Application of Computer-Aided Drug Design for Identification of *P. falciparum* inhibitors

A thesis submitted in fulfilment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

IN BIOINFORMATICS

of

RHODES UNIVERSITY, SOUTH AFRICA

Research Unit in Bioinformatics (RUBi)

DEPARTMENT OF BIOCHEMISTRY and MICROBIOLOGY

Faculty of Science

by

Bakary N'tji Diallo

March 2021

ABSTRACT

Malaria is a millennia-old disease with the first recorded cases dating back to 2700 BC found in Chinese medical records, and later in other civilizations. It has claimed human lives to such an extent that there are a notable associated socio-economic consequences. Currently, according to the World Health Organization (WHO), Africa holds the highest disease burden with 94% of deaths and 82% of cases with *P. falciparum* having ~100% prevalence. Chemotherapy, such as artemisinin combination therapy, has been and continues to be the work horse in the fight against the disease, together with seasonal malaria chemoprevention and the use of insecticides. Natural products such as quinine and artemisinin are particularly important in terms of their antimalarial activity. The emphasis in current chemotherapy research is the need for time and cost-effective workflows focussed on new mechanisms of action (MoAs) covering the target candidate profiles (TCPs). Despite a decline in cases over the past decades with, countries increasingly becoming certified malaria free, a stalling trend has been observed in the past five years resulting in missing the 2020 Global Technical Strategy (GTS) milestones. With no effective vaccine, a reduction in funding, slower drug approval than resistance emergence from resistant and invasive vectors, and threats in diagnosis with the pfhrp2/3 gene deletion, malaria remains a major health concern.

Motivated by these reasons, the primary aim of this work was a contribution to the antimalarial pipeline through *in silico* approaches focusing on *P. falciparum*. We first intended an exploration of malarial targets through a proteome scale screening on 36 targets using multiple metrics to account for the multi-objective nature of drug discovery. The continuous growth of structural data offers the ideal scenario for mining new MoAs covering antimalarials TCPs. This was combined with a repurposing strategy using a set of orally available FDA approved drugs. Further, use was made of time- and cost-effective strategies combining QVina-W efficiency metrics that integrate molecular properties, GRIM rescoring for molecular interactions and a hydrogen mass repartitioning (HMR) molecular dynamics (MD) scheme for accelerated development of antimalarials in the context of resistance. This pipeline further integrates a complex ranking for better drug-target selectivity, and normalization strategies to overcome docking scoring function bias. The different metrics, ranking, normalization strategies and their combinations were first assessed using their mean ranking error (MRE). A version combining all metrics was used to select 36 unique protein-ligand complexes, assessed in MD, with the final retention of 25. From the 16 *in vitro* tested hits of the 25, fingolimod, abiraterone, prazosin, and terazosin showed antiplasmodial activity with IC_{50} 2.21, 3.37, 16.67 and 34.72 µM respectively and of these, only fingolimod was found to be not safe with respect to human cell viability. These compounds were predicted active on different molecular targets, abiraterone was predicted to interact with a putative liver-stage essential target, hence promising as a transmission-blocking agent. The pipeline had a promising 25% hit rate considering the proteome-scale and use of cost-effective approaches.

Secondly, we focused on *Plasmodium falciparum* 1-deoxy-D-xylulose-5-phosphate reductoisomerase (PfDXR) using a more extensive screening pipeline to overcome some of the current *in silico* screening limitations. Starting from the ZINC lead-like library of ~3M, hierarchical ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS)

approaches with molecular docking and re-scoring using eleven scoring functions (SFs) were used. Later ranking with an exponential consensus strategy was included. Selected hits were further assessed through Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA), advanced MD sampling in a ligand pulling simulations and (Weighted Histogram Analysis Method) WHAM analysis for umbrella sampling (US) to derive binding free energies. Four leads had better predicted affinities in US than LC5, a 280 nM potent PfDXR inhibitor with ZINC000050633276 showing a promising binding of -20.43 kcal/mol. As shown with fosmidomycin, DXR inhibition offers fast acting compounds fulfilling antimalarials TCP1. Yet, fosmidomycin has a high polarity causing its short half-life and hampering its clinical use. These leads scaffolds are different from fosmidomycin and hence may offer better pharmacokinetic and pharmacodynamic properties and may also be promising for lead optimization. A combined analysis of residues' contributions to the free energy of binding in MM-PBSA and to steered molecular dynamics (SMD) F_{max} indicated GLU233, CYS268, SER270, TRP296, and HIS341 as exploitable for compound optimization.

