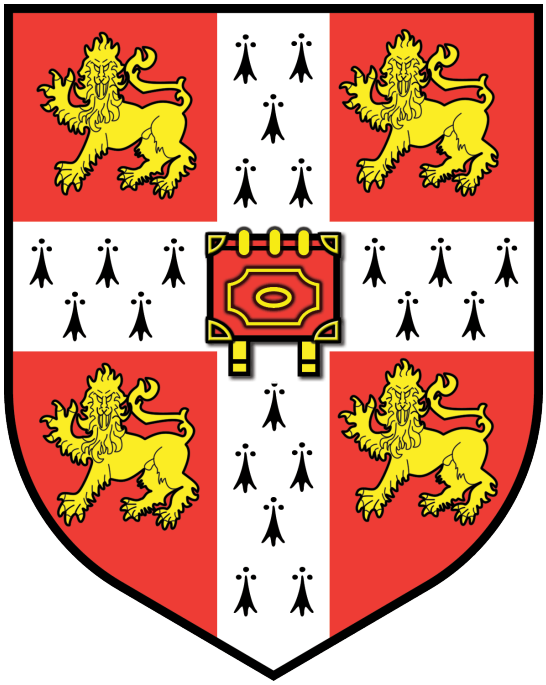
INTEGRATING CHEMICAL, BIOLOGICAL AND PHYLOGENETIC SPACES OF AFRICAN NATURAL PRODUCTS TO UNDERSTAND THEIR THERAPEUTIC ACTIVITY



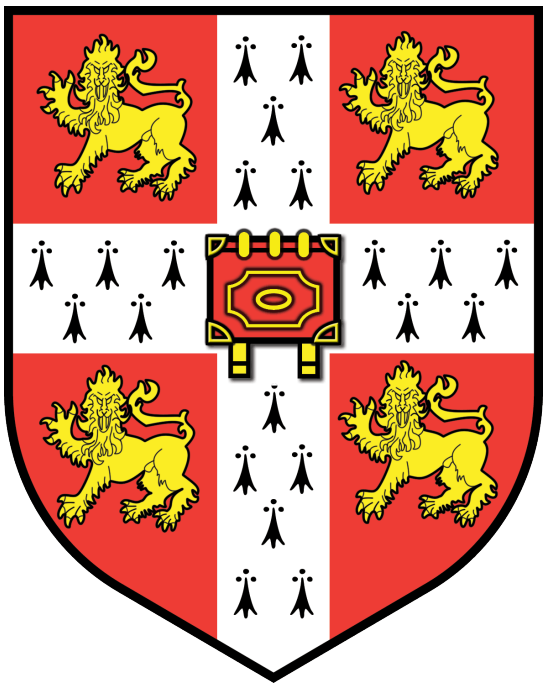
Fatima Magdi Hamza Baldo

Murray Edwards College

June 2018

This dissertation is submitted for the degree of Doctor of Philosophy

INTEGRATING CHEMICAL, BIOLOGICAL AND PHYLOGENETIC SPACES OF AFRICAN NATURAL PRODUCTS TO UNDERSTAND THEIR THERAPEUTIC ACTIVITY



Fatima Magdi Hamza Baldo

Ph.D.

June 2018

INTEGRATING CHEMICAL, BIOLOGICAL AND PHYLOGENETIC SPACES OF AFRICAN NATURAL PRODUCTS TO UNDERSTAND THEIR THERAPEUTIC ACTIVITY

Fatima Magdi Hamza Baldo

This research aims to utilise ligand-based target prediction to (i) understand the mechanism of action of African natural products (ANPs), (ii) help identify patterns of phylogenetic use in African traditional medicine and (iii) elucidate the mechanism of action of phenotypically active small molecules and natural products with anti-trypanosomal activity.

In Chapter 2 the objective was to utilise ligand-based target prediction to understand the mechanism of action of natural products (NPs) from African medicinal plants used against cancer. The Random Forest classifier used in this work compares the similarity of the input compounds from the natural product dataset with compound-target combinations in the training set. The more similar they are in structure, the more likely they are to modulate the same target. Natural products from plants used against cancer in Africa were predicted to modulate targets and pathways directly associated with the disease, thus understanding their mechanism of action e.g. “flap endonuclease 1” and “Mcl-1”. The “Keap1-Nrf2 Pathway” and “apoptosis modulation by HSP70”, two pathways previously linked to cancer (which are not currently targeted by marketed drugs, but have been of increasing interest in recent years) were predicted to be modulated by ANPs.

In Chapter 3, we aimed to identify phylogenetic patterns in medicinal plant use and the role this plays in predicting medicinal activity. We combined chemical, predicted target and phylogenetic information of the natural products to identify patterns of use for plant families containing plant species used against cancer in African, Malay and Indian (Ayurveda) traditional medicine. Plant families that are close phylogenetically were found to produce similar natural products that act on similar targets regardless of their origin. Additionally, phylogenetic patterns were identified for African traditional plant families with medicinal species used against cancer, malaria and human African trypanosomiasis (HAT). We identified plant families that have more medicinal species than would statistically be expected by chance and rationalised this by linking their activity to their unique phyto-chemistry e.g. the napthyl-isoquinoline alkaloids, uniquely produced by Acistrocladaceae and Dioncophyllaceae, are responsible for anti-malarial and anti-trypanosome activity.

In Chapter 4, information from target prediction and experimentally validated targets was combined with orthologue data to predict targets of phenotypically active small molecules and natural products screened against *Trypanosoma brucei*. The predicted targets were prioritised based on their essentiality for the survival of the *T. brucei* parasite. We predicted orthologues of targets that are essential for the survival of the trypanosome e.g. glycogen synthase kinase 3 (GSK3) and rhodesain. We also identified the biological processes predicted to be perturbed by the compounds e.g. “glycolysis”, “cell cycle”, “regulation of symbiosis, encompassing mutualism through parasitism” and “modulation of development of symbiont involved in interaction with host”.

In conclusion, *in silico* target prediction can be used to predict protein targets of natural products to understand their molecular mechanism of action. Phylogenetic information and phytochemical information of medicinal plants can be integrated to identify plant families with more medicinal species than would be expected by chance.

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit for the relevant Degree Committee.

The work has been carried out under the supervision of Dr Andreas Bender at the Centre for Molecular Informatics, Department of Chemistry, University of Cambridge.

ACKNOWLEDGEMENTS

First and foremost, I wish to thank my parents and siblings for their full and unconditional love, encouragement and support during this endeavour.

I would like to thank Dr Andreas Bender for his supervision during my study period.

I would also like to extend my gratitude to Dr Lucy Colwell for mentoring me and guiding me through my work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AMA	African Malay Ayurveda
ANP	African Natural Product
AU	Approximately Unbiased
BBB	Blood Brain Barrier
BGC	Biosynthetic Gene Cluster
BP	Biological Processes
CBIC	Chemical Bioactivity Information Centre
CC	Cellular component
CTD	Comparative Toxicogenomics Database
DNDi	Drugs for Neglected Diseases Initiative
DNA	Deoxyribonucleic Acid
EC_{50}	Half Maximal Effective Concentration
ECFP4	Extended Connectivity Fingerprint
FDA	Food and Drug Administration
GIT	Gastro-Intestinal
GO	Gene Ontology
GSK	GlaksoSmithKlein
HAT	Human African Trypanosomiases
IC_{50}	Half Maximal Inhibitory Concentration
K_d	Equilibrium dissociation constant
KEGG	Kyoto Encyclopedia of Genes and Genomes
K_i	Inhibitor constant
LDO	Least divereged orthologue
LS	Least squares
MDDR	MDL Drug Database Report
MDS	Multi-Dimensional Scaling
ME	Minimum Evolution
MF	Molecular Functions
MoA	Mechanism of Action
MoSS	Molecular Subsructure
MOST	Most-Similar ligand Target
NADI	Natural Product Discovery System
NCBI	National Centre for Biotechnology Information
NCI	National Cancer Institute
NIQ	Naphthylisoquinoline
NJ	Neighbout Joining
NP	Natural Product
PANTHER	Protein Analysis through Evolutionary Relationships
PASS	Prediction of Activity Spectra for Substances
PCM	Proteochemometric Modeliing
plogBB	Predicted brain/blood partition coefficient
QSAR	Quantitative Structure-Activity Relationship
RBS	Rubber Band Scaling

REVIGO	Reduce and Visualise Gene Ontology
RNA_i	Ribonucleic Acid Interference
SDF	Structure Data Format
SE	Shannon Entropy
SEA	Similarity Ensemble Approach
SH	Screen Hits Dataset
SOM	Self Organising Maps
SSE	Scaled Shannon Entropy
TAM	Traditional African Medicine
TAMOSIC	Targets Associated with its Most Similar Counterparts
TCM	Traditional Chinese Medicine
μM	Micro Molar
VSG	Variant Surface Glycoprotein
WHO	World Health Organisation

CHAPTER 1: INTRODUCTION

1.1 TRADITIONAL AFRICAN MEDICINES

Humans have used traditional medicines since before written history began, ranging from the Neanderthals in the Palaeolithic era[1], the Sumerians in Mesopotamia[2], Ancient Egyptians[2], to India since 4000 BC[3] and China since 2000 BC. Traditional medicines consist of: (a) entire organisms e.g. plant or animal, (b) part of an organism e.g. leaf or gland, (c) extracts, (d) exudates, (e) pure compounds or (f) venoms and toxins[4]. The early use of traditional medicines drives our aim to investigate the underlying mechanism of action of the NPs they contain.

The WHO[5] estimates that more than 30% of the population in developed countries and more than 80% of the population in developing countries use herbal medicines either to promote and maintain health or as treatments for diseases such as malaria, dysentery, and cancer. The significant use of traditional medicine by the population of developing countries may be attributed to the inadequate availability of pharmaceuticals in those areas, the low purchasing power of these communities, and that these natural products seem to work.

Of particular interest to our work are Traditional African Medicines (TAMs). The knowledge of African Traditional medicine is passed on from one generation of healers to the next by word of mouth, most often in the form of stories[6]. Prior to gaining access to this knowledge, apprentices are initiated into secret societies where they are educated in the aspects of TAM[7]. Traditional healers are known by different names in Africa e.g. *sangoma, n'anga,* and *inyanga*.

The main difference between TAM and western medicine is that TAM takes a holistic approach to treating illness[6], much like TCM[8] and Ayurveda[9]. In Africa a person is considered to exist in a balance of different aspects. These are shown in Figure 1:1 below:

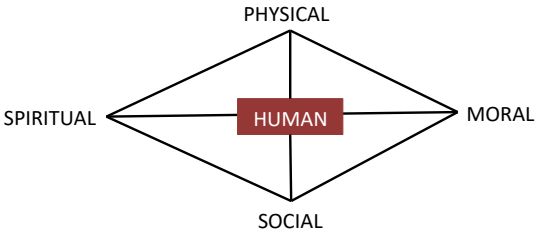


Figure 1:1 The different aspects that are in harmony in a healthy individual.

In TAM a person is considered to be in good health when the different aspects are in harmony. Alternatively, when one or more of these aspects is out of balance, a person becomes ill in health and "spirituality". In such cases, illness is considered to be caused by "supernatural ancestor spirit anger"[10]. During the healing procedure, all aspects ("spiritual", "moral", "social" or "physical disorders") are addressed. Thus, the holistic regimen taken in TAM includes not only the pharmacology of the medicine, but also physical characteristics of the medicine, e.g. aroma, taste, shape and colour, as well as attendant rituals, e.g. "incantation and song"[11]. The work in this thesis focuses on the pharmacological aspect of TAMs in the treatment of "physical disorders".

Plants that are the source of drugs (e.g. atropine, strychnine, the ergot alkaloids, physostigmine, d-tubocurarine), used in western medicine to selectively target the cause of the disease, are rarely used in TAM. This is due to a lack of precision technology in administering controlled doses of the plants[6]. Nevertheless, drugs have been discovered from African medicinal plants (see Table 1:1). This table shows the African medicinal plant and the natural products isolated from it that are responsible for medicinal activity. There are very few marketed drugs from these medicinal plants; these include Vincristine, Vinblastine and Reserpine, used as anti-cancer, anti-hypertensive and anti-psychotic respectively. The very small number of medicines isolated from the ~45,000 plants in the African flora[11] (5,000 of which have documented medicinal use[11]) allows a great deal of scope for further exploration and exploitation.

Table 1:1 Drugs from African medicinal plants

Plant	Activity	Natural Product
Catharanthus roseus L.	Anti-cancer	Vincristine Vinblastine
Combretum caffrum (Eckl. & Zeyh.) Kuntze	Anti-cancer	Combretastatins
Pausinystalia johimbe (K. Schum.) Pierre ex Beille	α-adrenergic agonist	Yohimbine
Physostigma venunosum Balf.	Cholinesterase inhibitor	Physostigmine
Rawolfia vomitoria Afzel.	Anti-hypertensive Anti-psychotic	Reserpine
Strophanthus gratus (Wall. &Hook.) Baill.	Cardiotonic	Ouabain
Tabernanthe iboga Baill.	Hallucigenic	Ibogaine

1.1.1 AFRICAN NATURAL PRODUCTS AND CANCER

As can be seen from the table above, natural products (which in the context of this work are "isolated and purified compounds from plant extracts" and their derivatives) isolated from medicinal plants from Africa play a role in the treatment of cancer. Cancer is a group of diseases characterised by an abnormal growth of cells. A cancerous growth is a malignant tumour whose cells continue to grow and divide uncontrollably and without coordination with normal tissues, and can then invade surrounding organs and other parts of the body[12]. According to the WHO, it is the second leading cause of death[13] after cardiovascular diseases, accounting for 8.8 million deaths in 2015. The most prevalent cancers in Africa are cervical cancer, breast cancer, liver cancer and prostate cancer, as well as Kaposi's sarcoma and non-Hodgkin's lymphoma[13].

The biological steps involved in the different stages of human cancer development, defined by Hanahan and Weinberg in 2000[12], are sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis. These were updated in 2011[14] to include deregulating cellular energetics and evading immune destruction. Modulation of targets involved in the pathways of these hallmarks forms the basis of targeted anti-cancer therapy. One of the main aspects to consider when studying NPs isolated from traditional medicines used against cancer is the definition of cancer as

used by traditional healers. “Cancer” in this case can include tumours, warts, excessive growths, excrescences, polyps, pustules and corns [15, 16]. These presentations may or may not be malignant and may present anywhere in the body. In the context of this work, the Western definition of cancer is used when considering plants to be studied.

Eighty-five of the 175 drugs (i.e. 49%) approved by the FDA for cancer treatment between 1940-2014 have either been NPs or derived from them[17]. The alkaloids 5-methoxymaculine, flindersiamine and 7-hydroxy-8-methoxydictamine from the plant *Oricia suaveolens* (Rutaceae) have shown significant cytotoxicity against lung adenocarcinoma cell lines, with IC_{50} values of 9.5, 7.9 and 8.9 μM respectively[18]. The vinca alkaloids from *Catharnathus rosaeus* are a prime example of African plants that are currently in the market as anti-cancer agents and are used for their antimicrotubule activity[19]. Terpenoids have been shown to supress NF-κB signalling[20], which is important in the pathogenesis of inflammatory diseases and cancer. These studies and others provided in the review by Simoben *et al*[21] and Nwodo *et al*[22] provide a promising start to the chance of finding novel NPs from TAMs with anti-cancer activity, as well as identifying the mechanism of action of those with activity.

1.1.2 HUMAN AFRICAN TRYPANOSOMIASIS

TAMs are also used to ameliorate or treat human African trypanosomiasis[23]. Human African trypanosomiasis (HAT) is a parasitic disease transmitted by the tsetse fly (Glossina sp) in 36 sub-Saharan African countries. It is caused by two kinetoplastids, namely, *Trypanosoma brucei gambiense* (west Africa) and *Trypanosoma brucei rhodesiense* (central and east Africa). A summary of the lifecycle of *Trypanosoma brucei* is shown in Figure 1:2.

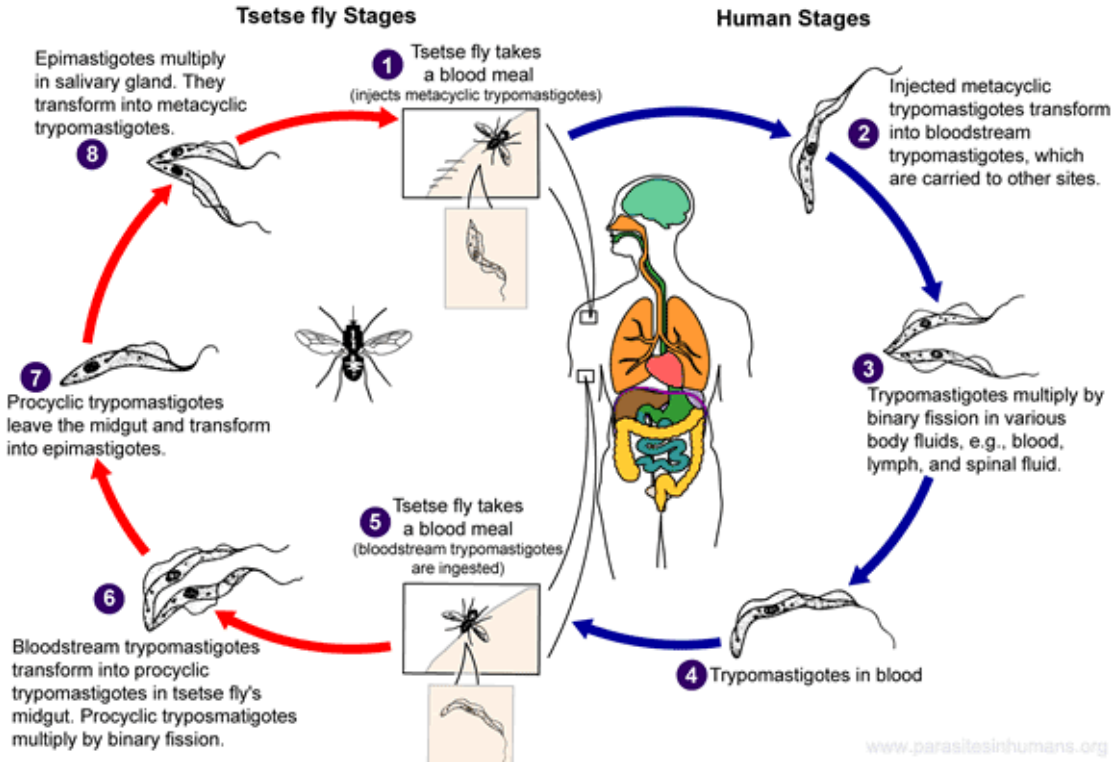


Figure 1:2 Life cycle of *Trypanosoms brucei*. Human stage: The tsetse fly bites a mammalian host delivering growth-arrested metacyclic trypomastigotes into the lymphatic system and eventually the blood stream. The metacyclic trypomastigotes differentiate into bloodstream trypomastigotes (long slender forms of the parasite), causing a bloodstream infection. They then penetrate the CNS by crossing the BB where they continue to replicate by binary fission. Tsetse fly stage: This is initiated when a tsetse fly takes short stumpy forms of the parasite in a blood meal from a mammalian host. They are transported to the midgut where they replicate by binary fission into procyclic trypomsatigotes and infect the midgut. The midgut procyclic trypomastigotes migrate within the fly to reach the proventiculus where they undergo differentiation and assymetric division to produce 1 long epimastigote and 1 short epimastigote. They then migrate onwards towards the salivary gland where the short epimastigote attaches to the salivary gland epithelium and undergoes assymetric division to metacyclic trypomastigotes, hence completing the cycle. (This image is a work of the Centers for Disease Control and Prevention, part of the United States Department of Health and Human Services, taken or made as part of an employee's official duties. As a work of the U.S. federal government, the image is in the public domain.)

Good control efforts have decreased the number of cases from 50,000 deaths in 2001 to fewer than 10,000 in 2009 and 2,804 reported cases in 2015. It is estimated that 11,000 people are currently affected.[24] The disease can be characterised by two stages: the symptoms of the first stage are fevers, headaches, itchiness and joint pains; and symptoms of the second stage (which occur when the parasite crosses the blood brain barrier (BBB) by expressing a parasite version of cysteine proteases causing an increase

in the oscillatory pattern of calcium ions in the cell) include confusion and trouble sleeping.

Treatment for the first stage is by Pentamidine and Suramin, whereas for the second stage, Melarospol, Elfornithine or a combination of Elfornithine and Nifutimox are used. Current drugs (i) have toxic side effects, e.g. encephalopathy (Melarospol)[25], (ii) require skilled workers for administration (all drugs)[26], (iii) have complex administration regimens (Elfonithine)[26] or (iv) are developing resistance (Melarospol)[26]. Potential new drugs for HAT in the drug discovery pipeline include Oxaboroles[27] (Preclinical), and Fexinidazole[28] which is currently in Phase III trials. Insufficient market forces could explain the lack of drive for more drugs entering the pipeline.

Current approaches to antitrypanosoid drug discovery include target-based screening[29] and phenotypic screening[30]. Both have advantages and drawbacks, discussed by Field *et al*[31] and Pink *et al*[32]. The recently published genome sequence of *Trypanosoma brucei* revealed approximately 1,500 genes coding for immunogenic Variant Surface Glycoproteins (VSG), which are the targets of vaccines[33]. For each individual organism, only one VSG is expressed at a time[34]. The VSG changes semi-randomly in 1 in every 100 cell divisions, meaning the antibodies generated by the body against the VSG no longer recognise the surface antigen of the parasite or its progeny as they have a new VSG coat[35], making it extremely difficult to develop a vaccine against HAT. It is thus important to identify new targets for anti-trypanosomals and to understand their MoA. The only validated target in HAT is ornithine decarboxylase[36], but others are suspected to be drug targets e.g. cysteine proteases, responsible for haemoglobin degradation, turnover of VSG and crossing of the BBB[37]; type II enoyl-acyl carrier protein reductase, responsible for fatty-acid biosynthesis[38]; and trypanothone reductase ,which is responsible for defence against oxidant and chemical stress[39]. Table 1:2 shows the important points as set out by the DDNDi[40] for selecting parasite molecular targets and ligands. In order to be considered a viable target, those targets identified to be modulated by compounds must be essential for the survival of the parasite. In this work, we consider this factor when predicting targets of phenotypically active anti-trypanosomal compounds. For the ligand, it is important to consider passage across the

blood brain barrier (important for stage 2 of the disease when the parasite crosses into the brain tissue). This is also a factor that we consider in this thesis when discussing the mode of action of phenotypically active anti-tryapanosomal compounds.

Table 1:2 Important points to consider in selecting parasite molecular targets and ligands

Target	Ligand
Target must be essential to the viability of the parasite (essential targets tend to be highly conserved among different species which causes a problem of selectivity)	Selectively active on parasite target – in some cases the host may have a survival mechanism e.g. the turnover rate of ornithine decarboxylase in humans is faster, thus allowing the ligand to be cidal to the parasite which is unable to regenerate the enzyme at a sufficient speed to overcome blockade
Must be amenable to modulation by drug-like ligands	Permeable in order to be able to access the target
Open to selective inhibition	Cross BBB
Structurally and chemically characterised	Orally active
Resistance potential e.g. single point mutations, over-expression of the target, efflux pumps gene amplification of target and inactivation of the drug	

1.1.2.1 NPs FROM AROUND THE WORLD WITH ANTI-TRYPANOSOME ACTIVITY

Promising results have been achieved by screening NPs from around the world against HAT. This is an important starting point to identify lead compounds for drugs against this disease. A review[41] outlines a range of natural product classes with activity against HAT and covers a period from the mid-1980s to 2003. Recently, eight plant extracts from North America were shown to have anti-trypanosomal activity with IC_{50} values < 1μg/ml and 125 plant extracts had activity with IC_{50} values < 10μg/ml, with none of the extracts showing toxicity towards THP1 cells. In 2017, Afrotryp, a public dataset of African compounds active against HAT, was released. The pharmacokinetic properties of NPs in this dataset were predicted and they were docked against six trypanosome targets, identifying nine compounds suitable for the treatment of stage two HAT, due to their low polar surface area. Taken together these results provide a promising start for investigating the target space in HAT from NP space, which is what we do in this thesis.

1.2 AFRICAN NATURAL PRODUCT DATABASES

Despite containing a plethora of unique phytochemicals with therapeutic value, knowledge of the mechanism of action of African phytochemicals remains largely unexploited; a point we wanted to address in this work. Recently several electronic databases have been created, which group the compounds according to their ethno-botanical uses and which can now be used together with novel computational tools to better understand the mechanism of action of African medicines (Table 1:3). These databases comprise in particular NANPDB, ConMedNP, AfroCancer and AfroMalaria and they form the basis of the current study. The North African Natural Products Database (NANAPDB) was published in 2017 and contains natural products from 4 Kingdoms (endophytes, animals, fungi, and bacteria) and 146 families with 98 reported activities. The North African region consists of Algeria, Egypt, Libya, Morocco, Sudan, South Sudan, Tunisia, Western Sahara, and parts of Northern Mali. The biggest classes of NP in this dataset are terpenoids (38%), flavonoids (22%) and alkaloids (10%). The compounds were collected from literature (links to literature are included in the database) as well as PhD theses from 1962-2016. Where available, information about the uses of the source species, experimentally verified activities and modes of action is also included. Compounds can be accessed via query searches and can be downloaded directly from the website. ConMedNP was published in 2014 and contains compounds from plants mainly from countries in the Congo Basin (Burundi, Central African Republic, Chad, Congo, Equatorial Guinea, Gabon, the Democratic Republic of Congo, Rwanda and the Republic of São Tomé and Príncipe). The NPs were collected from a literature search and PhD theses published between 1971-2013. These compounds come from 376 plant species and 79 plant families. AfroMalaria was also published in 2014 and contains 265 compounds from 131 species and 44 families from plants across Africa. The compounds were collected from literature sources and PhD theses published between 1971-2013. The major compound class in this dataset is the terpenoids (30.7%), followed by alkaloids (27.7%), flavonoids (12.9%), quinones (4.5%) and xanthones (4.5%). Twenty compounds in the datasets showed *in vivo* anti-malarial activities, while 278 compounds showed *in vitro* activities from moderate ($0.06\ \mu M \leq IC_{50} \leq 5\ \mu M$) to very high activities ($IC_{50} < 0.06\ \mu M$). AfroCancer,

published in 2014, contains compounds collected from literature sources and PhD theses between 1971-2014. The compound information, which is available on request from the authors, includes plant sources (species, genus and family), traditional uses of plant, region of collection of plant material, isolated metabolites, phytochemical class (e.g., flavonoid, alkaloid, etc.), and, where available, the measured biological activities of isolated compounds. The compounds are stored in SDF format.

It is important to note that these datasets are not complete. For example, not all compounds from all medicinal plants used against cancer, malaria or HAT have been characterized. Information about which NP is responsible for the activity of the medicinal plant is also not fully available. Furthermore, we don't have quantitative information or full bioactivity profiles of all the NPs in these datasets. While there is still much information missing, these datasets provide a starting point to understanding the mechanism of action of African medicinal plants.

Table 1:3 Datasets of African NPs analysed in the current study

Database	Number of Compounds	Notes
NANPDB[42]	4469	Natural products from 4 Kingdoms, available to download as SMILES, SDF-2D and SDF-3D
ConMedNP[43]	3,177	Compounds available in sdf format Annotated bioactivity can be obtained directly from the authors
AfroCancer[44]	364	Compounds available in .sdf format. Annotated bioactivity can be obtained directly from the authors. Some ligand-target information.
AfroMalaria[45]	265	Compounds available in .sdf format.

1.3 Advantages and Limitations of Natural Products in Drug Discovery

The advantages of using natural products isolated from traditional medicines as a starting point for drug discovery have been extensively reviewed[46]. One of the main advantages is that, since medicinal plants have been used for many generations to alleviate or treat symptoms, their tolerance levels and toxicities are relatively well

known[47]. This cannot be said for the NPs isolated from these plants, which is one of the reasons why it is important to elucidate their mechanism of action. Another advantage of natural products is that they are more likely than synthetic compounds to resemble endogenous metabolites and hence are more likely to enter the cells via active transport[47].

The limitations of natural products in the field of drug discovery were reviewed [48] and include (i) difficulty isolating and identifying the active compounds, (ii) synthesis of active constituents (too many chiral centres, rings, etc.) and (iii) elucidating and validating the mechanism of action. Another difficulty in the study of traditional medicines is that most are used in the form of an extract[49]. It is not yet understood whether the compounds act individually or in synergy with compounds that have little or no activity. A prime example of this is the case of the peroxysesquiterpene lactone, Artemesinin from the anti-malarial *Artemesia annua*, which has been found to be 30-60 times more active in leaf tea than when it is used alone[50]. The crude extract of the triterpenoid acids glycyrrhizin and glycyrrhetinic acid from *Glycyrrhiza glabra* root has been found to inhibit angiogenesis, whereas the isolated compounds promote angiogenesis[51]. Synergy may also play a role in the reduction of the toxicity of the isolated active compound as observed with the extract *of Rauwolfia serpentina*[52]. Another role that synergy may have is the enhancement of absorption of the active constituent when it is in the form of an extract. Phospholipids and polysaccharides found in the plant extracts may help in the absorption and hence increase in blood levels of the compounds compared to when they are administered individually (e.g. in the case of flavonoids)[53, 54]. This may be explained by the NMR studies which reveal that interaction occurs between the polar heads of the phospholipids and the phenolic groups of the flavonoids[50]. But most of the phospholipids and polysaccharides are removed during the early phases of extraction. Secondary metabolites also have the risk of being recognised as xenobiotics and being exported out of the cell. Despite these drawbacks, NPs are still an important source for drugs. The limitations need to be considered when making choices about which compounds are to be chosen as HIT/lead compounds.

Limitations for natural product drug discovery of particular relevance to Africa are as follows: low levels in investment in science and technology, lack of collaboration and

coordination between different research groups for various reasons including diplomatic and political issues, and limited drug discovery expertise, are among the problems faced in the continent[55]. Other issues highlighted by Ntie-Kang[56] include: results obtained in labs not being extrapolated to an industrial setting due to funding shortages, results obtained by various research groups in the same field not made public and curated in a single repository, and labs not equipped to carry out high throughput screens of the isolated and purified active components. The most that can be done in terms of research is the collection of information from the local healers, collection of the plants from their area of origin, taxonomic identification and extraction of fractions, which then need be sent away to European labs for isolation, purification and HTS.

To benefit from the advantages of natural products (NPs) and begin to address some of the limitations, it is important to understand their underlying mechanism of action. One of the ways of achieving this is to identify the targets and pathways modulated by these NPs. This will contribute towards understanding their medicinal activity, metabolic profiles as well as their toxicities. In this work we address their medicinal activity.

1.4 MODE OF ACTION ANALYSIS

Mode of action analysis comprises the study of biochemical and physiological mechanisms by which compounds or drugs elicit a response. This step is important for elucidating the mechanism of action (MoA) of a drug candidate. The importance of elucidating the mechanism of action of NP is two-fold. First, knowing the target, and hence the pathway that these NPs modulate, validates the use of the natural products by the herbalists and will inform authorities in the regulation of their use. MoA information allows medicinal chemists to understand the side effects and toxicity of the NPs. Here we will briefly mention the three main approaches to target identification, namely biochemical methods, genetic interaction methods and computational inference. Computational inference methods will then be discussed in greater detail, as this is the area on which the thesis mainly focuses.

Biochemical Methods: this involves labelling either the compound or target of interest and incubating them together for some time, followed by direct measurement of binding[57].

Genetic Interaction: this method involves altering the functions of putative targets by, for example, gene knockout, RNAi or small molecules[58, 59]. This allows for a target hypothesis to be generated.

Computational Inference: In these methods, pattern recognition is used to compare the effects of the tested compounds to those with known and validated activities. Here hypotheses are made about targets/pathways, but the results remain to be experimentally validated. Thus combining computational inference methods with direct measurements is a good approach for target convolution for mode of action analysis[57].

These methods detect interactions between ligands and targets, and not the actual mechanism of action of the ligands. Once a ligand is bound to a target/receptor it can act in a number of different ways to elicit a biological response, including but not limited to activating the receptor (agonist), blocking or reducing the biological response of the receptor (antagonist), or binding to the same receptor as an agonist but producing a pharmacological response opposite to that agonist (inverse agonist).

1.5 COMPUTATIONAL METHODS OF TARGET PREDICTION

1.5.1 LIGAND-BASED TARGET PREDICTION

There are several chemo-informatic approaches to investigating the potential targets of a natural product. These can be broadly divided into three categories, based on information used, into single ligand based, multiple ligand based and ligand-target based. Single-ligand mechanism of action studies include molecular similarity modelling and pharmacophore modelling. Multiple-ligand approaches include machine learning and quantitative structure activity relationship (QSAR) modelling. Target-ligand approaches to mechanism of action studies include proteochemometrics and docking studies. These methods are explored further in Table 1:4.

Table 1:4 *In-silico* methods for NP target identification. Examples of their use to identify targets of NPs are shown in the Notes column.

Method	Notes	Limitations
Molecular Similarity	Molecular fingerprints of an input ligand are compared to those of ligands known to modulate studied target. This method is based on the "similarity principle" where ligands with similar structure are predicted to bind similar targets.[60]	• Does not take into account any information about the activities of known modulators of the target • Molecule descriptors describe different classes of compounds differently • Compounds with similar structure may have different activity (activity cliff) [61]
Pharmacophore Modelling	Pharmacophores are the essential molecular features of ligands that are responsible for biological or pharmacological interactions. Pharmacophore models are built based on pharmacophore features of ligands known to bind a studied target. The method is discussed in detail in [62] This method was used to identify the NP solanidine as a potent sigma-1 receptor ligand [63]. This method was also used by to identify the targets of secondary metabolites from *Ruta graveolens*.[64]	• Dependent on pre-computed conformation database • Absence of good scoring metrics • No clear way to construct a pharmacophore query i.e. different pharmacophores can be created for similar targets
QSAR	This method involves calculating the correlation between properties of the ligand, e.g. physicochemical properties and the experimentally validated biological activity of the various studied drug targets. The method is discussed in detail in [65] This method has been used to identify the BH3-binding groove of the Bcl-2 protein to be the target of lonchocarpin, a chalcone isolated from *Pongamia pinnata*.[66]. 3d-QSAR was used to identify the glycosomal GDPH of *Trypanosoma cruzi* as the target of the coumarin, chalepin from Rutaceae species.[67]	• Success depends on selected molecular descriptors • Model must be trained on a dataset with enough activity data to be able to extract patterns
Proteo-chemometric Modelling (PCM)	PCM uses statistical modelling techniques to model ligand–target interactions. In PCM the ligand-descriptor matrix used for training the model in QSAR is extended to include protein descriptors[68]. There are currently no known studies using this method to identify targets of natural products.	• Success depends on selected molecular descriptors • Model must be trained on a dataset with enough activity data to be able to extract patterns
Docking	This method uses scoring functions to rank predictions of the predominant binding mode of a ligand with a target (protein) of known three-dimensional structure[69].	• The 3D structure of the target must be known • Scoring functions are not uniform

1.5.2 MACHINE LEARNING IN LIGAND-BASED TARGET PREDICTION

Machine learning approaches are currently one of the most important in computer-aided drug discovery[70]. Machine learning techniques use pattern recognition algorithms to detect mathematical relationships between empirical observations of small molecules[71]. The relationships are extrapolated to predict chemical, biological and physical properties of new compounds. Machine learning is also used to understand and exploit the relationship between chemical structures and their bioactivities[72], which is what we aim to do in this thesis.

Machine learning techniques can broadly be classified into two categories: supervised techniques, in which labels are assigned to training data and, after training, the model predicts labels for the input data; and unsupervised techniques, which involve learning patterns of molecular features directly from unlabelled data[73].

The main types of supervised machine-learning techniques used in target identification along with their advantages and disadvantages are shown in Table 1:5. In this work, a supervised machine-learning model (Random Forest) is used.

Table 1:5 Supervised machine-learning techniques of target prediction and their advantages and limitations

Method	Description	Advantages	Limitations
Support Vector Machines	Supervised learning model. SVM builds a model that assigns new samples to one of two predefined categories. Compound libraries where compounds are represented as descriptor vectors are projected onto a high dimensional feature space via kernels that involved polynomial, sigmoid or radial basis functions. Ideally the samples (compounds) become linearly separable by a hyperplane that maximises the distance between the two classes i.e. maximises the distance between the closest samples. If the two classes are inseparable, a soft margin is applied to maximise the distance between the two classes in a way that minimises the number of misclassified samples[74].	Has regulation parameters to avoid over-fitting Possible to include expert knowledge by modifying the kernel No local minima as it is defined by a convex optimisation problem	High algorithmic complexity and extensive memory use needed for the quadratic programming Selection of kernel function and associated parameters
Decision Tree	Rule based tool that allows the association of descriptor values with the activity under investigation. During splitting, the more important rules are at the root node of the tree.	Easy to understand and interpret Allows addition of new data	High variance Small changes in data may lead to different series of splits and thus different interpretations. The effect of a single error at the top of the tree travels down the whole tree

Method	Description	Advantages	Limitations
			Small datasets affect the learning process Large datasets carry the risk of overfitting
Naïve Bayes Classifier	This is a conditional probability classifier that calculates the probability that a compound x with features F_1, F_2,…F_n will fall in class w_m. This is calculated based on what the model has learned from the training set. It operates on the assumption that all the features are conditionally independent from each other (hence the term Naïve) given the class label. Bayes' theorem is used to predict the class of a novel instance by assigning it to the class with the highest probability[75].	Versatile, robust and easy to use	Not applicable in cases where there are strong conditional dependencies between the variables
K-nearest neighbours	Projects samples represented by features e.g. fingerprints into a feature space and predicts their class, property or rank. The new molecule is assigned to the class most common to its neighbours[76].	Sensitive to the local structure of the data Intuitive	Sensitive to noisy data Predicted value cannot be lower than or greater than the minimum or maximum in the training dataset Irrelevant descriptors risk giving false predictions
Random Forest	Random Forests identify new class labels for unlabelled instances by using an ensemble approach that uses a number of classifiers that	Compared to Decision trees, there is a reduction of overfitting due to averaging of multiple trees	Poor performance of unbalanced data

Method	Description	Advantages	Limitations
	work together [77]. It uses un-pruned classification or regression trees created by using bootstrap samples (with replacement) from the training data and random feature selection in tree induction. The final prediction is made by aggregating the predictions of the tree ensemble[78].	Less variance due to use of multiple trees	Lack of interpretable model i.e. for a given data point and prediction it is not possible to determine which variable (or combination of variables) explain this specific prediction.
Artificial Neural Networks	These are general, flexible, non-linear regression models. They are made up of connected layers of units known as neurons arranged in a specific topology that are connected to each other. The first layer (input layer) sends information to the second layer (hidden layer – of which there may be many) and this information is eventually passed onto the output neurons in the final layer. These artificial neurons have a weight that is adjusted during the learning process. The weights assigned increase or decrease the strength of the final signal[79].	Learns from observing datasets Uses samples of the data (not the whole input data) to arrive at solutions, thus saving time and money	"Black box" nature means that it is hard to understand how the model is solving the problem It is not possible to decipher the idiosyncrasies in the dataset that the model may over-fit Typically involves a very large number of parameters Extensive hardware requirements Needs a large dataset for training

Numerous target prediction approaches have been published, e.g. SEA (Similarity Ensemble Approach)[80] and PASS (Prediction of Activity Spectra for Substances)[81], that predict biological targets of a query ligand, including natural products. These methods rely on training models on bioactivity information of the ligands obtained from databases. Some of the open source databases include PubChem[82], ChEMBL[83] and WOMBAT[84].

SEA is a molecular similarity method that quantitatively relates proteins to one another based on the chemical similarity of their bound ligands[85]. This method was used to predict the anti-malarial activity of the physalins B, D, F and G (isolated from *Physalis angulata*), with B, F and G subsequently showing IC_{50} values of 2.8μm, 2.2μm and 6.7μm (respectively) against *Plasmodium falciparum*[86].

Another tool that allows targets to be predicted based on the Tanimoto similarity of ligands to ligands associated with the target is TargetHunter[87]. TargetHunter uses the Targets Associated with its MOst Similar Counterparts (TAMOSIC) algorithm to predict the biological targets of query ligands. When a query compound is input into TargetHunter, the TAMOSIC algorithm generates the fingerprints of the compound (chosen by the user, any of ECFP6, ECFP4 and ECFP2) and compares the Tanimoto similarity of this compound to compounds in a chemogenomics database, ChEMBL-11. Targets with the most similar compounds to the query compound are output as the predicted targets and ranked according to the similarity scores of the ligands to the input compound. TAMOSIC was trained on 117,535 unique compounds from ChEMBL and 794 targets. TargetHunter obtained 91.1% prediction accuracy of the top 3 targets, i.e. 91.1% of the compounds are assigned to their known targets in the top 3 predictions. TargetHunter was used to identify the mechanism of action of compound CID46907796 from the PubChem database. This compound was reported in PubChem to display cellular apoptosis with AC_{50} values of 0.4136 and 4.908 μM, but the mechanism of action of this compound was not known. TargetHunter predicted the nuclear factor erythroid 2-related factor 2 (Nrf2), which is known to have anti-apoptotic activity[88] as a likely target due to the Tanimoto similarity score of 0.78 and 0.63 to compounds in the dataset.

A machine learning method, PASS, predicts the biological activity of a compound based on its structure[81]. The principle underlying this method is that biological activity equates to structure. This method has been used to predict the anti-oxidant and anti-microbial activity of the acetogenin alkaloid, neoannonin, from an extract of *Annona reticulata*[89]. It has also been used by Goel *et al*[90] to predict the mechanism of action of natural products in five plants. The authors generated a prediction coefficient, P, which calculates the number of activities predicted by PASS over the number of reported activities for the compound. An average prediction co-efficient of 0.66 for the five compounds led to the conclusion by the authors that PASS can be applied to predicting the MOAs of natural products.

Another machine learning target prediction method, used by Nidhi *et al*[91], uses a Laplacian-modified Bayes model to predict biological targets of compounds in the MDDR Database. For *every* target class in the database, the authors built a Laplacian-modified Naïve Bayesian model. A query ligand is passed through each Laplacian-modified Naïve Bayesian model of each target class. The relative estimator score for each of the target classes is calculated. The most probable predicted target for that query compound is the target with the highest score. The model was trained on 103,735 compounds annotated to 964 target classes from World of Molecular BioAcTivity (WOMBAT)[84]. It was used to predict the top 3 most likely targets for compounds from the MDL Drug Database Report (MDDR) dataset[92]. The model predicted 77% correct targets for compounds from 10 target classes in MDDR.

Self-organising maps (SOMs) have been used to predict ligands of targets as well as selective ligands of targets. Self-organising maps are a type of artificial neural network developed by Kohonen in 1982[93]. SOMs learn through unsupervised learning. They are essentially a clustering technique used to visualise similarity in the data where geometric similarity between nodes indicates similarity. This method of SOM was used by Schneider *et al*[94] to correctly predict prosanoid E receptor 3 as a target for the anti-cancer natural product Doliculide. The authors trained the SOM model, on COBRA data, which contains 4,236 drugs and drug candidates[95]. They computed the p-values based on the background distribution of known ligands and drugs to rank the predicted targets[96].

In 2017, Huang *et al*[97] proposed the MOst-Similar ligand Target (MOST) based approach to predict targets. This method incorporates the explicit bioactivity of the most similar ligand. The method is also able to remove false positive predictions due to incorporating p-values associated with explicit bioactivity information as an index. The method involved training a combination of different machine learning algorithms including Naïve Bayes, Logistic Regression and Random Forest using compounds characterised by ECFP-4 Morgan-like fingerprints and FP-2 fingerprints. The model was trained on 61,937 compounds annotated with 173 targets from ChEMBL-19 and validated used 7-fold cross-validation. The dataset comprised 91.3% active compounds and 8.7% inactive compounds. The algorithm worked by calculating the Tanimoto similarity between the input compounds and annotated ligands of the targets. The Tanimoto similarities were then ranked and the most similar ligand was chosen. The Tanimoto similarity and pK_i (-log dissociation constant) of the most similar ligand were fed into the training model to generate the probability of how likely the input compound is to be inactive. If the probability of being active is greater than the probability of being inactive, then the query compound is classified as active and vice-versa. MOST was able to identify the mechanism of action of aloe-emodin by predicting acetylincholinesterase as the target, which was validated *in vitro*. MOST was also able to predict novel targets for the drug Fluanisone (not in ChEMBL), where MOST correctly predicted adrenoreceptor alpha 1B and adrenoreceptor alpha 1D as the second and third most likely targets. These targets were validated by literature to be human targets of Fluanisone.

Deep Learning (DL) is a class of machine learning that uses artificial neural networks (ANN) with a hierarchy of multiple layers whereby each layer transforms input data into more abstract representations[98]. They contain more layers and more nodes per layer than an ANN. The architecture of a DNN consists of many layers, each layer formed of a row of neurons. Neurons (or nodes) in each of the layers can either be fully connected or partially connected. Each successive layer of the DNN uses the output from the previous layer as input. A basic layer of a DNN consists of an:

(i) input layer, which receives large volumes of data (features) as input in different formats, e.g. target descriptors or drug descriptors.

(ii) hidden layer(s), which uses several activation functions e.g. rectified linear unit (ReLU), Sigmoid, Step function etc. to calculate weighted sums of input and add a bias to each layer. The output of the computation determines which nodes to fire. The predicted output is compared with the actual output and the difference in the output i.e. the error is back-propagated through the network and weights are adjusted accordingly. Error in the network is calculated using a cost function

(iii) output layer which generates the desired output i.e. predictions.

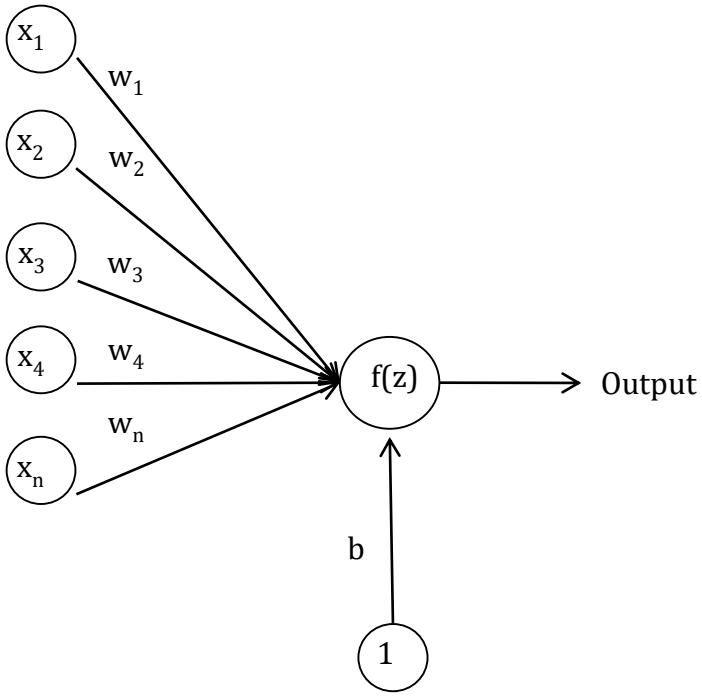


Figure 1:3 Architecture of a single neuron

Each input arrow is associated with a weight w_i. The neuron is also associated with a function $f(z)$ called the activation function and a default bias b. When a vector of features $\mathbf{X}=[x_1, x_2 \ldots x_N]^T$ is fed into a neuron, the output can be represented as shown in Equation 1:

$$O = f\left(\sum_{i=1}^{N} w_i x_i + b\right)$$

Equation 1 – Neuron Output

x_i refers to the input features, w_i, is the weight of input neuron and f is the cost function.

The number of layers, number of neurons in each layer and activation function in each neuron need to be pre-specified when training.

There are two main issues faced when training a DNN. The first involves vanishing gradients[99], exploding gradients and oscillating gradients[100]. Vanishing gradients occur when small weights are updated during training with repeated multiplication of values less than 1 leading to cost functions in lower layers approaching zero. This issue is mitigated by either (i) layer-by-layer unsupervised pre-training of the DNN [101], or (ii) using benchmark guidelines for choosing hyper parameters[102].

The second issue that occurs in DNN is "over-fitting" of the model. This can be overcome by "early termination"[103]. This process involves separating the dataset into a training set and a test set, followed by further separating the training set into a training set and an external validation set. The DNN is optimised on the training set while simultaneously tested on the validation set. As soon as the accuracy drops on the validation set, the training is terminated. " Dropout"[104] is another method that is used to mitigate "over-fitting" used. In "Dropout" randomly selected neurons in each layer vanish (by setting the activation function output=0) at each training round. This leads to the nodes competing to learn a general feature independently as they cannot rely on the presence of other neurons.

DNNs have been successfully used to predict compound target interaction and activity prediction[105, 106]. A study by Dahl *et al*[102] which applied DNN on the Merck Kaggle challenge dataset showed that DNNs can handle thousands of descriptors without prior feature selection. They were also able to optimise the performance of the DNN by optimising the hyper-parameters of the model (number of layers, number of nodes per layer and activation function used). In the study DNNs outperformed RFs in 13 out of 15 datasets. Other studies have also shown that DNNs outperform commonly used machine learning methods like RF, SVM and naïve Bayes regression models[107, 108]. Several studies have shown that multi-task DNN outperform single-task DNN in activity and toxicity predictions[109, 110]. Despite the successes of DNN in activity predictions, it will not be the method of choice in this work. This is because DNN require a large amount of data for training[111] and ANP databases in this study only have

between 364-3177 compounds, This is in addition to model training and validation being computationally expensive and hyper-parameter optimisation is subjective[111].

ECFP4 fingerprints used in the previously mentioned studies do not capture many of the important features required for NP activity, including stereochemistry, repeat units, etc. This is important because stereoisomers can have very different activities as exemplified by the stereomer alkaloids quinidine (antiarrhythmic) and quinine (anti-malarial). Another case of difference in enantiomer activity can be exemplified by cocaine. The naturally occurring (1R,2R,3S,5S)-(-)-cocaine is psychoactive whereas its enantiomer is inactive[112]. This is also the case with the atropine enantiomers: R-(+)-hyoscyamine is a fairly potent analgesic whereas S-(-)-hyoscyamine is completely devoid of such activity[113]. Nevertheless, it can be seen from the above-mentioned implementations that *in silico* target prediction has been successful in identifying the protein targets of ligands including natural products, without need for information about the target.

In this work ECFP4 fingerprints were used as the molecular descriptors to transform the chemical information of the compounds. ECFP4 fingerprints capture molecular features relative to their activity, which is useful for gaining information about activity [114]. The steps to generate ECFP4 fingerprints are:

1. Each atom in the molecule is assigned an integer identifier these might be e.g. atomic numbers of the atoms. These identifiers are collected into an initial fingerprint set.
2. Each atom identifier is iteratively updated to reflect the neighbouring atom identifiers. This includes updates identifying structural duplicates of existing features. Identifiers of the initial atom and its neighbours are collected into an array and a hash function is applied to this array to get a new single integer identifier. (ECFP4 hashing generates identifiers that are comparable across molecules). This occurs for all the atoms in the molecule. All the old identifiers are replaced by the new identifiers, which are subsequently added to the fingerprint set. In our case, this iteration is repeated four times (hence ECFP4)

3. Upon completion of the four iterations, any duplicate identifiers (multiple occurrences of the same feature) are removed. The final ECFP-4 fingerprint is comprised of the remaining integer identifiers in the fingerprint set.

ECFPs have been used for virtual screening [115], SAR modelling [116] and compound library analysis [117, 118].

1.5.3 DATABASES USED TO TRAIN LIGAND-BASED TARGET PREDICTION ALGORITHMS

Several databases are available to train ligand-based target prediction models. These include PubChem[119], ChEMBL[83], DrugBank[120] and WOMBAT[84]. These databases contain bioactivity data including phenotypic readouts and toxicity readouts as well structure information of compounds and drugs. PubChem contains 2,570,179 tested compounds, with information on 10,857 protein targets and 22,106 gene targets. ChEMBL-23 contains 1,735,442 unique compounds annotated to 11,538 targets and 14,675,320 bioactivities. DrugBank 5.0 contains information from 11,037 drug entries, 2,524 approved small molecule drugs with 4,913 proteins annotated to them. WOMBAT 2006.1 contains 136,091 unique SMILES and 1,320 unique targets annotated to 307,700 activites.

It is important to note however that the accuracy of predictions of a model is only as good as the training set, i.e. the accuracy of the model beyond the training set (outside the applicability domain) cannot be guaranteed. Several limitations have been highlighted by Kalliokoski *et al*[121] and these are:

1. Chemical structure related errors;
2. Transcription errors;
3. Inaccurate and insufficient target annotations;
4. Ineffective and incomplete archiving of original data;
5. Redundancy
6. Different lab conditions with different protocols for measurement.

1.5.4 APPLICABILITY DOMAIN OF TARGET PREDICTION MODELS TO NP CHEMISTRY

When extrapolating the predictions of a QSAR model to compounds outside of the training set, it is possible to get good predictions for compounds that are relatively similar to the training set[122]. Predictions may fail for those compounds that are very different to those in the training set[122]. This concept is known as the applicability domain of a model and is usually defined using the similarity of molecular structures or a similarity measure based on descriptors of the compounds[123]. Several chemoinformatic analyses comparing the properties of different sets of natural products and synthetic compounds[124-126] found that natural products differ considerably from synthetic compounds in several molecular properties. They found that natural products on average tend to have: more oxygen atoms, fewer nitrogen atoms, more stereogenic centres, more fused rings, but fewer aromatic rings and fewer rotatable bonds. These models also use fingerprint similarity to compare training sets to the test compounds. This is not the best representation since fingerprints lead to a loss of atom order as well as the fact that they do not capture many aspects important for NP activity, e.g. repeat units, stereochemistry etc as mentioned above. This leads to one of the limitations of utilising ligand-based target prediction trained on ChEMBL, WOMBAT etc. e.g. PIDGINv2. Natural products generally do not share the same chemical space as the training space of the algorithm. In PIDGIN (the model used in this thesis), the models perform better overall, with up to 96% probability scores achieving 0.98 (maximum of 1) true positive prediction when the similarity between the test and training sets are higher than 0.3[127]. To address this, in this work prediction results for natural products were filtered for compounds that fall below a specified (Tc = 0.3) similarity threshold to their nearest neighbour in the training set.

1.6 PATHWAY ANNOTATIONS

When attempting to understand mechanism of action of compounds, studying individual target/gene information does not give insight into the underlying mechanistic action. In this work we look at biological pathways. These consist of genes, proteins and small molecules interacting with each other in a cellular setting to elicit cellular change or creating products. The most well-known types of biological pathways are

signalling pathways, which move a signal from outside a cell to its interior; metabolic pathways, responsible for chemical reactions within the body; and gene-regulatory pathways which control the activation and deactivation of genes. We use pathway annotations to put the isolated targets into biological context[128]. This is carried out by combining information from databases with statistical testing[128]. This facilitates both the interpretation of isolated target information and generation of a hypothesis of mode of action. Pathway annotation can be used to identify the biological roles of candidate genes[129] (in this work: predicted targets). It has been applied to predicted targets of NPs and it improved the mechanistic understanding of the mechanism of action of the studied NPs[130]. Pathway annotation has also been used to identify important targets to be modulated in order to elicit a required response, e.g. stop a particular function or inhibit a particular mechanism[131].

Several databases are available that provide signalling, metabolic and gene-regulatory pathway annotations when provided with gene lists, e.g. Reactome[132], KEGG[133, 134], Gene Ontology (GO)[135, 136], PANTHER[137] and Comparative Toxicogenomics Database (CTD)[138]. In this work we use Gene Ontology (GO) and WikiPathways[139].

The Gene Ontology (GO)[135, 136] is a large resource that provides a standardised, structured and controlled vocabulary of terms for both gene and gene product functions across all species. It operates based on the observation that similar genes will often display conserved functions in different organisms[140]. The vocabulary of terms is divided into three non-overlapping ontologies, namely, Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). It associates each gene with the most specific set of terms that describes its functionality[140]. In this work we use the Biological Process ontology in an attempt to understand the mechanism of action of phenotypically active anti-trypanosome compounds by analysing the biological processes they are predicted to modulate.

WikiPathways is a database of biological pathways, which is maintained by the scientific community[141]. In this work we use WikiPathways due to its clear interpretability[142]. In this work it was accessed by PIDGIN[143], via the NCBI BioSystems Database[144].

1.7 PHYLOGENETIC AND ETHNOBOTANICAL BIO-PROSPECTING

One method of predicting the activity of medicinal plants used in traditional medicine is by incorporating *a priori* information of the phylogenetics of the medicinal plant. The rationale behind this is that evolutionarily similar plants will produce NPs with similar structure (as measured by the Tanimoto similarity of their ECFP4 fingerprints) and therefore (based on the similarity hypothesis) they will have similar activity profiles. In previous studies relating phylogeny of plants to their ethnobotanical use, Hawkins, J. A. *et al*[145] found that in the genus Pterocarpus (Leguminosease) related species were used for similar ethnobotanical uses in different geographical areas. This was demonstrated by studying plants from the genus Pterocarpus across Indomalaya, Tropical Africa and the Neotropics. Plants with medicinal activity were concentrated on specific clades, i.e. they were not randomly distributed. This provided a link between biogeography and phylogeny of the plant.

Some studies[146, 147] have also looked at the possibility of predicting medicinal potential of a plant using its phylogeny. The study by Hawkins *et al*[147] found that phylogenetic patterns were shared among the medicinal plant species of the flora of Nepal, New Zealand and Cape of South Africa. It was observed that for the geographic areas under study, traditional medicine use is not randomly distributed on the phylogenetic tree, but rather concentrated on specific nodes. Hot nodes comprised on average 133% more medicinal plants than a random sample of the studied floras. They found that on average "hot nodes" from one region contain 17% more medicinal plants from another region than would be expected by chance. Furthermore, on average, the "hot nodes" from one region contain 38% more disease specific medicinal plants from another region.

In a study by Duez *et al*[148] an ethnobotanical study of plants used by Burundian traditional healers to treat microbial disease was carried out. They attempted to compare the plants used to the plant distribution in the area of interest (Burundi). They found that 155 plant species, distributed in 51 families and 139 genera, were used, with the most common families being Asteraceae, Fabaceae, Lamiaceae, Rubiaceae, Solanaceae and Euphorbiaceae.

To our knowledge, none of the phylogenetic studies to date have incorporated or integrated the chemical information of the NPs produced in a plant to aid prediction of activity, which is what we are aiming to do in this thesis.

1.7.1 PHYLOGENETIC TREES

To study the evolutionary relationship between the medicinal plants in our datasets, and relate this to their activity, we used phylogenetic trees. Phylogenetics is defined as the study of the evolutionary history and relationships among individuals or a group of organisms[149]. These relationships are not observed but inferred through two steps: (i) identifying homologous characters, e.g. amino acid sequences, nucleotide sequences, biochemical pathways or any other homologous character and (ii) reconstructing the evolutionary history of the individuals using either distance-based or character-based methods.

In distance-based methods, the distance between every pair of sequences is calculated to produce a matrix and a cluster algorithm, e.g. neighbour joining (NJ)[150], minimum evolution (ME)[151, 152], or least squares method[153] is applied to this matrix to produce the resolved phylogeny: see Table 1:6. This relies on the assumption that all sequences are homologous.

Table 1:6 Cluster algorithms used to produce resolved phylogenies in disatance-based methods.

Cluster algorithm	Notes
Neighbour joining (NJ)	This algorithm starts with a star tree. Pairs of taxa are chosen based on their distances and successively joined together. This is repeated until a fully resolved tree is obtained. The pairs of taxa to be joined are chosen in order to minimise an estimate of tree length. The distance matrix is updated with the new joined taxa (represented by their ancestor) replacing the two original taxa.
Least squares (LS)	This method involves minimising the measure of differences between the calculated distances in the distance matrix and the expected distances in the tree. A score Q is given to the optimised branch lengths liking two species for a given tree. The least square estimate for the true tree is the tree with the smallest Q score.
Minimum evolution (ME)	This is similar to LS but uses the sum of branch lengths for tree selection. Shorter trees are more likely to be correct in ME than longer ones.

In character-based methods all sequences in the alignment are compared simultaneously, and one site of the alignment is considered at a time to calculate a score for the tree. Maximum parsimony[154, 155] assigns character states to nodes on the tree to minimise the number of changes on a phylogenetic tree. The maximum parsimony tree is the tree that minimises the tree score, which is the sum of character lengths (the minimum number of changes required for a particular site) over all sites. In brief, the maximum likelihood method[156] determines the topology of a tree, its branch lengths and the parameters of the evolutionary model that maximise the probability of observing the given homologous characters, e.g. nucleoside sequences. Bayesian inference[157-160] involves combining the prior probability of a phylogeny with the tree likelihood of the data to produce a posterior probability distribution on trees. The tree that best represents the true phylogenetic tree is the one with the highest posterior probability. The tree score for maximum parsimony[154, 155], maximum likelihood[156] and Bayesian inference[157-160] are "minimum number of changes", "likelihood value" and "posterior probability" respectively[161]. The advantages and limitations of these methods are shown in Table 1:7.

Table 1:7 Methods used for phylogenetic inference with their advantages and limitations. (Adapted from [161])

Method	Advantages	Limitations
Distance-based methods	Fast computational speed Can be applied to all data types Possible to choose models for distance calculation to fit available data	Variance of distance estimates not considered Divergent sequences and those with many alignment gaps are problematic Negative branch lengths hold no information
Maximum parsimony	Simple and intuitive	Poorly understood and implicit assumptions Knowledge of sequence evolution cannot be incorporated High substitution rates lead to underestimated branch lengths May suffer from long-branch attraction
Maximum likelihood	Biological reality can be represented using complete substitution models	Iteration is computationally expensive
Bayesian inference	Biological reality can be represented using complete substitution models Expert knowledge can be incorporated into model via prior probability Easy to interpret posterior probabilities of trees and clades	Computationally expensive Difficult to identify and rectify Markov Chain Monte Carlo convergence

In this work, we use a tree generated by Zanne et al[162], constructed using the maximum-likelihood method, to calculate the patristic distance between medicinal plant families and relate this information to their traditional use as well as the predicted activity and structural similarity of the NPs they contain.

1.8 AIMS OF THE THESIS

In this thesis several aims are explored. The first aim, explored in Chapter 2, is to explore the chemical space and biological space of African NPs used against cancer, and how this differs from the chemical space and biological space of other NPs used against cancer as well as cancer drugs in the market. In this chapter we utilise machine learning to understand the mechanism of action of NPs in an attempt to rationalise their ethno-botanical use.

In Chapter 3, methods borrowed from the ecology community are used to determine phylogenetic patterns of medicinal plant use in the African continent. We will examine whether plants closer together on the phylogenetic tree produce similar compounds that are predicted to act on similar targets and vice versa; i.e. where plants that are further away phylogenetically synthesize chemically diverse NPs with different predicted targets. Ultimately, this information is important to determine whether phylogeny, along with ligand-based target prediction and knowledge of ethno-botanic use, can be integrated to predict the activity of African NPs.

In Chapter 4 we attempt to understand the molecular mode of action of small molecules and NPs active against *Trypanosoma brucei*, the causative parasite of Human African Trypanosomiasis (HAT). We also identify which compounds are predicted to cross the BBB. Compounds with the ability to traverse the BBB are important because they are essential for the treatment of Stage 2 HAT, for which there are currently only 2 drugs, used in combination, in the market. Thirdly we explore the biological processes enriched in the predicted gene sets to better understand how these compounds exert their phenotypic activity.

CHAPTER 2: UTILISING TARGET AND PATHWAY PREDICTIONS TO SUGGEST MECHANISMS OF ACTION OF AFRICAN NATURAL PRODUCTS

2.1 INTRODUCTION

The WHO encourages African member states to incorporate traditional medicines into their health care systems[163]. In order to implement this, it is important to understand the efficacy, safety and mechanism of action of the natural products (NPs) in these traditional medicines.

In this work we will use computational approaches to attempt to understand the mechanism of action of African natural products that have recently been curated into databases, in particular, ConMedNP and AfroCancer. In our approach to understanding the mechanism of action we use ligand based machine learning, taking advantage of the "similarity principle", which says that similar molecules having similar physicochemical properties [164] will have similar biological activity"[61, 164-166].

Traditional medicines provide us with new starting points for drug discovery in both chemical and target space[17, 167], and the chemistry of African traditional medicines has not yet been explored in much detail. Scaffold diversity analysis of NP datasets has been carried out[168],whilst the diversity of the most frequent scaffolds of African NPs have not been previously analysed. To investigate the chemical space covered by traditional African medicines, we studied the scaffold diversity of the African compounds in both the datasets and related this to approved drugs in the market. We carried this analysis out to see if African NPs contain unique scaffolds and chemistry that are not found in drugs in the market, which may be exploited for further drug discovery experiments. We also compared the scaffold diversity of African NPs used against cancer to Malay and Ayurveda NPs used against cancer.

Secondly, we investigated the target space, and hence the mechanism of action of African NPs, compared to approved anti-cancer drugs in the market. In this part of the

study, ligand-based target prediction was used to investigate the possible mechanism of action of compounds from the AfroCancer dataset. Ligand-based target prediction involves comparing a novel ligand to a group of ligands that are known to bind to a target, either via similarity or machine learning methods[127]. Several approaches have been published, e.g. SEA (Similarity Ensemble Approach)[80] and PASS (Prediction of Activity Spectra for Substances)[81], and they have been used to rationalise the mechanism-of-action of NPs from Chinese, Malaysian and Indian traditional medicines [169-171] [169], but not yet to African medicines. These approaches and their limitations have been reviewed in[170-173]. It is important to note that predicting targets for natural products is difficult. This is because the training sets are often small synthetic molecules, whereas NPs are generally larger, more complex compounds. One way to overcome this problem is to fragment the natural products into smaller entities and predict their targets by comparing the smaller entities to synthetic drugs with known targets[174]. Another way to do this is to include natural products with experimentally validated results in the training set[175]. In this study, we use a target prediction algorithm that contains natural products in the training dataset. Since the fraction of natural products trained in PIDGIN is small, predictions were only kept for NPs with Tanimoto coefficient (Tc) ≥ 0.3 to compounds in the training set. Jasial *et al*[176] carried out Tc similarity value distributions to determine activity-relevant similarity value ranges. They found that there is a much higher probability that a comparison of active compounds to active compounds yielded a Tc value of at least 0.3 than a comparison of active vs. random or random vs. random compounds i.e. 38.2 % of ECFP-4 Tc values for a comparison of active compounds reached or exceeded 0.3 but only 0.3% of random vs. active reached or exceeded 0.3. This small percentage of 0.3% translates to a large number of false negatives since in reality, when searching a large database there are more random vs. active than active vs. active compounds. For example, when searching through the 1,828,820 compounds in ChEMBL, where for example only 500 compounds are active against a particular target, a cutoff of 0.3 will result in 191 true positive compounds and 549 false positive compounds. A cut-off Tc value of 0.3 is used in our work, as this is the smallest value where the maximum number of true positives and minimum number of false negatives can be obtained for active vs. random compounds.

Lastly, we annotate predicted targets with pathways, with the aim of understanding their effects of modulating particular proteins in the human body. Previous studies have utilised pathway annotations to understand the mechanism of action of TCM and Indian traditional medicines[130, 170] as target predictions alone do not provide information on downstream biological effects. The idea here is that if a NP from medicinal plant "m" is predicted to bind targets "x, y and z", which are involved in pathway "a", then we can infer that medicinal plant "m" mechanistically works by modulating pathway "a". It is important to mention that not all predicted targets will be linked to the mechanism of action of the NP. This method of pathway annotation will enable us to infer the pharmacological action of the natural products (at a pathway level), explain the molecular basis for their ethno-botanical use, and predict new mechanisms of action if they exist.

2.2 MATERIALS AND METHODS

2.2.1 DATASETS

2.2.1.1 DATASETS USED FOR COMPOUNDS IN AFRICAN NPS

The compounds analysed in this study were obtained from ConMedNP[43], which contains 3,177 compounds from 79 plant families comprising 376 species, and AfroCancer[44], which contains 390 compounds from 48 families comprising 102 species. All compounds in the datasets were used. Compounds from ConMedNP are annotated for a wide variety of diseases collected from traditional healers, while the compounds from AfroCancer are focused on cancer indications. Only compound structures were used from each of the datasets in subsequent analyses.

2.2.1.2 DATASETS USED AS BACKGROUND FOR AND COMPARISON TO AFRICAN COMPOUNDS

We next identified databases to be used as a background in order to assess the properties of African compounds with respect to them.

APPROVED DRUG DATASETS

As the first dataset to compare compounds used traditionally in African medicine (from the ConMedNP dataset) all 1,510 approved drugs from DrugBank 4.0[120] were retrieved. These compounds will be, hereafter, referred to as 'Approved DrugBank'. Secondly, a list of approved drugs used for cancer treatment was used as a comparison set (reference dataset) to AfroCancer. This list was obtained by request from the National Cancer Institute (NCI)[177]. The ChEMBL-19[178] database was queried to identify Simple Molecular Input-line Entries (SMILES) strings for each drug from the NCI, and this list of 185 compounds (shown in Supplementary Table 1) will from here onwards be referred to as 'NCI Cancer'. The experimentally validated targets of these drugs were also extracted from ChEMBL. An IC_{50} of 10μm was used for activity against any of the targets annotated in the database to assign whether a drug is active against a particular protein or not.

TRADITIONAL MEDICINES DATASETS

In addition, two traditional medicine datasets were used as background comparisons for the AfroCancer dataset, to determine the similarity of African natural products used against cancer to others used in different regions of the world. Firstly, 1,091 compounds from Malay traditional medicine with reported anti-cancer activity, using the query "cancer" and "tumour", were derived from the commercial database Natural Product Discovery System (NADI[179]), hereafter referred to as 'Malay Cancer' dataset. Furthermore, 1,043 compounds with reported anti-cancer properties from Ayurveda were obtained from Dr. Duke's Phytochemical and Ethnobotanical Databases[180] which will from here onwards be referred to as the 'Ayurveda Cancer' dataset.

2.2.2 STRUCTURAL PREPROCESSING

ChemAxon Standardizer[181] was used for structure canonicalization, transformation, and conversion of compounds from SD format to SMILES. To standardise the compounds in ChemAxon Standardizer, the following options were used: Clean 2D, Mesomerize, Neutralize, Remove Explicit Hydrogen and Remove Fragment. Duplicate structures in each dataset were removed, using ChemAxon JChem Software[181], using the command "remove duplicates". In total, this left us with 185 compounds in NCI Cancer, 1,510 compounds in Approved DrugBank, 1,015 compounds in Malay Cancer and 1,037 compounds in Ayurveda Cancer.

2.2.3 CHEMICAL SPACE ANALYSIS

2.2.3.1 MULTI-DIMENSIONAL SCALING

Multi-Dimensional Scaling (MDS), was used in R[182] using library(rgl) based on modified Tanimoto similarity matrices comprised of MOLPRINT2D fingerprints of the compounds. MOLPRINT_2D[183] fingerprints were generated using Canvas[184-186]. The modified Tanimoto coefficient[187] was used since both the ON and OFF bits are assessed. The Tanimoto coefficient considers only the ON bits and is sensitive to unwanted size dependent effects, i.e. the modified Tanimoto reduces size dependence of the similarity coefficient[187]. It was calculated as follows using Canvas[184]:

$$\text{Modified Tanimoto} = \propto \text{Tanimoto} + (1 - \propto))T0$$

Equation 2 – Modified Tanimoto coefficient

where:

$$\propto \equiv 2/3 - (a + b)/[6.\min(d, 10000) \text{ and,}$$

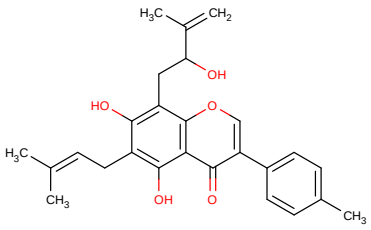
$$T0 \equiv d/(a + b - 2c + d)) = \text{Tanimoto of "off" bit}$$

Let:

- $a \equiv$ Count of "on" bits in bitset1.
- $b \equiv$ Count of "on" bits in bitset2.
- $c \equiv$ Count of bits that are "on" in both bitset1 and bitset2.
- $d \equiv$ Count of bits that are "off" in both bitset1 and bitset2.

2.2.3.2 SCAFFOLD DIVERSITY ANALYSIS

Next, the scaffold diversity was analysed for the three NP datasets related to cancer, namely AfroCancer, Malay Cancer and Ayurveda Cancer, and compared to NCI Cancer. The purpose of this part was to assess the diversity of the AfroCancer dataset in relation to other NPs from plants with anti-cancer activity. In this study the Bermis-Murcko[188] (BM) scaffolds were used to represent the scaffolds of the molecules. These frameworks are defined as "the union of rings plus the linker atoms", as shown in Figure 2:1 Bemis-Murcko scaffolds. ChemAxon JKlustor[181] command-line was used to generate the BM frameworks using the function "bm", then the compounds were sorted into clusters according to their scaffolds and written out in SD format, using the function "cluster_*.sdf". None of the compounds were sorted into more than one cluster.

a. 

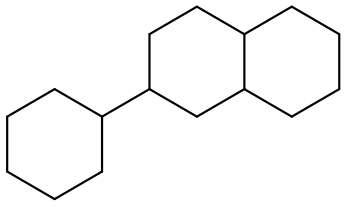
b. 

Figure 2:1 Bemis-Murcko scaffolds The cyclic systems in this study were obtained by removing the side chains from the entire molecule (a), and leaving the linkers between the rings to get the Bermis-Murcko scaffold (b).

The next step involved analysing the diversity of the most frequent scaffolds. The Shannon Entropy measure (SE) was used to quantify the distribution of compounds in the *n* most populated scaffolds[189, 190]. The SE is defined in Equation 3 as:

$$SE = -\sum_{i=1}^{n} p_i log_2 p_i: \qquad p_i = c_i/P$$

Equation 3 – Shannon Entropy

Where:

p_i is the relative frequency of the cyclic systems *i* in a dataset of *P* compounds that contain *n* distinct cyclic systems

c_i is the absolute number of compounds containing a cyclic system *i*

To normalise the results for varying values of *n*, a Scaled Shannon Entropy (SSE)[190] was used, which is defined as in Equation 4:

$$SSE = \frac{SE}{log_2 n}$$

Equation 4 – Scaled Shannon Entropy

For more diverse indications, the scaffold diversity of the African compounds from ConMedNP was also compared to the diversity of the Approved DrugBank database. For this purpose, Murcko scaffolds[188] were used as generated using DataWarrior, version 4.3[191]. The compounds were then sorted into clusters; according to their scaffolds, i.e. all compounds having the same scaffold were placed in the same cluster. The diversity of the entire datasets was studied by analysing the diversity of the most frequent scaffolds by using the Shannon Entropy (SE) measure, in order to quantify the distribution of compounds in the *n* most populated scaffolds[189, 190]. This measure indicates the global diversity of the datasets.

2.2.4 TARGET PREDICTION – PIDGINV2

In this work we make use of a Random Forest algorithm (PIDGINv2) trained by Mervin *et al*[143] using ECFP-4 for SAR modelling. Active compounds were extracted from

ChEMBL21[83] and are those with activity values ($IC_{50}/EC_{50}/K_i/K_d$) of 10 μM or lower, with a confidence score of 5 or greater for 'binding' or 'functional' human protein assays. The active dataset contains 2,089,404 bioactivities across 3,394 proteins. The inactive dataset contains 11,829,475 inactives derived from PubChem. PIDGIN v2 is therefore capable of predicting both the probability of activity and inactivity for orphan compounds against a range of biological targets. The parameters used were: Random Forest with 100 trees, class weight = "balanced", sample weight = ratio inactive:active.

The steps involved in a Random Forest algorithm are[192]:

Ensemble of Trees: A Random Forest is represented by $[T_1(X), ..., T_B(X)]$, with B being the number of trees and $X = [x_1, ..., xv]$ is a v-dimensional vector of molecular descriptors or properties associated with a molecule. The ensemble produces B outputs $[Y\hat{}_1 = T_1(X), ..., Y\hat{}_B)\ T_B(X)]$ where $Y\hat{}_b$, b = 1, ..., B, is the prediction for a molecule by the bth tree. $Y\hat{}$ is the final predicted class as predicted by the majority of the trees.

Training Procedure: For a dataset of n molecules, $D = [(X_1, Y_1), ..., (X_n, Y_n)]$, where X_i, i = 1, ..., n, is a vector of descriptors and Y_i is the class label (e.g., active:inactive) The algorithm functions as follows:

(1) From the training data of n molecules, draw a bootstrap sample (n samples chosen at random, with replacement).

