME prediction of top ten virtually screened compounds and
that of Dichapetalin A and albendazole as predicted by SwissADME............... 75

Table 4.6. Results of ADME prediction of top ten ranking compounds
from NANPDB.................................................................................................. 76

Table 4.7. Toxicological profile results of top ten ranking compounds from both set of
virtual library compounds as predicted by DataWarrior........................................ 78
Table 4.8. Scaffold diversity analysis of natural products and anthelmintics ................... 81

Table 4.9. Proteins and compounds descriptors used in the development of the model .. 82

Table 4.10. SVM model parameters and evaluation of classification performance ....... 87
LIST OF ABBREVIATIONS

ADME – Absorption, Distribution, Metabolism and Excretion
CADD – Computer-Aided Drug Design
CV – Cross Validation
DOPE – Discrete optimised potential energy
EST – Expressed Sequence TAG
FDA – Food and Drugs Authority
GTP – Guanosine-5-triphosphate
H-bond – Hydrogen bond
HTS – High-throughput Screening
LBDD – Ligand Based Drug Design
MDA – Mass Drug Administration
NANPDB – Northern African Natural Product Database
NMR – Nuclear Magnetic Resonance
PCM – Proteochemometric Modelling
PDB – Protein Data Bank
QSAR – Quantity Structure Activity Relationship
SBDD – Structure Based Drug Design
SVM – Support Vector Machine
CHAPTER 1

INTRODUCTION

1.1 Background

The need for development of new drugs cannot be overemphasised in light of current global burden of debilitating diseases such as hookworm infection. Drug discovery relies heavily on protein structures and the docking of potential compounds in the search for lead drugs. The latter is considered as a more deterministic approach for finding drugs against diseases and has led to what is called rational drug design [1].

Benzimidazoles including albendazole and mebendazole have been around for many years as potent therapeutic drugs for hookworm infections, but there are concerns about their low efficacies and drug resistance [2–5]. There is therefore, the need to develop alternative novel drugs with improved therapeutic efficacy. Natural products present an excellent opportunity for novel drug discovery. Researchers from Noguchi Memorial Institute for Medical Research and the Chemistry Department of the University of Ghana, for example, have found potential anthelminthic activities of a natural compound, Dichapetalin A., against hookworm infection highlighting the importance for studies to determine further potential anthelminthic drug from natural products [6]. Natural products possess highly diverse and novel scaffold structures that make them potentially better drug candidates than synthetic compounds [7–12]. The discovery of unique scaffolds different from the known anthelminthic drugs is key to the identification of different mechanisms of action [13]. In addition, a greater scaffold diversity may suggest a wide chemical space coverage thus;
increasing the chances of identifying compounds with interaction to more biological targets to elicit anthelminthic activity [13]. For this reason, scaffold diversity analysis is normally essential for exploring the presence of unique scaffolds and structures within compounds that show favourable binding than benzimidazoles.

*In silico* approaches are generally advantageous methods that can be used to provide more insight into the bioactivity profile of experimentally determined drugs and potentially predict additional drug targets implicated in various disease mechanisms and biological pathways. They are very useful strategies for shortening the time, reducing associated cost and effort required for drug development [1]. These computationally aided drug design approaches, including molecular docking and proteochemometric modelling can unravel the potential anthelminthic properties of natural compounds. In addition, it has the potential to identify new receptors, novel drug leads and different modes of action of potential natural product therapeutic agents and their implicated biological pathways.

Two major *in silico* approaches comprising molecular docking and proteochemometric modelling are the focus in this study. Molecular docking is an approach which involves the interaction between two or more molecules to give a stable complex with optimized conformation and less binding free energy [14]. Proteochemometric modelling (PCM), on the other hand, is a computational method that can predict the bioactivity relations between a series of ligands and a series of targets [15].
Overall, this thesis provides an in-depth look at computer-aided design for predicting anti-tubulin activities and computational techniques in optimizing leads from naturally derived drugs specific for hookworm. The computer-aided techniques cover structure-based drug design studies which includes homology modelling, binding site identification, docking to potential binding pockets, virtual screening, toxicity prediction, scaffold analysis and construction of a support vector machine based proteochemometric classification model.

1.2 Problem Statement, Rational and Overall Goal of the Study

1.2.1 Problem statement

A major problem with current anthelminthic treatments is drug resistance [16]. Development of varying degrees of drug tolerance among different species of nematodes including *Necator americanus* have been widely reported. This is largely due to the frequent and unnecessary use of anthelminthic drugs or increasing drug pressure especially in mass drug administration [2]. The increasing problem of drug resistance is a major concern because the older active drugs are becoming less effective, thereby drawing the attention of major stakeholders including the World Health Organisation (WHO) [17]. Drug development is a highly resource-intensive, time-consuming and expensive endeavour. Most research in hookworm drug discovery programs have employed low throughput methods to isolate, characterise and evaluate the anthelminthic activities of both synthetic and natural compounds. These techniques are expensive and laborious, and some isolated natural products that although demonstrated anthelminthic activities have not yielded the desired results for further drug development. Moreover, after an exhaustive
literature review, it was found that very little research has been done in the search for natural products or naturally inspired products against the human hookworm, *Necator americanus*.

Efforts in screening natural compounds both *in vivo* and *in vitro*, involve the purchase of millions of compounds as libraries from pharmaceutical companies. There appears to be diminishing investment by pharmaceutical companies in therapeutic research areas owing to the prospect of drug failure that could lead to huge financial losses. This is because these compounds sometimes fail the basic Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) testing and the cost of these compounds is so expensive that it becomes an economic burden in drug discovery thus serving as a rate limiting step. Moreover, most debilitating diseases such as the hookworm disease are endemic in resource-constrained countries which may not be able to afford the cost of newly developed drugs. There is, therefore, the onus on all key stakeholders in these countries is to identify novel potential anthelmintic agents especially from natural products.

### 1.2.2 Rationale of the study

In view of the widespread increasing resistance of helminths to current anthelmintic drugs and its concomitant high cost, complicated drug administration procedures and high-risk, and long-term endeavour of drug development, it has become important to identify alternative drug treatment methods and agents. This should be done by tailoring more natural products to current drug development pipelines using computational strategies.
Computational methods including rational drug design present advantages in the prioritisation of potential lead compounds for preclinical development within the drug design landscape. Computational methods have been used in hookworm drug discovery, however, the screening datasets need to be expanded with natural products or compounds with similar properties to natural products (naturally inspired). Limited use of computational techniques leads to the whole sale application of expensive laboratory techniques from drug screening to pre-clinical trial level.

In silico approaches that can be used include molecular docking, pharmacophore modelling, similarity search, Quantitative Structure Activity Relationship (QSAR) and an emerging area, Proteochemometric Modelling (PCM). However, these computational techniques have their successes and pitfalls. PCM has been shown from previous studies to be robust in predicting bioactivity profiles of untested compounds and protein targets [18–21]. Both molecular docking and PCM, however, have their limitations; molecular docking is sometimes challenged in terms of accuracy and speed while PCM is unable to predict compound bioactivity of unrelated targets [21]. After an exhaustive literature research, there were no reports on using PCM for anti-tubulin activity of chemical compounds and very limited study on application of computational docking of natural products against beta tubulins of nematodes including hookworms [22]. Given the limited exploration of molecular docking of natural products against beta tubulins from hookworm and the unexplored implementation of PCM in the prediction of anti-tubulin activity, there is the need to create an integrated approach to anthelminthic drug discovery combining
proteochemometric modelling and molecular docking for the prediction of novel active compounds that target tubulins.

1.2.3 Overall goal of the study

The main aims of this study were: (i) the application of computational methods in the identification of novel anthelminthic drugs from natural products, and (ii) the use of a support vector machine based proteochemometric modelling to predict the bioactivity profile of compounds to beta tubulin targets in hookworm.

1.2.3.1 Specific objectives

The specific objectives of the study are:

1. Homology modelling of the 3D structure of beta tubulin of *Necator americanus*.

2. Virtual screening of naturally derived compounds for the identification of potential anthelminthic agents.

3. *In silico* evaluation of pharmacological, drug-likeness and toxicity profiles of lead compounds.

4. Comparative analysis of scaffolds of the docked natural products and known synthetic anthelminthics, specifically albendazole and mebendazole.

5. Preliminary exploration of proteochemometric based machine learning model as a plausible technique for bioactivity profiling of beta tubulin receptors.
1.2.4 Expected outcome (Contribution to knowledge)

It is expected that this research will accelerate hookworm drug design effort by including and prioritising potential naturally derived lead compounds as alternative anthelmintic drugs. The support vector machine based proteochemometric predictive model that was built in this research can be further explored by leveraging experimental bioassay of hookworm activity for the development of an enhanced model for predicting anthelmintic activity.
CHAPTER 2

LITERATURE REVIEW

2.1 The Hookworm, *Necator americanus*

Hookworm infection remains a significant health burden globally. The worms currently infect over 700 million persons in resource limited countries and result in 135,000 deaths annually [23, 24]. Areas largely affected by hookworm include: South Asia, Latin America and the Caribbean, the Middle East, and North Africa [24 – 26]. The morbidity and mortality of helminths far exceeds that of other tropical diseases including African trypanosomiasis and dengue [27, 28]. Children and pregnant women are more susceptible. Hookworm infection causes stunted growth and diminished physical fitness as well as impaired memory and cognition in children. It also causes infant death in pregnant women. According to WHO, hookworm infections is classified as moderate in individuals when it produces from 2000 to 3999 eggs per gram (epg) of faeces and heavily burdened in individuals at 4,000 epg or more [28, 29].

2.1.1 Life cycle

Human hookworm infection is primarily caused by *Ancylostoma duodenale* and *Necator americanus* [27] with the latter, accounting for more than 85% of all hookworm infections [23]. Hookworm gains access into a human by penetrating the skin, undergoes growth and development, and tends to reside more in the duodenum. The life cycle of *N. americanus* begins with eggs embryonating in the soil where under favourable conditions, the first-stage larvae hatch and feed on environmental microbes. They then molt twice to become
infective third-stage larvae (IL3). These larvae penetrate the skin through the epidermis causing cutaneous larva migrants that invades the human circulatory system as shown in Figure 2.1. After skin penetration, they penetrate the pulmonary alveoli, migrate to the pharynx through the bronchial tree and locate the small intestines where they reside and mature into adults. Adult hookworms become attached to the intestinal lining and rupture capillaries to gain access to blood. Female worm can release about 10,000 eggs which when excreted may contaminate soil and water. The egg hatches in the soil, releasing a larva that undergoes various larval stages before a new host infection. Laceration of the capillaries can normally lead to anaemia. The life cycle of hookworm is illustrated in the Figure 2.1. Hookworms also play immunomodulatory roles suggesting a mechanism by which they can be used to suppress autoimmune, allergic and atopic dermatitis disease [31–33].

**Figure 2.1. Life cycle of Hookworm** [32]. Hookworm eggs transferred through faeces hatch in the faces. (1). The larva that is released mature in the faeces and become infective L3 after few days of 5 to 10 (2-3). L3 penetrates the skin upon contact and migrate to the heart and lungs through the blood vessel (4) L3 locates the pharynx though the pulmonary alveoli and bronchial tree. The larvae then migrate to the small intestines (5), and attaches itself to the intestinal wall.
2.1.2 Geographical distribution of nematode (hookworm) infections

The most affected regions are in the tropic and subtropic with *A. duodenale* and *N. americanus* geographically restricted somehow [34, 35]. An understanding of the epidemiology across countries and trends over time is of great importance to enable strategies of a cost-effective mass administration intervention programs. In Ghana, hookworm infection is seasonal with a high prevalence between April and August [34]. The disease is co-endemic with *Oesophagostomum biurcum* in Northern Togo and Ghana with a 50% greater prevalence [35].

![Figure 2.2. Global distribution of the human hookworm infection](image)

Figure 2.2. Global distribution of the human hookworm infection [191]. The sub-Saharan Africa and eastern Asia shows the highest prevalence of hookworm infection.
2.2 Hookworm Drug Targets

2.2.1 Beta-tubulin

Beta tubulin is a subunit of the microtubule which, plays a crucial role in cell division and maintenance of the cytoskeleton. It binds to two molecules of guanosine-5-triphosphate (GTP), at the positive end of microtubules \[36\]. Beta tubulin has so far been exploited as a crucial target for anthelminthic and as a target for several other compounds \[37\]. Tubulin can be selectively targeted by benzimidazole anthelmintics (including albendazole, fenbendazole) which inhibits microtubule polymerization \[38\]. Benzimidazoles confer their anthelmintic effect on susceptible nematodes by binding to their beta-tubulin resulting in the subsequent prevention of microtubule polymerisation causing destabilisation of the intracellular processes and cellular division within the parasite and an overall immobility effect \[39\]. Glu198, Phe167 and Phe200 are implicated in anthelmintic resistance within the colchicine binding site of the beta tubulin of hookworm \[40\]. Genetic changes such as single nucleotide polymorphisms (SNPs) in beta tubulin have been widely reported to convey the nematodic resistance in several parasitic nematodes and the human hookworm \textit{N. americanus}. These mutations have been found to occur at codons 167,198 and 200 \[40 – 42\]. In many cases, the changes lead to the substitution of phenylalanine with tyrosine at amino acid positions 167 and 200, and glutamate with alanine at amino acid position 198 \[43\]. These mutations or SNPs have been reported to be predominant in several benzimidazole-resistant nematodes.
Benzimidazoles have also been used as fungicides, bacteriostatics, insecticides, antivirals and anti-cancer agents due to low toxicity in mammals. They are believed to bind close to the colchicine-binding site on the beta tubulin molecule and disrupt microtubule function, but the precise mechanism of action of anthelminthic are poorly understood [30, 44 – 46]. They have been enormously successful drugs but their continued use as antiparasitics is being threatened by the development of resistance [46].