Finally, we updated the SANCDB library with new NPs and their commercially available analogs as a solution to NP availability. The library is extended to 1005 compounds from its initial 600 compounds and the database is integrated to Mcule and Molport APIs for analogs automatic update. The new set may contribute to virtual screening and to antimalarials as the most effective ones have NP origin.

DECLARATION

I, Bakary N'tji Diallo, declare that this thesis submitted to Rhodes University is wholly my own work and has not previously been submitted for a degree at this or any other institution.

The research described in this thesis was carried out for the degree Doctor of Philosophy in Bioinformatics at Rhodes University, from March 15th, 2018 to March 15th, 2021 under the supervision of Prof Kevin Lobb and Prof Ozlem Tastan Bishop.

Signature: 

Date: /01/2021

ACKNOWLEDGEMENTS

Firstly, I would like to thank the Developing Excellence in Leadership and Genetics Training for Malaria Elimination in sub-Saharan Africa (DELGEME) and Prof Djimdé for the scholarship to study bioinformatics and all the other financial support.

I acknowledge my supervisors Pr. Kevin Lobb and Pr. Özlem Tastan Bishop for all the help and support and the provided learning environment.

Finally, special thanks to my family, friends, my mentor Dr Kone, all RUBi friends and colleagues for invaluable moral and technical support.

Simulations were done on the Centre for High Performance Computing (CHPC), Cape Town, South Africa.

This work was supported through the DELTAS Africa Initiative [grant 107740/Z/15/Z]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [DELGEME grant 107740/Z/15/Z] and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD.

Table of content

List of figures

List of tables

RESEARCH OUTPUTS

Oral presentations:

Bakary N'tji Diallo, Kevin Lobb, and Özlem Taştan Bishop

In silico study of Plasmodium 1-deoxy-dxylulose 5-phosphate reductoisomerase (DXR) for identification of novel inhibitors from SANCDB

AAAMR 1st Congress (African Association for Research and Control of Antimicrobial Resistance) / ASTMH (American Society of Tropical Medicine & Hygiene) West Africa / University of Sciences, Techniques and Technologies of Bamako (USTTB), University of Bamako Mali

Bakary N'tji Diallo, Kevin Lobb, and Özlem Taştan Bishop

In silico study of Plasmodium 1-deoxy-dxylulose 5-phosphate reductoisomerase (DXR) for identification of novel inhibitors from SANCDB

1st DELGEME Mid-Term General meeting July 18th - July 20th, 2018 Bamako Mali

Poster presentations:

Bakary N'tji Diallo, Kevin Lobb, and Özlem Taştan Bishop. In silico study of Plasmodium 1-deoxy-dxylulose 5-phosphate reductoisomerase (DXR) for identification of novel inhibitors from SANCDB

ASTMH American Society of Tropical Medicine & Hygiene Sixty-Seventh Annual Meeting- October 28 – November 1, 2018

Diallo, Bakary N'tji, Tastan Bishop, Ö., & Lobb, K. (2019). Novel potential antimalarials through drug repurposing and multitargeting: a computational approach. *F1000Research*, *8*. https://doi.org/10.7490/f1000research.1117266.1

27th Conference On Intelligent Systems For Molecular Biology And The 18th European Conference On Computational Biology - ISMB/ECCB 2019 July 21-25

Publications and contributions:

Parts of this thesis have been published in the following papers:

Diallo, B. N., Swart, T., Hoppe, H. C., Tastan Bishop, Ö., & Lobb, K. (2021). Potential repurposing of four FDA approved compounds with antiplasmodial activity identified through proteome scale computational drug discovery and *in vitro* assay. *Scientific Reports*, *11*(1), 1413. https://doi.org/10.1038/s41598-020-80722-2

Contribution: All the computational experiments, their analysis, and wrote the first draft of manuscript under the guidancd of Pr. Özlem Tastan Bishop and Pr. Kevin Lobb.