(2) For each bootstrap sample, grow a tree and at each node, select m random variables out of all possible M variables. Select the best split on the selected m variables.

(3) Grow the trees until no further splits are possible or until a maximum depth specified at the start (in this case 100).

Repeat the steps until a specified number of B trees are grown. A simplified schematic of this process is shown in Figure 2:2.

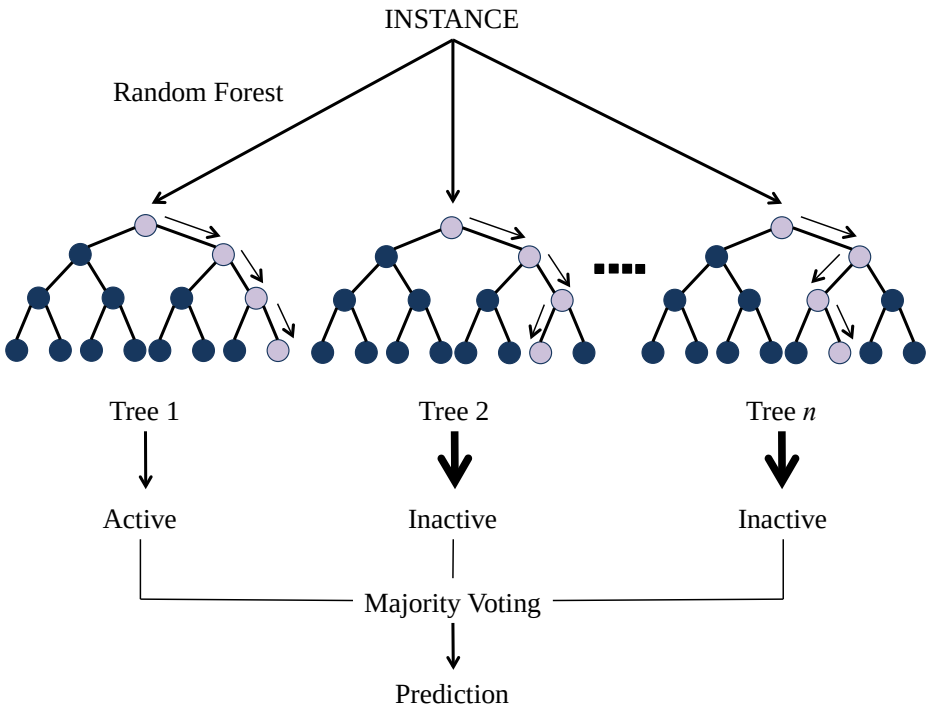


Figure 2:2 Schematic showing a simplified Random Forest model. In the Random Forest algorithm, each new data point goes from the root node to the bottom until it is classified in a leaf node. It visits all the different trees in the ensemble, which are grown using random samples of variables. In this classification model, the function used for aggregation is the mode or most frequent class predicted by the individual trees (also known as a majority vote).

The models were scaled using Platt Scaling[193]. Platt Scaling assigns true positive rate (TPR) values to the predictions by splitting the training set into a calibration and training set, thereby converting the Random Forest predictions into the corresponding TPR for a given threshold. In this work targets were predicted for the input compounds with a cut off of 0.9 (meaning that a 10% false positive rate is accepted, which is a rather stringent value).

To obtain the enriched targets, the list of predicted targets was compared by Mervin *et al*[143] to the predicted targets of a random sample of over 2,000,000 compounds obtained from PubChem. The Fisher's exact test and odds ratio[194, 195] were calculated using the contingency table for both sets[143]. A low odds ratio and p-value indicate a higher enrichment for a target when compared to targets from a random set. In this work the resulting list was filtered for an odd ratio of less than 0.1 and ranked by p-value.

2.2.5 PATHWAY ANNOTATION

The output from the PIDGIN v2[143] pathway prediction results contains pathways predicted from NCBI BioSystems pathways. The WikiPathways annotations in PIDGIN2 were used for pathway annotation due to their interpretability[141, 142].

Figure 2:3 shows a workflow of the work carried out in this chapter.

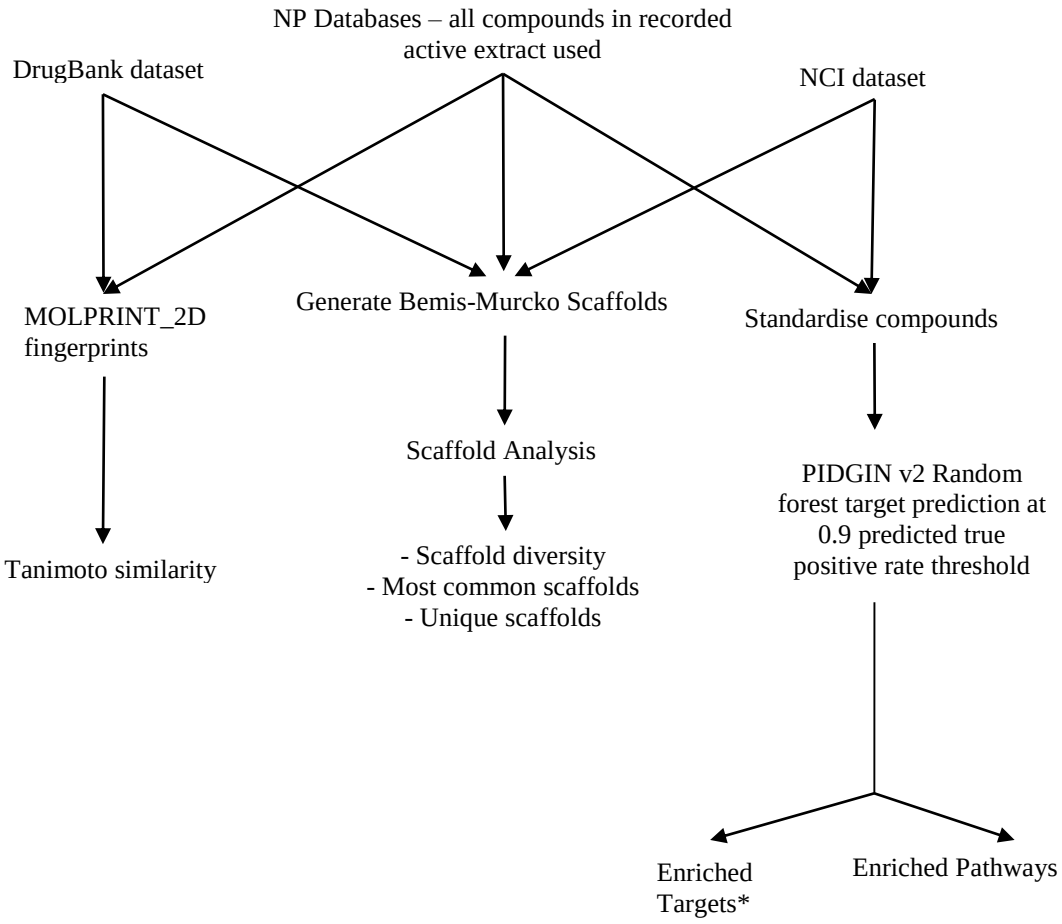


Figure 2:3 Workflow illustrating the work carried out in this study. (i) Chemical space of ANPs in ConMedNP was studied and compared of the chemical space coverage of approved drugs in DrugBank. (ii) Scaffold diversity of ANPs in the AfroCancer and ConMedNP datasets was studied and compared to the scaffold diversity of the approved drugs in DrugBank and approved drugs for cancer (NCI). (iii) Target and pathway prediction of ANPs was carried out to understand their mechanism of action. These predicted targets were compared to the experimentally validated targets of the NCI dataset. Enriched targets were only calculated for the ANPs and these were compared to experimentally validated targets of the NCI dataset. Enriched pathways were calculated for both the AfroCancer and NCI dataset compounds.

2.3 RESULTS AND DISCUSSION

2.3.1 COMPARATIVE CHEMICAL SPACE ANALYSIS

The distribution of the ConMedNP database in chemical space, in comparison with the approved DrugBank compounds, were first compared, using MOLPRINT2D fingerprints and multi-dimensional scaling (MDS, Figure 2:4). It can be seen that ConMedNP compounds dominate two areas, (shown in red), not populated by many compounds from the Approved DrugBank set. The approved drugs that did populate this area were also natural products such as Colchicine. The overlap in space may be due to the fact that marketed drugs (those in Approved DrugBank) might be influenced by natural products. This has been shown by Newman and Cragg[17] in their study of the sources of drugs between 1981-2014. They found that 26% of new drugs are either botanical drugs, unaltered natural product drugs or synthetic drugs derived from natural products. Our results indicate that the African compounds occupy a different chemical space. The space of Approved DrugBank is not as diverse in chemical space due to the fact that compounds are screened for Lipinski's rule violations early on in the drug discovery process. Lipinski's rule does not apply to natural products because they mostly enter the cells via transmembrane transporters and not passive diffusion.[196] In conclusion we have shown that African NPs and drugs occupy a different chemical space.

.

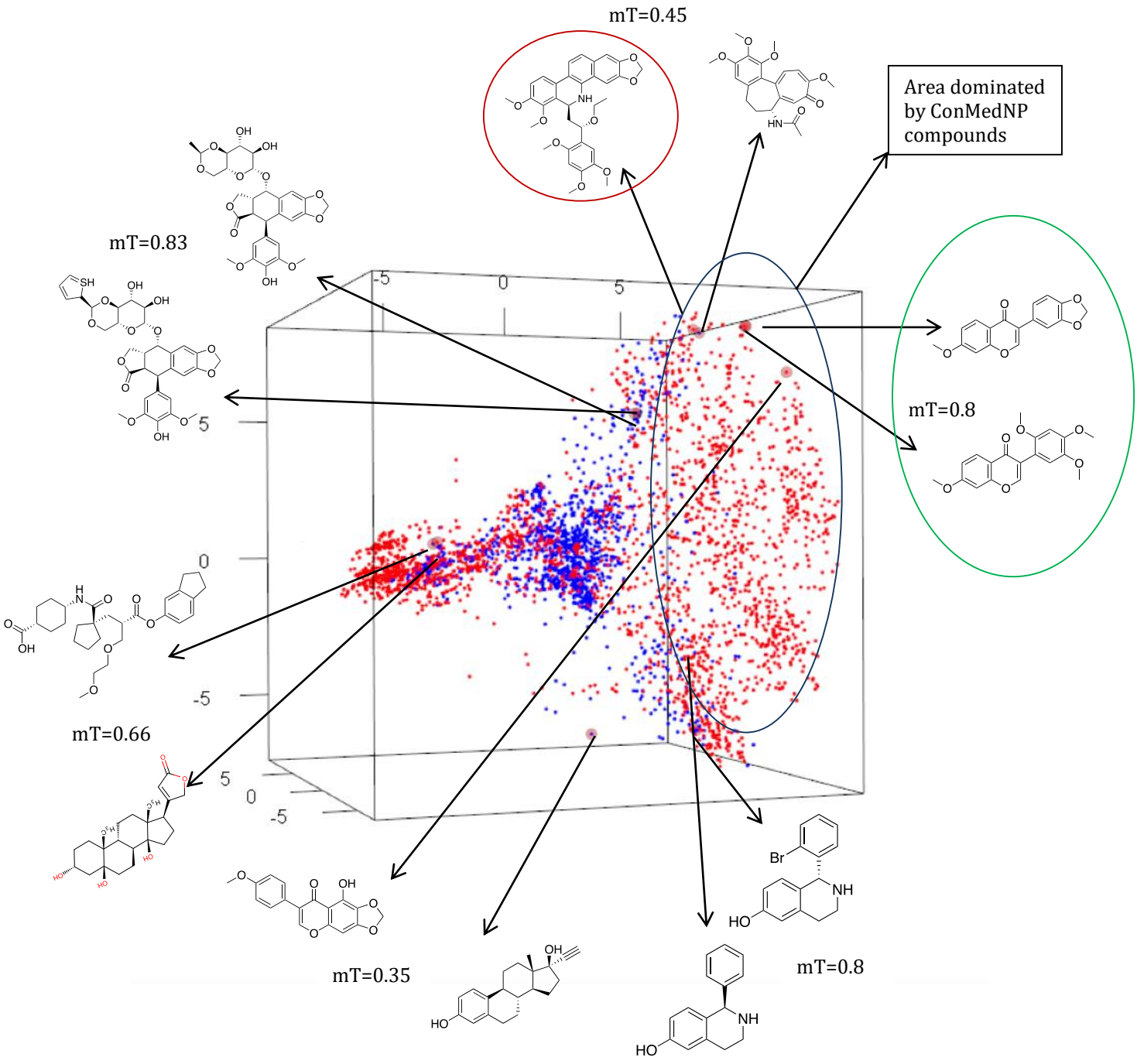


Figure 2:4 MDS of molprint2d fingerprints of compounds of the CONMEDNP (red) and Approved DrugBank (blue). Compounds with unique scaffolds from ConMedNP occupy a different chemical space to those occupied by Approved DrugBank compounds. The modified Tanimoto coefficients shown between the pairs of compounds range from 0.80 to 0.83 for similar compounds and about 0.35 for structurally dissimilar compounds. Examples of bioactive compounds from the ConMedNP dataset are shown in the red and green circles.

The range of molecular weights covered by the compounds in both datasets is also shown in Figure 2:5. It can be seen that there are more compounds in the ConMedNP dataset, with molecular weights raging from 84.16 to 1439.59 with a mean molecular weight of 241.75 Daltons. On the other hand, the compounds in the Approved DrugBank dataset have molecular weights which range from 17.00 to 1449.27 with a mean molecular weight of 354.73 Daltons.

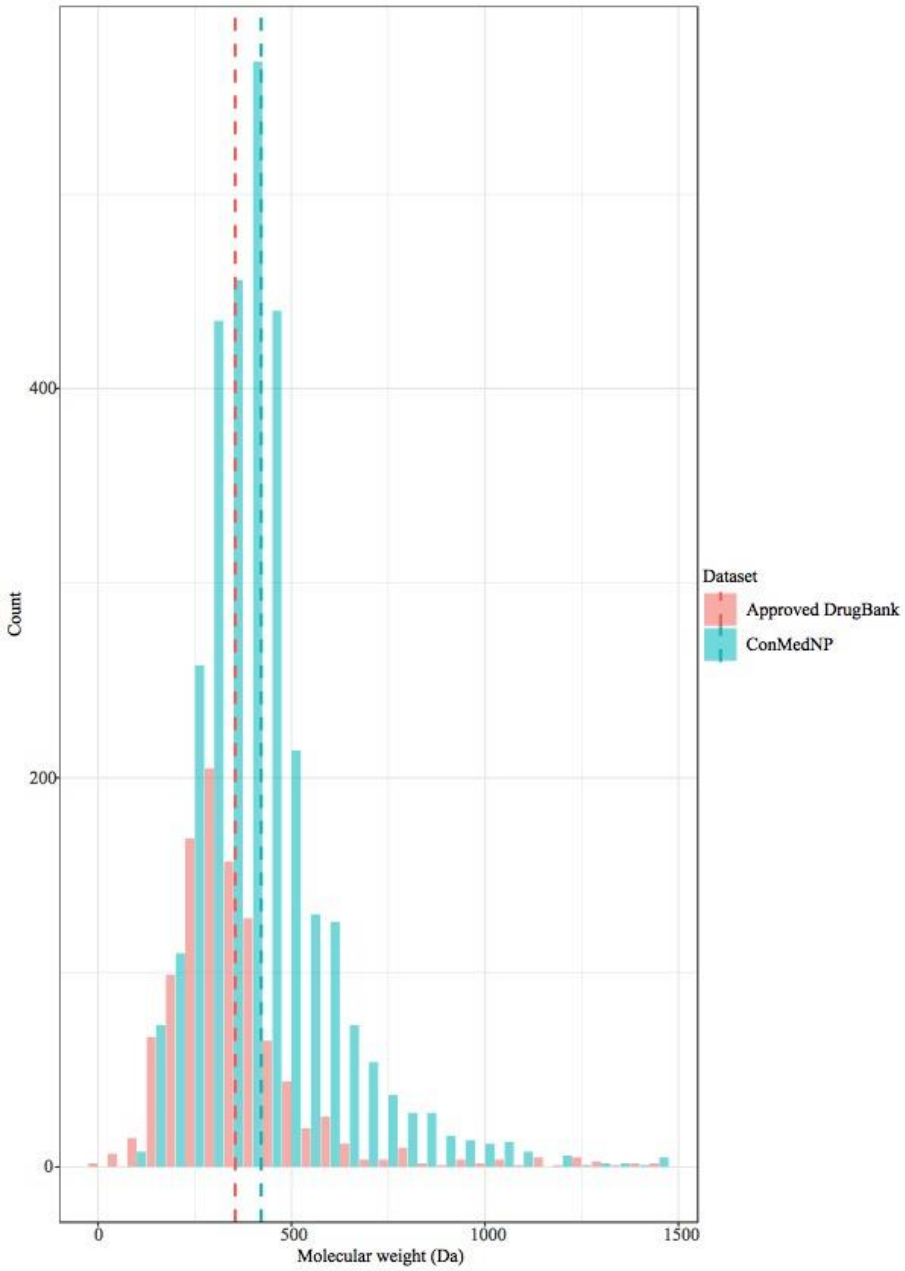


Figure 2:5 Range of molecular weights of compounds in the Approved DrugBank and ConMedNP datasets. Molecular weights for the ConMedNP dataset range from 84.16 to 1439.59 with a mean molecular weight of 241.75 Daltons. Molecular weights for the Approved DrugBank dataset range from 17.00 to 1449.27 with a mean molecular weight of 354.73 Daltons.

To understand the difference in chemical space covered in the different datasets further a scaffold analysis was carried out next.

2.3.2 SCAFFOLD DIVERSITY OF AFRICAN NATURAL COMPOUND DATASETS

Next, the scaffold diversity of African compounds isolated from traditional medicines was quantified. Scaled Shannon Entropy (SSE) was used to analyse the distribution of compounds in the scaffolds of the datasets, with a value of 0 indicating that all compounds are contained in one cyclic system (representing the lowest possible diversity) to 1, indicating that each cyclic system contains an equal number of compounds (representing highest possible diversity). Looking at Table 2:1 the SSE values for the AfroCancer and NCI Cancer dataset (of 0.94 and 0.89 respectively) demonstrate their higher diversity when compared to the other datasets (of 0.73 and 0.69 for Malaya Cancer and Ayurveda Cancer respectively). ConMedNP and Approved DrugBank showed relatively lower (but still high) diversity values of 0.87 and 0.85, respectively. These results are in line with results obtained by Schneider *et al*[197] who found a greater diversity of ring systems in natural product libraries compared to synthetic and combinatorial libraries. Yet this was not the same as results obtained in a previous study[168], where it was found that the diversity of a combinatorial library was higher than that of the natural products they studied. It is surprising that the more focused libraries (on cancer) have a larger diversity on this measure than the more diverse ones with respect to indications, i.e. ConMedP and Approved Drug Bank. This may be because the compounds in the smaller datasets, e.g. AfroCancer and NCI Cancer were synthesised through very different routes, e.g. from plants or *via* organic synthesis to specifically inhibit different targets with different roles in cancer.

Table 2:1 Scaled Shannon Entropy table for the 6 studied datasets NC: number of compounds in the database; NS: number of scaffolds; NS1: number of singletons; NS/NC and NS1/NC: number of scaffolds and number of singletons normalised by the number of compounds, respectively; NS1/NS: number of singletons in relation to the number of scaffolds; SSE5, SSE10, SSE20: scaled Shannon Entropy at 5, 10 and 20 most populated scaffolds, respectively; n5, n10, n20: fraction of compounds contained in the 5, 10 and 20 most populated scaffolds, respectively. It can be seen that the AfroCancer and NCI Cancer dataset are more diverse than the other datasets.

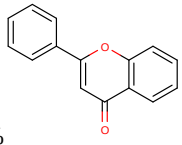
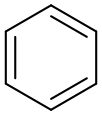
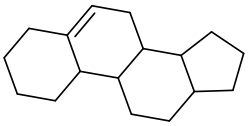
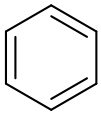
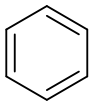
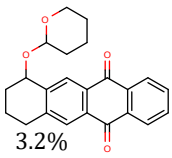
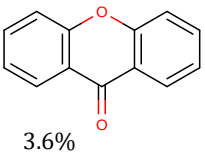
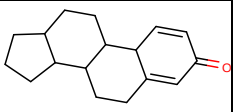
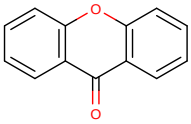
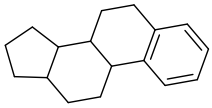
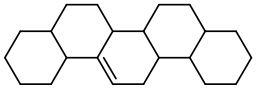
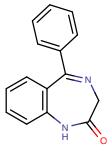
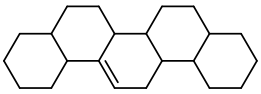
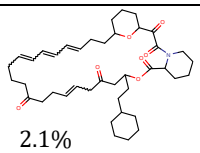
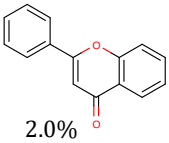
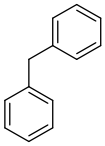
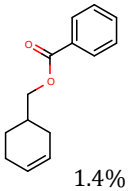
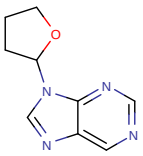
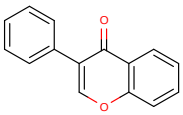
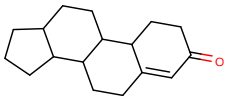
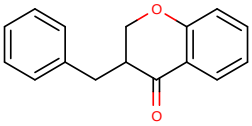
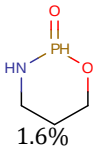
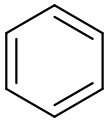
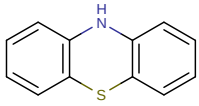
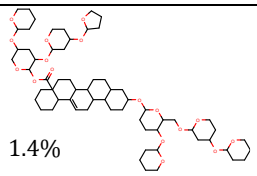
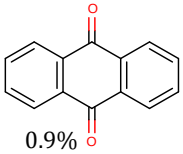
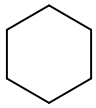
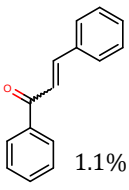
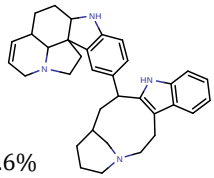
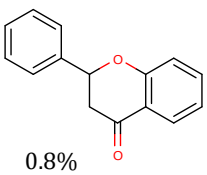
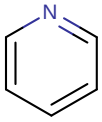
Database	NC	NS	NS/NC	NS_1	NS_1/NS	NS_1/NC	SSE	SSE_5	SSE_{10}	SSE_{20}	n_5	n_{10}	n_{20}
AfroCancer	364	226	0.62	164	0.73	0.45	0.94	0.97	0.96	0.96	0.13	0.19	0.28
Malay Cancer	1,043	425	0.41	322	0.76	0.31	0.73	0.80	0.81	0.82	0.18	0.22	0.28
Ayurveda Cancer	1,091	387	0.35	284	0.73	0.26	0.69	0.71	0.77	0.80	0.17	0.22	0.28
NCI Cancer	187	124	0.66	100	0.81	0.53	0.89	0.97	0.97	0.99	0.14	0.21	0.31
ConMedNP	647	1,128	0.43	758	0.67	0.29	0.87	0.97	0.92	0.92	0.13	0.17	0.23
Approved Drugs in DrugBank	510	892	0.59	740	0.83	0.49	0.85	0.61	0.69	0.77	0.12	0.15	0.19

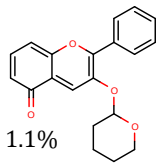
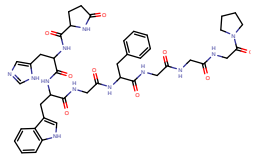
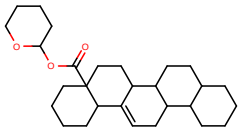
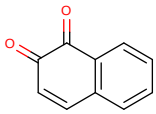
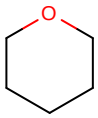
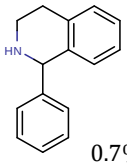
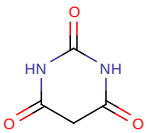
We next analysed the Murcko scaffolds of our datasets in order to be able to interpret scaffold diversity on the chemical level.

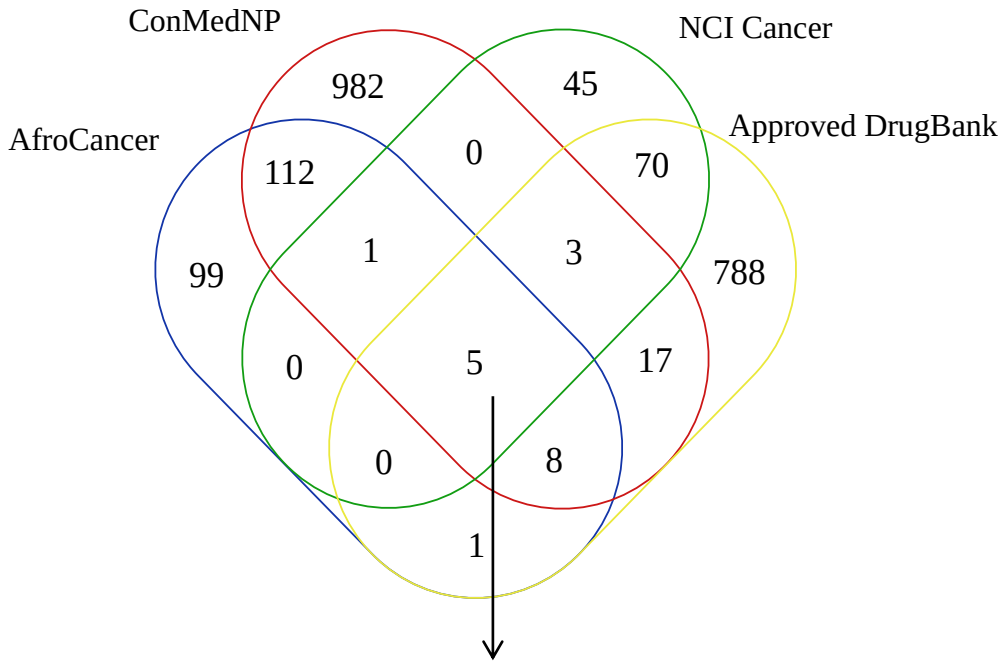
Table 2:2 shows the most frequent Murcko scaffolds in the AfroCancer, NCI Cancer, ConMedNP and Approved DrugBank datasets, along with the percentage of the dataset covered by that scaffold. The percentage of compounds not containing ring systems were 2.19%, 10%, 3.9% and 10% for the AfroCancer, NCI Cancer, ConMedNP and Approved DrugBank datasets, respectively. The benzene scaffold is the most populated scaffold in both the NCI Cancer and Approved DrugBank datasets, whereas it is less populated in both the African datasets (at second place for AfroCancer, and at rank 5 for ConMedNP, respectively). This was also observed in a recent study[198] of all drugs in DrugBank (we used only the approved drugs). Eight of the scaffolds in our top ten most populated scaffold were present in the top 12 scaffolds of this study. A further study identified the top five populated scaffolds of drug and drug-like compounds[199]. Of these, three are present in the top 10 most populated scaffolds of the Approved DrugBank dataset. The flavone (rank 1 in the AfroCancer dataset) and isoflavone (rank 5 in the ConMedNP dataset) scaffolds are also more populated in the African datasets, but absent in both drug datasets in total. This is expected because flavonoids fulfil many

important functions in plants including anti-oxidant activity and cell signalling[200]. The observation of the presence of flavone and isoflavone scaffolds in the top populated scaffolds of the natural product datasets is similar to the observation by Yongye *et al*[168, 201, 202] that flavones, coumarins and flavanones are the most frequent scaffolds in NP datasets. Our analysis shows that the frequent ring systems in African NP datasets are consistent with those found in NP datasets in the literature.

Table 2:2 Top 10 most common scaffolds in the AfroCancer, NCI Cancer, ConMedNP and Approved DrugBank datasets. It can be seen that the benzene scaffold appears in the top 10 most populated scaffolds in all datasets. The compounds in the AfroCancer and ChEMBL dataset are more evenly distributed across their scaffolds.

	AfroCancer	NCI Cancer	ConMedNP	Approved DrugBank
1	 3.6%	 4.3%	 3.6%	 8.5%
2	 2.7%	 3.2%	 3.6%	 1.0%
3	 2.7%	 2.1%	 2.2%	 0.9%
4	 2.2%	 2.1%	 2.0%	 0.8%
5	 1.4%	 2.1%	 1.7%	 0.7%
6	 1.4%	 1.6%	 0.9%	 0.6%
7	 1.4%	 1.6%	 0.9%	 0.6%
8	 1.1%	 1.6%	 0.8%	 0.6%

9	 1.1%	1.6% 	 0.8%	 0.5%
10	 1.1%	 1.1%	 0.7%	 0.5%



	i	ii	iii	iv	v
AfroCancer	2	10	1	2	1
NCI Cancer	1	8	2	2	3
ConMedNP	45	94	1	1	24
Approved DrugBank	1	129	1	1	3

Figure 2:6 Overlap of scaffolds between the different datasets and scaffolds representing them. AfroCancer and NCI Cancer share six Murcko scaffolds, while the ConMedNP and Approved DrugBank datasets share 33 Murcko scaffolds. The table shows the five scaffolds shared by all four datasets and the number of compounds in each dataset having that scaffold.

We next analysed the scaffolds that were present in all 4 datasets. From Figure 2:6 we can see that there are five structures shared between all four datasets. The AfroCancer dataset shares six scaffolds with the NCI Cancer dataset, and ConMedNP shares 33 scaffolds with the Approved DrugBank dataset. For the five-shared scaffolds, structures ii, iii and v are simple and we would expect to see them as scaffolds for small molecule

drug[s198]. The anthraquinone (i) and stilbene (iv) scaffolds are more common in the NP datasets due to the important role these secondary metabolites play in plants, e.g. colouring pigments (anthraquinone) and antimicrobial, cell signalling and antifungal roles (stilbenes)[203]. This finding shows that there is more structural variation and hence different chemical space being covered by the different datasets. This can be extrapolated to see how many unique scaffolds there are in the AfroCancer and NCI Cancer datasets. 99 out of 226 scaffolds (43.8%) were unique to the AfroCancer dataset (when compared to the 45 out of 124 scaffolds (36.3%) in the NCI Cancer dataset).

The totality of bioactive medicinal chemical space from African and approved drug origins were hence rather distinct on the scaffold level: out of the 1,128 scaffolds in ConMedNP, there were 1,095 scaffolds (97%) not present in the Approved DrugBank dataset. In turn, for the Approved DrugBank dataset 98% of the scaffolds were not present in ConMedNP. These percentages are similar to those obtained previously[199] where it was found that 85-92% of scaffolds are unique to a dataset and not found in other datasets. This was also similar to results obtained by Lee and Schneider[197]. They compared scaffolds of natural product libraries and drug libraries, but they cleaved the single bonds (we kept single bonds between rings). They found that 17% of NP scaffolds were present in the drug dataset, and 35% of the drug scaffolds were present in the NP datasets.

As an illustration we investigated one of the scaffolds which is only present in the African datasets, and describe its pharmacological activity in the following. This scaffold, the flavone scaffold, was the fifth most populated scaffold in the ConMedNP dataset (which was not present in Approved DrugBank). Two of the compounds possessing this scaffold (whose structures are shown in the green circle in Figure 2:4) come from the plant *Milletia griffonia*, which is used traditionally to relieve menopausal symptoms and limit bone resorption, i.e. treat osteoporosis[43, 204]. Another compound in this area is Buesgeniine (structure shown in the red circle in Figure 2:4), which was isolated from the stem bark of *Zanthoxylem buesegenii*, a plant used traditionally to treat convulsions[43]. This illustrates the novelty, as well as diversity, of the bioactivities that structures from the plant origins analysed here possess. Previous work had been carried out to identify small hetero-cycles (by enumerating all possible hetero-cycles)

due to their usefulness in the development of drugs[205], and our examples now provide a list of hetero-cycles with ethnobotanical evidence of bioactivity.

In general, these results show that the AfroCancer and ConMedNP databases exhibit more scaffold diversity than traditional compound databases and comparable scaffold diversity to approved drugs. The results also show that the African datasets contain unique scaffolds that are not represented in approved drug datasets, which are able in turn to convey very diverse bioactivities.

2.3.3 Predicting the Mechanism-of-Action of Traditional African Medicines

2.3.3.1 Target prediction for compounds in the AfroCancer dataset

In order to understand the mechanism of action of NPs used against cancer, we implemented a target prediction algorithm (PIDGINv2) on the AfroCancer dataset. We compared this to the experimentally validated targets of the NCI dataset to see if the two datasets shared any predicted target space or if the African NPs have a different mechanism of action.

Target prediction (with the software set to a 0.9 true positive rate, corresponding to a 90% confidence that positive predictions are true positives[127]) was carried out on the AfroCancer dataset and compared with the experimental targets of the NCI Cancer dataset. We found that there are 14 shared targets between the two datasets, with 134 targets uniquely predicted in the AfroCancer dataset, and 82 unique targets in the NCI Cancer dataset (see Supplementary Table 2 for details). We will first analyse trends in predicted targets on a higher level, before subsequently moving on to individual targets and pathways.

We first analysed targets predicted on the protein family level. Figure 2:7 shows the target classes that make up the predicted (AfroCancer) and experimental (NCI Cancer) targets in each dataset. All targets were counted with their classes and normalised with respect to the total number of predictions. In the NCI Cancer dataset kinases are the largest predicted target class (36% of all individual target predictions), in contrast to

only 7% of the targets targeted by the AfroCancer set. In turn, 20% of the target classes were oxidoreductases in the AfroCancer dataset, while only 6% of the NCI Cancer dataset were oxidoreductases. There is similarity in some smaller target classes, e.g. Lipases, which comprise 1% in both data sets. On the other hand, there is a distinct difference in the distributions between both datasets for the GPCRs and Other target classes.

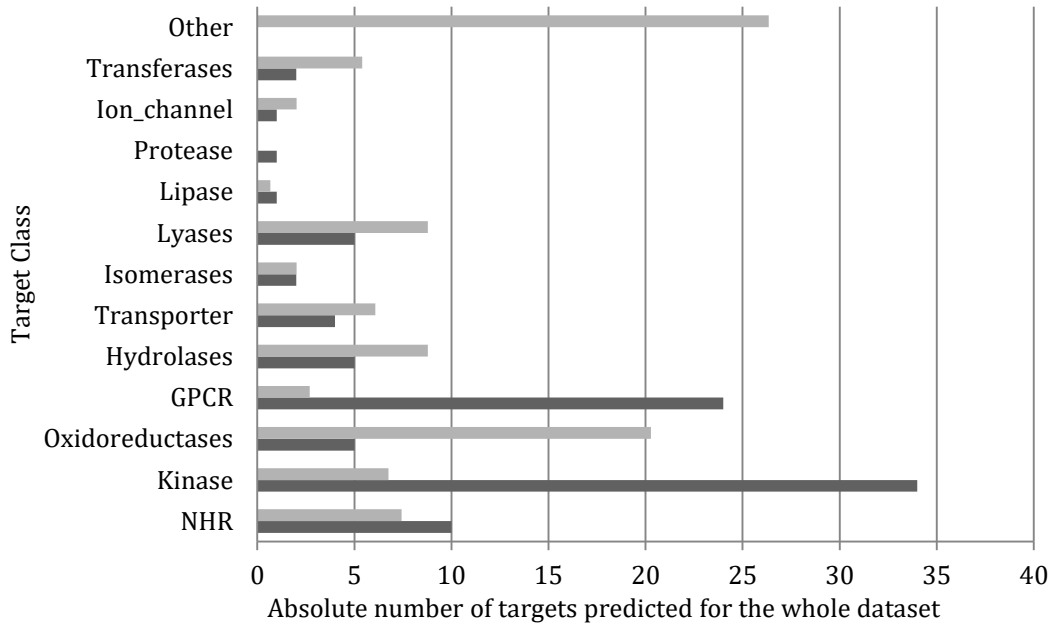
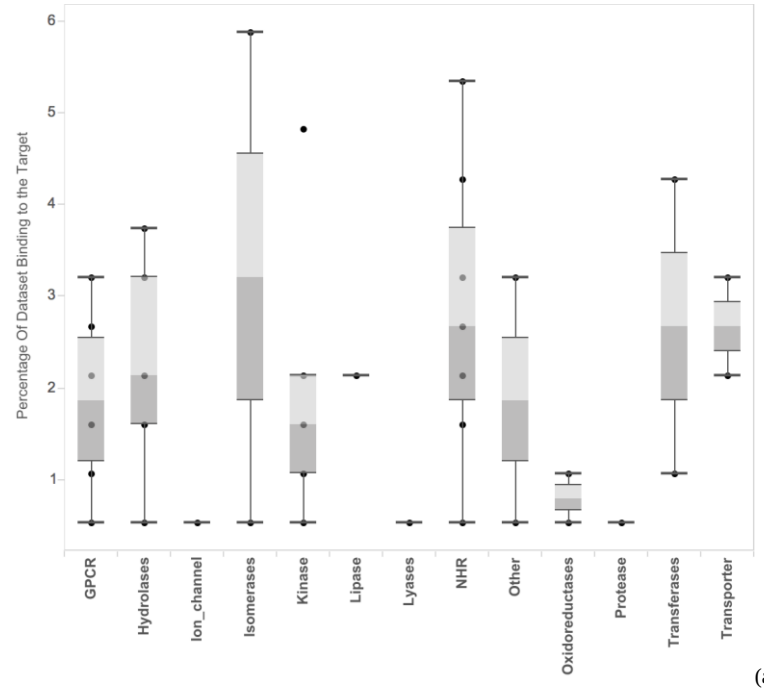
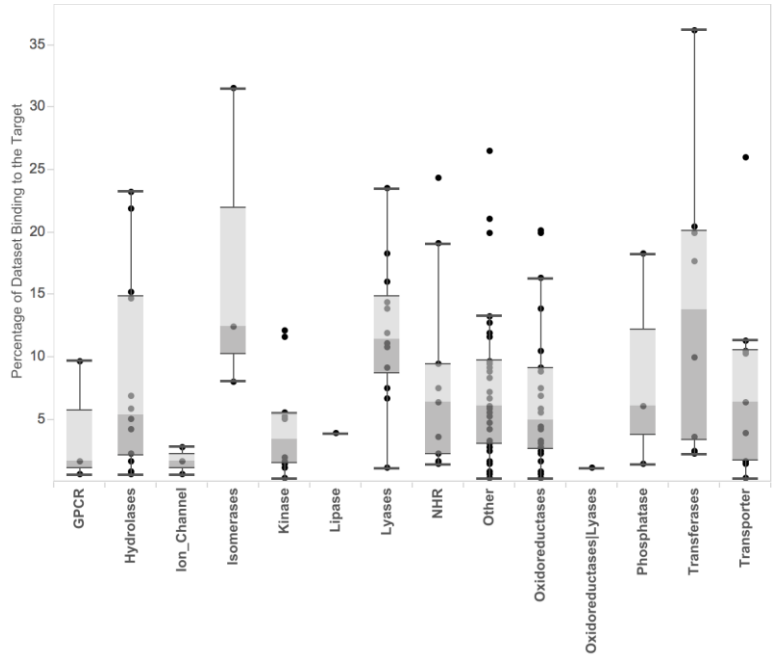


Figure 2:7 Target classes interacting with compounds in the AfroCancer and NCI Cancer datasets. 34% of the targets that bind compounds in the NCI Cancer (dark grey) dataset are kinases and 24% are GPCRs. 20% of the targets predicted to bind to the AfroCancer (light grey) compounds are oxidoreductases while only 7% are kinases and 3% are GPCRs.

We next analysed the number of times a target was predicted. This was carried out to identify and compare the popular targets in each dataset. Whereas Figure 2:7 shows the distribution between unique targets predicted, Figure 2:8 (a and b) takes the number of times a target was predicted into account.



(a) NCI Cancer



(b) AfroCancer

Figure 2:8 Number of targets per target class in the NCI Cancer and AfroCancer dataset. Each circle represents a target from the target class shown on the x-axis. (a) Shows the percentage of the dataset predicted to bind to the different target classes in the NCI Cancer dataset. Less than 10% of the dataset is predicted to bind to kinases, even though they make up 36% of the target classes in that dataset. (b) Transferases only make up 5% of the targets predicted to bind the AfroCancer compounds, yet we see that many compounds in the dataset were predicted to bind to them. More than 30% of the dataset is predicted to bind to isomerases, but they represent only 2% of the target classes for this dataset. (The binding frequency is higher in the AfroCancer dataset because these are predicted targets whereas in the NCI Cancer set, they are experimentally validated targets).

In Figure 2:8 (a) we see that even though kinases make up the majority of the target classes in the NCI Cancer dataset, only around 9% of the compounds in the dataset are predicted to actually target kinases, with the most targeted kinase being Tyrosine protein kinase JAK1. Only 26 small-molecule kinase inhibitors are currently approved by the FDA for cancer indications [206]. In contrast, almost 15% of the AfroCancer dataset are predicted to bind at least one of the kinases in the dataset, with the most targeted kinase being galactokinase. On the other hand, 21 FDA approved drugs with anti-neoplastic activity elicit their effect by modulating 14 GPCRs[207] as primary targets. From Figure 2:7 and Figure 2:8 it can be seen that just over 3% of the compounds in the NCI Cancer dataset modulate 24 GPCRs. Meanwhile, just under 10% of the compounds in the AfroCancer dataset are predicted to modulate not more than 3 GPCRs in total. There are 48 FDA approved Nuclear receptor agonists and antagonists[208]. In the NCI Cancer dataset, 10 Nuclear Hormone Receptors (NHRs) are targeted by over 5% of the compounds in the dataset. By comparison, ~25% of the compounds in the AfroCancer dataset are predicted to modulate only 7 NHRs. These results show that, despite the fact that some target classes known to be involved in cancerogenesis only represent a small percentage of the overall classes, e.g. kinases (7%) and isomerases (2%), the compounds in the AfroCancer dataset are predicted to bind a wide range of target classes.

In our next analysis, we attempt to understand the mechanism of action of NPs in plants used against cancer at the target level. To do this, the predicted targets for the AfroCancer dataset were arranged in order of decreasing enrichment, and the top 10 enriched targets were analysed, the results of which are shown in Table 2:3. The top 100 most enriched targets for the AfroCancer dataset are shown in Supplementary Table 3.

Table 2:3 Top 10 most enriched targets in the AfroCancer dataset and the roles they play in cancer

Target	AfroCancer Hits	PubChem Hits	P Value	Odds Ratio	Role of Target in Cancer
Heat shock protein beta-1	5	18	6.51E-15	6.46E-04	Overexpressed in range of cancers. Plays role in tumor cell proliferation, differentiation, invasion, metastasis, death, and recognition by the immune system[209].
NADPH oxidase 4	38	350	6.76E-91	1.50E-03	Expression is increased in pre-malignant states of lung and liver cancer, and plays a role in the production of reactive oxygen species by cancer cells[210].
Carbonyl reductase [NADPH] 1	32	473	2.12E-70	2.45E-03	CBR1 reduces apoptosis and promotes cell survival in pancreatic b cells by reducing the generation of reductive oxygen species[211].
Cytochrome P450 1B1	1	21	1.38E-43	3.57E-03	Role in cancer and effect of inhibition by NPs are reviewed in[212]
Aldehyde dehydrogenase	21	437	4.00E-03	3.81E-03	May play a role in differentiation and progression of cancer cells. Roles reviewed in[213]
Interleukin-2	1	26	4.90E-03	4.72E-03	Used as an immunotherapy agent to treat cancer[214]
Thioredoxin reductase 2	1	33	6.17E-03	5.99E-03	Involved in tumour oxygenation, roles reviewed in[215]
Multidrug resistance-associated protein 1	37	1790	1.15E-62	7.92E-03	Plays a role in reducing resistance to drugs[216].
Steroid hormone receptor	8	419	2.67E-14	9.32E-03	Overexpression is used as a prognostic marker in breast cancer. Roles reviewed in[217]
Potassium voltage-gated channel subfamily A member 3	6	18	5.09E-11	9.49E-03	Controls the cell resting membrane potential, cell proliferation and apoptosis. Potential new target in lymph node cancer, reviewed in[218]

For all the 10 most enriched targets, links to cancer were identified. The most enriched target in the AfroCancer dataset is Heat shock protein beta-1. Hsp27 is a chaperone of the small heat shock protein and provides cyto-protection and inhibition of apoptosis under stress conditions[219]. Hsp27 is induced by heat shock, hypoxia and DNA damage and is overexpressed in a wide range of cancers[209]. The NP Quercetin is an effective inhibitor of Hsp27[220] and sensitizes glioblastoma cells to temozolomide by increasing caspase-3 activity and inducing cell apoptosis[221]. The scaffold of Quercetin, which is a flavonoid, is the most abundant scaffold in the AfroCancer database (see Table 2:2). This indicates that several other African NPs may share this bioactivity with Quercetin as well as sharing the same scaffold.

Another target involved in cancer in the top 10 is Cytochrome P450 1B1, a mono-oxygenase of endogenous compounds and xenobiotics. It is the most efficient 17β-estradiol hydroxylase (4-hydroxylation of estrogens is considered to be an important step in hormonal carcinogenesis)[222]. Furthermore, Cyp450 1B1 is involved in the metabolism of some cancer drugs, e.g Docetaxel, which leads to drug resistance that is associated with the overexpression of CYP1B1[223, 224]. CYP1B1-null mice show no

obvious change in phenotype, which indicates CYP1B1 is not necessary for mammalian development[225]. It is highly expressed in cancers of the breast, colon, esophagus, skin, lymph nodes, brain and testicles compared to healthy tissues[222]. These observations indicate that CYP1B1 is a potential target of interest in cancer[226]. However, it is important to specifically inhibit CYP1B1 because CYP1A1, which displays 41% amino acid sequence similarity to CYP1B1, plays a role in the detoxication of environmental procarcinogens, and also contributes to the metabolic activation of dietary compounds with preventive activity against cancer[227]. NP inhibitors of this CYP1B1 include coumarins, flavonoids, stilbenes and anthraquinones[212]. The compounds predicted to bind to this enzyme from the AfroCancer dataset are stilbenes and flavonoids with a hydroxyl and/or methoxy substitution at the 3' and 4' positions, which is also required for selectivity over CYP1A[212]. This is beneficial as the CYPs, once modulated by the NPs, will be deactivated and hence the xenobiotics, (or other active NPs in this case) will not be detoxified in the cells[228].

There is evidence from the literature that several of the targets listed in Table 2:3 are modulated by natural products. Potassium voltage-gated channel subfamily A member 3 is important in setting the cell membrane potential and is currently being considered as a potentially new anti-cancer target[218]. It has been found to be overexpressed in various cancers including breast, colon, smooth and skeletal muscle and lymph node cancers[218]. Inhibition of this channel arrests the G1 phase of the cell cycle[229], thus stopping cell proliferation. A recent study found that the flavonoid, 8-prenylnaringenin (*Humulus lupulus*) inhibits the gate at micromolecular concentrations[230]. DNA topoisomerase I and IIα were also predicted to bind compounds from the AfroCancer dataset. These enzymes make incisions in the backbone of the DNA, thus catalysing the winding and unwinding of the DNA strands. Inhibitors of DNA topoisomerase I and DNA topoisomerase II α induce single and double strand breaks respectively, thus inhibiting the cell cycle at the G2 stage[231]. They are both validated anti-cancer targets currently inhibited by irinotecan, topotecan and camphotethecin (DNA topoisomerase I) and etoposide, doxorubicin and daunorubicin (DNA topoisomerase II). These enzymes are also inhibited by phytoalexins, namely genistein[232], quercetin[233] and resveratrol[234]. Estrogen receptor α and β were predicted to bind compounds in the AfroCancer dataset and they are both validated drug targets being inhibited by the pro-

drug Tamoxifen[235] and Fulvestrant[236]. The phytoestrogens genistein, kaempferol and liquiritigenin are known agonists of these proteins [237].

As can be seen from the above, there is broad literature support for NPs acting on the predicted targets. The AfroCancer NPs may be acting via the same mechanisms and this provides a promising insight into their mechanism of action.

We next investigated which *novel* proteins (i.e., those not currently targeted by anti-cancer drugs) are predicted to be targeted by the AfroCancer compounds, in order to firstly understand which novel mechanisms might already be used by African medicines to treat cancer, and secondly to make concrete suggestions which type of chemistry might be active against which target(s) in an *in vivo* setting.

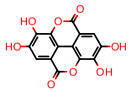
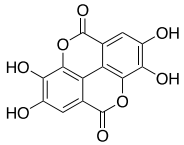
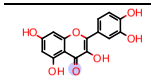
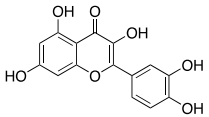
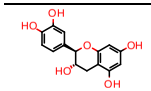
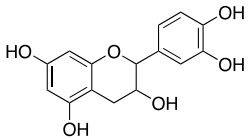
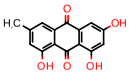
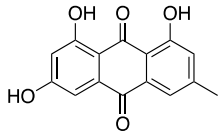
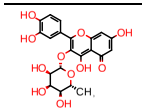
In this regard, Table 2:4 shows the unique enriched targets and structures (of either the top 5 most similar structures or with 0.3 or more similarity to those in the training set, whichever is more) that are predicted to bind to them.

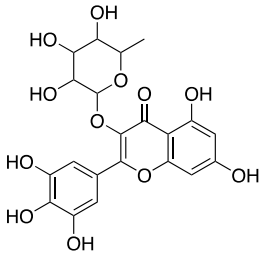
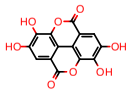
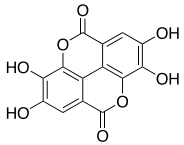
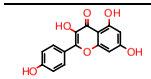
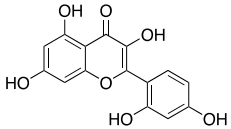
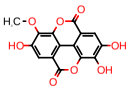
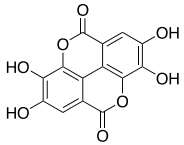
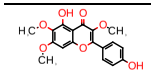
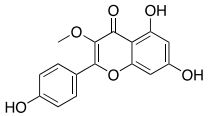
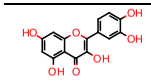
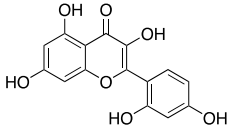
NPs in AfroCancer are predicted to bind Mcl-1, which is an anti-apoptotic member of the Bcl-2 family. Mcl-1 is a target of interest and inhibitors are being pursued as drugs [238]. Currently, Omacetaxine Mepesuccinate and Seliciclib are marketed drugs that inhibit the synthesis of Mcl-,1 but there are no drugs approved as of now that inhibit the function of the actual protein. Flap endonuclease is also predicted to be a target. It is overexpressed in breast[239], prostate[240], stomach[241], neuroblastoma[242], pancreatic[243] and lung cancer[244] and is responsible for inaccurate repair of double strand breaks in the DNA repair pathway[245]. Overexpression is associated with cancer because inaccurate DNA repair leads to a higher risk of mutations and thus an increased risk of cancer. Another unique target predicted for the AfroCancer dataset is HSP70 which plays a housekeeping role in conditions of stress. The role of this target and its potential as an anticancer target has been recognised and reviewed [246, 247]. To date, there are no drugs in the market targeting HSP70. NPs from traditional medicines such as the datasets analysed here, which are targeting HSP70, could thus be exploited for further experimentation to see if they are viable modulators. Tankyrase 1 is another of the novel targets predicted for the AfroCancer datasets. Tankyrase 1 binds to telomeric repeat

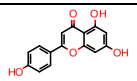
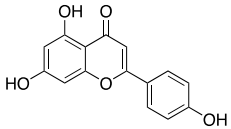
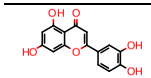
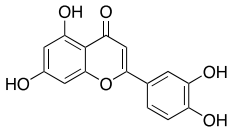
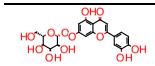
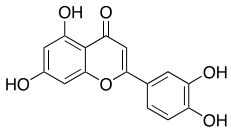
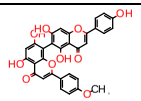
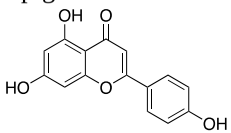
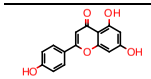
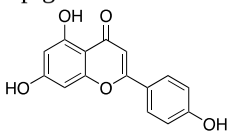
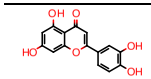
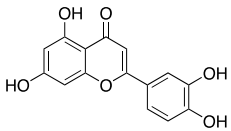
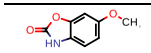
factor 1 (TRF1) which positively regulates telomere length[248, 249]. It does this by poly(ADP-ribosyl)ating TRF1 to release it from telomeres[250], hence allowing access of telomerase to telomeres. Telomeres consist of tandem repeats of a G base rich DNA sequence[251]. Shortening of the telomeres below a certain threshold leads to end-to-end chromosome fusions, cell cycle arrest and/or apoptosis[251]. Thus, the importance of Tankyrase 1 in cancer lies in prevention of telomere shortening and hence prevention of cancer cell cycle arrest and apoptosis. Tankyrase is suggested to be a therapeutic target[252, 253] with one of the reasons being that it was found that inhibition of Tankyrase 1 accentutated the ability of MST-312 to induce telomere shortening[254]. Also, over-expression of Tankyrase 1 was found to promote telomere elongation[249]. A reduction in Tankyrase 1 has been shown by Dynek and Smith[255] to cause cells to accumulate in the M phase of the cell cycle. It has been found by Chang *et al*[256] to cause abnormal spindle structures. Several compounds in the AfroCancer dataset were predicted to bind this target. They may be acting by allowing cell cycle arrest and death by inhibiting Tankyrase 1. Two mitotic-specific cyclins B2 and B3 were predicted for the AfroCancer compounds. They are also involved in control of the cell cycle at G2/M transition[257]. Overexpression of G2/mitotic specific cyclin B2 is associated with poor prognosis in patients with non-small cell lung cancer[258], whereas a decrease of expression leads to an inhibition of both invasion and metastasis in bladder cancer[259]. This means that their modulation by the AfroCancer compounds may lead to cell cycle arrest and better prognosis.

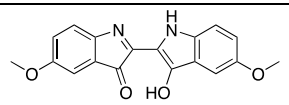
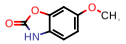
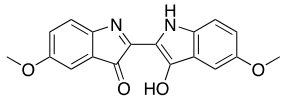
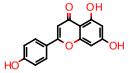
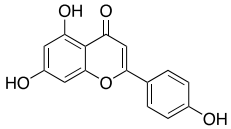
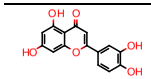
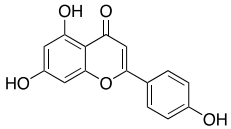
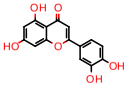
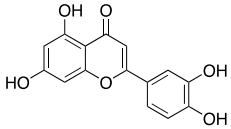
Hence, overall it can be seen that NPs from medicinal plants with anti-cancer activity are predicted to bind novel cancer-related targets. These results provide insight into the mechanism of action of these NPs.

Table 2:4 Unique cancer related targets in the AfroCancer dataset. These target predictions had an odds ratio below 0.1 and compounds with Tanimoto similarity of 0.3 or more to those in the training set. It can be seen that the unique targets are predicted to be modulated by NPs that are similar in structure to bioactive compounds from ChEMBL. pChEMBL values, when reported, are given in brackets.

Target	Structure of AfroCancer Compound	Similar compounds in ChEMBL	Reported activity of similar ChEMBL compound against target (μm)	Tanimoto Similarity
Flap endonuclease 1		Ellagic Acid – CHEMBL6246 	1.58	1
		Quercetin – CHEMBL50 	2.24	1
		Cianidanol – CHEMBL311498 	4.74	1
		Emodin - CHEMBL289277 	15	1
		CHEMBL1599224	2.69	0.84

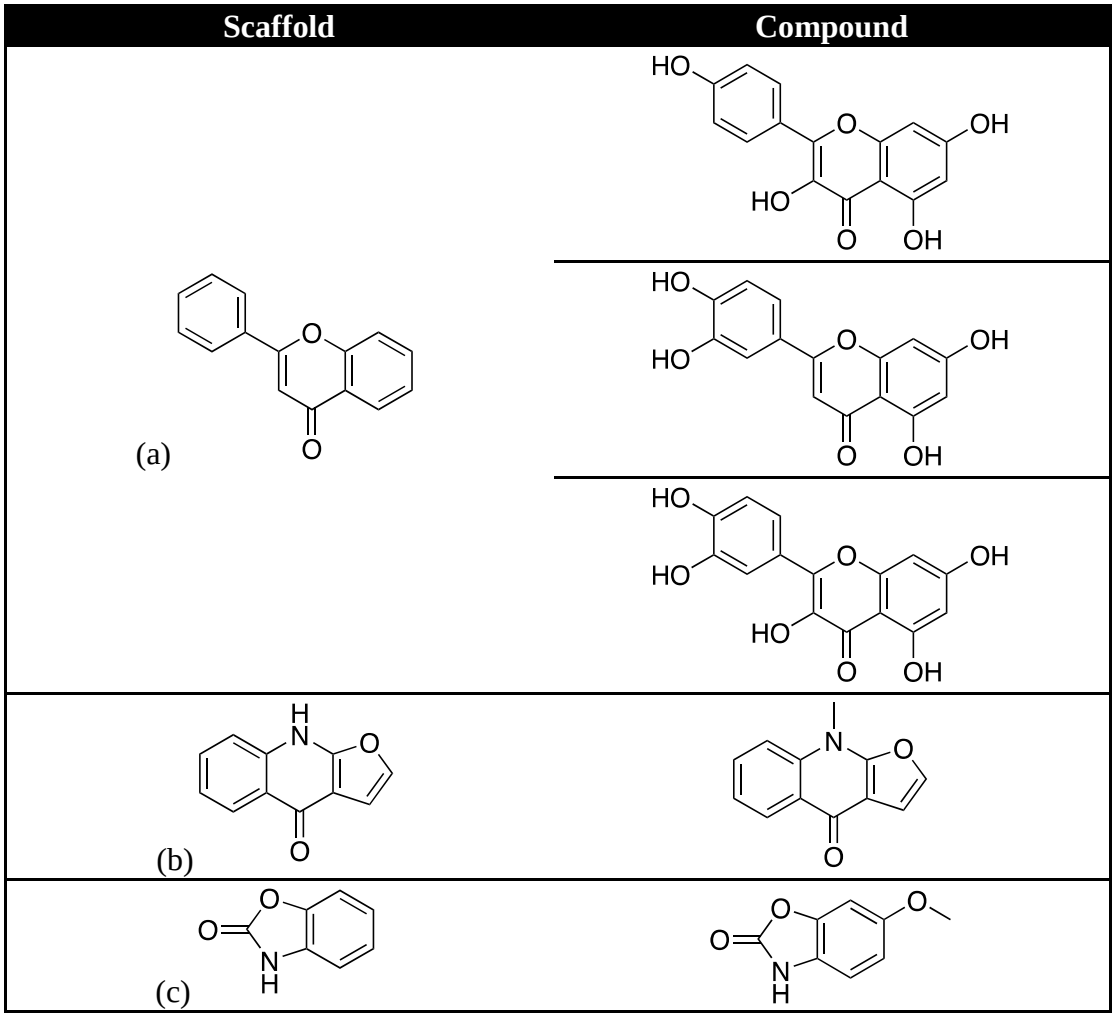
				
Induced myeloid leukemia cell differentiation protein Mcl-1		Ellagic Acid – CHEMBL6246 	54	1
		Morin - CHEMBL28626 	2.99 (5.53)	0.75
		Ellagic Acid – CHEMBL6246 	54	0.72
		Isokaempferide- CHEMBL165064 	5.41 (5.27)	0.69
		Morin - CHEMBL28626 	2.99 (5.53)	0.67

Tankyrase-1		Apigenin – CHEMBL28 	3.1 (5.51)	1
		Luteolin – CHEMBL151 	2.4 (5.62)	1
		Luteolin – CHEMBL151 	2.4 (5.62)	0.54
G2/mitotic-specific cyclin-B2		Apigenin – CHEMBL28 	4 .00 (5.4)	0.52
		Apigenin – CHEMBL28 	4 .00 (5.4)	1
		Luteolin – CHEMBL151 	6.2 0 (5.21)	1
		CHEMBL114396	1.20 (5.92)	0.38

				
G2/mitotic-specific cyclin-B3		CHEMBL114396 	1.20 (5.92)	0.333
		Apigenin – CHEMBL28 	4 .00 (5.4)	1
		Apigenin – CHEMBL28 	4 .00 (5.4)	1
		Luteolin – CHEMBL151 	6.2 0 (5.21)	1

To determine if there is a relationship between novel scaffolds and novel targets, we next study the relationship between the unique targets (predicted for AfroCancer compounds and not present in NCI Cancer targets) and the scaffolds of the compounds they were predicted to bind. Table 2:5 (a) shows the compounds (and their common Murcko scaffold) predicted to bind heat shock protein beta-1. One of these compounds is Quercetin. As this compound has been shown to inhibit heat shock protein by Western Blot analysis[260], we are confident in the prediction. It is important to note however that polyphenols pose as a problem in drug discovery because they bind targets promiscuously and have poor pharmacokinetic properties[261]. Similarly, the scaffold in Table 2:5 (b) was unique to the AfroCancer dataset as was the target it was predicted to bind, cyclic dependent kinase 14. In Table 2:5 (c) the scaffold from a compound predicted to bind G2/mitotic –specific cyclin-B3 was found in both the AfroCancer and the Approved DrugBank datasets (yet none of the drugs in the DrugBank dataset are known to modulate this target), but not the NCI Cancer dataset. This is relevant since we have identified important anti-cancer targets predicted to be modified by compounds with unique scaffolds. We have predicted the mechanism of action of NPs of plants used against cancer and they have a different mechanism of action to drugs in the market, as was predicted due to their different scaffolds.

Table 2:5 Scaffolds of compounds predicted to bind unique targets. (a) The flavonoid scaffold which is common to the three compounds shown in (a) is unique to the AfroCancer dataset. The compounds were predicted to bind heat shock protein beta-1. (b) This scaffold was also unique to the AfroCancer dataset. This compound was predicted to bind cyclic dependent kinase 14. (c) This scaffold was found in the AfroCancer dataset but not the NCI Cancer dataset. It was found in the Approved DrugBank dataset. This compound was predicted to bind G2/mitotic –specific cyclin-B3. These unique scaffolds (a and b) are occupied by compounds that are predicted to bind unique targets.



2.3.3.2 ANNOTATING TARGET PREDICTIONS FOR THE AFROCANCER DATASET WITH PATHWAYS

We next annotated the predicted targets for the AfroCancer dataset with pathways that they are involved in, to further understand their potential molecular mechanism of action. From this analysis 102 and 89 WikiPathways were obtained for the experimentally validated NCI Cancer targets and predicted AfroCancer targets

respectively, and 52 pathways were common to both datasets (see Supplementary Table 4).

Looking at the pathways modulated by the NCI Cancer compounds (Supplementary Table 4) we see that they modulate, among others, the cell cycle and cell death, e.g. DNA Replication, Senescence and Autophagy, and Regulation of Microtubule Cytoskeleton. The AfroCancer compounds are predicted to also act on the cell cycle, but they are predicted do so through different pathways, i.e. G1 to S cell cycle control by modulating cyclin-dependent kinase 6 and Apoptosis Modulation and Signalling by modulating Bcl2, MAPK3, MCL1 and NFKB1A.

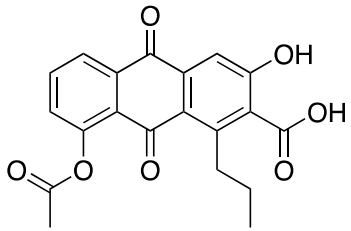
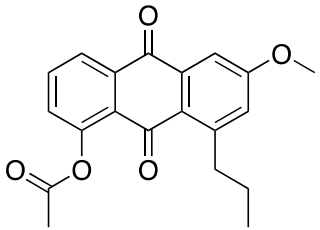
One of the 37 unique pathways modulated by the compounds in the AfroCancer dataset with evidence of involvement in cancer is the Keap1-Nrf2 Pathway. The compounds are predicted to act by modulating Nrf2 (nuclear factor erythroid 2-related factor 2). Nrf2 is highly expressed in pre-malignant and malignant cells and enhances both chemo-resistance and growth of tumour cells[262]. The NP brusatol (a quassinoid isolated from *Brucea javanica*) is known to inhibit Nrf2, and thus increase chemosensitivity and reduce tumour size[263]. The Nrf2 transcription factor is also responsible for cytoprotection against chemical and oxidative stress[264]. No known inhibitors are currently in the market due to structural similarity with other bZIP domain containing proteins[262], causing selectivity problems due to off-target effects. Only one other protein, CREBBP, containing the bZip domain was predicted to bind the compounds from the AfroCancer dataset and none of the compounds predicted to bind Nrf2 were predicted to bind CREBBP, hence indicating possible selectivity (though the target prediction models used do not necessarily provide the resolution needed to be certain about this type of predictions). These results suggest that the AfroCancer dataset comprises some chemopreventive activity through Nrf2 inhibition.

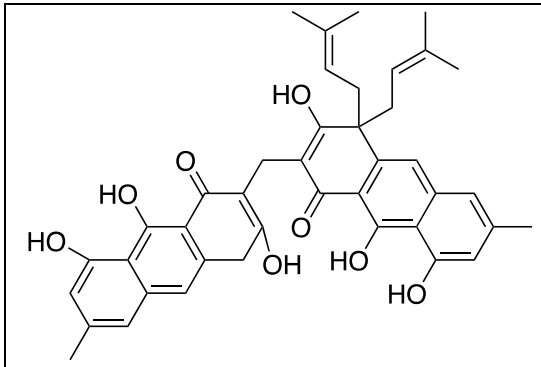
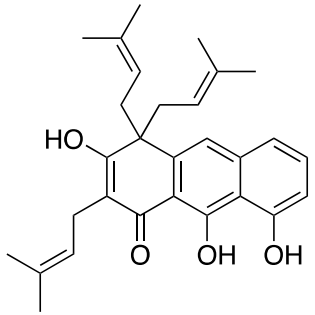
Taken together, these results lead us to believe that the AfroCancer compounds occupy a different yet pharmacologically relevant biological space compared to approved medicines in the NCI Cancer dataset, and that their activities appear to be achieved by somewhat different means, both in target and pathway space.

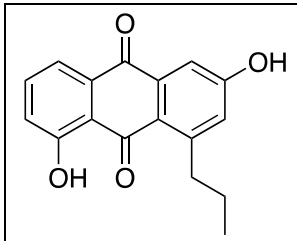
2.3.3.3 CASE STUDY: TARGET PREDICTION OF COMPOUNDS ISOLATED FROM PSOROSPERMUM AURANTIACUM

We then turned our attention to one particular plant in the ConMedNP dataset which has a variety of uses, in order to see whether our analyses can provide insight into its mode of action. Fruits of the plant *Psorospermum aurantiacum*, Family Hypericaceae are used in Cameroon and other parts of Africa for the treatment of cancer as well as gastrointestinal and urinary tract infections, skin infections, venereal diseases, gastrointestinal disorder, infertility, epilepsy and microbial infections[43]. The variety of indications seemed surprising to us at first, hence, target prediction was used to shed light on why this plant was used for such a wide range of seemingly unrelated indications. Results for the target predictions for the 5 NPs isolated and characterised from *Psorospermum aurantiacum*[265] are shown in Table 2:6.

Table 2:6 Predicted targets for compounds from *Psorospermum aurantiacum*. For each compound we can see that it is linked to one of the activities that the plant is traditionally used for.

Structure	Predicted Target/s	Link to plant indication (as found in ConMedNP dataset)	Biological Role of Predicted Target/s
 **C 8- (acetyloxy)-3-hydroxy-9, 10-dioxo-10propyl-9, 10-dihydroanthracene-2-carboxylic acid (Compound 1)**	Protein kinase C gamma type	Skin infections	Protein kinase C inhibitor ingenol mebutate used to treat actinic keratosis, which is a pre-cancerous condition of the skin that manifests as a patch of thick, scaly skin[266].
	Peroxisome proliferator-activated receptor gamma	GIT disorder	Regulates fatty acid storage and glucose metabolism[267]
	Induced myeloid leukemia cell differentiation protein Mcl-1	Cancer	Inhibits apoptosis[268, 269]
	Leukotriene B4 receptor 1	Microbial infections	A receptor for leukotriene B4 which is a potent chemoattractant[270] involved in inflammation and immune response[271]
 **6-methoxy-9, 10-dioxo-8-propyl-9, 10-dihydroanthracen-1-yl acetate**	Endothelin-1 receptor		Activation results in elevation of intracellular-free calcium, thus binding of endothelin increases vasoconstriction, mediates neurotransmission[272] – no link to indications
	Induced myeloid leukemia cell differentiation protein Mcl-1	Cancer	Inhibits apoptosis[268, 269]

 **Psorantin**	G-protein coupled receptor 55		Putative cannabinoid receptor. May be involved in hyperalgesia[273] – no link to indications
	Induced myeloid leukemia cell differentiation protein Mcl-1	Cancer	Inhibits apoptosis[268, 269]
 **Ferruginin B**	Estrogen receptor β	Infertility	Potent tumour suppressor, ERB knockout mice have been reported to be either infertile or subfertile[274]
	G-protein coupled receptor 55		Cannabinoid receptor, increases intracellular calcium[275], affects osteoclast function and bone mass[273]
	Estrogen receptor	Epilepsy	Plays a role in catamenial epilepsy[276]

 **Haronginanthrone**	Estradiol 17-β-dehydrogenase 2	Infertility	Catalyses the interconversion of testosterone and androstenedione, and estradiol and estrone, and also responsible for in utero embryonic and placenta development[277]
	Estrogen receptor β	Infertility	ERB knockout mice have been reported to be either infertile or subfertile[274]

Compound 1 is predicted to bind to Protein kinase C gamma, and it has previously been shown to be linked to keratosis. Neutrophilic cutaneous infiltrates are produced as a result of activating Protein kinase C, and they act to prevent relapse of the tumour by mediating antibody-dependent cellular toxicity against the tumour cells[266, 278]. It thus appears that the activity of Compound 1 against Protein kinase C gamma may be responsible for its utilization for treating skin infections. At the same time, Compound 1 is also predicted to bind to Leukotriene B4 receptor 1 and Induced myeloid leukaemia cell differentiation protein Mcl-1. Elevated levels of Leukotriene B4 receptor 1 have been found in a number of inflammatory diseases[271], so inhibition may explain the anti-inflammatory activity associated with this plant. Apart from annotated indications, an experimental leukotriene B4 inhibitor was found to inhibit proliferation and induce apoptosis in pancreatic cells[270], and suppression of induced myeloid leukemia cell differentiation protein Mcl-1 is known to induce apoptosis[268], so this compound may have the potential to also act as an anti-tumour agent. The targets predicted for Haronginanthrone can explain the seemingly unrelated annotated bioactivity of this plant in treating epilepsy and infertility. This compound was predicted to bind to the estrogen receptor and estrogen receptor beta, which play an important role in cancer [279, 280] and catamenial epilepsy[276]. More proof that estrogen receptors play a role in epilepsy is shown by the fact that reproductive dysfunction is associated with epilepsy[281] as well as anti-epileptic therapy[282]. A literature review revealed that estrogen receptor knockout mice have been shown to exhibit infertility as well as reduced fertility[274]. The fact that *Psorospermum aurantiacum* is used to treat infertility indicates that the compounds may bind to estrogen receptor β and act as agonists.

Looking at the bioactivity profiles of the compounds in *Psorospermum aurantiacum* and similar compounds in the ChEMBL database we find that these compounds are predicted to have activity in a variety of cell lines and targets e.g. Ferruginin C and Vismin (>80% similar to compounds in *P. aurantiacum*) have reported activities against cancer lines including MCF-7[283].

As we can see, target prediction has allowed us to develop a plausible mode of action hypothesis for this plant, despite the rather dissimilar indications for which it is being used. Furthermore, combining information from different sources, e.g. ethno-botanical

use with target prediction gives further validation and greater confidence in the traditional use of these plants as medicines.

2.4 CONCLUSION

In this study we have looked at the scaffolds and the diversity of the African compound libraries compared to other NP libraries and approved drugs. We showed that African compounds are structurally diverse and share only a proportion of structural space, namely 3.6% of scaffolds, with approved drugs. 97% of the scaffolds in ConMedNP were unique and not present in the Approved DrugBank compounds, and 43.8% scaffolds were unique to the AfroCancer dataset (when compared to the 36.3% scaffolds in the NCI Cancer dataset), representing unexplored chemical space with some evidence of therapeutically relevant biological activity. We also showed that these African compounds share 14 predicted targets with those of the NCI Cancer compounds, but that the remainder represents targets with potential novel therapeutic value.

Results obtained using the target prediction algorithm gave an indication of the mechanism-of-action of the compounds from the AfroCancer dataset. Three targets (MCL-1, bcl2 and Flap endonuclease) have been identified as targets that can be modulated and the compound-target predictions experimentally validated in further studies. Pathway analysis of the AfroCancer dataset revealed 14 cancer related pathways similar to those modulated by the cancer drugs in the market, though they appear to act via different mechanisms of action as shown by the different targets and stages of the pathway modulated. Novel pathways, e.g. the Keap1-Nrf2 Pathway and Apoptosis Modulation by HSP70, provide starting points both from the chemical and the biological side for future anti-cancer treatments, derived from traditional African medicines.

As a more detailed case study, the apparent variety of conditions against which *Psorospermum aurantiacum* is used was also explained using target prediction. This finding illustrates the benefit of target-prediction in shedding light into the mechanism of action of plants that have not previously been extensively studied. Furthermore, this approach can be used to guide the screening of African plants for which no molecular targets are currently known.

CHAPTER 3: INTEGRATING ETHNO-BOTANICAL AND PHYLOGENETIC INFORMATION OF MEDICINAL PLANTS WITH THEIR PREDICTED MECHANISMS OF ACTION TO IDENTIFY PHYLOGENTIC PATTERNS OF USE

3.1 INTRODUCTION

The combined knowledge of ethno-medicine and phylogenetic information of plants has been utilised to identify promising lineages of medicinal plants for lead identification, such as by Hawkins *et al*[145] in the genus Pterocarpus (Leguminosease). This was demonstrated by studying plants from the genus Pterocarpus across Indomalaya, Tropical Africa and the Neotropics. Plants with medicinal activity were concentrated on specific clades, i.e. they were not randomly distributed. This study provided a link between biogeography and phylogeny of the plant. Some studies [146, 147] have also looked at the possibility of predicting medicinal potential of a plant using its phylogeny. The study by Saslis-Lagoudakis *et al*[147] found that phylogenetic patterns were shared among the medicinal plant species of the flora of Nepal, New Zealand and Cape of South Africa. "Hot nodes", which are nodes in the phylogenetic tree (corresponding to plant families or genera) that are significantly over-represented in species with a given property, e.g. anti-Malarial, compared with the rest of the tree, comprised on average 133% more medicinal plants than a random sample of the studied flora, thus demonstrating that plants descended from related lineages are used for the same conditions across continents. A recent study used evolutionary tools to predict plant lineages with psychoactive properties[284] and narrowed down the prospect of psychoactive plants to 8.5% of all land plants. These approaches are based on the hypothesis that phylogenetic lineages with plants used in traditional medicine are more likely to contain plants with medicinally active products.

The ecology tools used in the studies mentioned above can be applied to TAM, since a need exists for more research into TAMs used against endemic disease, e.g. malaria and HAT. The hypothesis is that there is a relationship between phylogenetically related

plants and their medicinal use in African medicinal flora. Any such relationship is likely to be related to the unique metabolites produced by the phylogenetically related plants.

To test this hypothesis, we start by using the newly available NANPDB dataset [42] as a background dataset to explore whether plant species in the same family in North Africa produce chemically similar natural products (NPs). We then compare the similarity of NPs produced by the same family in African, Malay and Ayurveda plants used against cancer. Subsequently, we integrate phylogenetic information to determine whether the phylogenetic grouping of the plant is correlated with the chemistry and predicted targets of the NPs that these plants contain. Next, we identify plant families whose members are over-represented as remedies against cancer in African, Malay and Ayurveda traditional medicine, and then narrow down the search to the African medicinal flora, where we identify the over-represented families used for cancer, malaria and human African trypanosomiasis. This is followed by investigating the relationship between the unique metabolites produced by the plants, and whether or not this plays a role in their over-representation in the medicinal flora of Africa. If present, quantifying this type of relationship will have a two-fold benefit: (i) it will guide the phytochemist towards the type of plants to explore when looking for modulators of specific targets, and (ii) it will help to train prediction models on the type of target classes that compounds from this plant may modulate.