2.2.2 Other potential targets

Other potential drug targets include ion channels which are pore-forming membrane proteins and protein complexes that play important role in electrical signalling and fast synaptic transmission in cells [47]. Activation of ion channels makes them particularly useful as targets for anthelmintics [48]. Examples of ion channels in helminths include nicotinic Acetylcholine receptors (nAchR), choline receptor, slo-1 K⁺ channels; latrophilin receptors, voltage-gated Ca²⁺ channels and Gamma Aminobutyric Acid (GABA) receptors [48]. Glutamate-gated Chloride (GluCl) channels are also macrocyclic lactones targets, including some avermectin anthelmintics [49]. Cell signalling pathway targets such as G protein-couple receptors (GPCR) are implicated in the pathology of many diseases including hookworms. Neuropeptides, which are examples of G protein-couple receptors, are presumed to be either neurotransmitters or neurohormones involved in the regulation of both physiology and behaviour of nematodes [23]. Others include proteases [50], kinases [51], hydrolases, and catalases which are all required to help the adult worms feed on blood. Single-domain serine proteases in *N. americanus* for example are potential targets for immunomodulation since they play a key role in immunomodulation.
2.3 Existing Treatment Methods and their Molecular Targets

Efforts to clearly understand and combat hookworm disease dates to 1916 when the Department of Helminthology at the Johns Hopkins School of Hygiene and Public Health was set up to combat hookworm using quantitative methods, and thus providing a framework for understanding the pathogenesis of the disease. Their efforts in addition to many others [27] resulted in the understanding the disease and a renewed interest in its control with chemotherapeutic agents.

Recent years have seen notable advances in several control and treatment methods for improved therapeutic intervention. Mass Drug Administration (MDA) is currently used as a strategy for treating hookworm and it usually involves a combinatorial administration of benzimidazoles (albendazole or mebendazole) along with others including pyrantel pamoate, imidazothiazoles (levamisole), oxantel, avermectin etc. It is, however, noteworthy that drug treatment against hookworm whether through MDA or other treatment strategies do not prevent re-infection. This is because, existing drugs are becoming significantly less effective due to its repeated and excessive use leading to drug resistance and treatment failures [52]. The parasitic resistance to the current anthelmintics has been attributed to the following: changes in drug translocation; receptor modification or post receptor modification and mutation [4, 44]. Moreover, there are issues of unknown mechanisms of action of some of these drugs after their recurrent and frequent usage [53, 54]. Table 1 shows the most widely used anthelmintic drugs and their molecular targets.
Table 1.1 Existing therapeutic agents and their mechanism of action [2]. All known anthelmintic drugs are being resisted by the nematodes.

<table>
<thead>
<tr>
<th>Anthelmintic Group</th>
<th>Examples</th>
<th>Target</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benzimidazoles</td>
<td>Albendazole, mebendazole, fenbendazole</td>
<td>Nematodes, trematodes; β-tubulin</td>
<td>Resistance, Ineffective</td>
</tr>
<tr>
<td>Imidazothiazoles</td>
<td>Levamisole</td>
<td>Nematodes; nAchR agonists</td>
<td>Resistance, Ineffective</td>
</tr>
<tr>
<td>Tetrahydropyrimidines</td>
<td>Pyrantel, oxantel, morantel</td>
<td>Nematodes; nAchR agonists</td>
<td>Resistance, Ineffective</td>
</tr>
<tr>
<td>Amino-acetonitriles</td>
<td>Monepantel</td>
<td>Nematodes; choline receptor agonists</td>
<td>Resistance, Ineffective</td>
</tr>
<tr>
<td>Tribendimidine</td>
<td></td>
<td>Nematodes; nAchR agonist</td>
<td>Resistance, Ineffective</td>
</tr>
<tr>
<td>Spiroindoles</td>
<td>Derquantel</td>
<td>Nematodes; nAchR antagonist</td>
<td>Resistance, Ineffective</td>
</tr>
<tr>
<td>Macrocyclic lactones</td>
<td>Ivermectin, moxidectin</td>
<td>Nematodes; GluCl activation</td>
<td>Resistance, Ineffective</td>
</tr>
<tr>
<td>Piperazines</td>
<td>Piperazine</td>
<td>Nematodes; GABA receptor agonist</td>
<td>Resistance, Ineffective</td>
</tr>
</tbody>
</table>
2.4 Natural Products (NP) and their Utility as Anthelminthic Therapeutics

Nature has given us abundantly rich sources of medicinal agents that are used to treat a lot of diseases. Some naturally derived drugs that have become forerunner drugs in modern pharmaceutical care include quinone, cocaine, salicylic acid, *digitalis*, morphine, penicillin, ergotamine, reserpine, paclitaxel, digoxine, cyclosporine, and Vitamin A [55, 56]. Evidence of documented practice of medicine was recorded on ancient Egyptian “the Ebers Papyrus” as far back as 1500BC [57]. The papyrus listed over 700 drugs most of which were plant derived with detailed formulation and use [57]. The first natural product to be isolated and analysed was morphine in the 19\textsuperscript{th} century. Thousands of natural products described in that era are still relevant today although some are no longer in use. Natural products for their unique active ingredients are the important cornerstone in the pharmaceutical industry. They are known to possess enormous, exceptional structural and chemical diversity unmatched by any synthetic library [10]. Analysis of their chemical formulation has revealed that about more than 50\% of them pass Lipinski rule of five for drug-likeness [58, 59]. The remainder of NP, however, are characterised by higher molecular weights and more rotatable bonds with desirably low logP values. This makes natural products more absorbable than the synthetic counterparts. Interestingly, about 40\% of the chemical scaffolds found in natural products cannot be accounted by today’s drugs [60]. According to the Scripps Research Institute (2016) [61], from 1940s to date, 131 (74.8\%) out of 175 small molecule anticancer drugs are naturally inspired. Interestingly, half of the 20 approved small molecule New Chemical Entities (NCEs) in 2010 are natural products [62]. The perceived efforts geared towards exploring natural products as anti-cancer lead compounds must be extended to infectious diseases such as hookworm drug
discovery. Ghana, including the rest of the Africa continent, is endowed with vast natural flora and fauna, with the potential to exploit to identify new natural product derived lead compounds.

2.4.1 Contribution of natural products to anthelminthic therapy

Some natural products have for many years been used for the treatment of the human hookworm. The Chenopodium oil is an example of such natural products which was obtained from Jerusalem Oak (*Chenopodium ambrosoides*). It contains about 60% of a terpene peroxide known as ascaridol [63]. It is however, reported to have many side effects and not recommended for use in children and pregnant women. A few natural products have become the major fulcrum around which alternative therapies are being developed recently. Two of such are plant extracts namely Dalea ornate and Oemalaria cerasiformis that have been proven to show potential anthelminthic activity against *Anclystoma cecylanicum* [64]. Another example of use of natural products is highlighted in Ghana where researchers from the Noguchi Memorial institute of Medical Research and the chemistry department have found residual anthelminthic activity in a group of *Dichapetalin* compounds notably Dichapetalin A [6]. This shows how natural products can be exploited as anthelminthics aside the well-known benzimidazoles.
2.4.2 Other naturally derived compounds effective against hookworm

Halogenated hydrocarbons such as carbon tetrachlorides, hexachloroethane among others have also been recognised long ago to possess varying degrees of anthelminthic activity. However, only a few of these compounds have been used due to the wide range of side effects [63]. For example, Tetrachloroethylene (Nema®, Tetracap®), which have been used since 1925 as human anti-hookworm drug has many side effects including anaemia, somnolence, dizziness and headache. The phenols and their derivatives have also been shown to have marked activity against hookworm although they are no more used in clinical practice [63]. Some of the phenolic compounds that have been used include 1-Bromo-beta-napthol (III), Hexylresorcinal (IV), 2,4,5-Trichlorophenol (V) (Ranestol®), 2,6-Diido-4-nitrophenol (VI) (Disofen) and diospyrol (X) [63].

2.5 Computer-Aided Drug Design (CADD)

The traditional approach of high-throughput screening (HTS) in the late 1990s have paved the way for modern drug discovery techniques that comprise ligand and structure based drug design in the search for lead compounds as drug candidates. Modern drug development techniques using computational techniques have more advantages due to the less considerable time and effort that need to be invested in screening and searching for promising drug candidates compared to traditional drug development. Especially important is the use of computational tools for identifying potential candidates through virtual screening to pre-filter inactive and toxic compounds before performing clinical evaluation.
2.5.1 Economic significance and time factor of CADD

The process of designing drug before meeting the strict regulatory requirements of Food and Drugs Authority (FDA) and getting approval for marketing authorization in humans requires a lot of research, money and time [1]. This normally requires screening of millions of compounds to obtain a drug candidate which is taken through many years of experimental testing and pre-clinical studies. *In silico* methods such as virtual screening of small compounds against drug targets have proven advantageous over *in vivo* or *in vitro* experimental methods in terms of cost, effort and time by significantly decreasing the number of compounds and retaining only lead hits for further HTS.

2.5.2 Steps involved of CADD

The availability of a public library of compounds, bioactivity and target databases has aided in the application of *in silico* techniques to predict potential lead compounds and the binding affinities on therapeutically interesting targets. These techniques are successful in their prediction using good performance classification models and algorithms encoded within docking programs [1]. The drug discovery pipeline generally involves target identification, target validation, virtual screening, lead identification and subsequent optimization [1, 65, 66]. Figure 2.3 shows the drug design pipeline. Target identification is the process in which drug targets are identified through literature review and searching databases that have information of experimental results [66]. In target validation, identified targets are compared to each other based on their association with each other and their associated effect on the behaviour of disease cells and interaction with metabolites in the
body. Lead identification involves the identification of compounds with plausible potential to treat diseases often referred to as lead compound often using molecular docking techniques. Molecular docking of drugs to large libraries of proteins has the capability of identifying potential targets. Hits are generally compounds that exhibit favourable activity in the screening process using virtual screening [67, 68]. Virtual screening is an effective means of molecular docking to search for potential compounds against the target protein by using computational approaches. Hits with good binding affinities as represented by low binding energies are optimized by structural modification (normally using QSAR methods) to obtain improved potency and pharmacokinetic properties and desirably a reduced toxicity [69].

The major approaches to current drug design are (i) structure based and (ii) ligand based [67, 71]. In both cases, a library of compounds is virtually screened against the target of interest. Virtual screening of potential lead generates several conformations of the complex with different inhibitors and can be very essential for providing insights into the mechanism of interaction of the lead and receptor. It can also predict the occurrence of resistance, the identification of new binding sites, potential targets and the design and optimisation of lead compounds for therapeutic agents. Some CADD also employs a technique generally referred to as drug repurposing whereby drugs known to be efficacious against one disease is tested against other diseases [71].
These in silico approaches among others are clearly advantageous in pre-filtering potentially low in vivo activity drugs not showing ideal pharmacokinetic profile and mode of actions [69]. CADD can undoubtedly be used to provide more insight into the bioactivity profile of experimentally determined drugs and be employed to potentially predict additional drug targets implicated in various disease mechanisms and biological pathways. Its effectiveness for expediting drug discovery has been recognized for decades, without exception, as in the case of exploring natural products for CADD.

2.5.3 Structure based drug design

Structure based drug design (SBDD) is employed when the structure of the target protein is available. With an exponential increase in the number of protein structures deposited in
the protein data bank (PDB, www.rcsb.org/pdb/), the volume of research using SBDD has increased significantly [72]. The process of SBDD can be summarised in the three steps with the main goal of finding potential ligand binders to targets. The first step is target selection. There are several protein databases with huge information about proteins which have been solved experimentally by either X-ray crystallography or Nuclear Magnetic Resonance (NMR). Notable among these databases are Protein Data Bank and UniProt (www.uniprot.org). If the target has not been solved experimentally, it is obtained computationally using homology or comparative modelling. Homology modelling is a technique that is used to construct a three-dimensional (3D) model of an unknown structure of a target based on the structure of a suitable homologous template. It comprises four steps: (i) search and identification of template, (ii) alignment of target to template sequence, (iii) construction of models and (iv) model quality evaluation [73].

Beyond that, there are several machine learning techniques that can be used to predict the secondary structure of the target even before constructing the 3D model. PHD program predictor [74] uses neural network by taking evolutionary information and multiple sequence alignments to predict beta strand, PSIPRED [75] involves a feed forward neural network based on PSIBLAST [76] outputs and HHpred [77] uses Hidden Markov modelling for homology detection and a host of others. Figure 2.4 shows the structure based drug design pipeline.
Using the obtained structure of target, the second step is determining possible binding sites within the receptor. There are several computational tools that predict the binding site and/or druggability regions of druggable targets including Fpocket [78], DoGSiteScorer [79], PRANK [80], Meta Pocket [81]. These programs employ machine learning algorithms to identify cavities/pockets and/or “druggable” regions. For example, SitePredict [82] uses Random Forest algorithms to predict binding sites using information about the solvent accessible surface area, pocket volume, pocket principal components and nearby residue pair count.

An important third step after determining the druggable sites, is to study the structure of the ligand-target interactions. Here, the intermolecular interactions, binding conformations, conformational changes induced by ligands are studied. The most used technique for binding conformation is molecular docking which explores receptor-ligand interactions and conformation of some residues in the binding pocket. Then, potential bioactive ligands are identified, purchased and subjected to various pre-clinical biological tests.
2.5.4 Ligand based drug design

This technique is used when the structural information about the biological target is not available and therefore, cannot be experimentally determined or homology modelled. Most ligand based drug design methods include similarity search and construction of classification models using multivariate statistical analysis [84]. Similarity searches normally employ two-dimensional (2D) and three-dimensional (3D) descriptors. 2D descriptors include molecular fingerprints, topological descriptors and molecular properties whereas 3D descriptors may be molecular shapes and MACC fingerprints [85].
The goal of similarity searches is to calculate a similarity index (Tanimoto, Dice or Tversky coefficients) or a fitting score in the case of 3D descriptors such as QSAR models based on available information of active compounds to rank unknown compounds [86]. It is generally faster with lower computational cost than structure based drug design approaches. Virtual screening using ligand based drug design approach is founded on the principle that compounds that share structural similarities have similar biological activities [87]. There are several programs that have been developed for ligand based virtual screening. LigandScout is an example of ligand-based virtual screening tool [88]. Figure 2.5 illustrates ligand based drug design.