Subramaniyam, S., Yang, S., Diallo, B. N., Fanshu, X., Lei, L., Li, C., … Bhattacharyya, S. (2020). Oral Phyto-thymol ameliorates the stress induced IBS symptoms. Scientific Reports, 10(1), 13900. https://doi.org/10.1038/s41598-020-70420-4

Contribution: All the computational experiments, their analysis, and the writing of related sections.

Diallo, B. N., Glenister, M., Musyoka, T. M., Lobb, K. A., & Tastan Bishop, Ö. (2021). SANCDB: an update on South African natural compounds and their readily available analogs.

Submitted to the Journal of Cheminformatics, under review.

Contribution: Part of the database content update, addition of analogs, data analysis and wrote the first draft of manuscript under the guidancd of Pr. Özlem Tastan Bishop and Pr. Kevin Lobb.

List of amino acids and their abbreviations

Amino acid	Three letter code	Single letter code
Alanine	ALA	A
Cysteine	CYS	C
Aspartic acid	ASP	D
Glutamic acid	GLU	E
Phenylalanine	PHE	F
Glycine	GLY	G
Histidine	HIS	H
Isoleucine	ILE	I
Lysine	LYS	K
Leucine	LEU	L
Methionine	MET	M
Asparagine	ASN	N
Proline	PRO	P
Glutamine	GLN	Q
Arginine	ARG	R
Serine	SER	S
Threonine	THR	T
Valine	VAL	V
Tryptophan	TRP	W
Tyrosine	TYR	Y

List of web servers and applications

SANCDB: https://sancdb.rubi.ru.ac.za/.

Protein Data Bank RCSB: https://www.rcsb.org/

scPDB: http://bioinfo-pharma.u-strasbg.fr/scPDB/

DrugBank: https://www.drugbank.com/

ZINC: http://zinc15.docking.org/

SciFinder: https://scifinder-n.cas.org/

MCULE: https://mcule.com/

MOLPORT: https://www.molport.com/

List of acronyms

2D	Two dimensions
3D	Three dimensions
AChE	Acetylcholinesterase
ACT	Artemisinin-based Combination Therapy
ADMET	Absorption, distribution, metabolism, excretion and toxicity
ADP	Adenosine Diphosphate
ADT	AutoDockTools
AFM	Atomic Force Microscopy
AMP	Adenosine monophosphate
API	Application Programming Interface
ATP	Adenosine Triphosphate
BEI	Binding Efficiency Index
BFE	Binding Free energy
BINANA	BINding ANAlyzer
CAS	Chemical Abstracts Service
CDP-ME	Methylerythritol cytidyl diphosphate
CHPC	Center for High-Performance Computing
COM	Center of Mass
COPD	Chronic Obstructive Pulmonary Disorder
CPU	Central Processing Unit
CSAR	Community Structure-Activity Resource
CSR	Chiral Shape Recognition
CTP	Cytidine 5'-triphosphate
D3R	Drug Design Data Resource
D-GLP	D-glyceraldehyde-3-phosphate
DHNCA	3,7-dihydroxynaphthalene-2-carboxylic acid
DMAPP	Dimethylallyl diphosphate, Dimethylallyl diphosphate
DOI	Digital Object Identifier
DOPE	Discrete Optimized Protein Energy
DRL	1-deoxy-d-xylulose 5-phosphate reductoisomerase-like
DSX	DrugScore eXtended
DUD	Directory of Useful Decoy
DXP	1-deoxy-D-xylulose-5-phosphate, 1-deoxy-d-xylulose 5-phosphate
DXR	1-deoxy-d-xylulose 5-phosphate reductoisomerase
ECD	Extra Cellular Domain
ECFP	Extended-Connectivity FingerPrint