3.2 MATERIALS AND METHODS

3.2.1 DATASETS

3.2.1.1 ETHNOMEDICINAL INFORMATION - CANCER

The African compounds analysed in this study were obtained from AfroCancer [44] which contains 390 compounds in SD format. The annotations of each compound, containing compound name, plant origin and ethnobotanical use, are not freely available and were obtained by special permission from the CBIC (Chemical and Bioactivity Information Centre, University of Buea, Cameroon). 1,091 compounds from Malay traditional medicine, which have reported anti-cancer activity, were derived from the commercial database Natural Product Discovery System (NADI)[179], hereafter referred to as 'Malay

Cancer' dataset). 1,043 compounds with reported anti-cancer properties from Ayurveda were obtained from Dr. Duke's Phytochemical and Ethnobotanical Databases[180] and will henceforth be referred to as the 'Ayurveda Cancer' dataset. Structures for the compounds from Malay and Ayurveda traditional medicine were downloaded from PubChem[285], ChemSpider[180] or HMDB[286], by matching the SMILES strings. The combined NPs from plants from AfroCancer, Malay Cancer and Ayurveda Cancer will be referred to as the "AMA" dataset. Plants that have NPs currently used against cancer in the market were also added to this dataset, namely *Catharanthus roseus* and *Taxus brevifolia*.

3.2.1.2 ETHNOMEDICINAL INFORMATION – HAT

Medicinal plants used traditionally in Africa against HAT were obtained from a recent review article[23]. Here, activity is defined as traditional use against HAT or *in vitro* and *in vivo* studies. This dataset will be referred to as AfricaTryp.

3.2.1.3 ETHNOMEDICINAL INFORMATION – MALARIA

Medicinal plants used traditionally in Africa against Malaria were obtained from the AfroMalaria dataset by special permission from the CBIC. This dataset contains compounds from 95 species and all taxonomic information (family, genus, species) was used.

3.2.1.4 INTER- AND INTRA- FAMILY CHEMICAL SIMILARITY OF NPS IN NANPDB AND THE AMA DATASET

The Northern African Natural Products Database (NANPDB) is a global dataset of ~4500 North African natural products curated from literature between 1962-2016. In an attempt to investigate the relationship between the chemical similarity of compounds within plant families and across plant families we used all the NPs in this dataset extracted from the Plantae kingdom. Secondary metabolites produced by plants serve different purposes in the plant, ranging from defence to pollination. It is therefore expected that plants will produce the same or similar compounds regardless of the family. At the same time, we know from the study of chemotaxonomy (the process of

classifying plants according to the secondary metabolites that they produce and the biosynthetic pathways used to produce them) that some plant families and closely related families produce unique metabolites. (It is this second fact that we hope to take advantage of when using plant phylogeny to predict activity.) To view this information we constructed an MDS plot in R[182] of the fingerprints of the NANPDB compounds. Extended connectivity fingerprints (ECFP_6) [114] were generated using Canvas, version 1.5, Schrödinger[184-186]. The structural similarity between the compounds in the AMA dataset was visualised in DataWarrior[191], generated using 2D RBS (2-dimensional rubber band scaling applied as described in [191]).

3.2.2 PHYLOGENETIC ANALYSIS AND MANIPULATIONS

We next analysed the position of phylogenetic clumping in different medicinal groups, followed by identifying important families in the African flora for use against cancer, malaria and HAT. To this end, metrics from community ecology phylogenetics (described below) were used to explore the lineages where the clustering of medicinal use is present in the datasets.

To identify the position of phylogenetic clumping from different "medicinal groups" on the phylogeny, the "*nodesigl*" command in Phylocom v4.2[287] was used. This command identifies nodes that are significantly over-represented in genera having a specific medicinal use (i.e. belonging to a "medicinal group") compared with the rest of the tree. In "*nodesigl*", the observed pattern for each sample is compared to the pattern of random samples using a null model that draws "*s*" taxa from the phylogeny terminals where "*s*" is the number of taxa in a sample. The dataset was tested with the Zanne[162] tree as the background phylogeny to generate the random samples.

To measure the phylogenetic distance between plant families (internal nodes of a species tree) we used the *dist.nodes* function from the "ape" package[288] in R[182]. This measures the distances between nodes by computing the pairwise distances between the pairs of internal and external nodes from a phylogenetic tree using its branch lengths, which in this case each unit correspond to millions of years ago (*mya*).

The phylogenetic similarity between taxa (in this work, plant families) is measured by the patristic distance[289] calculated from the sum of branch lengths connecting the studied taxa. The branch lengths represent the amount of genetic change between the taxa studied (family, genus, species etc). Calculating the patristic distance is shown in Figure 3:1 below. The patristic distance between terminal taxa (A, B and C) is equal to the sum of the branches connecting those taxa. The larger the number, the larger is the distance between the taxa and hence the further away they are in evolutionary terms.

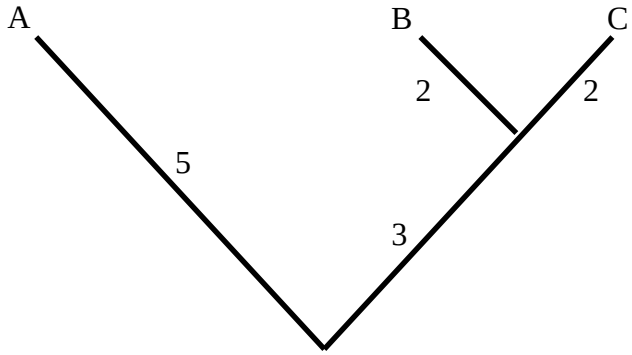


Figure 3:1 Calculating the patristic distance. The figure shows an example of a simplified phylogenetic tree showing arbitrary branch lengths. The pairwise patristic distance calculated between the tree tips A, B and C is shown in the table.

	A	**B**	**C**
A	0	10	10
B	10	0	4
C	10	4	0

3.2.2.1 Binomial analyses

We next analyse the distribution of medicinal plants in families across the African flora. Binomial analysis highlights families that depart from a uniform model of proportion of medicinal plants in a given flora- assess the patterns for medicinal plant usage across a flora. An exact randomisation Goodness of fit test approximated via Monte Carlo simulation was carried out on a contingency table for the African flora (medicinal and non-medicinal plants in a family for cancer, malaria and Trypanosoma) to test the deviation from the null hypothesis (a uniform proportion of medicinal species among families). A small p-value indicates the medicinal (anti-cancer, antimalarial and antitrypanosomal) species are not evenly distributed among families in the African flora.

The null hypothesis in our case is; H_0: $M_i = p_{\text{flora}} \times s_i$; i.e. plants belonging to family *I* are no more likely to be used medicinally than would be the case for the flora as a whole, i.e. the proportion of medicinal plants in family $i(p_i)$ equals the proportion of medicinal plants in the total flora ($p_{\text{flora}} = \Sigma m_i / \Sigma s_i$).

The binomial p-values for over -representation were calculated using the BINMODIST in Excel. BINOMDIST gives (a) probability that there are x or fewer successes, and (b) that there are exactly X number of successes. The probability of X or more successes is (1-(a)) + (b). In this case the number of successes is the number of medicinal species.

The number of species per family for the African flora was curated manually from The African Plants Database (version 3.4.0)[290].

3.2.3 STRUCTURAL PRE-PROCESSING

All compounds were obtained from their respective sources in SD format. ChemAxon Standardizer[181] was used for structure canonicalization, transformation, and conversion of compounds in SD format into SMILES. To standardise the compounds in ChemAxon Standardizer, the following options were used: Clean 2D, Mesomerize, Neutralize, Remove Explicit Hydrogen and Remove Fragment. Duplicate structures in each dataset were removed, using ChemAxon JChem Software[181], "remove duplicates".

3.2.4 TARGET AND PATHWAY PREDICTION

See Section 2.2.4.

3.2.5 CLUSTERING

The matrix of predicted targets was clustered using *pvclust*[291] in R[182]. An approximately unbiased (AU) p-value of 0.95 was chosen. The AU p-value is calculated by multi-scale bootstrap re-sampling. For a cluster with AU p-value > 0.95, the hypothesis that "the cluster does not exist" is rejected with significance level 0.05; i.e. it can be assumed that these clusters do not only "seem to exist" caused by chance or sampling error, but can also be observed if we increase the number of observations.

3.3 RESULTS AND DISCUSSION

3.3.1 INTER- AND INTRA- FAMILY STRUCTURAL SIMILARITY OF NATURAL PRODUCTS

We first analysed whether NPs produced in each plant family are more similar to each other than those not from the same family, the result of which are shown in Figure 3:2. Looking at Figure 3:2, we can see that in structural space NPs from the same family share the same structural space. This is important as it indicates that when plants from a family are used for a medicinal indication, then NPs from other genera and species in that family can be bio-screened, as they will likely occupy the same structural space. A few examples have been highlighted in the figure, e.g. NPs from Anacardiaceae (light blue) and Chenopodiaceae (black). In the case of the NPs from Chenopodiaceae it is interesting to note that they are in the periphery of the graph, due to their unusual chemistry, and all other compounds from that family are clustered in the same area.

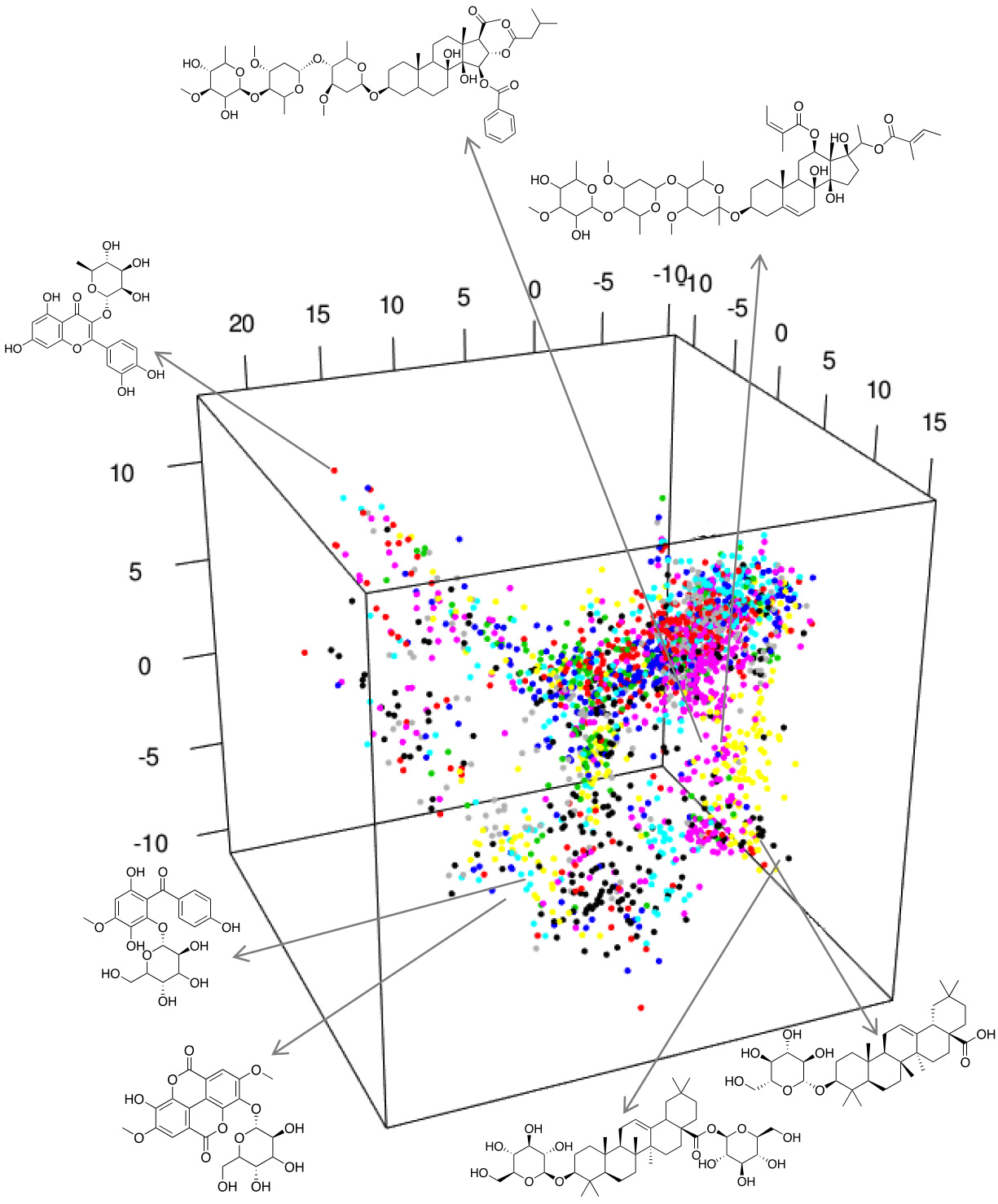


Figure 3:2 MDS plot of tanimoto similarity between NPs using their Morgan fingerprints. NPs are coloured based on plant family. Some NPS from the same families are similar to each other and are clustered together in space, whereas others are more diverse and spread out in space. All metabolites in the dataset were used.

Quantifying the results will allow us to draw conclusions about the similarity of natural products within a plant family and will aid the hypothesis of the next section of identifying patterns in TAM. To this end, we carried out a quantitative chemical similarity analysis to better represent this information, the results of which are shown in Figure 3:3 and Figure 3:4. The median Tanimoto similarity between compounds in each family is much higher, ranging from 0.11 to 0.47 within the same family, but a maximum median of only 0.15 for the random samples. A Tanimoto coefficient of 0.3 between compound pairs is generally accepted to indicate that compounds are structurally similar[176]. In ChEMBL, 95% of compounds had Tc > 0.424 to their active nearest neighbours[143]. These studies increase our confidence in the conclusion natural products are more similar within a family than would be expected by chance.

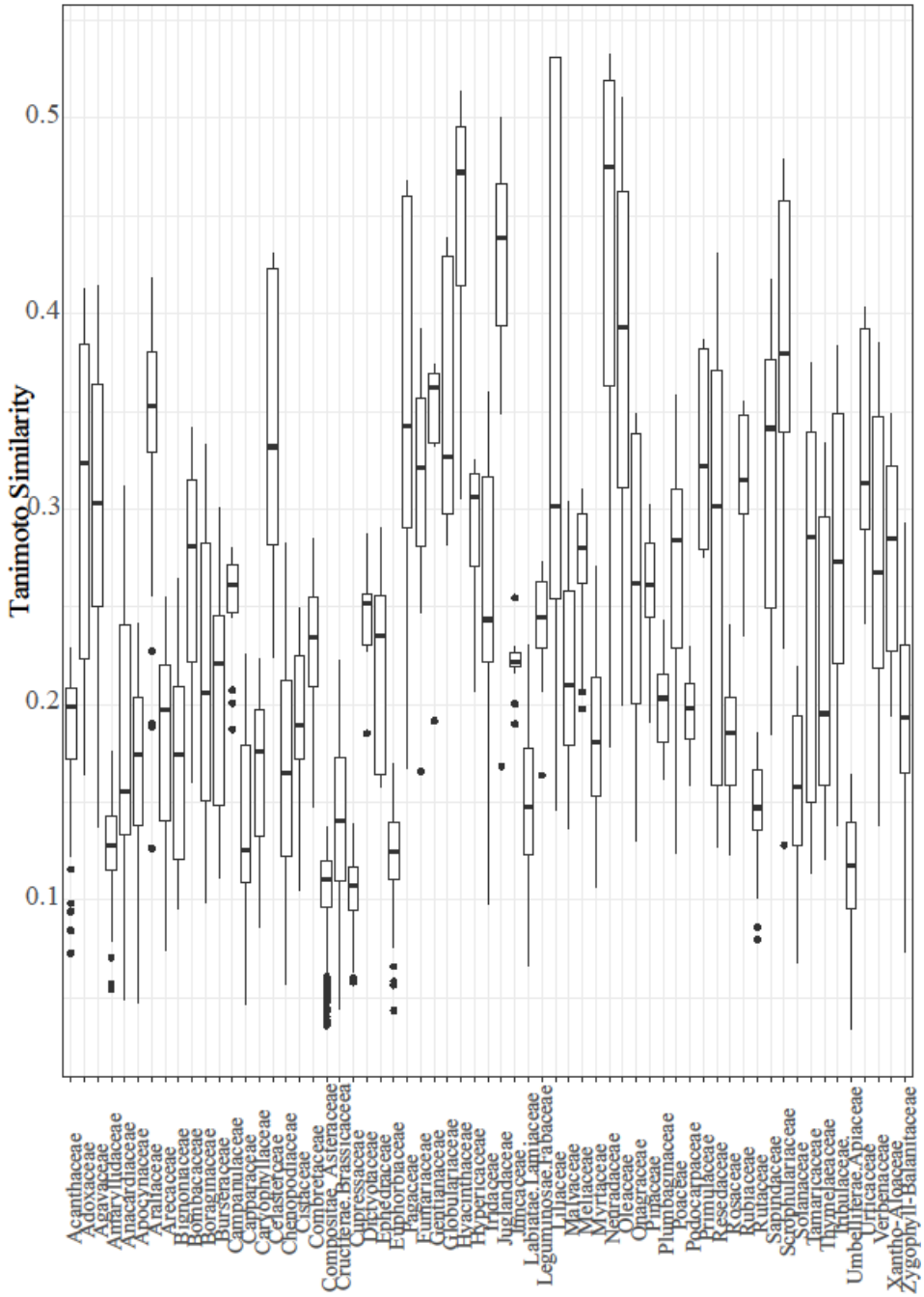


Figure 3:3 Average Tanimoto similarity of the compounds produced by the family to each other. The y-axis displays the average Tanimoto similarity of compounds within in each family to other compounds in that Family. From this figure we see that the median of average similarity ranges from 0.11 to 0.47. This represents NPs with high structural similarity within each plant.

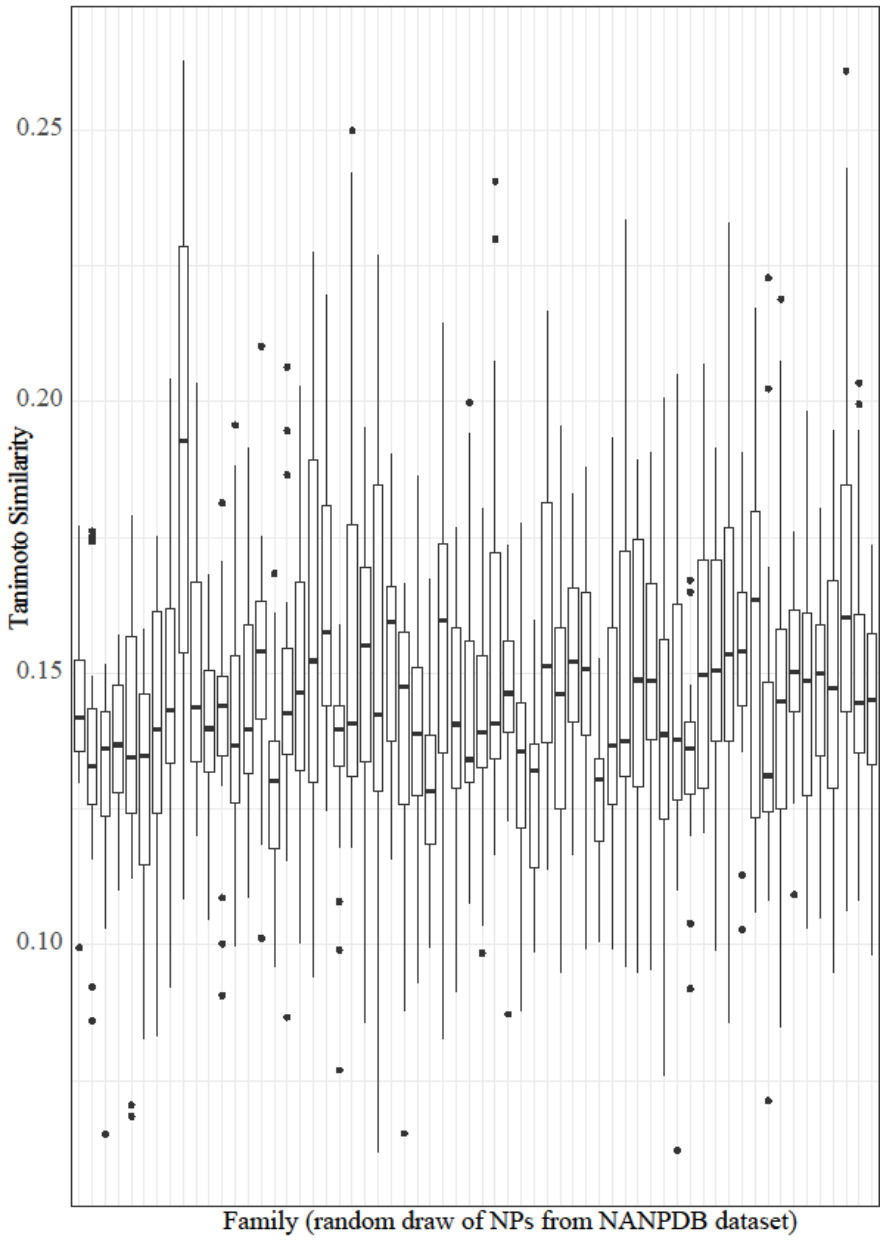


Figure 3:4 Here random selections of 25 NPs were drawn without replacement from the NANPDB and repeated 63 times (number of families studied), to represent a set of randomised plant families. Each boxplot represents a different randomised family. The y-axis is the average Tanimoto similarity of each NP in this specific randomised family to all other NPs in the other randomised families. From this figure we see that the median of average similarity ranges from 0.11 to 0.15. This represents NPs with low structural similarity within a group of random NPs representing a family.

We next investigated the bias of families with a higher number of NPs having lower Tanimoto coefficient scores and vice versa. We suspected that families with a high number of NPs compared to other families would intrinsically have lower Tanimoto similarity scores. This is because by chance the more NPs extracted and studied from a

plant family, the more structurally diverse they would be, since they would belong to different phytochemical classes. This turned out to not always be the case as can be seen from Figure 3:5. We found that in some cases families produced up to 75 NPs (The majority of the families contained around 25 compounds) and the similarity score was still above 0.3. When this was not the case, it was found that the families contained more than one phytochemical class with little structural similarity to each other e.g. Euphorbiaceae contains 295 natural products that included but were not limited to monoterpenes, triterpenoids, coumarins, lignans and flavones. Similarly, Umbelliferae-Apiaceae, contains 389 NPs that fall into several phytochemical groups including long chain unsaturated hydrocarbons, monoterpenes, sesquiterpenes and furocoumarins. The large numbers of NPs having different chemical structures within these families contributes to the low structural similarity between the natural products.

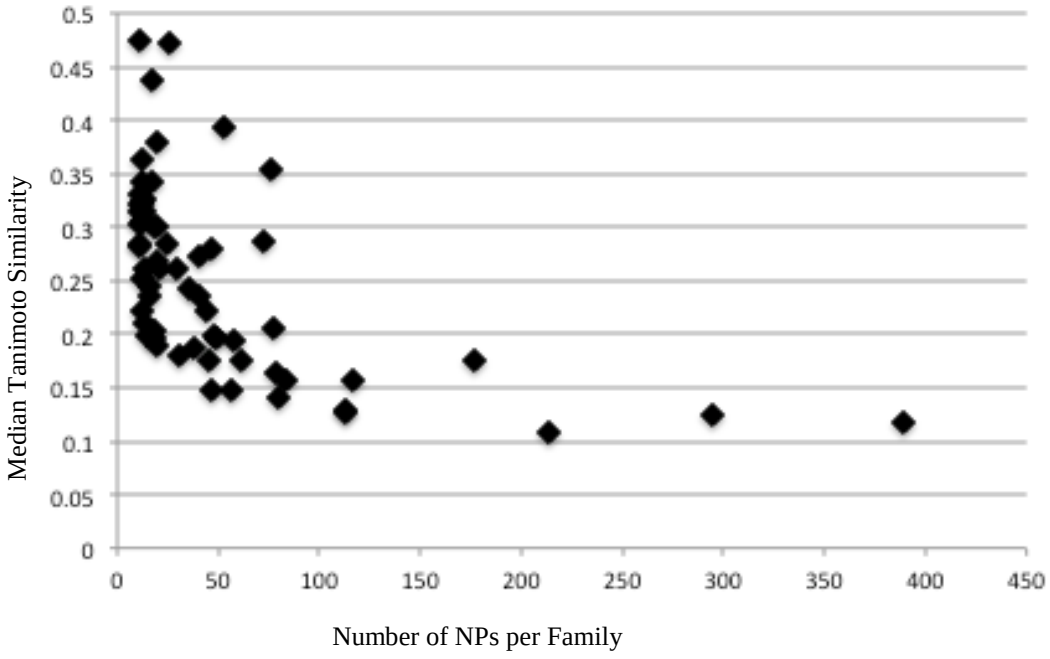


Figure 3:5 Number of NPs per family versus the mean Tanimoto similarity of NPs in each family. For most cases, the similarity was lowest when the number of NPs was above average (average number of compound per family is 74.25). However, in some cases, the Tanimoto similarity remained high despite there being 75 compounds in the family.

In this section we have shown that NPs in a plant family are structurally more similar to each other than to NPs produced in other plant families.

3.3.2 Role of Geographic Origin of Plant on Structure and Predicted Activity of NPs

Plant families with species that display anti-cancer activity are found in different parts of the world. To investigate whether use of a plant family in one region of the world could help inform use of plants from the same or closely related families in other areas of the world we examined whether the increased level of Tanimoto similarity between natural products (NPs) within a family, shown in the previous section, is also found for NPs produced by members of the same plant family that are found in different geographic areas. A global study of the African, Malay and Ayurvadic (AMA) dataset was carried out to identify relationships between the chemical structure of NPs, their predicted activity and the phylogeny of the plants they are produced by, taking into account information about their geographic origin.

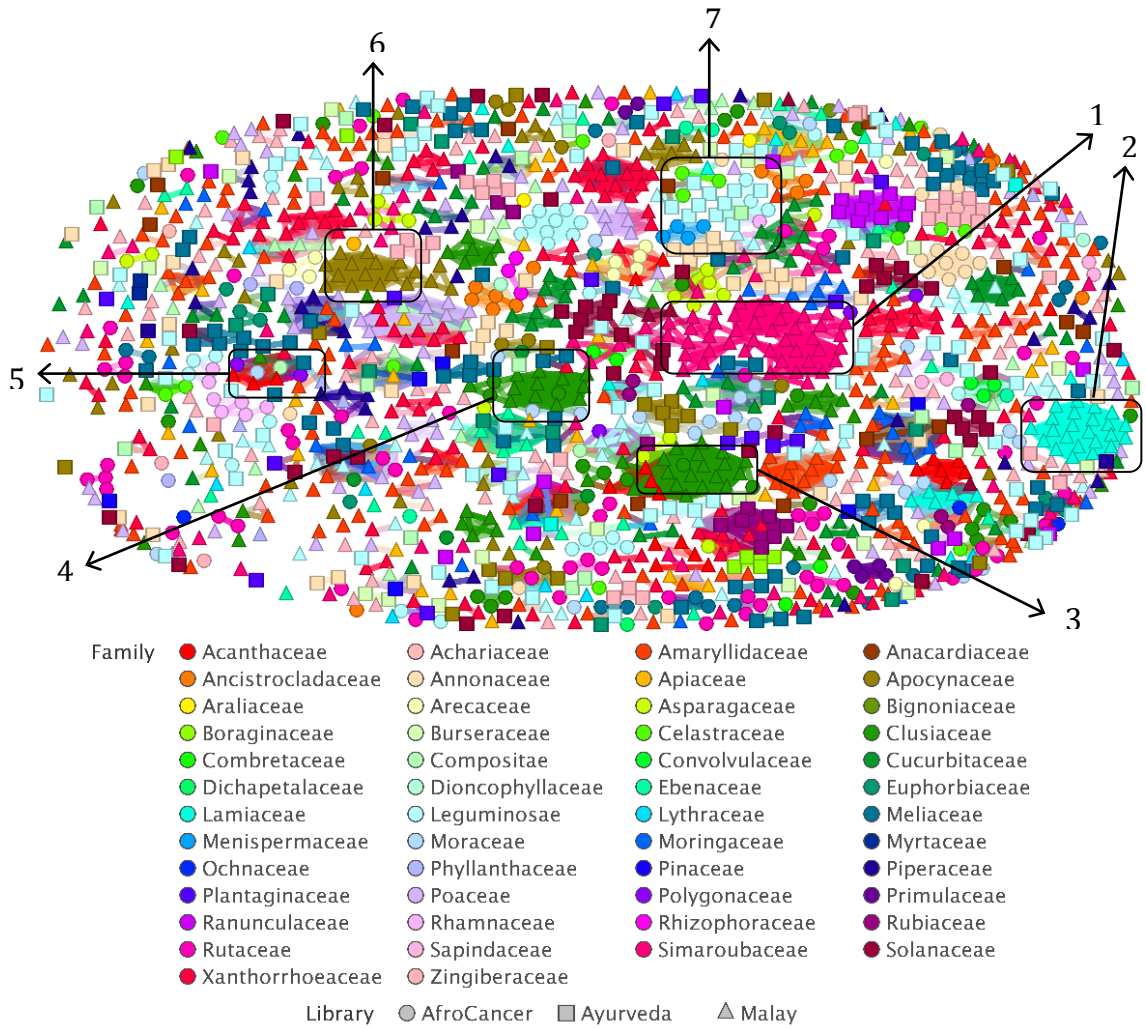


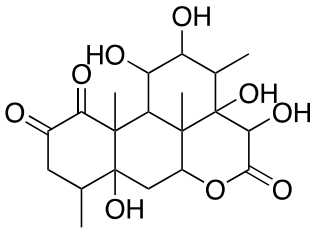
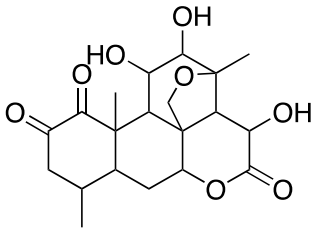
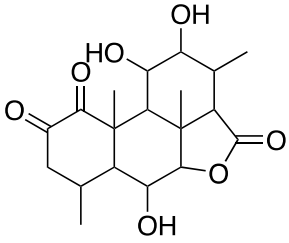
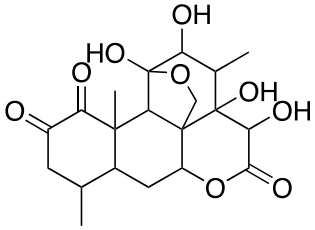
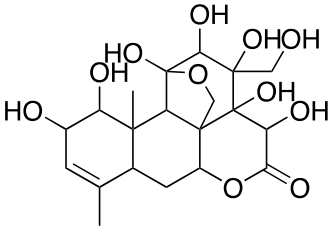
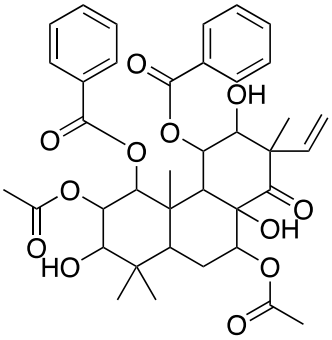
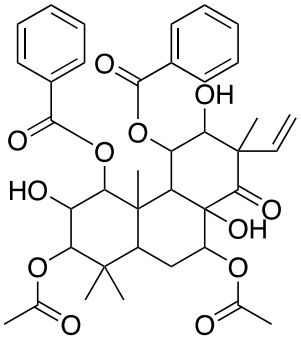
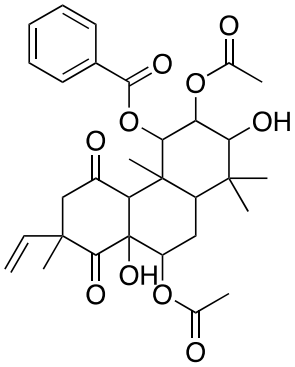
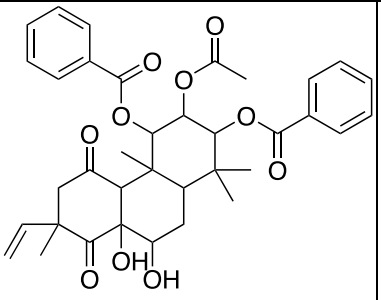
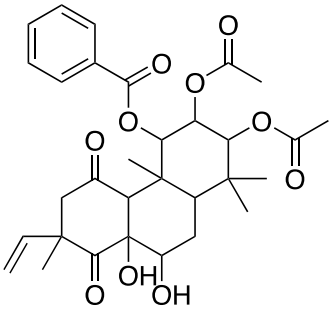
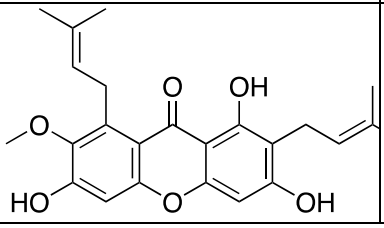
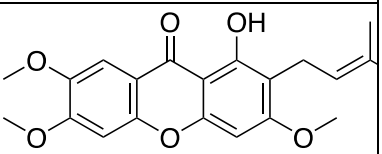
Figure 3:6 Structural similarities of compounds from families from AfroCancer, Malay and Ayurveda libraries. Each data-point represents a compound. The compounds are clustered by their structural similarity calculated using Tanimoto coefficients and coloured according to the family of the plant they come from. Similar compounds are clustered together, with connecting lines drawn between those having 95% or higher structural similarity. This figure shows that compounds within a family are structurally similar e.g. compounds from Simaroubaceae (cluster 1) are grouped together as are those from Lamiaceae (cluster 2). Clusters 3 and 4, contain similar compounds from Clusiaceae, while clusters 6 and 7 contain similar compounds from Apocynaceae and Leguminoseae respectively. On the other hand, cluster 5 contains compounds that are similar in structure to each other but come from more than 5 families. This information can on the one hand be used to suggest novel indications for particular plants, and on the other hand to improve the mode of action prediction of compounds from particular biological species.

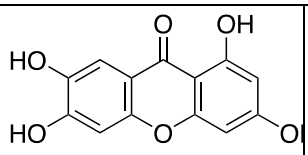
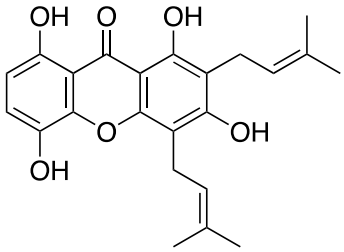
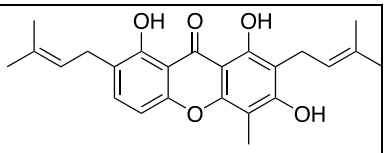
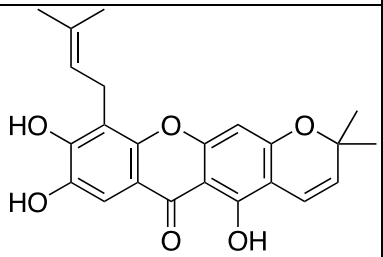
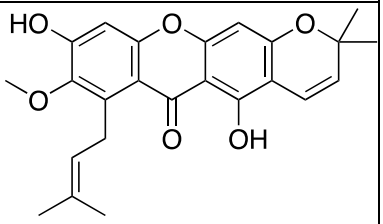
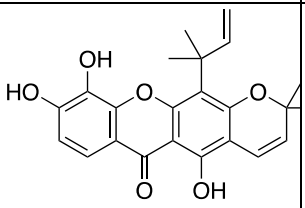
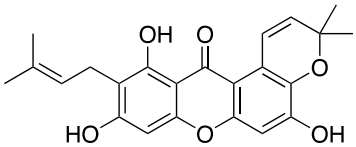
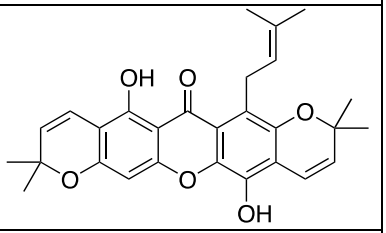
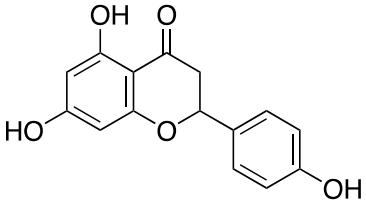
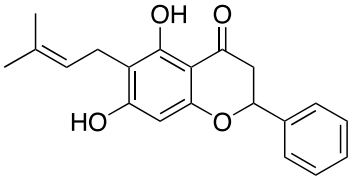
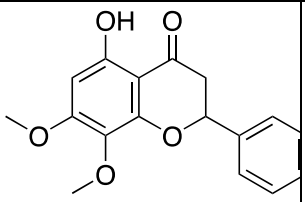
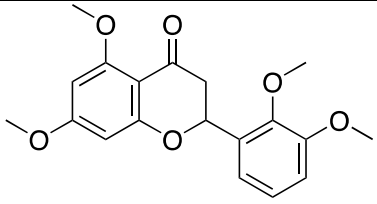
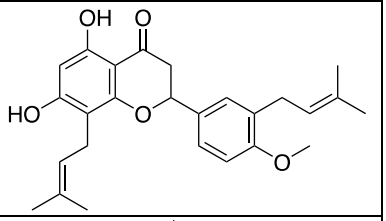
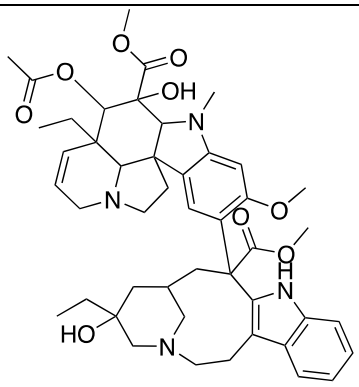
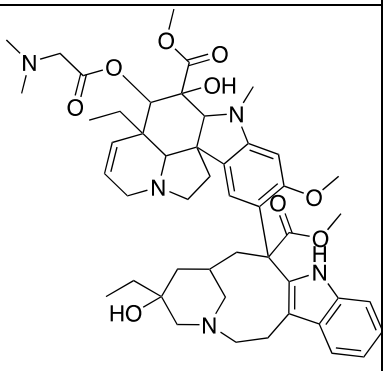
From the plot in Figure 3:6 we see that NPs tend to be grouped together and clustered by family regardless of the geographic region from which they originated, e.g. the Clusiaceae compounds in clusters 3 and 4 from both the AfroCancer and Malay datasets and the Leguminoseae compounds from AfroCancer, Malay and Ayurveda (cluster 7). We also see clusters of compounds that are similar in more than one family across all three regions, e.g. cluster 5 (structures shown in Table 2). Previously, it has been shown

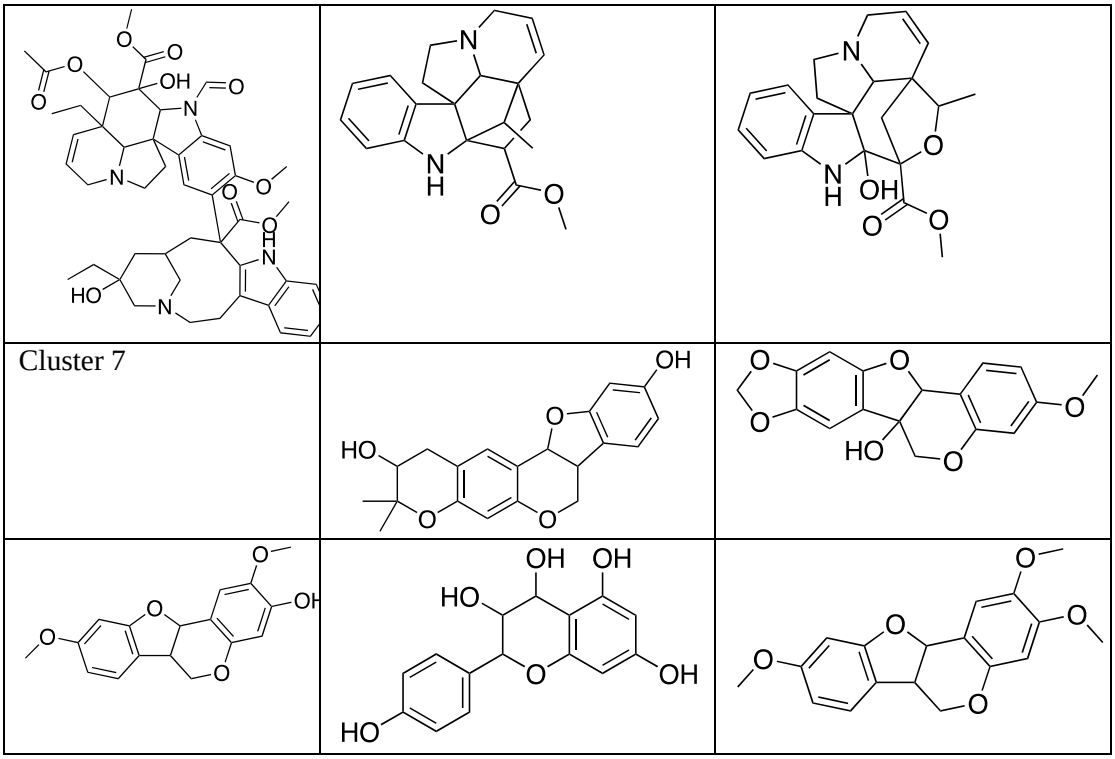
that compounds with Tc ≥ 0.3 display similar bioactivities[176]. Thus, compounds in the AMA dataset that are clustered together in Figure 3:6 are expected to display similar bioactivity profiles across the three studied regions. This finding agrees with previous studies that show that use of a medicinal plant family in one region of the world can predict activity of that family in another region[145].

The following Table 3:1 shows a comparison between the similarity of compounds in clusters, and those in the periphery of the 2D RBS[191] plot. The results from Figure 3:6 and Table 3:1 are discussed below.

Table 3:1 Clusters identified by the 2D RBS plot and the structures of the compounds within those clusters. This is not a comprehensive cluster list, but an illustration of structural similarities within clusters with similar compounds identified by connecting lines when Tanimoto similarity of ECFP4 fingerprints is over 0.95.

Cluster 1		
		
Cluster 2		
		
Cluster 3		

		
Cluster 4		
		
Cluster 5		
		
Cluster 6		



The alkaloids from *Catharanthus roseaus* (Family Apocynaceae) are grouped together with opaque connecting lines (Cluster 6 -Figure 3:6). These alkaloids exhibit anti-cancer activity by preventing cell division, eventually leading to apoptosis. They do this by binding at the "vinca domain" of the β-subunit of the tubulin protein thus disrupting microtubule assembly and preventing cell division during the metaphase stage of the cell cycle[292, 293]. The Simaroubaceae quassinoids (Cluster 1) are also closely clustered. They display chemo-protective activities through their inhibition of the carcinogenic CYP1A1 enzyme[294] and cytotoxicity of Simalikalactone D[295]. The anti-tumour properties of the Simaroubaceae quassinoids in our dataset display mainly anti-leukemic activity[296]. It is important to note that these compounds have not been taken into further stages of drug discovery due to their toxicity. This clustering is also true for many families whose plants are used traditionally for cancer but for which there is currently no clinical evidence of activity, e.g. Clusiaceaae NPs and their anti-inflammatory activity[297]. This clustering of families containing NPs that display similar activities also has the potential to help direct selection of isolated compounds for screening via the following workflow: if a compound falls within a cluster whose activity is previously known (traditionally or through *in vivo* and *in vitro* experiments), then it would be prioritised for screening. Alternatively, if a compound is found to be

too toxic, other compounds from the cluster can be screened to identify non-toxic forms. This suggests that the distance between NPs with known bioactivity and new NPs in chemical space can inform subsequent screening and characterization.

3.3.2.1 ROLE OF PHYLOGENY OF PLANT IN OBSERVED NP SIMILARITY IN PLANT FAMILIES

In the second part of this analysis, we aimed to investigate the relationship between plant phylogeny, NP structure and the predicted protein target activity of NPs from the AMA dataset. We proceed by analysing the clustering patterns of plant families that have been clustered according to both the Tanimoto similarity of the ECFP4 fingerprints of their NPs, and also the Tanimoto similarity of the corresponding sets of predicted targets (Figure 3:7 and Figure 3:8). The results in these figures allow us to analyse whether plants that cluster together in chemical space (with similarity defined according to their structural fingerprints) are either (i) also clustered together in the space of predicted targets as found in, e.g. the cluster containing Leguminoseae and Zingiberaceae as well as the cluster containing Ebenaceae, Rutaceae and Acanthaceae, or (ii) found in distinct clusters in predicted target space, e.g. Simaroubaceae, Paceae and Apocynaceae, which cluster together in structure space but not in predicted target space.

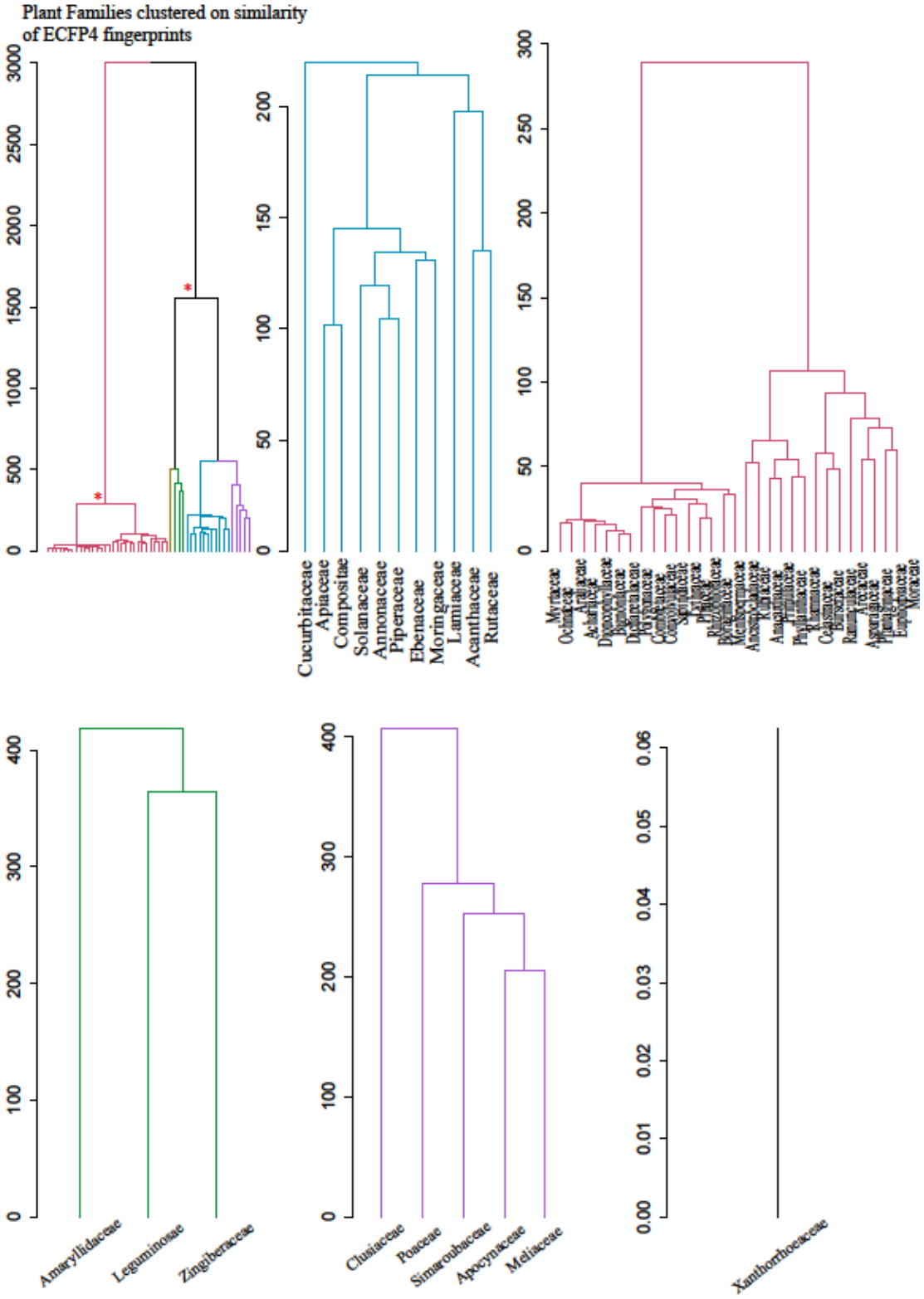


Figure 3:7 Plant families clustered according to the similarity of the ECFP4 fingerprints of the NPs that have been isolated from them. The vertical branch lengths are arbitrary and represent distances between tips. Here we set a significance threshold for cluster existence. The start on the top left clustergram is for a cluster with AU p-value > 0.95, where the hypothesis that "the cluster does not exist" is rejected with significance level 0.05.

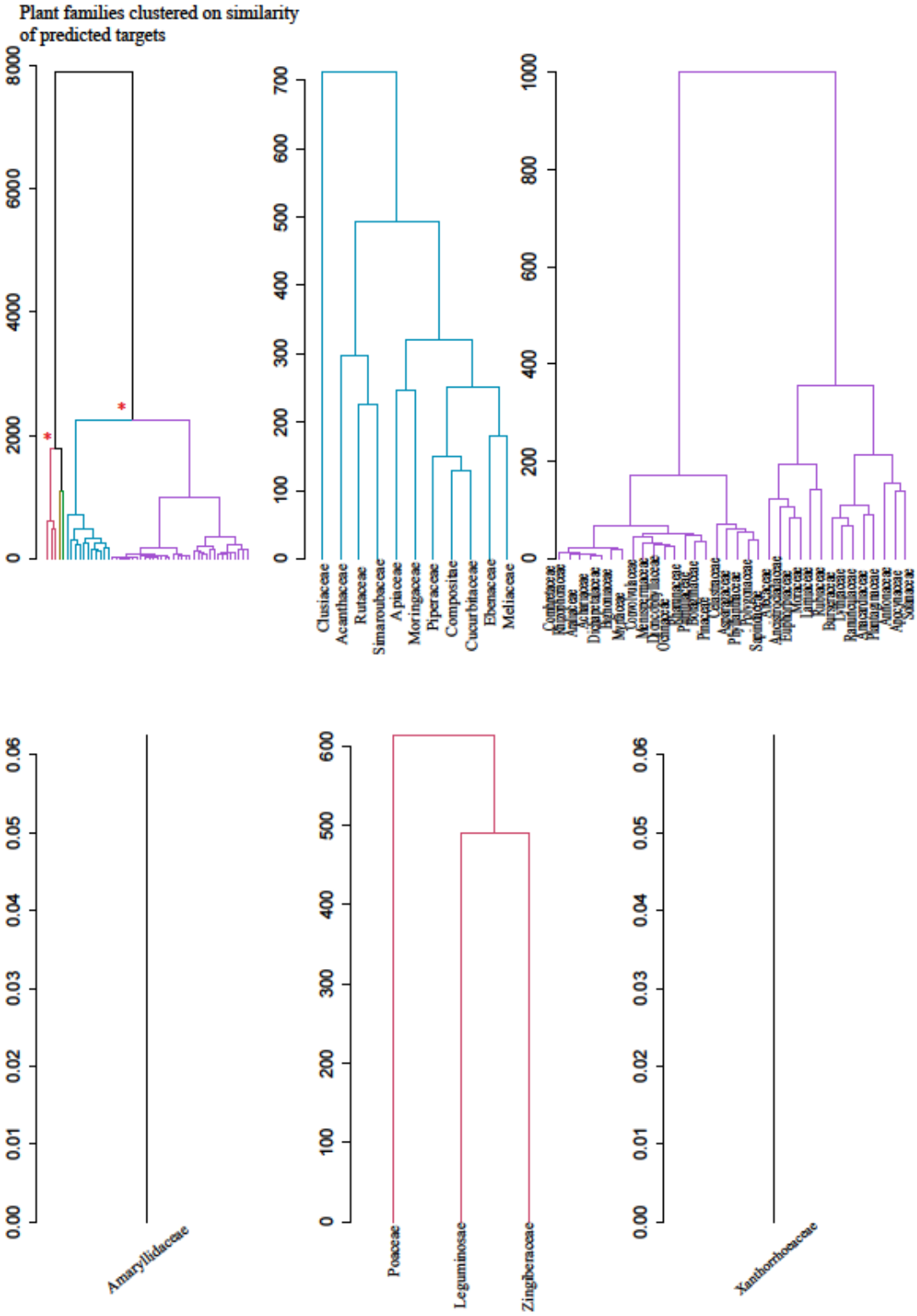


Figure 3:8 Plant families clustered according to the similarity of the predicted targets of compounds that have been isolated from them. The vertical branch lengths are arbitrary and represent distances between tips. i.e. the longer the branch, the more the distance is between the daughter clusters. Here we set a significance threshold for cluster existence. The stars on the top left clustergram is for a cluster with AU p-value > 0.95, where the hypothesis that "the cluster does not exist" is rejected with significance level 0.05.

From Figure 3:7 and Figure 3:8, we can see that Xanthorrhoaeceae does not cluster with any of the other families in either structure or predicted target space. Amaryllidaceae is also in its own cluster for predicted target space, and clusters with the large Leguminoseae and Zingiberaceae families in structure space. Clusiaceae, Simaroubaceae and Meliaceae cluster together in both predicted target space and structure space, as do, for example, Acanthaceae, Rutaceae, Apiaceae, Moringaceae, Piperacee, Ebenaceae, Compoisitae and Cucurbitaceae. Some plant families that do not cluster together in predicted target space and structure space include Solanaceae, Annonaceae, Lamiaceae and Meliaceae. Figure 3:9 shows the plant families coloured in their respective clusters and projected onto the phylogenetic tree of plants, to show the relative position of plants on the tree in relation to the predicted targets and structure of the NPs they contain. We can see that some clusters are localised to specific parts in the tree (indicated in Figure 3:9) leading to the conclusion that for this dataset, with the currently available information, these plants contain chemistries not found in other unrelated plants, e.g. the cluster containing Asteraceae and Apiaceae.

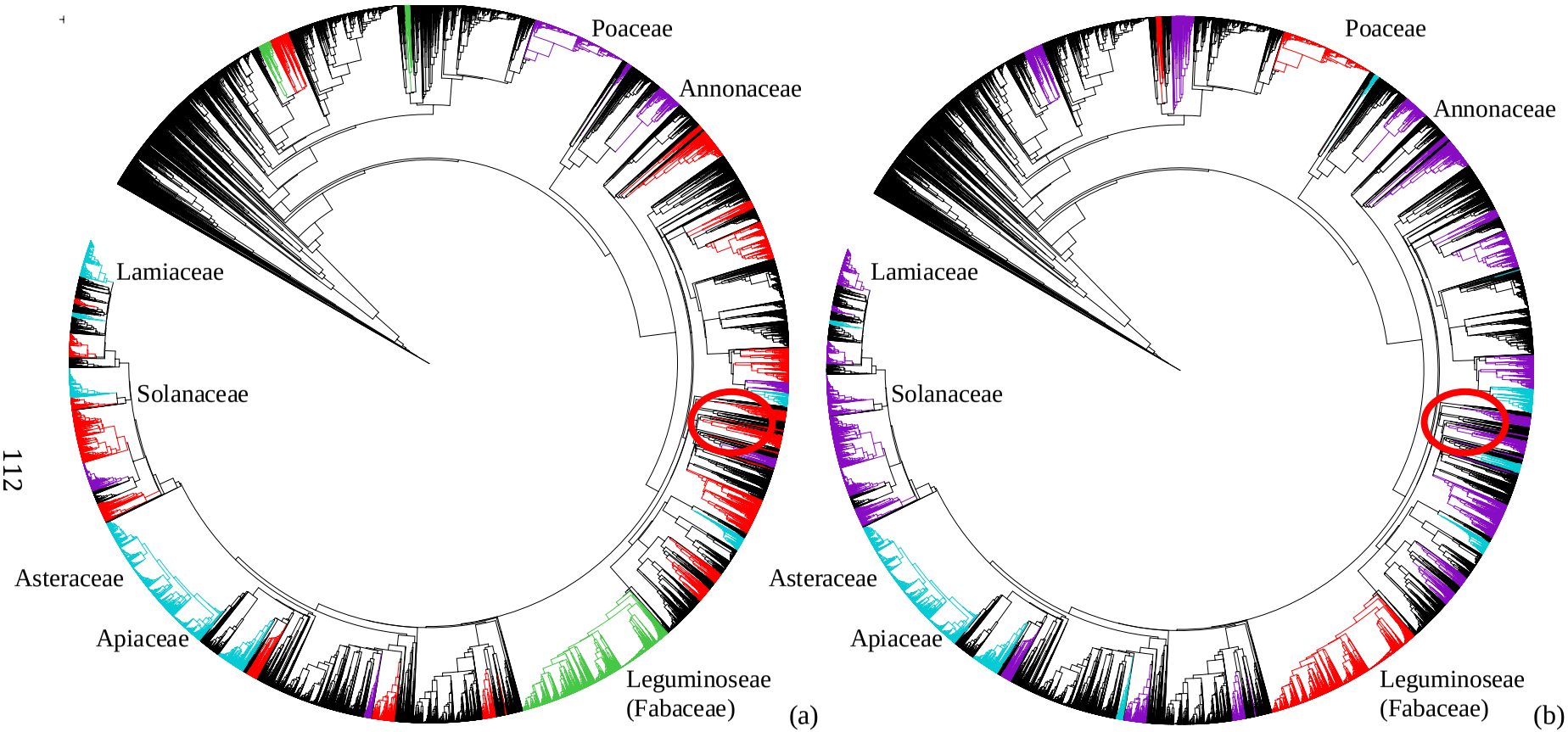


Figure 3:9 Phylogenetic tree of land plants showing the position of the clusters obtained from the clustering of the NPs in structure space (a) and predicted target space (b). For each tree, plant families in the same cluster are shown in the same colour. The plant families are coloured according to the clusters that they are found in from **Figure 3:7** for structural space and **Figure 3:8** for predicted target space. Asteraceae and Apiaceae are an example of families clustering together in target space and structure space. Families shown in the red circle, Ochnaceae, Phyllanthaceae and Dichapetalaceae, provide another example of families that cluster together in predicted target space, structure space and phylogeny space. On the other hand, Solanaceae, Annonaceae and Lamiaceae provide an example of families that cluster together in structural space but not predicted target space.

We next analyse, and incorporate into this analysis, the phylogenetic distances (as measured by branch lengths between families in divergence times of million years ago (mya)) between plant families and how this relates to both the structures of the NPs produced by each family, and the corresponding sets of predicted targets. This is inspired by work carried out by Liu *et al* [298] which classified plants based on their metabolite content. Despite working with incomplete data, they found that clustering plants according to their metabolite content produced clusters consistent with known evolutionary relations of the plant. Here we analyse the plant families in detail. In (Table 1 – CD) we report the patristic distances between all pairs of the plant families in the AMA dataset, calculated using the Zanne[162] tree of land plants, and we report the patristic distances between the plants discussed here in Table 3:2. The distances in (Table1 - CD) are further clustered in Figure 3:10, revealing the presence of several distinct phylogenetic clusters within this set of studied families.

The first cluster contains Rutaceae, Simaroubaceae and Meliaceae; another cluster contains Rhizophoraceae, clusiaceae and Achariaceae; a further cluster contains Ochnaceae Dichapetalaceae and Phyllanthaceae, with two additional clusters containing Anacardiaceae, Sapindaceae and Burseraceae in the first and Convuluvlaceae, Plantaginaceae, Rubiaceae, Acanthaceae and Solanacea in the second. For all the families in these phylogenetic clusters, we find they also cluster in predicted target space and structure space. Therefore, we conclude that among the set of studied families, plant families that have both similar compound structures and similar predicted targets tend to be closer together on the phylogenetic tree (see Figure 3:9 and Figure 3:10). In contrast, those families that have similar compound structures but different sets of predicted targets tend to be further apart on the phylogenetic tree.

A closer look at the phytochemical classes responsible for anti-cancer activity reveals a putative explanation for the differences and similarities in clustering. We find that Apocynaceae, Simaroubaceae and Meliaceae are clustered together according to structural similarity but not predicted activity. The high level of structural similarity can be explained by the phytochemistry of these plant families. Despite this, the families do not all cluster together in predicted target space: Simaroubaceae and Meliaceae do cluster together, with a divergence time of 42.82 (mya), as shown in

Figure 3:10; however, in the phylogenetic tree, Apocynaceae lies further away from Simaroubaceae and Meliaceae (see Figure 3:10). We find that Apocynacea and Meliaceae diverged 362.31 (mya), and Apocynaceae and Simaroubaceae 368.04 (mya). This observation likely reflects the fact that despite the structural similarity between the alkaloids of Apocynanceae and the terpenoids of Simaroubaceae and Meliaceae, they exert their activity via different mechanisms of action. The alkaloids inhibit cell division by interacting with tubulin and topoisomerase II[19, 292, 299], whereas the terpenoids and limonoids have been found to inhibit NF-κB [20, 300]. In accord with these reported findings, in our dataset 58 out of 89 compounds (30.7%) from Simaroubaceae and Meliaceae, but only 3 out of 81 (3.7%) of the compounds from Apocynaceae, were predicted to modulate NF-κB1. In contrast, 22% of Apocynaceae compounds but just 4.7% Simaroubaceae and Meliaceae compounds were predicted to modulate tubulin α-1B chain, and 19 out of 81 (23.5%) of Apocynaceae and 10 out of 189 Simaroubaceae and Meliaceae compounds (5.3%) were predicted to bind tubulin α-3C/D chain.

This finding of NPs from Meliaceae and Simaroubaceae acting via different mechanisms of action to the NPs in the phylogenetically distant Apocynaceae, despite having similar structures, supports the hypothesis that plant families phylogenetically further away from each other on the tree are predicted to act by modulating different targets.

Table 3:2 The phylogenetic distances between the families discussed above. The numbers represent the pairwise distances between each node, calculated by summing up the branch lengths (divergence times in mya) to their MRCA. The boxes are coloured in a heatmap fashion corresponding to the distances to the distances. The closer nodes are red, whereas those further away are yellow and green.

	Zingiberaceae	Poaceae	Amaryllidaceae	Xanthorrhoeaceae	Apocynaceae	Meliaceae	Simaroubaceae	Calastraceae	Burseraceae	Ranunculaceae
Zingiberaceae	0	229.1754	336.9787	371.3506	216.292	362.3557	368.0833	359.4797	342.1194	226.9202
Poaceae	229.1754	0	338.0009	372.3728	229.128	363.3779	369.1055	360.5019	343.1416	223.0059
Amaryllidaceae	336.9787	338.0009	0	357.2506	336.9313	348.2557	353.9833	345.3797	328.0194	335.7457
Xanthorrhoeaceae	371.3506	372.3728	357.2506	0	371.3032	260.4819	266.2095	257.6059	240.2456	370.1175
Apocynaceae	216.292	229.128	336.9313	371.3032	0	362.3083	368.036	359.4323	342.072	226.8728
Meliaceae	362.3557	363.3779	348.2557	260.4819	362.3083	0	42.81856	141.6724	209.2483	361.1227
Simaroubaceae	368.0833	369.1055	353.9833	266.2095	368.036	42.81856	0	147.4	214.9759	366.8503
Calastraceae	359.4797	360.5019	345.3797	257.6059	359.4323	141.6724	147.4	0	206.3722	358.2466
Burseraceae	342.1194	343.1416	328.0194	240.2456	342.072	209.2483	214.9759	206.3722	0	340.8864
Ranunculaceae	226.9202	223.0059	335.7457	370.1175	226.8728	361.1227	366.8503	358.2466	340.8864	0

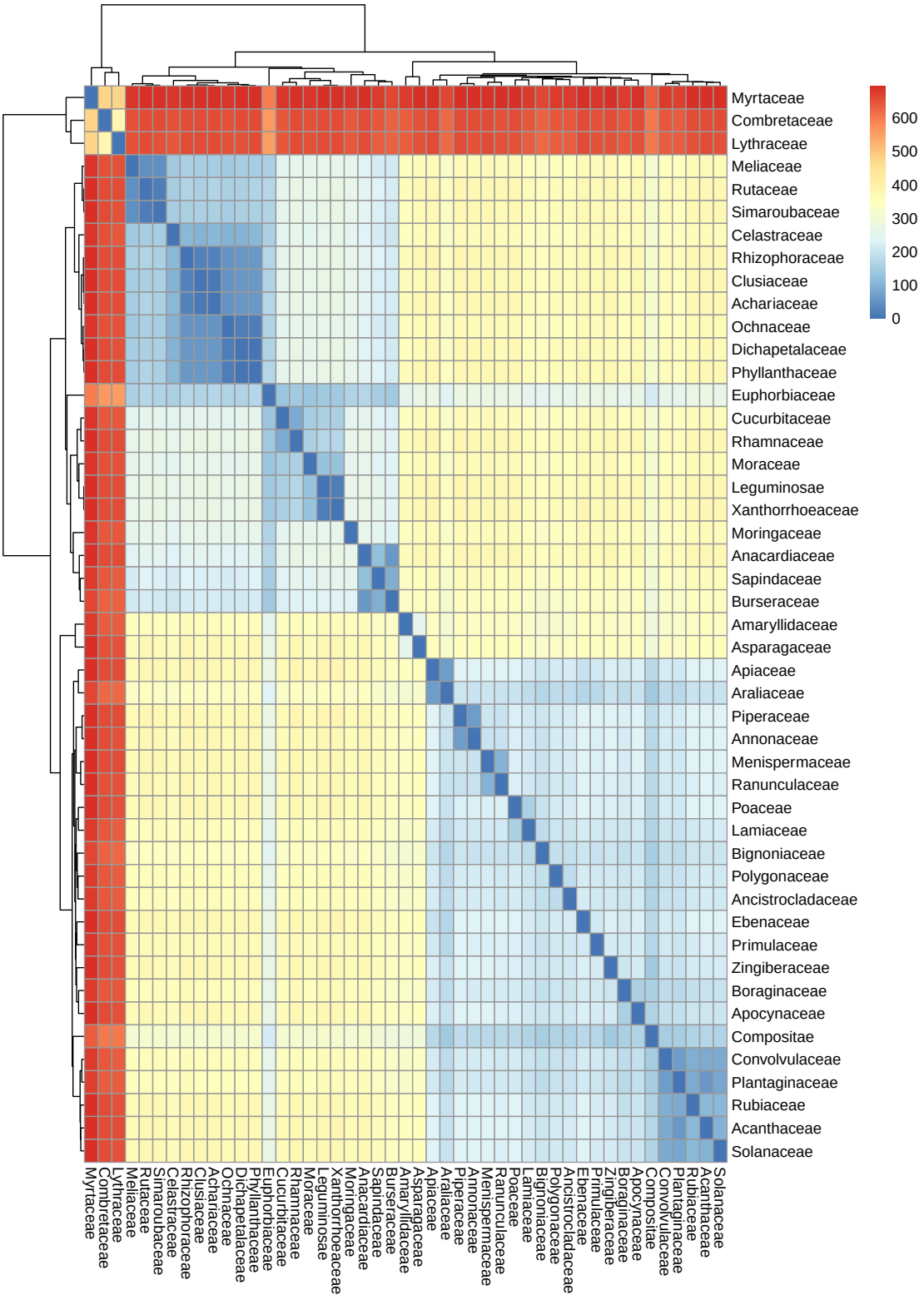


Figure 3:10 Heatmap of the families discussed above; clustered by the phylogenetic distances between them. This heatmap shows that the distance between families in the phylogenetic tree can be represented as clusters e.g. the Rutaceae and Simaroubaceae cluster. The dark blue represents plant families that are close together phylogenetically, while those in yellow and red are further away phylogenetically. The numbers on the coloured bar represent divergence in millions of years ago.

This analysis supports the hypothesis that the position of a plant on the phylogenetic tree, relative to other plants, can help predict the activity of its natural products. When plants are close phylogenetically, they produce similar compounds that are predicted to act via similar mechanisms of action, as shown by the clustering in Figure 3:7 and Figure 3:8. When they are further away phylogenetically, but produce structurally similar compounds, they are predicted to act via different mechanisms of action. Common processes that occur in different plant species that need to be carried out require secondary metabolites to be adapted to the local environment. Our findings suggest that these NPs tend to still be structurally similar, while having different mechanisms of action. This wide range of chemical space likely gives rise to the different predicted mechanisms of action displayed by the AMA plants.

3.3.3 Cross-cultural patterns in medicinal floras with traditional anti-cancer activity

In this section we aim to detect plant families whose members are used significantly more often in medicinal flora than would be expected by random draw, i.e. families that are over-represented in traditional use. First, we wanted to show that NPs within a family in the AMA dataset are structurally similar to each other. This would mean that when a plant family is identified as over-utilised, then species in that family are prioritised for screening because they would produce structurally similar NPs with similar activities. In Figure 3:3 we showed that for the NANPDB dataset, species within a family tend to produce more similar NPs to each other than would be expected at random. The AMA dataset is analysed in detail in Figure 3:6, where NPs produced by species within the same family also tend to cluster in structure space. To analyse this in more detail, Figure 3:11 displays boxplots that illustrate the Tanimoto similarity of natural products within each family of the AMA dataset, and Figure 3:12 shows the median Tanimoto similarity of NPs within a family as a function of the number of NPs in a plant family.

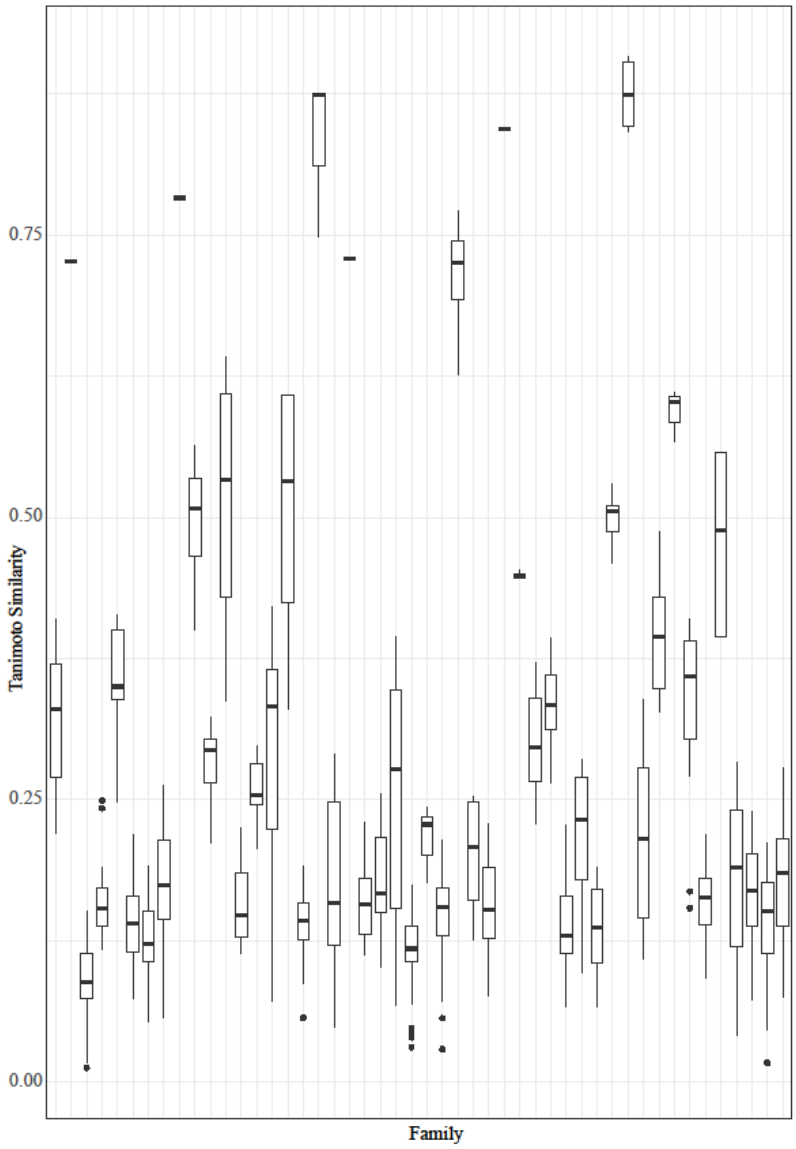


Figure 3:11 Tanimoto similarity of NPs to each other in each family. Each boxplot represents a plant family. The y-axis is the average Tanimoto similarity of the compounds produced by the family to each other. From this figure we see that the median of average similarity of compounds in the AMA dataset ranges from 0.08 to 0.88. This plot shows that NPs within some families have very high Tanimoto similarities to each other, compared to those NPs and families analysed above in the NANPDB dataset, where NPs within a family showed a median similarity between 0.11 and 0.47, and the random draws only showed median similarity between 0.11 - 0.15.

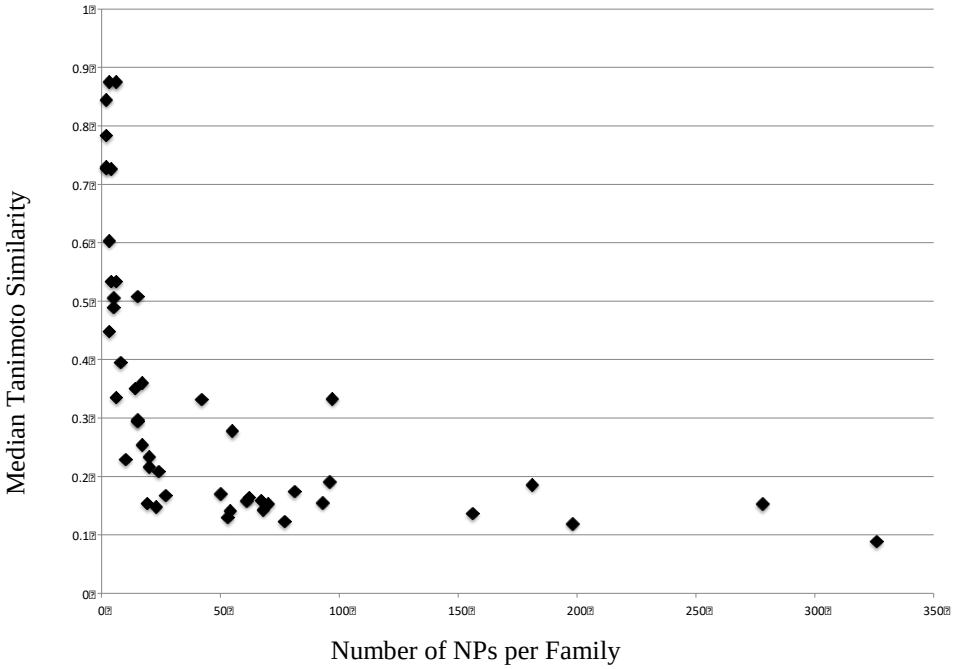


Figure 3:12 Number of NPs per family versus the mean Tanimoto similarity of NPs in each family. For most cases, the similarity was lower (less than 0.3) when the number of NPs was above average (average number of compound per family is 51.29). However, in some cases, the Tanimoto similarity remained high (0.33) despite there being 97 compounds in the family.

The results show that NPs within a plant family are structurally similar to each other (Tanimoto coefficient up to 0.88), e.g. Primulaceae contains 6 NPs with median Tanimoto similarity of 0.87 to each other and Convulvulacea contains only 3 NPs that have a median similarity of 0.87. We also see that the greater the number of NPs per family, the lower the similarity, e.g. the 326 NPs in Amaryllidaceae and 278 NPs in Xanthorrhoeaceae share a median Tc of only 0.09 and 0.15 respectively. This is not always the case, as can be seen for the family Clusiaceae. This family contains 97 NPs which have a median Tanimoto similarity of 0.33 to each other. Here we have shown that NPs in a family are somewhat structurally similar to each other.

Hot Nodes

Previous studies[147, 284, 301], have looked at the over-representation of medicinal plants in different flora across different continents by identifying 'hot nodes" in plant lineages, in order to guide medicinal chemists' choice of lineages to pursue for drug discovery. This previous work has focused on ethno-botanical studies rather than chemistry. Moreover, this type of study has not been carried out for African medicinal flora or for

medicinal flora from different parts of the world used against cancer. To address these questions, we compiled and curated a completely novel dataset of 51 plant families (129 species, see Methods), that have been reported to have activity and are used against cancer in Africa, Malaysia and India (the AMA dataset). We then analysed this dataset to identify phylogenetic lineages in which anti-cancer properties are over-represented. Figure 3:13 shows the phylogenetic tree of land plants, with the clades descending from the hot nodes coloured in red. In this case we hypothesise that if a node is identified, its descendants (terminal taxa) are more likely to belong to the "medicinal use" group (in this case against cancer) than would be expected by random chance.