Figure 2.5 Ligand based drug design [83]. This involves similarity search of ligands that bind to same binding site in target molecules represented using pharmacophore models. After leads have been identified, they are usually optimised by modification of their moieties.
2.5.5 Molecular docking and virtual screening

Molecular docking studies are used to determine protein-protein and/or protein-ligand interactions and evaluate their binding affinities. The two most widely used approaches in the case of SBDD are protein-ligand docking and virtual screening. Most docking programs employ conformational sampling of the protein complex with potential ligands or predicting the probability of the activity of the interacting protein receptor with several small compound ligands [14]. The protein ligand interactions are based on different approximations/objectives: force fields, search and optimization algorithms as well as score functions [14]. The degrees of freedom of both ligand and target are also considered in molecular docking simulations. In most cases, it is desirable for the receptor to be less flexible and the ligand more flexible to aid in the docking simulations and avoid producing false positive results [89]. The scoring function and optimizations/search algorithms are used to evaluate the performance of a docking simulation and they serve as the major determinants of the efficiency of the docking algorithm. An efficient docking algorithm is generally considered to be one that has a good and fast score function, and a good search or optimization function [14]. Most docking programs rely on the energy scoring function as a way of evaluating the quality of a docking simulation and they all have similar accuracies making energy scoring function an ideal choice for evaluating docking simulations [89, 90]. One of the challenges faced by most docking programs is the susceptibility of the receptor undergoing conformational changes [90]. Regardless of any energy minimisation technique that can be employed in reducing conformational changes within targets, the complete flexibility of the targets used in the docking study remains the major challenge faced by most docking programs. Some docking programs have addressed
this issue by achieving partially flexible targets [91]. The approach normally used is a combination of Monte Carlo methods with molecular dynamics, simulated cooling and others. Some docking programs and the scoring functions employed are mentioned here. Popular molecular docking programs include GSADOCK [92], Glide [93], Fred [94], AutoDock [95], AutoDock Vina [96], GOLD [97] and FlexX [98]. GOLD and AutoDock, for example, use genetic algorithms as search or optimization function. AutoDock Vina uses a hybrid scoring function that combines knowledge-based and empirical scoring functions [96].

The protein-ligand docking described in this section predicts the probability of interaction of a ligand to a target but do not provide pIC$_{50}$ or pKi that are retrieved from experimental bioassays. Besides approaches employed by various docking programs to overcome the challenge of conformational changes inherent in proteins, there is the still need to increase the performance of docking program by resorting to other approaches. This way, the results of bioactivity prediction are totally dependent on activity values in terms of pIC$_{50}$ or pKi.

2.5.6 Proteochemometric modelling (PCM)

PCM is a quantitative bioactivity prediction technique that is used to predict the bioactivity of compound-target pairs, usually reported by pIC$_{50}$ or pKi values which come directly from experimental bioassay as the true binding [99]. PCM uses compounds and related targets information in the construction of a single machine learning model [99], allowing the simultaneous prediction of compound affinities across multiple targets. In terms of cost,
PCM is not as computationally expensive as molecular docking. However, PCM is limited by its inability to account for the bioactivity of compounds against unrelated targets [21]. In PCM, the descriptions of both the ligand and protein, and an additional term called ligand protein cross term could be correlated to the binding interactions. To create a PCM model, the binding interaction of a series of ligand and targets is needed to train the model to enable to prediction and exploration of known targets. In PCM, compounds can be described by structural descriptors including molecular fingerprints, topological descriptors, geometrical descriptors and three-dimensional grid-independent descriptors (GRINDs) [99]. Description of receptors can be determined by calculating the receptors' amino acid sequence compositions. PCM can be applied whether 3D information of the target is available or not. PCM like QSAR implements a wide range of machine learning techniques (including both linear and non-linear methods) to develop models. Proteochemometric modelling, is used to alleviate the limitations associated with QSAR.

**Figure 2.6 Proteochemometric modelling [102].** The bioactivity profiles of multiple targets and ligands are used to construct a single model that can be used to predict the bioactivity between untested targets say target A and ligand, say compound 2.
2.5.6.1 Advantages of PCM over QSAR

Quantitative structure activity relationship (QSAR), normally employed in LBDD, has over the last decades been one of the mainstream computational methods in addition to molecular docking in the search of viable lead compounds. The basic assumption underpinning the success of QSAR is that compounds that share similar chemical activity should share similar targets and targets sharing similar ligands should share similar properties \[87\]. It is a method that is used to quantitatively determine the relationship between the structure and biological activity of a compound using statistical analytical methods. With QSAR, a model of the output variable can be constructed based on computed molecular descriptors using statistical method \[100\]. There are, however, some notable drawbacks and limitations with QSAR. One of the drawbacks of QSAR is that it only considers the interaction of groups of compounds with a single target and thus this requires sufficiently enough data about the target before a meaningful model can be constructed which rarely should be the case when searching for hits for previously identified targets. Conventional QSAR approaches are limited in terms of finding new ligand classes or binding interactions for a set of new compounds. This is because, in the strictest sense, multiple ligands that bind to targets are not determined only by the chemical structure but also binding interactive residues. Further pitfalls with QSAR is that it is not be able to describe all aspects of binding interactions in the case where the model was trained on descriptors of certain class of compounds. That is, it will fail to predict anything outside its applicability domain.
PCM outperforms QSAR in many ways and these findings are corroborated by many literature reports [21, 15]. One of the main advantages of PCM over QSAR is that it does not only model similar targets but also dissimilar ones allowing scientists to explore the extensive applicability domain of PCM for highly distant targets [99]. In terms of bioactivity, however, models that are built using PCM techniques are difficult to account for when the dataset covers unrelated targets. *In silico* prediction algorithms that have been used over the years include Naive Bayesian classifiers, Support Vector Machine (SVM), neural network, Random Forest (RF), and regression analysis.

2.5.6.2 Application of machine learning in PCM

Machine learning algorithm employed in PCM include SVM, Naive Bayesian classifiers, and decision tree algorithms. Prior to constructing PCM models, the data should be pre-processed based on a description given by Andersson *et al* [101] and van Westen *et al* [100]. Following that, chemical and protein descriptors are calculated based on which feature selection is done and the subsequent construction of models. Three popular machine algorithms are discussed here, namely SVM, RF and Gaussian Processes.

2.5.6.2.1 Support Vector Machine (SVM)

Support vector machines are a group of non-linear machine learning techniques that have gained a lot of popularity in PCM [102]. It is a type of machine learning technique for classification and/or regression that uses linear or non-linear kernel-functions to project data into a high-dimensional feature space [103]. SVMs are able to produce high
performance models and efficiently able to deal with large dataset with high dimensional space \([103, 104]\). Interpretability is normally the major challenge faced by SVM but accuracies of models are improved by fine-tuning using the so-called hyper parameters, the most important being the kernel function parameter, \(\gamma\) and the error penalty parameter, \(C\). SVMs generally use internal kernel methods, the Radial Basis Function (RBF) Kernels being the most dominant \([99]\). RBF have been shown to produce some reliable results on the performance of PCM. Wu et al \([106]\) improved the mapping power of their PCM models for a set of histone deacetylases (HDAC's) by using a (Pearson function-based Universal Kernel) PUK kernel. Various authors have applied different types of the classical SVM including the Dual Component SVMs (DC-SVM), Transductive SVMs and Relevance Vector Machines (RVMs). DC-SVM based PCM were shown to outperform classical SVM based QSAR \([102]\). Notable is the RVM where the authors demonstrated how well it performed by employing binary classifiers trained on some dataset from the MDL Drug Data Report (MDDR) database and concluded that it must be applied in future PCM studies \([102]\). SVMs have contributed enormously as a useful algorithm in several PCM studies.

2.5.6.2.2 Random Forest (RF)

Random Forests form a unique group of nonlinear machine learning techniques which have a comparable performance to SVM \([107]\). RF generally constitute a decision tree comprising of nodes and branches. Each node represents a point where dataset is divided based on a selected attribute value so that instances of different classes are moved to different branches. RF classification is performed starting at the root node along the tree to
the leaf nodes. The collective result of all trees is used as an estimate of the performance of the classification. Unlike SVMs, they involve relatively short training times with less hyper parameter tuning. Although highly interpretable, it suffers from its inability to output error estimates which are tremendously important due to the level of error and noise annotations associated with public bioactivity databases [108]. This is normally fixed by applying Quantile Regression Forests (QRF) based on quantile inferences from the conditional distribution of the class variable [108].

### 2.5.6.2.3 Gaussian Process

Gaussian Processes (GP) are a group of kernel-based non-parametric machine learning method based upon Bayesian framework. As there are huge concerns with errors or the so-called “noise” in bioactivity databases arising from data curation and experimental inaccuracies, GP aims to address these concerns by constructing probabilistic models using the uncertainties contained in the data as input [89]. For a given compound-target combination, the GP predicts using a Gaussian distribution whose variance defines confidence interval as a measure of the distance of the compound-target pair to the training set. GP models can be generally validated by the conventional statistical metrics, square of the correlation coefficient ($R^2$ or $Q^2$) [108, 109] but has also internal validations and assessments. GP has seen many applications in the chemogenomic space [110 – 112]. The downside with GPs however is the longer training time due to the algorithm of $O(N^3)$ time complexity [114].
2.6 Recent Efforts in Hookworm Drug Discovery

Herein, previous efforts geared towards hookworm drug discovery are enumerated. A new cysteine protease inhibitor, oral single-dose anthelmintic that is active in an animal model of hookworm infection and demonstrated a distinct mechanism of action from current anthelmintic was discovered [115]. Drug repositioning and pharmacophore identification was utilised in the discovery of hookworm MIF Inhibitors by targeting AceMIF [116]. About 1600 FDA approved library of compounds were screened against laboratory models of human intestinal nematode infections. Hits that were identified were suggested to serve as a starting point for drug discovery for soil transmitted helminths [117]. Also, lead chemotherapeutic agents from medicinal plants were identified against blood flukes and whipworms [118]. As reported elsewhere [119], a set of compounds that were known to show activity against parasitic nematodes were collated from various literature sources including PubChem while the inactive dataset was retrieved from DrugBank database based on Tanimoto cutoff range of 0.25 to 0.75. An SVM algorithm was used to construct a model and stratified 10-fold cross validation was used to evaluate the performance of each classifier using the radial basis function kernel. An accuracy of 81.79% was achieved for the model when an external independent test set was applied. The results reported were remarkable. The model was then used to identify novel compounds with potential anthelmintic activity. In another work, Ponce-Marrero et al. [120] used a linear discriminant analysis to obtain a quantitative model for classification of anthelminthics and non anthelminthics. This novel approach resulted in a model that correctly classified 88.18% of the compounds in external test set. Virtual screening was used to validate the performance of the model where it identified several compounds annotated as
anthelminthic in the Merck Index and Negwer’s handbook. Train-Match-Fit-Streamline (TMFS), novel rapid computational proteo-chemometric method were used to map new interaction space and map new drug targets. The method combined shape, topology and chemical signatures, including docking score and functional contact points of the ligand, to predict potential drug-target interactions. Extensive molecular fit computations were performed on 3,671 FDA approved drugs across 2,335 human protein crystal structures. The algorithm predicted drug-target associations with 91% accuracy for most drugs. Over 58% of the known best ligands for each target were correctly predicted as top ranked [121]. Furthermore, TMFS method was used to discover that mebendazole had the structural potential to inhibit EGFR2. In another work [120], support vector machine approach was employed to predict compounds active against parasitic nematodes, suggesting the importance of employing computational approaches for anti-parasitic drug discovery. The method presented an alternative approach to the existing traditional methods and may be useful for predicting hitherto novel anthelmintic compounds.
CHAPTER 3

METHODS

The methods used which were homology modelling of the protein of interest, molecular dynamics simulations, virtual screening, Absorption, Distribution, Metabolism, Excretion (ADME) and toxicity predictions, scaffold analysis of the most favourable docked compounds are presented in this chapter. The proteochemometric modelling techniques for anti-tubulin bioactivity are also presented in this chapter. The entire workflow of homology modelling to scaffold analysis is shown in Figure 3.1 and details explained subsequently.

3.1 Template Identification and Homology Modelling of Proteins

A search in PDB (http://www.rcsb.org/) revealed that the tertiary structure of none of the beta tubulins of *N. americanus* was publicly available. The primary sequence of the beta tubulin protein with Gene ID: NECAME_01536 was retrieved from UniProt [122] (Accession number, W2T758, length: 449 amino acid). The sequence was submitted for template and binding site identification via the Iterative Threading ASSEembly Refinement (I-TASSER) server [123]. The I-TASSER server is an online server for automated protein structure prediction and structure-based function annotation. I-TASSER predicted a number of plausible templates. The D chain, which is present in the subunit of the multi-meric structure of tubulin tyrosine ligase (T2R-TTL) (PDB ID: 5c8y), was selected as the most plausible template based on the presence of amino acid residues associated with nematode resistance in the binding site as found by I-TASSER as well as a high sequence
Figure 3.1. **Workflow of protein modelling to scaffold analysis.** The target of interest is modelled using homology modelling. The model is subjected to molecular dynamics simulation and binding site identification. Virtual screening of small compounds from AfroDB and the North African Natural Product Database against the target are used for the identification of potential lead compounds. Binding affinity scores are used for ranking the docked compounds. The pharmacokinetic and pharmacological properties of the top-ranking compounds are identified by predicting their ADMET properties. Scaffold analysis is used to compare the scaffold diversity and/or similarity between the top ranking natural products and anthelmintics.
identity to the template. The crystallographic structure of tubulin tyrosine ligase was downloaded from PDB (PDB ID: 5c8y, resolution: 2.59 Å) and used as a template in modelling. MODELLER [124] is a software that is used to generate three dimensional models of proteins which are known as homology models. MODELLER aligns the target sequence with the template structure and builds 3D models based on a target/template alignment. The major characteristic of MODELLER is the extraction of spatial constraints such as template Cα-Cα distances, backbone dihedrals (φ/ψ), sidechain dihedrals and van der Waals contacts from the template which are applied to target sequence to generate the modelled target protein [124]. The align2d function in MODELLER v.9.16 was used to align the sequence of target with the template (files in Appendix A). Once a target-template alignment was constructed, MODELLER 9.17 was used to compute a five candidate 3-D models of the target using the whole sequence of the target protein. The best model was selected based on the lowest value of the MODELLER 9.17 objective function or the Discrete Optimized Potential Energy (DOPE) and high GA341 score [124]. DOPE and GA341 are in-built assessment scores used to assess the quality of the protein model generated by MODELLER. Protein models constructed using homology modelling normally produce unfavourable bond lengths, bond angles, torsion angles and contacts. The model was therefore refined to fix steric clashes and bumps by submitting the model in protein data bank (pdb) format to WHAT IF server [125] and energy minimized to correct local bond and angle geometry, and to relax close contacts in the geometric chain using Swiss-PdbViewer 4.10 [126]. The WHAT IF server implements WHAT_CHECK program on its server to check and fix steric clashes based on the overlap of two non-bonding atoms of distance cutoff set at 0.4 Å [125]. Swiss-Pdb Viewer is an application
that can be used for visualization, homology modelling and 3D structural analysis of proteins and it includes GROMOS43B1 force for minimization of the protein structures. The refined model was visualized using educational version of PYMOL 1.74 [127] software and further subjected to molecular dynamics simulation described in subsequent steps. PYMOL is an open-source software for interactive visualization and analysis of the molecular structures.