ES	ElectroShape
ETM-DB	Integrated Ethiopian Traditional Herbal Medicine and Phytochemicals Database
FAD	Flavin adenine dinucleotide
FDA	Food and Drug Administration
Fos	Fosmidomycin
FS	Full Structure
FXR	Farnesoid X Receptor
GBIF	Global Biodiversity Information Facility
GI	Gastrointestinal
G_{MM}	Molecular Mechanics Energy
GNU	Gnu's Not Unix
GRIM	GRaph Interaction Matching
GROMACS	GROningen MAchine for Chemical Simulations
G_{TIE}	Total Interactions Energy
G_{US}	BFE from Umbrella Sampling
HBA	Hydrogen Bond Acceptor
Hbond	Hydrogen bonds
HIV	Human Immunodeficiency Virus
HMBPP	4-hydroxy-3-methylbutenyl 1-diphosphate
HMR	Hydrogen mass repartitioning
HYDE	HYdrogen bond and DEhydration
IBS	Irritable Bowel Syndrome
IMP	Inosine monophosphate
IPP	Isopentenyl diphosphate
LBVS	Ligand-Based Virtual Screening
LEC	Lowest Energy Conformations
LJ	Lennard-Jones
LR	Linker Region
LSH	Local Sensitive Hashing
LBVS	Ligand-Based Virtual Screening
MAOI	Monoamine Oxidase Inhibitor
MD	Molecular Dynamics
MEP	2-C-methyl-d-erythritol 4-phosphate
MG	Magnesium
MHFP	MinHash FingerPrint
MHFP6	MinHash fingerprint, up to six bonds
ML	Machine Learning
MM	Molecular Mechanics
MMFF94	Merck Molecular Force Field

MM-PBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
MoA	Mechanism of Action
MOAD	Mother Of All Databases
MQN	Molecular Quantum Numbers
MMV	Medicines for Malaria Venture
MVA	Mevalonate pathway
MW	Molecular weight
NAD	Nicotinamide Adenine Dinucleotide Oxidized
NADH	Nicotinamide Adenine Dinucleotide Reduced
NADPH	Nicotinamide adenine dinucleotide phosphate
NAG	N-acetyl-D-Glucosamine
NaPLeS	Natural products likeness scorer
NCBI	National Center for Biotechnology Information
nHA	Number of hydrogen bond acceptor
nHD	Number of hydrogen bond donor
NP	Natural Product
NPASS	Natural Product Activity and Species Source
NPT	Isothermal–isobaric ensemble
nRing	Number of rings
nRot	Number of rotatable bonds
NuBBE	Nuclei of Bioassays, Ecophysiology and Biosynthesis of Natural Products Database
OBSPEC	Obsepectrophore
ODDT	Open Drug Discovery Toolkit
PAINS	Pan-assay interference compounds
p-ANAPL	Pan-African Natural Products Library
PB	Poisson-Boltzmann
PBS	Portable Batch System
PBSA	Poisson-Boltzmann Solvent Accessible
PCA	Principal Component Analysis
PDB	Protein Data Bank
PES	Potential Energy Surface
PLEC	Protein-ligand extended connectivity
PLIE	Protein Ligand Interaction Energy
PMF	Potential of Mean Force
PPI-like	Protein-protein inhibitor like
PSA	Polar Surface Area
QED	Quantitative Estimate of Druggability
QSAR	Quantitative structure–activity relationship
Rg	Radius of gyration
RMSD	root-mean-square deviation
RMSF	root-mean-square fluctuation
RNA	Ribonucleic Acid