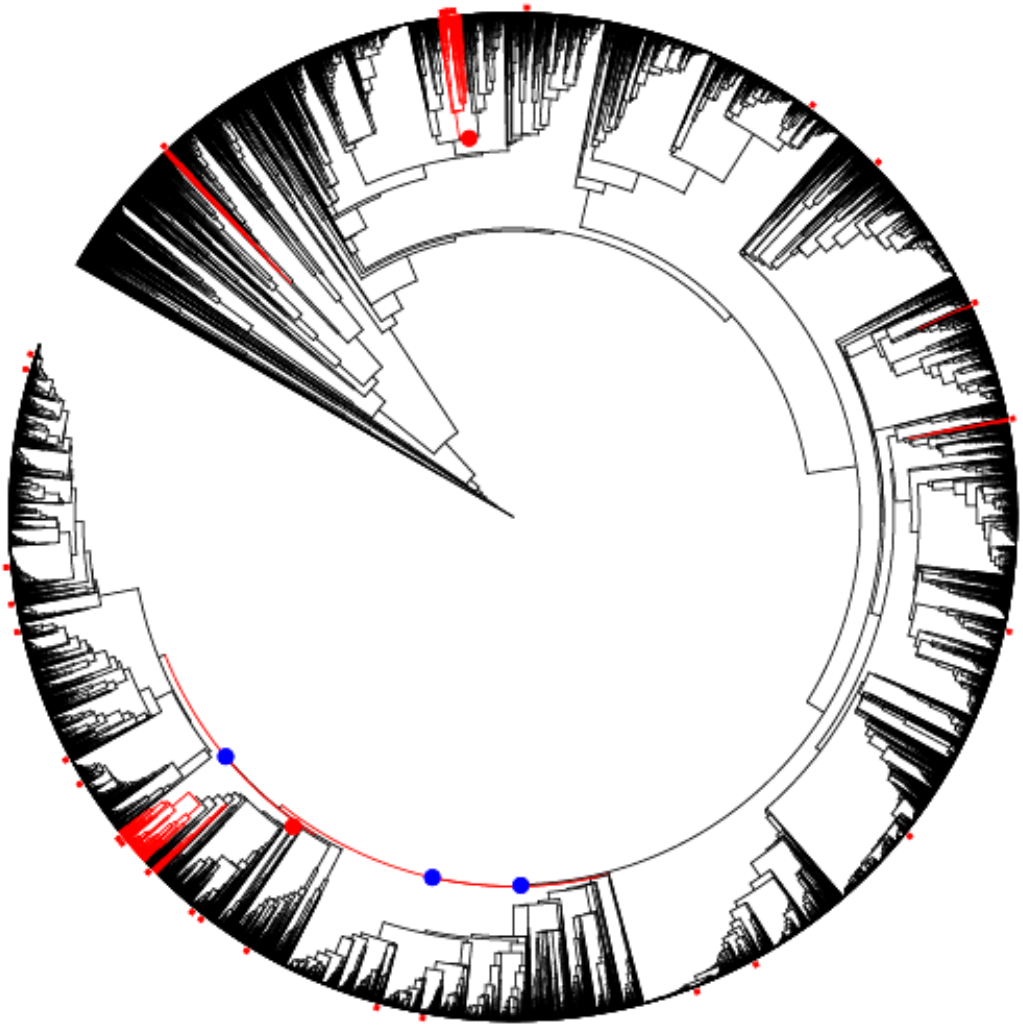


Figure 3:13 Distribution of AMA plants and hot nodes on the angiosperm phylogeny. There are 51 plants used against cancer in Africa, Malaysia and India (red dots). The hot nodes (red clades) represent lineages that are over-represented in cancer use. The blue dots represent plant orders that were identified as hot nodes. Blue dots: clades

have not been colored in so as not to obscure the families that are over-represented within each clade. This is because not all families in the over-represented clade are over-represented for use against cancer in the AMA dataset.

Table 3:3 Hot nodes identified from the AMA dataset. These families are significantly over-represented in genera having anti-cancer activity compare with the rest of the tree. The table shows each "hot-node" and the number of genera and species within that node. The total species represent 8.5% of land plants that are expected to be of greater medicinal value for use against cancer than would be expected by random chance.

Hot nodes	Genera in node	Species in node
Zingiberaceae	52	1,587
Clusiaceae	24	1,047
Ancistrocladaceae	1	21
Nyssaceae	5	37
Taxaceae	6	31

Using the 129 cancer-remedying plant species from the AMA dataset, our hot node analysis identified 2,574 novel species, which represents ~8.58% of all land plants (see Table 3:3). Table 3:3 also contains Nyssaceae and Taxaceae, which do not have any plants in the AMA dataset, but were included in the analysis to demonstrate the efficacy of the tool and for retrospective validation of the method. We note that Taxol (a chemotherapeutic agent) is isolated from various *Taxus* species that belong to Taxaceae, while the quinoline alkaloid Camptothecin (chemotherapeutic agent) is produced by both *Camptotheca acuminata* and *Camptothecin lowreyana*[302] that belong to Nyssaceae. The identification of just 2,574 novel species suggests a greatly reduced set of plants that could be prioritized for screening for potential anti-cancer activity.

We next look more closely at the taxa identified by the analysis presented above to see if there is any primary literature support for anti-cancer activity, since these taxa were not represented in the AMA dataset of known anti-cancer plant species compiled from existing database resources. This would provide additional confidence in the ability of our analysis to identify of taxa with suspected medicinal activity. Table 3:4 shows the families identified as "hot-nodes" and other plants in those families that were not contained in the AMA dataset, but for which anti-cancer activities have been reported in the literature. For example, despite the fact that alkaloids produced by Acistrocladaceae member *Ancistrocladus korupensis* have shown anti-HIV activity[303] as well as anti-malarial activity[304], the 30 family members that are not contained in the

AMA dataset have not been screened for anti-cancer activity, suggesting that Acistrocladaceae alkaloids should be prioritised for anti-cancer screening campaigns.

Table 3:4 Plants with reported anti-cancer activities that were identified by the "hot-nodes", but were not in the input data.

Family	Plants not in dataset with reported anti-cancer activity
Zingiberaceae	*Alpinia galanga*[305]
	Alpinia officinarum[306]
	Curcuma caesia[307]
	Curcuma kwangsiensis[308]
	Curcuma purpurascens[309]
Clusiaceae	*Mesua beccariana*[310, 311]
	Allanblackia gabonensis[312]
	Garcinia nervosa[313]
	Garcinia achachairu[314]
Acistrocladaceae	*Ancistrocladus korupensis (anti-HIV and anti-malarial properties)*[303, 304]

3.3.4 PATTERNS OF AFRICAN MEDICINAL PLANT USE – CANCER, MALARIA AND HAT

The AfroCancer (subset of the AMA dataset - 17 families), AfroMalaria and AfricaTryp (see Methods), contain plants with reported activity against endemic diseases in Africa, for which the population is highly dependent on traditional medicine. We utilised two statistical approaches to better understand the distribution of medicinal flora across families and to identify which families are important for cancer, malaria and human African Trypanosomiases (HAT) use in Africa. We first establish whether specific families are significantly enriched for plant species that are used against cancer, malaria and HAT. The goodness of fit test on the collected families showed a significant departure of the anti-cancer, anti-malarial and anti-trypanosomal species from

homogeneity (for p-values see Table 3:5). The high level of statistical significance means that we can now assess the distribution of the medicinal plants within families.

Table 3:5 χ^2 Test for the 3 datasets. We calculated whether the AfroCancer, AfroMalaria and AfricaTryp species can be distinguished from the flora as a whole. The chi-square goodness of fit test on the collected families showed a significant departure of species from homogeneity (shown by the p-values), i.e. statistically more medicinal (cancer, malaria, HAT) species than in the flora as a whole.

	χ^2	p-value
AfroCancer	69.28	$2.07e^{-12}$
AfroMalaria	1039.71	$2.20e^{-16}$
AfricaTryp	1378.9	$2.20e^{-16}$

For the second part of this study, we assessed the distribution of medicinal plants within families and how they deviate from a homologous null model of distribution using Binomial analysis. This method evaluates the statistical significance of numerical deviation from the expected norm. For each of the datasets, the families that statistically deviate from the null hypothesis are shown in Figure 3:14, Figure 3:15 and Figure 3:16. Significance values (p-values) are only shown for those families that contain more medicinal plants than would be expected under the null hypothesis. This is because in this study we are looking at plant families that are used medicinally more than would be expected by random chance.

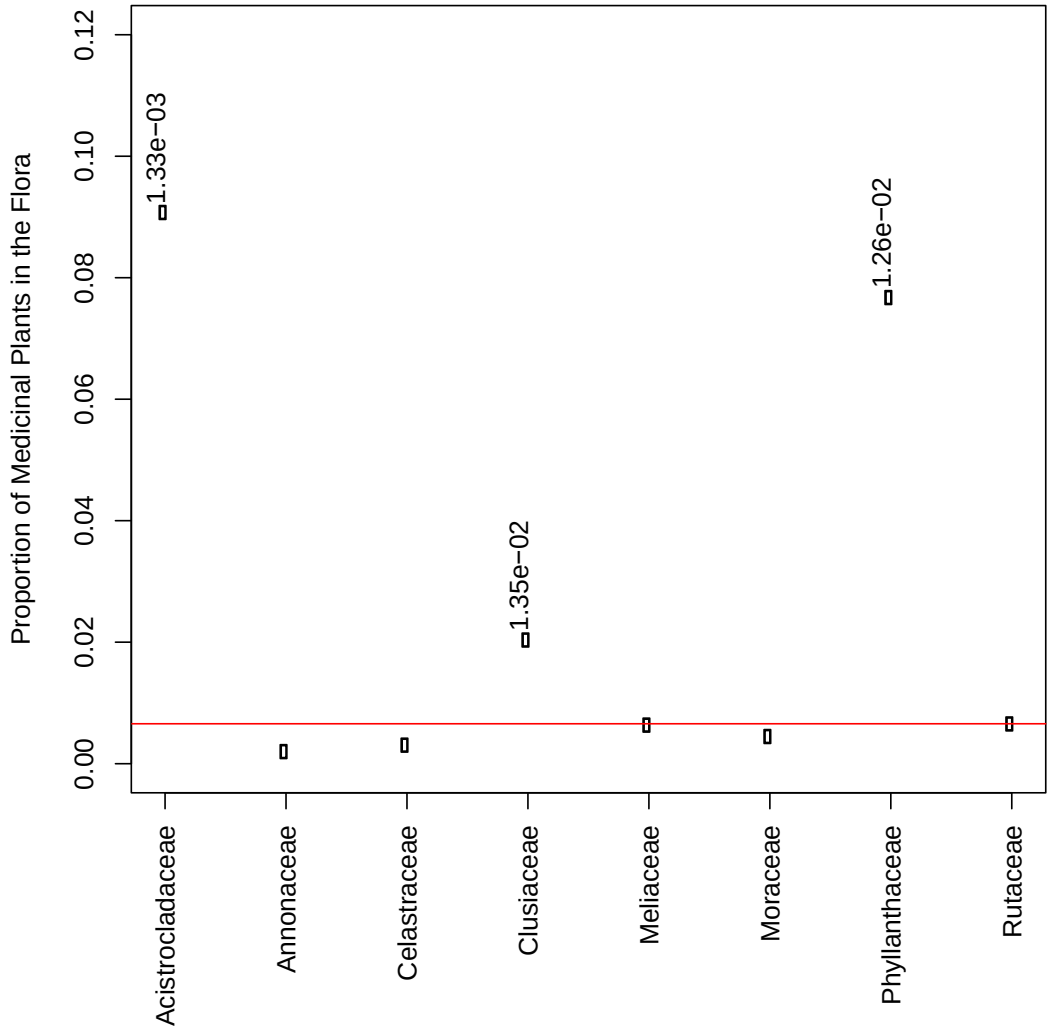


Figure 3:14 Results of the binomial test on the families in the AfroCancer dataset. p-values less than 0.05 for families that are used more than would be expected by random chance are shown. Here, Acistrocladaceae, Clusiaceae and Phyllanthaceae depart from a uniform model (over-represented) of proportion of medicinal plants in the African flora.

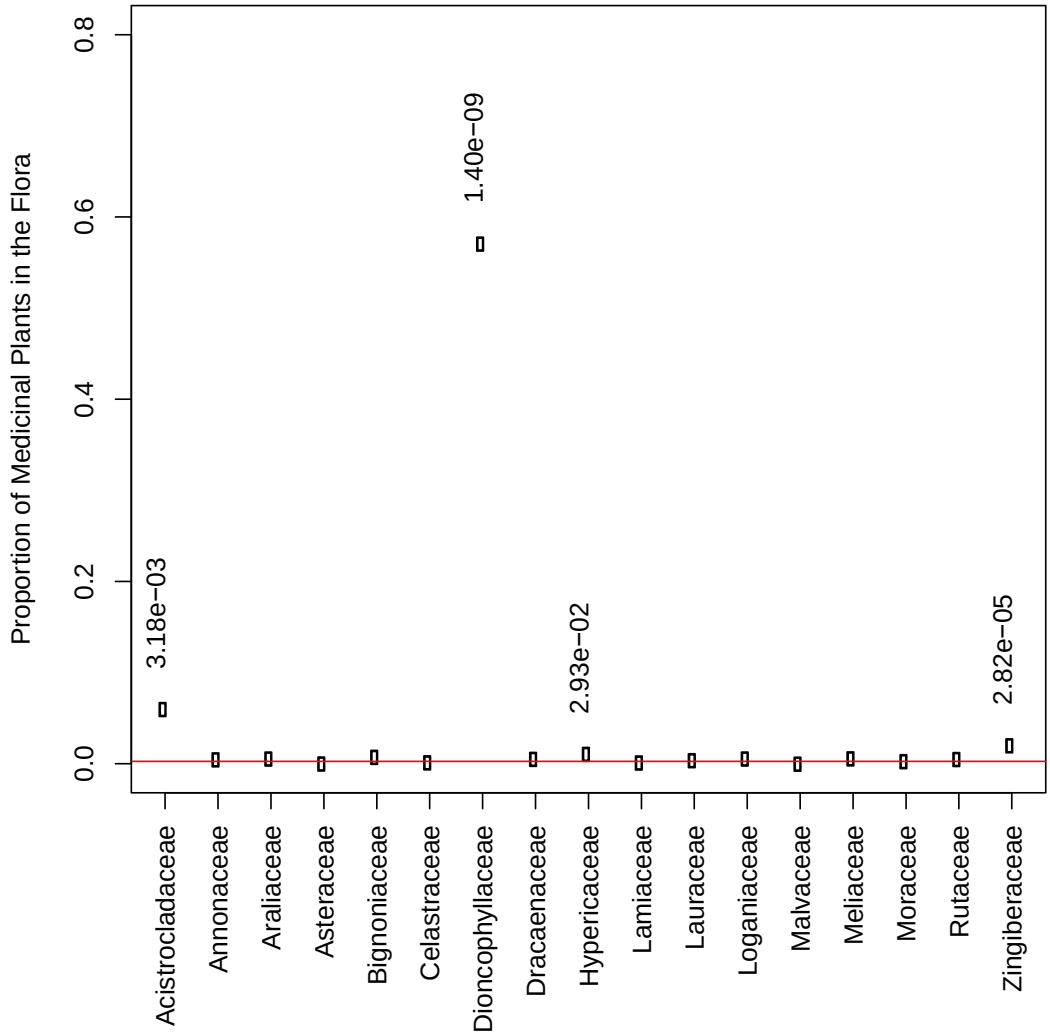


Figure 3:15 Results of the binomial test on the families in the AfroMalaria dataset. In this figure p-values less than 0.05 for families that are used more than would be expected by random chance are shown. Here, Acistrocladaceae, Dioncophyllaceae, Hypericaceae and Zingiberaceae depart from a uniform model (over-represented) of proportion of medicinal plants in the African flora.

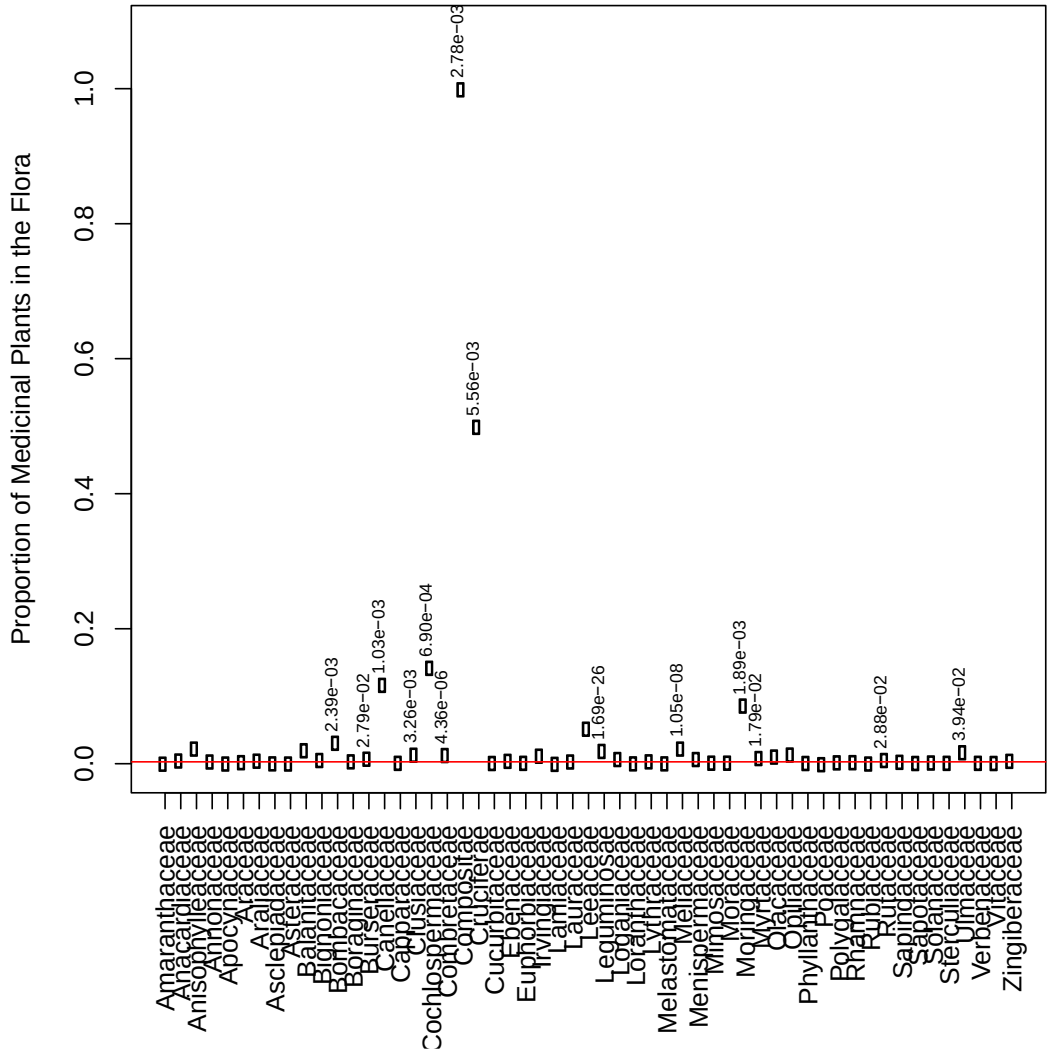


Figure 3:16 Results of the binomial test on the families in the AfricaTryp dataset. p-values less than 0.05 for families that are used more than would be expected by random chance are shown. Here, Bombaceae, Burseraceae, Canellaceae, Clusiaceae, Cochlospermaceae, Combretaceae, Compositae, Cruciferae, LEguminoseae, Meliaceae, Moringaceae, Myrtaceae, Rutaceae and Ulmaceae depart from a uniform model (over-represented) of proportion of medicinal plants in the African flora.

For the medicinal plant families in the AfroCancer dataset, Acitrocladaceae, Clusiaceae and Phyllanthaceae were identified by the Binomial method to be over-represented medicinally as they differ significantly ($p<0.05$) from null expectations (Figure 3:14). Acistrocladaceae, Dioncophyllaceae, Hypericaceae and Zingiberaceae were identified by the Binomial method to be over-represented medicinally for the plant families in the AfroMalaria dataset (Figure 3:15). For the AfricaTryp dataset, 14 families were over-represented medicinally (Figure 3:16). In addition to families that are well known for their cancer, malaria and HAT use, e.g. the annonaceous acetognenins from Annonaceae are known to display anti-plasmodial activity[315 316], and Meliaceae for HAT[317, 318], we identified others that are less well known, e.g. Acistrocladaceae for cancer. Of the identified families, we also found that other species in these plant families produce the same active metabolites that are responsible for activity and these are discussed in further detail in section 1.3.4.1. This provides a form of validation for our approach of narrowing down the search for potential medicinal activity of the African flora to the over-represented families and the ones that seem to be overlooked, i.e. they contain chemistry with the desired activity, but they are not annotated with a potentially suitable use, e.g. Dioncophyllaceae.

Our results are similar to previous studies carried out to investigate phylogeny patterns in medicinal plants. Phylogenetic clustering was found when inspecting medicinal properties of Plectranthus[319] where similar uses were found among the related species. This was also the case for Pterocarpus[145], Aloes[320] and Euphorbia[321] genera, where a phylogenetic signal was found for medicinal use. Similarly, studies were carried out concentrating on plant use for a specific activity, e.g. studies on psychoactive plants[284] and those used against snakebite[322]. Here the researchers found plant lineages displaying over-abundance of plants having psychoactive activity and anti-snakebite activity respectively. However, apart from the study on psychoactive plants, no direct link was drawn between the chemistry of the secondary metabolites of the plants and the observed phylogenetic signal. In the study by Halse-Gramkow *et al*[284] a link between the tropane alkaloids produced by the Solanaceae plants and the identified psychoactive "hot-node" is made but this is not explored in detail. Following on from these studies our results show that for the African medicinal flora, the distribution of

medicinal plants departs significantly from a homogenous null model (binomial analysis results).

It is important to note here that our analysis relies on data collected from the published literature, which suffers from reporting bias. This means that the data is not complete, i.e. not all plants in Africa with anti-malarial, anti-trypanosome or anti-cancer activity are included in our datasets.

3.3.4.1 Relationship between Unique metabolites and biosynthetic pathways in over-represented plant families and activity

From the binomial analysis in the previous section, we recovered plant families that are over-represented, i.e. used more than would be expected by chance. Acistrocladaceae, Rutaceae and Clusiaceae appeared in the results for all three datasets; therefore, a literature review was carried out to assess the relationship between the secondary metabolites produced by these plant families and their medicinal use.

Ancistrocladaceae and Dioncophyllaceae

Plants (from our database) in the two small, closely related, tropical Ancistrocladaceae and Dioncophyllaceae produce a unique class of compounds called the naphthylisoquinolines (NIQs). They are characterised by a biogenetically unique scaffold of acetate origin, which has a methyl substituent at C3 and a meta-oxygentation pattern at C6 and C8[315]. The active NPs of these two families, e.g. dioncophyllin A is produced by a different biosynthetic pathway to all other isoquinolones in nature and, in the plant, the pathway is initiated in response to stress e.g. chemical stress, biotic stress or physical stress[323]. Activity of these NIQs have been reviewed previously, e.g. *in vivo* anti-tumour activity of dioncophylline A[22], anti-plasmodial activity of dioncophylline C and dioncopeltine A[324] and *in vitro* activity against *Trypanosoma brucei rhodesiense* and *Trypanosoma brucei brucei*[325]. The unique chemistry is responsible for activity in plants over-represented in medicinal flora. Other species in the family producing this unique chemistry are expected to have similar activity.

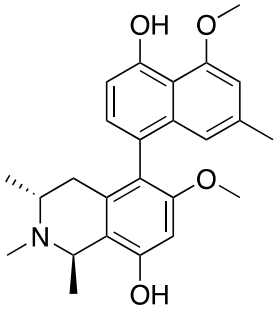
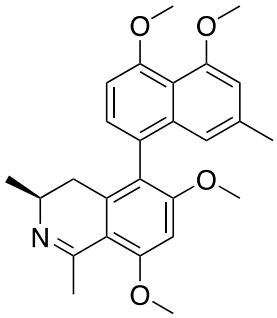
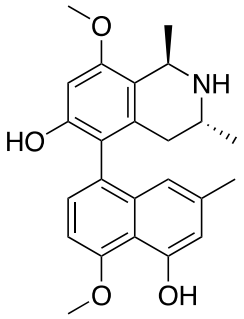
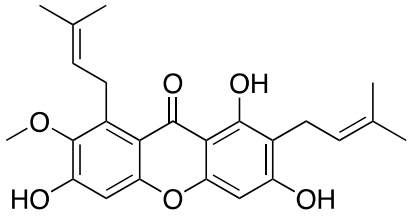
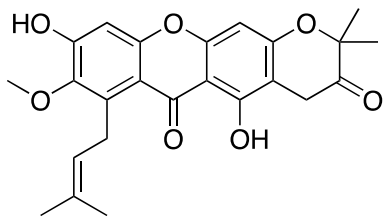
Clusiaceae

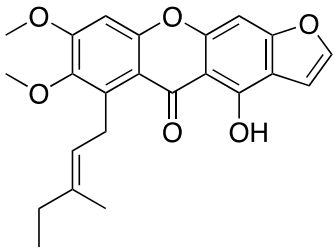
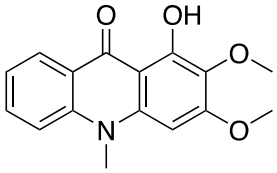
Plants in the Clusiaceae family, as well as the closely related Gentianceae and Bannetiaceae (native to neotropics i.e. not present in Africa), uniquely produce distinctive xanthones. African plants from Clusiaceae producing xanthones have shown *in silico* binding against tryapanosomal targets, including: xanthchymol[326]. The *in vivo* anti-tumour activity of xanthones from African NPs have been reviewed[21], as has their anti-plasmodial activity[327]. Furthermore, xanthones produced by plants in different continents have also been shown to display anti-plasmodial activity, e.g. *Swertia alata* (Gentianceae) in Pakistan[328]. These findings further increase our confidence in our finding that these plant families are used more than average due to their unique chemistry.

Rutaceae

The Rutaceae family is unique in producing C3 substituted coumarins with a 1,1-dimethylallyl functional group as well as the acridones. The activity of acridones isolated from African plants has been reviewed previously, e.g. their potential as cancer therapeutics[22], and anti-protazoal activity[315]. Acridone alkaloids isolated from plants not indigenous to Africa, e.g. *Swinglea glutinosa* from the Phillipines[329], displayed IC_{50} activity of 0.3 to 11.6 μM against *Plasmodum falciparum*, and five of the acridone compounds had $IC_{50} < 10$ μM against *Trypanosoma brucei rhodesiense*. This shows that the plant families that we have identified to be used more than average do display anti-protozoal activity in other geographic regions of the world. This is useful because, knowing that plants from the same families are likely to have NPs with similar structure and activity, we can exploit information of known activities from plant families around the world and apply this to the African medicinal flora.

Table 3:6 Some known NPs from plants in the dataset from the Ancistrocaldaceae, Dioncophyllaceae, Clusiaceae and Rutaceae and their predicted or experimental activities.

Family	NP	Activity
Acistrocladaceae (NIQs)	  	Moderate anti-protozaol activity [330]
Clusiaceae (Xanthones)	 	Apoptotic and antiproliferative activities [331]

		
Rutaceae (Acridones)		Anti-plasmodial activity [332]

From the examples in Table 3:6 we can see that NPs that are uniquely produced by the over-represented families are responsible for activity and may be responsible for the family being over-represented and over-utilised. As such, we can validate our recommendation of further investigating plant species in over-represented families in the hope of finding novel bioactive NPs. In our study we have shown a phylogenetic correlation between African medicinal plants, their secondary metabolites and their predicted activities/known activities. To our knowledge this correlation has not been made before and none of the previous predictive phylogenetic studies have been carried out on African medicinal flora.

3.4 LIMITATIONS OF THE STUDY

Despite the apparent connection that we have found in this study between the phylogeny of a plant and its predicted activity, several limitations exist. Firstly, as in many studies involving natural product datasets, we are limited by the information available in the datasets. In this type of database (e.g. NANPDB) not all secondary metabolites in a plant are recorded, as, through bioactivity-guided fractionation, only bioactive phyto-constituents are characterised and analysed. Thus, it is not possible to predict protein targets for compounds in a plant (which may be active) beyond what is available in the dataset. Also, when trying to extrapolate the mode of action of a NP from the predicted target, we must keep in mind that ethno-botanic use includes ameliorative effects of symptoms associated with a disease and not necessarily treatment or cure of a disease. It is also important to note how data for each medicinal plant is presented. For each plant (in one of the African medicinal plant datasets) annotated with an activity e.g. cancer, all compounds isolated from the active extract are included. The active constituent cannot be determined by querying the dataset. Furthermore, diverse assays are presented within a database, e.g. for the AfroCancer dataset, more than 40 assays are recorded that determine anticancer activity with recorded activities being anti-proliferative, cytotoxic etc. on different cell lines including ovarian cancer cell line, human colon cancer cell line, fibrosarcoma and melanoma. Several activity values e.g. IC_{50} and ED_{50} are used. It is thus important to determine the mechanism of action of these NPs.

Secondly, for the regression and binomial analysis to identify "hot nodes" it is important to remember that not all plants with medicinal activity are recorded and used traditionally. This may be due to several reasons, e.g. the plant not being accessible geographically. Therefore, lack of ethno-botanic use of a plant does not indicate that the NPs in the said plant are inactive for a disease.

Thirdly, we are using the current Angiosperm Phylogney Group (APG) III classification system of plants. This classification is constantly being reviewed and updated and plants from one family are moved to another family upon discovery of new molecular data. As such, the relationships that we draw between phylogeny and

predicted activity are not exact; rather, they are based on the currently accessible material and the accuracy of the present classification system.

3.5 CONCLUSION

In this chapter we have shown that for the NANPDB dataset, compounds share higher Tanimoto similarity to compounds in the same family than they do to compounds in other families. We have also shown that plant families produce these similar compounds regardless of the geographic origin of the plant, where we see that, e.g. Leguminoseae in Africa, Malay and Indian traditional medicine produced structurally similar compounds having Tanimoto coefficient 0.95 or more. We have shown that compounds that are closely related to each other phylogenetically produce compounds that are similar to each other and these compounds bind similar targets. Plants further away in the phylogeny tree produce diverse compounds that act on different targets. We were able to rationalise when this was not the case. Furthermore, we have statistically identified plants that are over-represented and under-represented in African traditional medicine for use against cancer, malaria and human African trypanosomiases. These families are known to produce unique metabolites via unique biosynthetic pathways, e.g. the napthoisoquinolones in Ancistrocladaceae and Dioncophyllaceae, and we have made the connection between these unique metabolites, bioactivity and over-utilisation. Based on our initial finding that plant families produce similar compounds and have similar predicted activities, we recommend that these plant families be prioritised for screening for bioactive metabolites.

CHAPTER 4: INTEGRATING STRUCTURAL AND CHEMOGENOMIC SPACE TO PREDICT THE MECHANISM OF ACTION OF PHENOTYPICALLY ACTIVE SMALL MOLECULES AND NATURAL PRODUCTS IN *TRYPANOSOMA BRUCEI*

4.1 INTRODUCTION

Human African trypanosomiasis (HAT) is a parasitic disease caused by the protozoan parasite *Trypanosoma brucei*. This disease is fatal if left untreated[333]. A need exists, in particular in Africa, to identify effective drugs for this disease. The current approaches to trypanosome drug discovery include:

(i) Development of new molecules inspired by known anti-trypanosomal agents, e.g. the bisamidines that were developed based on the structure of Pentamidine. These compounds failed in clinical trials due to nephrotoxicity[334].

(ii) Target based screening, used to identify, e.g. DDD85646, which showed activity against N-myristoyltransferase (NMT)[335]. This compound could not penetrate into the CNS therefore was not deemed useful for Stage 2 of the disease.

(iii) Phenotypic screening, used to identify candidates such as Fexinidazole[28], which entered Phase 2 and Phase 3 trials in 2012 and Oxaborole[336], which entered clinical trials in March 2012.

Previous work has been carried out to identify targets of compounds that have shown activity against several infectious diseases by integrating publicly available structural, chemical and bioassay data. Martínez-Jiménez *et al*[337] identified 139 target proteins modulated by compounds from an HTS against *Mycobacterium tuberculosis* by integrating bioinformatics and cheminformatics. A study by Spitzmüller and Mestres[338] on results from an HTS on *Plasmodium falciparum* identified 39 putative targets by using computational target prediction to predict protein targets followed by statistical

analysis to detect enrichment of the compounds in *Plasmodium falciparum* targets. Ekins *et al*[339] used a Bayesian machine learning algorithm to identify 11 compounds which had an EC_{50} below 10µM on *Trypanosoma cruzi*, and their potential targets.

Recently, HTS results have been deposited in ChEMBL-NTD to mediate drug discovery for HAT. In 2011, the Drugs for Neglected Diseases Initiative (DNDi) released the results of a screening and optimization of specific chemical series against human African Trypanosomiasis containing 4,927 compounds, 1,415 of which have an IC_{50} below 10µM (IC_{50} values for the currently marketed HAT drugs Pentamidine, Nifurtimox, Eflornithine and Melarsoprol are 0.01 µM[26], 5 µM[26], 81-693 µM[340] and 2.1ng/ml[26] respectively). In 2015 DNDi released the results of an antiprotozoal activity profiling of approved drugs. Similarly, the Swiss Tropical and Public Health Institute (SwissTPH) released a screening hits dataset containing 28 compounds in 2016. In March 2015 GSK deposited the GSK TCAKS Dataset (hits from *Leishmania donovani*, *Trypanosoma cruzi* and *Trypanosoma brucei brucei* phenotypic screening). Combining the wealth of information in these datasets with literature searches of the results of natural product screens against *Trypanosoma brucei* and target-based assays provides a promising starting point for *in silico* target identification.

In this study we introduce an approach in which bioinformatics and orthology information are integrated to predict and prioritise putative targets in *Trypanosoma brucei*. The goal is to elucidate the mechanism of action of these phenotypically active compounds. We start by characterising the structural and chemical features of the compounds from both the high throughput screens and literature search. A preliminary search for activity of the screened compounds on other organisms is carried out, and in the case of experimental activity or predicated activity, *Trypanosoma brucei* orthologues (if they exist) are identified. We then predict the protein targets of compounds using a ligand-based machine-learning algorithm that uses a Random Forest. The predicted protein targets from non-Trypanosomal organisms are then projected by orthology onto the *Trypanosoma brucei* genome to identify putative targets within this species. In addition, experimentally validated targets of the compounds from the screening datasets and literature review are obtained from ChEMBL and their trypanosomal orthologues identified. The biological processes

modulated by these predicted targets will also be studied and the difference between the NPs and SHs will be compared.

4.2 MATERIALS AND METHODS

4.2.1 DATASETS

To identify the target space of trypanocidal agents, several datasets comprised of phenotypic screen hits were used. These datasets were made up of 3 different datasets of compounds screened against Trypanosoma, which were downloaded from ChEMBL-NTD (www.ebi.ac.uk/chemblntd). These datasets contain results of phenotypic screens carried out on the bloodstream form of *Trypanosoma brucei*. All active and inactive data were pooled together as they were tested on members of the same species of Trypanosoma. The resazurin fuorescent T. brucei whole-cell viability assay[341] was used to screen compounds in two of the datasets, namely, "DNDi_T.b.brucei Dataset" and "GSK TCAKS Dataset". The cut-off for this assay was >80% growth inhibition of the *T. brucei* parasite. The STIB900 acute mouse model[342] was used to screen the compounds in the "DNDi Dataset: Antiprotozoal activity profiling of approved drugs". Mice in this assay are considered cured when there is no parasitaemia relapse detected in tail-blood over a 60-day observation period. For the purpose of this study, compounds (from these three datasets) with IC_{50} <10μm were considered active. The datasets are shown in Table 4:1 below and will collectively be referred to as the Screen Hits Dataset (SH).

Table 4:1 Datasets downloaded from ChEMBL-NTD for the Screen Hits Dataset (SH)

Name of Dataset	Active Compounds	Inactive Compounds	Reference
DNDi_T.b.brucei Dataset	638	511	[343]
DNDi Dataset: Antiprotozoal activity profiling of approved drugs	39	66	[342]
GSK TCAKS Dataset	249	343	[344]

A second dataset comprised of NPs screened or traditionally used against HAT was also studied. A literature review of articles with the keywords "Africa", "natural product", "medicinal plant", "HAT", "trypanosoma" and "parasite" revealed 862 compounds [23, 41, 315, 345, 346]. Plants used with activity against *Trypanosoma brucei* from

the NPASS[347] dataset were also downloaded and added to this dataset. This dataset will be referred to as the NP dataset.

A schematic showing the steps taken to identify the predicted target space of *Trypanosoma brucei* interacting with the phenotypically active compounds from the NP and SH dataset is shown in Figure 4:1.

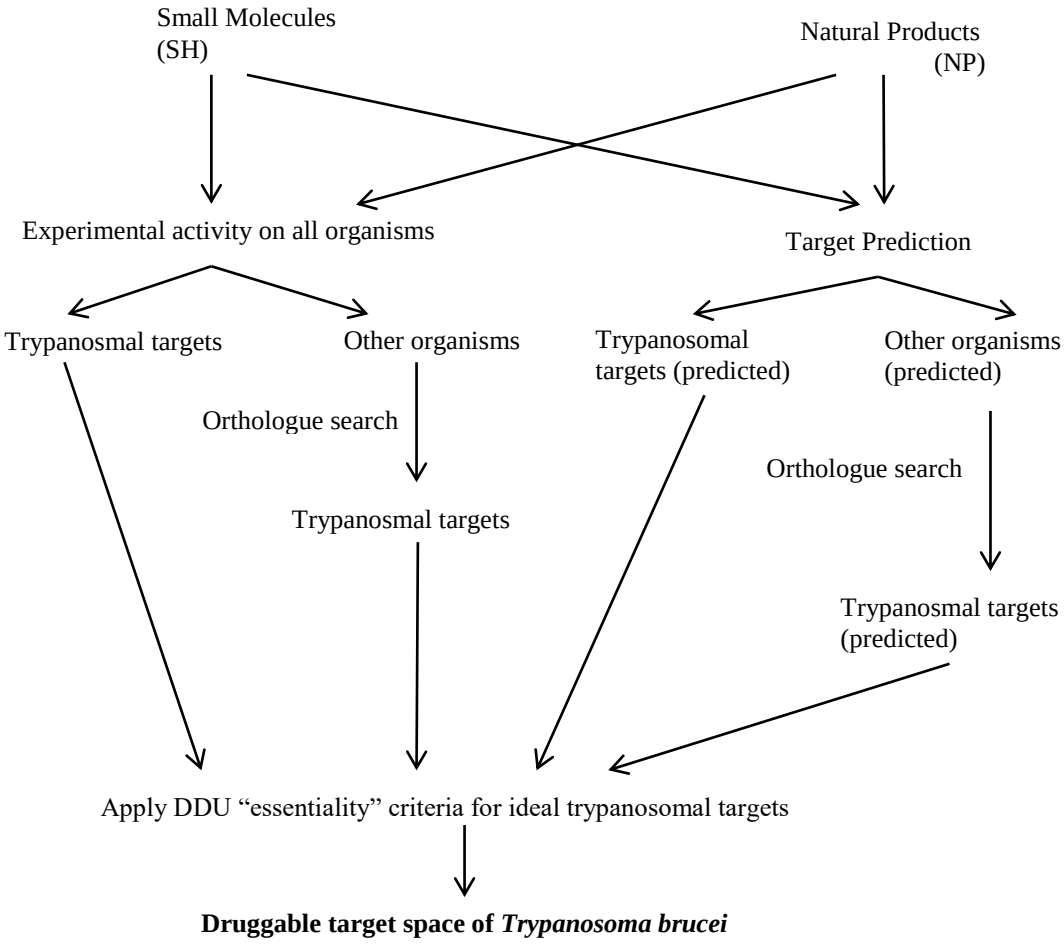


Figure 4:1 Schematic of the steps used to identify the targets of *Trypanosoma brucei.* For each of the two datasets (SH and NP) experimental activity of the active compounds was extracted directly from ChEMBL. In addition, targets were also predicted using PIDGIN v2. Trypanosomal targets as well as targets from different organisms were identified. Orthologues of the non-trypanosomal targets were obtained from PantherDB. Targets essential for the survival of the trypanosome were obtained from TriTrypDB. We obtained the phenotypically relevant target space of Trypanosoma brucei by overlapping the predicted and experimental targets with the essential targets.

4.2.2 STRUCTURAL PRE-PROCESSING

Compounds in the SH dataset were downloaded in SD Format and converted into canonical SMILES using OpenBabel[348]. Compounds in the NP dataset were converted from PDF to canonical smiles using Document2Structure JChem 17.21.0, ChemAxon (http://www.chemaxon.com). ChemAxon Standardizer[181] was used for structure canonicalization, transformation, and conversion of compounds from SD format to SMILES. To standardise the compounds in ChemAxon Standardizer, the following options were used: Clean 2D, Mesomerize, Neutralize, Remove Explicit Hydrogen and Remove Fragment. Duplicate structures in each dataset were removed, using ChemAxon JChem Software[181], "remove duplicates".

4.2.3 CHEMICAL SPACE ANALYSIS

4.2.3.1 FINGERPRINT CALCULATION AND MDS VISUALISATION

To visualise the chemical space of the two datasets, we calculated the Morgan fingerprints[349] (radius 2) on KNIME version 3.3.1[350] and projected this information on an MDS plot in R[182]. To quantify this information further, the average Tanimoto similarity of each compound to all was calculated in KNIME and a density plot was constructed to visualise the results using R[182].

4.2.3.2 SCAFFOLD GENERATION

The Murcko scaffolds of the compounds in both datasets were generated using Datawarrior[191]. The options "Analyse Scaffolds" followed by "Murcko Scaffold" were used. This was carried out to compare the scaffolds in both datasets to scaffolds of compounds with known anti-trypanosomal activity. This gave a table of the generated scaffolds and the number of compounds populating those scaffolds. The top 10 scaffolds were retained and are shown in the results.

4.2.3.3 FRAGMENT ENRICHMENT ANALYSIS

To obtain the fragments that are more common in one active dataset over the other, the compounds were decomposed into fragments using the MoSS[351] node in KNIME[350]. The options "ignore pure carbon fragments" and "use ring mining" were used. To

extract the most common fragments in the SH dataset compared to the NP dataset, a fragment had to occur in a minimum of 10% of the fragments of the SH dataset and a maximum of 5% in the NP dataset. To extract the most common fragments in the NP dataset compared to the SH dataset, a fragment had to occur in a minimum of 10% of the fragments of the NP dataset and a maximum of 5% in the SH dataset.

4.2.3.4 PLOGBB CALCULATION

The log BB, which is a prediction of blood brain barrier permeation, was calculated for the active compounds in the SH dataset to predict which of these compounds will cross the blood brain barrier, as this is an important feature for drugs required for the phase two part of HAT. Compounds with log BB >0.3 cross the blood brain barrier readily, whereas those with log BB <-1 do not cross so readily[352, 353]. The predicted log BB, which defined as the logarithm of the ratio of the concentration of a drug in the brain and in the blood, measured at equilibrium, was predicted using the QikProp plogBB function on Canvas Schrodinger software[354].

4.2.4 CHEMOGENOMIC SPACE ANALYSIS

4.2.4.1 TARGET PREDICTION AND ORTHOLOGUE SEARCH

PIDGIN v2 [355] was used to predict enriched targets and pathways for both the NP and the SH datasets. Target enrichment is calculated in PIDGIN v2 using the prediction ratio, defined in Equation 5:

$$\text{Prediction ratio} = \frac{\text{Ft/Nt}}{\text{Fb/Nb}}$$

Equation 5 – Prediction Ratio

Where:

F_t = Frequency of prediction in the test set i.e. the number of active predictions [p(activity) above threshold] across the entire set of input molecules.

N_t = Number of predictions in test set

F_b = Frequency of prediction in a background distribution set

N_b = Number of predictions in a background distribution set

The lower the prediction ratio, the more enriched the target is in the phenotypic library. Since the models vary in size, chemical space and ratio of active:inactive molecules, an Odds ratio and Fisher's exact test is carried out by PIDGIN v2 to correct for over and under prediction of promiscuous and/or selective models. The targets and pathways with an Odds ratio below 0.1 were kept for all organisms. Orthologues of the predicted targets to Trypanosoma targets were identified using PantherDB[137]. The "least diverged orthologues" (LDO), i.e. genes in two different organisms that have diverged the least since their most common recent ancestor (expected to retain similar functional activity across organisms), were kept.

4.2.4.2 KNOWN BIOACTIVITY

For both datasets, the experimentally validated targets against all organisms were obtained from ChEMBL[356] and the actives were determined using the following criteria: IC_{50} < 10µm, target_type = single protein, confidence score > 8 activity_comment = active or inhibitor. *Trypanosoma brucei* orthologues of these targets were also obtained from the PantherDB[137] and the LDO, i.e. most nearly equivalent were kept.

4.2.4.3 TARGET ESSENTIALITY

In order to identify essential *Trypanosoma brucei* targets the TDR Targets database version 5[357] was used and the following criteria were applied: for all queries, target organism = *Trypanosoma brucei*, target is an enzyme, target is a receptor, target is a transporter and evidence that target is essential in any species, resulting in 1,809 genes.

4.2.4.4 NETWORK CONSTRUCTION

This was carried out to visualise the multi-target prediction of compounds in the small molecules hits dataset. Cytoscape[358] was used to construct the target-compound network of the active screening hits. The target-compound pairs were obtained from the PIDGINv2 predictions of the active compounds of the SH dataset as well as the bioactivity data of these compounds which was extracted from ChEMBL as described above in 4.3.4.2. The network was analysed using the "Network Analyser" function of

Cytoscape. The average node degree of the generated network is 22.305, so any node with a higher degree (i.e. more connectedness) was considered a "hub" in our analysis.

4.2.4.5 ENRICHED BIOLOGICAL PROCESSES

For each dataset, the list of trypanosomal orthologues of predicted targets was used to obtain a list of enriched *Trypanosoma brucei* GO biological processes. This was carried out using TriTrypDB[359], with the option "Analyse Results" to find the biological process terms enriched in the gene list at a p-value < 0.05. The p-value is a statistical measure of the likelihood that a certain GO term appears among the input genes, more often than it appears in the set of all genes in *Trypansoma brucei* (background). A tree map representation of the enriched GO terms was generated using REVIGO[360] and plotted using R[182].

4.3 RESULTS AND DISCUSSION

4.3.1 CHEMICAL SPACE ANALYSIS

We first analysed the chemical space of the two datasets (SH, 872 molecules and NP, 826 molecules) to see if they share chemical space or if they occupy different regions of chemical space. The results are shown in Figure 4:2. Despite the difficulty of representing chemical diversity in low dimensions, it can be seen that compounds from the two datasets have little overlap in chemical space, and that the NPs occupy space that is not occupied by the SMs. Furthermore, Figure 4:2 suggests that the NPs are more spread out in chemical space and are thus more chemically and structurally diverse than the SMs. We therefore expect these two datasets to display different mechanisms of action since their chemistries appear to be different.

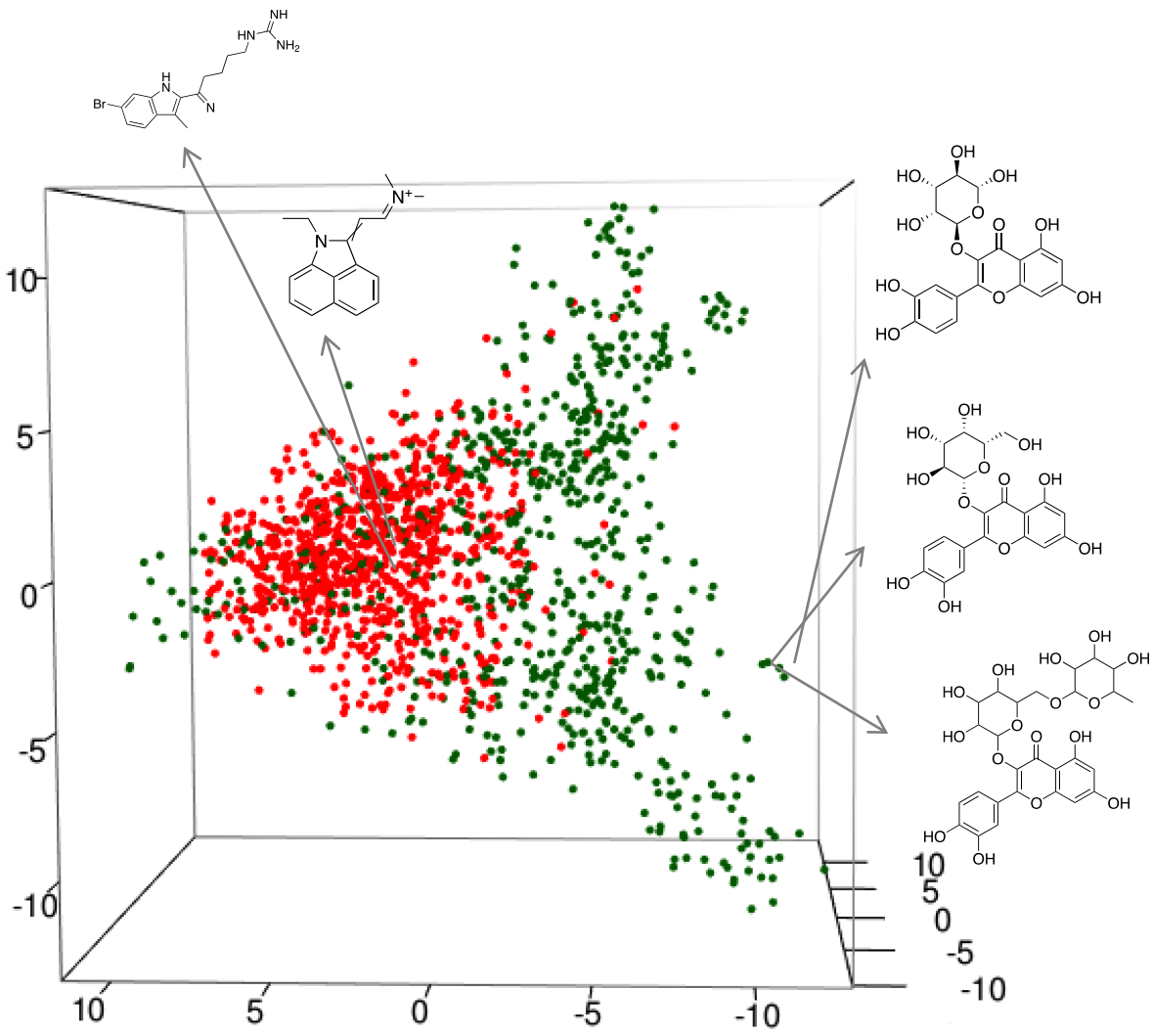


Figure 4:2 MDS plot calculated from the Euclidean distance between Morgan fingerprints of the NP and SM datasets. Each data point corresponds to a compound in the datasets. Green data points are NP compounds and red data points are small molecule hit (SH) compounds. It can be seen from the MDS image that compounds from the two libraries share a small amount of chemical space and the NPs expand into an area of space not covered by the small molecules.

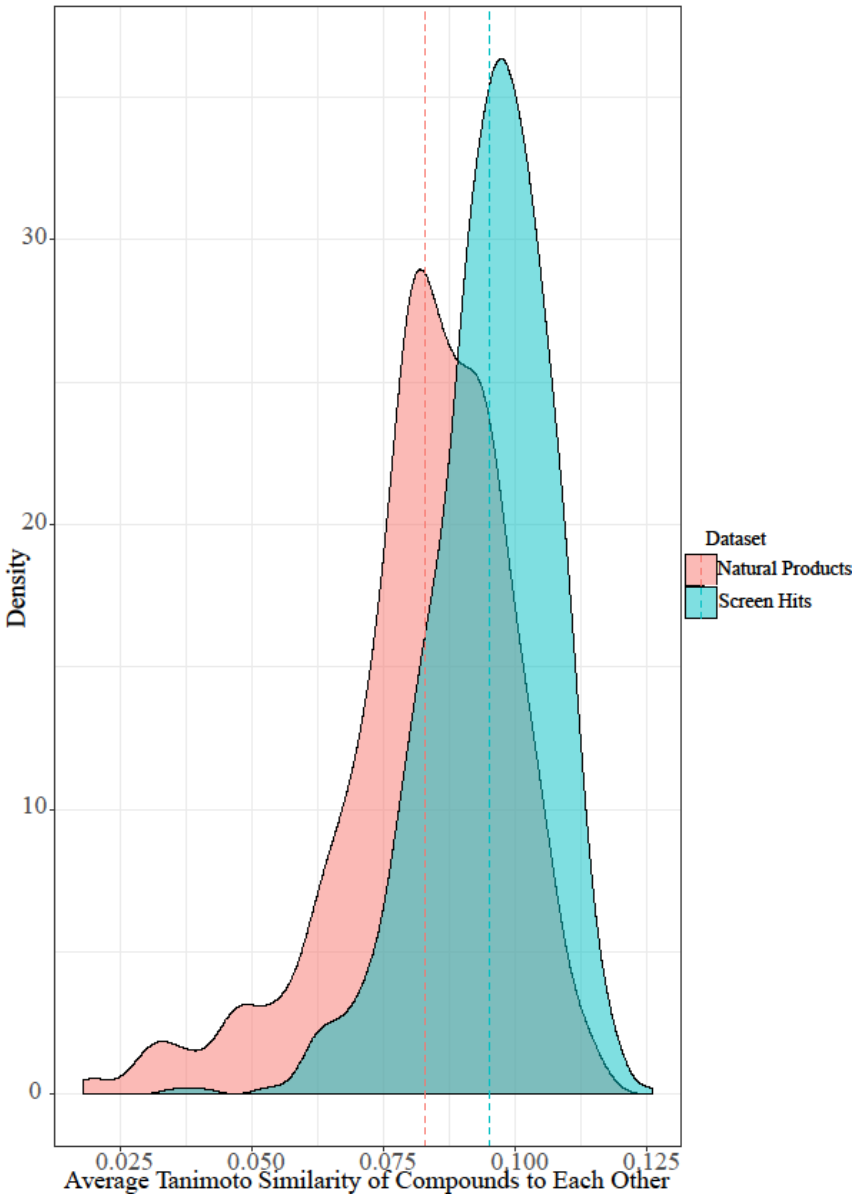


Figure 4:3 Density plot of Tanimoto similarity of all vs all compounds in each dataset. Natural products have the lowest intra library similarity compared to the small molecule hits and the combined library. The mean similarity for the SH dataset is 0.095 ±0.012, while the mean similarity for the NP dataset is 0.0827 ±0.016.

To quantify this further, the average Tanimoto similarities of each compound to all compounds in both its corresponding dataset were calculated and projected onto a density plot; see Figure 4:3. Here we see clearly that NPs have lower intra-similarity (mean average Tanimoto coefficient of 0.0827 ±0.016) when compared to the SH dataset. This indicates that they occupy a broader chemical space and this can also be seen in Figure 4:2 where the NPs are more spread out in chemical space. The SH dataset has a mean average intra-similarity Tanimoto coefficient of 0.095 ±0.012 indicating that they are more similar to each other than members of the NP dataset.

A previous study[176] identified an activity-relevant Tanimoto coefficient value of ≥ 0.3 for bioactive compound pairs. The Tanimoto similarities that we have calculated suggest both that these two datasets are each chemically diverse, and moreover that the datasets occupy different regions of chemical space.

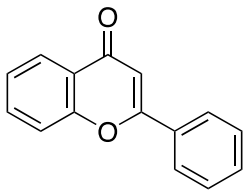
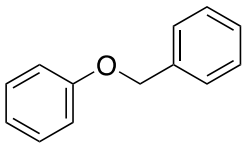
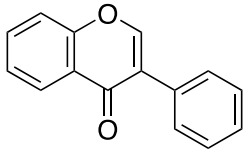
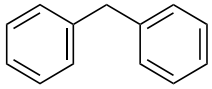
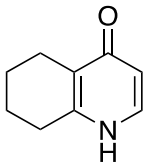
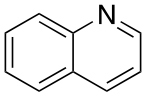
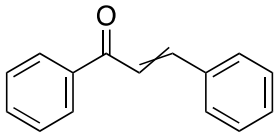
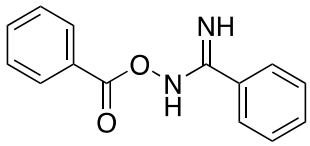
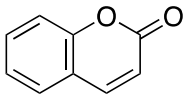
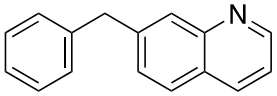
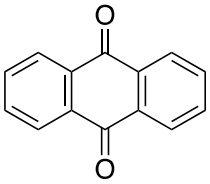
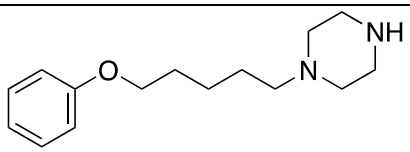
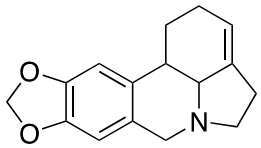
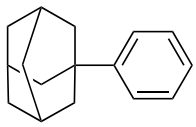
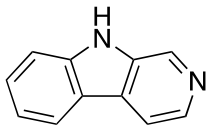
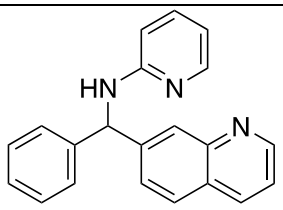
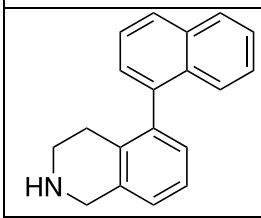
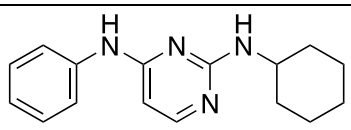
In an effort to understand the chemical diversity present in these datasets in greater detail, we decomposed the compounds in the two datasets into their Murcko[188] scaffolds. The NPs were found to have 413 scaffolds whereas the SHs were found to occupy 712 scaffolds, with 25 scaffolds shared between the two datasets.

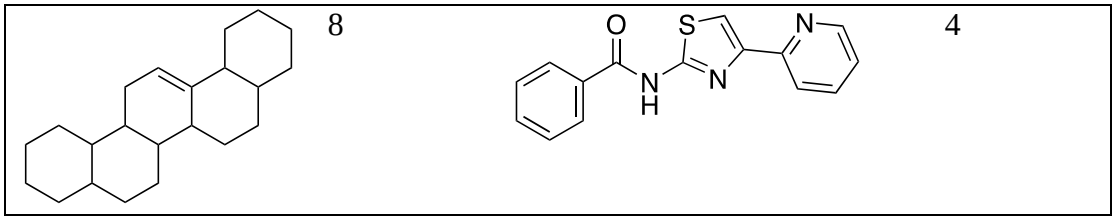
Table 4:2 shows the ten most populated scaffolds in each dataset (active compounds). It is important to note that none of the top ten populated scaffolds overlap between the two datasets, further illustrating the chemical diversity of the two datasets shown in the MDS plot (Figure 4:2). There was an overlap of just 25 scaffolds in total between the databases. Ten of these 25 overlapping scaffolds were in the top 100 most populated NP scaffolds with only seven of the overlapping scaffolds appearing in the 100 most populated SH scaffolds. Further insight into the diversity within the libraries was shown by the percentage of scaffolds that were represented by only one compound. This was 73.4% for the NPs and 87.2% for the SMD. This also explains why the compounds are so spread out in the MDS plot. As we have shown above in the Tanimoto coefficient density plot, the compounds in the SH dataset are structurally more similar to each other, than the NPs are to each other. So, despite having more scaffolds occupied by only one compound than the NP dataset, the SH dataset is still less chemically diverse than the NP dataset. One reason for this may be that the SH dataset contains compound series intentionally designed to have similar scaffolds for high throughput screening.

No evidence of anti-parasitic activity was found for the third, fourth, fifth, sixth, ninth and tenth most populated scaffolds in the NP dataset. The top two most populated scaffolds in the NP dataset, (flavonoids and isoflavonoids) are occupied by compounds that are known to have activity against trypanomastids as well as other parasites, e.g. Plasmodium parasites. These are reviewed extensively by Schmidt *et al*[315]. Indolizidine alkaloids (the seventh most populated scaffold in the NP dataset) have shown activity against chloroquine-sensitive and chloroquine-resistant strains of *Plasmodium*

falciparum with IC_{50} values of between 39-120 ng/ml respectively. (Chloroquine standard activity is 17 and 140 ng/ml respectively). The eighth most populated scaffold in the NP dataset is also known to have potent anti-trypanosomal activity[361]. Evidence for anti-parasitic activity was found for the ninth and tenth most populated scaffolds in the SH dataset. We found that the ninth most populated scaffold in the SH dataset is similar in structure to 4-anilinoquinazoline which is a scaffold of compounds known to display anti-trypanosomal activity[362]. The tenth most populated scaffold is an amino thiazole, also known to have potent anti-trypanosomal activity[361]. Taken together these results show that even though the scaffolds in the two datasets are different, they are nonetheless associated with compounds with known anti-trypanosomal activity. We expect them to act by modulating different targets due to the differences in their structures.

Table 4:2 Top 10 populated Murcko scaffolds in the NP library and small molecule hits library (SH). It can be seen that none of the top 10 populated scaffolds are shared between the two datasets. Benzene was not included as a scaffold.

NP Scaffold	Frequency	SH Scaffold	Frequency
	60		7
	23		7
	10		6
	10		6
	9		6
	9		5
	9		4
	8		4
	8		4



To explore the diversity of the chemistry of these compounds even further, the molecules in each dataset were decomposed into molecular fragments, and the fragments that were most enriched in each dataset were identified (see Methods for details of the analysis). Supplementary Table 5 shows the fragments enriched in the SH dataset against the NP dataset and Supplementary Table 6 shows the fragments enriched in the NP dataset against the SM dataset. Supplementary Table 7 shows the fragments enriched in all the active compounds against all the inactive compounds. From Supplementary Tables 5 and 6 it is clear that the fragments are different in both datasets. The NP fragments tend to be the polyphenols whereas the small molecules are amines. This further illustrates the different chemical space of these datasets. The enriched fragments shown in Supplementary Table 7 are important because knowing which fragments are enriched in phenotypically active compounds can help with the design of new screening libraries.

4.3.2 CHEMOGENOMIC SPACE ANALYSIS

4.3.2.1 TARGET PREDICTION

In an effort to identify the target space of *Trypanosoma brucei*, *in silico* target prediction and a database search were carried out on 871 active screen compounds, 569 inactive screen compounds and 826 active NPs. The NP compounds were enriched against a background of over 2,000,000 compounds from PubMed using PIDGIN v2[143] to produce 186 enriched targets (see Methods for details). At least one target was predicted for 772 compounds of the active screen dataset and 391 compounds for the NPs, thus 11.4% of the screen compounds and 2.5% of the NPs were outside the applicability domain of the model. 1,544 proteins were predicted for the active screen compounds of which 1,539 were predicted for non-trypanosome organisms and 1,646 for the NPs, 1,640 of which were for non-trypanosome organisms. An orthologue search for these proteins was carried out on Panther DB to map 101 proteins of the 131

enriched predicted proteins for the screening compounds and 275 for the NPs; this mapping was six for the bioactivity search for the screen compounds and 109 for the NP bioactivity search. To obtain a prioritised set of targets, an enrichment calculation was carried out between the predicted targets of the active dataset and the inactive dataset of the screen compounds. This approach takes advantage of the negative dataset, to further prioritise the targets. The active compounds enriched 131 proteins over the inactive compounds and only 5 proteins were predicted to bind only the active compounds. A full list of enriched targets is provided in the Appendix of the thesis, Supplementary Table 11 and 13.

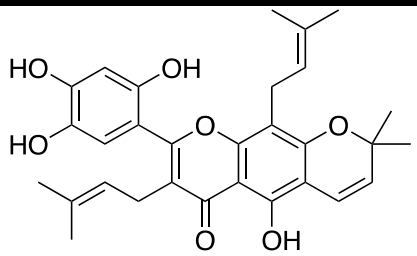
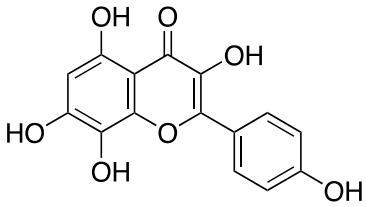
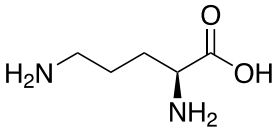
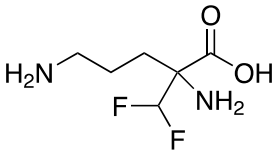
In this work, we assume shared bioactivity between targets from different organisms, e.g *Homo sapiens*, *Rattus norvegicus*, *Bos taurus* etc, and their orthologues in *T. brucei*. A recent study[143] found that annotations across orthologues are overall compatible, where it was found that only 1,363 of 124,540 (1.2%) orthologue bioactivities have conflicting annotations with the corresponding compounds inactive in humans. At the target level, 75.9% human-orthologue HomoloGene target pairs were found to be not conflicting. Another study[363] on ChEMBL bioactivity data found a statistically significant relationship between bioactivity in human and rat targets ($R=0.71$, $p<2e^{-16}$). Furthermore, a recent study involved carrying out a systematic search for bioactive small molecules shared by orthologous targets[364]. This study identified compound-orthologue pairs, covering 938 orthologues, 358 unique targets across 98 organisms. Of these, a total of 158 orthologous target pairs involving human orthologues were identified.[364] On the other hand, a study[365] which generated both phylogenetic and bioactivity tree representations of kinases found that in 57% of the studied cases, kinases that cluster together in protein structure space do not necessarily cluster together in bioactivity space. This study highlights that implicit assumptions of bioactivity across orthologues cannot be assumed based on protein structure similarity. Taken together, the results from these studies show that bioactivity data can be extrapolated with some confidence from one organism to another. This has been demonstrated by a study[337] that mined *Homo sapiens* bioactivity data in combination with structural and historical assay space searches to identify active target-compound links. The data from the targets identified from *Homo sapiens* targets was used to propose the equivalent *Mycobacterium tuberculosis* targets via an orthology search.

4.3.2.1.1 Checking confidence of predictions before analysing target prediction results

We carried out a retrospective confidence check of our predictions in three ways. Firstly, we compared the predicted results to the known targets of trypanocidal drugs in the market used against HAT (Eflornithine, Pentamidine, Suramin, Melarsoprol and Nifurtimox), shown in Supplementary Table 8. Currently, the only validated drug target of any of the drugs is ornithine decarboxylase, the target of eflornithine, which was correctly predicted for Eflornithine at a tpr >0.9.

Secondly, we compared the chemical similarity of NP in the dataset to experimentally validated trypansomal target inhibitors. We looked at two validated trypanosomal targets and compared their experimental inhibitors with those predicted from the NP dataset. First, we looked at ornithine decarboxylase (ODC), which is the drug target of the suicide inhibitor Elfornithine, in which 2 drug moieties irreversibly bind ODC and physically block ornithine from binding. Ornithine decarboxylase is an enzyme that catalyses the first reaction in polyamine synthesis[366]. Polyamines are (i) responsible for stabilising the structure of DNA, (ii) responsible for the DNA double-strand-break repair pathway, and (iii) antioxidants. Lack of ornithine decarboxylase leads to DNA damage induced apoptosis. Ornithine decarboxylase is also targeted by the molecule Heterophyllin, found in the NP dataset. The flavanol Herbacetin has previously been shown to allosterically inhibit ornithine decarboxylase[367] and we hypothesise that this is also the case for Heterophyllin, since the compounds are similar in structure. The structures of these compounds are shown in Table 4:3.

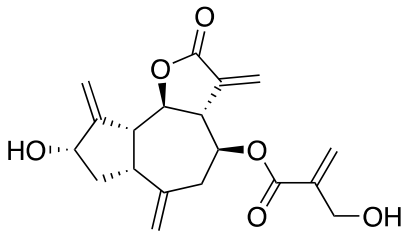
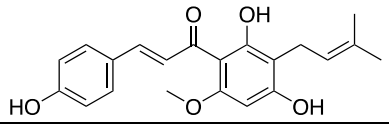
Table 4:3 Structure of Ornithine (natural ligand of ornithine decarboxylase), Eflornithine (Stage 2 Human African trypanosomiases drug) Heterophyllin (NP previously shown to bind ornithtine decarboxylase) and Herbacetin (NP predicted to bind ornithime decarboxylase).

Name	Structure
Heterophyllin	
Herbacetin	
Ornithine	
Eflornithine	

We identified mitochondrial Trypanothione reductase (TryR) (Tb10.406.0520) as a potential target of the NPs. Trypanosomes have a unique metabolic redox metabolism called the trypanothione redox metabolism. This has been investigated as a potential therapeutic target due to its essentiality for the survival of the trypanosome[39]. Trypanothione is responsible for defence against oxidative stress, therefore enzymes that make and use it can be targeted. Despite numerous efforts, no clinical compounds have been developed due to the presence of a large hydrophobic site in TryR. The NP cynaropicrin interacts with targets in the trypanosome redox pathway including ornithine decarboxylase, trypanothione reductase, trypanothione synthase and glutathione-S-transferase to produce anti-trypanosomal activity[368]. Cynaropicrin has been shown to exhibit this anti-trypanosomal activity via its α,β-unsaturated methylene moiety which acts as Michael acceptor for glutathione and trypanothione, thus depleting intracellular glutathione and trypanothione[369]. Michael addition is the nucleophilic addition of a carbanion or another nucleophile to an α,β-unsaturated

carbonyl compound. Xanthohumol also contains an α,β-unsaturated methylene moiety and may thus act via the same mechanism as Cynaropicrin. Xanthohumol has previously been shown to modify both NF-κB[370] and Keap1[371] protein by acting as a Michael acceptor to cysteine residues in these proteins. The structures of the compounds Xanthohumol (identified from the NP dataset) and Cynaropicrin (experimentally validated trypanothione redox modulator) are shown in Table 4:4. This suggests the mechanism of action of the phenotypically active NP (Xanthohumol) from our dataset.

Table 4:4 Structure of Cynaropicrin, a known modulator of the trypanothione redox pathway, and Xanthohumol, a NP predicted to modulate trypanothione reductase.

Name	Structure
Cynaropicrin	
Xanthohumol	

These two examples demonstrate that we can use the target prediction results with some confidence in our further analysis, i.e. biological process and mechanism of action elucidation.

In order to prioritise the predicted targets for both datasets, and understand their underlying molecular mechanism of action, we proceeded to carry out target predictions for the five marketed trypanocidal drugs. This analysis allows us to ascertain whether potential new molecules have predicted targets that overlap with the predicted targeted of these validated drugs, providing greater insight into their mechanism of action. We also constructed a list of essential targets of the whole *T.brucei* genome and compared the overlap with our predicted targets from both datasets as well as the 5 drugs.

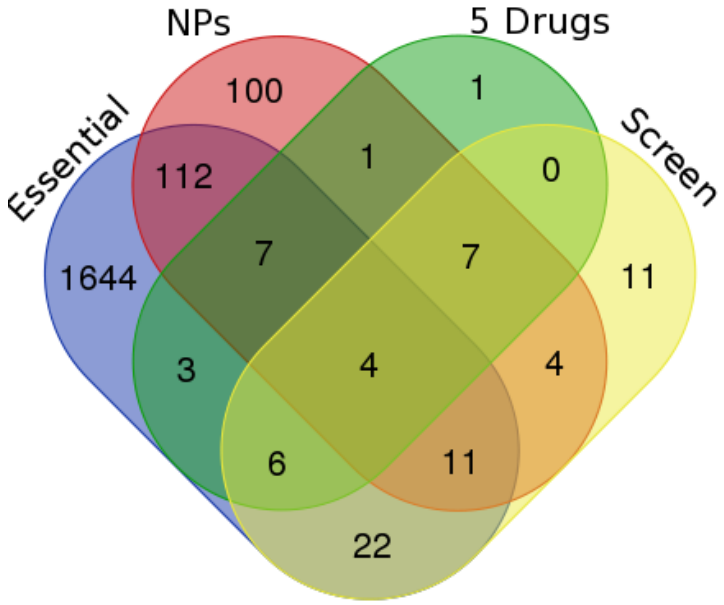


Figure 4:4 Overlap of genes between the different datasets. It can be seen that NPs share 134 predicted targets with the TDR Essential dataset. The Screen Hits targets share 43 targets with the TDR Essential targets. There is an overlap of 19 targets between the 5 HAT drugs and the NP targets, and 17 targets between the Screen HITS targets and the 5 Drugs.

From Figure 4:4 we see that 19 of the targets predicted for the natural products (NPs) are shared with the 5 HAT drugs and of these, 11 are in the TDR Essential set. This overlap between the predicted targets of the 5 marketed drugs, the TDR essential drugs and the predicted targets of the compounds from our datasets serves to increase confidence in the predictions. Of the 134 targets shared between NPs and TDR Essential, there were targets essential for tryapanosome survival. One of these targets shared between NPs and the TDR essential list is rhodesain. This cysteine protease plays a role in impassivity (allows BBB crossing), immune evasion (turnover of VSGs) and is responsible for degrading host immunoglobulins[37]. Allicin has been found to bind this target with a K_i value of 5.31µm from the NP Bioactivity dataset. The primary carbon atom in the vicinity of the thio-sulphinate sulphur atom is attracted by the cysteine residue in the active site of rhodesain. Another protein identified as a target for the phenotypically active compounds is the flagellum surface protein, voltage-dependent calcium channel type A subunit alpha-1 (Tb10.70.4750) identified in both datasets (including the screen compounds). This protein is present in the flagellar attachment zone (FAZ), which is a unique feature of *Trypanosoma brucei*

trypanomastid. Knockdown studies of this gene resulted in flagellar detachment and deficient growth of the trypanomastid[372]. Another family of targets predicted that are also essential are the heat shock proteins. The HAT drug Suramin acts on HSPs. With the exception of ornithine decarboxylase, and the heat shock proteins, none of the other trypanosomal genes predicted for the NPs or SHs are targeted by drugs in the market.

Here we have predicted that the phenotypically active compounds from both the NP dataset and SH datasets exert their anti-trypanosomal activity by modulating targets that are essential to the survival of the parasite.

4.3.2.2 Suitability of phenotypically active compounds from SM and NP datasets for Stage 2 HAT

We proceeded to look into the possibility that these compounds have the potential to be active against Stage 2 HAT. Stage 2 HAT is characterised by the traversal of the BBB by the parasite. Unlike other parasitic diseases, trypanosome traversal is not dependent on the level of parasitaemia[373], rather, it depends on the host immune response[373]. In order to achieve Stage 2 activity, the compounds must (i) be able to cross the BBB and modulate targets that are essential for the survival of the parasite and/or (ii) act on human targets involved in the host response system facilitating parasite traversal, e.g. chemokines, TNF- α and interferons[374]. Supplementary Table 14 shows a list of the compounds with favourable plogBB values and the essential *Trypanosoma brucei* targets they are predicted to modulate. We therefore deem these compounds favourable for prioritisation for *in vivo* Stage 2 HAT models.

4.3.2.3 Analysing Multi-target activity of phenotypically active small molecule hits compounds

It is generally advantageous for therapeutic drugs to target multiple proteins involved in multiple stages of the life cycle [375] as trypanosomes have multiple host life cycles [376]. To visualise if the phenotypically active compounds from the SH dataset display multiple protein modulation we constructed a target network. The target network showing the connections of the 871 phenotypically active SH dataset compounds to the enriched set of targets is shown in Figure 4:5. The figure also shows the compounds

that are predicted to penetrate the BBB (coloured in red). This is important for the treatment of the second stage of the disease, as the parasite crosses into the brain and causes mental deterioration followed by the induction of a coma and eventually leading to death. Eflornithine is predicted to have a log BB of -0.328 and is one of the drugs suitable for use in the second stage of the disease.