3.1.1 Model assessment and refinement
Further assessments of the selected best model was done by generating a Ramachandran plot [128] using the PROCHECK 3.5.4 software. Other programs such as ERRAT, VERIFY3D and Qmean [128, 129] were used to corroborate the PROCHECK results. Homology models of proteins are usually subject to prediction errors. Therefore, PROCHECK 3.5.4 was used to assess the stereochemical qualities of the three-dimensional homology model. PROCHECK [131] a suite of C and Fortran programs, provides a way to check the stereochemistry of a protein by a detailed residue-by-residue listing with an assessment of the overall quality of the structure compared to refined structures produced at the same resolution.

3.2 Molecular Dynamic Simulations of Modelled Protein
The modelled structure of the tubulin may show good accuracy but to use it for virtual screening, it is required to show good molecular dynamics behaviour as well. To evaluate the stability and folding, and obtain insights into the conformational changes as well as the dynamics of the modelled protein in solution, a 1 nanosecond (ns) molecular dynamics
simulation was performed. The molecular dynamics (MD) simulations of modelled tubulin receptor was carried out with the Linux version of GROMACS 5.1.4 [132] software package by employing GROMOS 96_43a1 force field and the flexible Simple Point Charge (SPC) water model by passing “-water spce” command. The modelled structure was first immersed in a periodic water box of cubic shape (1 nm thick). After solvating the receptor, the net charge on the protein was +8e. *Genion* command in GROMACS was used to add 8Cl\textsuperscript{-} ions to neutralise the net charge on the protein. Electrostatic energy was calculated using the particle mesh Ewald method with a computational load of 0.19. Cutoff distance for the calculation of the coulomb and van der Waals interaction was 1.0 Å. The Cutoff scheme used was Verlet.

After energy minimization using a steepest descent for 50000 steps, the system was subjected to equilibration at 300k and normal pressure for two pico-seconds (ps) under the conditions of position restraints for backbone atoms. Linear Constraint Solver (LINCS) constraints were performed for all bonds, keeping the whole protein molecule fixed and allowing only the water molecule to move to equilibrate with respect to the protein structure. The final molecular dynamic calculations were performed for 1 ns under the same conditions. The results were analysed using GROMACS 5.1.4 [132] and GRACE 5.1.4 [190] plot software using the command *xmgrace* in a Linux terminal. The stabilised receptor file in gro format was uploaded as frames and saved in pdb format using Visual Molecular Dynamics (VMD) software 1.9.3 version [133]. VMD is a molecular graphics and visualisation program of molecular structures.
3.3 Prediction and Analysis of Binding Site

The potential colchicine binding site of the receptor or protein containing the amino acid residues of interest was predicted with MetaPocket 2.0 server [134], complemented with Computed Atlas of Surface Topography of proteins (CASTp) server [135] and analysed with AutoDock/Vina v2.2.0 plugin [136] in educational version of PYMOL 1.74 [127] before undertaking molecular docking. MetaPocket and CASTp are online servers for the prediction of ligand-binding sites.

3.4 Protein Preparation

AutoDockTools 4.2.6 version [137] was used to prepare both receptors and ligands. Gasteiger charges were calculated and polar hydrogens added with non-polar hydrogens merged using AutoDockTools. All water or solvent molecules were removed to eliminate the influence of solvent interactions in the protein-ligand docking. The receptor file was converted to protein file in pdbqt format which was used as input receptor file for AutoDock Vina (Vina). Receptor energy grid and parameters were generated using AutoDockTools. The grid box was set to dimensions; 22.5 Å x 22.5 Å x 22.5 Å for the receptor with coordinates of -18.35, -8.23, -22.48, and centered around amino acids Glu198, Phe167 and Phe200.

3.5 Ligands Preparation

AfroDb [138] subset of natural compounds from ZINC [139] database was downloaded as single batch file in Structure Data File (SDF) format on 17th November, 2016. ZINC
contains millions of free collections of small molecules that can be used for virtual screening. AfroDB is a collection of highly potent natural products isolated from African medicinal plants and a subset of the ZINC database [138]. The file retrieved from the AfroDb subset of ZINC contained a total of 885 molecules. Dichapetalin A and albendazole were added to make it up to 887. A different set of virtual compound library was retrieved from the Northern African Natural Product Database (NANPDB) [140] on 12th June, 2017. The NANPDB contains a large collection of over 4,500 annotated natural products originating from North Africa [141]. The retrieved file was a file containing a single 3D structure of all the compounds in an SDF format. The file into 2297 molecules with custom bash script using Open Babel 2.3.1 [142] (Appendix A). All ligand files were first optimized and energy minimised using PRODRUG server [143] and Open Babel 2.3.1 within the Pyrx 0.8 interface [142]. They were then converted to pdbqt files using Pyrx 0.8 [144]. Pyrx is a computer software that can be used for small molecule virtual screening.

3.6 Virtual Screening Analysis

To find out the preferred binding modes of the ligands in the active site of the receptor, molecular docking analysis was performed using AutoDock Vina 1.1.1 [145] via a four core Intelcore-I7 processor Linux operating system machine. Docking involves 3 main steps, (i) protein preparation and grid box specification, (ii) ligand preparation and (iii) docking of ligand against protein. Protein and ligand preparation had been previously performed which meant that the next step involved virtual screening. The docking simulation of each compound for the first set of AfroDb virtual library compounds was conducted and the different binding conformations of the docking ligands were generated.
and scored. Lastly, the top-ranking results were selected based on their binding energies in the final output log files. Virtual screening analysis was conducted separately for the NANPDB virtual library compounds and the different binding conformations results were obtained as well. Conformational analysis of the ligands was employed to fit ligand molecules into the receptor using AutoDock Vina 1.1.1 version with details of the docking generated as docking log files (Appendix A). The log files were analysed and tabulated. AutoDock Vina is a molecular docking software that uses an empirical scoring function to calculate the binding affinities of protein-ligand complex by summing up contributions of the energies of the protein-ligand binding (measured as the sum of the distance-dependent atom pair interactions) [112]. The lowest binding affinity score is normally considered as the compound that exhibited the strongest binding. Docking of the ligands to receptor for the AfroDb compounds in addition to Dichapetalin A and albendazole completed in 1 day, 15 hours, 8 minutes, and 45.327 seconds and that of NANPDB due to its large quantity completed in 3 days, 2 hours and 45 minutes. All protein ligand binding affinities were expressed in Kcal/mol.

3.7 Interaction Profiling using LIGPLOT

The protein-ligand complex interactions were computed using LIGPLOT 1.4.5 [146]. LIGPLOT provides a schematic 2-D representation of the hydrogen and hydrophobic interactions between ligand and active site residues of the protein-ligand complex. Hydrogen bond interactions were represented by dashed green lines, while hydrophobic interactions were represented by arc with spokes radiating towards the ligands and the number of hydrogen bonds with the active site residues.
3.8 Absorption, Distribution, Metabolism and Excretion (ADME) Prediction

ADME profiling was carried out on SwissADME server [147] which predicted the relevant ADME properties. The latter constitute the pharmacokinetic profile of drugs which has a direct effect on the pharmacodynamics of the drug molecule immediately after the drug is orally administered. SwissADME [147] is an online webserver that allows the calculation of several physicochemical descriptors and the prediction of ADME parameters, pharmacokinetic properties and drug-likeness of small molecules. It requires the user to upload the SMILES format of the query molecule. SwissADME was used for the calculation of pharmacokinetic properties such as ESOL logS, molecular weight, lipinski rule (drug-likeness), Gastrointestinal (GI) absorption, Blood Brain Barrier (BBB) permeant and bioavailability score.

3.9 Toxicity Prediction using OSIRIS Property Explorer in DataWarrior

The toxicity profile of the top ten virtually screened compounds (in addition to Dichapetalin A and albendazole) and that of the NANPDB whose interaction profile have been previously investigated were further analysed using the OSIRIS property explorer embedded in DataWarrior 4.5.2 [148] in order to assess the toxicity of the drug candidates. This explorer gives drug relevant properties such as mutagenicity, irritancy and reproductive effect. The top ten compounds from the first set of screened AfroDb compounds in addition to Dichapetalin A and albendazole and the second set of screened NANPDB compounds were subjected to toxicity prediction using DataWarrior 4.5.2 by submitting the compounds in SMILES format and the results were investigated.
3.10 Scaffold Analysis

A scaffold analysis was conducted to compare the scaffold diversity and/or similarity between the docked natural products from AfroDb and NANPDB with Dichapetalin A included (hereafter referred to as “list A”) and 16 anthelmintics including albendazole, mebendazole, febendazole, levamisole and piperazine (hereafter referred to as “list B”) (files in Appendix A). List A included Dichapetalin A, and top 100 compounds each from AfroDB and NANPDB. The scaffolds of list A were compared with that of list B. Another analysis was conducted between list A and a dataset of only albendazole with mebendazole (hereafter referred to as “list C”).

There are several ways to represent scaffolds [147, 148]. One of such representations is the Murcko framework/scaffold as proposed by Bemis and Murcko [149]. This framework has been used to analyse the structures of known drugs and to identify the similarities in screening compound library [11, 153]. The Murcko framework preserves only the molecular topology of the ring systems and removes any substituents that do not contain ring systems linked to the ring or ring side chains. It contains no three-dimensional structure or any stereochemistry [152]. Murcko scaffold analysis was used for exploration of similarity and/or diversity amongst the scaffolds in the datasets of list A and list B and another analysis between list A and list C. This was conducted using DataWarrior v.4.5.2 by using the SMILES notation of the different datasets and then analysed the scaffold architecture using the Murcko framework. Scaffold frequency and counts were used to measure the distribution of compounds over unique scaffold present in the data subsets. Comparative studies were performed between the subsets of natural products and anthelmintics. Bar charts were used to characterise the
distribution, diversity and/or similarity of the scaffolds within the datasets using DataWarrior.

3.11 Proteo-Chemometric Predictive Model of Anti-Tubulin Activity

3.11.1 Data collection

Since beta tubulins are the primary targets of benzimidazoles, the keyword “beta tubulin” was used to query a chemogenomic database, BindingDB [153]. The query produced a dataset comprising a bioassay of tubulins that was subsequently retrieved from BindingDB on 20th of December 2016. The dataset comprised active and inactive ZINC compounds against mostly beta tubulins and a few other tubulins with Uniprot IDs: Q25270, P02554 and Q6B856. BindingDB is a publicly accessible database currently containing over 20,000 experimentally determined binding affinities of protein–ligand complexes [153]. A dataset of four hundred and thirty-seven (437) bioactivity data on tubulins was obtained from BindingDB. The retrieved bioassay data reported 129 compounds with potency against beta tubulins and other tubulins ranging from 100 nM and 810,000 nM for inhibition constants, Kd (nM), IC\textsubscript{50} (nM), Ki (nM) and EC\textsubscript{50} (nM), covering beta tubulins from \textit{Leishmania donovani}, \textit{Sus scrofa} and \textit{Bos taurus}. Due to the different assay conditions of the experimental dataset, the bioactivity values provided in the dataset were not used in developing the PCM-SVM model. The dataset was instead labelled as actives (Ki, Kd, IC\textsubscript{50} or EC\textsubscript{50} bioactivity values equal to or lower than 1\textmu{}M or 1000nM) and inactives (all remaining observations/protein–ligand combinations) [154] as explained in subsequent section of the chapter (section 3.11.2). The schematic workflow for the PCM modelling is summarised in Figure 3.2.
Figure 3.2. Workflow for PCM modelling of beta tubulin bioactivity profiling. A bioassay dataset is first retrieved from a chemogenomic database, BindingDB. The dataset comprises a bioassay of small compounds tested against beta tubulin variants. Ligand and compound descriptors are computed which are used in the construction of a single SVM based PCM. The PCM could be used for predicting the bioactivity between untested or new compounds and beta tubulins.
3.11.2 Pre-processing of dataset

Ligand structures based on SMILES notations were processed with the R package `camb` version 2.0 [155] using the function, `StandardiseMolecules`, which enables the depiction of molecular structures in the same (standardised) form [155]. The function also allows the removal of inorganic molecules. All the data was then further pre-processed by removing salts within the software, PaDEL 6.0 version [156]. After calculation of descriptors (described in sections 3.11.3 and 3.11.4), constant and near constant predictors (called zero and near-zero variance predictors respectively [157]) which can sometimes be found in the dataset and do not add any significance to the data were checked for removal with the `nearZeroVar` function of the R package `caret` [158]. The function removes predictors that have one unique value across samples (zero variance predictors), but also removes predictors that are few unique values relative to the number of samples. In general, a predictor is classified as near-zero variance if the percentage of unique values in the samples is less than 10% and when the frequency ratio mentioned above is greater than 19 (95/5) [159]. The `cor` function from the R package `stats` 3.50 version [160] was then used to compute the pairwise correlation between descriptors. Pearson’s correlation coefficient greater than a threshold of 0.7 was performed to check and filter out similar descriptors using the `findCorrelation` function from the R package `caret` 6.0-76 version, set at a cutoff of 0.7. The rationale behind normalising the dataset by the removal of highly correlated features and near zero variance features was to avoid bias in the final model that would be built. The feature elimination pipeline, however, retained the features/descriptors that were computed (described in sections 3.11.3 and 3.11.4).
The dataset was subdivided into two groups based on their experimental anti-tubulin bioactivities using the rule: “active” (IC\text{50}: < 1\mu M) and “inactive” (IC\text{50}: \geq 1\mu M). This is because, approximately 1000 nM activity values are normally considered as a right threshold for IC\text{50} to differentiate a bioactivity or not [153, 160]. A drug candidate is generally considered to have low nanomolar on target with a concentration, IC\text{50}, required to reduce the target activity or biological process by half [154].