ROC	receiver operating characteristic
ROCS	Rapid overlay of chemical structures
RUBi	Research Unit in Bioinformatics
SA	Solvent accessible
SANBI	South African National Biodiversity Institute
SANCDB	South African natural compound database
SAR	Structure-Activity Relationship
SASA	Solvent Accessible Surface Area
SBVS	Structure-Based Virtual Screening
SD	Standard Deviations
SEI	Surface Efficiency Index
SF	Scoring Function
SLURM	Simple Linux Utility for Resource Management
SMC	Seasonal Malaria Chemoprevention
SMD	Steered Molecular Dynamics
SMILES	Simplified Molecular Input Line Entry Specification
SPC	Simple Point-Charge
Tc	Tanimoto coefficient
TCP	Target Candidate Profile
TPP	Targe Product Profile
TPSA	Total Polar Surface Area
t-SNE	t-Distributed Stochastic Neighbor Embedding
US	Umbrella Sampling
USRCAT	Ultrafast Shape Recognition with CREDO Atom Types
VS	Virtual Screening
WHAM	Weighted Histogram Analysis Method
WHO	World Health Organization

Chapter 1: Literature Review

1.1 Malaria

Malaria is a disease caused by the parasite *Plasmodium*, where infection of this parasite is through female *Anopheles* mosquitoes bites [1]. Ten to fifteen days after infection, its symptoms: fever, vomiting, headache, diarrhea, nausea, and abdominal pain appear [2]. The parasite is a protozoan with mainly four species (*P. falciparum*, *P. vivax*, *P. ovale,* and *P. malariae*), and rarely *P. knowlesi* causing the disease in humans with health consequences [3,4].

Malaria carries a serious health burden. The disability-adjusted life years (DALYs), is a composite metric capturing both premature mortality and prevalence and severity of ill-health [5] allowing a disease burden assessment. According to the latest Global Burden of Disease 2017 statistics, the top five DALYs causes were communicable diseases (lower respiratory infections, malaria, diarrheal diseases, HIV/AIDS, and tuberculosis) and neonatal disorders. Neglected tropical diseases including malaria cause approximately 62,300 global disability-adjusted life years (DALYs). Malaria contributed up to 72% of that number [5]. In the context of other infectious diseases, due to its DALY malaria was the third most funded disease with ($125 per DALY) behind HIV/AIDS and tuberculosis with $772 and $156 per DALY respectively. The fourth most funded one was pneumonia with $33 per DALY [6].

The 2020 World Malaria Report indicated 229 million cases worldwide and 409 000 deaths in 2019, with a global mortality rate (deaths per 100 000 population at risk) of 10 in 2019. The World Health Organization (WHO) African Region recorded around 94% (215 million) of all cases in 2019. Malaria has an associated burden with a particularly weak segment of the population (pregnant women and children under the age of five) representing 70% of deaths. 35% of pregnancies (12 million) were exposed to malaria infection-causing 822 000 children with low birth weight in the WHO African Region [7]. This health burden has economic consequences locking some families in a vicious poverty circle and being a major obstacle to socio-economic development. Models associated a 10% decrease in malaria incidence with a nearly 0.3% increase in income per capita and a 0.11 percentage point faster per capita growth per annum [8].

Despite these statistics, malaria had been declining thanks to advancements, but a stalling trend has been observed in recent years (2015-2020). The global case incidence (cases per 1000 population at risk) fell from 80 to 57 between 2000 and 2019. Deaths were reduced from 736 000 to 409 000 and the mortality rate from 25 to 10 during the same period. More countries moved toward elimination. The number of malaria-endemic countries reporting fewer than 10 000 cases increased from 26 to 46 and those with fewer than 100 indigenous cases from six to

27. All malaria-free certified countries remained transmission-free. These advancements prevented 7.6 million deaths and 1.5 billion cases[7].

In vaccine research RTS,S (Mosquirix) remains the only approved vaccine since 2015 [9]. In 2019, RTS,S/AS01 was launched in a pilot study in three African countries. A broader use may be considered by WHO based on results. Other promising vaccine research projects are undergoing: Multi-Stage Malaria Vaccine Consortium, MIMVaC-Africa, and the PfTBV consortium [10]. Successful results from these projects could open a new paradigm for disease control. These advancements allow a shift towards elimination with the Global technical strategy (GTS) for malaria. The GTS aim for during the period 2016–2030: at least 90% reduction in case incidence and mortality rate from a 2015 baseline, elimination in 35 countries, and preventing resurgence in malaria-free countries [7].