From the network we identified several multi-target compound hubs, the most prominent of which are labelled in Figure 4:5. In this case, a hub is a node for which the number of links exceeds the average number of degrees (connectedness), which in this case is 22.305. There were 78 targets that exceeded the average number of degrees, ranging from 22.3 to 90 degrees (connections to compounds). The transporters had the highest number of degrees at 90. The identified hubs show the most promiscuous targets binding to the active compounds in the SH dataset. We predict that compounds from the SH dataset are exerting their activity by modulating the targets in these hubs. We can also see that unlike the compounds predicted to bind transporters and oxidoreductases, some of the compounds binding the kinases have poor distribution in the brain.

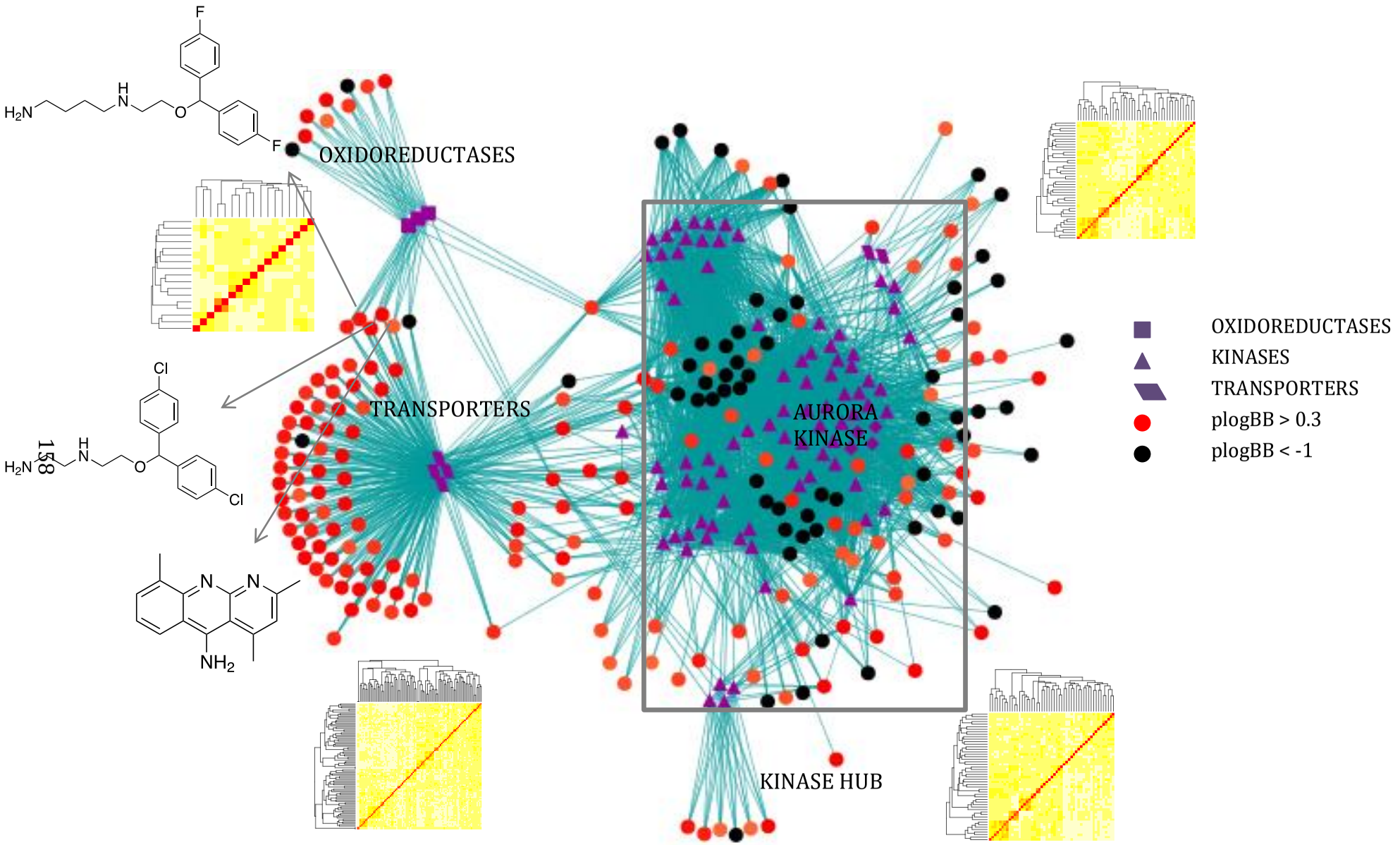

OXIDOREDUCTASES
TRANSPORTERS
AURORA KINASE
KINASE HUB
OXIDOREDUCTASES
KINASES
TRANSPORTERS
plogBB > 0.3
plogBB < -1

Figure 4:5 Visualisation of the compound-target network of predicted targets of the Screen HITS compounds. Node shapes encode protein class. All target nodes are shown in purple. The circular nodes represent the compound nodes and their colour represents those compounds' ability to: readily cross the BBB plogBB > 0.3 (red), somewhat cross the BBB plogBB > -1 (orange) and those with poor distribution in the brain (black). The heatmaps that are shown for each target class represent the structural similarity of the compounds predicted to bind to these targets. They have low similarity (as calculated from the Tanimoto similarity of their Morgan fingerprints) indicating that these targets can be modulated by compounds having diverse structures and hence diverse ADME properties. This indicates that diverse compounds modulate kinases in the network.

The first multi-target hub identified is of the compounds predicted to bind transporters. Several glucose transporters were identified as targets from the phenotypically active compounds, shown in Supplementary Tables 12 and 13, namely glucose transporter (Tb10.6k15.2030), glucose transporter, putative (Tb927.4.2290) and a hexose transporter (Tb10.6k15.2040), all belonging to the family "facilitated glucose transporter protein 1". This is significant because one of the most important pathways in African trypanosomes is glycolysis, the enzymes of which are packaged in the glycosome. Blood stage African trypanosomes rely solely on glycolysis for the production of ATP, thus enzymes and transporters involved in this pathway represent viable drug targets. It is ideal to have compounds, some of which are shown in Figure 4:4, with the ability to cross the BBB [377] that will target the transporters.

Another set of multi-target compounds of particular relevance to HAT were identified in the kinase hub of Figure 4:5. It has been suggested that it would be advantageous to develop kinase inhibitors that target multiple kinases within a family[378] to reduce resistance that arises from point mutations at the residues involved in the binding site between the compound and the kinase[379]. In our predictions we have identified targets including serine/threonine kinases, Mitogen activated kinases, and more (shown in Supplementary Tables 12 and 13). Similar targets were predicted for *Plasmodium falciparum* [338], including a number of putative protein kinases, serine/threonine kinases and MAP kinases. A potential therapeutic kinase target that was predicted (for compounds shown in Table 3 – CD) in the kinase hub was Aurora Kinase A, which plays an important role in metaphase-anaphase transition and the initiation of cytokinesis[380]. Knockdown studies showed an essential contribution to infection in mice[380]. Small molecule inhibitors, e.g. hesperidin, have been shown to inhibit this protein[380]. 276 phenotypically active compounds from the screen dataset were predicted to bind Aurora A; these molecules are shown using the smiles representation in Table 3 – CD. The fact that they inhibit Aurora A is relevant because it means that we can begin to understand the putative mechanisms of action of these compounds.

Glycogen synthase kinase 3 (GSK3) was also identified as a target of 37 compounds. This target is present in the list of highly prioritised targets for trypanosoma drug discovery. It is expressed in the bloodstream form of Trypanomastids. RNAi

knockdown studies of this gene resulted in defects in mitosis and cytokinesis, thus leading to GSK3 being investigated as a therapeutic drug target[381, 382]. During host infection, i.e. bloodstream form, the trypanosome relies on the glycolysis of host sugar for the production of ATP[383]. Consequently, enzymes involved in the process have been investigated as potential therapeutic targets. Phosphofructokinases catalyse the phosphorylation of fructose-6-phosphate to fructose-6-biphosphate, which is the key rate-limiting step of glycolysis (the sole pathway for ATP production in African trypanosomes). As a consequence, these kinases are being actively pursued as targets[384], and it is promising that they are predicted by our analysis as members of the kinase hub.

It is important to note that there are limitations to targeting kinases. Humans also have numerous kinases and the kinome is important for human survival, hence it is important to avoid cross-reactivity with members of the human kinome. We are attempting to identify targets in *Trypanosoms brucei*, which is evolutionarily very distant from humans. This evolutionary distance suggests that there may be differences between the human and trypanosome kinome that can be exploited. For example, the selectivity of Eflornithine to trypanosomal ornithine decarboxylase (not a kinase) arises from the difference in turnover for this enzyme between humans and the parasite[385]. In humans the turnover is much faster[385], and so suicidal binding[386] of Eflorntihine to the tryapnsome enzyme affects the trypanosome more than the human enzyme. Human enzyme activity is maintained by a protein whose turnover is significantly different from the trypanosome version. These evolutionary differences will become important at the lead optimisation stage, when these compounds would need to be modified in such a way that they do not interact in a detrimental way with the human kinome. In summary our analysis shows that individual phenotypically active compounds are predicted to modulate multiple targets in the trypanosome parasite. These results suggest that a number of these compounds may make good potential therapeutic candidates.

4.3.2.4 ENRICHED BIOLOGICAL PROCESSES

Target prediction alone does not provide the full picture with respect to understanding the mode of action of phenotypically active compounds. To address this question in more detail we analysed the enriched *Trypanosoma brucei* GO biological processes of the predicted targets from the two datasets. The results are represented in Figure 4:6.

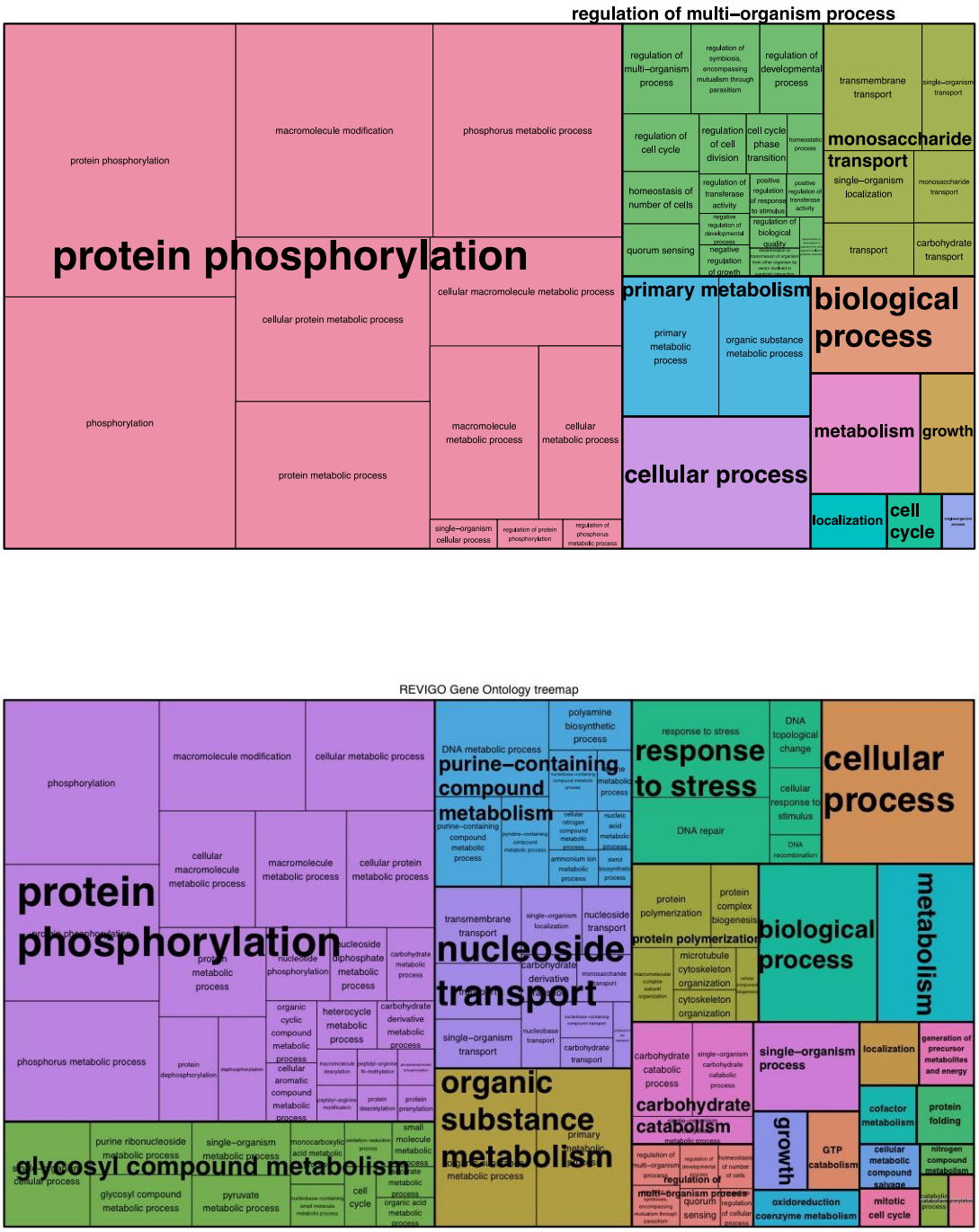


Figure 4:6 Treemap showing the biological processes GO term clusters. Each rectangle is a single cluster representative. The size of the rectangle represents the p-value of the GO term (all levels were considered). The top figure corresponds to the biological processes of the small molecule gene list whereas the bottom figure corresponds to the biological processes enriched in the NP genelist. It can be seen that the major difference between the two is the absence of "response to stress" and "response to stimulus" in the SM biological process profile. "Biological process" as a GO term refers to annotation of gene products whose biological process is unknown. The p-values corresponding to these biological processes are shown in Supplementary Tables 15 and

Biological processes involved in protein phosphorylation are enriched in both sets of gene lists (these gene lists contain the predicted targets and known biological targets of both datasets). Protein kinases, which are the enzymes responsible for protein phosphorylation, play an important role in many cellular processes. These include, but are not limited to, cell cycle propagation and differentiation as well as transcription control. Kinases have been the focus of current kinetoplastid drug discovery programs[361, 387]. The "kinase library" subset of the SH dataset contained compounds that were likely to hit kinases[388]. Here we have identified the genes they are predicted to modulate (Supplementary Tables 10-13) and the biological processes that they enrich in order to have the observed cidal and static effects reported by Thompson *et al*[387].

We can see from Figure 4:5 that processes which are essential to the survival of the parasite such as glucose and folate metabolism[389, 390] as well as cell death are predicted to be modulated by compounds in both datasets. This is shown by the presence of the GO terms "growth", "metabolism" and the "cell cycle" in both the top and bottom panels of Figure 4:5. The cell cycle has been identified as an important target for exploitation in other parasitic diseases caused by Plasmodia, Trypanosoma and Leishmania[391]. An interesting set of biological processes that was found for both sets is "regulation of symbiosis, encompassing mutualism through parasitism" and "modulation of development of symbiont involved in interaction with host". Disruption of these genes has been shown to be vital for quorum sensing signalling in the trypanosome[392]. Deletion of one of these genes, the differentiation inhibitory kinase (Tb927.11.9270), has been shown to increase the rate of differentiation between the bloodstream form (pathogenic form) to the stumpy non-dividing form[393]. The genes in each dataset responsible for this observation in the small molecules dataset are:

- CMGC/DYRK protein kinase, putative (Tb927.10.15020),
- NEK family Serine/threonine-protein kinase, putative (Tb927.10.5940),
- Serine/threonine-protein kinase NEK17, putative (Tb927.10.5950),
- Differentiation inhibitory kinase (Tb927.11.9270),
- 5'-AMP-activated protein kinase catalytic subunit alpha, putative (Tb927.3.4560) ,
- Repressor of differentiation kinase 2 (Tb927.4.5310).

In the natural product (NP) dataset, the genes were:

- 5'-AMP-activated protein kinase catalytic subunit alpha, putative (Tb927.3.4560),
- Serine threonine-protein phosphatase PP1, putative (Tb927.4.3620),
- Serine threonine-protein phosphatase PP1, putative (Tb927.4.3630),
- Serine threonine-protein phosphatase PP1, putative (Tb927.4.3640).

Note that all these targets are putative kinases, further reinforcing the importance of kinases as drug targets for *Trypanosoma brucei*. The most common scaffolds of the compounds predicted to bind the genes enriched in these processes is shown in Table 4 CD. We can see that the compounds share no scaffolds and that they modulate a completely different set of trypanosomal genes, yet they are involved in the same biological process. An approach to modulate different kinases simultaneously has been suggested to be an advantage for kinetoplastid drug discovery as discussed above. A combination of molecules from the NP and SH datasets will likely achieve this as we have shown they target different genes involved in the same biological processes.

Other processes that are of particular importance to trypanosome integrity were only enriched in the SH gene list, e.g. “homeostasis” and “monosaccharide transport”, the importance of which was discussed above. In particular, Ca^{++} homeostasis, which is an essential messenger, is important as its disruption causes apoptosis and necrosis (i.e. disruption is lethal to the trypanosome). We have already shown that small molecules at the molecular level are predicted to regulate transport mechanisms, and here we see the biological processes that are affected. Other biological processes were only enriched for the NP gene list, e.g. “cytoskeleton organisation” and “sterol metabolism”. This finding is in agreement with previous studies, where sterol metabolism has been investigated as a target for the closely related *Trypanosoma cruzi*[394] and Leishmania species[395]. Psilostacyn C, a NP sesquiterpene lactone, induced cell death by apoptosis in *T. cruzi* by interfering with sterol synthesis[396]. One of the drugs used in Leishmaniasis is the NP Amphotericin B, which has a high affinity for egosterol and thus inhibits sterol metabolism[397]. The process of cytoskeleton organisation is important as it is responsible for orchestrating the extreme changes in cellular

morphology of the trypanosome, during both its life cycle and various different cell cycles[372, 398]. These morphological changes require a high level of integration and coordination. Disrupting these processes will therefore lead to growth defects[372].

Beyond these biological processes, we note that "response to stress", "organic substance metabolism" and "purine-containing compound metabolism" are absent from the SH map in Figure 4:6. One reason that these processes are present in the NP map is that those compounds responsible for predicting targets enriched in these processes are also responsible for similar processes in the plant that the compounds have been extracted from. It is well documented that secondary metabolites in plants are responsible for defence against stress[399]. For example, compounds responsible for "cellular response to DNA damage stimulus" included flavonoids, glucosides and flavonoids, e.g. caffeic acid, sennoside A, sennoside B, shickimic acid and tannic acid. These secondary metabolites protect plant cells from UV-a and UV-B stress via various mechanisms[400]. Furano-coumarins were responsible for "DNA repair response". The furano-coumarins are produced in plants in response to UV-A damage. They are activated by UV-A and lead to cell death by blocking transcription through inserting themselves into the DNA double-helix and binding to the pyramidine bases[400].

Here we have shown that phenotypically active compounds are predicted to modulate biological processes that are essential for the survival of the *Trypanosoma brucei* parasite. We have predicted their target genes, and identified the biological processes that are enriched for those genes. This analysis provides a deeper layer of understanding regarding the mechanism of action of the phenotypically active anti-trypanosomal compounds in the two datasets.

4.4 CONCLUSION

In this study we have shown that the chemical space spanned by our small molecule (SH) and natural product (NP) datasets are quite different, as shown by the spread of the compounds on the MDS plot. The density plot in Figure 4:3 reveals that the NP molecules have less chemical similarity to each other than the small molecule dataset compounds do, despite the fact that the SH datasets compounds also display low similarity to one another. This means that while both datasets are highly diverse in chemical space, the NP dataset is more diverse than the small molecule one.

We have made use of negative and orthologue data to identify therapeutic targets of phenotypically active HITS and NPs. Our analysis identified overlaps between the sets of predicted targets and those genes that are essential for the trypanosome, identifying these genes as targets that should be highly prioritised. This analysis is timely and relevant, in particular because none of these genes are currently targeted by drugs in the market.

We predicted the activity of the small compounds in second stage HAT by predicting their logBB values, and found not only that some compounds are predicted to penetrate the blood brain barrier (important for Stage 2 of the disease) but also that they are predicted to modulate multiple targets. We identified multi-target activity of the small molecules, where we show that they are predicted to modulate multiple orthologues of the same kinase, e.g. Aurora Kinase A and the MAP Kinases. We have identified distinct but therapeutically relevant target spaces for members of both the NP and SH datasets of compounds.

We have taken this further and explored the biological processes modulated by these compounds. The most interesting finding was the enrichment of processes affecting host-parasite interaction, namely “regulation of symbiosis, encompassing mutualism through parasitism” and “modulation of development of symbiont involved in interaction with host”. We found that compounds from both datasets modulate this process through their predicted activity against different targets.

CONCLUSION

Throughout the studies in this thesis the aim was to utilise *in silico* target prediction to understand the mechanism of action of African natural products.

In Chapter 2, a Random Forest algorithm was used to predict targets and pathways modulated by natural products from African medicinal plants with anti-cancer activity. From our study we were able to establish several links between the suggested MOAs of the natural products with experimental evidence. Compounds from plants used in cancer were predicted to bind primary cancer targets, e.g. the apoptosis regulator Bcl2, as well as targets involved in the metabolism and hence resistance to cancer drugs, e.g. CYP1B1. We also identified targets that may exhibit novel mechanisms of action that are not currently targeted by drugs in the market, e.g. Induced myeloid leukaemia cell differentiation protein Mcl-1 and Tankyrase 1. Furthermore, we identified the pathways that these compounds modulated and not only were some directly involved in cancer, e.g. Apoptosis pathways, but some were not modulated by drugs in the market, e.g. the Kaep1-Nrf2 pathway. Similar results were obtained in our case study for *Psorospermum aurantiacum*, where the primary targets associated with the medicinal uses of the plant were predicted, e.g. protein kinase C gamma type (linked to skin infections), oestrogen receptor β (linked to infertility), Mcl-1 (linked to cancer). However, there are limitations of applying an *in silico* target prediction algorithm trained on ChEMBL data to natural products, specifically the applicability domain (the physic-chemical structural or biological domain for which it is valid to make predictions for new compounds). A target prediction algorithm trained on natural product data will go towards addressing this problem. This will require access to complete (not sparse) databases containing natural product structure data as well as bioactivity data. This should be possible in the near future as more databases are curated and updated, e.g. NANPDB, NPASS, etc.

In our novel approach to integrate phylogenetic data with *in silico* target prediction we were able to identify relationships between the phylogeny of a medicinal plant and its medicinal use. Plant families in this study mostly cluster together in structure space, predicted target space and phylogeny space. Furthermore, over-represented plant

families were identified for medicinal uses (cancer, malaria and HAT) in Africa including Acistrocladaceae, Clusiaceae and Rutaceae. The over-represented presence of medicinal plants within these families was directly linked to the unique phytochemicals produced by these plants. Taken together these findings provide a basis for predicting the use of a medicinal plant based on its phylogenetic relationship to other medicinal plants. Limitations to this study include incomplete data and experimental annotations of the NPs from the medicinal species of Africa. As more information is curated and added to the databases, more concrete results can be obtained.

Phenotypic studies have identified both natural products and small molecules with anti-tryapnosomal activity. In Chapter 4, *in silico* target prediction, combined with an orthology search, revealed predicted *Trypanosoma brucei* targets and biological processes modulated by the phenotypic compounds that are essential for the survival of the trypanosome including glycogen synthase kinase. In addition to predicting the targets of these compounds, we were able to predict their activity in stage 2 HAT by predicting their ability to cross the blood brain barrier. This study elucidates the mode of action of ANPs used in HAT as well as identifying the difference in target space between NPs and small molecules. The major limitation in this study is the extrapolation of orthologue data from different species to *Trypansoma brucei.* To address this limitation, a similarity cut-off of the predicted target to the tryapnosomal target can be applied, e.g. >80% sequence similarity.

Despite the limitations discussed above, *in silico* target prediction can be used to elucidate and understand the mechanism of action of African natural products.

FUTURE WORK

In Chapter 2 we attempted to elucidate the mechanism of action of NP from TAMs by showing that targets predicted to be modulated by NPs from TAM play a role in cancer. We have also shown the molecular pathways that these NP are predicted to modulate. In order to fully understand the mechanism of action of these compounds it is vital to experimentally validate these predictions. The next step after these predictions is to predict binding and interaction of the compounds to the targets. Molecular docking of NPs to targets will give us a score of the predicted binding affinity and binding mode of NPs to their target proteins. Further steps would be taken for the most promising of these predictions to experimentally validate binding. Should these opportunities become available the two targets to prioritise would be Tankyrase 1 and Thioredoxin reductase 2. Assays to validate NP binding to Tankyrase 1 include those developed by Thomson *et al*[401] and assays that can be carried out for Thioredoxin reductase 2 include those described in [402-404].

As more NP datasets are curated and made public it will be valuable to build a DNN model trained on NP data. DNNs have been shown to outperform RF[102, 108] on QSAR based protocols when trained on ChEMBL data[108] but they require a large amount of descriptors and features to train a model, which are not available yet for NPs. The applicability domain of the model will improve, as the model will be trained on compounds with similar chemical space to the test compounds.

In Chapter 3, we identified over-represented families in Africa used against cancer malaria and HAT. This knowledge can be use to prioritize screening of plants from over-represented families e.g. Acistrocladaceae and Dioncophyllacea for anti-trypanosomal activity. It would also be useful to identify biosynthetic gene clusters (BGC) in plants where it was found that medicinal activity was due to unique metabolites e.g. NIQs in Ancistrocladaceae, Xanthones in Clusiaceae and Acridones in Rutaceae. Identifying these BGCs in one of the plants known to produce a medicinally active NPs is useful when genome mining other plants in the same family for novel medicinally active NPs[405, 406].

In Chapter 4 we predicted the biological targets of phenotypically active compounds. The next step would be to validate these predictions. This can be achieved by first carrying out docking studies to determine binding affinities and binding modes of compounds predicted to bind two prioritized targets. Compounds to be prioritized would be those that are predicted to cross the BBB, have good ADMET properties and predicted to modulate targets essential to the *Trypanosoma bruceii* parasite. The two targets that can be prioritized from our results are Glucose transporter and hexose transporter inhibitor[407, 408] and GSK3[382, 409].

Another extension of this work would be to expand on the network analysis carried out in this Chapter to uncover novel druggable target space in *Trypanosoma bruceii*. To carry this out, a network of the *Trypanosoma bruceii* proteome, obtained from STRING[410] database will be constructed. The STRING database contains information predicted and known protein-protein interactions. This network can be mapped out, visualized and analyzed using Cytoscape[358]. Several functions scores can be calculated including the Betweeness Centrality, which shows which nodes (in this case targets) are more likely to be in communication paths between other nodes. It is useful in determining which targets can be modulated where the protein network would break apart i.e. modulating a target with high Betweeness Centrality will cause a larger downstream effect. This measure demonstrates how likely the target is to be the most direct route between two targets in the network. Other useful measures that can be calculated include Eigenvectors, which determine how well a target is connected to other well connected targets and Degree which shows how many targets a particular target interacts with directly.

The next step would be to identify which of the compounds in our SH and NP dataset are predicted to (or experimentally known to) modulate these targets (targets with high Betweeness Centrality, Eigenvectors and Degree). It would also be useful find out if the compounds in the SH and NP datasets bind orthologues in other kinetoplastids of these targets. Experimentally validating these predictions will potentially uncover novel target space in *Trypanosoma bruceii*.

REFERENCES

1. Solecki, R. S., Shanidar IV, a Neanderthal Flower Burial in Northern Iraq. *Science* **1975**, 190, 880-881.
2. Sumner, J., *The natural history of medicinal plants*. Timber Press: 2000.
3. Ebbell, B., The Ebers Papyrus. *The Greatest Egyptian Medical Document. London: H. Milford and Oxford University Press* **1937**, 17, 123.
4. Elufioye, T. O.; Badal, S. Chapter 1 - Background to Pharmacognosy. In *Pharmacognosy*; Academic Press: Boston, 2017, 3-13.
5. Bannerman, R. H., In; Geneva, Switzerland, World Health Organization, 1983: Switzerland, 1983, p. 318-327.
6. Okpako, D. T., Traditional African medicine: theory and pharmacology explored. *Trends in Pharmacological Sciences* **1999**, 20, 482-485.
7. Weiss, B., African Divination Systems: Ways of Knowing. PHILIP M. PEEK. *American Ethnologist* **1994**, 21, 951-952.
8. Chan, K., Progress in traditional Chinese medicine. *Trends in Pharmacological Sciences* **1995**, 16, 182-187.
9. Obeyesekere, G., The theory and practice of psychological medicine in the Ayurvedic tradition. *Culture, medicine and psychiatry* **1977**, 1, 155-81.
10. Wilson, B., *Rationality*. Wiley-Blackwell: 1991.
11. Rasamiravaka, T.; J, K.; Pn, O.; Amuri, B.; L, B.; Jb, K.; Kiendrebeogo, M.; C, R.; El Jaziri, M.; Williamson, E.; Duez, P., *Traditional African medicine: From ancestral knowledge to a modern integrated future*. 2015; Vol. 350, p S61-S63.
12. Hanahan, D.; Weinberg, R. A., The hallmarks of cancer. *Cell* **2000**, 100, 57-70.
13. World Health Organisation http://www.afro.who.int/health-topics/cancer (23.03.2018),
14. Hanahan, D.; Weinberg, R. A., Hallmarks of cancer: the next generation. *Cell* **2011**, 144, 646-74.
15. ; Hartwell, J. L., *Plants Used Against Cancer: A Survey (Bioactive Plants, Vol 2)*. Quarterman Publications: 1984.
16. Graham, J. G.; Quinn, M. L.; Fabricant, D. S.; Farnsworth, N. R., Plants used against cancer - an extension of the work of Jonathan Hartwell. *J Ethnopharmacol* **2000**, 73, 347-77.
17. Newman, D. J.; Cragg, G. M., Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products* **2016**, 79, 629-661.
18. Wansi, J. D.; Mesaik, M. A.; Chiozem, D. D.; Devkota, K. P.; Gaboriaud-Kolar, N.; Lallemand, M.-C.; Wandji, J.; Choudhary, M. I.; Sewald, N., Oxidative Burst Inhibitory and Cytotoxic Indoloquinazoline and Furoquinoline Alkaloids from Oricia suaveolens. *Journal of Natural Products* **2008**, 71, 1942-1945.
19. Jordan, M. A.; Wilson, L., Microtubules as a target for anticancer drugs. *Nat Rev Cancer* **2004**, 4, 253-65.
20. Salminen, A.; Lehtonen, M.; Suuronen, T.; Kaarniranta, K.; Huuskonen, J., Terpenoids: natural inhibitors of NF-kappaB signaling with anti-inflammatory and anticancer potential. *Cellular and molecular life sciences : CMLS* **2008**, 65, 2979-99.
21. Simoben, C. V.; Ibezim, A.; Ntie-Kang, F.; Nwodo, J. N.; Lifongo, L. L., Exploring Cancer Therapeutics with Natural Products from African Medicinal Plants, Part I: Xanthones, Quinones, Steroids, Coumarins, Phenolics and other Classes of Compounds. *Anti-cancer agents in medicinal chemistry* **2015**, 15, 1092-111.

22. Nwodo, J. N.; Ibezim, A.; Simoben, C. V.; Ntie-Kang, F., Exploring Cancer Therapeutics with Natural Products from African Medicinal Plants, Part II: Alkaloids, Terpenoids and Flavonoids. *Anti-cancer agents in medicinal chemistry* **2016**, 16, 108-27.
23. Ibrahim, M. A.; Mohammed, A.; Isah, M. B.; Aliyu, A. B., Anti-trypanosomal activity of African medicinal plants: a review update. *J Ethnopharmacol* **2014**, 154, 26-54.
24. World Health Organisation Trypanosomiasis, human African (sleeping sickness). http://www.who.int/mediacentre/factsheets/fs259/en/ (June 2017),
25. Blum, J.; Nkunku, S.; Burri, C., Clinical description of encephalopathic syndromes and risk factors for their occurrence and outcome during melarsoprol treatment of human African trypanosomiasis. *Tropical medicine & international health : TM & IH* **2001**, 6, 390-400.
26. Barrett, M. P.; Boykin, D. W.; Brun, R.; Tidwell, R. R., Human African trypanosomiasis: pharmacological re-engagement with a neglected disease. *British Journal of Pharmacology* **2007**, 152, 1155-1171.
27. Jacobs, R. T.; Nare, B.; Wring, S. A.; Orr, M. D.; Chen, D.; Sligar, J. M.; Jenks, M. X.; Noe, R. A.; Bowling, T. S.; Mercer, L. T.; Rewerts, C.; Gaukel, E.; Owens, J.; Parham, R.; Randolph, R.; Beaudet, B.; Bacchi, C. J.; Yarlett, N.; Plattner, J. J.; Freund, Y.; Ding, C.; Akama, T.; Zhang, Y. K.; Brun, R.; Kaiser, M.; Scandale, I.; Don, R., SCYX-7158, an orally-active benzoxaborole for the treatment of stage 2 human African trypanosomiasis. *PLoS Negl Trop Dis* **2011**, 5, e1151.
28. Torreele, E.; Bourdin Trunz, B.; Tweats, D.; Kaiser, M.; Brun, R.; Mazué, G.; Bray, M. A.; Pécoul, B., Fexinidazole – A New Oral Nitroimidazole Drug Candidate Entering Clinical Development for the Treatment of Sleeping Sickness. *PLOS Neglected Tropical Diseases* **2010**, 4, e923.
29. Jefferson, T.; McShan, D.; Warfield, J.; Ogungbe, I. V., Screening and Identification of Inhibitors of Trypanosoma brucei Cathepsin L with Antitrypanosomal Activity. *Chem Biol Drug Des* **2016**, 87, 154-8.
30. Woodland, A.; Thompson, S.; Cleghorn, L. A. T.; Norcross, N.; De Rycker, M.; Grimaldi, R.; Hallyburton, I.; Rao, B.; Norval, S.; Stojanovski, L.; Brun, R.; Kaiser, M.; Frearson, J. A.; Gray, D. W.; Wyatt, P. G.; Read, K. D.; Gilbert, I. H., Discovery of Inhibitors of Trypanosoma brucei by Phenotypic Screening of a Focused Protein Kinase Library. *ChemMedChem* **2015**, 10, 1809-1820.
31. Field, M. C.; Horn, D.; Fairlamb, A. H.; Ferguson, M. A. J.; Gray, D. W.; Read, K. D.; De Rycker, M.; Torrie, L. S.; Wyatt, P. G.; Wyllie, S.; Gilbert, I. H., Anti-trypanosomatid drug discovery: an ongoing challenge and a continuing need. *Nat Rev Micro* **2017**, 15, 217-231.
32. Pink, R.; Hudson, A.; Mouries, M.-A.; Bendig, M., Opportunities and Challenges in Antiparasitic Drug Discovery. *Nat Rev Drug Discov* **2005**, 4, 727-740.
33. Cornelissen, A. W.; Bakkeren, G. A.; Barry, J. D.; Michels, P. A.; Borst, P., Characteristics of trypanosome variant antigen genes active in the tsetse fly. *Nucleic acids research* **1985**, 13, 4661-76.
34. Horn, D., Antigenic variation in African trypanosomes. *Molecular and biochemical parasitology* **2014**, 195, 123-129.
35. Donelson, J. E., Antigenic variation and the African trypanosome genome. *Acta Tropica* **2003**, 85, 391-404.
36. Gilbert, I. H., Drug Discovery for Neglected Diseases: Molecular Target-Based and Phenotypic Approaches. *Journal of Medicinal Chemistry* **2013**, 56, 7719-7726.

37. Ettari, R.; Tamborini, L.; Angelo, I. C.; Micale, N.; Pinto, A.; De Micheli, C.; Conti, P., Inhibition of rhodesain as a novel therapeutic modality for human African trypanosomiasis. *J Med Chem* **2013**, 56, 5637-58.
38. Paul, K. S.; Bacchi, C. J.; Englund, P. T., Multiple Triclosan Targets in Trypanosoma brucei. *Eukaryotic Cell* **2004**, 3, 855-861.
39. Patterson, S.; Alphey, M. S.; Jones, D. C.; Shanks, E. J.; Street, I. P.; Frearson, J. A.; Wyatt, P. G.; Gilbert, I. H.; Fairlamb, A. H., Dihydroquinazolines as a novel class of Trypanosoma brucei trypanothione reductase inhibitors: discovery, synthesis, and characterization of their binding mode by protein crystallography. *J Med Chem* **2011**, 54, 6514-30.
40. Frearson, J. A.; Wyatt, P. G.; Gilbert, I. H.; Fairlamb, A. H., Target assessment for antiparasitic drug discovery. *Trends in Parasitology* **2007**, 23, 589-595.
41. Hoet, S.; Opperdoes, F.; Brun, R.; Quetin-Leclercq, J., Natural products active against African trypanosomes: a step towards new drugs. *Nat Prod Rep* **2004**, 21, 353-64.
42. Ntie-Kang, F.; Telukunta, K. K.; Doring, K.; Simoben, C. V.; AF, A. M.; Malange, Y. I.; Njume, L. E.; Yong, J. N.; Sippl, W.; Gunther, S., NANPDB: A Resource for Natural Products from Northern African Sources. *J Nat Prod* **2017**, 80, 2067-2076.
43. Ntie-Kang, F.; Onguene, P. A.; Scharfe, M.; Owono Owono, L. C.; Megnassan, E.; Mbaze, L. M. a.; Sippl, W.; Efange, S. M. N., ConMedNP: a natural product library from Central African medicinal plants for drug discovery. *RSC Advances* **2014**, 4, 409-419.
44. Ntie-Kang, F.; Nwodo, J. N.; Ibezim, A.; Simoben, C. V.; Karaman, B.; Ngwa, V. F.; Sippl, W.; Adikwu, M. U.; Mbaze, L. M. a., Molecular Modeling of Potential Anticancer Agents from African Medicinal Plants. *Journal of Chemical Information and Modeling* **2014**, 54, 2433-2450.
45. Onguéné, P. A.; Ntie-Kang, F.; Mbah, J. A.; Lifongo, L. L.; Ndom, J. C.; Sippl, W.; Mbaze, L. M. a., The potential of anti-malarial compounds derived from African medicinal plants, part III: an in silico evaluation of drug metabolism and pharmacokinetics profiling. *Organic and Medicinal Chemistry Letters* **2014**, 4.
46. Mangal, M.; Sagar, P.; Singh, H.; Raghava, G. P.; Agarwal, S. M., NPACT: Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database. *Nucleic acids research* **2013**, 41, D1124-9.
47. Ganesan, A., The impact of natural products upon modern drug discovery. *Current opinion in chemical biology* **2008**, 12, 306-17.
48. Corson, T. W.; Crews, C. M., Molecular Understanding and Modern Application of Traditional Medicines: Triumphs and Trials. *Cell* **2007**, 130, 769-774.
49. Chen, X.; Zhou, H.; Liu, Y. B.; Wang, J. F.; Li, H.; Ung, C. Y.; Han, L. Y.; Cao, Z. W.; Chen, Y. Z., Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation. *British Journal of Pharmacology* **2006**, 149, 1092-1103.
50. Gilbert, B.; Alves, L. F., Synergy in plant medicines. *Current medicinal chemistry* **2003**, 10, 13-20.
51. Kimura, M.; Kimura, I.; Guo, X.; Luo, B.; Kobayashi, S., Combined effects of Japanese-Sino medicine 'Kakkon-to-ka-senkyu-shin'i' and its related combinations and component drugs on adjuvant-induced inflammation in mice. *Phytotherapy Research* **1992**, 6, 209-216.

52. Cott, J.; Misra, R., Medicinal plants: a potential source of new psychotherapeutic drugs. *Herbal Medicines for Neuropsychiatric Diseases: Current Developments and Research* **2013**, 51.
53. Ohta, A.; Uehara, M.; Sakai, K.; Takasaki, M.; Adlercreutz, H.; Morohashi, T.; Ishimi, Y., A combination of dietary fructooligosaccharides and isoflavone conjugates increases femoral bone mineral density and equol production in ovariectomized mice. *The Journal of nutrition* **2002**, 132, 2048-54.
54. Gottlieb, O. R.; Borin, M. R. d. M. B.; Bosisio, B. M., Trends of plant use by humans and nonhuman primates in Amazonia. *American Journal of Primatology* **1996**, 40, 189-195.
55. Nyasse, B. Overview of Current Drug Discovery Activities in Africa and Their Links to International Efforts to Combat Tropical Infectious Diseases. In *Drug Discovery in Africa*, Chibale, K.; Davies-Coleman, M.; Masimirembwa, C., Eds.; Springer Berlin Heidelberg: 2012; Chapter 1, 1-28.
56. Ntie-Kang, F.; Lifongo, L. L.; Simoben, C. V.; Babiaka, S. B.; Sippl, W.; Mbaze, L. M. a., The uniqueness and therapeutic value of natural products from West African medicinal plants. Part I: uniqueness and chemotaxonomy. *RSC Advances* **2014**, 4, 28728-28755.
57. Schenone, M.; Dancik, V.; Wagner, B.; Clemons, P., Target identification and mechanism of action in chemical biology and drug discovery. *Nat Chem Biol* **2013**, 9, 232 - 240.
58. Zheng, X. S.; Chan, T. F.; Zhou, H. H., Genetic and genomic approaches to identify and study the targets of bioactive small molecules. *Chem Biol* **2004**, 11, 609-18.
59. Boutros, M.; Ahringer, J., The art and design of genetic screens: RNA interference. *Nat Rev Genet* **2008**, 9, 554-66.
60. Klopmand, G., Concepts and applications of molecular similarity, by Mark A. Johnson and Gerald M. Maggiora, eds., John Wiley & Sons, New York, 1990, 393 pp. Price: $65.00. *Journal of Computational Chemistry* **1992**, 13, 539-540.
61. Martin, Y. C.; Kofron, J. L.; Traphagen, L. M., Do Structurally Similar Molecules Have Similar Biological Activity? *Journal of Medicinal Chemistry* **2002**, 45, 4350-4358.
62. Lin, S.-K., Pharmacophore Perception, Development and Use in Drug Design. Edited by Osman F. Güner. *Molecules* **2000**, 5, 987.
63. Laggner, C.; Schieferer, C.; Fiechtner, B.; Poles, G.; Hoffmann, R. D.; Glossmann, H.; Langer, T.; Moebius, F. F., Discovery of high-affinity ligands of sigma1 receptor, ERG2, and emopamil binding protein by pharmacophore modeling and virtual screening. *J Med Chem* **2005**, 48, 4754-64.
64. Rollinger, J. M.; Schuster, D.; Danzl, B.; Schwaiger, S.; Markt, P.; Schmidtke, M.; Gertsch, J.; Raduner, S.; Wolber, G.; Langer, T.; Stuppner, H., In silico target fishing for rationalized ligand discovery exemplified on constituents of Ruta graveolens. *Planta Med* **2009**, 75, 195-204.
65. Jitender, V.; Vijay, M. K.; Evans, C. C., 3D-QSAR in Drug Design - A Review. *Current topics in medicinal chemistry* **2010**, 10, 95-115.
66. Chen, G.; Zhou, D.; Li, X.-Z.; Jiang, Z.; Tan, C.; Wei, X.-Y.; Ling, J.; Jing, J.; Liu, F.; Li, N., A natural chalcone induces apoptosis in lung cancer cells: 3D-QSAR, docking and an in vivo/vitro assay. *Scientific Reports* **2017**, 7, 10729.
67. Menezes, I. R.; Lopes, J. C.; Montanari, C. A.; Oliva, G.; Pavao, F.; Castilho, M. S.; Vieira, P. C.; Pupo, M. T., 3D QSAR studies on binding affinities of coumarin

natural products for glycosomal GAPDH of Trypanosoma cruzi. *Journal of computer-aided molecular design* **2003**, 17, 277-90.
68. Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J. E., Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochimica et biophysica acta* **2001**, 1525, 180-90.
69. Morris, G. M.; Lim-Wilby, M., Molecular docking. *Methods in molecular biology (Clifton, N.J.)* **2008**, 443, 365-82.
70. Varnek, A.; Baskin, I., Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis? *Journal of Chemical Information and Modeling* **2012**, 52, 1413-1437.
71. Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B., Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today* **2018**.
72. Ali, S. M.; Hoemann, M. Z.; Aube, J.; Georg, G. I.; Mitscher, L. A.; Jayasinghe, L. R., Butitaxel analogues: synthesis and structure-activity relationships. *J Med Chem* **1997**, 40, 236-41.
73. Nasrabadi, N. M. Pattern Recognition and Machine Learning. 2007; SPIE: 2007; Vol. 16.
74. Cortes, C.; Vapnik, V., Support-vector networks. *Machine Learning* **1995**, 20, 273-297.
75. Nigsch, F.; Bender, A.; Jenkins, J.; Mitchell, J., Ligand-target prediction using Winnow and naive Bayesian algorithms and the implications of overall performance statistics. *J Chem Inf Model* **2008**, 48, 2313 - 2325.
76. Altman, N. S., An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* **1992**, 46, 175-185.
77. Breiman, L., Random Forests. *Mach. Learn.* **2001**, 45, 5-32.
78. Breiman, L., Random Forests. *Machine Learning* **2001**, 45, 5-32.
79. Schmidhuber, J., Deep learning in neural networks: An overview. *Neural Networks* **2015**, 61, 85-117.
80. Keiser, M.; Roth, B.; Armbruster, B.; Ernsberger, P.; Irwin, J.; Shoichet, B., Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* **2007**, 25, 197 - 206.
81. Lagunin, A.; Stepanchikova, A.; Filimonov, D.; Poroikov, V., PASS: prediction of activity spectra for biologically active substances. *Bioinformatics (Oxford, England)* **2000**, 16, 747 - 748.
82. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H., PubChem Substance and Compound databases. *Nucleic acids research* **2016**, 44, D1202-D1213.
83. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P., The ChEMBL bioactivity database: an update. *Nucleic acids research* **2014**, 42, D1083-90.
84. Marius Olah, M. M., Liliana Ostopovici, Ramona Rad, Alina Bora, Nicoleta Hadaruga, Ionela Olah, Magdalena Banda, Zeno Simon, Mirceae Mracec, Tudor I. Oprea, WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*.
85. Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K., Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* **2007**, 25, 197-206.

86. Sá, M. S.; de Menezes, M. N.; Krettli, A. U.; Ribeiro, I. M.; Tomassini, T. C. B.; Ribeiro dos Santos, R.; de Azevedo, W. F.; Soares, M. B. P., Antimalarial Activity of Physalins B, D, F, and G. *Journal of Natural Products* **2011**, 74, 2269-2272.
87. Wang, L.; Ma, C.; Wipf, P.; Liu, H.; Su, W.; Xie, X.-Q., TargetHunter: An In Silico Target Identification Tool for Predicting Therapeutic Potential of Small Organic Molecules Based on Chemogenomic Database. *AAPS J* **2013**, 15, 395-406.
88. Tan, K. P.; Yang, M.; Ito, S., Activation of nuclear factor (erythroid-2 like) factor 2 by toxic bile acids provokes adaptive defense responses to enhance cell survival at the emergence of oxidative stress. *Molecular pharmacology* **2007**, 72, 1380-90.
89. Jamkhande, P. G.; Wattamwar, A. S.; Pekamwar, S. S.; Chandak, P. G., Antioxidant, antimicrobial activity and in silico PASS prediction of Annona reticulata Linn. root extract. *Beni-Suef University Journal of Basic and Applied Sciences* **2014**, 3, 140-148.
90. Goel, R. K.; Singh, D.; Lagunin, A.; Poroikov, V., PASS-assisted exploration of new therapeutic potential of natural products. *Medicinal Chemistry Research* **2011**, 20, 1509-1514.
91. Nidhi; Glick, M.; Davies, J.; Jenkins, J., Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* **2006**, 46, 1124 - 1133.
92. MDDR licensed by Molecular Design, L., San Leandro, CA. www.mdli.com., In.
93. Kohonen, T., Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **1982**, 43, 59-69.
94. Schneider, G.; Reker, D.; Chen, T.; Hauenstein, K.; Schneider, P.; Altmann, K. H., Deorphaning the Macromolecular Targets of the Natural Anticancer Compound Doliculide. *Angewandte Chemie International Edition* **2016**, 55, 12408-12411.
95. Schneider, P.; Schneider, G., Collection of Bioactive Reference Compounds for Focused Library Design. *QSAR & Combinatorial Science* **2003**, 22, 713-718.
96. Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G., Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proceedings of the National Academy of Sciences* **2014**, 111, 4067.
97. Huang, T.; Mi, H.; Lin, C.-y.; Zhao, L.; Zhong, L. L. D.; Liu, F.-b.; Zhang, G.; Lu, A.-p.; Bian, Z.-x., MOST: most-similar ligand based approach to target prediction. *BMC Bioinformatics* **2017**, 18, 165.
98. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T., The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, 23, 1241-1250.
99. Gawehn, E.; Hiss, J. A.; Schneider, G., Deep Learning in Drug Discovery. *Molecular Informatics* **2015**, 35, 3-14.
100. Pascanu, R.; Mikolov, T.; Bengio, Y., In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*; JMLR.org: Atlanta, GA, USA, 2013, III-1310-III-1318.
101. Schmidhuber, J., Learning to Control Fast-Weight Memories: An Alternative to Dynamic Recurrent Networks. *Neural Computation* **1992**, 4, 131-139.
102. Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V., Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *Journal of Chemical Information and Modeling* **2015**, 55, 263-274.
103. Tropsha, A., Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, 29, 476-488.

104. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R., Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, 15, 1929-1958.
105. Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A., Computational Modeling of beta-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. *J Chem Inf Model* **2016**, 56, 1936-1949.
106. Aliper, A.; Plis, S.; Artemov, A.; Ulloa, A.; Mamoshina, P.; Zhavoronkov, A., Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Molecular pharmaceutics* **2016**, 13, 2524-30.
107. Koutsoukas, A.; Monaghan, K. J.; Li, X.; Huan, J., Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J Cheminform* **2017**, 9, 42.
108. Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; Ijzerman, A. P.; van Westen, G. J. P., Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of Cheminformatics* **2017**, 9, 45.
109. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S., DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science* **2016**, 3, 80.
110. Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V., Is Multitask Deep Learning Practical for Pharma? *Journal of Chemical Information and Modeling* **2017**, 57, 2068-2076.
111. LeCun, Y.; Bengio, Y.; Hinton, G., Deep learning. *Nature* **2015**, 521, 436.
112. Leffingwell, J., *Chirality & Bioactivity I.: Pharmacology*. 2003; Vol. 3, p 1-27.
113. Dei, S.; Bartolini, A.; Bellucci, C.; Ghelardini, C.; Gualtieri, F.; Manetti, D.; Romanelli, M. N.; Scapecchi, S.; Teodori, E., Differential analgesic activity of the enantiomers of atropine derivatives does not correlate with their muscarinic subtype selectivity. *European Journal of Medicinal Chemistry* **1997**, 32, 595-605.
114. Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, 50, 742-754.
115. Rogers, D.; Brown, R. D.; Hahn, M., Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *Journal of biomolecular screening* **2005**, 10, 682-6.
116. Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A., New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *Journal of Chemical Information and Modeling* **2006**, 46, 462-470.
117. Yoon, S.; Smellie, A.; Hartsough, D.; Filikov, A., Surrogate docking: structure-based virtual screening at high throughput speed. *Journal of computer-aided molecular design* **2005**, 19, 483-97.
118. Steindl, T. M.; Schuster, D.; Laggner, C.; Langer, T., Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *J Chem Inf Model* **2006**, 46, 2146-57.
119. Bolton, E.; Wang, Y.; Thiessen, P.; Bryant, S., PubChem: integrated platform of small molecules and biological activities. *Annu Rep Comput Chem* **2008**, 4, 217 - 241.
120. Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.;

Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S., DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research* **2014**, 42, D1091-7.
121. Kalliokoski, T.; Kramer, C.; Vulpetti, A., Quality Issues with Public Domain Chemogenomics Data. *Molecular Informatics* **2013**, 32, 898-905.
122. Gedeck, P.; Kramer, C.; Ertl, P. 4 - Computational Analysis of Structure–Activity Relationships. In *Progress in Medicinal Chemistry*, Lawton, G.; Witty, D. R., Eds.; Elsevier: 2010; Vol. 49, 113-160.
123. Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A., Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model* **2008**, 48, 1733-46.
124. Kristina, G.; Gisbert, S., Properties and Architecture of Drugs and Natural Products Revisited. *Current Chemical Biology* **2007**, 1, 115-127.
125. Henkel, Statistical investigation into the structural complementarity of natural products and synthetic compounds. *ANGEWANDTE CHEMIE-INTERNATIONAL EDITION* **1999**, 38, 643-647.
126. Feher, M.; Schmidt, J. M., Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *Journal of Chemical Information and Computer Sciences* **2003**, 43, 218-227.
127. Mervin, L.; Afzal, A.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A., Target prediction utilising negative bioactivity data covering large chemical space. *Journal of Cheminformatics* **2015**, 7, 51.
128. García-Campos, M. A.; Espinal-Enríquez, J.; Hernández-Lemus, E., Pathway Analysis: State of the Art. *Frontiers in Physiology* **2015**, 6, 383.
129. Folger, O.; Jerby, L.; Frezza, C.; Gottlieb, E.; Ruppin, E.; Shlomi, T., Predicting selective drug targets in cancer through metabolic networks. *Mol Syst Biol* **2011**, 7, 501.
130. Mohamad Zobir, S. Z.; Mohd Fauzi, F.; Liggi, S.; Drakakis, G.; Fu, X.; Fan, T.-P.; Bender, A., Global Mapping of Traditional Chinese Medicine into Bioactivity Space and Pathways Annotation Improves Mechanistic Understanding and Discovers Relationships between Therapeutic Action (Sub)classes. *Evidence-Based Complementary and Alternative Medicine* **2016**, 2016, 25.
131. Liggi, S.; Drakakis, G.; Koutsoukas, A.; Cortes-Ciriano, I.; Martinez-Alonso, P.; Malliavin, T.; Velazquez-Campoy, A.; Brewerton, S.; Bodkin, M.; Evans, D.; Glen, R.; Carrodeguas, J.; Bender, A., Extending in silico mechanism-of-action analysis by annotating targets with pathways: application to cellular cytotoxicity readouts. *Future Med Chem* **2014**, 6, 2029 - 2056.
132. Croft, D.; Mundo, A. F.; Haw, R.; Milacic, M.; Weiser, J.; Wu, G.; Caudy, M.; Garapati, P.; Gillespie, M.; Kamdar, M. R.; Jassal, B.; Jupe, S.; Matthews, L.; May, B.; Palatnik, S.; Rothfels, K.; Shamovsky, V.; Song, H.; Williams, M.; Birney, E.; Hermjakob, H.; Stein, L.; D'Eustachio, P., The Reactome pathway knowledgebase. *Nucleic acids research* **2014**, 42, D472-7.
133. Kanehisa, M.; Goto, S., KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **2000**, 28, 27-30.
134. Kanehisa, M., The KEGG database. *Novartis Found Symp* **2002**, 247, 91 - 101.
135. Ashburner, M.; Ball, C.; Blake, J.; Botstein, D.; Butler, H.; Cherry, J.; Davis, A.; Dolinski, K.; Dwight, S.; Eppig, J.; Harris, M.; Hill, D.; Issel-Tarver, L.; Kasarskis, A.; Lewis, S.; Matese, J.; Richardson, J.; Ringwald, M.; Rubin, G.; Sherlock, G.,

Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **2000**, 25, 25 - 29.
136. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic acids research* **2017**, 45, D331-d338.
137. Mi, H.; Huang, X.; Muruganujan, A.; Tang, H.; Mills, C.; Kang, D.; Thomas, P. D., PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic acids research* **2017**, 45, D183-D189.
138. Davis, A. P.; Grondin, C. J.; Johnson, R. J.; Sciaky, D.; King, B. L.; McMorran, R.; Wiegers, J.; Wiegers, T. C.; Mattingly, C. J., The Comparative Toxicogenomics Database: update 2017. *Nucleic acids research* **2017**, 45, D972-D978.
139. Slenter, D. N.; Kutmon, M.; Hanspers, K.; Riutta, A.; Windsor, J.; Nunes, N.; Mélius, J.; Cirillo, E.; Coort, S. L.; Digles, D.; Ehrhart, F.; Giesbertz, P.; Kalafati, M.; Martens, M.; Miller, R.; Nishida, K.; Rieswijk, L.; Waagmeester, A.; Eijssen, L. M. T.; Evelo, C. T.; Pico, A. R.; Willighagen, E. L., WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research* **2018**, 46, D661-D667.
140. du Plessis, L.; Škunca, N.; Dessimoz, C., The what, where, how and why of gene ontology—a primer for bioinformaticians. *Briefings in Bioinformatics* **2011**, 12, 723-735.
141. Kelder, T.; van Iersel, M. P.; Hanspers, K.; Kutmon, M.; Conklin, B. R.; Evelo, C. T.; Pico, A. R., WikiPathways: building research communities on biological pathways. *Nucleic acids research* **2012**, 40, D1301-D1307.
142. Kutmon, M.; Riutta, A.; Nunes, N.; Hanspers, K.; Willighagen, Egon L.; Bohler, A.; Mélius, J.; Waagmeester, A.; Sinha, Sravanthi R.; Miller, R.; Coort, S. L.; Cirillo, E.; Smeets, B.; Evelo, Chris T.; Pico, A. R., WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic acids research* **2016**, 44, D488-D494.
143. Mervin, L. H.; Bulusu, K. C.; Kalash, L.; Afzal, A. M.; Svensson, F.; Firth, M. A.; Barrett, I.; Engkvist, O.; Bender, A., Orthologue chemical space and its influence on target prediction. *Bioinformatics (Oxford, England)* **2017**, btx525-btx525.
144. Geer, L. Y.; Marchler-Bauer, A.; Geer, R. C.; Han, L.; He, J.; He, S.; Liu, C.; Shi, W.; Bryant, S. H., The NCBI BioSystems database. *Nucleic acids research* **2010**, 38, D492-D496.
145. Saslis-Lagoudakis, C. H.; Klitgaard, B. B.; Forest, F.; Francis, L.; Savolainen, V.; Williamson, E. M.; Hawkins, J. A., The Use of Phylogeny to Interpret Cross-Cultural Patterns in Plant Use and Guide Medicinal Plant Discovery: An Example from Pterocarpus (Leguminosae). *PLoS ONE* **2011**, 6, e22275.
146. Ronsted, N.; Symonds, M. R.; Birkholm, T.; Christensen, S. B.; Meerow, A. W.; Molander, M.; Molgaard, P.; Petersen, G.; Rasmussen, N.; van Staden, J.; Stafford, G. I.; Jager, A. K., Can phylogeny predict chemical diversity and potential medicinal activity of plants? A case study of Amaryllidaceae. *BMC evolutionary biology* **2012**, 12, 182.
147. Saslis-Lagoudakis, C. H.; Savolainen, V.; Williamson, E. M.; Forest, F.; Wagstaff, S. J.; Baral, S. R.; Watson, M. F.; Pendry, C. A.; Hawkins, J. A., Phylogenies reveal predictive power of traditional medicine in bioprospecting. *Proceedings of the National Academy of Sciences* **2012**, 109, 15835-15840.

148. Ngezahayo, J.; Havyarimana, F.; Hari, L.; Stevigny, C.; Duez, P., Medicinal plants used by Burundian traditional healers for the treatment of microbial diseases. *J Ethnopharmacol* **2015**, 173, 338-51.
149. Colless, D., Phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology* **1982**, 31.
150. Saitou, N.; Nei, M., The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **1987**, 4, 406-425.
151. Rzhetsky, A.; Nei, M., A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution* **1992**, 9, 945-967.
152. Desper, R.; Gascuel, O., Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of computational biology : a journal of computational molecular cell biology* **2002**, 9, 687-705.
153. Cavalli-Sforza, L. L.; Edwards, A. W. F., Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics* **1967**, 19, 233-257.
154. Farris, J. S., Methods for Computing Wagner Trees. *Systematic Zoology* **1970**, 19, 83-92.
155. Fitch, W. M., Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology* **1971**, 20, 406-416.
156. Felsenstein, J., Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **1981**, 17, 368-76.
157. Rannala, B.; Yang, Z., Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol* **1996**, 43, 304-11.
158. Yang, Z.; Rannala, B., Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol* **1997**, 14, 717-24.
159. Mau, B.; Newton, M. A., Phylogenetic Inference for Binary Data on Dendograms Using Markov Chain Monte Carlo. *Journal of Computational and Graphical Statistics* **1997**, 6, 122-131.
160. Li, S.; Pearl, D. K.; Doss, H., Phylogenetic Tree Construction Using Markov Chain Monte Carlo. *Journal of the American Statistical Association* **2000**, 95, 493-508.
161. Yang, Z.; Rannala, B., Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* **2012**, 13, 303.
162. Zanne, A. E.; Tank, D. C.; Cornwell, W. K.; Eastman, J. M.; Smith, S. A.; FitzJohn, R. G.; McGlinn, D. J.; O'Meara, B. C.; Moles, A. T.; Reich, P. B.; Royer, D. L.; Soltis, D. E.; Stevens, P. F.; Westoby, M.; Wright, I. J.; Aarssen, L.; Bertin, R. I.; Calaminus, A.; Govaerts, R.; Hemmings, F.; Leishman, M. R.; Oleksyn, J.; Soltis, P. S.; Swenson, N. G.; Warman, L.; Beaulieu, J. M., Three keys to the radiation of angiosperms into freezing environments. *Nature* **2014**, 506, 89-92.
163. Regional Committee for Africa *Promoting the role of traditional medicine in health systems: a strategy for the African Region*; Brazzaville, 2000.
164. M. A. Johnson and G. M. Maggiora, *Concepts and Applications of Molecular Similarity*. 1st edn ed.; Wiley-Interscience: New York, 1990.
165. Walters, W. P.; Stahl, M. T.; Murcko, M. A., Virtual screening—an overview. *Drug Discovery Today* **1998**, 3, 160-178.
166. Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E., Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *Journal of Medicinal Chemistry* **1996**, 39, 3049-3059.
167. Newman, D. J.; Cragg, G. M., Natural Products as Sources of New Drugs over the Last 25 Years⊥. *Journal of Natural Products* **2007**, 70, 461-477.

168. Yongye, A. B.; Waddell, J.; Medina-Franco, J. L., Molecular scaffold analysis of natural products databases in the public domain. *Chem Biol Drug Des* **2012**, 80, 717-24.
169. Mohd Fauzi, F.; Koutsoukas, A.; Lowe, R.; Joshi, K.; Fan, T.-P.; Glen, R. C.; Bender, A., Chemogenomics Approaches to Rationalizing the Mode-of-Action of Traditional Chinese and Ayurvedic Medicines. *Journal of Chemical Information and Modeling* **2013**, 53, 661-673.
170. Lagunin, A. A.; Goel, R. K.; Gawande, D. Y.; Pahwa, P.; Gloriozova, T. A.; Dmitriev, A. V.; Ivanov, S. M.; Rudik, A. V.; Konova, V. I.; Pogodin, P. V.; Druzhilovsky, D. S.; Poroikov, V. V., Chemo- and bioinformatics resources for in silico drug discovery from medicinal plants beyond their traditional use: a critical review. *Natural Product Reports* **2014**, 31, 1585-1611.
171. Barlow, D. J.; Buriani, A.; Ehrman, T.; Bosisio, E.; Eberini, I.; Hylands, P. J., In-silico studies in Chinese herbal medicines' research: Evaluation of in-silico methodologies and phytochemical data sources, and a review of research to date. *Journal of Ethnopharmacology* **2012**, 140, 526-534.
172. Liggi, S.; Drakakis, G.; Hendry, A. E.; Hanson, K. M.; Brewerton, S. C.; Wheeler, G. N.; Bodkin, M. J.; Evans, D. A.; Bender, A., Extensions to In Silico Bioactivity Predictions Using Pathway Annotations and Differential Pharmacology Analysis: Application to Xenopus laevis Phenotypic Readouts. *Molecular Informatics* **2013**, 32, 1009-1024.
173. Bajorath, J., Chemoinformatics methods for systematic comparison of molecules from natural and synthetic sources and design of hybrid libraries. *Molecular diversity* **2002**, 5, 305-13.
174. Reker, D.; Perna, A. M.; Rodrigues, T.; Schneider, P.; Reutlinger, M.; Mönch, B.; Koeberle, A.; Lamers, C.; Gabler, M.; Steinmetz, H.; Müller, R.; Schubert-Zsilavecz, M.; Werz, O.; Schneider, G., Revealing the macromolecular targets of complex natural products. *Nat Chem* **2014**, 6, 1072-1078.
175. Ehrman, T. M.; Barlow, D. J.; Hylands, P. J., Virtual screening of Chinese herbs with Random Forest. *J Chem Inf Model* **2007**, 47, 264-78.
176. Jasial, S.; Hu, Y.; Vogt, M.; Bajorath, J., Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Research* **2016**, 5, Chem Inf Sci-591.
177. http://cactus.nci.nih.gov/download/nci/ (8/9/2014),
178. In *Nucleic acids research*; 2011; Vol. 40, D1100-D1107.
179. Wahab, H. A., In; Pulau Pinang, Malaysia: Pharmaceutical Design and Simulation Laboratory, Universiti Sains Malaysia, 2007.
180. ChemSpider Synthetic Pages, A. C., In; 2001.
181. ChemAxon http://www.chemaxon.com
182. R Development Core Team *R: A language and environment for statistical computing*, R Foundation for statistical computing: Vienna, Austria, 2013.
183. Xing, L.; Glen, R. C., Novel Methods for the Prediction of logP, pKa, and logD. *Journal of Chemical Information and Computer Sciences* **2002**, 42, 796-805.
184. Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W., Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *Journal of Molecular Graphics and Modelling* **2010**, 29, 157-170.
185. Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W., Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *Journal of Chemical Information and Modeling* **2010**, 50, 771-784.

186. Shrodinger LLC *OikProp 3.4 User Manual*, 3.4; Schrodinger Press LLC: New York, NY, 2011.
187. Kelly, M. D.; Mancera, R. L., Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *J Chem Inf Comput Sci* **2004**, 44, 1942-51.
188. Bemis, G. W.; Murcko, M. A., The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, 39, 2887-2893.
189. Shannon, C. E., A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* **2001**, 5, 3-55.
190. Godden, J. W.; Bajorath, J., Analysis of chemical information content using shannon entropy. *Reviews in Computational Chemistry* **2007**, 23, 263.
191. Sander, T.; Freyss, J.; von Korff, M.; Rufener, C., DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *Journal of Chemical Information and Modeling* **2015**, 55, 460-473.
192. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P., Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* **2003**, 43, 1947-58.
193. Platt, J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In Advances In Large Margin Classifiers, 1999; 1999; 61-74.
194. Edwards, A. W. F., The Measure of Association in a 2 × 2 Table. *Journal of the Royal Statistical Society. Series A (General)* **1963**, 126, 109-114.
195. Mosteller, F., Association and Estimation in Contingency Tables. *Journal of the American Statistical Association* **1968**, 63, 1-28.
196. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* **2001**, 46, 3-26.
197. Lee, M.-L.; Schneider, G., Scaffold Architecture and Pharmacophoric Properties of Natural Products and Trade Drugs: Application in the Design of Natural Product-Based Combinatorial Libraries. *Journal of Combinatorial Chemistry* **2001**, 3, 284-289.
198. Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L., Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J Chem Inf Model* **2009**, 49, 1010-24.
199. Grabowski, K.; Baringhaus, K.-H.; Schneider, G., Scaffold diversity of natural products: inspiration for combinatorial library design. *Natural Product Reports* **2008**, 25, 892-904.
200. Mansuri, M. L.; Parihar, P.; Solanki, I.; Parihar, M. S., Flavonoids in modulation of cell survival signalling pathways. *Genes & Nutrition* **2014**, 9, 400.
201. Aran, V. J.; Kaiser, M.; Dardonville, C., Discovery of nitroheterocycles active against African trypanosomes. In vitro screening and preliminary SAR studies. *Bioorganic & medicinal chemistry letters* **2012**, 22, 4506-16.
202. He, X.; Chen, X.; Lin, S.; Mo, X.; Zhou, P.; Zhang, Z.; Lu, Y.; Yang, Y.; Gu, H.; Shang, Z.; Lou, Y.; Wu, J., Diversity-Oriented Synthesis of Natural-Product-like Libraries Containing a 3-Methylbenzofuran Moiety for the Discovery of New Chemical Elicitors. *ChemistryOpen* **2017**, 6, 102-111.
203. Chong, J.; Poutaraud, A.; Hugueney, P., Metabolism and roles of stilbenes in plants. *Plant Science* **2009**, 177, 143-155.

204. Magloire Ketcha Wanda, G. J.; Njamen, D.; Tagatsing, F. M.; Yankep, E.; Vollmer, G., Regulation of CD1, Ki-67, PCNA mRNA expression, and Akt activation in estrogen-responsive human breast adenocarcinoma cell line, MCF-7 cells, by griffonianone C, an isoflavone derived from Millettia griffoniana. *Pharmaceutical biology* **2011**, 49, 341-7.
205. Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R., Heteroaromatic Rings of the Future. *Journal of Medicinal Chemistry* **2009**, 52, 2952-2963.
206. Wu, P.; Nielsen, T. E.; Clausen, M. H., Small-molecule kinase inhibitors: an analysis of FDA-approved drugs. *Drug Discov Today* **2016**, 21, 5-10.
207. Hauser, A. S.; Attwood, M. M.; Rask-Andersen, M.; Schiöth, H. B.; Gloriam, D. E., Trends in GPCR drug discovery: new agents, targets and indications. *Nature Reviews Drug Discovery* **2017**, 16, 829.
208. Yenerall, P.; Kittler, R., Harnessing the nuclear receptor PPARγ to inhibit the growth of lung adenocarcinoma by rewiring metabolic circuitries. *Molecular & Cellular Oncology* **2015**, 2, e980660.
209. Ciocca, D. R.; Calderwood, S. K., Heat shock proteins in cancer: diagnostic, prognostic, predictive, and treatment implications. *Cell Stress & Chaperones* **2005**, 10, 86-103.
210. Roy, K.; Wu, Y.; Meitzler, J. L.; Juhasz, A.; Liu, H.; Jiang, G.; Lu, J.; Antony, S.; Doroshow, J. H., NADPH oxidases and cancer. *Clinical science (London, England : 1979)* **2015**, 128, 863-75.
211. Rashid, M. A.; Lee, S.; Tak, E.; Lee, J.; Choi, T. G.; Lee, J. W.; Kim, J. B.; Youn, J. H.; Kang, I.; Ha, J.; Kim, S. S., Carbonyl reductase 1 protects pancreatic beta-cells against oxidative stress-induced apoptosis in glucotoxicity and glucolipotoxicity. *Free radical biology & medicine* **2010**, 49, 1522-33.
212. Liu, J.; Sridhar, J.; Foroozesh, M., Cytochrome P450 Family 1 Inhibitors and Structure-Activity Relationships. *Molecules (Basel, Switzerland)* **2013**, 18, 14470-14495.
213. Ma, I.; Allan, A. L., The role of human aldehyde dehydrogenase in normal and cancer stem cells. *Stem cell reviews* **2011**, 7, 292-306.
214. Antony, G. K.; Dudek, A. Z., Interleukin 2 in cancer therapy. *Current medicinal chemistry* **2010**, 17, 3297-302.
215. Karlenius, T. C.; Tonissen, K. F., Thioredoxin and Cancer: A Role for Thioredoxin in all States of Tumor Oxygenation. *Cancers* **2010**, 2, 209-232.
216. Endicott, J. A.; Ling, V., The biochemistry of P-glycoprotein-mediated multidrug resistance. *Annual review of biochemistry* **1989**, 58, 137-171.
217. Osborne, C. K., Steroid hormone receptors in breast cancer management. *Breast cancer research and treatment* **1998**, 51, 227-38.
218. Comes, N.; Bielanska, J.; Vallejo-Gracia, A.; Serrano-Albarrás, A.; Marruecos, L.; Gómez, D.; Soler, C.; Condom, E.; Ramón y Cajal, S.; Hernández-Losa, J.; Ferreres, J. C.; Felipe, A., The voltage-dependent K+ channels Kv1.3 and Kv1.5 in human cancer. *Frontiers in Physiology* **2013**, 4.
219. Kaigorodova, E. V.; Zavyalova, M. V.; Bogatyuk, M. V.; Tarabanovskaya, N. A.; Slonimskaya, E. M.; Perelmuter, V. M., Relationship between the expression of phosphorylated heat shock protein beta-1 with lymph node metastases of breast cancer. *Cancer biomarkers : section A of Disease markers* **2015**, 15, 143-50.
220. Zanini, C.; Giribaldi, G.; Mandili, G.; Carta, F.; Crescenzio, N.; Bisaro, B.; Doria, A.; Foglia, L.; Di Montezemolo, L. C.; Timeus, F.; Turrini, F., Inhibition of heat shock proteins (HSP) expression by quercetin and differential doxorubicin

sensitization in neuroblastoma and Ewing's sarcoma cell lines. *Journal of Neurochemistry* **2007**, 103, 1344-1354.
221. Sang, D.-p.; Li, R.-j.; Lan, Q., Quercetin sensitizes human glioblastoma cells to temozolomide in vitro via inhibition of Hsp27. *Acta Pharmacologica Sinica* **2014**, 35, 832-838.
222. Peter Guengerich, F.; Chun, Y. J.; Kim, D.; Gillam, E. M.; Shimada, T., Cytochrome P450 1B1: a target for inhibition in anticarcinogenesis strategies. *Mutation research* **2003**, 523-524, 173-82.
223. Cui, J.; Meng, Q.; Zhang, X.; Cui, Q.; Zhou, W.; Li, S., Design and Synthesis of New α-Naphthoflavones as Cytochrome P450 (CYP) 1B1 Inhibitors To Overcome Docetaxel-Resistance Associated with CYP1B1 Overexpression. *Journal of Medicinal Chemistry* **2015**, 58, 3534-3547.
224. McFadyen, M. C.; McLeod, H. L.; Jackson, F. C.; Melvin, W. T.; Doehmer, J.; Murray, G. I., Cytochrome P450 CYP1B1 protein expression: a novel mechanism of anticancer drug resistance. *Biochem Pharmacol* **2001**, 62, 207-12.
225. Buters, J. T. M.; Sakai, S.; Richter, T.; Pineau, T.; Alexander, D. L.; Savas, U.; Doehmer, J.; Ward, J. M.; Jefcoate, C. R.; Gonzalez, F. J., Cytochrome P450 CYP1B1 determines susceptibility to 7,12-dimethylbenz[a]anthracene-induced lymphomas. *Proceedings of the National Academy of Sciences of the United States of America* **1999**, 96, 1977-1982.
226. Gajjar, K.; Martin-Hirsch, P. L.; Martin, F. L., CYP1B1 and hormone-induced cancer. *Cancer letters* **2012**, 324, 13-30.
227. Androutsopoulos, V. P.; Tsatsakis, A. M.; Spandidos, D. A., Cytochrome P450 CYP1A1: wider roles in cancer progression and prevention. *BMC cancer* **2009**, 9, 187.
228. Anzenbacher, P.; Anzenbacherova, E., Cytochromes P450 and metabolism of xenobiotics. *Cellular and molecular life sciences : CMLS* **2001**, 58, 737-47.
229. Rao, V. R.; Perez-Neut, M.; Kaja, S.; Gentile, S., Voltage-Gated Ion Channels in Cancer Cell Proliferation. *Cancers* **2015**, 7, 849-875.
230. Gasiorowska, J.; Teisseyre, A.; Uryga, A.; Michalak, K., The influence of 8-prenylnaringenin on the activity of voltage-gated Kv1.3 potassium channels in human Jurkat T cells. *Cellular & molecular biology letters* **2012**, 17, 559-70.
231. Binaschi, M.; Zunino, F.; Capranico, G., Mechanism of action of DNA topoisomerase inhibitors. *Stem cells (Dayton, Ohio)* **1995**, 13, 369-79.
232. Schmidt, F.; Knobbe, C. B.; Frank, B.; Wolburg, H.; Weller, M., The topoisomerase II inhibitor, genistein, induces G2/M arrest and apoptosis in human malignant glioma cell lines. *Oncology reports* **2008**, 19, 1061-6.
233. Cantero, G.; Campanella, C.; Mateos, S.; Cortes, F., Topoisomerase II inhibition and high yield of endoreduplication induced by the flavonoids luteolin and quercetin. *Mutagenesis* **2006**, 21, 321-5.
234. Leone, S.; Basso, E.; Polticelli, F.; Cozzi, R., Resveratrol acts as a topoisomerase II poison in human glioma cells. *International journal of cancer* **2012**, 131, E173-8.
235. Goodsell, D. S., The Molecular Perspective: Tamoxifen and the Estrogen Receptor. *The Oncologist* **2002**, 7, 163-164.
236. Osborne, C. K.; Wakeling, A.; Nicholson, R. I., Fulvestrant: an oestrogen receptor antagonist with a novel mechanism of action. *British journal of cancer* **2004**, 90, S2-S6.