SVM training and testing sets require normalised class data input as binary values; hence the dataset was subsequently labelled by normalizing it as active dataset with a “1” and the non-active dataset with a “0”. The dataset was split into training and testing set with 67% for training and validation and the remaining 33% for held out or independent test set (not included in the training set). Thus, the training data contained 292 data samples, while the test set comprised 145 as listed in Table 3.1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Active</th>
<th>Non-actives</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>32</td>
<td>260</td>
<td>292</td>
</tr>
<tr>
<td>Test</td>
<td>13</td>
<td>132</td>
<td>145</td>
</tr>
</tbody>
</table>

**Table 3.1. Complete dataset used for PCM.** Training dataset comprise 292 samples

3.11.3 Ligand descriptions (Compound descriptors)

In this work, circular molecular fingerprints and physicochemical descriptors were used to represent the ligands. Compounds were described with unhashed Morgan fingerprints
using the *MorganFPs* function of the package *camb* version 2.0 in R 3.4.0 version. R is a statistical programming language with a set of inbuilt functions and object-oriented features for software development [162]. Morgan fingerprints encode chemical structures by considering atom neighbourhoods [119]. Maximal user-defined bond diameter is normally assigned to each substructure. In this study, the maximum diameter of the substructures considered was set to 4, whereas the length of the fingerprints was set to 128. Physicochemical descriptors were also computed using the *GeneratePadelDescriptors* function from the R package *camb*, version 2.0 [119] which invokes the software PaDEL. The unhashed circular fingerprints were used because they are interpretable, very important for the inhibition against the tubulins with good performance and widely adopted [84, 101, 162].

3.11.4 Target descriptions (Protein descriptors)

To describe the target space in PCM models, whole protein sequence descriptors were calculated using composition/transition/distribution (CTD) descriptors. CTD descriptors such as relative hydrophobicity, predicted secondary structure, and predicted solvent exposure were employed because they are widely used and shown to be interpretable as well [164]. CTD amino acid descriptors were calculated with the function *SeqDesc* from the R package *camb* version 2.0 [155].
3.11.5 Exploratory principal component analysis (PCA) of compounds and target datasets

Principal Component Analysis (PCA) was performed as a way of reducing the high dimensional dataset to a low-dimensional set of variables using R. Principal component analysis (PCA) is a mathematical method for dimensionality reduction that allows for multidimensional datasets to be visualized using two or three-dimensional plots with minimal loss of information or variance [164 – 166]. It is a method for finding the linear combination of a set of observations with the most possible variance, and can reveal important characteristics of the data structures including the similarities and differences within the dataset based on calculated descriptors, which are otherwise difficult to distinguish. PCA identifies the features or descriptors that show as much variation across the data. The `prcomp` and `autoplot` functions of the R package FactoMineR version 1.2.4 [168] were used to perform PCA and the result was plotted by clustering the dataset with ellipses. PCA was performed on the descriptors of the compounds and proteins separately.

3.11.6 Model development

Among the multitude of available machine learning binary classification algorithms, Support Vector Machine learning algorithm was employed because it is highly effective, robust and has been extensively successful in the field of drug discovery [169–171].

Support Vector Machine (SVM) developed by Vapnik [172] is a statistical learning method known to be popular owing to its effectiveness and robust performance. The downside of
using SVM or any other machine learning algorithm to the user is the backend statistical and computer algorithms that are used which the user generally does not have control over or not preview to. The way SVM learns is by finding the maximal hyperplane to differentiate data points in a vector feature space. Optimisation of the models is done by fine tuning optimal hyperparameters (i.e., the gamma ($\gamma$) and C parameters). The dataset was first mean centered and scaled to unit variance using a preprocessing module (StandardScaler) from a python library, Scikit-learn 0.17. The dataset was then split into 67% and 33% for internal training/validation and hold-out (test) set respectively. The module, SVM estimator, of the python library Scikit-learn 0.17 [173] was used to train a SVM model using radial basis kernel function (rbf) with 10-fold Stratified Cross validation due to the paucity of the dataset. The performance of the model was optimised by fine tuning the hyperparameters of C and gamma ($\gamma$) using the values: {
"C": [0.1, 1, 10, 100, 1000],
"gamma": [0.1, 0.01, 0.001, 0.0001, 0.00001]} (See Appendix A for scripts).

The basic mathematics underpinning SVM for classification carried out is as follows [174]: Suppose we are given a set of samples, that is, a series of input vectors $x_i \in \mathbb{R}_d$ ($i=1, ..., N$) with corresponding targets $(x_1, y_1), ..., (x_m, y_m), ..., y \in \{-1, +1\}$; where -1 and +1 are used to represent the two classes respectively; in our case “non-active” and “inactive”. The goal in SVM is to construct a binary classifier or derive a decision function from the available data sample, which has a low probability of misclassification given a new data sample. It accomplishes that using an optimized linear separator, that is, construct a hyperplane, $w^T x + b = 0$ that separates the 2 classes (can be extended to multi-class problem) [174]. Different
mappings construct different SVM. The mapping \( x_i \in \mathbb{R}^d \) \((i = 1, ..., N)\) is performed by a kernel function:

\[
K(x_i, x_j) = \Phi(x_i)\Phi(x_j) \quad [174]
\]

which is an inner dot product in the feature space, \( H \) mapped by \( \Phi \). In the case of radial basis kernel function (RBF) where \( \gamma \) is the width parameter, the equation is given as:

\[
K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \quad [174]
\]

3.11.7 Validation of model performance

In this work, the robustness of the model was assessed by a 10-fold cross validation and the prediction accuracy or predictability of the models evaluated by internal validation set. The internal validation set (i.e., the 67% data subset) was subjected to training and 10-fold cross-validation (10-fold CV). In a 10-fold CV scheme, one-fold (10%) of the data was left out as the test set, while the remaining 90% were used as the training set for constructing the predictive model. This technique was repeated iteratively until all folds were left out once. SVM model was internally validated by 10-fold cross-validation in the scikit-learn package. In addition, the model was evaluated using the counts of True Positive (TP) and False Positives (FP) or over-predictions, True Negatives (TN) and False Negatives (FN) or missed predictions. Specificity and Sensitivity were calculated based on the latter values. The binary classification model was also evaluated by the receiver operating characteristic (ROC).
Accuracy was calculated as the percentage of correctly classified instances and computed as

\[
\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{3}
\]

Due to unequal size of the classes used and certain errors that may potentially be considered in the model as more serious than others (e.g., false negatives compared to false positives), accuracy did not serve as an optimal measure of model performance. Instead, the Area Under the Receiver Operating Characteristic (ROC) curve (AUC) was used and the ROC curve is basically as a measure of the discriminatory power that is insensitive to changes in class distribution usually in the test dataset. It was obtained by calculating sensitivity and specificity at various discrimination threshold levels (chance level). Sensitivity is the fraction of true positives (the true positive rate (TPR)) and Specificity is the fraction of the true negative rate (TNR).

\[
\text{Sensitivity} = \frac{TP}{TP+TN} \tag{4}
\]

\[
\text{Specificity} = \frac{TN}{TN+FP} \tag{5}
\]

\[
\text{MCC} = \frac{TN \times TN - FN \times FP}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}} \tag{6}
\]

In addition, a balanced measure like MCC, was also computed. An AUC usually greater than 50% indicates a good prediction, and an MCC equal to 1 indicates a perfect prediction while MCC equal to 0 indicates a random prediction.
CHAPTER 4

RESULTS AND DISCUSSION

The results of the homology modelling, molecular dynamics simulation, molecular docking, active site interaction profiling, ADMET prediction and PCM undertaken in this study are presented and discussed in this chapter.

4.1 Template Identification, Homology Modelling of Proteins and Validation

The D chain of the crystal structure of tubulin tyrosine ligase (T2R-TTL) with PDB ID 5c8y was chosen as the best template for homology modelling based on the presence of amino acid residues comprising Phe167, Glu198 and Phe200, which are associated with nematode resistance within the binding site. A sequence alignment with the template also demonstrated that the beta tubulin is highly homologous to the subunit of the T2R-TTL with 96% sequence identity (Figure 4.1). In addition, the results of the sequence alignment indicated that beta tubulin contained conserved residues Phe167, Phe200 and Glu198 within the active site of interest, the colchicine binding site (Figure 4.2).

The best modelled protein of *N. americanus* (UniProt ID W2T75) produced using MODELLER was selected based on low DOPE score and high GA341 score (Figure 4.3). The modelled protein is a monomer, folded into a β domain consisting of 11-stranded β-sheets and 11 α-helices (Figure 4.3).
Figure 4.1. A pairwise sequence alignment between the beta tubulin sequence of *N. americanus* and D chain of the crystal structure with PDB ID, 5c8y. The initials represent the amino acid residues. The PDB ID of the homologous template and the accession number of the beta tubulin of *N. americanus* are provided on the left. The highlighted residues supported with an asterisk (*) show the conserved residues between the sequences of the homologous template and *N. americanus* beta tubulin sequence.
Figure 4.2. Predicted binding site from I-TASSER [88] and rendered in PYMOL. The D chain of the crystal structure of the template, PDB ID, 5c8y is represented in gray. The binding pocket is represented as a green surface.

Figure 4.3. 3D model of the beta tubulin of *N. americanus* (Uniprot ID, W2T75, where helices are shown in red, beta sheets in yellow and loops in green. Monomer subunit of tetramer template with PDB ID, 5c8y.
The quality of the generated 3D model was evaluated via Ramachandran plot (Figure 4.4) using PROCHECK software. Ramachandran plots highlight the most favoured, allowed, generously allowed and disallowed regions of the modelled protein structure. Ideally, a model of reasonably high quality should have at least 90% residues in the core regions [128]. The Ramachandran plot for the predicted model showed that 92.3% of residues were within the most favourable region whilst 4.9% were in the allowed region, suggestive that the predicted model was of reasonably high quality. In addition, the overall quality factor predicted by the ERRAT server for the model was 89.327 (Figure 4.5), which corroborates the quality of the model. ERRAT [175] provides the overall quality factor for non-bonded atomic interactions and the generally accepted range is greater than 50 for a high-quality model [131]. When the model was further validated using VERIFY 3D server [176], 88.77% of the residues were predicted as having an average 3D-1D score greater than 0.2.
Figure 4.4. Ramachandran plot of beta tubulin model from *N. americanus* obtained by PROCHECK: 92.3% residues in favourable regions (A, B, L); 7.7% residues in additional allowed region (a, b, l, p); 0.0% residues in generously allowed regions (-a, -b, -p, -l); 0% residues in disallowed regions.
The results of the MD simulation of the receptor obtained using GROMACS indicated that the Root Mean Square deviation (RMSD) increased from the beginning but after a period of 0.5 ns, it remained almost constant for the rest of the duration of the simulation (Figure 4.6). This suggests that the model has very low RMSD for the backbone with less Root Mean Square (RMS) fluctuations and flexibility, indicating that the model had a stable structure during the MD simulations.

**Figure 4.5. Errat plot.** Black bars identify the misfolded region located distantly from the active site, gray bars demonstrate the error region between 95% and 99%, and white bars indicate the region with a lower error rate for protein folding.

4.2 Molecular Dynamics Simulation

The results of the MD simulation of the receptor obtained using GROMACS indicated that the Root Mean Square deviation (RMSD) increased from the beginning but after a period of 0.5 ns, it remained almost constant for the rest of the duration of the simulation (Figure 4.6). This suggests that the model has very low RMSD for the backbone with less Root Mean Square (RMS) fluctuations and flexibility, indicating that the model had a stable structure during the MD simulations.
The predicted binding pocket was found to contain all the residues whose mutation is associated with anthelminthic resistance (Figure 4.7A). The binding site had 41 residues which form the putative binding pocket. These 41 residues include Phe167, Glu198 and Phe200 (Figure 4.7B).

**Figure 4.6. RMSD plot of the molecular dynamics simulation using GROMACS.** A plot of RMSD in nanometres (nm) against time in nanoseconds (ns). The RMSD increased from 0ns to 0.5ns and levels off with slight fluctuations to the end of 1ns.

### 4.3 Prediction and Analysis of Binding Site

The predicted binding pocket was found to contain all the residues whose mutation is associated with anthelminthic resistance (Figure 4.7A). The binding site had 41 residues which form the putative binding pocket. These 41 residues include Phe167, Glu198 and Phe200 (Figure 4.7B).
Figure 4.7 Predicted colchicine binding site of beta tubulin from *N. americanus*. **A.** The binding pocket indicated in blue mesh surface and enclosed in a box is depicted in PYMOL. **B.** The amino acid residues of the colchicine binding site in the modelled beta tubulin receptor of hookworm are shown in green.

### 4.4 Virtual Screening Analysis results

The top 20 ligands from AfroDb in addition to Dichapetalin A and albendazole (Table 4.1) and that of NANPDB (Table 4.2) after virtual screening were ranked according to the decreasing order of negative binding affinity scores. The complexes of the top ranked
ligands show how firmly fitted the ligands are within the binding pockets of the protein (Figures 4.8, 4.9, 4.10 and Appendix B).

Table 4.1 Results of the molecular docking scores of the top 20 ligands from AfroDB library plus Dichapetalin A and albendazole. Two variants of RMSD metrics are also provided, rmsd/lb (RMSD lower bound) and rmsd/ub (RMSD upper bound).