Many challenges are to be addressed despite the above advancements. Malaria burden decrease has slowed since 2015, and most of the set GTS 2020 milestones were off-track (Figure 1-1). Mortality rates stagnated between 2015 and 2020 in seven countries (8%), augmented in 12 (13%) with six of them having more than 40% increase. Case incidence only decreased from 58 to 57 between 2015 and 2019. The global mortality rate was 12 in 2015 and 10 in 2019. Comoros, Costa Rica, Ecuador, and Suriname recorded more cases in 2019 than in 2018. Moreover, there has been a decrease in investment in malaria programmes and research. Expenses required and thus the invested gap increased to 1.3, 2.3, and 2.6 US$ billion for 2017, 2018, and 2019 respectively [7].

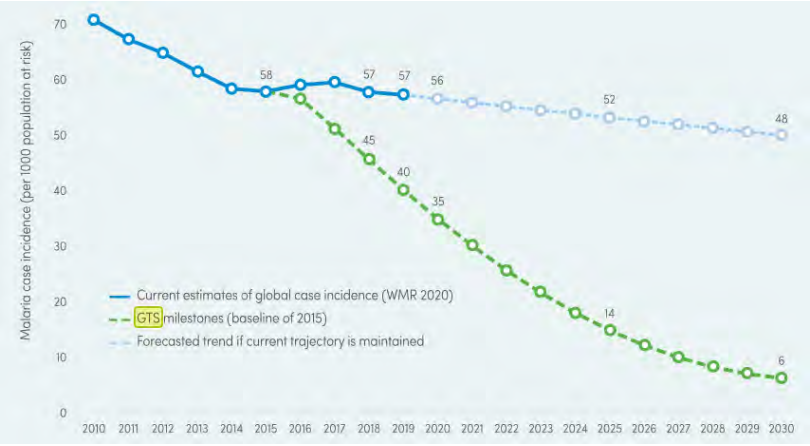


Figure 1-1 Case incidence in two scenarios: blue and light blue: current trajectory and its forecasting respectively, green: GTS achieved goals (Source: World Malaria Report: 20 years of global progress and challenges [7])

1.2 Current biological threats

The above trend combined with the potential threats portrays a preoccupying situation. The malaria threat map highlights four current challenges: invasive vectors and vector insecticide resistance, deletions in *pfhrp2/3* genes, and parasite drug efficacy and resistance [11].

Chemotherapy has been key in the fight against malaria by treatment and prevention. Artemisinin-based combination therapies (ACTs) treatment, preventive chemoprophylaxis for travelers, 3 doses of intermittent with sulfadoxine-pyrimethamine for pregnant women, seasonal malaria chemoprevention (SMC), and routine vaccinations are currently recommended by the WHO [7]. 3.1 billion ACTs were globally sold in 2010–2019 and 0.2 to 21.5 million children received at least one SMC dose in 2012-2019. 36% of malaria R&D funding between 2007 and 2018 went into drugs followed by lower shares in basic research, vaccines, vector control, and diagnostics products [7]. In the elimination era, threats related to the parasite are mainly fought through adapting intervention strategies to its biology. The parasite showed remarkable ability to develop drug resistance and resisted all classes of drugs used in malaria treatment: atovaquone, quinine, proguanil, chloroquine, mefloquine and sulfadoxine-pyrimethamine [12–14]. The current WHO recommendation treatment regimen: ACTs is now threatened. Cases of plasmodium resistance to artemisinin were reported in Southeast Asia: Thailand, Laos, Myanmar, Wet Nam, and Cambodia. Its spread to other areas can hamper past progresses. More than resisting a single drug, the parasite can also resist drug combinations thus driving strategies involving more than two combinations of drugs such as artesunate-lumefantrine-amodiaquine and dihydroartemisinin-piperaquine-mefloquine [7,15–17].