237. Hajirahimkhan, A.; Dietz, B. M.; Bolton, J. L., Botanical modulation of menopausal symptoms: Mechanisms of action? *Planta medica* **2013**, 79, 538-553.
238. Brumatti, G.; Ekert, P. G., Seeking a MCL-1 inhibitor. *Cell Death And Differentiation* **2013**, 20, 1440.
239. Singh, P.; Yang, M.; Dai, H.; Yu, D.; Huang, Q.; Tan, W.; Kernstine, K.; Lin, D.; Shen, B., Over-expression and hypomethylation of flap endonuclease 1 gene in breast and other cancers. *Molecular cancer research : MCR* **2008**, 6, 1710-1717.
240. Lam, J. S.; Seligson, D. B.; Yu, H.; Li, A.; Eeva, M.; Pantuck, A. J.; Zeng, G.; Horvath, S.; Belldegrun, A. S., Flap endonuclease 1 is overexpressed in prostate cancer and is associated with a high Gleason score. *BJU international* **2006**, 98, 445-51.
241. Wang, K.; Xie, C.; Chen, D., Flap endonuclease 1 is a promising candidate biomarker in gastric cancer and is involved in cell proliferation and apoptosis. *International journal of molecular medicine* **2014**, 33, 1268-74.
242. Krause, A.; Combaret, V.; Iacono, I.; Lacroix, B.; Compagnon, C.; Bergeron, C.; Valsesia-Wittmann, S.; Leissner, P.; Mougin, B.; Puisieux, A., Genome-wide analysis of gene expression in neuroblastomas detected by mass screening. *Cancer letters* **2005**, 225, 111-20.
243. Iacobuzio-Donahue, C. A.; Maitra, A.; Olsen, M.; Lowe, A. W.; van Heek, N. T.; Rosty, C.; Walter, K.; Sato, N.; Parker, A.; Ashfaq, R.; Jaffee, E.; Ryu, B.; Jones, J.; Eshleman, J. R.; Yeo, C. J.; Cameron, J. L.; Kern, S. E.; Hruban, R. H.; Brown, P. O.; Goggins, M., Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. *The American journal of pathology* **2003**, 162, 1151-62.
244. Nikolova, T.; Christmann, M.; Kaina, B., FEN1 is overexpressed in testis, lung and brain tumors. *Anticancer research* **2009**, 29, 2453-9.
245. Sharma, S.; Javadekar, S. M.; Pandey, M.; Srivastava, M.; Kumari, R.; Raghavan, S. C., Homology and enzymatic requirements of microhomology-dependent alternative end joining. *Cell Death & Disease* **2015**, 6, e1697.
246. Evans, C. G.; Chang, L.; Gestwicki, J. E., Heat Shock Protein 70 (Hsp70) as an Emerging Drug Target. *Journal of medicinal chemistry* **2010**, 53, 4585-4602.
247. Sherman, M. Y.; Gabai, V. L., Hsp70 in cancer: back to the future. *Oncogene* **2015**, 34, 4153-4161.
248. Smith, S.; Giriat, I.; Schmitt, A.; de Lange, T., Tankyrase, a poly(ADP-ribose) polymerase at human telomeres. *Science* **1998**, 282, 1484-7.
249. Seimiya, H.; Muramatsu, Y.; Smith, S.; Tsuruo, T., Functional subdomain in the ankyrin domain of tankyrase 1 required for poly(ADP-ribosyl)ation of TRF1 and telomere elongation. *Mol Cell Biol* **2004**, 24, 1944-55.
250. Chang, W.; Dynek, J. N.; Smith, S., TRF1 is degraded by ubiquitin-mediated proteolysis after release from telomeres. *Genes & development* **2003**, 17, 1328-33.
251. Donate, L. E.; Blasco, M. A., Telomeres in cancer and ageing. *Philosophical Transactions of the Royal Society B: Biological Sciences* **2011**, 366, 76-84.
252. McCabe, N.; Cerone, M. A.; Ohishi, T.; Seimiya, H.; Lord, C. J.; Ashworth, A., Targeting Tankyrase 1 as a therapeutic strategy for BRCA-associated cancer. *Oncogene* **2009**, 28, 1465-1470.
253. Seimiya, H., The telomeric PARP, tankyrases, as targets for cancer therapy. *British journal of cancer* **2006**, 94, 341-5.
254. Seimiya, H.; Muramatsu, Y.; Ohishi, T.; Tsuruo, T., Tankyrase 1 as a target for telomere-directed molecular cancer therapeutics. *Cancer Cell* **2005**, 7, 25-37.

255. Dynek, J. N.; Smith, S., Resolution of sister telomere association is required for progression through mitosis. *Science* **2004**, 304, 97-100.
256. Chang, W.; Dynek, J. N.; Smith, S., NuMA is a major acceptor of poly(ADP-ribosyl)ation by tankyrase 1 in mitosis. *The Biochemical journal* **2005**, 391, 177-84.
257. Re, F.; Braaten, D.; Franke, E. K.; Luban, J., Human immunodeficiency virus type 1 Vpr arrests the cell cycle in G2 by inhibiting the activation of p34cdc2-cyclin B. *Journal of Virology* **1995**, 69, 6859-6864.
258. Takashima, S.; Saito, H.; Takahashi, N.; Imai, K.; Kudo, S.; Atari, M.; Saito, Y.; Motoyama, S.; Minamiya, Y., Strong expression of cyclin B2 mRNA correlates with a poor prognosis in patients with non-small cell lung cancer. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* **2014**, 35, 4257-65.
259. Lei, C. Y.; Wang, W.; Zhu, Y. T.; Fang, W. Y.; Tan, W. L., The decrease of cyclin B2 expression inhibits invasion and metastasis of bladder cancer. *Urologic oncology* **2016**, 34, 237.e1-10.
260. Wang, R. E.; Kao, J. L. F.; Hilliard, C. A.; Pandita, R. K.; Roti Roti, J. L.; Hunt, C. R.; Taylor, J.-S., Inhibition of Heat Shock Induction of Heat Shock Protein 70 and Enhancement of Heat Shock Protein 27 Phosphorylation by Quercetin Derivatives. *Journal of Medicinal Chemistry* **2009**, 52, 1912-1921.
261. Ingólfsson, H. I.; Thakur, P.; Herold, K. F.; Hobart, E. A.; Ramsey, N. B.; Periole, X.; de Jong, D. H.; Zwama, M.; Yilmaz, D.; Hall, K.; Maretzky, T.; Hemmings, H. C.; Blobel, C.; Marrink, S. J.; Koçer, A.; Sack, J. T.; Andersen, O. S., Phytochemicals Perturb Membranes and Promiscuously Alter Protein Function. *ACS Chemical Biology* **2014**, 9, 1788-1798.
262. Kansanen, E.; Kuosmanen, S. M.; Leinonen, H.; Levonen, A.-L., The Keap1-Nrf2 pathway: Mechanisms of activation and dysregulation in cancer. *Redox Biology* **2013**, 1, 45-49.
263. Ren, D.; Villeneuve, N. F.; Jiang, T.; Wu, T.; Lau, A.; Toppin, H. A.; Zhang, D. D., Brusatol enhances the efficacy of chemotherapy by inhibiting the Nrf2-mediated defense mechanism. *Proc Natl Acad Sci U S A* **2011**, 108, 1433-8.
264. Sporn, M. B.; Liby, K. T., NRF2 and cancer: the good, the bad and the importance of context. *Nat Rev Cancer* **2012**, 12, 564-571.
265. Kouam, S. F.; Njonkou, Y. L. N.; Kuigoua, G. M.; Ngadjui, B. T.; Hussain, H.; Green, I. R.; Schulz, B.; Krohn, K., Psorantin, a unique methylene linked dimer of vismin and kenganthranol E, two anthranoid derivatives from the fruits of Psorospermum aurantiacum (Hypericaceae). *Phytochemistry Letters* **2010**, 3, 185-189.
266. Challacombe, J. M.; Suhrbier, A.; Parsons, P. G.; Jones, B.; Hampson, P.; Kavanagh, D.; Rainger, G. E.; Morris, M.; Lord, J. M.; Le, T. T. T.; Hoang-Le, D.; Ogbourne, S. M., Neutrophils Are a Key Component of the Antitumor Efficacy of Topical Chemotherapy with Ingenol-3-Angelate. *The Journal of Immunology* **2006**, 177, 8123-8132.
267. Jones, J. R.; Barrick, C.; Kim, K.-A.; Lindner, J.; Blondeau, B.; Fujimoto, Y.; Shiota, M.; Kesterson, R. A.; Kahn, B. B.; Magnuson, M. A., Deletion of PPARγ in adipose tissues of mice protects against high fat diet-induced obesity and insulin resistance. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, 102, 6207-6212.

268. Bingle, C. D.; Craig, R. W.; Swales, B. M.; Singleton, V.; Zhou, P.; Whyte, M. K. B., Exon Skipping in Mcl-1 Results in a Bcl-2 Homology Domain 3 Only Gene Product That Promotes Cell Death. *Journal of Biological Chemistry* **2000**, 275, 22136-22146.
269. Maurer, U.; Charvet, C.; Wagman, A. S.; Dejardin, E.; Green, D. R., Glycogen Synthase Kinase-3 Regulates Mitochondrial Outer Membrane Permeabilization and Apoptosis by Destabilization of MCL-1. *Molecular Cell* **2006**, 21, 749-760.
270. Tong, W.-G.; Ding, X.-Z.; Hennig, R.; Witt, R. C.; Standop, J.; Pour, P. M.; Adrian, T. E., Leukotriene B4 Receptor Antagonist LY293111 Inhibits Proliferation and Induces Apoptosis in Human Pancreatic Cancer Cells. *Clinical Cancer Research* **2002**, 8, 3232-3242.
271. Crooks, S. W.; Stockley, R. A., Leukotriene B4. *The International Journal of Biochemistry & Cell Biology* **1998**, 30, 173-178.
272. Sokolovsky, M., Endothelin receptor subtypes and their role in transmembrane signaling mechanisms. *Pharmacology & therapeutics* **1995**, 68, 435-471.
273. Whyte, L. S.; Ryberg, E.; Sims, N. A.; Ridge, S. A.; Mackie, K.; Greasley, P. J.; Ross, R. A.; Rogers, M. J., The putative cannabinoid receptor GPR55 affects osteoclast function in vitro and bone mass in vivo. *Proc Natl Acad Sci U S A* **2009**, 106, 16511-6.
274. Dupont, S.; Krust, A.; Gansmuller, A.; Dierich, A.; Chambon, P.; Mark, M., Effect of single and compound knockouts of estrogen receptors alpha (ERalpha) and beta (ERbeta) on mouse reproductive phenotypes. *Development* **2000**, 127, 4277-4291.
275. Lauckner, J. E.; Jensen, J. B.; Chen, H.-Y.; Lu, H.-C.; Hille, B.; Mackie, K., GPR55 is a cannabinoid receptor that increases intracellular calcium and inhibits M current. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, 105, 2699-2704.
276. Fucic, A.; Miškov, S.; Želježić, D.; Bogdanovic, N.; Katić, J.; Gjergja, R.; Karelson, E.; Gamulin, M., Is the role of estrogens and estrogen receptors in epilepsy still underestimated? *Medical Hypotheses* **2009**, 73, 703-705.
277. Mindnich, R.; Möller, G.; Adamski, J., The role of 17 beta-hydroxysteroid dehydrogenases. *Molecular and cellular endocrinology* **2004**, 218, 7-20.
278. Le, T. T. T.; Gardner, J.; Hoang-Le, D.; Schmidt, C. W.; MacDonald, K. P.; Lambley, E.; Schroder, W. A.; Ogbourne, S. M.; Suhrbier, A., Immunostimulatory cancer chemotherapy using local ingenol-3-angelate and synergy with immunotherapies. *Vaccine* **2009**, 27, 3053-3062.
279. Nelson, A. W.; Tilley, W. D.; Neal, D. E.; Carroll, J. S., Estrogen receptor beta in prostate cancer: friend or foe? *Endocrine-related cancer* **2014**, 21, T219-T234.
280. Palmieri, C.; Cheng, G. J.; Saji, S.; Zelada-Hedman, M.; Warri, A.; Weihua, Z.; Van Noorden, S.; Wahlstrom, T.; Coombes, R. C.; Warner, M.; Gustafsson, J.-A., Estrogen receptor beta in breast cancer. *Endocrine-Related Cancer* **2002**, 9, 1-13.
281. Bauer, J.; Isojarvi, J. I.; Herzog, A. G.; Reuber, M.; Polson, D.; Tauboll, E.; Genton, P.; van der Ven, H.; Roesing, B.; Luef, G. J.; Galimberti, C. A.; van Parys, J.; Flugel, D.; Bergmann, A.; Elger, C. E., Reproductive dysfunction in women with epilepsy: recommendations for evaluation and management. *Journal of neurology, neurosurgery, and psychiatry* **2002**, 73, 121-5.
282. Calabro, R. S.; Marino, S.; Bramanti, P., Sexual and reproductive dysfunction associated with antiepileptic drug use in men with epilepsy. *Expert review of neurotherapeutics* **2011**, 11, 887-95.

283. Hussein, A. A.; Bozzi, B.; Correa, M.; Capson, T. L.; Kursar, T. A.; Coley, P. D.; Solis, P. N.; Gupta, M. P., Bioactive Constituents from Three Vismia Species. *Journal of Natural Products* **2003**, 66, 858-860.
284. Halse-Gramkow, M.; Ernst, M.; Rønsted, N.; Dunn, R. R.; Saslis-Lagoudakis, C. H., Using evolutionary tools to search for novel psychoactive plants. *Plant Genetic Resources* **2016**, 14, 246-255.
285. Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*, Ralph, A. W.; David, C. S., Eds.; Elsevier: 2008; Vol. Volume 4, 217-241.
286. Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A., HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic acids research* **2013**, 41, D801-7.
287. Webb, C. O.; Ackerly, D. D.; Kembel, S. W., Phylocom: software for the analysis of phylogenetic community structure and trait evolution. *Bioinformatics (Oxford, England)* **2008**, 24, 2098-2100.
288. Paradis, E.; Claude, J.; Strimmer, K., APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England)* **2004**, 20.
289. Patristic Distance. In *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*; Springer Netherlands: Dordrecht, 2008, 1454-1454.
290. In; Conservatoire et Jardin botaniques de la Ville de Genève and South African National Biodiversity Institute ,Pretoria, June 2017.
291. Suzuki, R.; Shimodaira, H., Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics (Oxford, England)* **2006**, 22, 1540-1542.
292. Kingston, D. G., Tubulin-interactive natural products as anticancer agents. *J Nat Prod* **2009**, 72, 507-15.
293. Kavallaris, M., Microtubules and resistance to tubulin-binding agents. *Nat Rev Cancer* **2010**, 10, 194-204.
294. Shields, M.; Niazi, U.; Badal, S.; Yee, T.; Sutcliffe, M. J.; Delgoda, R., Inhibition of CYP1A1 by Quassinoids found in Picrasma excelsa. *Planta Med* **2009**, 75, 137-41.
295. Houel, E.; Bertani, S.; Bourdy, G.; Deharo, E.; Jullian, V.; Valentin, A.; Chevalley, S.; Stien, D., Quassinoid constituents of Quassia amara L. leaf herbal tea. Impact on its antimalarial activity and cytotoxicity. *J Ethnopharmacol* **2009**, 126, 114-8.
296. Kupchan, S. M.; Streelman, D. R., Quassimarin, a new antileukemic quassinoid from Quassia amara. *The Journal of Organic Chemistry* **1976**, 41, 3481-3482.
297. de Melo, M.; #xf4; Santos, n.; Quintans, J. d. S. S.; Ara; #xfa; jo, A. A. d. S.; Duarte, M. C.; Bonjardim, L. R.; Nogueira, P. C. d. L.; Moraes, V.; #xe9; Souza, r. R. d.; Ara; #xfa; jo-J; #xfa; nior, J.; #xe3; de, o. X.; Ribeiro; #xca; Ad, u.; #xe9; Nogueira, l.; Quintans-J; #xfa; nior, L. J.; #xe9, A Systematic Review for Anti-Inflammatory Property of Clusiaceae Family: A Preclinical Approach. *Evidence-Based Complementary and Alternative Medicine* **2014**, 2014, 10.
298. Liu, K.; Abdullah, A. A.; Huang, M.; Nishioka, T.; Altaf-Ul-Amin, M.; Kanaya, S., Novel Approach to Classify Plants Based on Metabolite-Content Similarity. *BioMed Research International* **2017**, 2017, 12.

299. Baikar, S.; Malpathak, N., Secondary metabolites as DNA topoisomerase inhibitors: A new era towards designing of anticancer drugs. *Pharmacognosy Reviews* **2010**, 4, 12-26.
300. Gupta, S. C.; Prasad, S.; Sethumadhavan, D. R.; Nair, M. S.; Mo, Y.-Y.; Aggarwal, B. B., Nimbolide, a Limonoid Triterpene, Inhibits Growth of Human Colorectal Cancer Xenografts by Suppressing the Proinflammatory Microenvironment. *Clinical cancer research : an official journal of the American Association for Cancer Research* **2013**, 19, 4465-4476.
301. Saslis-Lagoudakis, C. H.; Williamson, E. M.; Savolainen, V.; Hawkins, J. A., Cross-cultural comparison of three medicinal floras and implications for bioprospecting strategies. *J Ethnopharmacol* **2011**, 135, 476-87.
302. Li, S., Camptotheca Lowreyana, A New Species of Anti-Cancer Happytrees. *NCPC Publications and Patents* **1997**, Paper 50.
303. Boyd, M. R.; Hallock, Y. F.; Cardellina, J. H.; Manfredi, K. P.; Blunt, J. W.; McMahon, J. B.; Buckheit, R. W.; Bringmann, G.; Schäffer, M.; Cragg, G. M.; Thomas, D. W.; Jato, J. G., Anti-HIV Michellamines from Ancistrocladus korupensis. *Journal of Medicinal Chemistry* **1994**, 37, 1740-1745.
304. Hallock, Y. F.; Manfredi, K. P.; Blunt, J. W.; Cardellina, J. H.; Schaeffer, M.; Gulden, K.-P.; Bringmann, G.; Lee, A. Y.; Clardy, J., Korupensamines A-D, Novel Antimalarial Alkaloids from Ancistrocladus korupensis. *The Journal of Organic Chemistry* **1994**, 59, 6349-6355.
305. Samarghandian, S.; Hadjzadeh, M.-A.-R.; Afshari, J. T.; Hosseini, M., Antiproliferative activity and induction of apoptotic by ethanolic extract of Alpinia galanga rhizhome in human breast carcinoma cell line. *BMC Complementary and Alternative Medicine* **2014**, 14, 192-192.
306. Ghil, S., Antiproliferative activity of Alpinia officinarum extract in the human breast cancer cell line MCF-7. *Molecular medicine reports* **2013**, 7, 1288-92.
307. Karmakar, I.; Dolai, N.; Suresh Kumar, R. B.; Kar, B.; Roy, S. N.; Haldar, P. K., Antitumor activity and antioxidant property of Curcuma caesia against Ehrlich's ascites carcinoma bearing mice. *Pharmaceutical biology* **2013**, 51, 753-9.
308. Chen, S. D.; Gao, J. T.; Liu, J. G.; Liu, B.; Zhao, R. Z.; Lu, C. J., Five new diarylheptanoids from the rhizomes of Curcuma kwangsiensis and their antiproliferative activity. *Fitoterapia* **2015**, 102, 67-73.
309. Rouhollahi, E.; Moghadamtousi, S. Z.; Hajiaghaalipour, F.; Zahedifard, M.; Tayeby, F.; Awang, K.; Abdulla, M. A.; Mohamed, Z., Curcuma purpurascens BI. rhizome accelerates rat excisional wound healing: involvement of Hsp70/Bax proteins, antioxidant defense, and angiogenesis activity. *Drug design, development and therapy* **2015**, 9, 5805-13.
310. Teh, S. S.; Ee, G. C.; Mah, S. H.; Lim, Y. M.; Ahmad, Z., Cytotoxicity and structure-activity relationships of xanthone derivatives from Mesua beccariana, Mesua ferrea and Mesua congestiflora towards nine human cancer cell lines. *Molecules* **2013**, 18, 1985-94.
311. Teh, S. S.; Cheng Lian Ee, G.; Mah, S. H.; Lim, Y. M.; Rahmani, M., Mesua beccariana (Clusiaceae), a source of potential anti-cancer lead compounds in drug discovery. *Molecules* **2012**, 17, 10791-800.
312. Fankam, A. G.; Das, R.; Mallick, A.; Kuiate, J. R.; Hazra, B.; Mandal, C.; Kuete, V., Cytotoxicity of the extracts and fractions from Allanblackia gabonensis (Clusiaceae) towards a panel of cancer cell lines. *South African Journal of Botany* **2017**, 111, 29-36.

313. Seruji, N. M. U.; Khong, H. Y.; Kutoi, C. J., Antioxidant, Anti-Inflammatory, and Cytotoxic Activities of Garcinia nervosa (Clusiaceae). *Journal of Chemistry* **2013**, 2013, 5.
314. Mariano, L. N. B.; Vendramini-Costa, D. B.; Ruiz, A. L. T. G.; de Carvalho, J. E.; Corrêa, R.; Cechinel Filho, V.; Delle Monache, F.; Niero, R., In vitro antiproliferative activity of uncommon xanthones from branches of Garcinia achachairu. *Pharmaceutical biology* **2016**, 54, 1697-1704.
315. Schmidt, T. J.; Khalid, S. A.; Romanha, A. J.; Alves, T. M.; Biavatti, M. W.; Brun, R.; Da Costa, F. B.; de Castro, S. L.; Ferreira, V. F.; de Lacerda, M. V.; Lago, J. H.; Leon, L. L.; Lopes, N. P.; das Neves Amorim, R. C.; Niehues, M.; Ogungbe, I. V.; Pohlit, A. M.; Scotti, M. T.; Setzer, W. N.; de, N. C. S. M.; Steindel, M.; Tempone, A. G., The potential of secondary metabolites from plants as drugs or leads against protozoan neglected diseases - part II. *Current medicinal chemistry* **2012**, 19, 2176-228.
316. Rakotomanga, M.; Razakantoanina, V.; Raynaud, S.; Loiseau, P. M.; Hocquemiller, R.; Jaureguiberry, G., Antiplasmodial activity of acetogenins and inhibitory effect on Plasmodium falciparum adenylate translocase. *Journal of chemotherapy (Florence, Italy)* **2004**, 16, 350-6.
317. Ambrozin, A. R. P.; Vieira, P. C.; Fernandes, J. B.; Silva, M. F. d. G. F. d.; Albuquerque, S. d., Trypanocidal activity of Meliaceae and Rutaceae plant extracts. *Memórias do Instituto Oswaldo Cruz* **2004**, 99, 227-231.
318. Setzer, W. N.; Ogungbe, I. V., In-silico Investigation of Antitrypanosomal Phytochemicals from Nigerian Medicinal Plants. *PLoS Neglected Tropical Diseases* **2012**, 6, e1727.
319. Lukhoba, C. W.; Simmonds, M. S.; Paton, A. J., Plectranthus: a review of ethnobotanical uses. *J Ethnopharmacol* **2006**, 103, 1-24.
320. Grace, O. M.; Buerki, S.; Symonds, M. R.; Forest, F.; van Wyk, A. E.; Smith, G. F.; Klopper, R. R.; Bjora, C. S.; Neale, S.; Demissew, S.; Simmonds, M. S.; Ronsted, N., Evolutionary history and leaf succulence as explanations for medicinal use in aloes and the global popularity of Aloe vera. *BMC evolutionary biology* **2015**, 15, 29.
321. Ernst, M.; Saslis-Lagoudakis, C. H.; Grace, O. M.; Nilsson, N.; Simonsen, H. T.; Horn, J. W.; Ronsted, N., Evolutionary prediction of medicinal properties in the genus Euphorbia L. *Sci Rep* **2016**, 6, 30531.
322. Molander, M.; Saslis-Lagoudakis, C. H.; Jager, A. K.; Ronsted, N., Cross-cultural comparison of medicinal floras used against snakebites. *J Ethnopharmacol* **2012**, 139, 863-72.
323. Bringmann, G.; Feineis, D., Stress-related polyketide metabolism of Dioncophyllaceae and Ancistrocladaceae. *Journal of Experimental Botany* **2001**, 52, 2015-2022.
324. Francois, G.; Timperman, G.; Eling, W.; Assi, L. A.; Holenz, J.; Bringmann, G., Naphthylisoquinoline alkaloids against malaria: evaluation of the curative potentials of dioncophylline C and dioncopeltine A against Plasmodium berghei in vivo. *Antimicrob Agents Chemother* **1997**, 41, 2533-9.
325. Bringmann, G.; Hertlein-Amslinger, B.; Kajahn, I.; Dreyer, M.; Brun, R.; Moll, H.; Stich, A.; Ioset, K. N.; Schmitz, W.; Ngoc, L. H., Phenolic analogs of the N,C-coupled naphthylisoquinoline alkaloid ancistrocladinium A, from Ancistrocladus cochinchinensis (Ancistrocladaceae), with improved antiprotozoal activities. *Phytochemistry* **2011**, 72, 89-93.

326. Ibezim, A.; Debnath, B.; Ntie-Kang, F.; Mbah, C. J.; Nwodo, N. J., Binding of anti-Trypanosoma natural products from African flora against selected drug targets: a docking study. *Medicinal Chemistry Research* **2017**, 26, 562-579.
327. Ntie-Kang, F.; Onguene, P.; Lifongo, L.; Ndom, J.; Sippl, W.; Mbaze, L., The potential of anti-malarial compounds derived from African medicinal plants, part II: a pharmacological evaluation of non-alkaloids and non-terpenoids. *Malaria Journal* **2014**, 13, 81.
328. Karan, M.; Bhatnagar, S.; Wangtak, P.; Vasisht, K. Phytochemical and antimalarial studies on *Swertia alata* Royle. 2005; International Society for Horticultural Science (ISHS), Leuven, Belgium: 2005; 139-145.
329. Santos, D. A. P. d.; Vieira, P. C.; Silva, M. F. d. G. F. d.; Fernandes, J. B.; Rattray, L.; Croft, S. L., Antiparasitic activities of acridone alkaloids from Swinglea glutinosa (Bl.) Merr. *Journal of the Brazilian Chemical Society* **2009**, 20, 644-651.
330. Bringmann, G.; Saeb, W.; Rückert, M.; Mies, J.; Michel, M.; Mudogo, V.; Brun, R., Ancistrolikokine D, a 5,8′-coupled naphthylisoquinoline alkaloid, and related natural products from Ancistrocladus likoko. *Phytochemistry* **2003**, 62, 631-636.
331. Azebaze, A. G.; Menasria, F.; Noumi, L. G.; Nguemfo, E. L.; Tchamfo, M. F.; Nkengfack, A. E.; Kolb, J. P.; Meyer, M., Xanthones from the seeds of Allanblackia monticola and their apoptotic and antiproliferative activities. *Planta Med* **2009**, 75, 243-8.
332. Kamdem Waffo, A. F.; Coombes, P. H.; Crouch, N. R.; Mulholland, D. A.; El Amin, S. M. M.; Smith, P. J., Acridone and furoquinoline alkaloids from Teclea gerrardii (Rutaceae: Toddalioideae) of southern Africa. *Phytochemistry* **2007**, 68, 663-667.
333. Brun, R.; Blum, J.; Chappuis, F.; Burri, C., Human African trypanosomiasis. *The Lancet* **2010**, 375, 148-159.
334. Paine, M. F.; Wang, M. Z.; Generaux, C. N.; Boykin, D. W.; Wilson, W. D.; De Koning, H. P.; Olson, C. A.; Pohlig, G.; Burri, C.; Brun, R.; Murilla, G. A.; Thuita, J. K.; Barrett, M. P.; Tidwell, R. R., Diamidines for human African trypanosomiasis. *Current opinion in investigational drugs (London, England : 2000)* **2010**, 11, 876-83.
335. Brand, S.; Cleghorn, L. A.; McElroy, S. P.; Robinson, D. A.; Smith, V. C.; Hallyburton, I.; Harrison, J. R.; Norcross, N. R.; Spinks, D.; Bayliss, T.; Norval, S.; Stojanovski, L.; Torrie, L. S.; Frearson, J. A.; Brenk, R.; Fairlamb, A. H.; Ferguson, M. A.; Read, K. D.; Wyatt, P. G.; Gilbert, I. H., Discovery of a novel class of orally active trypanocidal N-myristoyltransferase inhibitors. *J Med Chem* **2012**, 55, 140-52.
336. Ding, D.; Zhao, Y.; Meng, Q.; Xie, D.; Nare, B.; Chen, D.; Bacchi, C. J.; Yarlett, N.; Zhang, Y.-K.; Hernandez, V.; Xia, Y.; Freund, Y.; Abdulla, M.; Ang, K.-H.; Ratnam, J.; McKerrow, J. H.; Jacobs, R. T.; Zhou, H.; Plattner, J. J., Discovery of Novel Benzoxaborole-Based Potent Antitrypanosomal Agents. *ACS medicinal chemistry letters* **2010**, 1, 165-169.
337. Martínez-Jiménez, F.; Papadatos, G.; Yang, L.; Wallace, I. M.; Kumar, V.; Pieper, U.; Sali, A.; Brown, J. R.; Overington, J. P.; Marti-Renom, M. A., Target Prediction for an Open Access Set of Compounds Active against Mycobacterium tuberculosis. *PLOS Computational Biology* **2013**, 9, e1003253.
338. Spitzmüller, A.; Mestres, J., Prediction of the P. falciparum Target Space Relevant to Malaria Drug Discovery. *PLOS Computational Biology* **2013**, 9, e1003257.

339. Ekins, S.; Lage de Siqueira-Neto, J.; McCall, L.-I.; Sarker, M.; Yadav, M.; Ponder, E. L.; Kallel, E. A.; Kellar, D.; Chen, S.; Arkin, M.; Bunin, B. A.; McKerrow, J. H.; Talcott, C., Machine Learning Models and Pathway Genome Data Base for Trypanosoma cruzi Drug Discovery. *PLOS Neglected Tropical Diseases* **2015**, 9, e0003878.
340. Zweygarth, E.; Kaminsky, R., Evaluation of DL-alpha-difluoromethylornithine against susceptible and drug-resistant Trypanosoma brucei brucei. *Acta Trop* **1991**, 48, 223-32.
341. Sykes, M. L.; Avery, V. M., Development of an Alamar Blue viability assay in 384-well format for high throughput whole cell screening of Trypanosoma brucei brucei bloodstream form strain 427. *The American journal of tropical medicine and hygiene* **2009**, 81, 665-74.
342. Kaiser, M.; Mäser, P.; Tadoori, L. P.; Ioset, J.-R.; Brun, R., Antiprotozoal Activity Profiling of Approved Drugs: A Starting Point toward Drug Repositioning. *PLOS ONE* **2015**, 10, e0135556.
343. Sykes, M. L.; Baell, J. B.; Kaiser, M.; Chatelain, E.; Moawad, S. R.; Ganame, D.; Ioset, J.-R.; Avery, V. M., Identification of Compounds with Anti-Proliferative Activity against Trypanosoma brucei brucei Strain 427 by a Whole Cell Viability Based HTS Campaign. *PLOS Neglected Tropical Diseases* **2012**, 6, e1896.
344. Peña, I.; Pilar Manzano, M.; Cantizani, J.; Kessler, A.; Alonso-Padilla, J.; Bardera, A. I.; Alvarez, E.; Colmenarejo, G.; Cotillo, I.; Roquero, I.; de Dios-Anton, F.; Barroso, V.; Rodriguez, A.; Gray, D. W.; Navarro, M.; Kumar, V.; Sherstnev, A.; Drewry, D. H.; Brown, J. R.; Fiandor, J. M.; Julio Martin, J., New Compound Sets Identified from High Throughput Phenotypic Screening Against Three Kinetoplastid Parasites: An Open Resource. *Scientific Reports* **2015**, 5, 8771.
345. Schmidt, T. J.; Khalid, S. A.; Romanha, A. J.; Alves, T. M.; Biavatti, M. W.; Brun, R.; Da Costa, F. B.; de Castro, S. L.; Ferreira, V. F.; de Lacerda, M. V.; Lago, J. H.; Leon, L. L.; Lopes, N. P.; das Neves Amorim, R. C.; Niehues, M.; Ogungbe, I. V.; Pohlit, A. M.; Scotti, M. T.; Setzer, W. N.; de, N. C. S. M.; Steindel, M.; Tempone, A. G., The potential of secondary metabolites from plants as drugs or leads against protozoan neglected diseases - part I. *Current medicinal chemistry* **2012**, 19, 2128-75.
346. Ntie-Kang, F.; Zofou, D.; Babiaka, S. B.; Meudom, R.; Scharfe, M.; Lifongo, L. L.; Mbah, J. A.; Mbaze, L. M. a.; Sippl, W.; Efange, S. M. N., AfroDb: A Select Highly Potent and Diverse Natural Product Library from African Medicinal Plants. *PLoS ONE* **2013**, 8, e78085.
347. Zeng, X.; Zhang, P.; He, W.; Qin, C.; Chen, S.; Tao, L.; Wang, Y.; Tan, Y.; Gao, D.; Wang, B.; Chen, Z.; Chen, W.; Jiang, Y. Y.; Chen, Y. Z., NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic acids research* **2018**, 46, D1217-D1222.
348. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, 3, 33.
349. Morgan, H. L., The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, 5, 107-113.
350. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; K\, T.; \#246; tter; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B., KNIME - the Konstanz information miner: version 2.0 and beyond. *SIGKDD Explor. Newsl.* **2009**, 11, 26-31.

351. Borgelt, C.; Berthold, M. R., In *Proceedings of the 2002 IEEE International Conference on Data Mining*; IEEE Computer Society: 2002, 51.
352. Clark, D. E., Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *J Pharm Sci* **1999**, 88, 815-21.
353. Abraham, M. H.; Takacs-Novak, K.; Mitchell, R. C., On the partition of ampholytes: application to blood-brain distribution. *J Pharm Sci* **1997**, 86, 310-5.
354. QikProp *QikProp, version 3.5, Schrödinger, LLC, New York, NY, 2012.* , 2017.
355. Lewis H. Mervin, A. M. A., Krishna C. Bulusu, Leen Kalash, Fredrik Svensson, Mike A. Firth, Ian P. Barrett, Ola Engkvist, Andreas Bender, Analysis of Orthologue Chemical Space and Extending In Silico Target Prediction by Including Orthologue Bioactivity Data. *Under preparation* **2017**.
356. Gaulton, A.; Bellis, L.; Bento, A.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **2012**, 40, D1100 - D1107.
357. Aguero, F.; Al-Lazikani, B.; Aslett, M.; Berriman, M.; Buckner, F. S.; Campbell, R. K.; Carmona, S.; Carruthers, I. M.; Chan, A. W. E.; Chen, F.; Crowther, G. J.; Doyle, M. A.; Hertz-Fowler, C.; Hopkins, A. L.; McAllister, G.; Nwaka, S.; Overington, J. P.; Pain, A.; Paolini, G. V.; Pieper, U.; Ralph, S. A.; Riechers, A.; Roos, D. S.; Sali, A.; Shanmugam, D.; Suzuki, T.; Van Voorhis, W. C.; Verlinde, C. L. M. J., Genomic-scale prioritization of drug targets: the TDR Targets database. *Nat Rev Drug Discov* **2008**, 7, 900-907.
358. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **2003**, 13, 2498-504.
359. Aslett, M.; Aurrecoechea, C.; Berriman, M.; Brestelli, J.; Brunk, B. P.; Carrington, M.; Depledge, D. P.; Fischer, S.; Gajria, B.; Gao, X.; Gardner, M. J.; Gingle, A.; Grant, G.; Harb, O. S.; Heiges, M.; Hertz-Fowler, C.; Houston, R.; Innamorato, F.; Iodice, J.; Kissinger, J. C.; Kraemer, E.; Li, W.; Logan, F. J.; Miller, J. A.; Mitra, S.; Myler, P. J.; Nayak, V.; Pennington, C.; Phan, I.; Pinney, D. F.; Ramasamy, G.; Rogers, M. B.; Roos, D. S.; Ross, C.; Sivam, D.; Smith, D. F.; Srinivasamoorthy, G.; Stoeckert, C. J., Jr.; Subramanian, S.; Thibodeau, R.; Tivey, A.; Treatman, C.; Velarde, G.; Wang, H., TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic acids research* **2010**, 38, D457-62.
360. Supek, F.; Bošnjak, M.; Škunca, N.; Šmuc, T., REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE* **2011**, 6, e21800.
361. Faria, J.; Moraes, C. B.; Song, R.; Pascoalino, B. S.; Lee, N.; Siqueira-Neto, J. L.; Cruz, D. J.; Parkinson, T.; Ioset, J. R.; Cordeiro-da-Silva, A.; Freitas-Junior, L. H., Drug discovery for human African trypanosomiasis: identification of novel scaffolds by the newly developed HTS SYBR Green assay for Trypanosoma brucei. *Journal of biomolecular screening* **2015**, 20, 70-81.
362. Behera, R.; Thomas, S. M.; Mensa-Wilmot, K., New chemical scaffolds for human african trypanosomiasis lead discovery from a screen of tyrosine kinase inhibitor drugs. *Antimicrob Agents Chemother* **2014**, 58, 2202-10.
363. Kruger, F. A.; Overington, J. P., Global Analysis of Small Molecule Binding to Related Protein Targets. *PLOS Computational Biology* **2012**, 8, e1002333.

364. Dimova, D.; Stumpfe, D.; Bajorath, J., Identification of orthologous target pairs with shared active compounds and comparison of organism-specific activity patterns. *Chem Biol Drug Des* **2015**, 86, 1105-14.
365. Paricharak, S.; Klenka, T.; Augustin, M.; Patel, U. A.; Bender, A., Are phylogenetic trees suitable for chemogenomics analyses of bioactivity data sets: the importance of shared active compounds and choosing a suitable data embedding method, as exemplified on Kinases. *J Cheminform* **2013**, 5, 49.
366. Smithson, D. C.; Lee, J.; Shelat, A. A.; Phillips, M. A.; Guy, R. K., Discovery of potent and selective inhibitors of Trypanosoma brucei ornithine decarboxylase. *The Journal of biological chemistry* **2010**, 285, 16771-81.
367. Kim, D. J.; Roh, E.; Lee, M. H.; Oi, N.; Lim, D. Y.; Kim, M. O.; Cho, Y. Y.; Pugliese, A.; Shim, J. H.; Chen, H.; Cho, E. J.; Kim, J. E.; Kang, S. C.; Paul, S.; Kang, H. E.; Jung, J. W.; Lee, S. Y.; Kim, S. H.; Reddy, K.; Yeom, Y. I.; Bode, A. M.; Dong, Z., Herbacetin Is a Novel Allosteric Inhibitor of Ornithine Decarboxylase with Antitumor Activity. *Cancer Res* **2016**, 76, 1146-1157.
368. Zimmermann, S.; Oufir, M.; Leroux, A.; Krauth-Siegel, R. L.; Becker, K.; Kaiser, M.; Brun, R.; Hamburger, M.; Adams, M., Cynaropicrin targets the trypanothione redox system in Trypanosoma brucei. *Bioorganic & medicinal chemistry* **2013**, 21, 7202-9.
369. Elsebai, M. F.; Mocan, A.; Atanasov, A. G., Cynaropicrin: A Comprehensive Research Review and Therapeutic Potential As an Anti-Hepatitis C Virus Agent. *Frontiers in Pharmacology* **2016**, 7, 472.
370. Harikumar, K. B.; Kunnumakkara, A. B.; Ahn, K. S.; Anand, P.; Krishnan, S.; Guha, S.; Aggarwal, B. B., Modification of the cysteine residues in IκBα kinase and NF-κB (p65) by xanthohumol leads to suppression of NF-κB–regulated gene products and potentiation of apoptosis in leukemia cells. *Blood* **2009**, 113, 2003-2013.
371. Dietz, B. M.; Kang, Y.-H.; Liu, G.; Eggler, A. L.; Yao, P.; Chadwick, L. R.; Pauli, G. F.; Farnsworth, N. R.; Mesecar, A. D.; van Breemen, R. B.; Bolton, J. L., Xanthohumol Isolated from Humulus lupulus Inhibits Menadione-Induced DNA Damage through Induction of Quinone Reductase. *Chemical Research in Toxicology* **2005**, 18, 1296-1305.
372. Docampo, R.; Huang, G., Calcium signaling in trypanosomatid parasites. *Cell calcium* **2015**, 57, 194-202.
373. Masocha, W.; Amin, D. N.; Kristensson, K.; Rottenberg, M. E., Differential Invasion of Trypanosoma brucei brucei and Lymphocytes into the Brain of C57BL/6 and 129Sv/Ev Mice. *Scandinavian Journal of Immunology* **2008**, 68, 484-491.
374. Elsheikha, H. M.; Khan, N. A., Protozoa traversal of the blood–brain barrier to invade the central nervous system. *FEMS Microbiology Reviews* **2010**, 34, 532-553.
375. Babokhov, P.; Sanyaolu, A. O.; Oyibo, W. A.; Fagbenro-Beyioku, A. F.; Iriemenam, N. C., A current analysis of chemotherapy strategies for the treatment of human African trypanosomiasis. *Pathogens and Global Health* **2013**, 107, 242-252.
376. Matthews, K. R., The developmental cell biology of Trypanosoma brucei. *Journal of cell science* **2005**, 118, 283-290.

377. Wyatt, P. G.; Gilbert, I. H.; Read, K. D.; Fairlamb, A. H., Target validation: linking target and chemical properties to desired product profile. *Current topics in medicinal chemistry* **2011**, 11, 1275-83.
378. Naula, C.; Parsons, M.; Mottram, J. C., Protein kinases as drug targets in trypanosomes and Leishmania. *Biochimica et biophysica acta* **2005**, 1754, 151-159.
379. Shah, N. P.; Tran, C.; Lee, F. Y.; Chen, P.; Norris, D.; Sawyers, C. L., Overriding imatinib resistance with a novel ABL kinase inhibitor. *Science* **2004**, 305, 399-401.
380. Jetton, N.; Rothberg, K. G.; Hubbard, J. G.; Wise, J.; Li, Y.; Ball, H. L.; Ruben, L., The cell cycle as a therapeutic target against Trypanosoma brucei: Hesperadin inhibits Aurora kinase-1 and blocks mitotic progression in bloodstream forms. *Mol Microbiol* **2009**, 72, 442-58.
381. Jones, N. G.; Thomas, E. B.; Brown, E.; Dickens, N. J.; Hammarton, T. C.; Mottram, J. C., Regulators of Trypanosoma brucei Cell Cycle Progression and Differentiation Identified Using a Kinome-Wide RNAi Screen. *PLOS Pathogens* **2014**, 10, e1003886.
382. Ojo, K. K.; Gillespie, J. R.; Riechers, A. J.; Napuli, A. J.; Verlinde, C. L. M. J.; Buckner, F. S.; Gelb, M. H.; Domostoj, M. M.; Wells, S. J.; Scheer, A.; Wells, T. N. C.; Van Voorhis, W. C., Glycogen Synthase Kinase 3 Is a Potential Drug Target for African Trypanosomiasis Therapy. *Antimicrobial Agents and Chemotherapy* **2008**, 52, 3710-3717.
383. Albert, M.-A.; Haanstra, J. R.; Hannaert, V.; Van Roy, J.; Opperdoes, F. R.; Bakker, B. M.; Michels, P. A. M., Experimental and in Silico Analyses of Glycolytic Flux Control in Bloodstream Form Trypanosoma brucei. *Journal of Biological Chemistry* **2005**, 280, 28306-28315.
384. Brimacombe, K. R.; Walsh, M. J.; Liu, L.; Vasquez-Valdivieso, M. G.; Morgan, H. P.; McNae, I.; Fothergill-Gilmore, L. A.; Michels, P. A.; Auld, D. S.; Simeonov, A.; Walkinshaw, M. D.; Shen, M.; Boxer, M. B., Identification of ML251, a Potent Inhibitor of T. brucei and T. cruzi Phosphofructokinase. *ACS medicinal chemistry letters* **2014**, 5, 12-7.
385. Brun, R.; Don, R.; Jacobs, R. T.; Wang, M. Z.; Barrett, M. P., Development of novel drugs for human African trypanosomiasis. *Future microbiology* **2011**, 6, 677-91.
386. Priotto, G.; Pinoges, L.; Fursa, I. B.; Burke, B.; Nicolay, N.; Grillet, G.; Hewison, C.; Balasegaram, M., Safety and effectiveness of first line eflornithine for Trypanosoma brucei gambiense sleeping sickness in Sudan: cohort study. *BMJ (Clinical research ed.)* **2008**, 336, 705-8.
387. Woodland, A.; Thompson, S.; Cleghorn, L. A.; Norcross, N.; De Rycker, M.; Grimaldi, R.; Hallyburton, I.; Rao, B.; Norval, S.; Stojanovski, L.; Brun, R.; Kaiser, M.; Frearson, J. A.; Gray, D. W.; Wyatt, P. G.; Read, K. D.; Gilbert, I. H., Discovery of Inhibitors of Trypanosoma brucei by Phenotypic Screening of a Focused Protein Kinase Library. *ChemMedChem* **2015**, 10, 1809-20.
388. Brenk, R.; Schipani, A.; James, D.; Krasowski, A.; Gilbert, I. H.; Frearson, J.; Wyatt, P. G., Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *Chemmedchem* **2008**, 3, 435-444.
389. Gilbert, I. H., Inhibitors of dihydrofolate reductase in Leishmania and trypanosomes. *Biochimica et biophysica acta* **2002**, 1587, 249-57.

390. Sienkiewicz, N.; Jarosławski, S.; Wyllie, S.; Fairlamb, A. H., Chemical and genetic validation of dihydrofolate reductase–thymidylate synthase as a drug target in African trypanosomes. *Molecular Microbiology* **2008**, 69, 520-533.
391. Hammarton, T. C.; Mottram, J. C.; Doerig, C., The cell cycle of parasitic protozoa: potential for chemotherapeutic exploitation. *Progress in cell cycle research* **2003**, 5, 91-101.
392. Mony, B. M.; MacGregor, P.; Ivens, A.; Rojas, F.; Cowton, A.; Young, J.; Horn, D.; Matthews, K., Genome-wide dissection of the quorum sensing signalling pathway in Trypanosoma brucei. *Nature* **2014**, 505, 681-685.
393. Vassella, E.; Krämer, R.; Turner, C. M. R.; Wankell, M.; Modes, C.; Van Den Bogaard, M.; Boshart, M., Deletion of a novel protein kinase with PX and FYVE-related domains increases the rate of differentiation of Trypanosoma brucei. *Molecular Microbiology* **2001**, 41, 33-46.
394. Stuart, K.; Brun, R.; Croft, S.; Fairlamb, A.; Gürtler, R. E.; McKerrow, J.; Reed, S.; Tarleton, R., Kinetoplastids: related protozoan pathogens, different diseases. *The Journal of Clinical Investigation* **2008**, 118, 1301-1310.
395. Chawla, B.; Madhubala, R., Drug targets in Leishmania. *Journal of Parasitic Diseases: Official Organ of the Indian Society for Parasitology* **2010**, 34, 1-13.
396. Sülsen, V. P.; Puente, V.; Papademetrio, D.; Batlle, A.; Martino, V. S.; Frank, F. M.; Lombardo, M. E., Mode of Action of the Sesquiterpene Lactones Psilostachyin and Psilostachyin C on Trypanosoma cruzi. *PLOS ONE* **2016**, 11, e0150526.
397. Saravolatz, L. D.; Bern, C.; Adler-Moore, J.; Berenguer, J.; Boelaert, M.; den Boer, M.; Davidson, R. N.; Figueras, C.; Gradoni, L.; Kafetzis, D. A.; Ritmeijer, K.; Rosenthal, E.; Royce, C.; Russo, R.; Sundar, S.; Alvar, J., Liposomal Amphotericin B for the Treatment of Visceral Leishmaniasis. *Clinical Infectious Diseases* **2006**, 43, 917-924.
398. Lacomble, S.; Vaughan, S.; Gadelha, C.; Morphew, M. K.; Shaw, M. K.; McIntosh, J. R.; Gull, K., Three-dimensional cellular architecture of the flagellar pocket and associated cytoskeleton in trypanosomes revealed by electron microscope tomography. *Journal of cell science* **2009**, 122, 1081-90.
399. Selmar, D.; Kleinwächter, M., Stress Enhances the Synthesis of Secondary Plant Products: The Impact of Stress-Related Over-Reduction on the Accumulation of Natural Products. *Plant and Cell Physiology* **2013**, 54, 817-826.
400. Mazid M, K. T., Mohammad F, Role of secondary metabolites in defense mechanisms of plants. *Biology and Medicine* **2011**, 3, 232-249.
401. Thomson, D. W.; Wagner, A. J.; Bantscheff, M.; Benson, R. E.; Dittus, L.; Duempelfeld, B.; Drewes, G.; Krause, J.; Moore, J. T.; Mueller, K.; Poeckel, D.; Rau, C.; Salzer, E.; Shewchuk, L.; Hopf, C.; Emery, J. G.; Muelbaier, M., Discovery of a Highly Selective Tankyrase Inhibitor Displaying Growth Inhibition Effects against a Diverse Range of Tumor Derived Cell Lines. *J Med Chem* **2017**, 60, 5455-5471.
402. Holmgren, A.; Bjornstedt, M. [21] Thioredoxin and thioredoxin reductase. In *Methods in Enzymology*; Academic Press: 1995; Vol. 252, 199-208.
403. Arner, E. S.; Sarioglu, H.; Lottspeich, F.; Holmgren, A.; Bock, A., High-level expression in Escherichia coli of selenocysteine-containing rat thioredoxin reductase utilizing gene fusions with engineered bacterial-type SECIS elements and co-expression with the selA, selB and selC genes. *Journal of molecular biology* **1999**, 292, 1003-16.

404. Arner, E. S.; Holmgren, A., Measurement of thioredoxin and thioredoxin reductase. *Current protocols in toxicology* **2001**, Chapter 7, Unit 7.4.
405. Nützmann, H.-W.; Huang, A.; Osbourn, A., Plant metabolic clusters – from genetics to genomics. *New Phytologist* **2016**, 211, 771-789.
406. Kjærbølling, I.; Vesth, T. C.; Frisvad, J. C.; Nybo, J. L.; Theobald, S.; Kuo, A.; Bowyer, P.; Matsuda, Y.; Mondo, S.; Lyhne, E. K.; Kogle, M. E.; Clum, A.; Lipzen, A.; Salamov, A.; Ngan, C. Y.; Daum, C.; Chiniquy, J.; Barry, K.; LaButti, K.; Haridas, S.; Simmons, B. A.; Magnuson, J. K.; Mortensen, U. H.; Larsen, T. O.; Grigoriev, I. V.; Baker, S. E.; Andersen, M. R., Linking secondary metabolites to gene clusters through genome sequencing of six diverse Aspergillus species. *Proceedings of the National Academy of Sciences* **2018**, 115, E753.
407. Voyton, C. M.; Morris, M. T.; Ackroyd, P. C.; Morris, J. C.; Christensen, K. A., FRET Cytometric Method for High Throughput Screening of Potential Metabolic Inhibitors in Trypanosoma brucei. *The FASEB Journal* **2016**, 30, lb143-lb143.
408. Schmidl, S.; Iancu, C. V.; Choe, J.-Y.; Oreb, M., Ligand Screening Systems for Human Glucose Transporters as Tools in Drug Discovery. *Frontiers in chemistry* **2018**, 6, 183-183.
409. Yeow, K.; Novo-Perez, L.; Gaillard, P.; Page, P.; Gotteland, J. P.; Scheer, A.; Lang, P., A cellular assay for measuring the inhibition of glycogen synthase kinase-3 via the accumulation of beta-catenin in Chinese hamster ovary clone K1 cells. *Assay and drug development technologies* **2006**, 4, 451-60.
410. Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; Kuhn, M.; Bork, P.; Jensen, L. J.; von Mering, C., STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research* **2015**, 43, D447-52.

SUPPLEMENTARY MATERIAL

Supplementary Table 1 List of cancer drugs obtained from NCI (NCI Cancer)

	ChEMBL ID	Drug Name	Canonical SMILES
1	CHEMBL34259	Methotrexate (BAN, FDA, INN, JAN, USAN, USP); Methotrexate Sodium (FDA);	CN (Cc1cnc2nc (N)nc (N)c2n1)c3ccc (cc3)C (=O)N[C@@H] (CCC (=O)O)C (=O)O
2	CHEMBL428647	Paclitaxel (BAN, FDA, INN, USAN, USP);	CC (=O)O[C@H]1C (=O)[C@]2 (C)[C@@H] (O)C[C@H]3OC[C@@]3 (OC (=O)C)[C@H]2[C@H] (OC (=O)c4ccccc4)[C@]5 (O)C[C@H] (OC (=O)[C@H] (O)[C@@H] (NC (=O)c6ccccc6)c7ccccc7)C (=C1C5 (C)C)C
3	CHEMBL1742994	Brentuximab Vedotin (FDA, INN, USAN);	
4	CHEMBL1398373	Pirarubicin (INN, JAN, MI);	COc1cccc2C (=O)c3c (O)c4C[C@] (O) (C[C@H] (O[C@H]5C[C@H] (N)[C@H] (O[C@H]6CCCCO6)[C@H] (C)O5)c4c (O)c3C (=O)c12)C (=O)CO
5	CHEMBL185	Fluorouracil (BAN, FDA, INN, JAN, USAN, USP);	FC1=CNC (=O)NC1=O
6	CHEMBL1908360	Everolimus (FDA, INN, USAN);	CO[C@@H]1C[C@H] (C[C@@H] (C)[C@@H]2CC (=O)[C@H] (C)\C=C (/C)\[C@@H] (O)[C@@H] (OC)C (=O)[C@H] (C)C[C@H] (C)\C=C\C=C\C=C (/C)\[C@H] (C[C@@H]3CC[C@@H] (C)[C@@] (O) (O3)C (=O)C (=O)N4CCCC[C@H]4C (=O)O2)OC)CC[C@H]1OCCO
7	CHEMBL1201258	Pemetrexed (BAN, INN); Pemetrexed Disodium (FDA, USAN);	NC1=Nc2[nH]cc (CCc3ccc (cc3)C (=O)N[C@H] (CCC (=O)O)C (=O)O)c2C (=O)N1
8	CHEMBL852	Melphalan HCl (FDA); Melphalan (BAN, FDA, INN, JAN, USAN, USP);	N[C@@H] (Cc1ccc (cc1)N (CCCl)CCCl)C (=O)O
9	CHEMBL834	Pamidronate Disodium (FDA, JAN, USAN); Pamidronic Acid (BAN, INN, MI);	NCCC (O) (P (=O) (O)O)P (=O) (O)O
10	CHEMBL1399	Anastrozole (BAN, FDA, INN, USAN);	CC (C) (C#N)c1cc (Cn2cncn2)cc (c1)C (C) (C)C#N

11	CHEMBL1200374	Exemestane (BAN, FDA, INN, USAN);	C[C@]12CC[C@H]3[C@@H] (CC (=C)C4=CC (=O)C=C[C@]34C)[C@@H]1CCC2=O
12	CHEMBL1201112	Nelarabine (BAN, FDA, INN, USAN);	COc1nc (N)nc2c1ncn2[C@@H]3O[C@H] (CO)[C@@H] (O)[C@@H]3O
13	CHEMBL1201836	Ofatumumab (FDA, INN, USAN);	
14	CHEMBL1201583	Bevacizumab (FDA, INN);	
15	CHEMBL1201604	Tositumomab (FDA, INN);	
16	CHEMBL513	Carmustine (BAN, FDA, INN, USAN);	ClCCNC (=O)N (CCCl)N=O
17	CHEMBL3039590	Bleomycin (INN); Bleomycin Sulfate (FDA, USAN, USP);	C[C@@H] (O)[C@H] (NC (=O)[C@@H] (C)[C@H] (O)[C@@H] (C)NC (=O)[C@@H] (NC (=O)c1nc (nc (N)c1C)[C@H] (CC (=O)N)NC[C@H] (N)C (=O)N)[C@@H] (O[C@@H]2O[C@@H] (CO)[C@@H] (O)[C@H] (O)[C@@H]2O[C@H]3O[C@H] (CO)[C@@H] (O)[C@H] (OC (=O)N)[C@@H]3O)c4c[nH]cn4)C (=O)NCCc5nc (cs5)c6ncc (s6)C (=O)NCCC[S+] (C)C
18	CHEMBL288441	Bosutinib (INN, USAN); Bosutinib Monohydrate (FDA);	COc1cc (Nc2c (cnc3cc (OCCCN4CCN (C)CC4)c (OC)cc23)C#N)c (Cl)cc1Cl
19			
20	CHEMBL1201587	Alemtuzumab (BAN, FDA, INN, USAN);	
21	CHEMBL481	Irinotecan (BAN, INN); Irinotecan HCl (FDA, JAN, USAN);	CCc1c2CN3C (=O)C4=C (C=C3c2nc5ccc (OC (=O)N6CCC (CC6)N7CCCCC7)cc15)[C@@] (O) (CC)C (=O)OC4
22	CHEMBL24828	Vandetanib (BAN, FDA, INN, USAN);	COc1cc2c (Nc3ccc (Br)cc3F)ncnc2cc1OCC4CCN (C)CC4
23	CHEMBL409	Bicalutamide (BAN, FDA, INN, USAN);	CC (O) (CS (=O) (=O)c1ccc (F)cc1)C (=O)Nc2ccc (C#N)c (c2)C (F) (F)F
24	CHEMBL514	Lomustine (BAN, FDA, INN, USAN);	ClCCN (N=O)C (=O)NC1CCCCC1
25	CHEMBL178	Daunorubicin (BAN, INN); Daunorubicin Citrate (FDA); Daunorubicin HCl (FDA, JAN, USAN, USP);	COc1cccc2C (=O)c3c (O)c4C[C@] (O) (C[C@H] (O[C@H]5C[C@H] (N)[C@H] (O)[C@H] (C)O5)c4c (O)c3C (=O)c12)C (=O)C

26	CHEMBL1750	Clofarabine (BAN, FDA, INN, USAN);	Nc1nc (Cl)nc2c1ncn2[C@@H]3O[C@H] (CO)[C@@H] (O)[C@@H]3F
27	CHEMBL2105717	Cabozantinib (INN, USAN); Cabozantinib S-Malate (FDA, USAN);	COc1cc2nccc (Oc3ccc (NC (=O)C4 (CC4)C (=O)Nc5ccc (F)cc5)cc3)c2cc1OC
28	CHEMBL1554	Dactinomycin (BAN, FDA, INN, USAN, USP); Actinomycin D (JAN);	CC (C)[C@H]1NC (=O)[C@@H] (NC (=O)C2=C (N)C (=O)C (=C3Oc4c (C)ccc (C (=O)N[C@H]5[C@@H] (C)OC (=O)[C@H] (C (C)C)N (C)C (=O)CN (C)C (=O)[C@@H]6CCCN6C (=O)[C@H] (NC5=O)C (C)C)c4N=C23)C)[C@@H] (C)OC (=O)[C@H] (C (C)C)N (C)C (=O)CN (C)C (=O)[C@@H]7CCCN7C1=O
29	CHEMBL1743062	Ramucirumab (FDA, INN, USAN);	
30	CHEMBL803	Cytarabine (BAN, FDA, INN, JAN, USAN, USP); Cytarabine HCl (USAN);	NC1=NC (=O)N (C=C1)[C@@H]2O[C@H] (CO)[C@@H] (O)[C@@H]2O
31	CHEMBL88	Cyclophosphamide (BAN, JAN, USP, BAN, FDA, INN, JAN, USP);	ClCCN (CCCl)P1 (=O)NCCCO1
32	CHEMBL476	Dacarbazine (BAN, FDA, INN, JAN, USAN, USP);	CN (C)N=Nc1[nH]cnc1C (=O)N
33	CHEMBL1201129	Decitabine (BAN, FDA, INN, USAN);	NC1=NC (=O)N (C=N1)[C@H]2C[C@H] (O)[C@@H] (CO)O2
35	CHEMBL1201302	Dexamethasone Sodium Phosphate (BAN, FDA, JAN, USP);	C[C@@H]1C[C@H]2[C@@H]3CCC4=CC (=O)C=C[C@]4 (C)[C@@]3 (F)[C@@H] (O)C[C@]2 (C)[C@@]1 (O)C (=O)COP (=O) (O)O
39	CHEMBL92	Docetaxel (BAN, FDA, INN, USAN);	CC (=O)O[C@@]12CO[C@@H]1C[C@H] (O)[C@]3 (C)[C@@H]2[C@H] (OC (=O)c4ccccc4)[C@]5 (O)C[C@H] (OC (=O)[C@H] (O)[C@@H] (NC (=O)OC (C) (C)C)c6ccccc6)C (=C ([C@@H] (O)C3=O)C5 (C)C)C
40	CHEMBL53463	Doxorubicin (BAN, INN, USAN); Doxorubicin HCl (FDA, JAN, USP);	COc1cccc2C (=O)c3c (O)c4C[C@] (O) (C[C@H] (O[C@H]5C[C@H] (N)[C@H] (O)[C@H] (C)O5)c4c (O)c3C (=O)c12)C (=O)CO
41	CHEMBL467	Hydroxycarbamide (INN); Hydroxyurea	NC (=O)NO

		(BAN, FDA, USAN, USP);	
42	CHEMBL1201199	Leuprolide Acetate (FDA, USAN); Leuprorelin (BAN, INN);	CCNC (=O)[C@@H]1CCCN1C (=O)[C@H] (CCCNC (=N)N)NC (=O)[C@H] (CC (C)C)NC (=O)[C@@H] (CC (C)C)NC (=O)[C@H] (Cc2ccc (O)cc2)NC (=O)[C@H] (CO)NC (=O)[C@H] (Cc3c[nH]c4ccccc34)NC (=O)[C@H] (Cc5c[nH]cn5)NC (=O)[C@@H]6CCC (=O)N6
43	CHEMBL417	Epirubicin (BAN, INN); Epirubicin HCl (FDA, JAN, USAN);	COc1cccc2C (=O)c3c (O)c4C[C@] (O) (C[C@H] (O[C@H]5C[C@H] (N)[C@@H] (O)[C@H] (C)O5)c4c (O)c3C (=O)c12)C (=O)CO
44	CHEMBL414804	Oxaliplatin (BAN, FDA, INN, USAN);	
45	CHEMBL2108989	Asparaginase (FDA, USAN); Colaspase (BAN); L-Asparaginase (JAN);	
46	CHEMBL1575	Estramustine (BAN, INN, USAN);	C[C@]12CC[C@H]3[C@@H] (CCc4cc (OC (=O)N (CCCl)CCCl)ccc34)[C@@H]1CC[C@@H]2O
47	CHEMBL1201577	Cetuximab (FDA, INN, USAN);	
48	CHEMBL473417	Vismodegib (FDA, INN, USAN);	CS (=O) (=O)c1ccc (C (=O)Nc2ccc (Cl)c (c2)c3ccccn3)c (Cl)c1
49	CHEMBL1863514	Asparaginase Erwinia Chrysanthemi (FDA, USAN);	
50	CHEMBL1006	Amifostine (BAN, FDA, INN, USAN, USP);	NCCCNCCSP (=O) (O)O
51	CHEMBL44657	Etoposide (BAN, FDA, INN, JAN, USAN, USP);	COc1cc (cc (OC)c1O)[C@H]2[C@@H]3[C@H] (COC3=O)[C@H] (O[C@@H]4O[C@@H]5CO[C@@H] (C)O[C@H]5[C@H] (O)[C@H]4O)c6cc7OCOc7cc26
52	CHEMBL806	Flutamide (BAN, FDA, INN, USAN, USP);	CC (C)C (=O)Nc1ccc (c (c1)C (F) (F)F)[N+] (=O)[O-]
53	CHEMBL917	Floxuridine (FDA, INN, USAN, USP);	OC[C@H]1O[C@H] (C[C@@H]1O)N2C=C (F)C (=O)NC2=O
54	CHEMBL1655	Toremifene Citrate (FDA, USAN); Toremifene (BAN, INN);	CN (C)CCOc1ccc (cc1)\C (=C (\CCCl)/c2ccccc2)\c3ccccc3

55	CHEMBL1358	Fulvestrant (BAN, FDA, INN, USAN);	C[C@]12CC[C@H]3[C@@H] ([C@H] (CCCCCCCCC[S+] ([O-])CCCC (F) (F)C (F) (F)F)Cc4cc (O)ccc34)[C@@H]1CC[C@@H]2O
56	CHEMBL1444	Letrozole (BAN, FDA, INN, USAN, USP);	N#Cc1ccc (cc1)C (c2ccc (cc2)C#N)n3cncn3
57	CHEMBL415606	Degarelix (INN, USAN); Degarelix Acetate (FDA, USAN);	CC (C)C[C@H] (NC (=O)[C@@H] (Cc1ccc (NC (=O)N)cc1)NC (=O)[C@H] (Cc2ccc (NC (=O)[C@@H]3CC (=O)NC (=O)N3)cc2)NC (=O)[C@H] (CO)NC (=O)[C@@H] (Cc4cccnc4)NC (=O)[C@@H] (Cc5ccc (Cl)cc5)NC (=O)[C@@H] (Cc6ccc7ccccc7c6)NC (=O)C)C (=O)N[C@@H] (CCCCNC (C)C)C (=O)N8CCC[C@H]8C (=O)N[C@H] (C)C (=O)N
58	CHEMBL1096882	Fludarabine (INN); Fludarabine Phosphate (BAN, FDA, USAN, USP);	Nc1nc (F)nc2c1ncn2[C@@H]3O[C@H] (COP (=O) (O)O)[C@@H] (O)[C@@H]3O
60	CHEMBL1201746	Pralatrexate (FDA, INN, USAN);	Nc1nc (N)c2nc (CC (CC#C)c3ccc (cc3)C (=O)N[C@@H] (CCC (=O)O)C (=O)O)cnc2n1
61	CHEMBL1743048	Obinutuzumab (INN, USAN);	
62	CHEMBL888	Gemcitabine HCl (FDA, USAN, USP); Gemcitabine (BAN, INN, USAN);	NC1=NC (=O)N (C=C1)[C@@H]2O[C@H] (CO)[C@@H] (O)C2 (F)F
63	CHEMBL1173655	Afatinib Dimaleate (FDA, USAN); Afatinib (INN, USAN);	CN (C)C\C=C\C (=O)Nc1cc2c (Nc3ccc (F)c (Cl)c3)ncnc2cc1O[C@H]4CCOC4
64	CHEMBL941	Imatinib mesylate (FDA); Imatinib (BAN, INN);	CN1CCN (Cc2ccc (cc2)C (=O)Nc3ccc (C)c (Nc4nccc (n4)c5cccnc5)c3)CC1
66	CHEMBL1683590	Eribulin mesylate (FDA, USAN); Eribulin (INN);	CO[C@H]1[C@@H] (C[C@H] (O)CN)O[C@H]2C[C@H]3O[C@@H] (CC[C@@H]4O[C@@H] (CC[C@@]56C[C@H]7O[C@@H]8[C@@H] (O[C@H]9CC[C@H] (CC (=O)C[C@H]12)O[C@@H]9[C@@H]8O5)[C@H] 7O6)CC4=C)C[C@@H] (C)C3=C
67	CHEMBL1201585	Trastuzumab (BAN, FDA, INN);	
68	CHEMBL1455	Altretamine (BAN, FDA, INN, USAN, USP);	CN (C)c1nc (nc (n1)N (C)C)N (C)C
69	CHEMBL84	Topotecan (BAN, INN);	CC[C@@]1 (O)C (=O)OCC2=C1C=C3N (Cc4cc5c (CN (C)C)c (O)ccc5nc34)C2=O

		Topotecan HCl (FDA, USAN);	
71	CHEMBL1171837	Ponatinib HCl (FDA, USAN); Ponatinib (INN, USAN);	CN1CCN (Cc2ccc (NC (=O)c3ccc (C)c (c3)C#Cc4cnc5cccnn45)cc2C (F) (F)F)CC1
72	CHEMBL1117	Idarubicin (BAN, INN); Idarubicin HCl (FDA, INN, USAN, USP);	C[C@@H]1O[C@H] (C[C@H] (N)[C@@H]1O)O[C@H]2C[C@@] (O) (Cc3c (O)c4C (=O)c5ccccc5C (=O)c4c (O)c23)C (=O)C
73	CHEMBL1024	Ifosfamide (BAN, FDA, INN, JAN, USAN, USP);	ClCCNP1 (=O)OCCCN1CCCl
74	CHEMBL1873475	Ibrutinib (FDA, INN, USAN);	Nc1ncnc2c1c (nn2[C@@H]3CCCN (C3)C (=O)C=C)c4ccc (Oc5ccccc5)cc4
75	CHEMBL1289926	Axitinib (FDA, INN, USAN);	CNC (=O)c1ccccc1Sc2ccc3c (\C=C\c4ccccn4)n[nH]c3c2
76	CHEMBL1201561	Peginterferon alfa-2b (BAN, FDA, INN);	
77	CHEMBL939	Gefitinib (BAN, FDA, INN, USAN);	COc1cc2ncnc (Nc3ccc (F)c (Cl)c3)c2cc1OCCCN4CCOCC4
78	CHEMBL1213490	Romidepsin (FDA, INN, USAN);	C\C=C\1/NC (=O)[C@@H] (CS)NC (=O)[C@H] (CC (=O)C[C@H] (OC (=O)[C@@H] (NC1=O)C (C)C)\C=C\CCS)C (C)C
79	CHEMBL1201752	Ixabepilone (FDA, INN, USAN);	C[C@H]1CCC[C@@]2 (C)O[C@H]2C[C@H] (NC (=O)C[C@H] (O)C (C) (C)C (=O)[C@H] (C)[C@H]1O)\C (=C\c3csc (C)n3)\C
80	CHEMBL1789941	Ruxolitinib Phosphate (FDA, USAN); Ruxolitinib (INN, USAN);	N#CC[C@H] (C1CCCC1)n2cc (cn2)c3ncnc4[nH]ccc34
81	CHEMBL1201748	Cabazitaxel (FDA, INN, USAN);	CO[C@H]1C[C@H]2OC[C@@]2 (OC (=O)C)[C@H]3[C@H] (OC (=O)c4ccccc4)[C@]5 (O)C[C@H] (OC (=O)[C@H] (O)[C@@H] (NC (=O)OC (C) (C)C)c6ccccc6)C (=C ([C@@H] (OC)C (=O)[C@]13C)C5 (C)C)C
82	CHEMBL1743082	Ado-Trastuzumab Emtansine (FDA); Trastuzumab Emtansine (INN, USAN);	
83	CHEMBL451887	Carfilzomib (FDA, INN, USAN);	CC (C)C[C@H] (NC (=O)[C@H] (CCc1ccccc1)NC (=O)CN2CCOCC2)C (=O)N[C@@H] (Cc3ccccc3)C (=O)N[C@@H] (CC (C)C)C (=O)[C@@]4 (C)CO4
84	CHEMBL515	Chlorambucil (BAN, FDA, INN, USP);	OC (=O)CCCc1ccc (cc1)N (CCCl)CCCl
85	CHEMBL1201670	Sargramostim (BAN, FDA,	

		INN, USAN, USP);	
86	CHEMBL1619	Cladribine (BAN, FDA, INN, USAN);	Nc1nc (Cl)nc2c1ncn2[C@H]3C[C@H] (O)[C@@H] (CO)O3
88	CHEMBL1670	Mitotane (FDA, INN, JAN, USAN, USP);	ClC (Cl)C (c1ccc (Cl)cc1)c2ccccc2Cl
89	CHEMBL90555	Vincristine Sulfate (FDA, JAN, USAN, USP); Vincristine (BAN, INN);	CC[C@]1 (O)C[C@H]2CN (CCc3c ([nH]c4ccccc34)[C@@] (C2) (C (=O)OC)c5cc6c (cc5OC)N (C=O)[C@H]7[C@] (O) ([C@H] (OC (=O)C)[C@]8 (CC)C=CCN9CC[C@]67[C@H]89)C (=O)OC)C1
90	CHEMBL1321	Procarbazine (BAN, INN); Procarbazine HCl (FDA, JAN, USAN, USP);	CNNCc1ccc (cc1)C (=O)NC (C)C
91	CHEMBL1201139	Megestrol (BAN, INN); Megestrol Acetate (FDA, USAN, USP);	CC (=O)O[C@@]1 (CC[C@H]2[C@@H]3C=C (C)C4=CC (=O)CC[C@]4 (C)[C@H]3CC[C@]12C)C (=O)C
92	CHEMBL2103875	Trametinib (INN, USAN); Trametinib Dimethyl Sulfoxide (FDA, USAN);	CN1C (=O)C (=C2N (C (=O)N (C3CC3)C (=O)C2=C1Nc4ccc (I)cc4F)c5cccc (NC (=O)C)c5)C
96	CHEMBL427	Chlormethine (BAN, INN); Nitrogen Mustard N-Oxide HCl (JAN); Mechlorethamine HCl (FDA, USP);	CN (CCCl)CCCl
97	CHEMBL105	Mitomycin (BAN, FDA, INN, USAN, USP); Mitomycin C (JAN);	CO[C@]12[C@H]3N[C@H]3CN1C4=C ([C@H]2COC (=O)N)C (=O)C (=C (C)C4=O)N
98	CHEMBL820	Busulfan (BAN, FDA, INN, JAN, USP);	CS (=O) (=O)OCCCCOS (=O) (=O)C
99	CHEMBL1201506	Gemtuzumab Ozogamicin (FDA, INN, USAN);	
100	CHEMBL553025	Vinorelbine (BAN, INN); Vinorelbine Tartrate (FDA, USAN, USP);	CCC1=C[C@@H]2CN (C1)Cc3c ([nH]c4ccccc34)[C@@] (C2) (C (=O)OC)c5cc6c (cc5OC)N (C)[C@H]7[C@] (O) ([C@H] (OC (=O)C)[C@]8 (CC)C=CCN9CC[C@]67[C@H]89)C (=O)OC

102	CHEMBL1201568	Pegfilgrastim (BAN, FDA, INN, USAN);	
103	CHEMBL1201567	Tbo-Filgrastim (FDA); Filgrastim (BAN, FDA, INN, USAN); Filgrastim-sndz (FDA);	
104	CHEMBL1336	Sorafenib (INN, USAN); Sorafenib Tosylate (FDA, USAN);	CNC (=O)c1cc (Oc2ccc (NC (=O)Nc3ccc (Cl)c (c3)C (F) (F)F)cc2)ccn1
105	CHEMBL1274	Nilutamide (BAN, FDA, INN, MI, USAN);	CC1 (C)NC (=O)N (C1=O)c2ccc (c (c2)C (F) (F)F)[N+] (=O)[O-]
106	CHEMBL1580	Pentostatin (BAN, FDA, INN, JAN, USAN);	OC[C@H]1O[C@H] (C[C@@H]1O)n2cnc3[C@H] (O)CNC=Nc23
107	CHEMBL83	Tamoxifen (BAN, INN); Tamoxifen citrate (FDA, JAN, USAN, USP);	CC\C (=C (/c1ccccc1)\c2ccc (OCCN (C)C)cc2)\c3ccccc3
108	CHEMBL58	Mitoxantrone (INN); Mitoxantrone HCl (FDA, JAN, USAN, USP); Mitozantrone (BAN);	OCCNCCNc1ccc (NCCNCCO)c2C (=O)c3c (O)ccc (O)c3C (=O)c12
109	CHEMBL2108546	Pegaspargase (FDA, INN, USAN);	
111	CHEMBL1201550	Denileukin diftitox (BAN, FDA, INN, USAN);	
113	CHEMBL705	Alitretinoin (BAN, FDA, INN, USAN);	C\C (=C\C=C\C (=C\C (=O)O)\C)\C=C\C1=C (C)CCCC1 (C)C
115	CHEMBL2007641	Pertuzumab (BAN, FDA, INN, USAN);	
116	CHEMBL11359	Cisplatin (BAN, FDA, INN, JAN, USAN, USP);	
118	CHEMBL43452	Pomalidomide (FDA, INN, USAN);	Nc1cccc2C (=O)N (C3CCC (=O)NC3=O)C (=O)c12