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Binding Affinity (Kcal/mol)</th>
<th>rmsd/ub</th>
<th>rmsd/lb</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZINC14760755</td>
<td>-8.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95485927</td>
<td>-8.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95486082</td>
<td>-8.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95486263</td>
<td>-8.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC14780716</td>
<td>-8.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95485922</td>
<td>-8.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95486052</td>
<td>-8.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95485928</td>
<td>-8.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC13480348</td>
<td>-8.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC28462577</td>
<td>-8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95486072</td>
<td>-8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95486073</td>
<td>-8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95486081</td>
<td>-7.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC33833639</td>
<td>-7.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95485992</td>
<td>-7.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95486074</td>
<td>-7.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC95486075</td>
<td>-7.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC13365959</td>
<td>-7.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC13485435</td>
<td>-7.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ZINC15120680</td>
<td>-7.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dichapetalin A</td>
<td>-5.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Albendazole</td>
<td>-5.6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

ZINC14760755 had the strongest binding with a more negative binding affinity score of -8.6 Kcal/mol (Table 4.1) from the first library screening. 504 ligands had more negative binding affinity than albendazole (Appendix A). Also, Dichapetalin A had a more negative binding affinity score than albendazole. Higher negative binding affinity is an indication of stronger binding to the receptors and perhaps the ligands could be potential
anthelminthic leads as well as inhibitors of beta tubulin receptor of *N. americanus*. Notably, Dichapetalin A exhibited higher negative binding affinity than albendazole with scores of -5.8Kcal/mol and -5.6Kcal/mol, respectively.

### Table 4.2: Results of the molecular docking scores of the top 20 ligands from the Northern African Natural Product Database

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Binding Affinity (Kcal/mol)</th>
<th>rmsd/ub</th>
<th>rmsd/lb</th>
</tr>
</thead>
<tbody>
<tr>
<td>S,5Z,8Z,11Z,13E,17Z)-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylica-5,8,11,13,17-pentaen-1-one</td>
<td>-8.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>campesterol</td>
<td>-8.4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>orthidine_A</td>
<td>-8.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>robustaflavone</td>
<td>-8.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tetrahydrorobustaflavone</td>
<td>-8.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>siphenellinol_C</td>
<td>-8.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6,10-dimethyl-9-methylene-2-(4-methyl-1,2-dioxabicyclo[2.2.2]oct-5-en-1-yl) undec-5-ene</td>
<td>-7.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>spinescen</td>
<td>-7.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>eupholhelionon</td>
<td>-7.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>anchinopeptolide_A</td>
<td>-7.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>uzarigenin</td>
<td>-7.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>isorhamnetin_3- [3'''-feruloylhamnosyl (16) galactoside</td>
<td>-7.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1,2,3,6-tetra-O-galloy1-beta-D-glucose</td>
<td>-7.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(+)-silychristin</td>
<td>-7.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>scopofofarnol</td>
<td>-7.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>isoquercitrin_6''-O-p-hydroxybenzoate</td>
<td>-7.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(-)-(R,R)-7''-O-methylcupidaline</td>
<td>-7.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>quercetin-3-rutinoside</td>
<td>-7.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>auraptene</td>
<td>-7.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rutin</td>
<td>-7.7</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Results of the top 20 ligands from NANPDB as listed in Table 4.2 revealed the compound with structural formula S,5Z,8Z,11Z,13E,17Z)-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-
15-methyllicos-5,8,11,13,17-pentaen-1-one as showing the strongest binding to the beta tubulin due to its more negative binding affinity score of -8.7 Kcal/mol. This suggest that it could also be a potential lead compound.

**Figure 4.8. Docking pose of ZINC14760755, beta-tubulin receptor complex.** Pose shows how well fitted the ligand is in the binding pocket as visualised in PYMOL. The structure in cyan represents the receptor and the region encircled in black shows the docked ligand in the binding pocket.
Figure 4.9. Docking pose of Dichapetalin A and albendazole beta-tubulin receptor complex. A. Dichapetalin A due to its long molecular chain has side chains projecting out of the pocket. B. Albendazole is well fitted inside the pocket. Image was rendered in PYMOL. The structure in cyan represents the receptor and the region encircled in black shows the docked ligand in the binding pocket.
The interactions within the protein-ligand complexes were further analysed using the program LIGPLOT. The number of hydrogen bonds formed, bond distances and the interacting residues are shown in Tables 4.3 and 4.4. The hydrogen bond distances contribute to the stability of the ligands within the active site. Even though, ZINC14760755

Figure 4.10. Docking pose of S,5Z,8Z,11Z,13E,17Z-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylicos-5,8,11,13,17-pentaen-1-one from North African Database. Image was rendered in PYMOL. The structure in red, blue and white representation is the receptor and the region encircled in black shows the docked ligand in the binding pocket.

4.5 Interaction Profile using LIGPLOT

The interactions within the protein-ligand complexes were further analysed using the program LIGPLOT. The number of hydrogen bonds formed, bond distances and the interacting residues are shown in Tables 4.3 and 4.4. The hydrogen bond distances contribute to the stability of the ligands within the active site. Even though, ZINC14760755
had the stronger binding with the receptor, it had fewer hydrogen bond interactions and more hydrophobic interactions with residues of the active site as compared to ZINC28462577, which had its complex stabilised with four hydrogen bonds through GLU198, GLN134, ASN256 and LYS350 (Figure 4.11). Remarkably, ZINC28462577 was found to have stabilised with GLU198, which is a key residue associated with anthelminthic activity with relatively strong hydrogen bonding (shorter bond length). The bond length of the interaction between ZINC28462577 and GLU198 is 3.05 Å and this happens to be the shortest length compared to other compounds in Table 4.3. This suggests that both ZINC28462577 and ZINC14760755 are potential inhibitors. The results also indicate that most of the ligands are stabilized inside the binding site mainly by a hydrogen bond with ASN247 and GLN134 (Table 4.3). Albendazole and Dichapetalin did not form hydrogen bonds with Phe167, Phe200 and Glu198, which are key residues associated with anthelminthic activity but stabilised with hydrophobic interactions instead (Figure 4.12). Dichapetalin A was involved in hydrophobic interaction with residues Val255, Lys350, Asn247 and Ala315 while albendazole had hydrophobic contacts with Ala 315, Glu198 and Phe200. It is therefore tempting to suggest that perhaps Dichapetalin A, a potential anthelminthic compound may have different mechanisms of interaction from albendazole, even though, both are predicted to bind Ala315 through hydrophobic interaction bond within the colchicine binding site of the beta tubulin target in \textit{N. americanus}.

The interaction profile of ligand $S,5Z,8Z,11Z,13E,17Z)-15$-hydroxy-$1-(2,4,6$-trihydroxyphenyl)$-15$-methylcosa-$5,8,11,13,17$-pentaen-$1$-one (Table 4.4) reveal stabilization of the ligand inside the binding site mainly by hydrogen bonds with VAL236,
ASN256 and GLN134 with relatively short bond lengths of 2.87, 3.21 and 0.00 respectively (Figure 4.13). The results also show that the docked ligands in Table 4.4 make hydrophobic contacts with Phe200, Phe167 and Glu198 which suggests a weak stabilisation with those key residues (Appendix C). ZINC14760755 and $S,5Z,8Z,11Z,13E,17Z\)-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylcicosa-5,8,11,13,17-pentaen-1-one do not form hydrogen bonds with Phe167, Phe200 and Glu198 but rather hydrophobic bonds. As reported, these key residues of beta tubulin protein of hookworm are implicated in drug resistance [46, 178, 179]. Perhaps, ZINC14760755 and $S,5Z,8Z,11Z,13E,17Z\)-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylcicosa-5,8,11,13,17-pentaen-1-one ligands have alternative binding modes.
Figure 4.11. Interaction profile of ZINC14760755 and ZINC28462577. Interaction profile of A. ZINC14760755 and B. ZINC28462577 from AfroDb. The green dotted lines indicate hydrogen bond interactions between the residues in green and ligand in blue. The residues behind red radiating spokes are involved in hydrophobic interaction with the ligand.
Figure 4.12. Interaction profile of the Dichapetalin A. and albendazole ligands. Interaction profile of A. Dichapetalin A and B. Albendazole. The green dotted lines indicate hydrogen bond (H-bond) interactions between the residues and ligand in blue. The residues behind red radiating spokes are involved in hydrophobic interaction with the ligand.
Figure 4.13. Interaction profile of S,5Z,8Z,11Z,13E,17Z-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylcosa-5,8,11,13,17-pentaen-1-one from North African Database as predicted by LIGPLOT [146]. The green dotted lines indicate H-bond interactions between the residues in green and ligand in blue. The residues behind red radiating spokes are involved in hydrophobic interaction with the ligand.
Table 4.3. Results of number of hydrogen-bonds/hydrophobic-bonds and contact residues of top ten ligands from AfroDB, and that of Dichapetalin A and albendazole.

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Binding Affinity (kcal/mol)</th>
<th>Number of hydrogen bonds</th>
<th>Key residue contact residues and hydrophobic residues</th>
<th>Bond distance (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZINC14760755</td>
<td>-8.6</td>
<td>2</td>
<td>ALA315, LYS350</td>
<td>2.77, 3.01</td>
</tr>
<tr>
<td>ZINC95485927</td>
<td>-8.5</td>
<td>0</td>
<td>Hydrophobic contacts</td>
<td>none</td>
</tr>
<tr>
<td>ZINC95486082</td>
<td>-8.5</td>
<td>2</td>
<td>ASN247</td>
<td>3.07, 3.32</td>
</tr>
<tr>
<td>ZINC95486263</td>
<td>-8.5</td>
<td>3</td>
<td>GLN134, GLU198, ASN247</td>
<td>2.89, 3.10, 3.25</td>
</tr>
<tr>
<td>ZINC14780716</td>
<td>-8.3</td>
<td>2</td>
<td>GLU198, ALA315</td>
<td>3.16, 2.95</td>
</tr>
<tr>
<td>ZINC95485922</td>
<td>-8.3</td>
<td>3</td>
<td>ASN247, ASN256, LYS350</td>
<td>2.86, 2.97, 3.13</td>
</tr>
<tr>
<td>ZINC95486052</td>
<td>-8.3</td>
<td>0</td>
<td>Hydrophobic</td>
<td></td>
</tr>
<tr>
<td>ZINC95485928</td>
<td>-8.2</td>
<td>2</td>
<td>CYS239, ALA315</td>
<td>3.34, 3.00</td>
</tr>
<tr>
<td>ZINC13480348</td>
<td>-8.1</td>
<td>1</td>
<td>ASN247</td>
<td>3.00</td>
</tr>
<tr>
<td>ZINC28462577</td>
<td>-8.0</td>
<td>4</td>
<td>GLU198, GLN134, ASN256, LYS350</td>
<td>3.05, 2.75, 3.31, 3.32</td>
</tr>
<tr>
<td>DICHAPETALIN A</td>
<td>-5.8</td>
<td>0</td>
<td>Hydrophobic (VAL255, LYS350, ASN247, ALA315)</td>
<td>none</td>
</tr>
<tr>
<td>ALBENDAZOLE</td>
<td>-5.6</td>
<td>0</td>
<td>Hydrophobic (PHE200, GLU198, ALA315)</td>
<td>none</td>
</tr>
</tbody>
</table>
Table 4.4. Results of number of hydrogen-bonds/hydrophobic-bonds and contact residues of top ten ligands from NANPDB.

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Binding Affinity (kcal/mol)</th>
<th>Number of hydrogen bonds</th>
<th>Key residue contact residues and hydrophobic residues</th>
<th>Bond Distance (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S,5Z,8Z,11Z,13E,17Z-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylicosa-5,8,11,13,17-pentaen-1-one</td>
<td>-8.7</td>
<td>3</td>
<td>GLN134, VAL236, ASN256</td>
<td>2.87, 3.21, 0.00</td>
</tr>
<tr>
<td>campesterol</td>
<td>-8.4</td>
<td>0</td>
<td>Hydrophobic contacts (GLU198)</td>
<td>none</td>
</tr>
<tr>
<td>orthidine_A</td>
<td>-8.2</td>
<td>2</td>
<td>ASN256, ASN247, THR351, ASN348</td>
<td>3.02, 3.12, 3.13</td>
</tr>
<tr>
<td>robustaflavone</td>
<td>-8.2</td>
<td>3</td>
<td>ASN256, GLN134</td>
<td>3.21, 2.64</td>
</tr>
<tr>
<td>tetrahydrorobustaflavone</td>
<td>-8.2</td>
<td>2</td>
<td>ASN256, GLN134</td>
<td>3.22, 2.64</td>
</tr>
<tr>
<td>siphonellinol_C</td>
<td>-8.1</td>
<td>3</td>
<td>ASN256, VAL313, ASN348</td>
<td>2.69, 3.26</td>
</tr>
<tr>
<td>6,10-dimethyl-9-methylene-2-(4-methyl-1,2-dioxabicyclo [2.2.2] oct-5-en-1-yl) undec-5-ene</td>
<td>-7.9</td>
<td>0</td>
<td>Hydrophobic (GLU198, PHE200, PHE167)</td>
<td>none</td>
</tr>
<tr>
<td>spinescen</td>
<td>-7.9</td>
<td>2</td>
<td>Hydrophobic (GLU198, PHE200)</td>
<td>3.34, 3.00</td>
</tr>
<tr>
<td>euphohelionon</td>
<td>-7.9</td>
<td>1</td>
<td>ASN256</td>
<td>3.19</td>
</tr>
<tr>
<td>anchinopeptolid_A</td>
<td>-7.9</td>
<td>3</td>
<td>ASN256, ASN247, THR351</td>
<td>3.15, 3.17, (2.95,3.10)</td>
</tr>
</tbody>
</table>
4.6 ADME Prediction and Pharmacokinetic Properties

The results of the ADME properties (Tables 4.5 and 4.6) revealed that most of the virtually screened compounds had relatively low ESOL LogS which gave the indication of their poor solubility class of compounds. ZINC14760755 from Table 4.1 had a high gastrointestinal (GI) absorption which suggests that the compound can be absorbed into the intestinal tract when administered orally. The Blood Brain Barrier (BBB) penetration as the name implies gives an indication of the likelihood of the drug being delivered to the central nervous system (CNS). The top ten ligands in addition to Dichapetalin A and albendazole from AfroDB were found to have no permeation into the blood brain barrier with exception of ZINC9548608 (Table 4.5). In terms of distribution, the P-glycoprotein (P-gp) are important members of ATP transporters for active efflux through membranes. Knowledge about whether a compound is a substrate or not to P-gp provides an indication of how well it will be distributed. ZINC14760755, ZINC28462577, Dichapetalin A and albendazole were all found as a non-substrate/inhibitor of P-gp suggesting desirable distribution of the compounds in the circulatory system when administered. In terms of metabolism, Dichapetalin A was found to be relatively a better non-inhibitor of the CYP450 proteins while the others had at least one inhibition to a CYP450 protein (Table 4.5). The CYP450 are a superfamily of iso-enzymes and key players in drug elimination [180]. Any inhibition to the CYP450 will lead to accumulation and drug to drug interactions due to low clearance of the drugs [180]. In terms of drug-likeness, the Lipinski rule of five (ro5) [181] was used as a measure which has the following criteria: the molecular weight should be less than 500, the lipophilicity, LogP (the logarithm of the partition coefficient between water and 1-octanol) should be less than 5, the number of
hydrogen bond donor atoms in the molecules should be less than 5, and the number of hydrogen bond acceptors should be less than 10. The physicochemical descriptors computed using DataWarrior and SwissADME were used in the determination of whether a drug violated the ro5 or not. From the results provided in Table 4.5, ZINC14760755 and albendazole passed all Lipinski ro5 while ZINC28462577 and Dichapetalin A failed the Lipinski’s ro5 with one and two violations respectively. This suggests that ZINC14760755 exhibit high druglikeness and is potentially relevant in hookworm drug discovery [181].