Resistance can also occur in the vector. Vector control is another key element for disease control and elimination. 2.2 billion insecticide-treated mosquito nets were globally supplied in 2004–2019 and this was accompanied by indoor residual spraying. On 2010–2019 data from 82 countries, 28 had observed resistance to all four of the most frequently utilized insecticide classes in at least one malaria vector and one collection site. 73 had recorded resistance to at least one class [7]. More than their resistance, new invasive species settlement in new ecosystems is threatening. *Anopheles stephensi* is a southern Asia vector of *Plasmodium falciparum* and *P. vivax* was recorded in Djibouti in 2012 and linked with an unusual outbreak of urban *P. falciparum* malaria [18]. Another threat is related to disease diagnosis. Malaria rapid diagnostic tests (RDTs) contribute to proper treatment. 2.7 billion tests were sold in 2010-2019. Parasite deletions in *pfhrp2/3* genes make them undetectable by RDTs based on histidine-rich protein 2 (HRP2). This is a major biological threat given limited alternatives. Its real prevalence remains unknown, ranging from 0% to 100% also undermined by variable methods in sample selection and laboratory analysis means [19].

1.3 Biology of plasmodium and drug discovery opportunities

Plasmodium spp. are eukaryotes unicellular and belong to the apicomplexan of the protozoan phylum. The *Apicomplexa* are identified by the apicoplast, an essential organelle producing important compounds for parasite growth such as isoprenoids and fatty acids. *Plasmodium spp.* are obligate intracellular parasites with multiple hosts throughout their life-cycle [20,21]. The life-cycle of these parasites has three phases between human and mosquito hosts: the liver, blood, and the mosquito phase (Figure 1-2) [3]. Many groups active in antimalarial drug discovery, have organized their effort within a framework of molecule type (Target Candidate Profiles (TCP)), corresponding to chemotherapy strategies around the parasite life-cycle [22].

Malaria infection starts with parasite inoculation into the human bloodstream through a female anopheline bite (TCP6). These infecting parasite cells are sporozoites. During this liver stage, they migrate to the liver, invade the hepatocytes through the Kupffer cells, and start schizogony. This endogenous asexual multiplication lasts 5-15 days depending on the species [23]. *P. vivax* and *P. ovale* sporozoites can differentiate into a latent form in the liver, the hypnozoites (TCP3). They can multiply days to years later leading to a new infection [24,25]. The schizogony results in mature sporozoites, schizonts containing thousand of merozoites released into the bloodstream by hepatocytes rupture (TCP4).

During the blood phase or erythrocytic cycle, the merozoites invade and multiply asexually in the erythrocytes for 48 to 72 hours. They evolve into diverse forms: rings, trophozoites, and schizonts with each schizont containing about 6 to 36 merozoites[26]. This multiplication results in erythrocytes rupture and release of merozoites, which can infect new erythrocytes (TCP1). The symptoms in humans occur at this stage [26,27]. Some parasites further differentiate into gametocytes male and female [26].

These gametes can then be absorbed by a female anopheline [27,28]. During this mosquito phase, the parasites evolve into their sexual forms (male microgametes and female macrogametes) (TCP5) [28]. Their fertilization results in the ookinete formation [3] which mature into oocysts, in which sporogonic replication takes place for about two weeks. New infective sporozoites are hence formed and migrate to the salivary glands. They will be released into the human dermis during the mosquito blood meal, hence closing the cycle [28].