119	CHEMBL635	Prednisone (BAN, FDA, INN, USP);	C[C@]12CC (=O)[C@H]3[C@@H] (CCC4=CC (=O)C=C[C@]34C)[C@@H]1CC[C@]2 (O)C (=O)CO
120	CHEMBL1201438	Aldesleukin (BAN, FDA, INN, USAN);	
121	CHEMBL1425	Mercaptopurine (BAN, JAN, USP, BAN, FDA, INN, JAN, USP);	Sc1ncnc2nc[nH]c12
122	CHEMBL924	Zoledronate Trisodium (USAN); Zoledronate Disodium (USAN); Zoledronic acid (BAN, FDA, INN, USAN);	OC (Cn1ccnc1) (P (=O) (O)O)P (=O) (O)O
123	CHEMBL848	Lenalidomide (BAN, FDA, INN, USAN);	Nc1cccc2C (=O)N (Cc12)C3CCC (=O)NC3=O
125	CHEMBL1201576	Rituximab (BAN, FDA, INN, USAN);	
126	CHEMBL2108508	Interferon Alfa-2A (BAN, FDA, INN, USAN); Interferon Alfa-2A (Genetical Recombination) (JAN);	
128	CHEMBL1680	Octreotide Pamoate (USAN); Octreotide (BAN, INN, USAN); Octreotide hydrochloride (USAN); Octreotide Acetate (FDA, JAN, USAN);	C[C@@H] (O)[C@@H] (CO)NC (=O)[C@@H]1CSSC[C@H] (NC (=O)[C@H] (N)Cc2ccccc2)C (=O)N[C@@H] (Cc3ccccc3)C (=O)N[C@H] (Cc4c[nH]c5ccccc45)C (=O)N[C@@H] (CCCCN)C (=O)N[C@@H] ([C@@H] (C)O)C (=O)N1
130	CHEMBL1421	Dasatinib (FDA, INN, USAN);	Cc1nc (Nc2ncc (s2)C (=O)Nc3c (C)cccc3Cl)cc (n1)N4CCN (CCO)CC4
132	CHEMBL1946170	Regorafenib (FDA, INN, USAN);	CNC (=O)c1cc (Oc2ccc (NC (=O)Nc3ccc (Cl)c (c3)C (F) (F)F)c (F)c2)ccn1
133	CHEMBL1201255	Histrelin Acetate (FDA); Histrelin (INN, USAN);	CCNC (=O)[C@@H]1CCCN1C (=O)[C@H] (CCCNC (=N)N)NC (=O)[C@H] (CC (C)C)NC (=O)[C@@H] (Cc2cn (Cc3ccccc3)cn2)NC (=O)[C@H] (Cc4ccc (O)cc4)NC (=O)[C@H] (CO)NC (=O)[C@H] (Cc5c[nH]c6ccccc56)NC (=O)[C@H] (Cc7cnc[nH]7)NC (=O)[C@@H]8CCC (=O)N8

134	CHEMBL535	Sunitinib (INN); Sunitinib Malate (FDA);	CCN (CC)CCNC (=O)c1c (C)[nH]c (\C=C\2/C (=O)Nc3ccc (F)cc23)c1C
136	CHEMBL1743070	Siltuximab (FDA, INN, USAN);	
137	CHEMBL46286	Omacetaxine Mepesuccinate (FDA, INN, USAN);	COC (=O)C[C@] (O) (CCCC (C) (C)O)C (=O)O[C@H]1[C@H]2c3cc4OCOc4cc3CCN5CCC [C@]25C=C1OC
139	CHEMBL727	Thioguanine (FDA, USAN, USP); Tioguanine (BAN, INN);	NC1=Nc2[nH]cnc2C (=S)N1
140	CHEMBL2028663	Dabrafenib Mesylate (FDA, USAN); Dabrafenib (INN, USAN);	CC (C) (C)c1nc (c2cccc (NS (=O) (=O)c3c (F)cccc3F)c2F)c (s1)c4ccnc (N)n4
141	CHEMBL553	Erlotinib HCl (FDA, INN, USAN); Erlotinib (INN);	COCCOc1cc2ncnc (Nc3cccc (c3)C#C)c2cc1OCCOC
142	CHEMBL1023	Bexarotene (BAN, FDA, INN, USAN);	Cc1cc2c (cc1C (=C)c3ccc (cc3)C (=O)O)C (C) (C)CCC2 (C)C
143	CHEMBL255863	Nilotinib (INN, USAN); Nilotinib Hydrochloride Monohydrate (FDA);	Cc1cn (cn1)c2cc (NC (=O)c3ccc (C)c (Nc4nccc (n4)c5cccnc5)c3)cc (c2)C (F) (F)F
146	CHEMBL810	Temozolomide (BAN, FDA, INN, USAN);	CN1N=Nc2c (ncn2C1=O)C (=O)N
147	CHEMBL671	Thiotepa (BAN, FDA, INN, JAN, USP);	S=P (N1CC1) (N2CC2)N3CC3
148	CHEMBL468	Thalidomide (BAN, FDA, INN, USAN, USP);	O=C1CCC (N2C (=O)c3ccccc3C2=O)C (=O)N1
152	CHEMBL1201182	Temsirolimus (FDA, INN, USAN);	CO[C@@H]1C[C@H] (C[C@@H] (C)[C@@H]2CC (=O)[C@H] (C)\C=C (/C)\[C@@H] (O)[C@@H] (OC)C (=O)[C@H] (C)C[C@H] (C)\C=C\C=C\C=C (/C)\[C@H] (C[C@@H]3CC[C@@H] (C)[C@@] (O) (O3)C (=O)C (=O)N4CCCC[C@H]4C (=O)O2)OC)CC[C@H]1OC (=O)C (C) (CO)CO
153	CHEMBL487253	Bendamustine HCl (FDA, USAN); Bendamustine (INN);	Cn1c (CCCC (=O)O)nc2cc (ccc12)N (CCCl)CCCl
154	CHEMBL1201334	Triptorelin (BAN, INN, USAN);	CC (C)C[C@H] (NC (=O)[C@@H] (Cc1c[nH]c2ccccc12)NC (=O)[C@H] (Cc3ccc (O)cc3)NC (=O)[C@H] (CO)NC (=O)[C@H]

		Triptorelin Pamoate (FDA, USAN);	(Cc4c[nH]c5ccccc45)NC (=O)[C@H] (Cc6c[nH]cn6)NC (=O)[C@@H]7CCC (=O)N7)C (=O)N[C@@H] (CCCNC (=N)N)C (=O)N8CCC[C@H]8C (=O)NCC (=O)N
156	CHEMBL462019	Trofosfamide (INN, MI);	ClCCN (CCCl)P1 (=O)OCCCN1CCCl
157	CHEMBL1200978	Arsenic Trioxide (FDA, JAN, USAN);	O=[As]O[As]=O
158	CHEMBL554	Lapatinib (INN); Lapatinib Ditosylate (FDA, USAN);	CS (=O) (=O)CCNCc1oc (cc1)c2ccc3ncnc (Nc4ccc (OCc5cccc (F)c5)c (Cl)c4)c3c2
159	CHEMBL1096885	Valrubicin (FDA, INN, USAN, USP);	CCCCC (=O)OCC (=O)[C@@]1 (O)C[C@H] (O[C@H]2C[C@H] (NC (=O)C (F) (F)F)[C@H] (O)[C@H] (C)O2)c3c (O)c4C (=O)c5c (OC)cccc5C (=O)c4c (O)c3C1
161	CHEMBL1201827	Panitumumab (FDA, INN, USAN);	
162	CHEMBL159	Vinblastine Sulfate (FDA, JAN, USAN, USP); Vinblastine (BAN, INN);	CC[C@]1 (O)C[C@@H]2CN (CCc3c ([nH]c4ccccc34)[C@@] (C2) (C (=O)OC)c5cc6c (cc5OC)N (C)[C@H]7[C@] (O) ([C@H] (OC (=O)C)[C@]8 (CC)C=CCN9CC[C@]67[C@H]89)C (=O)OC)C1
163	CHEMBL325041	Bortezomib (BAN, FDA, INN, USAN);	CC (C)C[C@H] (NC (=O)[C@H] (Cc1ccccc1)NC (=O)c2cnccn2)B (O)O
165	CHEMBL38	Tretinoin (BAN, FDA, INN, USAN, USP);	C\C (=C/C=C/C (=C/C (=O)O)/C)\C=C\C1=C (C)CCCC1 (C)C
166	CHEMBL1489	Azacitidine (FDA, INN, USAN);	NC1=NC (=O)N (C=N1)[C@@H]2O[C@H] (CO)[C@@H] (O)[C@H]2O
169	CHEMBL477772	Pazopanib HCl (FDA, USAN); Pazopanib (INN);	CN (c1ccc2c (C)n (C)nc2c1)c3ccnc (Nc4ccc (C)c (c4)S (=O) (=O)N)n3
170	CHEMBL452231	Teniposide (BAN, FDA, INN, USAN);	COc1cc (cc (OC)c1O)[C@H]2[C@@H]3[C@H] (COC3=O)[C@H] (O[C@@H]4O[C@@H]5CO[C@H] (O[C@H]5[C@H] (O)[C@H]4O)c6cccs6)c7cc8OCOc8cc27
171	CHEMBL1908841	Levoleucovorin Calcium (FDA, USAN); Calcium Levofolinate (BAN, INN);	NC1=NC (=O)C2=C (NC[C@H] (CNc3ccc (cc3)C (=O)N[C@@H] (CCC (=O)O)C (=O)O)N2C=O)N1
172	CHEMBL601719	Crizotinib (FDA, INN, USAN);	C[C@@H] (Oc1cc (cnc1N)c2cnn (c2)C3CCNCC3)c4c (Cl)ccc (F)c4Cl
173	CHEMBL1773	Capecitabine (BAN, FDA, INN, USAN);	CCCCCOC (=O)NC1=NC (=O)N (C=C1F)[C@@H]2O[C@H] (C)[C@@H] (O)[C@H]2O

174	CHEMBL1082407	Enzalutamide (FDA, INN, USAN);	CNC (=O)c1ccc (cc1F)N2C (=S)N (C (=O)C2 (C)C)c3ccc (C#N)c (c3)C (F) (F)F
175	CHEMBL1789844	Ipilimumab (FDA, INN, USAN);	
176	CHEMBL1742982	Ziv-Aflibercept (FDA); Aflibercept (FDA, INN, USAN);	
177	CHEMBL1651906	Streptozocin (FDA, INN, USAN);	CN (N=O)C (=O)N[C@H]1C (O)O[C@H] (CO)[C@@H] (O)[C@@H]1O
178	CHEMBL1229517	Vemurafenib (FDA, INN, USAN);	CCCS (=O) (=O)Nc1ccc (F)c (C (=O)c2c[nH]c3ncc (cc23)c4ccc (Cl)cc4)c1F
179	CHEMBL1201606	Ibritumomab tiuxetan (BAN, INN, USAN);	
180	CHEMBL1201247	Goserelin (BAN, INN, USAN); Goserelin Acetate (FDA, JAN, JAN);	CC (C)C[C@H] (NC (=O)[C@@H] (COC (C) (C)C)NC (=O)[C@H] (Cc1ccc (O)cc1)NC (=O)[C@H] (CO)NC (=O)[C@H] (Cc2c[nH]c3ccccc23)NC (=O)[C@H] (Cc4cnc[nH]4)NC (=O)[C@@H]5CCC (=O)N5)C (=O)N[C@@H] (CCCNC (=N)N)C (=O)N6CCC[C@H]6C (=O)NNC (=O)N
181	CHEMBL98	Vorinostat (FDA, INN, USAN);	ONC (=O)CCCCCCC (=O)Nc1ccccc1
184	CHEMBL2403108	Ceritinib (FDA, INN, USAN);	CC (C)Oc1cc (C2CCNCC2)c (C)cc1Nc3ncc (Cl)c (Nc4ccccc4S (=O) (=O)C (C)C)n3
185	CHEMBL254328	Abiraterone (BAN, INN);	C[C@]12CC[C@H]3[C@@H] (CC=C4C[C@@H] (O)CC[C@]34C)[C@@H]1CC=C2c5cccnc5

Supplementary Table 2 The predicted and experimental targets for both the AfroCancer and NCI Cancer datasets. Targets in bold are shared between both datasets

AfroCancer predicted target	NCI Cancer experimental target
Bile acid receptor	**Bile acid receptor FXR**
Carbonic anhydrase I	**Carbonic anhydrase I**
Carbonic anhydrase II	**Carbonic anhydrase II**
Carbonic anhydrase VII	**Carbonic anhydrase VII**
Carbonic anhydrase XII	**Carbonic anhydrase XII**
DNA topoisomerase I	**DNA topoisomerase I**
DNA topoisomerase II alpha	**DNA topoisomerase II alpha**
Estrogen receptor alpha	**Estrogen receptor alpha**
Estrogen receptor beta	**Estrogen receptor beta**
Glucocorticoid receptor	**Glucocorticoid receptor**
Receptor-type tyrosine-protein kinase FLT3	**Receptor-type tyrosine-protein kinase FLT3**

Retinoic acid receptor beta	**Retinoic acid receptor beta**
Retinoic acid receptor gamma	**Retinoic acid receptor gamma**
Steroid 17-alpha-hydroxylase/17,20 lyase	**Steroid 17-alpha-hydroxylase/17,20 lyase**
1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase gamma-1	Acetylcholinesterase
15-hydroxyprostaglandin dehydrogenase [NAD	Adenosine A2a receptor
3-hydroxyacyl-CoA dehydrogenase type-2	Adenosine deaminase
3-oxo-5-alpha-steroid 4-dehydrogenase 1	ALK tyrosine kinase receptor
3-oxo-5-alpha-steroid 4-dehydrogenase 2	Androgen Receptor
5'-AMP-activated protein kinase catalytic subunit alpha-2	Beta-1 adrenergic receptor
7-dehydrocholesterol reductase	Beta-3 adrenergic receptor
Acid ceramidase	Carbonic anhydrase IV
Aldehyde dehydrogenase, mitochondrial	Cholecystokinin A receptor
Aldo-keto reductase family 1 member B10	Cytochrome P450 11B1
Aldose reductase	Cytochrome P450 19A1
Alkaline phosphatase, placental-like	Dihydrofolate reductase
Alkaline phosphatase, tissue-nonspecific isozyme	Discoidin domain-containing receptor 2
Amine oxidase [flavin-containing] A	Dopamine D1 receptor
Amyloid beta A4 protein	Dopamine D2 receptor
Apoptosis regulator Bcl-2	Dopamine D3 receptor
Arachidonate 12-lipoxygenase, 12S-type	Dopamine transporter
Arachidonate 15-lipoxygenase	Dual specificity mitogen-activated protein kinase kinase 1
Arachidonate 5-lipoxygenase	Dual specificity mitogen-activated protein kinase kinase 2
Aryl hydrocarbon receptor	Ephrin type-A receptor 2
ATP-binding cassette sub-family G member 2	Epidermal growth factor receptor erbB1
ATP-dependent DNA helicase Q1	Farnesyl diphosphate synthase
ATPase family AAA domain-containing protein 5	Fibroblast growth factor receptor 1
Beta-glucuronidase	Fibroblast growth factor receptor 2
Bloom syndrome protein	Fibroblast growth factor receptor 3
Carbonic anhydrase 13	Glutathione reductase
Carbonic anhydrase 14	Gonadotropin-releasing hormone receptor
Carbonic anhydrase 3	Growth hormone-releasing hormone receptor
Carbonic anhydrase 5A, mitochondrial	Hepatocyte growth factor receptor
Carbonic anhydrase 5B, mitochondrial	HERG
Carbonic anhydrase 6	Histamine H2 receptor
Carbonic anhydrase 9	Histone deacetylase 1
Carbonyl reductase [NADPH] 1	Histone deacetylase 2
Casein kinase II subunit beta	Histone deacetylase 3
Catechol O-methyltransferase	Histone deacetylase 6
Cocaine esterase	Insulin receptor
Corticosteroid 11-beta-dehydrogenase isozyme 2	Insulin-like growth factor I receptor
Cyclin-dependent kinase 14	Kappa opioid receptor

Cyclin-dependent kinase 5 activator 1	Macrophage colony stimulating factor receptor
Cyclin-dependent kinase 6	MAP kinase p38 beta
Cytochrome P450 1A1	Matrix metalloproteinase-1
Cytochrome P450 1B1	Mineralocorticoid receptor
Cytosolic phospholipase A2	Mu opioid receptor
D-amino-acid oxidase	Muscarinic acetylcholine receptor M1
DNA polymerase beta	Muscarinic acetylcholine receptor M2
DNA polymerase eta	Muscarinic acetylcholine receptor M3
DNA polymerase iota	Neurokinin 1 receptor
DNA polymerase kappa	Norepinephrine transporter
DNA- (apurinic or apyrimidinic site) lyase	Platelet-derived growth factor receptor beta
Dual specificity protein kinase CLK1	Progesterone receptor
Dual specificity protein kinase CLK3	Receptor protein-tyrosine kinase erbB-2
Dual specificity tyrosine-phosphorylation-regulated kinase 2	Receptor protein-tyrosine kinase erbB-4
Estradiol 17-beta-dehydrogenase 1	Serine/threonine-protein kinase B-raf
Estradiol 17-beta-dehydrogenase 2	Serine/threonine-protein kinase RAF
Flap endonuclease 1	Serotonin 1a (5-HT1a) receptor
G-protein coupled bile acid receptor 1	Serotonin 2a (5-HT2a) receptor
G-protein coupled receptor 35	Serotonin 2b (5-HT2b) receptor
G2/mitotic-specific cyclin-B2	Serotonin 2c (5-HT2c) receptor
G2/mitotic-specific cyclin-B3	Serotonin transporter
Galactokinase	Sigma opioid receptor
Gamma-aminobutyric acid receptor subunit beta-1	Smoothened homolog
Glutaminase kidney isoform, mitochondrial	Somatostatin receptor 1
Heat shock 70 kDa protein 1A	Somatostatin receptor 2
Heat shock protein beta-1	Somatostatin receptor 3
Heat shock protein HSP 90-alpha	Somatostatin receptor 5
Heat shock protein HSP 90-beta	Stem cell growth factor receptor
Hydroxycarboxylic acid receptor 2	Thymidylate synthase
Induced myeloid leukemia cell differentiation protein Mcl-1	Thyroid stimulating hormone receptor
Interleukin-2	Tyrosine-protein kinase ABL
Intestinal-type alkaline phosphatase	Tyrosine-protein kinase BRK
L-lactate dehydrogenase A chain	Tyrosine-protein kinase BTK
Lactoylglutathione lyase	Tyrosine-protein kinase FRK
Lethal (3)malignant brain tumor-like protein 1	Tyrosine-protein kinase HCK
Lysine-specific demethylase 4E	Tyrosine-protein kinase ITK/TSK
Lysosomal alpha-glucosidase	Tyrosine-protein kinase JAK1
Macrophage migration inhibitory factor	Tyrosine-protein kinase JAK2
Major prion protein	Tyrosine-protein kinase LCK
Maltase-glucoamylase, intestinal	Tyrosine-protein kinase Lyn
Microtubule-associated protein tau	Tyrosine-protein kinase receptor RET
Mitogen-activated protein kinase 3	Tyrosine-protein kinase SRC

Multidrug resistance protein 1	Tyrosine-protein kinase TIE-2
Multidrug resistance-associated protein 1	Vascular endothelial growth factor receptor 2
NAD (P)H dehydrogenase [quinone] 1	
NADPH oxidase 4	
NF-kappa-B inhibitor alpha	
Nuclear factor erythroid 2-related factor 2	
Oxoeicosanoid receptor 1	
P-selectin	
P2X purinoceptor 1	
Peroxisome proliferator-activated receptor gamma	
PH domain leucine-rich repeat-containing protein phosphatase 2	
Poly (ADP-ribose) glycohydrolase	
Potassium voltage-gated channel subfamily A member 3	
Prostaglandin G/H synthase 1	
Receptor tyrosine-protein kinase erbB-3	
Retinoic acid receptor alpha	
Ribosyldihydronicotinamide dehydrogenase [quinone]	
Sex hormone-binding globulin	
Sodium/glucose cotransporter 1	
Sodium/glucose cotransporter 2	
Solute carrier family 22 member 3	
Squalene monooxygenase	
Steroid hormone receptor ERR1	
Steroid hormone receptor ERR2	
Tankyrase-1	
Tankyrase-2	
Telomerase reverse transcriptase	
Testosterone 17-beta-dehydrogenase 3	
Thioredoxin reductase 2, mitochondrial	
Thyroid hormone receptor beta	
Transcription factor p65	
Transthyretin	
Troponin C, slow skeletal and cardiac muscles	
Tubulin alpha-1A chain	
Tubulin alpha-1B chain	
Tubulin alpha-1C chain	
Tubulin alpha-3C/D chain	
Tubulin alpha-3E chain	
Tubulin alpha-4A chain	
Tubulin beta chain	
Tubulin beta-1 chain	

Tubulin beta-2A chain
Tubulin beta-2B chain
Tubulin beta-3 chain
Tubulin beta-4A chain
Tubulin beta-4B chain
Tubulin beta-6 chain
Tubulin beta-8 chain
Tumor necrosis factor
Tumor susceptibility gene 101 protein
Tyrosinase
Tyrosine-protein phosphatase non-receptor type 1
Tyrosine-protein phosphatase non-receptor type 2
Xanthine dehydrogenase/oxidase

Supplementary Table 3 Top 100 most enriched targets in the AfroCancer dataset

Uniprot	Pref_Name	Afro CancerHits	Afro Cancer % Hits	PubChem Hits	PubChem % Hits	Odds_Ratio	Fishers Test p-value	Prediction Ratio
P04792	Heat shock protein beta-1	5	0.014	18	0	6.46E-04	6.51E-15	0.001
Q9NPH5	NADPH oxidase 4	38	0.104	350	0	1.50E-03	6.76E-91	0.002
P16152	Carbonyl reductase [NADPH] 1	32	0.088	473	0	2.45E-03	2.12E-70	0.003
P05091	Aldehyde dehydrogenase, mitochondrial	1	0.003	21	0	3.81E-03	0.003995653	0.004
Q16678	Cytochrome P450 1B1	21	0.058	437	0	3.57E-03	1.38E-43	0.004
P60568	Interleukin-2	1	0.003	26	0	4.72E-03	0.004901533	0.005
Q9NNW7	Thioredoxin reductase 2, mitochondrial	1	0.003	33	0	5.99E-03	0.006168385	0.006
P33527	Multidrug resistance-associated protein 1	37	0.102	1790	0.001	7.92E-03	1.15E-62	0.009
O95718	Steroid hormone receptor ERR2	8	0.022	419	0	9.32E-03	2.67E-14	0.01
P22001	Potassium voltage-gated channel subfamily A member 3	6	0.016	318	0	9.49E-03	5.09E-11	0.01
Q14534	Squalene monooxygenase	3	0.008	162	0	9.75E-03	4.30E-06	0.01
P04798	Cytochrome P450 1A1	1	0.003	61	0	1.11E-02	0.011219729	0.011
P0DMV8	Heat shock 70 kDa protein 1A	1	0.003	63	0	1.14E-02	0.011579558	0.011
P37058	Testosterone 17-beta-dehydrogenase 3	11	0.03	722	0	1.16E-02	4.22E-18	0.012
P37059	Estradiol 17-beta-dehydrogenase 2	10	0.027	635	0	1.12E-02	1.02E-16	0.012
P14061	Estradiol 17-beta-dehydrogenase 1	12	0.033	853	0	1.25E-02	3.24E-19	0.013
P15559	NAD (P)H dehydrogenase	27	0.074	1996	0.001	1.25E-02	3.97E-41	0.013
P47989	Xanthine dehydrogenase/oxidase	33	0.091	2343	0.001	1.18E-02	1.36E-50	0.013

	[Includes: Xanthine dehydrogenase							
P08236	Beta-glucuronidase	3	0.008	234	0	1.41E-02	1.27E-05	0.014
O95067	G2/mitotic-specific cyclin-B2	6	0.016	480	0	1.43E-02	5.74E-10	0.015
P14920	D-amino-acid oxidase	8	0.022	718	0	1.60E-02	1.83E-12	0.016
P19174	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase gamma-1	21	0.058	1831	0.001	1.50E-02	8.65E-31	0.016
P54646	5'-AMP-activated protein kinase catalytic subunit alpha-2	19	0.052	1852	0.001	1.68E-02	4.42E-27	0.018
Q8N1Q1	Carbonic anhydrase 13	66	0.181	6604	0.003	1.50E-02	3.36E-91	0.018
Q9UBM7	7-dehydrocholesterol reductase	6	0.016	670	0	2.00E-02	4.07E-09	0.02
P15121	Aldose reductase	73	0.201	8866	0.004	1.77E-02	5.31E-95	0.022
Q86W56	Poly(ADP-ribose) glycohydrolase	3	0.008	365	0	2.20E-02	4.69E-05	0.022
P10415	Apoptosis regulator Bcl-2	5	0.014	640	0	2.30E-02	1.62E-07	0.023
P46063	ATP-dependent DNA helicase Q1	79	0.217	10749	0.005	1.95E-02	3.65E-99	0.025
Q9UNQ0	ATP-binding cassette sub-family G member 2	41	0.113	5547	0.003	2.19E-02	2.05E-51	0.025
P10696	Alkaline phosphatase, placental-like	15	0.041	2174	0.001	2.53E-02	3.84E-19	0.026
P14679	Tyrosinase	20	0.055	2836	0.001	2.44E-02	2.93E-25	0.026
P08183	Multidrug resistance protein 1	94	0.258	13754	0.007	1.99E-02	1.06E-115	0.027
P80365	Corticosteroid 11-beta-dehydrogenase isozyme 2	15	0.041	2234	0.001	2.60E-02	5.72E-19	0.027
O60218	Aldo-keto reductase family 1 member B10	9	0.025	1384	0.001	2.73E-02	8.42E-12	0.028
P04278	Sex hormone-binding globulin	72	0.198	11279	0.006	2.30E-02	6.59E-86	0.029
P11387	DNA topoisomerase 1	45	0.124	7117	0.004	2.53E-02	2.10E-53	0.029
P13866	Sodium/glucose cotransporter 1	23	0.063	3608	0.002	2.68E-02	7.13E-28	0.029
Q96RI1	Bile acid receptor	27	0.074	4505	0.002	2.82E-02	8.35E-32	0.03
Q8TDS5	Oxoeicosanoid receptor 1	2	0.005	348	0	3.15E-02	0.001934041	0.032
P11388	DNA topoisomerase 2-alpha	114	0.313	20434	0.01	2.26E-02	1.03E-131	0.033
P16083	Ribosyldihydronicotinamide dehydrogenase [quinone]	6	0.016	1095	0.001	3.27E-02	7.19E-08	0.033
P23280	Carbonic anhydrase 6	58	0.159	10421	0.005	2.76E-02	1.13E-65	0.033
P07900	Heat shock protein HSP 90-alpha	34	0.093	6306	0.003	3.07E-02	2.96E-38	0.034
Q92731	Estrogen receptor beta	88	0.242	17068	0.009	2.70E-02	1.39E-97	0.035
Q9H2K2	Tankyrase-2	9	0.025	1812	0.001	3.58E-02	8.82E-11	0.037
O43451	Maltase-glucoamylase, intestinal [Includes: Maltase	18	0.049	3864	0.002	3.72E-02	9.97E-20	0.039
Q6ZVD8	PH domain leucine-rich repeat-containing protein phosphatase 2	5	0.014	1085	0.001	3.90E-02	2.10E-06	0.039
P35218	Carbonic anhydrase 5A, mitochondrial	52	0.143	11420	0.006	3.45E-02	1.74E-54	0.04
Q8WWL7	G2/mitotic-specific cyclin-B3	5	0.014	1159	0.001	4.16E-02	2.88E-06	0.042

Q92630	Dual specificity tyrosine-phosphorylation-regulated kinase 2	6	0.016	1370	0.001	4.09E-02	2.63E-07	0.042
P03372	Estrogen receptor	69	0.19	16604	0.008	3.58E-02	8.29E-70	0.044
P16050	Arachidonate 15-lipoxygenase	50	0.137	12163	0.006	3.84E-02	2.93E-50	0.044
Q99714	3-hydroxyacyl-CoA dehydrogenase type-2	59	0.162	14187	0.007	3.69E-02	1.34E-59	0.044
O75751	Solute carrier family 22 member 3	1	0.003	248	0	4.50E-02	0.044305221	0.045
O43570	Carbonic anhydrase 12	43	0.118	10869	0.005	4.08E-02	1.34E-42	0.046
P21964	Catechol O-methyltransferase	8	0.022	2036	0.001	4.53E-02	6.01E-09	0.046
Q9ULX7	Carbonic anhydrase 14	39	0.107	9912	0.005	4.15E-02	1.30E-38	0.046
P49761	Dual specificity protein kinase CLK3	4	0.011	1036	0.001	4.66E-02	4.50E-05	0.047
O95271	Tankyrase-1	6	0.016	1585	0.001	4.73E-02	6.10E-07	0.048
P43166	Carbonic anhydrase 7	50	0.137	13497	0.007	4.27E-02	4.29E-48	0.049
Q8TDU6	G-protein coupled bile acid receptor 1	6	0.016	1616	0.001	4.82E-02	6.82E-07	0.049
Q9HC97	G-protein coupled receptor 35	35	0.096	9609	0.005	4.54E-02	1.28E-33	0.05
P18054	Arachidonate 12-lipoxygenase, 12S-type	59	0.162	16424	0.008	4.28E-02	5.30E-56	0.051
P31639	Sodium/glucose cotransporter 2	5	0.014	1424	0.001	5.12E-02	7.74E-06	0.052
P68371	Tubulin beta-4B chain	22	0.06	6317	0.003	4.93E-02	3.69E-21	0.052
Q00534	Cyclin-dependent kinase 6	20	0.055	5754	0.003	4.96E-02	2.45E-19	0.052
Q3ZCM7	Tubulin beta-8 chain	32	0.088	9174	0.005	4.78E-02	2.98E-30	0.052
Q9UNA4	DNA polymerase iota	64	0.176	18394	0.009	4.35E-02	6.04E-60	0.052
P27695	DNA-(apurinic or apyrimidinic site)	85	0.234	24806	0.012	4.12E-02	1.33E-79	0.053
Q9UBT6	DNA polymerase kappa	131	0.36	38150	0.019	3.46E-02	6.61E-126	0.053
P10276	Retinoic acid receptor alpha	5	0.014	1480	0.001	5.32E-02	9.30E-06	0.054
P54132	Bloom syndrome protein	84	0.231	24829	0.012	4.19E-02	3.45E-78	0.054
Q13748	Tubulin alpha-3C/D chain	30	0.082	8871	0.004	4.96E-02	4.80E-28	0.054
P47712	Cytosolic phospholipase A2	2	0.005	612	0	5.54E-02	0.005773856	0.056
Q9BVA1	Tubulin beta-2B chain	32	0.088	9829	0.005	5.12E-02	2.43E-29	0.056
Q9Y2D0	Carbonic anhydrase 5B, mitochondrial	33	0.091	10155	0.005	5.12E-02	3.36E-30	0.056
P06746	DNA polymerase beta	72	0.198	22605	0.011	4.64E-02	6.22E-65	0.057
P17706	Tyrosine-protein phosphatase non-receptor type 2	22	0.06	6832	0.003	5.33E-02	1.90E-20	0.057
P23219	Prostaglandin G/H synthase 1	20	0.055	6330	0.003	5.46E-02	1.50E-18	0.058
P11474	Steroid hormone receptor ERR1	5	0.014	1626	0.001	5.84E-02	1.46E-05	0.059
Q16236	Nuclear factor erythroid 2-related factor 2	12	0.033	3871	0.002	5.69E-02	1.41E-11	0.059
P00338	L-lactate dehydrogenase A chain	2	0.005	664	0	6.01E-02	0.006752025	0.06
P13631	Retinoic acid receptor gamma	5	0.014	1653	0.001	5.94E-02	1.57E-05	0.06

Q13509	Tubulin beta-3 chain	24	0.066	7901	0.004	5.62E-02	1.29E-21	0.06
Q13510	Acid ceramidase	6	0.016	2018	0.001	6.03E-02	2.42E-06	0.061
Q9BQE3	Tubulin alpha-1C chain	35	0.096	11755	0.006	5.56E-02	1.04E-30	0.061
P08238	Heat shock protein HSP 90-beta	22	0.06	7497	0.004	5.85E-02	1.31E-19	0.062
Q71U36	Tubulin alpha-1A chain	35	0.096	11903	0.006	5.63E-02	1.57E-30	0.062
Q9Y468	Lethal(3)malignant brain tumor-like protein 1	48	0.132	16453	0.008	5.46E-02	1.90E-41	0.062
B2RXH2	Lysine-specific demethylase 4E	76	0.209	26131	0.013	5.02E-02	8.93E-66	0.063
P68363	Tubulin alpha-1B chain	20	0.055	7081	0.004	6.11E-02	1.24E-17	0.064
Q04760	Lactoylglutathione lyase	4	0.011	1410	0.001	6.35E-02	0.000146121	0.064
P39748	Flap endonuclease 1	96	0.264	34065	0.017	4.84E-02	1.43E-82	0.065
O14746	Telomerase reverse transcriptase	36	0.099	13044	0.007	5.98E-02	1.92E-30	0.066
O94921	Cyclin-dependent kinase 14	1	0.003	362	0	6.57E-02	0.063930888	0.066
P14174	Macrophage migration inhibitory factor	29	0.08	10498	0.005	6.10E-02	9.69E-25	0.066
P36888	Receptor-type tyrosine-protein kinase FLT3	7	0.019	2542	0.001	6.49E-02	5.77E-07	0.066
P00915	Carbonic anhydrase 1	27	0.074	9981	0.005	6.26E-02	6.96E-23	0.067
P07451	Carbonic anhydrase 3	39	0.107	14977	0.007	6.29E-02	5.55E-32	0.07

Supplementary Table 4 Pathways predicted for the AfroCancer and NCI Cancer datasets based on the predicted targets of AfroCancer compounds and experimental targets of NCI Cancer (Pathways in bold are common to both datasets)

Pathway_Name	**NCI Cancer % Hits**	**PubChem % Hits**	**χ^2 p-value**	**Pathway Name**	**AfroCancer % Hits**	**PubChem % Hits**	**χ^2 p-value**
AhR pathway	4.88E-05	2.96E-05	0.001758152	**AhR pathway**	0.000131079	8.48E-05	1.86E-06
Allograft Rejection	2.38E-06	1.31E-06	0.706957257	**Allograft Rejection**	5.74E-06	5.56E-06	0.887966376
BDNF signaling pathway	0.000362852	0.0003578886	0.832235449	**Androgen receptor signaling pathway**	9.38E-05	7.71E-05	0.073412904
Corticotropin-releasing hormone	0.000158227	0.0001559442	0.901231586	**B Cell Receptor Signaling Pathway**	0.000189442	0.00017122	0.187033132

Cytoplasmic Ribosomal Proteins	3.33E-05	3.33E-05	0.92856810 9	**BDNF signaling pathway**	0.0003 22434	0.000 300105	0.21 888 2315
Delta-Notch Signaling Pathway	0.0001 17778	8.78E-05	0.00 404 665 2	**Cell cycle**	0.0001 14813	8.44E-05	0.00 169 192 5
Estrogen signaling pathway	9.04E-05	6.35E-05	0.00 251 693 8	**Corticotropin-releasing hormone**	0.0002 06664	0.000 192515	0.33 743 664 9
Focal Adhesion	0.0003 92594	0.000 297355	4.87 E-07	**Cytokines and Inflammatory Response**	5.74E-06	5.56E-06	0.88 796 637 6
FSH signaling pathway	9.28E-05	7.23E-05	0.03 174 451 3	**Cytoplasmic Ribosomal Proteins**	1.05E-05	8.55E-06	0.63 019 766 1
Gastric cancer network 2	1.67E-05	9.98E-06	0.07 765 154	**Delta-Notch Signaling Pathway**	0.0001 23424	0.000 111267	0.28 203 785 5
IL-1 Signaling Pathway	2.02E-05	1.54E-05	0.32 845 190 6	**Endochondral Ossification**	3.35E-05	2.22E-05	0.02 745 047 7
IL-5 Signaling Pathway	0.0001 82021	0.000 176455	0.73 157 457 5	**Estrogen metabolism**	5.36E-05	3.35E-05	0.00 125 467 4
IL-6 Signaling Pathway	0.0003 21214	0.000 256197	0.00 022 455	**Estrogen signaling pathway**	0.0001 08116	0.000 104425	0.76 159 597 6
IL-7 Signaling Pathway	0.0001 65366	0.000 128397	0.00 325 574 8	**Fluoropyrimidine Activity**	6.22E-05	5.56E-05	0.42 677 733 9
IL17 signaling pathway	6.66E-05	5.95E-05	0.44 040 790 3	**Focal Adhesion**	7.56E-05	6.53E-05	0.23 728 502 3
Integrated Pancreatic Cancer Pathway	0.0004 00922	0.000 329409	0.00 034 073 3	**FSH signaling pathway**	7.18E-05	6.51E-05	0.45 688 853 1
MicroRNAs in cardiomyocyte hypertrophy	0.0002 2485	0.000 190007	0.02 274 993 8	**Gastric cancer network 2**	1.72E-05	1.65E-05	0.96 263 463 6
Mitochondrial Gene Expression	3.57E-06	7.10E-07	0.01 386 274 7	**Id Signaling Pathway**	9.76E-05	7.63E-05	0.02 119 616
Neural Crest Differentiation	0.0001 17778	9.21E-05	0.01 663	**IL-1 Signaling Pathway**	1.91E-06	5.99E-07	0.35 500

			2172				7904
Notch Signaling Pathway	1.67E-05	9.27E-06	0.04104129 3	**IL-2 Signaling pathway**	0.000136819	0.0001234 97	0.26127060 4
Nuclear Receptors	0.000140382	0.0001223 56	0.14859227 5	**IL-5 Signaling Pathway**	0.000209534	0.0001986 73	0.47335703 6
Parkin-Ubiquitin Proteasomal System pathway	7.38E-05	6.27E-05	0.22475709	**IL-6 Signaling Pathway**	0.000243022	0.0002210 8	0.15986689
Pathogenic Escherichia coli infection	0.000136813	0.0001079 88	0.01276706 1	**IL-7 Signaling Pathway**	0.00011003	1.00E-04	0.35246331 3
RANKL/RANK Signaling Pathway	0.000152279	0.0001178 6	0.00429029 5	**IL17 signaling pathway**	0.000111943	0.0001004 06	0.28377489 3
RB in Cancer	0.000105882	7.74E-05	0.00372552 4	**Integrated Pancreatic Cancer Pathway**	0.000628603	0.0005478 69	0.00084535 4
Regulation of Microtubule Cytoskeleton	0.000157038	0.0001227 55	0.00531743 6	**Interleukin-11 Signaling Pathway**	0.000106202	0.0001031 42	0.80678142 5
Signaling of Hepatocyte Growth Factor Receptor	0.000102313	7.95E-05	0.02228826 5	**MAPK Cascade**	5.74E-05	5.38E-05	0.68060627 4
Signaling Pathways in Glioblastoma	0.000285523	0.0001951 76	3.88E-09	**MicroRNAs in cardiomyocyte hypertrophy**	0.000138733	0.0001317 93	0.58471747 8
Steroid Biosynthesis	8.33E-06	1.82E-06	6.06E-05	**Mitochondrial Gene Expression**	4.78E-06	2.48E-06	0.28521302 2
TCR Signaling Pathway	0.000191539	0.0001624 83	0.04068637	**mRNA processing**	2.20E-05	9.15E-06	0.00015584 6
Androgen receptor signaling pathway	0.000121347	9.47E-05	0.01426924 4	**Neural Crest Differentiation**	8.71E-05	6.91E-05	0.04198447 6
B Cell Receptor Signaling Pathway	0.000354525	0.0002809 71	6.60E-05	**NOD pathway**	7.18E-05	6.05E-05	0.17811168 4
Cell cycle	0.000141572	0.0001317 45	0.46091001 5	**Notch Signaling Pathway**	8.61E-06	6.41E-06	0.52276 36

Cytokines and Inflammatory Response	2.38E-06	1.04E-06	0.507187054	**Nuclear Receptors**	0.000360705	0.0003585818	0.931316642
Endochondral Ossification	9.04E-05	7.71E-05	0.183327338	**Osteoblast Signaling**	3.83E-06	2.65E-06	0.6988853691
Estrogen metabolism	5.95E-06	3.52E-06	0.371734661	**Osteopontin Signaling**	0.000107159	9.71E-05	0.343523651
Fluoropyrimidine Activity	3.21E-05	1.43E-05	3.01E-05	**Oxidative Stress**	0.000125338	7.55E-05	6.22E-08
Id Signaling Pathway	9.87E-05	6.77E-05	0.000680484	**Parkin-Ubiquitin Proteasomal System pathway**	0.000353051	0.0002446	3.12E-11
IL-2 Signaling pathway	0.000158227	0.0001184112	0.000953705	**Pathogenic Escherichia coli infection**	0.000363576	0.0002623 03	1.98E-09
Interleukin-11 Signaling Pathway	0.00014752	0.0001088 28	0.000815407	**Prostate Cancer**	0.000183701	0.0001812 26	0.8869846 37
MAPK Cascade	7.73E-05	5.77E-05	0.021884571	**RANKL/RANK Signaling Pathway**	0.000147344	0.0001296 55	0.1419129 85
mRNA processing	3.57E-05	3.55E-05	0.945414678	**RB in Cancer**	7.46E-05	6.30E-05	0.1749876 03
NOD pathway	4.16E-05	3.89E-05	0.748537975	**Regulation of Microtubule Cytoskeleton**	8.71E-05	5.64E-05	0.0001231 59
Osteoblast Signaling	3.81E-05	2.38E-05	0.009965659	**Serotonin Receptor 4/6/7 and NR3C Signaling**	7.65E-05	6.78E-05	0.3319177 21
Osteopontin Signaling	4.64E-05	4.45E-05	0.856367283	**Signaling of Hepatocyte Growth Factor Receptor**	7.37E-05	6.82E-05	0.5623316 21
Oxidative Stress	2.50E-05	2.26E-05	0.723564292	**Signaling Pathways in Glioblastoma**	0.000147344	0.0001435 95	0.7921358 79
Prostate Cancer	0.000182021	0.0001703 28	0.435652086	**Steroid Biosynthesis**	3.54E-05	1.62E-05	1.35E-05
Serotonin Receptor 4/6/7 and NR3C Signaling	7.26E-05	6.68E-05	0.55979	**TCR Signaling Pathway**	0.000136819	0.0001341 02	0.85275

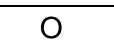
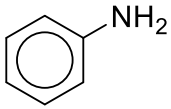
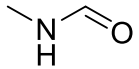
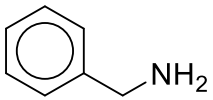
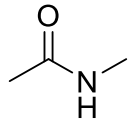
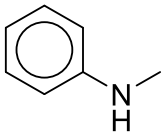
			5009				7452
TNF-alpha/NF-kB Signaling Pathway	0.000239126	0.0002112116	0.08489771 9	**TNF-alpha/NF-kB Signaling Pathway**	0.000216232	0.0001843 05	0.02443662 1
TOR signaling	4.64E-05	4.57E-05	0.98569789 1	**TOR signaling**	2.58E-05	2.45E-05	0.86632063 3
TSH signaling pathway	0.000161797	0.0001226 05	0.00139743 8	**TSH signaling pathway**	7.75E-05	6.67E-05	0.22115936 2
TWEAK Signaling Pathway	9.99E-05	9.53E-05	0.70718973 2	**TWEAK Signaling Pathway**	0.00013969	0.000 12726	0.30305200 6
Vitamin A and Carotenoid Metabolism	2.74E-05	1.31E-05	0.00052379	Adipogenesis	0.00019614	0.0001825 95	0.34655635 4
Vitamin A and carotenoid metabolism	2.74E-05	1.31E-05	0.00052379	Apoptosis	0.00013 2035	0.0001140 89	0.11233806 1
Alpha6-Beta4 Integrin Signaling Pathway	0.000202246	0.0001393 66	1.32E-06	Apoptosis Modulation and Signaling	0.00021511	0.0001103 26	0.32201289 2
Cardiac Progenitor Differentiation	7.85E-05	6.57E-05	0.16549876 4	ATM Signaling Pathway	3.83E-06	1.80E-06	0.29110293 8
DNA damage response (only ATM dependent)	0.000116589	9.63E-05	0.06626551 8	Benzo (a)pyrene metabolism	2.77E-05	1.76E-05	0.02845791
DNA Replication	1.43E-05	9.16E-06	0.17142930 4	Biogenic Amine Synthesis	5.07E-05	4.80E-05	0.75531041 2
ErbB signaling pathway	0.00014871	0.0001158 93	0.00607644 9	Codeine and morphine metabolism	8.99E-05	6.28E-05	0.00120924 3
Eukaryotic Transcription Initiation	1.43E-05	9.66E-06	0.23614723 8	Diurnally regulated genes with circadian orthologs	2.30E-05	1.68E-05	0.18075684 5
IL-9 Signaling Pathway	8.80E-05	6.74E-05	0.02514327 5	DNA damage response	9.09E-05	7.18E-05	0.03410743
Inflammatory Response Pathway	2.14E-05	1.06E-05	0.00386400 2	Dopamine metabolism	6.98E-05	5.35E-05	0.03686353 7

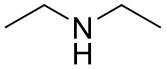
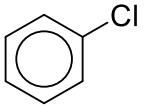
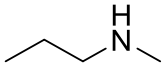
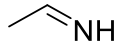
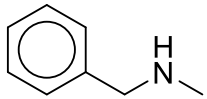
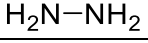
Insulin Signaling	0.000349766	0.0003347118	0.46889487 1	Drug Induction of Bile Acid Pathway	0.000166479	0.00016147	0.72967349 3
Integrin-mediated cell adhesion	0.000217711	0.0001775 18	0.00644098 2	Energy Metabolism	9.57E-05	8.99E-05	0.58705135 9
Kit Receptor Signaling Pathway	0.000233177	0.0001541 39	6.97E-09	FAS pathway and Stress induction of HSP regulation	2.97E-05	2.41E-05	0.32111864 9
MAPK signaling pathway	0.000312886	0.0002637 23	0.00611163 2	Folate Metabolism	4.59E-05	4.43E-05	0.87129036 9
miRNAs involved in DDR	3.33E-05	2.70E-05	0.31523530 7	G1 to S cell cycle control	6.22E-05	5.56E-05	0.42677733 9
Nifedipine Activity	2.26E-05	1.28E-05	0.01890822 6	G13 Signaling Pathway	2.87E-06	2.22E-06	0.93504249
NLR proteins	2.26E-05	1.80E-05	0.38456537 2	Glucocorticoid & Mineralcorticoid Metabolism	2.87E-05	1.68E-05	0.00854394 7
Nucleotide Metabolism	1.07E-05	1.34E-06	3.55E-12	Glycogen Metabolism	6.70E-06	2.57E-06	0.03798366 6
Oncostatin M Signaling Pathway	0.000198677	0.0001677 81	0.03199985 4	Hedgehog Signaling Pathway	2.10E-05	1.17E-05	0.01457435 2
One Carbon Metabolism	1.78E-05	1.97E-06	2.40E-23	Influenza A virus infection	4.78E-06	3.85E-06	0.83776983 8
Physiological and Pathological Hypertrophy of the Heart	3.81E-05	3.48E-05	0.67804559 5	Irinotecan Pathway	9.19E-05	6.29E-05	0.00055534 7
pilocytic astrocytoma	4.76E-06	3.00E-06	0.53650875 3	Keap1-Nrf2 Pathway	3.64E-05	2.58E-05	0.05777093 6
Proteasome Degradation	9.52E-06	7.54E-06	0.64571576 3	Matrix Metalloproteinases	1.44E-05	7.95E-06	0.04805673
Senescence and Autophagy	0.00011302	9.03E-05	0.03306617 8	metapathway biotransformation	0.000132035	0.0001004 91	0.00276133 3

AGE/RAGE pathway	0.000230798	0.00021548	0.357856419	Nicotine Activity on Dopaminergic Neurons	1.63E-05	1.49E-05	0.8272288825
Angiogenesis	0.0001499	9.47E-05	2.72E-07	Nuclear receptors in lipid metabolism and toxicity	0.000227713	0.0002133833	0.3557502999
Cholesterol biosynthesis	5.95E-06	5.54E-06	0.941053793	Ovarian Infertility Genes	0.00013969	0.0001282827	0.3483919233
EBV LMP1 signaling	9.04E-05	8.14E-05	0.3928929952	Phase I, non P450	1.34E-05	1.26E-05	0.9337621116
EGFR1 Signaling Pathway	0.000234367	0.0002207556	0.421924784	Polyol pathway	6.98E-05	4.11E-05	2.83E-05
EPO Receptor Signaling	8.80E-05	6.65E-05	0.018829752	Prostaglandin Synthesis and Regulation	9.19E-05	7.18E-05	0.0259474666
Gastric cancer network 1	2.02E-05	1.34E-05	0.12049207	Selenium Metabolism and Selenoproteins	5.36E-05	5.02E-05	0.6938068588
Heart Development	1.19E-05	6.87E-06	0.121073425	Selenium Pathway	0.000210491	0.0001863588	0.0919003711
IL-3 Signaling Pathway	0.000310507	0.0002382331	2.07E-05	Striated Muscle Contraction	8.90E-05	8.46E-05	0.6804074633
IL-4 signaling Pathway	0.00022247	0.0001867877	0.018597748	TGF Beta Signaling Pathway	7.27E-05	6.78E-05	0.6040224544
Integrated Breast Cancer Pathway	0.000281954	0.00023868	0.011324098	Tryptophan metabolism	4.31E-05	3.70E-05	0.3790039044
Interferon type I	9.52E-05	7.03E-05	0.007935041	Vitamin B12 Metabolism	4.59E-05	4.42E-05	0.8612828744
Leptin signaling pathway	0.000195108	0.00016005	0.01252039	Wnt Signaling Pathway NetPath	0.00013969	0.0001335044	0.6316838044
Nucleotide GPCRs	2.50E-05	1.84E-05	0.19748477				
Prolactin Signaling Pathway	0.000231988	0.0002204077	0.49753				

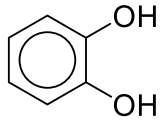
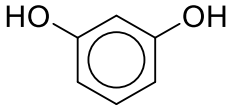
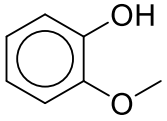
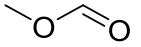
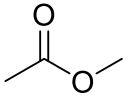
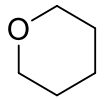
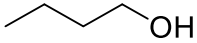
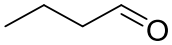
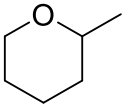
			5573				
Regulation of toll-like receptor signaling pathway	0.000183211	0.00017844	0.774459608				
SRF and miRs in Smooth Muscle Differentiation and Proliferation	1.07E-05	8.59E-06	0.63338314				
TFs Regulate miRNAs related to cardiac hypertrophy	2.86E-05	2.07E-05	0.14486143 9				
TGF-beta Receptor Signaling Pathway	0.000187969	0.000174122	0.35722186 4				
Toll-like receptor signaling pathway	0.000183211	0.00017844	0.774459608				
TP53 network	2.38E-05	1.48E-05	0.04612208 7				
Trans-sulfuration and one carbon metabolism	1.43E-05	1.14E-06	5.34E-27				
Trans-sulfuration pathway	1.19E-06	4.30E-07	0.81770729 5				
TSLP Signaling Pathway	0.000134434	0.000130933	0.81590106 9				
Type II diabetes mellitus	4.88E-05	3.85E-05	0.15196427				
Type III interferon signaling	1.78E-05	1.59E-05	0.75278930 8				

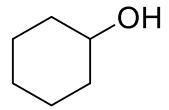
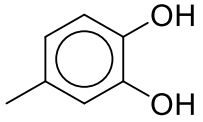
Supplementary Table 5 Fragments enriched in SM dataset (enriched over NP fragments)

Atom count	Bond count	Support in focus (abs)	Support in complement (abs)	Support in focus (rel)	Support in complement (rel)	Fragment (Molecule)
1	0	265	39	3.04E-01	4.73E-02	
4	3	255	40	2.93E-01	4.85E-02	
7	7	197	18	2.26E-01	2.18E-02	
4	3	189	32	2.17E-01	3.88E-02	
8	8	180	10	2.07E-01	1.21E-02	
5	4	178	30	2.04E-01	3.64E-02	
2	1	174	24	2.00E-01	2.91E-02	
1	0	174	25	2.00E-01	3.03E-02	
2	1	160	16	1.84E-01	1.94E-02	
8	8	133	10	1.53E-01	1.21E-02	

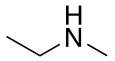
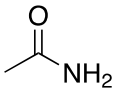
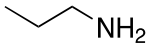
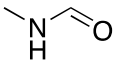
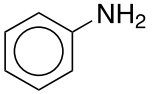
Atom count	Bond count	Support in focus (abs)	Support in complement (abs)	Support in focus (rel)	Support in complement (rel)	Fragment (Molecule)
5	4	133	30	1.53E-01	3.64E-02	
7	7	128	20	1.47E-01	2.42E-02	
5	4	128	37	1.47E-01	4.48E-02	
3	2	106	7	1.22E-01	8.48E-03	
2	1	100	5	1.15E-01	6.06E-03	
2	1	97	20	1.11E-01	2.42E-02	
9	9	96	6	1.10E-01	7.27E-03	
2	1	92	4	1.06E-01	4.85E-03	

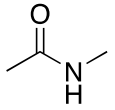
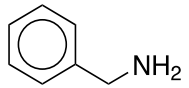
Supplementary Table 6 Fragments enriched in NP dataset (enriched over SM fragments)

Atom count	Bond count	Support in focus (abs)	Support in complement (abs)	Support in focus (rel)	Support in complement (rel)	Fragment (Molecule)
8	8	165	42	2.00E-01	4.82E-02	
8	8	160	14	1.94E-01	1.61E-02	
9	9	110	34	1.33E-01	3.90E-02	
4	3	106	41	1.28E-01	4.71E-02	
5	4	103	33	1.25E-01	3.79E-02	
6	6	100	26	1.21E-01	2.99E-02	
5	4	99	37	1.20E-01	4.25E-02	
4	3	98	34	1.19E-01	3.90E-02	
5	4	94	31	1.14E-01	3.56E-02	
7	7	93	7	1.13E-01	8.04E-03	

Atom count	Bond count	Support in focus (abs)	Support in complement (abs)	Support in focus (rel)	Support in complement (rel)	Fragment (Molecule)
7	7	91	11	1.10E-01	1.26E-02	
9	9	90	36	1.09E-01	4.13E-02	

Supplementary Table 7 Fragments enriched in the all active compounds (enriched over all inactive compounds)

Atom count	Bond count	Support in focus (abs)	Support in complement (abs)	Support in focus (rel)	Support in complement (rel)	Fragment (Molecule)
4	3	356	205	2.10E-01	4.86E-02	
4	3	295	197	1.74E-01	4.67E-02	
4	3	257	153	1.52E-01	3.63E-02	
4	3	221	145	1.30E-01	3.44E-02	
7	7	215	164	1.27E-01	3.89E-02	
6	6	212	107	1.25E-01	2.54E-02	

Atom count	Bond count	Support in focus (abs)	Support in complement (abs)	Support in focus (rel)	Support in complement (rel)	Fragment (Molecule)
5	4	208	128	1.23E-01	3.04E-02	
1	0	199	157	1.17E-01	3.72E-02	HCl
2	1	198	157	1.17E-01	3.72E-02	—Cl
8	8	190	122	1.12E-01	2.89E-02	
2	1	176	151	1.04E-01	3.58E-02	=NH

Supplementary Table 8 The predicted targets of the currently marketed HAT drugs at tpr > 0.9. Note that ornithine decarboxylase is correctly predicted for only elfornithine.

Gene Name	Pentamidine (-2.75)	Suramin (-14.89)	Nifurtimox (-1.47)	Elfornithine (-0.33)	Melarsoprol
Cell-division control protein 2 homolog 6, putative;Tb11.47.0031	0	0	0	0	1
Cell division control protein 2 homolog 2;Tb927.7.7360	0	0	0	0	1
Cell division related protein kinase 2, putative;Tb10.70.2210	0	0	0	0	1
Protein kinase, putative;Tb10.70.1760	0	0	0	0	1
Serine/threonine-protein kinase, putative;Tb10.70.5890	0	0	0	0	1
Serine/threonine protein kinase, putative;Tb927.3.4560	0	0	0	0	1
Protein kinase, putative;Tb09.160.0570	0	0	0	0	1
Protein kinase, putative;Tb11.01.0330	0	0	0	0	1
Protein kinase, putative;Tb927.7.1900	0	0	0	0	1
Pteridine reductase, putative;Tb927.8.2210	0	0	0	0	1
Protein kinase, putative;Tb09.211.2260	0	0	0	0	1
Glucose transporter;Tb10.6k15.2030	1	0	0	0	0
Glucose transporter, putative;Tb927.4.2290	1	0	0	0	0
Hexose transporter;Tb10.6k15.2040	1	0	0	0	0
Glucose transporter;Tb10.6k15.2020	1	0	0	0	0
Ornithine decarboxylase;Tb11.01.5300	0	0	0	1	0

Gene Name	Pentamidine (-2.75)	Suramin (-14.89)	Nifurtimox (-1.47)	Elfornithine (-0.33)	Melarsoprol
Cyclin 3;Tb927.6.1460	0	0	0	0	1
Cyclin 6;Tb11.01.8460	0	0	0	0	1
NADPH--cytochrome p450 reductase, putative;Tb09.211.4110	0	0	0	0	1
NADPH--cytochrome P450 reductase, putative;Tb11.01.0170	0	0	0	0	1
NADPH--cytochrome p450 reductase, putative;Tb11.02.5420	0	0	0	0	1
NADPH--cytochrome p450 reductase, putative;Tb09.211.4110	1	0	0	0	0
NADPH--cytochrome P450 reductase, putative;Tb11.01.0170	1	0	0	0	0
NADPH--cytochrome p450 reductase, putative;Tb11.02.5420	1	0	0	0	0
NADPH--cytochrome p450 reductase, putative;Tb09.211.4110	0	0	0	1	0
NADPH--cytochrome P450 reductase, putative;Tb11.01.0170	0	0	0	1	0
NADPH--cytochrome p450 reductase, putative;Tb11.02.5420	0	0	0	1	0
Protein kinase, putative;Tb09.211.2260	0	0	0	0	1
Protein kinase, putative;Tb10.329.0030	0	0	0	0	1
Mitogen-activated protein kinase;Tb927.8.3550	0	0	0	0	1
Protein kinase, putative;Tb927.7.1900	0	0	0	0	1
Protein kinase, putative;Tb11.01.4130	0	0	0	0	1
Protein kinase;Tb11.01.1030	0	0	0	0	1
Rac serine-threonine kinase, putative;Tb927.6.2250	0	0	0	0	1

Gene Name	Pentamidine (-2.75)	Suramin (-14.89)	Nifurtimox (-1.47)	Elfornithine (-0.33)	Melarsoprol
Ubiquitin-conjugating enzyme E2, putative;Tb09.211.0050	1	0	0	0	0
Cyclin 3;Tb927.6.1460	0	0	0	0	1
Cyclin 6;Tb11.01.8460	0	0	0	0	1
Protein kinase, putative;Tb10.70.1760	0	0	0	0	1
Serine/threonine-protein kinase, putative;Tb10.70.5890	0	0	0	0	1
Serine/threonine protein kinase, putative;Tb927.3.4560	0	0	0	0	1
NAD-dependent protein deacetylase SIR2rp1;SIR2rp1	1	0	0	0	0
Protein kinase, putative;Tb09.160.0570	0	0	0	0	1
Protein kinase, putative;Tb11.01.0330	0	0	0	0	1
Protein kinase, putative;Tb927.7.3210	0	0	0	0	1

Supplementary Table 9 Predicted targets common to 5 known drugs and compounds in SH dataset

Gene ID	Gene Name	Biological Process
Tb11.01.8460	Cyclin 6	cell cycle
Tb11.01.4130	Protein kinase, putative	biosynthetic process;cell cycle;nitrogen compound metabolic process;phosphate-containing compound metabolic process;regulation of transcription from RNA polymerase II promoter;transcription elongation from RNA polymerase II promoter
Tb927.4.2290	Glucose transporter, putative	
Tb10.6k15.2040	Hexose transporter	
Tb11.02.5420	NADPH--cytochrome p450 reductase, putative	
Tb10.70.1760	Non-specific serine/threonine protein kinase	intracellular signal transduction;phosphate-containing compound metabolic process;regulation of biological process;response to stimulus
Tb09.211.4110	NADPH--cytochrome p450 reductase, putative	
Tb10.6k15.2030	Glucose transporter	
Tb927.7.3210	Protein kinase, putative	cell cycle
Tb11.01.1030	Protein kinase	
Tb11.01.0170	NADPH--cytochrome p450 reductase, putative	

Gene ID	Gene Name	Biological Process
Tb11.01.0330	Protein kinase, putative	chromatin organization;cytokinesis;cytoskeleton organization;phosphate-containing compound metabolic process;regulation of cell cycle
Tb927.3.4560	Non-specific serine/threonine protein kinase	intracellular signal transduction;phosphate-containing compound metabolic process;regulation of biological process;response to stimulus
Tb927.7.1900	Protein kinase, putative	biosynthetic process;cell cycle;nitrogen compound metabolic process;phosphate-containing compound metabolic process;regulation of transcription from RNA polymerase II promoter
Tb10.6k15.2020	Glucose transporter	
Tb10.70.5890	Serine/threonine-protein kinase, putative	intracellular signal transduction;phosphate-containing compound metabolic process;regulation of biological process;response to stimulus
Tb927.6.1460	Cyclin 3	cell cycle

Supplementary Table 10 Bioactivity values extracted from ChEMBL for the NP dataset.

CMPD_CHEMBLID	PCHEMBL_VALUE	TARGET_CHEMBLID	PROTEIN_ACCESSION	PREF_NAME	ORGANISM
CHEMBL6246	5	CHEMBL4331	P68871	Hemoglobin beta chain	*Homo sapiens*
CHEMBL1553072	5	CHEMBL1293231	P51450	Nuclear receptor ROR-gamma	*Mus musculus*
CHEMBL168	5.02	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL168	5.02	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL400074	5.02	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL6246	5.03	CHEMBL3594	Q16790	Carbonic anhydrase IX	*Homo sapiens*
CHEMBL6246	5.03	CHEMBL4362	Q6P6U0	Tyrosine-protein kinase FGR	*Rattus norvegicus*
CHEMBL169	5.03	CHEMBL1741186	P51449	Nuclear receptor ROR-gamma	*Homo sapiens*
CHEMBL221543	5.03	CHEMBL2581	P07339	Cathepsin D	*Homo sapiens*
CHEMBL6246	5.04	CHEMBL3729	P22748	Carbonic anhydrase IV	*Homo sapiens*
CHEMBL463665	5.04	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL6246	5.05	CHEMBL3510	Q9ULX7	Carbonic anhydrase XIV	*Homo sapiens*

CMPD_CHEMBLID	PCHEMBL_VALUE	TARGET_CHEMBLID	PROTEIN_ACCESSION	PREF_NAME	ORGANISM
CHEMBL169	5.05	CHEMBL4696	P00489	Glycogen phosphorylase, muscle form	*Oryctolagus cuniculus*
CHEMBL169	5.05	CHEMBL4696	P00489	Glycogen phosphorylase, muscle form	*Oryctolagus cuniculus*
CHEMBL1553072	5.05	CHEMBL1293231	P51450	Nuclear receptor ROR-gamma	*Mus musculus*
CHEMBL6246	5.06	CHEMBL2326	P43166	Carbonic anhydrase VII	*Homo sapiens*
CHEMBL8659	5.07	CHEMBL2916	O14746	Telomerase reverse transcriptase	*Homo sapiens*
CHEMBL169	5.07	CHEMBL4343	P06766	DNA polymerase beta	*Rattus norvegicus*
CHEMBL6246	5.09	CHEMBL3242	O43570	Carbonic anhydrase XII	*Homo sapiens*
CHEMBL168	5.1	CHEMBL2366517	Q9YQ12	Protease	*Human immunodeficiency virus 1*
CHEMBL169	5.1	CHEMBL2366517	Q9YQ12	Protease	*Human immunodeficiency virus 1*
CHEMBL443146	5.1	CHEMBL4078	O42275	Acetylcholinesterase	*Electrophorus electricus*
CHEMBL1555307	5.1	CHEMBL4372	P15917	Anthrax lethal factor	*Bacillus anthracis*
CHEMBL6246	5.12	CHEMBL4789	P35218	Carbonic anhydrase VA	*Homo sapiens*
CHEMBL168	5.12	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL6246	5.13	CHEMBL262	P49841	Glycogen synthase kinase-3 beta	*Homo sapiens*
CHEMBL82293	5.13	CHEMBL3979	Q03181	Peroxisome proliferator-activated receptor delta	*Homo sapiens*
CHEMBL168	5.13	CHEMBL4343	P06766	DNA polymerase beta	*Rattus norvegicus*
CHEMBL169	5.14	CHEMBL4903	P24666	Low molecular weight phosphotyrosine protein phosphatase	*Homo sapiens*
CHEMBL6246	5.15	CHEMBL1293226	B2RXH2	Lysine-specific demethylase 4D-like	*Homo sapiens*
CHEMBL6246	5.15	CHEMBL3025	P23280	Carbonic anhydrase VI	*Homo sapiens*
CHEMBL463665	5.15	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL269277	5.19	CHEMBL4343	P06766	DNA polymerase beta	*Rattus norvegicus*
CHEMBL169	5.19	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL6246	5.2	CHEMBL2326	P43166	Carbonic anhydrase VII	*Homo sapiens*
CHEMBL365375	5.2	CHEMBL1293231	P51450	Nuclear receptor ROR-gamma	*Mus musculus*
CHEMBL168	5.22	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL168	5.22	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.22	CHEMBL5983	O60218	Aldo-keto reductase family 1 member B10	*Homo sapiens*
CHEMBL1783810	5.22	CHEMBL2366517	Q9YQ12	Protease	*Human immunodeficiency virus 1*
CHEMBL168	5.25	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL168	5.26	CHEMBL3807	P17706	T-cell protein-tyrosine phosphatase	*Homo sapiens*
CHEMBL8659	5.28	CHEMBL3979	Q03181	Peroxisome proliferator-activated receptor delta	*Homo sapiens*

CMPD_CHEMBLID	PCHEMBL_VALUE	TARGET_CHEMBLID	PROTEIN_ACCESSION	PREF_NAME	ORGANISM
CHEMBL169	5.28	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL6246	5.29	CHEMBL2366505	Q76353	Integrase	*Human immunodeficiency virus 1*
CHEMBL6246	5.29	CHEMBL3471	Q7ZJM1	Human immunodeficiency virus type 1 integrase	*Human immunodeficiency virus 1*
CHEMBL82293	5.3	CHEMBL4163	O60603	Toll-like receptor 2	*Homo sapiens*
CHEMBL169	5.32	CHEMBL4343	P06766	DNA polymerase beta	*Rattus norvegicus*
CHEMBL168	5.35	CHEMBL3807	P17706	T-cell protein-tyrosine phosphatase	*Homo sapiens*
CHEMBL6246	5.37	CHEMBL3912	Q8N1Q1	Carbonic anhydrase XIII	*Homo sapiens*
CHEMBL6246	5.37	CHEMBL4364	Q64725	Tyrosine-protein kinase SYK	*Rattus norvegicus*
CHEMBL168	5.37	CHEMBL3807	P17706	T-cell protein-tyrosine phosphatase	*Homo sapiens*
CHEMBL168	5.38	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL8659	5.39	CHEMBL235	P37231	Peroxisome proliferator-activated receptor gamma	*Homo sapiens*
CHEMBL169	5.39	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL6246	5.4	CHEMBL1293234	O97447	Putative fructose-1,6-bisphosphate aldolase	*Giardia intestinalis*
CHEMBL168	5.4	CHEMBL5983	O60218	Aldo-keto reductase family 1 member B10	*Homo sapiens*
CHEMBL169	5.4	CHEMBL5983	O60218	Aldo-keto reductase family 1 member B10	*Homo sapiens*
CHEMBL169	5.4	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL1783811	5.4	CHEMBL2366517	Q9YQ12	Protease	*Human immunodeficiency virus 1*
CHEMBL1783814	5.4	CHEMBL2366517	Q9YQ12	Protease	*Human immunodeficiency virus 1*
CHEMBL1783815	5.4	CHEMBL2366517	Q9YQ12	Protease	*Human immunodeficiency virus 1*
CHEMBL6246	5.41	CHEMBL4822	P56817	Beta-secretase 1	*Homo sapiens*
CHEMBL168	5.41	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.41	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL202496	5.41	CHEMBL5077	Q9N1N9	Butyrylcholinesterase	*Equus caballus*
CHEMBL168	5.42	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.42	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.42	CHEMBL3521	P10586	Receptor-type tyrosine-protein phosphatase F (LAR)	*Homo sapiens*
CHEMBL168	5.43	CHEMBL4343	P06766	DNA polymerase beta	*Rattus norvegicus*
CHEMBL169	5.43	CHEMBL5983	O60218	Aldo-keto reductase family 1 member B10	*Homo sapiens*
CHEMBL202496	5.43	CHEMBL5077	Q9N1N9	Butyrylcholinesterase	*Equus caballus*
CHEMBL169	5.44	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.44	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*

CMPD_CHEMBLID	PCHEMBL_VALUE	TARGET_CHEMBLID	PROTEIN_ACCESSION	PREF_NAME	ORGANISM
CHEMBL169	5.44	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.44	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.44	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.44	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL6246	5.45	CHEMBL1824	P04626	Receptor protein-tyrosine kinase erbB-2	*Homo sapiens*
CHEMBL169	5.46	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL168	5.47	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.47	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.47	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL6246	5.48	CHEMBL3024	P53350	Serine/threonine-protein kinase PLK1	*Homo sapiens*
CHEMBL6246	5.48	CHEMBL4282	P31749	Serine/threonine-protein kinase AKT	*Homo sapiens*
CHEMBL269277	5.5	CHEMBL1293231	P51450	Nuclear receptor ROR-gamma	*Mus musculus*
CHEMBL6246	5.51	CHEMBL2695	Q05397	Focal adhesion kinase 1	*Homo sapiens*
CHEMBL169	5.51	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.51	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.51	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.51	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.51	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL168	5.52	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL168	5.52	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL168	5.52	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL168	5.52	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL6246	5.53	CHEMBL3788	O00444	Serine/threonine-protein kinase PLK4	*Homo sapiens*
CHEMBL6246	5.53	CHEMBL1293267	Q9HC97	G-protein coupled receptor 35	*Homo sapiens*
CHEMBL6246	5.54	CHEMBL4363	Q07014	Tyrosine-protein kinase Lyn	*Rattus norvegicus*
CHEMBL8659	5.54	CHEMBL5738	P02692	Fatty acid-binding protein, liver	*Rattus norvegicus*
CHEMBL169	5.54	CHEMBL3967	P00599	Phospholipase A2 isozyme DE-I	*Naja melanoleuca*
CHEMBL169	5.55	CHEMBL5983	O60218	Aldo-keto reductase family 1 member B10	*Homo sapiens*
CHEMBL169	5.56	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL365375	5.57	CHEMBL1293228	P10520	Streptokinase A	*Streptococcus pyogenes serotype M1*
CHEMBL82293	5.59	CHEMBL3344	P05413	Fatty acid binding protein muscle	*Homo sapiens*
CHEMBL168	5.59	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL6246	5.6	CHEMBL2185	Q96GD4	Serine/threonine-protein kinase Aurora-B	*Homo sapiens*
CHEMBL6246	5.6	CHEMBL1293226	B2RXH2	Lysine-specific demethylase 4D-like	*Homo sapiens*

CMPD_CHEMBLID	PCHEMBL_VALUE	TARGET_CHEMBLID	PROTEIN_ACCESSION	PREF_NAME	ORGANISM
CHEMBL169	5.6	CHEMBL5022	P59264	Phospholipase A2 isozyme PLA-A	*Trimeresurus flavoviridis*
CHEMBL6246	5.61	CHEMBL4722	O14965	Serine/threonine-protein kinase Aurora-A	*Homo sapiens*
CHEMBL169	5.62	CHEMBL3807	P17706	T-cell protein-tyrosine phosphatase	*Homo sapiens*
CHEMBL6246	5.63	CHEMBL4899	P41279	Mitogen-activated protein kinase kinase kinase 8	*Homo sapiens*
CHEMBL6246	5.64	CHEMBL261	P00915	Carbonic anhydrase I	*Homo sapiens*
CHEMBL169	5.64	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.64	CHEMBL4195	Q7T3S7	Phospholipase A2	*Echis carinatus*
CHEMBL6246	5.66	CHEMBL205	P00918	Carbonic anhydrase II	*Homo sapiens*
CHEMBL169	5.68	CHEMBL5983	O60218	Aldo-keto reductase family 1 member B10	*Homo sapiens*
CHEMBL6246	5.7	CHEMBL4241	P52020	Squalene monooxygenase	*Rattus norvegicus*
CHEMBL6246	5.7	CHEMBL1293234	O97447	Putative fructose-1,6-bisphosphate aldolase	*Giardia intestinalis*
CHEMBL6246	5.7	CHEMBL1293237	P54132	Bloom syndrome protein	*Homo sapiens*
CHEMBL269277	5.7	CHEMBL5983	O60218	Aldo-keto reductase family 1 member B10	*Homo sapiens*
CHEMBL168	5.7	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.7	CHEMBL5983	O60218	Aldo-keto reductase family 1 member B10	*Homo sapiens*
CHEMBL168	5.72	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.72	CHEMBL4235	P28845	11-beta-hydroxysteroid dehydrogenase 1	*Homo sapiens*
CHEMBL6246	5.74	CHEMBL5147	P54760	Ephrin type-B receptor 4	*Homo sapiens*
CHEMBL6246	5.75	CHEMBL5145	P15056	Serine/threonine-protein kinase B-raf	*Homo sapiens*
CHEMBL169	5.75	CHEMBL1293228	P10520	Streptokinase A	*Streptococcus pyogenes serotype M1*
CHEMBL82293	5.77	CHEMBL4879	P12104	Fatty acid binding protein intestinal	*Homo sapiens*
CHEMBL6246	5.81	CHEMBL1955	P35916	Vascular endothelial growth factor receptor 3	*Homo sapiens*
CHEMBL8659	5.82	CHEMBL2083	P15090	Fatty acid binding protein adipocyte	*Homo sapiens*
CHEMBL82293	5.82	CHEMBL239	Q07869	Peroxisome proliferator-activated receptor alpha	*Homo sapiens*
CHEMBL6246	5.89	CHEMBL1913	P09619	Platelet-derived growth factor receptor beta	*Homo sapiens*
CHEMBL1610940	5.9	CHEMBL4261	Q16665	Hypoxia-inducible factor 1 alpha	*Homo sapiens*
CHEMBL82293	5.92	CHEMBL3674	Q01469	Fatty acid binding protein epidermal	*Homo sapiens*
CHEMBL168	5.93	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL6246	5.95	CHEMBL1293227	O75604	Ubiquitin carboxyl-terminal hydrolase 2	*Homo sapiens*
CHEMBL6246	5.95	CHEMBL1293237	P54132	Bloom syndrome protein	*Homo sapiens*
CHEMBL168	5.96	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	5.98	CHEMBL1741186	P51449	Nuclear receptor ROR-gamma	*Homo sapiens*
CHEMBL168	5.99	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*

CMPD_CHEMBLID	PCHEMBL_VALUE	TARGET_CHEMBLID	PROTEIN_ACCESSION	PREF_NAME	ORGANISM
CHEMBL168	6	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	6	CHEMBL1293231	P51450	Nuclear receptor ROR-gamma	*Mus musculus*
CHEMBL168	6.01	CHEMBL4804	P30305	Dual specificity phosphatase Cdc25B	*Homo sapiens*
CHEMBL82293	6.03	CHEMBL2083	P15090	Fatty acid binding protein adipocyte	*Homo sapiens*
CHEMBL6246	6.05	CHEMBL1293255	P15428	15-hydroxyprostaglandin dehydrogenase [NAD+]	*Homo sapiens*
CHEMBL6246	6.05	CHEMBL1293255	P15428	15-hydroxyprostaglandin dehydrogenase [NAD+]	*Homo sapiens*
CHEMBL6246	6.1	CHEMBL267	P12931	Tyrosine-protein kinase SRC	*Homo sapiens*
CHEMBL6246	6.1	CHEMBL279	P35968	Vascular endothelial growth factor receptor 2	*Homo sapiens*
CHEMBL6246	6.1	CHEMBL1293226	B2RXH2	Lysine-specific demethylase 4D-like	*Homo sapiens*
CHEMBL6246	6.12	CHEMBL5460	P0DMV8	Heat shock 70 kDa protein 1	*Homo sapiens*
CHEMBL168	6.14	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	6.15	CHEMBL1293231	P51450	Nuclear receptor ROR-gamma	*Mus musculus*
CHEMBL6246	6.16	CHEMBL203	P00533	Epidermal growth factor receptor erbB1	*Homo sapiens*
CHEMBL168	6.16	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL168	6.16	CHEMBL335	P18031	Protein-tyrosine phosphatase 1B	*Homo sapiens*
CHEMBL169	6.17	CHEMBL1741186	P51449	Nuclear receptor ROR-gamma	*Homo sapiens*
CHEMBL8659	6.22	CHEMBL239	Q07869	Peroxisome proliferator-activated receptor alpha	*Homo sapiens*
CHEMBL8659	6.22	CHEMBL239	Q07869	Peroxisome proliferator-activated receptor alpha	*Homo sapiens*
CHEMBL6246	6.24	CHEMBL3717	P08581	Hepatocyte growth factor receptor	*Homo sapiens*
CHEMBL6246	6.25	CHEMBL1293236	P46063	ATP-dependent DNA helicase Q1	*Homo sapiens*
CHEMBL6246	6.27	CHEMBL1293308	Q5A7N4	Likely tRNA 2'-phosphotransferase	*Candida albicans (strain SC5314 / ATCC MYA-2876) (Yeast)*
CHEMBL6246	6.29	CHEMBL5784	O60285	NUAK family SNF1-like kinase 1	*Homo sapiens*
CHEMBL169	6.3	CHEMBL1293231	P51450	Nuclear receptor ROR-gamma	*Mus musculus*
CHEMBL1553072	6.3	CHEMBL340	P08684	Cytochrome P450 3A4	*Homo sapiens*
CHEMBL506814	6.39	CHEMBL4822	P56817	Beta-secretase 1	*Homo sapiens*
CHEMBL3122152	6.4	CHEMBL3344	P05413	Fatty acid binding protein muscle	*Homo sapiens*
CHEMBL6246	6.47	CHEMBL1981	P06213	Insulin receptor	*Homo sapiens*
CHEMBL6246	6.59	CHEMBL4128	Q02763	Tyrosine-protein kinase TIE-2	*Homo sapiens*
CHEMBL6246	6.6	CHEMBL1957	P08069	Insulin-like growth factor I receptor	*Homo sapiens*
CHEMBL6246	6.6	CHEMBL4159	Q99714	Endoplasmic reticulum-associated amyloid beta-peptide-binding protein	*Homo sapiens*

CMPD_CHEMBLID	PCHEMBL_VALUE	TARGET_CHEMBLID	PROTEIN_ACCESSION	PREF_NAME	ORGANISM
CHEMBL6246	6.7	CHEMBL1900	P15121	Aldose reductase	*Homo sapiens*
CHEMBL8659	6.75	CHEMBL5738	P02692	Fatty acid-binding protein, liver	*Rattus norvegicus*
CHEMBL6246	6.8	CHEMBL2392	P06746	DNA polymerase beta	*Homo sapiens*
CHEMBL3122151	6.8	CHEMBL3344	P05413	Fatty acid binding protein muscle	*Homo sapiens*
CHEMBL6246	6.85	CHEMBL1293258	P84022	Mothers against decapentaplegic homolog 3	*Homo sapiens*
CHEMBL169	6.89	CHEMBL1741186	P51449	Nuclear receptor ROR-gamma	*Homo sapiens*
CHEMBL6246	6.9	CHEMBL1293255	P15428	15-hydroxyprostaglandin dehydrogenase [NAD+]	*Homo sapiens*
CHEMBL6246	6.9	CHEMBL1293236	P46063	ATP-dependent DNA helicase Q1	*Homo sapiens*
CHEMBL6246	6.96	CHEMBL1293267	Q9HC97	G-protein coupled receptor 35	*Homo sapiens*
CHEMBL6246	7	CHEMBL1075138	Q9NUW8	Tyrosyl-DNA phosphodiesterase 1	*Homo sapiens*
CHEMBL6246	7	CHEMBL1293267	Q9HC97	G-protein coupled receptor 35	*Homo sapiens*
CHEMBL168	7.05	CHEMBL5983	O60218	Aldo-keto reductase family 1 member B10	*Homo sapiens*
CHEMBL3122153	7.13	CHEMBL3344	P05413	Fatty acid binding protein muscle	*Homo sapiens*
CHEMBL6246	7.15	CHEMBL1075138	Q9NUW8	Tyrosyl-DNA phosphodiesterase 1	*Homo sapiens*
CHEMBL6246	7.4	CHEMBL3629	P68400	Casein kinase II alpha	*Homo sapiens*
CHEMBL6246	7.85	CHEMBL5619	P27695	DNA-(apurinic or apyrimidinic site) lyase	*Homo sapiens*

Supplementary Table 11 Enriched targets predicted for the NP dataset

PROTEIN_ACCESSION	PREF_NAME	ORGANISM	*Trypanosoma brucei* orthologue	Gene Name
P43166	Carbonic anhydrase VII	*Homo sapiens*	Tb11.01.0290	Carbonic anhydrase-like protein
O14746	Telomerase reverse transcriptase	*Homo sapiens*	Tb11.01.1950	Telomerase reverse transcriptase, putative
P22748	Carbonic anhydrase IV	*Homo sapiens*	Tb11.01.0290	Carbonic anhydrase-like protein
Q9ULX7	Carbonic anhydrase XIV	*Homo sapiens*	Tb11.01.0290	Carbonic anhydrase-like protein
Q16790	Carbonic anhydrase IX	*Homo sapiens*	Tb11.01.0290	Carbonic anhydrase-like protein
O43570	Carbonic anhydrase XII	*Homo sapiens*	Tb11.01.0290	Carbonic anhydrase-like protein
Q8N1Q1	Carbonic anhydrase XIII	*Homo sapiens*	Tb11.01.0290	Carbonic anhydrase-like protein
P35218	Carbonic anhydrase VA	*Homo sapiens*	Tb11.01.0290	Carbonic anhydrase-like protein
P23280	Carbonic anhydrase VI	*Homo sapiens*	Tb11.01.0290	Carbonic anhydrase-like protein
P49841	Glycogen synthase kinase-3 beta	*Homo sapiens*	Tb927.10.13780	Glycogen synthase kinase 3
O60218	Aldo-keto reductase family 1 member B10	*Homo sapiens*	Tb11.02.3040	Aldo/keto reductase, putative
P10586	Receptor-type tyrosine-protein phosphatase F (LAR)	*Homo sapiens*	Tb10.70.0070	Tyrosine specific protein phosphatase, putative
P53350	Serine/threonine-protein kinase PLK1	*Homo sapiens*	Tb927.7.3210	Protein kinase, putative

PROTEIN_ACCESSION	PREF_NAME	ORGANISM	*Trypanosoma brucei* orthologue	Gene Name
P53350	Serine/threonine-protein kinase PLK1	*Homo sapiens*	Tb927.6.5100	Serine/threonine-protein kinase, putative
P53350	Serine/threonine-protein kinase PLK1	*Homo sapiens*	Tb927.7.6310	Serine/threonine-protein kinase PLK
P41279	Mitogen-activated protein kinase kinase kinase 8	*Homo sapiens*	Tb11.46.0003	Protein kinase, putative
Q96GD4	Serine/threonine-protein kinase Aurora-B	*Homo sapiens*	Tb11.01.0330	Protein kinase, putative
O14965	Serine/threonine-protein kinase Aurora-A	*Homo sapiens*	Tb11.01.0330	Protein kinase, putative
P00915	Carbonic anhydrase I	*Homo sapiens*	Tb11.01.0290	Carbonic anhydrase-like protein
P54132	Bloom syndrome protein	*Homo sapiens*	Tb927.8.6690	ATP-dependent DEAD/H DNA helicase recQ, putative
P00918	Carbonic anhydrase II	*Homo sapiens*	Tb11.01.0290	Carbonic anhydrase-like protein
P0DMV8	Heat shock 70 kDa protein 1	*Homo sapiens*	Tb09.160.3090	Heat shock protein, putative
P27695	DNA-(apurinic or apyrimidinic site) lyase	*Homo sapiens*	Tb927.8.5510	DNA-(apurinic or apyrimidinic site) lyase
P15121	Aldose reductase	*Homo sapiens*	Tb11.02.3040	Aldo/keto reductase, putative
Q9NUW8	Tyrosyl-DNA phosphodiesterase 1	*Homo sapiens*	Tb927.2.5750	Tyrosyl-DNA Phosphodiesterase (Tdp1), putative
P68400	Casein kinase II alpha	*Homo sapiens*	Tb927.2.2430	Casein kinase II, alpha chain
P68400	Casein kinase II alpha	*Homo sapiens*	Tb09.211.4890	Casein kinase II, putative
P06746	DNA polymerase beta	*Homo sapiens*	Tb927.5.2780	Mitochondrial DNA polymerase beta
P06746	DNA polymerase beta	*Homo sapiens*	Tb927.5.2790	Mitochondrial DNA polymerase beta-PAK
Q6P6U0	Tyrosine-protein kinase FGR	*Rattus norvegicus*	Tb927.5.2780	Mitochondrial DNA polymerase beta
Q6P6U0	Tyrosine-protein kinase FGR	*Rattus norvegicus*	Tb927.5.2790	Mitochondrial DNA polymerase beta-PAK
P06766	DNA polymerase beta	*Rattus norvegicus*	Tb927.5.2780	Mitochondrial DNA polymerase beta
P06766	DNA polymerase beta	*Rattus norvegicus*	Tb927.5.2790	Mitochondrial DNA polymerase beta-PAK
Q07014	Tyrosine-protein kinase Lyn	*Rattus norvegicus*	Tb11.02.0780	Squalene monooxygenase, putative
P52020	Squalene monooxygenase	*Rattus norvegicus*	Tb11.02.0780	Squalene monooxygenase, putative

Supplementary Table 12 Bioactivity values extracted from ChEMBL for the small molecule hits (SH) dataset.