The results of the ADME properties computed by SwissADME for the top ten NANPDB compounds (Table 4.6) revealed that most of the compounds, similar to the top ten compounds in Table 4.5, had relatively a low ESOL Log S of -5.55 (poor solubility). The potential lead compound, S,5Z,8Z,11Z,13E,17Z-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylicosa-5,8,11,13,17-pentaen-1-one had a low GI absorption. Results of the prediction indicate favourable distribution and excretion predictions for most of the compounds in terms of the Pgp inhibition and inhibition to CYP450 isoenzymes. In terms of Lipinski ro5, S,5Z,8Z,11Z,13E,17Z-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylicosa-5,8,11,13,17-pentaen-1-one satisfied all ro5 criteria and therefore may indicate that it is the more druglike than its counterparts [181]. All top ten compounds from NANPDB were non-permeant into the blood brain barrier. Table 4.6 provides a comprehensive list of the ADME prediction results for the top ten compounds from NANPDB.
Table 4.5. Results of ADME prediction of top ten virtually screened compounds and that of Dichapetalin A and albendazole.

Abbreviations: GI, gastrointestinal; BBB, blood brain barrier, ESOL, estimated solubility; Pgp, P-glycoprotein, CYP, cytochrome P450.

<table>
<thead>
<tr>
<th>Compound</th>
<th>GI absorption</th>
<th>BBB permeant</th>
<th>ESOL Log S</th>
<th>ESOL Class</th>
<th>Lipinski # violations</th>
<th>Bioavailability Score</th>
<th>Pgp substrate</th>
<th>CYP1A2 inhibitor</th>
<th>CYP2C19 inhibitor</th>
<th>CYP2C9 inhibitor</th>
<th>CYP2D6 inhibitor</th>
<th>CYP3A4 inhibitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZINC14760755</td>
<td>High</td>
<td>No</td>
<td>-6.52</td>
<td>Poorly soluble</td>
<td>0</td>
<td>0.55</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>ZINC95485927</td>
<td>Low</td>
<td>No</td>
<td>-10.55</td>
<td>Insoluble</td>
<td>2</td>
<td>0.17</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>ZINC95486082</td>
<td>High</td>
<td>Yes</td>
<td>-5.82</td>
<td>Moderately soluble</td>
<td>0</td>
<td>0.55</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>ZINC95486263</td>
<td>Low</td>
<td>No</td>
<td>-6.99</td>
<td>Poorly soluble</td>
<td>3</td>
<td>0.17</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>ZINC14780716</td>
<td>High</td>
<td>No</td>
<td>-6.55</td>
<td>Poorly soluble</td>
<td>0</td>
<td>0.55</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>ZINC95485922</td>
<td>Low</td>
<td>No</td>
<td>-7.48</td>
<td>Poorly soluble</td>
<td>0</td>
<td>0.55</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>ZINC95486052</td>
<td>High</td>
<td>No</td>
<td>-6.03</td>
<td>Poorly soluble</td>
<td>0</td>
<td>0.55</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>ZINC95485928</td>
<td>Low</td>
<td>No</td>
<td>-7.19</td>
<td>Poorly soluble</td>
<td>0</td>
<td>0.55</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>ZINC13480348</td>
<td>High</td>
<td>No</td>
<td>-6.26</td>
<td>Poorly soluble</td>
<td>0</td>
<td>0.55</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>ZINC28462577</td>
<td>Low</td>
<td>No</td>
<td>-7.14</td>
<td>Poorly soluble</td>
<td>1</td>
<td>0.55</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Dichapetalin A</td>
<td>High</td>
<td>No</td>
<td>-7.29</td>
<td>Poorly soluble</td>
<td>2</td>
<td>0.17</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Albendazole</td>
<td>High</td>
<td>No</td>
<td>-3.23</td>
<td>Soluble</td>
<td>0</td>
<td>0.55</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Table 4.6. Results of ADME prediction of top ten ranking compounds from NANNPDB. Abbreviations: GI, gastrointestinal; BBB, blood brain barrier, ESOL, estimated solubility; Pgp, P-glycoprotein, CYP, cytochrome P450.

<table>
<thead>
<tr>
<th>Compound</th>
<th>GI absorption</th>
<th>BBB permeant</th>
<th>ESOL Log S</th>
<th>ESOL Class</th>
<th>Lipinski violations</th>
<th>Bioavailability Score</th>
<th>Pgp substrate</th>
<th>CYP1A2 inhibitor</th>
<th>CYP2C19 inhibitor</th>
<th>CYP2C9 inhibitor</th>
<th>CYP2D6 inhibitor</th>
<th>CYP3A4 inhibitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>S,5Z,8Z,11Z,13E,17Z)-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylcos-5,8,11,13,17-pentaen-1-one</td>
<td>Low</td>
<td>No</td>
<td>-5.55</td>
<td>Moderately soluble</td>
<td>0</td>
<td>0.55</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>campesterol</td>
<td>Low</td>
<td>No</td>
<td>-7.54</td>
<td>Poorly soluble</td>
<td>1</td>
<td>0.55</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>orthidine_A</td>
<td>Low</td>
<td>No</td>
<td>-2.5</td>
<td>Soluble</td>
<td>1</td>
<td>0.55</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>robustaflavone</td>
<td>Low</td>
<td>No</td>
<td>-6.75</td>
<td>Poorly soluble</td>
<td>2</td>
<td>0.17</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>tetrahydrorobustaflavone</td>
<td>Low</td>
<td>No</td>
<td>-6.75</td>
<td>Poorly soluble</td>
<td>2</td>
<td>0.17</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>siphonellinol_C</td>
<td>High</td>
<td>No</td>
<td>-4.76</td>
<td>Moderately soluble</td>
<td>0</td>
<td>0.55</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>6,10-dimethyl-9-methylene-2-(4-methyl-1,2-dioxabicyclo [2.2.2] oct-5-en-1-yl) undec-5-ene</td>
<td>High</td>
<td>No</td>
<td>-5.33</td>
<td>Moderately soluble</td>
<td>1</td>
<td>0.55</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>spinescens</td>
<td>Low</td>
<td>No</td>
<td>-7.42</td>
<td>Poorly soluble</td>
<td>1</td>
<td>0.55</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>euphoheionon</td>
<td>Low</td>
<td>No</td>
<td>-8.95</td>
<td>Poorly soluble</td>
<td>2</td>
<td>0.17</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>anchinopeptolide_A</td>
<td>Low</td>
<td>No</td>
<td>-2.51</td>
<td>Soluble</td>
<td>2</td>
<td>0.17</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
4.7 Toxicity Prediction Analysis

The results of the toxicity study summarised in Table 4.7 suggest that most of the drug candidates were found not to be tumorigenic and irritant. Notably, ZINC28462577 was found to be pre-eminent in terms of mutagenicity, tumorigencity, reproductive effect and irritation since it was predicted to be safe under all conditions. This suggests that ZINC28462577 may have minimal and tolerable harmful effects when administered [182]. Dichapetalin A was predicted to show irritation while albendazole, a well-known anthelminthic drug [3, 186] was predicted to be safe under all conditions. Most compounds from the second set (NANPDB) including S,5Z,8Z,11Z,13E,17Z-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylicosa-5,8,11,13,17-pentaen-1-one were predicted to be toxicologically safe except euphohelionon which exhibited high tumerogencity and irritation effects. Anchinopeptolide A was also predicted to possess high irritation effects but safe under the other conditions of mutagenicity, tumorigenicity and reproductive effects (Table 4.7).
Table 4.7. Toxicological profile results of top ten ranking compounds from both set of virtual library compounds as predicted by DataWarrior.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Mutagenic</th>
<th>Tumorigenic</th>
<th>Reproductive Effective</th>
<th>Irritant</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZINC14760755</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>ZINC95485927</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>ZINC95486082</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>ZINC95486263</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>ZINC14780716</td>
<td>low</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>ZINC95485922</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>ZINC95486052</td>
<td>high</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>ZINC95485928</td>
<td>high</td>
<td>none</td>
<td>none</td>
<td>high</td>
</tr>
<tr>
<td>ZINC13480348</td>
<td>none</td>
<td>none</td>
<td>high</td>
<td>none</td>
</tr>
<tr>
<td>ZINC28462577</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Dichapetalin A</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>high</td>
</tr>
<tr>
<td>Albendazole</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Campesterol</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Orthidine A</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Robustaflavone</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Tetrahydrorobustaflavone</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Siphonellinol C</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>6,10-dimethyl-9-methylene-2-(4-methyl-1,2-dioxabicyclo[2.2.2]oct-5-en-1-yl) undec-5-ene</td>
<td>none</td>
<td>high</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Spinescens</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Euphohelionon</td>
<td>none</td>
<td>high</td>
<td>none</td>
<td>high</td>
</tr>
<tr>
<td>Anchinopeptolide A</td>
<td>none</td>
<td>none</td>
<td>none</td>
<td>high</td>
</tr>
</tbody>
</table>

4.8 Scaffold Analysis

We conducted a scaffold analysis between list A and list B and another analysis between the list A and list C (Section 3.10). This was done to assess scaffold diversity or otherwise
similarity to currently used anthelminthics. In addition, the analysis was performed to determine how unique the scaffolds are within the natural products because such findings may give an indication of unique mechanisms of action owing to unique affinity for biological targets [184]. Murcko scaffolds were used to characterise the compounds and scaffold counts/frequency were used to characterise the distribution of molecules over unique scaffolds. The results of the scaffold analysis of list A compared to the list B showed that 142 unique scaffolds were identified in list A whereas 13 unique scaffolds were identified in list B but a single scaffold overlap between list A and list B (Figure 4.17). The scaffold percentage was 71% and that of the known anthelminthics in list B was 81% uniquely represented by Mucko framework (Table 4.8). This gives an indication of the high scaffold diversity between the docked natural compounds dataset and anthelminthics. Analysis of the list A as compared to list C clearly shows different ring systems present suggesting a high diversity between the docked natural products and albendazole with mebendazole as illustrated in Figure 4.14B. The results of both analyses suggest a high diversity within the natural compounds when compared to their anthelminthic counterparts. These findings appears to support the recognition of the high chemical diversity in natural products when compared to synthetic libraries [183, 184].
Figure 4.14. A bar plot of scaffold counts versus the ring systems of 201 compounds from list A compared to list B and list C. A. Comparison between list A and list B. B. Comparison between list A and list C. The plot with an asterisk * against it is a plot of compounds that failed to generate Murcko scaffold.
Table 4.8. **Scaffold diversity analysis of natural products and anthelminthics.** Ns/M represents ratio of number of unique scaffold to total number of compounds.

<table>
<thead>
<tr>
<th>Category list</th>
<th>Ns/M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural Products (AfroDb, NANPDB, Dichapetalin A)</td>
<td>0.71</td>
</tr>
<tr>
<td>Anthelminthics</td>
<td>0.81</td>
</tr>
</tbody>
</table>

### 4.9 Proteochemometric Modelling

The dataset retrieved from BindingDB comprised bioactivity profiles of 3 tubulins with Uniprot IDs: Q25270, P02554, Q6B856 with a sample size of 437. Although, the dataset had a low sample size, the performance of the model was complemented by the choice of algorithm and optimisation strategies used in achieving the best model. The bioactivity assays within the dataset were labelled differently due to the different assay conditions. Additionally, the binding affinity values that have been provided as different inhibitory constants (Kd (nM), IC\textsubscript{50} (nM), Ki (nM) and EC\textsubscript{50} (nM), makes it difficult to correlate the bioactivity prediction by the PCM to a precise inhibition constant. Thus, it was labelled as active and inactive using the criteria provided in section 3.11.1. Figure 4.15 is a plot of the distribution of active subset and inactive subset of the labelled dataset (class variables). It can also be observed from the plot that there was a high imbalance within the dataset.

For the development of a PCM model, it required protein and ligand descriptors. The highly interpretable compound and protein descriptors that were computed included solvent accessibility, polarizability, relative hydrophobicity, predicted secondary structure, ring
counts, substructure counts and electro-topological state descriptors. The total number of compound descriptors that were computed was 859 and that of the protein was 146 (Table 4.9). This resulted in a high dimensional dataset of 437 sample containing 1005 descriptors.

![Distribution of response variable (class) in the dataset.](image)

**Figure 4.15.** Distribution of response variable (class) in the dataset. A bar plot of frequencies of the response variables or class labels within the dataset.

**Table 4.9.** Proteins and compounds descriptors used in the development of the model

<table>
<thead>
<tr>
<th>Category</th>
<th>Descriptor Category</th>
<th>Total number of descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compounds</td>
<td>Circular Morgan fingerprints and Padel descriptors</td>
<td>859</td>
</tr>
<tr>
<td>Proteins</td>
<td>composition/distribution/transition descriptors</td>
<td>146</td>
</tr>
</tbody>
</table>
4.9.1 Exploratory principal component analysis (PCA) of compounds and target datasets

The high dimensional descriptors of the dataset comprising of 437 interacting compounds within the dataset were analysed using PCA to determine how groups of the dataset are substantially different from each other in terms of the biological and chemical space respectively. A PCA analysis of the compounds showed three clustering of the compounds (blue, green and red). The closeness of the clusters explains that most of the compounds occupy the same chemical space [186]. The computed descriptors differentiated the compounds into three clusters, suggesting variability within the dataset. Therefore, these descriptors were included in the construction of the PCM [186]. However, there were some outliers as shown in the PCA plot (Figure 4.16A) indicating these compounds do not share similar chemical space with the rest. The distribution of the beta tubulin protein variants suggest that they occupy a widened biological space and gives an indication of the high variability within the dataset (Figure 4.16B). The first principal components (PC1) explained the maximum variance of 95% for the compounds and 97% variance for protein. The first components of the PCAs based on the variance were significantly large to describe the variability within the dataset. The PCA plots thus have been used to provide a visualisation of how the descriptors separate the compounds and the proteins into clusters or groups.
Figure 4.16. Chemical and biological space (compound–target interaction space) of beta tubulin- inhibitor dataset. A. The PCA analysis of chemical descriptors shows an overlap of compound descriptors in PC1 and PC2 space. B. The widened distribution within the amino acid descriptors that suggests a wide variation within the orthologues information.
4.9.2 Model development

An SVM model was developed for prediction of antitubulin activity of compounds based on the bioassay dataset from BindingDb. The SVM model was constructed using a combination of the protein and ligand descriptors. Non-linear radial basis kernel function with stratified 10-fold CV was used to train the dataset while performing a grid search with the hyperparameters C and gamma to optimise the model. The model was then predicted on the held-out test set that was not used in training the model. The CV was used to select the best kernel parameters and to evaluate the best performance of the model. The choice of stratified CV was as a result of imbalance dataset (Figure 4.15) within the class variables which could result in overfitting of the model for the majority class because of more non-active than active datasets.

4.9.2.1 Model validation

To determine the predictive ability of the model on the test set, the MCC, AUC and classification error metrics were used. The AUC value for the model (87%) indicates that the PCM model achieved a good predictive ability (Figure 4.17). The best hyperparameters achieved were: {'C': 1000, 'gamma': 0.001}. The model was able to correctly classify 129 non-actives (97%) and 10 actives (77%) with an overall accuracy of 96% and MCC of 0.75 (Table 4.10). Due to the number of inactive being more than active compounds in the training set, a stratified cross validation technique which was used yielded a model with good sensitivity and specificity, although the specificity was rather higher (Table 4.10). The classification error of the model was 0.04, which shows a better-
balanced performance of the model. Based on the performance metrics, it can be concluded that the model yielded an overall good performance when tested on the independent test set (Table 4.10).

After exhaustive search of literature, it appears that this is the first time a PCM has been applied to the bioactivity profiling of beta tubulin receptors. The performance of the model was compared to others [19, 186]. Cao et al. [188] trained a random forest classifier on 13,079 data samples retrieved from BindingDB and PDSP Ki database. The target space was described with CTD descriptors and ligand space described with hashed circular morgan fingerprints. The classifier performed well with an AUC of 0.96. Fernandez et al [189] reported a SVM based PCM for ligand-target modelling trained on a total of 8,235 inhibitors for 95 sequences of kinase. The SVM could classify 82% of data to be stable or unstable indicating a reasonably high performance of the model. Lapins et al [187] developed a unified PCM model which showed excellent predictive ability with internal AUC of 0.923 and an external AUC of 0.940 for predicting the inhibition of five major drug metabolizing CYP isoforms (CYP1A2, CYP2C9, CYP2C19, CYP2D6 and CYP3A4) using 63,391 data samples. Comparing the SVM based PCM model built in this study to the aforementioned, it could be suggested that although the model was trained on a relatively smaller experimental dataset of 437 for 3 beta tubulin variants, it performed reasonably well with an AUC of 87%. It is recommended that in order to enhance the performance of the PCM model and be able to significantly compare it to other models, especially in the case of Cao et al [188] and Fernandez et al [189], a large bioactive dataset...
must be generated from an experimental high through-put screening of beta tubulin receptors to retrain the classifier.

**Figure 4.17. Area under Receiver Operating Curve (AUC).** A plot of TPR against false positive rate (FPR) constitutes the ROC curve shown in blue. The red dashes represent the chance level. A ROC curve lying above the chance level has usually has an AUC greater than 0.5.

**Table 4.10. SVM model parameters and evaluation of classification performance**

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Kernel</th>
<th>Parameter</th>
<th>AUC (test set)</th>
<th>Overall Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>MCC</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>RBF</td>
<td>C=1000 gamma= 0.001</td>
<td>87%</td>
<td>96%</td>
<td>77%</td>
<td>97%</td>
<td>75%</td>
<td>96%</td>
</tr>
</tbody>
</table>
CHAPTER 5

CONCLUSION AND RECOMMENDATION

This study was intended to apply computational methods in the identification of novel anthelminthic drugs from natural products as well as using support vector machine based -proteochemometric modelling to predict the bioactivity of compounds to beta tubulin targets. The specific aims were homology modelling of 3D structure of beta tubulin of *Necatar americanus*; virtual screening of naturally derived compounds for the identification of potential anthelminthic agents; evaluation of pharmacological, drug-likeness and toxicity profile of lead compounds; comparative analysis of scaffolds of the docked natural products and known synthetic anthelminthics, specifically albendazole and mebendazole; and preliminary exploration of proteochemometric based machine learning model as a plausible technique for bioactivity profiling of beta tubulin receptors. The main conclusions are presented herein:

- Homology modelling was used to model a monomeric protein that is folded into a \( \beta \) domain consisting of 11-stranded \( \beta \)-sheet and 11 \( \alpha \)-helices.
- Analysis of the modelled protein using molecular dynamic simulation showed a good dynamic behaviour for a period of 1ns which allowed us to subject the stabilised receptor or protein to molecular docking.
- Molecular docking and computational modelling techniques have been utilised for the identification of potential natural product-derived compounds against hookworm from AfroDB and NANPDB databases. ZINC28462577 from AfroDB, and S,5Z,8Z,11Z,13E,17Z-15-hydroxy-1-(2,4,6-trihydroxyphenyl)-15-methylicosa-5,8,11,13,17-pentaen-1-one, from NANPDB, were selected as the most favourable...
potential inhibitors when binding energy, interaction profile and pharmacological properties were considered.

- Analysis of the scaffold of the docked natural compounds as compared to albendazole and mebendazole revealed different ring systems present and therefore led us infer a high diversity between the docked natural products and the known anthelminthics.

- In addition, this study developed a PCM model for the prediction of the inhibition of compounds against beta tubulin variants using a curated experimental dataset. The chemical compounds retrieved from the curated dataset were represented with circular or morgan fingerprints while the proteins were described by CTD descriptors. The training dataset comprising 437 data samples and 1005 total descriptors were still retained after pre-processing. The PCM model was built using support vector machine based on the radial basis kernel method with stratified 10-fold CV and this yielded a model with a good overall predictive performance with AUC of 87%, MCC of 75%, overall accuracy of 96% and a classification error of 4%.

One of the challenges that was encountered with regards to the virtual screening was the huge computational cost involved. It is recommended that future virtual screening could be done on high performance computing clusters. Furthermore, the compounds identified in this study must be experimentally characterised for possible pre-clinical trials. PCM has significant advantage because its predictive ability can be extrapolated to other related targets that were not used to train the model. Due to the paucity of experimental dataset on beta tubulin specifically for hookworm bioactivity assays, a future direction can include
high through-put screening of compounds in the wet lab to generate larger dataset on hookworm. The larger datasets can then be used to train the PCM-SVM model to enhance the performance and increase its reliability for prediction pertaining to hookworm.
REFERENCES


[190] Turner PJ. XMGRACE, Version 5.1.19. Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology, Beaverton, OR; 2005

APPENDICES

APPENDIX I. REPOSITORY OF SUPPORTING FILES

All python scripts related to this research have been deposited into a github repository. The repository is available at https://github.com/odam23/Hookworm-Drug-Discovery.git. The fasta sequence of the tubulin that was used to build a homology model has been also stored in the repository along with the related python scripts that were used to build it and the pdb format of the homology model itself. Shell scripts that were used for performing molecular docking with Vina have also been stored along with the resulting protein-ligand pdb complexes and docking results. The dataset and all the PCM model scripts described in the PCM predictive server section have been stored in the folder “PCM” in the repository along with the related python script and instructions to install required libraries. The SVM model that is used to predict a new set of data resides in the “model” subfolder in PCM directory within the repository.
APPENDIX II

Pymol visualisation of protein-ligand complexes for the natural products including Dichapetalin A and albendazole
Complexes of docked compounds from the first set of screened virtual library
ZINC14760755, ZINC95485927, ZINC95486082, ZINC95486263, ZINC14780716, ZINC95485922, ZINC95486052, ZINC95485928, ZINC1348034, ZINC28462577, Dichapetalin A and albendazole respectively
APPENDIX III

Interaction profile of the protein-ligand complexes using LIGPLOT

Binding interactions of ZINC95485927, ZINC95486082, ZINC95486263, ZINC14780716, ZINC95485922, ZINC95486052 accordingly
Binding interactions of ZINC95485928, ZINC13480348, robustaflavone and tetrahydrorobustaflavone respectively
Binding interactions of anchinopeptolide_A and tetrahydrorobustaflavone
Binding interactions of campesterol and tetrahydrorobustaflavone
Binding interactions of orthidine_A and tetrahydrorobustaflavone
Binding interactions of eupholion and tetrahydronobustafavone
## APPENDIX IV

**Chemical formula of the top 20 compounds from AfroDB, Dichapetalin A, and albendazole**

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Chemical formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZINC14760755</td>
<td>3-[(2E)-3,7-dimethylocta-2,6-dienoxy]-1,8-dihydroxy-6-methyl-10H-anthracen-9-one</td>
</tr>
<tr>
<td>ZINC95485927</td>
<td>[(3S,4aR,6aR,6bS,8aR,12aS,14aR,14bR)-4,4,6a,6b,8a,11,11,14b-octamethyl-1,2,3,4a,5,6,7,8,9,10,12,12a,13,13a,13b,14b-1,2,3,4,5,6,7,8,9,10,11,12,13,13a,13b,14b-diheptamethylnaphthacen-9-one]</td>
</tr>
<tr>
<td>ZINC95486082</td>
<td>(2S)-2-[2,2-dimethyl-8-(3-methylbut-2-enyl)chroman-6-yl]-7-hydroxy-chroman-4-one</td>
</tr>
<tr>
<td>ZINC95486263</td>
<td>2-[(5,7-dihydroxy-4-oxochroman-2-yl)-2-hydroxyphenoxy]-3-hydroxy-phenyl]-5,7-dihydroxy-chrome</td>
</tr>
<tr>
<td>ZINC14780716</td>
<td>Stipulin</td>
</tr>
<tr>
<td>ZINC95485922</td>
<td>2-[(2E)-3,7-dimethylocta-2,6-dienyl]-1,3,5,8-tetrahydroxy-4-(3-methylbut-2-enyl)xanthan-9-one</td>
</tr>
<tr>
<td>ZINC95486052</td>
<td>(2S)-2-[2,2-dimethyl-8-(3-methylbut-2-enyl)chroman-6-yl]-5,7-dihydroxy-chroman-4-one</td>
</tr>
<tr>
<td>ZINC95485928</td>
<td>10-[(2Z)-3,7-dimethylocta-2,6-dienyl]-5,9,11-trihydroxy-3,3-dimethyl-pyran-3,2-[3,2-a]xanthan-12-one</td>
</tr>
<tr>
<td>ZINC13480348</td>
<td>[(2R)-7-[(2E)-3,7-dimethylocta-2,6-dienoxy]-5,10-dihydroxy-2-methyl-4-oxo-1,3-dihydroanthracen-2-yl]</td>
</tr>
<tr>
<td>ZINC28462577</td>
<td>DNC006449</td>
</tr>
<tr>
<td>ZINC95486072</td>
<td>heptamethylBLAHdione</td>
</tr>
<tr>
<td>ZINC95486073</td>
<td>hydroxy(heptamethyl)BLAHone</td>
</tr>
<tr>
<td>ZINC95486081</td>
<td>(2S)-7-hydroxy-2-[(2R,3S)-2-hydroxy-3-(3-methylbut-2-enyl)chroman-6-yl]chroman-4-one</td>
</tr>
<tr>
<td>ZINC33833639</td>
<td>(4aS,6aS,6aS,6bR,8aR,10S,12aR,14bS)-10-hydroxy-4a-(hydroxymethyl)-2,2,6a,6b,9,9,12a-heptamethyl-3,4,</td>
</tr>
<tr>
<td>ZINC95485992</td>
<td>(E)-1-[2,4-dihydroxy-5-[(3S)-3-hydroxy-4-methyl-pent-4-enyl]phenyl]-3-[4-hydroxy-3-[(E)-3-methylpent</td>
</tr>
<tr>
<td>ZINC95486074</td>
<td>heptamethylBLAHdiol</td>
</tr>
<tr>
<td>ZINC95486075</td>
<td>(3S,4aR,6aR,6bS,8R,8aS,12aS,14aS,14bR)-8a-(hydroxymethyl)-4,4,6a,6b,11,11,14b-heptamethyl-1,2,3,4a,5</td>
</tr>
<tr>
<td>ZINC13365959</td>
<td>3-[(1S)-1-(1H-indol-6-yl)-3-methyl-but-2-enyl]-6-(3-methyl-but-2-enyl)-1H-indole</td>
</tr>
<tr>
<td>ZINC13485435</td>
<td>Erybraedine C</td>
</tr>
<tr>
<td>ZINC15120680</td>
<td>DNC014426</td>
</tr>
<tr>
<td>Dichapetalin A</td>
<td>C_{38}H_{48}O_{5}</td>
</tr>
<tr>
<td>Albendazole</td>
<td>(5-(propylthio)-1H-benzimidazol-2-yl)carbamic acid methyl ester</td>
</tr>
</tbody>
</table>