COMPOUND_KEY	PUBLISHED_VALUE(uM)	TARGET_CHEMBLID	PROTEIN_ACCESSION	PREF_NAME	ORGANISM
SID92764752	6.657	CHEMBL4218	P06492	Alpha trans-inducing protein (VP16)	*Herpes simplex virus (type 1 / strain 17)*
SID56463673	0.987	CHEMBL1741207	Q96LD8	Sentrin-specific protease 8	*Homo sapiens*
SID56463673	0.935	CHEMBL1741207	Q96LD8	Sentrin-specific protease 8	*Homo sapiens*
SID87225754	0.125	CHEMBL1075322	Q9Y2T6	G-protein coupled receptor 55	*Homo sapiens*
SID87225754	7.66	CHEMBL1293267	Q9HC97	G-protein coupled receptor 35	*Homo sapiens*
SID17415722	2.534457	CHEMBL1075322	Q9Y2T6	G-protein coupled receptor 55	*Homo sapiens*

COMPOUND_KEY	PUBLISHED_VALUE(uM)	TARGET_CHEMBLID	PROTEIN_ACCESSION	PREF_NAME	ORGANISM
SID857745	1.3	CHEMBL237	P41145	Kappa opioid receptor	*Homo sapiens*
SID856938	2.6	CHEMBL1859	O95180	Voltage-gated T-type calcium channel alpha-1H subunit	*Homo sapiens*
SID49649053	6.84	CHEMBL1741179	P31941	Probable DNA dC->dU-editing enzyme APOBEC-3A	*Homo sapiens*
SID56373536	4.82	CHEMBL5979	P05186	Alkaline phosphatase, tissue-nonspecific isozyme	*Homo sapiens*
SID56373536	2.16	CHEMBL3402	P10696	Alkaline phosphatase placental-like	*Homo sapiens*
SID4244892	3.14	CHEMBL3402	P10696	Alkaline phosphatase placental-like	*Homo sapiens*
SID11532948	2.41	CHEMBL1741208	Q96P20	NACHT, LRR and PYD domains-containing protein 3	*Homo sapiens*
SID56322618	2.226	CHEMBL1741164	O60240	Perilipin-1	*Homo sapiens*
SID17414218	2.32	CHEMBL237	P41145	Kappa opioid receptor	*Homo sapiens*
SID22411930	7.62	CHEMBL1741208	Q96P20	NACHT, LRR and PYD domains-containing protein 3	*Homo sapiens*
SID14737257	6.76	CHEMBL237	P41145	Kappa opioid receptor	*Homo sapiens*
SID49679708	3.179	CHEMBL1293249	Q13887	Kruppel-like factor 5	*Homo sapiens*
SID24781162	1.67	CHEMBL4158	P49327	Fatty acid synthase	*Homo sapiens*
SID24781162	0.859	CHEMBL4158	P49327	Fatty acid synthase	*Homo sapiens*
SID858026	7.12	CHEMBL237	P41145	Kappa opioid receptor	*Homo sapiens*
SID858026	6.831	CHEMBL227	P30556	Type-1 angiotensin II receptor	*Homo sapiens*
SID858026	0.56	CHEMBL5023	Q00987	p53-binding protein Mdm-2	*Homo sapiens*
SID49645303	9.9	CHEMBL1741213	Q9BQF6	Sentrin-specific protease 7	*Homo sapiens*
SID24822843	9.63	CHEMBL5573	P09923	Intestinal alkaline phosphatase	*Homo sapiens*
SID49649021	4.86	CHEMBL237	P41145	Kappa opioid receptor	*Homo sapiens*
SID56373639	3.34	CHEMBL5979	P05186	Alkaline phosphatase, tissue-nonspecific isozyme	*Homo sapiens*
SID92764752	7.073	CHEMBL4374	Q9Y5X4	Photoreceptor-specific nuclear receptor	*Homo sapiens*
SID49827024	4.529	CHEMBL227	P30556	Type-1 angiotensin II receptor	*Homo sapiens*
SID49678979	2.8	CHEMBL237	P41145	Kappa opioid receptor	*Homo sapiens*
SID22412622	1.864	CHEMBL1293249	Q13887	Kruppel-like factor 5	*Homo sapiens*
SID857745	2.761	CHEMBL5313	P38532	Heat shock factor protein 1	*Mus musculus*
SID856938	0.2667	CHEMBL1741219	Q9QUQ5	Short transient receptor potential channel 4	*Mus musculus*
SID17414218	9.825	CHEMBL5313	P38532	Heat shock factor protein 1	*Mus musculus*
SID17403305	3.3493	CHEMBL1741219	Q9QUQ5	Short transient receptor potential channel 4	*Mus musculus*
SID857659	0.668	CHEMBL1741219	Q9QUQ5	Short transient receptor potential channel 4	*Mus musculus*
SID14732424	5.308	CHEMBL1741219	Q9QUQ5	Short transient receptor potential channel 4	*Mus musculus*
SID17432288	6.6827	CHEMBL1741219	Q9QUQ5	Short transient receptor potential channel 4	*Mus musculus*
SID24781162	3.637	CHEMBL5313	P38532	Heat shock factor protein 1	*Mus musculus*
SID24809545	5.308	CHEMBL1741219	Q9QUQ5	Short transient receptor potential channel 4	*Mus musculus*

COMPOUND_KEY	PUBLISHED_VALUE(uM)	TARGET_CHEMBLID	PROTEIN_ACCESSION	PREF_NAME	ORGANISM
SID49649021	5.926	CHEMBL5313	P38532	Heat shock factor protein 1	*Mus musculus*
SID49665200	0.8413	CHEMBL1741219	Q9QUQ5	Short transient receptor potential channel 4	*Mus musculus*
SID24781888	10	CHEMBL2146304	P35639	DNA damage-inducible transcript 3 protein	*Mus musculus*
SID24781162	4.21	CHEMBL2146304	P35639	DNA damage-inducible transcript 3 protein	*Mus musculus*
SID857745	2.82	CHEMBL5567	P08659	Luciferin 4-monooxygenase	*Photinus pyralis*
SID49667183	3.418	CHEMBL5567	P08659	Luciferin 4-monooxygenase	*Photinus pyralis*
SID49728456	3.651	CHEMBL5567	P08659	Luciferin 4-monooxygenase	*Photinus pyralis*
SID49668938	9.27	CHEMBL1741194	P87108	Mitochondrial import inner membrane translocase subunit TIM10	*Saccharomyces cerevisiae S288c*
SID856443	1.98	CHEMBL1741194	P87108	Mitochondrial import inner membrane translocase subunit TIM10	*Saccharomyces cerevisiae S288c*
SID856443	8.24	CHEMBL1741194	P87108	Mitochondrial import inner membrane translocase subunit TIM10	*Saccharomyces cerevisiae S288c*
SID856443	4.44	CHEMBL1741180	P32897	Mitochondrial import inner membrane translocase subunit TIM23	*Saccharomyces cerevisiae S288c*
SID4244892	3.71	CHEMBL1075257	P03070	Large T antigen	*Simian virus 40*
SID17401675	7.45	CHEMBL2146295	P65502	Probable nicotinate-nucleotide adenylyltransferase	*Staphylococcus aureus (strain N315)*

Supplementary Table 13 Enriched targets predicted for the small molecule hits (SH) dataset

PROTEIN_ACCESSION	PREF_NAME	ORGANISM	*Trypanosoma brucei* orthologue	Gene Name
O95180	Voltage-gated T-type calcium channel alpha-1H subunit	*Homo sapiens*	Tb10.70.4750	Calcium channel protein, putative
Q96P20	NACHT, LRR and PYD domains-containing protein 3	*Homo sapiens*	TB927.1.4180	Uncharacterized protein
Q96P20	NACHT, LRR and PYD domains-containing protein 3	*Homo sapiens*	Tb927.7.1430	Putative uncharacterized protein
Q96P20	NACHT, LRR and PYD domains-containing protein 3	*Homo sapiens*	Tb11.02.4230	Putative uncharacterized protein
P87108	Mitochondrial import inner membrane translocase subunit TIM10	*Saccharomyces cerevisiae* S288c	Tb927.7.2200	Putative uncharacterized protein

Supplementary Table 14 Compounds from the SH dataset showing compounds that are predicted to readily cross the BBB (plogBB >0.3) and the targets they are predicted to bind.

Compound	Target	plogBB
CC1(C)CC(CC(C)(C)N1)N1COC2=C(C1)C=C(Cl)C1=CC=CN=C21	Protein kinase_ putative;Tb09.211.2260	1.23
CC1(C)CC(CC(C)(C)N1)N1COC2=C(C1)C=C(Cl)C1=CC=CN=C21	Protein kinase_ putative;Tb10.70.1760	1.23
CC1(C)CC(CC(C)(C)N1)N1COC2=C(C1)C=C(Cl)C1=CC=CN=C21	Serine/threonine-protein kinase_ putative;Tb10.70.5890	1.23
CC1(C)CC(CC(C)(C)N1)N1COC2=C(C1)C=C(Cl)C1=CC=CN=C21	Serine/threonine protein kinase_ putative;Tb927.3.4560	1.23
CC1(C)CC(CC(C)(C)N1)N1COC2=C(C1)C=C(Cl)C1=CC=CN=C21	Protein kinase_ putative;Tb927.8.5730	1.23
CC1(C)CC(CC(C)(C)N1)N1COC2=C(C1)C=C(Cl)C1=CC=CN=C21	Protein kinase_ putative;Tb927.2.2120	1.23
CC1(C)CC(CC(C)(C)N1)N1COC2=C(C1)C=C(Cl)C1=CC=CN=C21	Protein kinase_ putative;Tb927.7.5220	1.23
CC1(C)CC(CC(C)(C)N1)N1COC2=C(C1)C=C(Cl)C1=CC=CN=C21	Protein kinase_ putative;Tb927.3.5650	1.23
CC1(C)CC(CC(C)(C)N1)N1COC2=C(C1)C=C(Cl)C1=CC=CN=C21	Protein kinase_ putative;Tb10.61.2490	1.23
CC1(C)CC(CC(C)(C)N1)N1COC2=C(C1)C=C(Cl)C1=CC=CN=C21	Protein kinase_ putative;Tb10.61.1520	1.23
CC1(C)CC(CC(C)(C)N1)N1COC2=C(C1)C=C(Cl)C1=CC=CN=C21	Protein kinase_ putative;Tb10.6k15.0770	1.23
CN[C@H]1CC[C@@H](C2=CC=C(Cl)C(Cl)=C2)C2=CC=CC=C12	Protein kinase_ putative;Tb927.2.2120	1.14
CN[C@H]1CC[C@@H](C2=CC=C(Cl)C(Cl)=C2)C2=CC=CC=C12	Protein kinase_ putative;Tb927.7.5220	1.14
CN[C@H]1CC[C@@H](C2=CC=C(Cl)C(Cl)=C2)C2=CC=CC=C12	Protein kinase_ putative;Tb927.3.5650	1.14
CN[C@H]1CC[C@@H](C2=CC=C(Cl)C(Cl)=C2)C2=CC=CC=C12	Protein kinase_ putative;Tb10.61.2490	1.14
ClC1=CC(Cl)=C(OCCCCCN2CCNCC2)C=C1	Glucose transporter;Tb10.6k15.2030	0.93
ClC1=CC(Cl)=C(OCCCCCN2CCNCC2)C=C1	Glucose transporter_ putative;Tb927.4.2290	0.93
ClC1=CC(Cl)=C(OCCCCCN2CCNCC2)C=C1	Hexose transporter;Tb10.6k15.2040	0.93
ClC1=CC(Cl)=C(OCCCCCN2CCNCC2)C=C1	Glucose transporter;Tb10.6k15.2020	0.93
CC1=CC(Cl)=CC(Cl)=C1OCCCCCN1CCNCC1	Glucose transporter;Tb10.6k15.2030	0.93
CC1=CC(Cl)=CC(Cl)=C1OCCCCCN1CCNCC1	Glucose transporter_ putative;Tb927.4.2290	0.93
CC1=CC(Cl)=CC(Cl)=C1OCCCCCN1CCNCC1	Hexose transporter;Tb10.6k15.2040	0.93
CC1=CC(Cl)=CC(Cl)=C1OCCCCCN1CCNCC1	Glucose transporter;Tb10.6k15.2020	0.93
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Glucose transporter;Tb10.6k15.2030	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Glucose transporter_ putative;Tb927.4.2290	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Hexose transporter;Tb10.6k15.2040	0.90

CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Glucose transporter;Tb10.6k15.2020	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Flap endonuclease 1;FEN1	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb10.329.0030	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Mitogen-activated protein kinase;Tb927.8.3550	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb927.3.1610	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb11.01.4250	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb11.01.4230	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb927.7.1900	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase;Tb11.01.1030	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Rac serine-threonine kinase_ putative;Tb927.6.2250	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb09.160.0450	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb927.5.2820	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb10.70.0960	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb10.70.0970	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Serine/threonine-protein kinase_ putative;Tb11.01.6650	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Serine/threonine-protein kinase A_ putative;Tb927.8.7110	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Serine/threonine protein kinase_ putative;Tb09.160.1090	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;TB927.1.3130	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Serine/threonine-protein kinase_ putative;Tb927.8.1670	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Putative uncharacterized protein;Tb11.01.1050	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Serine/threonine-protein kinase_ putative;Tb927.8.1690	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb927.7.3580	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Serine/threonine-protein kinase_ putative;Tb10.70.6680	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb11.01.2900	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb927.5.3320	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Serine/threonine-protein kinase A_ putative;Tb927.4.5310	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb927.8.5730	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb09.160.0570	0.90

CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb11.01.0330	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb10.61.1520	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb10.6k15.0770	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Serine/threonine-protein kinase_ putative;Tb927.7.4090	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Serine/threonine-protein kinase_ putative;Tb927.5.1650	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb10.70.1760	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Serine/threonine-protein kinase_ putative;Tb10.70.5890	0.90
CN1CCN(CC1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Serine/threonine protein kinase_ putative;Tb927.3.4560	0.90
ClC1=CC=C2OC3=CC=CC=C3N=C(N3CCNCC3)C2=C1	Protein kinase_ putative;Tb09.211.2260	0.84
CN1CCN(CC1)C1=CC(C)=CC2=C1NC1=C2C=CC=C1	Protein kinase_ putative;Tb09.211.2260	0.83
CN1CCN(CC2=C(O)C3=NC=CC=C3C(Br)=C2)CC1	Glucose transporter;Tb10.6k15.2030	0.82
CN1CCN(CC2=C(O)C3=NC=CC=C3C(Br)=C2)CC1	Glucose transporter_ putative;Tb927.4.2290	0.82
CN1CCN(CC2=C(O)C3=NC=CC=C3C(Br)=C2)CC1	Hexose transporter;Tb10.6k15.2040	0.82
CN1CCN(CC2=C(O)C3=NC=CC=C3C(Br)=C2)CC1	Glucose transporter;Tb10.6k15.2020	0.82
CC1=CC=CC(CN2CCN(CC3=CNC4=CC=CC=C34)CC2)=C1	Protein kinase_ putative;Tb09.211.2260	0.81
CN1C(N(C)C2=CC=CC=C12)C1=CC=C(O)C=C1	Flap endonuclease 1;FEN1	0.80
CN(C)CCCN1C2=CC=CC=C2SC2=CC=CC=C12	Flap endonuclease 1;FEN1	0.75
BrC1=CC=C(C=C1)C1=NC(N2CCNCC2)=C2C=CC=CC2=N1	Glucose transporter;Tb10.6k15.2030	0.73
BrC1=CC=C(C=C1)C1=NC(N2CCNCC2)=C2C=CC=CC2=N1	Glucose transporter_ putative;Tb927.4.2290	0.73
BrC1=CC=C(C=C1)C1=NC(N2CCNCC2)=C2C=CC=CC2=N1	Hexose transporter;Tb10.6k15.2040	0.73
BrC1=CC=C(C=C1)C1=NC(N2CCNCC2)=C2C=CC=CC2=N1	Glucose transporter;Tb10.6k15.2020	0.73
BrC1=CC=C(C=C1)C1=NC(N2CCNCC2)=C2C=CC=CC2=N1	Protein kinase_ putative;Tb927.7.1900	0.73
BrC1=CC=C(C=C1)C1=NC(N2CCNCC2)=C2C=CC=CC2=N1	Protein kinase_ putative;Tb09.211.2260	0.73
CC1=CC(Br)=C(OCCCCCCN2CCNCC2)C(C)=C1	Glucose transporter;Tb10.6k15.2030	0.69
CC1=CC(Br)=C(OCCCCCCN2CCNCC2)C(C)=C1	Glucose transporter_ putative;Tb927.4.2290	0.69
CC1=CC(Br)=C(OCCCCCCN2CCNCC2)C(C)=C1	Hexose transporter;Tb10.6k15.2040	0.69
CC1=CC(Br)=C(OCCCCCCN2CCNCC2)C(C)=C1	Glucose transporter;Tb10.6k15.2020	0.69
NC1C(CC2=C(Cl)C=C(Cl)C=C2)CC2=CC=CC=C12	NADPH--cytochrome p450 reductase_ putative;Tb09.211.4110	0.67

NC1C(CC2=C(Cl)C=C(Cl)C=C2)CC2=CC=CC=C12	NADPH--cytochrome P450 reductase_ putative;Tb11.01.0170	0.67
NC1C(CC2=C(Cl)C=C(Cl)C=C2)CC2=CC=CC=C12	NADPH--cytochrome p450 reductase_ putative;Tb11.02.5420	0.67
CCC1=C(Cl)C=CC(OCCCCCN2CCNCC2)=C1	Glucose transporter;Tb10.6k15.2030	0.66
CCC1=C(Cl)C=CC(OCCCCCN2CCNCC2)=C1	Glucose transporter_ putative;Tb927.4.2290	0.66
CCC1=C(Cl)C=CC(OCCCCCN2CCNCC2)=C1	Hexose transporter;Tb10.6k15.2040	0.66
CCC1=C(Cl)C=CC(OCCCCCN2CCNCC2)=C1	Glucose transporter;Tb10.6k15.2020	0.66
ClC1=C(OCCCCNCC=C)C=CC(Br)=C1	Glucose transporter;Tb10.6k15.2030	0.66
ClC1=C(OCCCCNCC=C)C=CC(Br)=C1	Glucose transporter_ putative;Tb927.4.2290	0.66
ClC1=C(OCCCCNCC=C)C=CC(Br)=C1	Hexose transporter;Tb10.6k15.2040	0.66
ClC1=C(OCCCCNCC=C)C=CC(Br)=C1	Glucose transporter;Tb10.6k15.2020	0.66
CNC(C)(C)CC1=C2CCCC2=CC2=C1CCC2	Glucose transporter;Tb10.6k15.2030	0.65
CNC(C)(C)CC1=C2CCCC2=CC2=C1CCC2	Glucose transporter_ putative;Tb927.4.2290	0.65
CNC(C)(C)CC1=C2CCCC2=CC2=C1CCC2	Hexose transporter;Tb10.6k15.2040	0.65
CNC(C)(C)CC1=C2CCCC2=CC2=C1CCC2	Glucose transporter;Tb10.6k15.2020	0.65
CC(C)(C)C(N)CCC1=C(Cl)C=C(Cl)C=C1	Glucose transporter;Tb10.6k15.2030	0.62
CC(C)(C)C(N)CCC1=C(Cl)C=C(Cl)C=C1	Glucose transporter_ putative;Tb927.4.2290	0.62
CC(C)(C)C(N)CCC1=C(Cl)C=C(Cl)C=C1	Hexose transporter;Tb10.6k15.2040	0.62
CC(C)(C)C(N)CCC1=C(Cl)C=C(Cl)C=C1	Glucose transporter;Tb10.6k15.2020	0.62
CN(C)CCCNCCCC1=CC=C(C=C1)C(C)(C)C	Glucose transporter;Tb10.6k15.2030	0.59
CN(C)CCCNCCCC1=CC=C(C=C1)C(C)(C)C	Glucose transporter_ putative;Tb927.4.2290	0.59
CN(C)CCCNCCCC1=CC=C(C=C1)C(C)(C)C	Hexose transporter;Tb10.6k15.2040	0.59
CN(C)CCCNCCCC1=CC=C(C=C1)C(C)(C)C	Glucose transporter;Tb10.6k15.2020	0.59
CC1=CC=CC2=CC(C3=CC=CC=C3)=C(N=C12)N1CCNCC1	Glucose transporter;Tb10.6k15.2030	0.57
CC1=CC=CC2=CC(C3=CC=CC=C3)=C(N=C12)N1CCNCC1	Glucose transporter_ putative;Tb927.4.2290	0.57
CC1=CC=CC2=CC(C3=CC=CC=C3)=C(N=C12)N1CCNCC1	Hexose transporter;Tb10.6k15.2040	0.57
CC1=CC=CC2=CC(C3=CC=CC=C3)=C(N=C12)N1CCNCC1	Glucose transporter;Tb10.6k15.2020	0.57
CC1=CC=CC2=CC(C3=CC=CC=C3)=C(N=C12)N1CCNCC1	Protein kinase_ putative;Tb09.211.2260	0.57
CC1=CC=CC2=CC(C3=CC=CC=C3)=C(N=C12)N1CCNCC1	Protein kinase_ putative;Tb927.8.5730	0.57
CC1=CC=CC2=CC(C3=CC=CC=C3)=C(N=C12)N1CCNCC1	Protein kinase_ putative;Tb10.61.1520	0.57
CC1=CC=CC2=CC(C3=CC=CC=C3)=C(N=C12)N1CCNCC1	Protein kinase_ putative;Tb10.6k15.0770	0.57
CC1=CC=CC2=CC(C3=CC=CC=C3)=C(N=C12)N1CCNCC1	Protein kinase_ putative;Tb10.70.1760	0.57

CC1=CC=CC2=CC(C3=CC=CC=C3)=C(N=C12)N1CCNCC1	Serine/threonine-protein kinase_ putative;Tb10.70.5890	0.57
CC1=CC=CC2=CC(C3=CC=CC=C3)=C(N=C12)N1CCNCC1	Serine/threonine protein kinase_ putative;Tb927.3.4560	0.57
ClC1=CC(Cl)=C(CCNCC2=CNC3=CC=CC=C23)C=C1	Glucose transporter;Tb10.6k15.2030	0.57
ClC1=CC(Cl)=C(CCNCC2=CNC3=CC=CC=C23)C=C1	Glucose transporter_ putative;Tb927.4.2290	0.57
ClC1=CC(Cl)=C(CCNCC2=CNC3=CC=CC=C23)C=C1	Hexose transporter;Tb10.6k15.2040	0.57
ClC1=CC(Cl)=C(CCNCC2=CNC3=CC=CC=C23)C=C1	Glucose transporter;Tb10.6k15.2020	0.57
CN1C2=NC3=CC=CC=C3C2=C(C)C=C1Cl	Glucose transporter;Tb10.6k15.2030	0.57
CN1C2=NC3=CC=CC=C3C2=C(C)C=C1Cl	Glucose transporter_ putative;Tb927.4.2290	0.57
CN1C2=NC3=CC=CC=C3C2=C(C)C=C1Cl	Hexose transporter;Tb10.6k15.2040	0.57
CN1C2=NC3=CC=CC=C3C2=C(C)C=C1Cl	Glucose transporter;Tb10.6k15.2020	0.57
CN1C2=NC3=CC=CC=C3C2=C(C)C=C1Cl	Flap endonuclease 1;FEN1	0.57
CN1C2=NC3=CC=CC=C3C2=C(C)C=C1Cl	Protein kinase_ putative;Tb10.70.1760	0.57
CN1C2=NC3=CC=CC=C3C2=C(C)C=C1Cl	Serine/threonine-protein kinase_ putative;Tb10.70.5890	0.57
CN1C2=NC3=CC=CC=C3C2=C(C)C=C1Cl	Serine/threonine protein kinase_ putative;Tb927.3.4560	0.57
CCCCNCC1=CC(Br)=C(OCC2=CC=C(F)C=C2)C(OC)=C1	Glucose transporter;Tb10.6k15.2030	0.56
CCCCNCC1=CC(Br)=C(OCC2=CC=C(F)C=C2)C(OC)=C1	Glucose transporter_ putative;Tb927.4.2290	0.56
CCCCNCC1=CC(Br)=C(OCC2=CC=C(F)C=C2)C(OC)=C1	Hexose transporter;Tb10.6k15.2040	0.56
CCCCNCC1=CC(Br)=C(OCC2=CC=C(F)C=C2)C(OC)=C1	Glucose transporter;Tb10.6k15.2020	0.56
CNCCCC1SC2=C(C=CC=C2)C(C)C2=C1C=CC=C2	Glucose transporter;Tb10.6k15.2030	0.55
CNCCCC1SC2=C(C=CC=C2)C(C)C2=C1C=CC=C2	Glucose transporter_ putative;Tb927.4.2290	0.55
CNCCCC1SC2=C(C=CC=C2)C(C)C2=C1C=CC=C2	Hexose transporter;Tb10.6k15.2040	0.55
CNCCCC1SC2=C(C=CC=C2)C(C)C2=C1C=CC=C2	Glucose transporter;Tb10.6k15.2020	0.55
CC1=CC(NCCN2CCCCC2)=C2C=CC3=CC=CC=C3C2=N1	Glucose transporter;Tb10.6k15.2030	0.53
CC1=CC(NCCN2CCCCC2)=C2C=CC3=CC=CC=C3C2=N1	Glucose transporter_ putative;Tb927.4.2290	0.53
CC1=CC(NCCN2CCCCC2)=C2C=CC3=CC=CC=C3C2=N1	Hexose transporter;Tb10.6k15.2040	0.53
CC1=CC(NCCN2CCCCC2)=C2C=CC3=CC=CC=C3C2=N1	Glucose transporter;Tb10.6k15.2020	0.53
CC1=CC(NCCN2CCCCC2)=C2C=CC3=CC=CC=C3C2=N1	Flap endonuclease 1;FEN1	0.53
CC(C)C1NC(CC(=N1)C1=CC2=C(OCO2)C=C1)C1=C(O)C=CC(Cl)=C1	NADPH--cytochrome p450 reductase_ putative;Tb09.211.4110	0.53
CC(C)C1NC(CC(=N1)C1=CC2=C(OCO2)C=C1)C1=C(O)C=CC(Cl)=C1	NADPH--cytochrome P450 reductase_ putative;Tb11.01.0170	0.53
CC(C)C1NC(CC(=N1)C1=CC2=C(OCO2)C=C1)C1=C(O)C=CC(Cl)=C1	NADPH--cytochrome p450 reductase_ putative;Tb11.02.5420	0.53

CC1=CC(NCCN2CCOCC2)=C2C=CC3=CC=CC=C3C2=N1	Glucose transporter;Tb10.6k15.2030	0.53
CC1=CC(NCCN2CCOCC2)=C2C=CC3=CC=CC=C3C2=N1	Glucose transporter_ putative;Tb927.4.2290	0.53
CC1=CC(NCCN2CCOCC2)=C2C=CC3=CC=CC=C3C2=N1	Hexose transporter;Tb10.6k15.2040	0.53
CC1=CC(NCCN2CCOCC2)=C2C=CC3=CC=CC=C3C2=N1	Glucose transporter;Tb10.6k15.2020	0.53
CCOC1=C(OC)C=CC(CNCCC2=CC=CC(Cl)=C2)=C1	Glucose transporter;Tb10.6k15.2030	0.53
CCOC1=C(OC)C=CC(CNCCC2=CC=CC(Cl)=C2)=C1	Glucose transporter_ putative;Tb927.4.2290	0.53
CCOC1=C(OC)C=CC(CNCCC2=CC=CC(Cl)=C2)=C1	Hexose transporter;Tb10.6k15.2040	0.53
CCOC1=C(OC)C=CC(CNCCC2=CC=CC(Cl)=C2)=C1	Glucose transporter;Tb10.6k15.2020	0.53
CCCCNCC1=C(Cl)C2=CC=CC=C2N1	Glucose transporter;Tb10.6k15.2030	0.52
CCCCNCC1=C(Cl)C2=CC=CC=C2N1	Glucose transporter_ putative;Tb927.4.2290	0.52
CCCCNCC1=C(Cl)C2=CC=CC=C2N1	Hexose transporter;Tb10.6k15.2040	0.52
CCCCNCC1=C(Cl)C2=CC=CC=C2N1	Glucose transporter;Tb10.6k15.2020	0.52
COC1=CC2=C3CNCC3=C(C)N=C2C(OC)=C1	Protein kinase_ putative;Tb10.70.1760	0.51
COC1=CC2=C3CNCC3=C(C)N=C2C(OC)=C1	Serine/threonine-protein kinase_ putative;Tb10.70.5890	0.51
COC1=CC2=C3CNCC3=C(C)N=C2C(OC)=C1	Serine/threonine protein kinase_ putative;Tb927.3.4560	0.51
NC1CCCC2=C1NC1=CC=C(Cl)C=C21	NADPH--cytochrome p450 reductase_ putative;Tb09.211.4110	0.46
NC1CCCC2=C1NC1=CC=C(Cl)C=C21	NADPH--cytochrome P450 reductase_ putative;Tb11.01.0170	0.46
NC1CCCC2=C1NC1=CC=C(Cl)C=C21	NADPH--cytochrome p450 reductase_ putative;Tb11.02.5420	0.46
CC(C)(C)C1=CC=C(OCCCCCCN2CCNCC2)C=C1	Glucose transporter;Tb10.6k15.2030	0.45
CC(C)(C)C1=CC=C(OCCCCCCN2CCNCC2)C=C1	Glucose transporter_ putative;Tb927.4.2290	0.45
CC(C)(C)C1=CC=C(OCCCCCCN2CCNCC2)C=C1	Hexose transporter;Tb10.6k15.2040	0.45
CC(C)(C)C1=CC=C(OCCCCCCN2CCNCC2)C=C1	Glucose transporter;Tb10.6k15.2020	0.45
OC1=C2N=CC=CC2=C(CN2CCOCC2)C=C1Br	Flap endonuclease 1;FEN1	0.44
CN1C2=C(C=CC=C2)C2=C1C=C1C=C(C)N=C(C)C1=C2	Protein kinase_ putative;Tb927.2.2120	0.44
CN1C2=C(C=CC=C2)C2=C1C=C1C=C(C)N=C(C)C1=C2	Protein kinase_ putative;Tb927.7.5220	0.44
CN1C2=C(C=CC=C2)C2=C1C=C1C=C(C)N=C(C)C1=C2	Protein kinase_ putative;Tb927.3.5650	0.44
CN1C2=C(C=CC=C2)C2=C1C=C1C=C(C)N=C(C)C1=C2	Protein kinase_ putative;Tb10.61.2490	0.44
CN1C2=C(C=CC=C2)C2=C1C=C1C=C(C)N=C(C)C1=C2	Protein kinase_ putative;Tb10.70.1760	0.44
CN1C2=C(C=CC=C2)C2=C1C=C1C=C(C)N=C(C)C1=C2	Serine/threonine-protein kinase_ putative;Tb10.70.5890	0.44

CN1C2=C(C=CC=C2)C2=C1C=C1C=C(C)N=C(C)C1=C2	Serine/threonine protein kinase_ putative;Tb927.3.4560	0.44
CN1C2=C(C=CC=C2)C2=C1C=C1C=C(C)N=C(C)C1=C2	Protein kinase_ putative;Tb09.211.2260	0.44
NC1=C(SC=C1)C=CC1=CC=CS1	NADPH--cytochrome p450 reductase_ putative;Tb09.211.4110	0.44
NC1=C(SC=C1)C=CC1=CC=CS1	NADPH--cytochrome P450 reductase_ putative;Tb11.01.0170	0.44
NC1=C(SC=C1)C=CC1=CC=CS1	NADPH--cytochrome p450 reductase_ putative;Tb11.02.5420	0.44
OC1=C2N=CC=CC2=C(CN2CCOCC2)C=C1Cl	Flap endonuclease 1;FEN1	0.42
CCN1CCCC(C1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Glucose transporter;Tb10.6k15.2030	0.42
CCN1CCCC(C1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Glucose transporter_ putative;Tb927.4.2290	0.42
CCN1CCCC(C1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Hexose transporter;Tb10.6k15.2040	0.42
CCN1CCCC(C1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Glucose transporter;Tb10.6k15.2020	0.42
CCN1CCCC(C1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb10.329.0030	0.42
CCN1CCCC(C1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Mitogen-activated protein kinase;Tb927.8.3550	0.42
CCN1CCCC(C1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb09.160.0570	0.42
CCN1CCCC(C1)NC1=C2C=C3C=CC=CC3=CC2=NC=C1	Protein kinase_ putative;Tb11.01.0330	0.42
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Glucose transporter;Tb10.6k15.2030	0.41
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Glucose transporter_ putative;Tb927.4.2290	0.41
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Hexose transporter;Tb10.6k15.2040	0.41
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Glucose transporter;Tb10.6k15.2020	0.41
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Protein kinase_ putative;Tb927.7.1900	0.41
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Protein kinase_ putative;Tb09.160.0570	0.41
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Protein kinase_ putative;Tb11.01.0330	0.41
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Protein kinase_ putative;Tb10.329.0030	0.41
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Mitogen-activated protein kinase;Tb927.8.3550	0.41
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Cell-division control protein 2 homolog 6_ putative;Tb11.47.0031	0.41
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Cell division control protein 2 homolog 2;Tb927.7.7360	0.41
CN1CCN(CC1)C1=NC(NC2=CC=C(F)C=C2)=NC(NC2=CC=C(F)C=C2)=N1	Cell division related protein kinase 2_ putative;Tb10.70.2210	0.41
CC1CCN(CC2=C3C=CC=NC3=C(O)C(Br)=C2)CC1	Glucose transporter;Tb10.6k15.2030	0.40
CC1CCN(CC2=C3C=CC=NC3=C(O)C(Br)=C2)CC1	Glucose transporter_ putative;Tb927.4.2290	0.40

CC1CCN(CC2=C3C=CC=NC3=C(O)C(Br)=C2)CC1	Hexose transporter;Tb10.6k15.2040	0.40
CC1CCN(CC2=C3C=CC=NC3=C(O)C(Br)=C2)CC1	Glucose transporter;Tb10.6k15.2020	0.40
OC1=C2N=CC=CC2=C(CN2CCCC2)C=C1Cl	Glucose transporter;Tb10.6k15.2030	0.40
OC1=C2N=CC=CC2=C(CN2CCCC2)C=C1Cl	Glucose transporter_ putative;Tb927.4.2290	0.40
OC1=C2N=CC=CC2=C(CN2CCCC2)C=C1Cl	Hexose transporter;Tb10.6k15.2040	0.40
OC1=C2N=CC=CC2=C(CN2CCCC2)C=C1Cl	Glucose transporter;Tb10.6k15.2020	0.40
OC1=C2N=CC=CC2=C(CN2CCCC2)C=C1Cl	Flap endonuclease 1;FEN1	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	Glucose transporter;Tb10.6k15.2030	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	Glucose transporter_ putative;Tb927.4.2290	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	Hexose transporter;Tb10.6k15.2040	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	Glucose transporter;Tb10.6k15.2020	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	NADPH--cytochrome p450 reductase_ putative;Tb09.211.4110	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	NADPH--cytochrome P450 reductase_ putative;Tb11.01.0170	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	NADPH--cytochrome p450 reductase_ putative;Tb11.02.5420	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	Flap endonuclease 1;FEN1	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	Protein kinase_ putative;Tb927.2.2120	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	Protein kinase_ putative;Tb927.7.5220	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	Protein kinase_ putative;Tb927.3.5650	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	Protein kinase_ putative;Tb10.61.2490	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	Serine/threonine-protein kinase_ putative;Tb927.7.4090	0.40
CC1=CC=C2C=CC3=CC=C(C)N=C3C2=N1	Serine/threonine-protein kinase_ putative;Tb927.5.1650	0.40
CC(NC1=C2C=CC=CC2=NC(=N1)N1CCNCC1)C1=CC=CC=C1	Protein kinase_ putative;Tb927.7.1900	0.40
CC(NC1=C2C=CC=CC2=NC(=N1)N1CCNCC1)C1=CC=CC=C1	Protein kinase_ putative;Tb09.211.2260	0.40
CC(NC1=C2C=CC=CC2=NC(=N1)N1CCNCC1)C1=CC=CC=C1	Protein kinase_ putative;Tb10.70.1760	0.40
CC(NC1=C2C=CC=CC2=NC(=N1)N1CCNCC1)C1=CC=CC=C1	Serine/threonine-protein kinase_ putative;Tb10.70.5890	0.40
CC(NC1=C2C=CC=CC2=NC(=N1)N1CCNCC1)C1=CC=CC=C1	Serine/threonine protein kinase_ putative;Tb927.3.4560	0.40
COC1=C(OCCCCCN2CCNCC2)C=CC(CC=C)=C1	Glucose transporter;Tb10.6k15.2030	0.39
COC1=C(OCCCCCN2CCNCC2)C=CC(CC=C)=C1	Glucose transporter_ putative;Tb927.4.2290	0.39
COC1=C(OCCCCCN2CCNCC2)C=CC(CC=C)=C1	Hexose transporter;Tb10.6k15.2040	0.39
COC1=C(OCCCCCN2CCNCC2)C=CC(CC=C)=C1	Glucose transporter;Tb10.6k15.2020	0.39
CN(C)CCCNC1=C2C=CC=CC2=NC(=N1)C1=CC=C(Cl)C=C1	Glucose transporter;Tb10.6k15.2030	0.38

CN(C)CCCNC1=C2C=CC=CC2=NC(=N1)C1=CC=C(Cl)C=C1	Glucose transporter_ putative;Tb927.4.2290	0.38
CN(C)CCCNC1=C2C=CC=CC2=NC(=N1)C1=CC=C(Cl)C=C1	Hexose transporter;Tb10.6k15.2040	0.38
CN(C)CCCNC1=C2C=CC=CC2=NC(=N1)C1=CC=C(Cl)C=C1	Glucose transporter;Tb10.6k15.2020	0.38
NCCNCCOC(C1=CC=C(Cl)C=C1)C1=CC=C(Cl)C=C1	Glucose transporter;Tb10.6k15.2030	0.35
NCCNCCOC(C1=CC=C(Cl)C=C1)C1=CC=C(Cl)C=C1	Glucose transporter_ putative;Tb927.4.2290	0.35
NCCNCCOC(C1=CC=C(Cl)C=C1)C1=CC=C(Cl)C=C1	Hexose transporter;Tb10.6k15.2040	0.35
NCCNCCOC(C1=CC=C(Cl)C=C1)C1=CC=C(Cl)C=C1	Glucose transporter;Tb10.6k15.2020	0.35
NCCNCCOC(C1=CC=C(Cl)C=C1)C1=CC=C(Cl)C=C1	NADPH--cytochrome p450 reductase_ putative;Tb09.211.4110	0.35
NCCNCCOC(C1=CC=C(Cl)C=C1)C1=CC=C(Cl)C=C1	NADPH--cytochrome P450 reductase_ putative;Tb11.01.0170	0.35
NCCNCCOC(C1=CC=C(Cl)C=C1)C1=CC=C(Cl)C=C1	NADPH--cytochrome p450 reductase_ putative;Tb11.02.5420	0.35
CC1=CC2=C(NC3=CC4=C(OCCO4)C=C3)C=C(C)N=C2C(C)=C1	Glucose transporter;Tb10.6k15.2030	0.34
CC1=CC2=C(NC3=CC4=C(OCCO4)C=C3)C=C(C)N=C2C(C)=C1	Glucose transporter_ putative;Tb927.4.2290	0.34
CC1=CC2=C(NC3=CC4=C(OCCO4)C=C3)C=C(C)N=C2C(C)=C1	Hexose transporter;Tb10.6k15.2040	0.34
CC1=CC2=C(NC3=CC4=C(OCCO4)C=C3)C=C(C)N=C2C(C)=C1	Glucose transporter;Tb10.6k15.2020	0.34
COC1=C(C=CC=C1)N1CCN(CC2=C(O)C3=NC=CC=C3C(Cl)=C2)CC1	Glucose transporter;Tb10.6k15.2030	0.34
COC1=C(C=CC=C1)N1CCN(CC2=C(O)C3=NC=CC=C3C(Cl)=C2)CC1	Glucose transporter_ putative;Tb927.4.2290	0.34
COC1=C(C=CC=C1)N1CCN(CC2=C(O)C3=NC=CC=C3C(Cl)=C2)CC1	Hexose transporter;Tb10.6k15.2040	0.34
COC1=C(C=CC=C1)N1CCN(CC2=C(O)C3=NC=CC=C3C(Cl)=C2)CC1	Glucose transporter;Tb10.6k15.2020	0.34
CC1=CC(Cl)=C2C(N)=C(C)C=C(C)C2=N1	Glucose transporter;Tb10.6k15.2030	0.32
CC1=CC(Cl)=C2C(N)=C(C)C=C(C)C2=N1	Glucose transporter_ putative;Tb927.4.2290	0.32
CC1=CC(Cl)=C2C(N)=C(C)C=C(C)C2=N1	Hexose transporter;Tb10.6k15.2040	0.32
CC1=CC(Cl)=C2C(N)=C(C)C=C(C)C2=N1	Glucose transporter;Tb10.6k15.2020	0.32
CC1=CC(Cl)=C2C(N)=C(C)C=C(C)C2=N1	NADPH--cytochrome p450 reductase_ putative;Tb09.211.4110	0.32
CC1=CC(Cl)=C2C(N)=C(C)C=C(C)C2=N1	NADPH--cytochrome P450 reductase_ putative;Tb11.01.0170	0.32
CC1=CC(Cl)=C2C(N)=C(C)C=C(C)C2=N1	NADPH--cytochrome p450 reductase_ putative;Tb11.02.5420	0.32
CC1=CC(Cl)=C2C(N)=C(C)C=C(C)C2=N1	Flap endonuclease 1;FEN1	0.32
OC1=C(CN2CCCCC2)C=CC2=CC=CN=C12	Glucose transporter;Tb10.6k15.2030	0.32
OC1=C(CN2CCCCC2)C=CC2=CC=CN=C12	Glucose transporter_ putative;Tb927.4.2290	0.32
OC1=C(CN2CCCCC2)C=CC2=CC=CN=C12	Hexose transporter;Tb10.6k15.2040	0.32
OC1=C(CN2CCCCC2)C=CC2=CC=CN=C12	Glucose transporter;Tb10.6k15.2020	0.32

OC1=C(CN2CCCCC2)C=CC2=CC=CN=C12	Protein kinase_ putative;Tb927.7.1900	0.32
CCN(CC)CCCC(C)NC1=C2C=CC(Cl)=CC2=NC=C1	Glucose transporter;Tb10.6k15.2030	0.31
CCN(CC)CCCC(C)NC1=C2C=CC(Cl)=CC2=NC=C1	Glucose transporter_ putative;Tb927.4.2290	0.31
CCN(CC)CCCC(C)NC1=C2C=CC(Cl)=CC2=NC=C1	Hexose transporter;Tb10.6k15.2040	0.31
CCN(CC)CCCC(C)NC1=C2C=CC(Cl)=CC2=NC=C1	Glucose transporter;Tb10.6k15.2020	0.31
NCC(O)(C1=CC=C(Cl)C=C1)C1=CC=C(Cl)C=C1	Glucose transporter;Tb10.6k15.2030	0.31
NCC(O)(C1=CC=C(Cl)C=C1)C1=CC=C(Cl)C=C1	Glucose transporter_ putative;Tb927.4.2290	0.31
NCC(O)(C1=CC=C(Cl)C=C1)C1=CC=C(Cl)C=C1	Hexose transporter;Tb10.6k15.2040	0.31
NCC(O)(C1=CC=C(Cl)C=C1)C1=CC=C(Cl)C=C1	Glucose transporter;Tb10.6k15.2020	0.31
CC1=C(C=CC=C1)N1SC2=CC=CC=C2C1=O	Flap endonuclease 1;FEN1	0.31
NCC(C1=CC=C(Cl)C=C1)C1=CC=C(C=C1)C1=CC=CC=C1	Glucose transporter;Tb10.6k15.2030	0.31
NCC(C1=CC=C(Cl)C=C1)C1=CC=C(C=C1)C1=CC=CC=C1	Glucose transporter_ putative;Tb927.4.2290	0.31
NCC(C1=CC=C(Cl)C=C1)C1=CC=C(C=C1)C1=CC=CC=C1	Hexose transporter;Tb10.6k15.2040	0.31
NCC(C1=CC=C(Cl)C=C1)C1=CC=C(C=C1)C1=CC=CC=C1	Glucose transporter;Tb10.6k15.2020	0.31
ClC1=CC(Cl)=C(CN2C3=C(CCC3)C(=N)C3=C2CCCC3)C=C1	Glucose transporter;Tb10.6k15.2030	0.30
ClC1=CC(Cl)=C(CN2C3=C(CCC3)C(=N)C3=C2CCCC3)C=C1	Glucose transporter_ putative;Tb927.4.2290	0.30
ClC1=CC(Cl)=C(CN2C3=C(CCC3)C(=N)C3=C2CCCC3)C=C1	Hexose transporter;Tb10.6k15.2040	0.30
ClC1=CC(Cl)=C(CN2C3=C(CCC3)C(=N)C3=C2CCCC3)C=C1	Glucose transporter;Tb10.6k15.2020	0.30

Supplementary Table 15 P-values of GO biological process fold enrichment of small molecule hits that are represented in Figure 4:6

ID	Name	P-value
GO:0006468	protein phosphorylation	2.89E-65
GO:0016310	phosphorylation	7.30E-60
GO:0036211	protein modification process	1.68E-48
GO:0006464	cellular protein modification process	1.68E-48
GO:0043412	macromolecule modification	2.34E-44
GO:0006796	phosphate-containing compound metabolic process	4.97E-43
GO:0006793	phosphorus metabolic process	1.11E-42
GO:0044267	cellular protein metabolic process	3.59E-35
GO:0019538	protein metabolic process	3.19E-32
GO:0044260	cellular macromolecule metabolic process	3.46E-31

GO:1901564	organonitrogen compound metabolic process	1.95E-28
GO:0043170	macromolecule metabolic process	8.74E-24
GO:0006807	nitrogen compound metabolic process	2.47E-20
GO:0044237	cellular metabolic process	1.24E-19
GO:0008152	metabolic process	4.01E-19
GO:0044238	primary metabolic process	5.43E-19
GO:0071704	organic substance metabolic process	2.91E-18
GO:0009987	cellular process	1.57E-16
GO:0008150	biological process	5.26E-14
GO:0044145	modulation of development of symbiont involved in interaction with host	2.35E-07
GO:0043900	regulation of multi-organism process	2.35E-07
GO:0043903	regulation of symbiosis, encompassing mutualism through parasitism	2.35E-07
GO:0050793	regulation of developmental process	5.67E-07
GO:0051726	regulation of cell cycle	1.60E-06
GO:0007346	regulation of mitotic cell cycle	5.44E-06
GO:0000278	mitotic cell cycle	7.31E-06
GO:0010564	regulation of cell cycle process	1.51E-05
GO:0015758	glucose transport	6.75E-05
GO:0052106	quorum sensing involved in interaction with host	9.96E-05
GO:0052097	interspecies quorum sensing	9.96E-05
GO:0007049	cell cycle	1.04E-04
GO:0048874	homeostasis of number of cells in a free-living population	1.14E-04
GO:0048872	homeostasis of number of cells	1.14E-04
GO:0009372	quorum sensing	1.14E-04
GO:0044764	multi-organism cellular process	1.14E-04
GO:1903047	mitotic cell cycle process	2.83E-04
GO:0044772	mitotic cell cycle phase transition	4.01E-04
GO:0044770	cell cycle phase transition	6.65E-04
GO:0051225	spindle assembly	1.38E-03
GO:0022402	cell cycle process	1.98E-03
GO:1902412	regulation of mitotic cytokinesis	2.34E-03
GO:0071900	regulation of protein serine/threonine kinase activity	2.91E-03
GO:0043549	regulation of kinase activity	2.91E-03
GO:0051338	regulation of transferase activity	2.91E-03
GO:0045859	regulation of protein kinase activity	2.91E-03
GO:0001932	regulation of protein phosphorylation	3.54E-03
GO:0031399	regulation of protein modification process	4.23E-03
GO:0032465	regulation of cytokinesis	4.23E-03
GO:0051302	regulation of cell division	4.23E-03
GO:0042325	regulation of phosphorylation	4.23E-03
GO:0007051	spindle organization	4.97E-03

GO:0051174	regulation of phosphorus metabolic process	4.97E-03
GO:0019220	regulation of phosphate metabolic process	4.97E-03
GO:0042592	homeostatic process	5.46E-03
GO:1901988	negative regulation of cell cycle phase transition	8.28E-03
GO:0042327	positive regulation of phosphorylation	8.28E-03
GO:0043410	positive regulation of MAPK cascade	8.28E-03
GO:2000045	regulation of G1/S transition of mitotic cell cycle	8.28E-03
GO:0000082	G1/S transition of mitotic cell cycle	8.28E-03
GO:0009967	positive regulation of signal transduction	8.28E-03
GO:1902806	regulation of cell cycle G1/S phase transition	8.28E-03
GO:0033674	positive regulation of kinase activity	8.28E-03
GO:0043406	positive regulation of MAP kinase activity	8.28E-03
GO:0023056	positive regulation of signaling	8.28E-03
GO:0071902	positive regulation of protein serine/threonine kinase activity	8.28E-03
GO:0015749	monosaccharide transmembrane transport	8.28E-03
GO:0034219	carbohydrate transmembrane transport	8.28E-03
GO:0044131	negative regulation of development of symbiont in host	8.28E-03
GO:1902807	negative regulation of cell cycle G1/S phase transition	8.28E-03
GO:0001934	positive regulation of protein phosphorylation	8.28E-03
GO:2000134	negative regulation of G1/S transition of mitotic cell cycle	8.28E-03
GO:0044147	negative regulation of development of symbiont involved in interaction with host	8.28E-03
GO:0010647	positive regulation of cell communication	8.28E-03
GO:1902533	positive regulation of intracellular signal transduction	8.28E-03
GO:0051093	negative regulation of developmental process	8.28E-03
GO:0010562	positive regulation of phosphorus metabolic process	8.28E-03
GO:0045860	positive regulation of protein kinase activity	8.28E-03
GO:0031401	positive regulation of protein modification process	8.28E-03
GO:0048584	positive regulation of response to stimulus	8.28E-03
GO:0044127	regulation of development of symbiont in host	8.28E-03
GO:0051347	positive regulation of transferase activity	8.28E-03
GO:0045937	positive regulation of phosphate metabolic process	8.28E-03
GO:0090307	mitotic spindle assembly	8.28E-03
GO:0044843	cell cycle G1/S phase transition	8.28E-03
GO:1901991	negative regulation of mitotic cell cycle phase transition	8.28E-03
GO:0043405	regulation of MAP kinase activity	8.28E-03
GO:0043408	regulation of MAPK cascade	8.28E-03
GO:0045926	negative regulation of growth	8.28E-03
GO:0043901	negative regulation of multi-organism process	8.28E-03
GO:0008645	hexose transmembrane transport	8.28E-03
GO:0000281	mitotic cytokinesis	1.40E-02
GO:0061640	cytoskeleton-dependent cytokinesis	1.40E-02

GO:0043085	positive regulation of catalytic activity	1.65E-02
GO:1901990	regulation of mitotic cell cycle phase transition	1.65E-02
GO:0044093	positive regulation of molecular function	1.65E-02
GO:1901987	regulation of cell cycle phase transition	1.65E-02
GO:0051821	dissemination or transmission of organism from other organism involved in symbiotic interaction	1.65E-02
GO:0044008	dissemination or transmission of symbiont from host by vector	1.65E-02
GO:0045930	negative regulation of mitotic cell cycle	1.65E-02
GO:0051822	dissemination or transmission of organism from other organism by vector involved in symbiotic interaction	1.65E-02
GO:0044007	dissemination or transmission of symbiont from host	1.65E-02
GO:0000910	cytokinesis	2.22E-02
GO:0032268	regulation of cellular protein metabolic process	2.22E-02
GO:0018105	peptidyl-serine phosphorylation	2.47E-02
GO:0018209	peptidyl-serine modification	2.47E-02
GO:0035404	histone-serine phosphorylation	2.47E-02
GO:0032270	positive regulation of cellular protein metabolic process	2.47E-02
GO:0043987	histone H3-S10 phosphorylation	2.47E-02
GO:0051247	positive regulation of protein metabolic process	2.47E-02
GO:0010948	negative regulation of cell cycle process	2.47E-02
GO:0044839	cell cycle G2/M phase transition	2.47E-02
GO:0000086	G2/M transition of mitotic cell cycle	2.47E-02
GO:0016572	histone phosphorylation	2.47E-02
GO:0050794	regulation of cellular process	2.48E-02
GO:0065008	regulation of biological quality	2.52E-02
GO:0051301	cell division	2.86E-02
GO:0051246	regulation of protein metabolic process	3.02E-02
GO:0046777	protein autophosphorylation	3.27E-02
GO:0050790	regulation of catalytic activity	3.37E-02
GO:0050789	regulation of biological process	3.50E-02
GO:0065009	regulation of molecular function	3.55E-02
GO:0055085	transmembrane transport	3.99E-02
GO:0040008	regulation of growth	4.08E-02
GO:0007088	regulation of mitotic nuclear division	4.08E-02
GO:0045786	negative regulation of cell cycle	4.87E-02
GO:0051783	regulation of nuclear division	4.87E-02

Supplementary Table 16 P-values of GO biological process fold enrichment of NPs that are represented in Figure 4:6

ID	Name	P-value
GO:0006259	DNA metabolic process	1.32E-16

GO:0016310	phosphorylation	1.64E-16
GO:0009987	cellular process	3.10E-15
GO:0044238	primary metabolic process	3.74E-15
GO:0071704	organic substance metabolic process	8.72E-15
GO:0008152	metabolic process	1.16E-14
GO:0044260	cellular macromolecule metabolic process	2.64E-14
GO:0044237	cellular metabolic process	1.25E-13
GO:0006807	nitrogen compound metabolic process	8.95E-13
GO:0006796	phosphate-containing compound metabolic process	9.16E-12
GO:0006793	phosphorus metabolic process	1.47E-11
GO:0043170	macromolecule metabolic process	1.66E-11
GO:0036211	protein modification process	6.04E-10
GO:0006464	cellular protein modification process	6.04E-10
GO:0006468	protein phosphorylation	6.99E-10
GO:0033554	cellular response to stress	8.87E-10
GO:0006281	DNA repair	9.48E-10
GO:0006974	cellular response to DNA damage stimulus	1.28E-09
GO:0006950	response to stress	1.35E-09
GO:1901360	organic cyclic compound metabolic process	4.72E-08
GO:0008150	biological process	4.73E-08
GO:0043412	macromolecule modification	4.85E-08
GO:0006265	DNA topological change	1.10E-07
GO:0006139	nucleobase-containing compound metabolic process	1.55E-07
GO:0042866	pyruvate biosynthetic process	2.38E-07
GO:0009135	purine nucleoside diphosphate metabolic process	2.38E-07
GO:0006096	glycolytic process	2.38E-07
GO:0046031	ADP metabolic process	2.38E-07
GO:0006757	ATP generation from ADP	2.38E-07
GO:0009185	ribonucleoside diphosphate metabolic process	2.38E-07
GO:0009179	purine ribonucleoside diphosphate metabolic process	2.38E-07
GO:0006725	cellular aromatic compound metabolic process	2.40E-07
GO:0046483	heterocycle metabolic process	2.96E-07
GO:0009166	nucleotide catabolic process	4.27E-07
GO:0006090	pyruvate metabolic process	4.27E-07
GO:0016052	carbohydrate catabolic process	5.60E-07
GO:1901292	nucleoside phosphate catabolic process	5.60E-07
GO:1901564	organonitrogen compound metabolic process	8.04E-07
GO:0006165	nucleoside diphosphate phosphorylation	1.19E-06
GO:0046939	nucleotide phosphorylation	1.19E-06
GO:0019363	pyridine nucleotide biosynthetic process	1.50E-06
GO:0019359	nicotinamide nucleotide biosynthetic process	1.50E-06

GO:0072525	pyridine-containing compound biosynthetic process	1.87E-06
GO:0046434	organophosphate catabolic process	1.87E-06
GO:0090304	nucleic acid metabolic process	3.47E-06
GO:0006754	ATP biosynthetic process	3.49E-06
GO:0009132	nucleoside diphosphate metabolic process	4.24E-06
GO:0009168	purine ribonucleoside monophosphate biosynthetic process	5.11E-06
GO:0009127	purine nucleoside monophosphate biosynthetic process	5.11E-06
GO:0009156	ribonucleoside monophosphate biosynthetic process	6.14E-06
GO:0009124	nucleoside monophosphate biosynthetic process	7.32E-06
GO:0006091	generation of precursor metabolites and energy	1.01E-05
GO:0009145	purine nucleoside triphosphate biosynthetic process	1.03E-05
GO:0009206	purine ribonucleoside triphosphate biosynthetic process	1.03E-05
GO:0051716	cellular response to stimulus	1.03E-05
GO:0019362	pyridine nucleotide metabolic process	1.21E-05
GO:0046496	nicotinamide nucleotide metabolic process	1.21E-05
GO:0072524	pyridine-containing compound metabolic process	1.41E-05
GO:0009150	purine ribonucleotide metabolic process	1.49E-05
GO:0009201	ribonucleoside triphosphate biosynthetic process	1.65E-05
GO:0009142	nucleoside triphosphate biosynthetic process	1.65E-05
GO:0019538	protein metabolic process	1.91E-05
GO:0072330	monocarboxylic acid biosynthetic process	2.22E-05
GO:0034655	nucleobase-containing compound catabolic process	2.22E-05
GO:0009152	purine ribonucleotide biosynthetic process	2.55E-05
GO:0046034	ATP metabolic process	2.55E-05
GO:0009259	ribonucleotide metabolic process	2.68E-05
GO:0006163	purine nucleotide metabolic process	2.68E-05
GO:0006733	oxidoreduction coenzyme metabolic process	2.93E-05
GO:0009126	purine nucleoside monophosphate metabolic process	3.36E-05
GO:0009167	purine ribonucleoside monophosphate metabolic process	3.36E-05
GO:0044270	cellular nitrogen compound catabolic process	3.84E-05
GO:0019439	aromatic compound catabolic process	3.84E-05
GO:0009161	ribonucleoside monophosphate metabolic process	3.84E-05
GO:0044267	cellular protein metabolic process	4.34E-05
GO:0046700	heterocycle catabolic process	4.36E-05
GO:0006164	purine nucleotide biosynthetic process	4.36E-05
GO:1901361	organic cyclic compound catabolic process	4.36E-05
GO:0009123	nucleoside monophosphate metabolic process	4.95E-05
GO:0046390	ribose phosphate biosynthetic process	4.95E-05
GO:0009260	ribonucleotide biosynthetic process	4.95E-05
GO:0009144	purine nucleoside triphosphate metabolic process	5.60E-05
GO:0009205	purine ribonucleoside triphosphate metabolic process	5.60E-05

GO:0034404	nucleobase-containing small molecule biosynthetic process	5.60E-05
GO:0046854	phosphatidylinositol phosphorylation	6.10E-05
GO:0046834	lipid phosphorylation	6.10E-05
GO:0019693	ribose phosphate metabolic process	6.84E-05
GO:0034641	cellular nitrogen compound metabolic process	7.20E-05
GO:0009199	ribonucleoside triphosphate metabolic process	7.98E-05
GO:0009141	nucleoside triphosphate metabolic process	7.98E-05
GO:0006732	coenzyme metabolic process	9.94E-05
GO:0032787	monocarboxylic acid metabolic process	1.11E-04
GO:0072521	purine-containing compound metabolic process	1.30E-04
GO:0005975	carbohydrate metabolic process	1.30E-04
GO:0009108	coenzyme biosynthetic process	1.52E-04
GO:0072522	purine-containing compound biosynthetic process	1.86E-04
GO:0017144	drug metabolic process	1.97E-04
GO:0046394	carboxylic acid biosynthetic process	2.25E-04
GO:0051188	cofactor biosynthetic process	2.25E-04
GO:0016053	organic acid biosynthetic process	2.25E-04
GO:0006996	organelle organization	2.37E-04
GO:0016043	cellular component organization	2.53E-04
GO:0006102	isocitrate metabolic process	3.43E-04
GO:0051186	cofactor metabolic process	4.16E-04
GO:0006006	glucose metabolic process	4.85E-04
GO:0006310	DNA recombination	5.33E-04
GO:0051276	chromosome organization	7.09E-04
GO:0071103	DNA conformation change	9.00E-04
GO:0071840	cellular component organization or biogenesis	9.03E-04
GO:0030258	lipid modification	9.27E-04
GO:0032006	regulation of TOR signaling	1.02E-03
GO:0019752	carboxylic acid metabolic process	1.25E-03
GO:0006470	protein dephosphorylation	1.38E-03
GO:0043436	oxoacid metabolic process	1.38E-03
GO:0044283	small molecule biosynthetic process	1.44E-03
GO:0006082	organic acid metabolic process	1.45E-03
GO:0048583	regulation of response to stimulus	1.96E-03
GO:0006476	protein deacetylation	2.01E-03
GO:0098732	macromolecule deacylation	2.01E-03
GO:0035601	protein deacylation	2.01E-03
GO:0071496	cellular response to external stimulus	2.01E-03
GO:0031668	cellular response to extracellular stimulus	2.01E-03
GO:0009991	response to extracellular stimulus	2.01E-03
GO:0052097	interspecies quorum sensing	2.15E-03

GO:0052106	quorum sensing involved in interaction with host	2.15E-03
GO:0044248	cellular catabolic process	2.30E-03
GO:0009372	quorum sensing	2.43E-03
GO:0048874	homeostasis of number of cells in a free-living population	2.43E-03
GO:0048872	homeostasis of number of cells	2.43E-03
GO:0044764	multi-organism cellular process	2.43E-03
GO:1901137	carbohydrate derivative biosynthetic process	2.94E-03
GO:0044145	modulation of development of symbiont involved in interaction with host	3.07E-03
GO:0043903	regulation of symbiosis, encompassing mutualism through parasitism	3.07E-03
GO:0043900	regulation of multi-organism process	3.07E-03
GO:0018342	protein prenylation	3.30E-03
GO:0097354	prenylation	3.30E-03
GO:0009056	catabolic process	3.51E-03
GO:0019318	hexose metabolic process	3.52E-03
GO:0055114	oxidation-reduction process	4.11E-03
GO:0006457	protein folding	4.42E-03
GO:0016126	sterol biosynthetic process	4.89E-03
GO:0016311	dephosphorylation	4.95E-03
GO:0050793	regulation of developmental process	5.15E-03
GO:1901135	carbohydrate derivative metabolic process	5.16E-03
GO:0005996	monosaccharide metabolic process	6.53E-03
GO:0016125	sterol metabolic process	6.77E-03
GO:1901575	organic substance catabolic process	8.69E-03
GO:1902531	regulation of intracellular signal transduction	8.91E-03
GO:0072350	tricarboxylic acid metabolic process	1.13E-02
GO:0010646	regulation of cell communication	1.40E-02
GO:0009966	regulation of signal transduction	1.40E-02
GO:0019637	organophosphate metabolic process	1.60E-02
GO:1901362	organic cyclic compound biosynthetic process	1.64E-02
GO:0023051	regulation of signaling	1.69E-02
GO:1903939	regulation of TORC2 signaling	1.86E-02
GO:0042149	cellular response to glucose starvation	1.86E-02
GO:0009432	SOS response	1.86E-02
GO:0007131	reciprocal meiotic recombination	1.86E-02
GO:0030952	establishment or maintenance of cytoskeleton polarity	1.86E-02
GO:0035825	homologous recombination	1.86E-02
GO:1903046	meiotic cell cycle process	1.86E-02
GO:0030950	establishment or maintenance of actin cytoskeleton polarity	1.86E-02
GO:0061982	meiosis I cell cycle process	1.86E-02
GO:0047484	regulation of response to osmotic stress	1.86E-02
GO:0006278	RNA-dependent DNA biosynthetic process	1.86E-02

GO:0010833	telomere maintenance via telomere lengthening	1.86E-02
GO:0007127	meiosis I	1.86E-02
GO:0034063	stress granule assembly	1.86E-02
GO:0140013	meiotic nuclear division	1.86E-02
GO:0045041	protein import into mitochondrial intermembrane space	1.86E-02
GO:0007004	telomere maintenance via telomerase	1.86E-02
GO:0034214	protein hexamerization	1.86E-02
GO:0009117	nucleotide metabolic process	1.87E-02
GO:0042592	homeostatic process	2.16E-02
GO:0055086	nucleobase-containing small molecule metabolic process	2.16E-02
GO:0006753	nucleoside phosphate metabolic process	2.29E-02
GO:0006694	steroid biosynthetic process	2.34E-02
GO:1901293	nucleoside phosphate biosynthetic process	2.36E-02
GO:0009165	nucleotide biosynthetic process	2.36E-02
GO:0008202	steroid metabolic process	2.69E-02
GO:1901617	organic hydroxy compound biosynthetic process	2.69E-02
GO:0006629	lipid metabolic process	3.09E-02
GO:0051258	protein polymerization	3.47E-02
GO:0016569	covalent chromatin modification	3.47E-02
GO:0034248	regulation of cellular amide metabolic process	3.47E-02
GO:0006417	regulation of translation	3.47E-02
GO:0016570	Histone modification	3.47E-02
GO:0006273	lagging strand elongation	3.68E-02
GO:1903432	regulation of TORC1 signaling	3.68E-02
GO:0051321	meiotic cell cycle	3.68E-02
GO:0009298	GDP-mannose biosynthetic process	3.68E-02
GO:0034250	positive regulation of cellular amide metabolic process	3.68E-02
GO:0022616	DNA strand elongation	3.68E-02
GO:0006271	DNA strand elongation involved in DNA replication	3.68E-02
GO:0045727	positive regulation of translation	3.68E-02
GO:0065002	intracellular protein transmembrane transport	3.88E-02
GO:0044743	protein transmembrane import into intracellular organelle	3.88E-02
GO:0071806	protein transmembrane transport	3.88E-02
GO:0065008	regulation of biological quality	3.94E-02
GO:0007010	cytoskeleton organization	4.07E-02
GO:0050896	response to stimulus	4.44E-02
GO:0034654	nucleobase-containing compound biosynthetic process	4.56E-02
GO:0006626	protein targeting to mitochondrion	4.77E-02
GO:1990542	mitochondrial transmembrane transport	4.77E-02