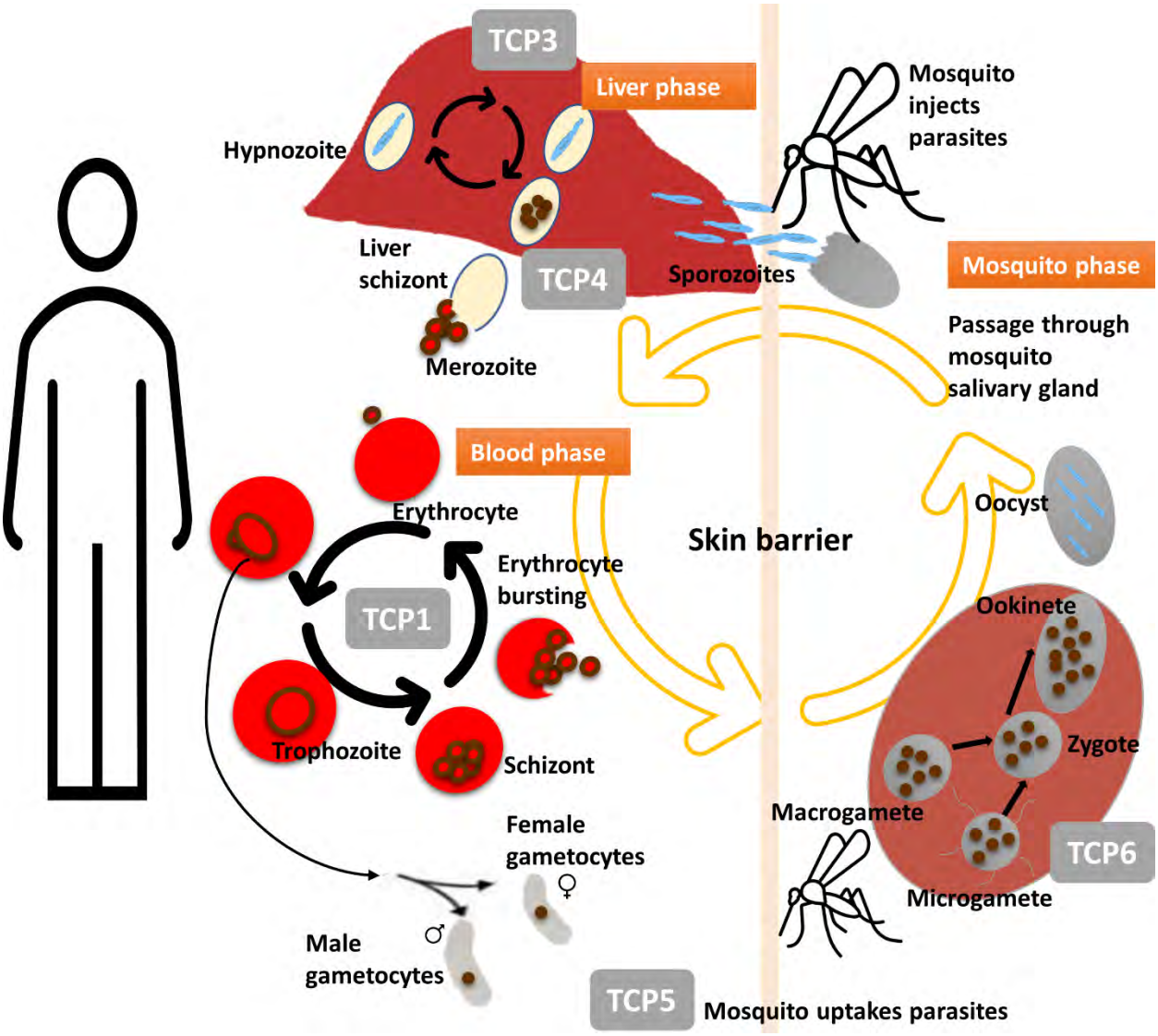


Figure 1-2 Malaria parasite life cycle through the lenses of drug discovery. The TCPs highlight the different possible drug discovery intervention strategies.

With current eradication targets and to anticipate future threats, key intervention strategies have been identified for antimalarial development and formulation in TCPs (Figure 1-2 and Table 1-1) [22]. Combining drugs with different mechanisms of action (MoA) decreases the chances of resistance occurring [29]. New drugs, with novel scaffolds and MoAs, are ideal [22]. New drug combinations (three molecules and more) and repurposing are also considered [30].

Table 1-1 Some drugs currently used and clinical candidates covering different TCPs. Adapted from [22].

TCPs	Targets	Example of current drugs	Clinical candidates
TCP1 Symptomatic treatment	Asexual blood stages	ACTs atovaquone-proguanil mefloquine [31]	KAE609 KAF156 [32] SJ733[33] DDD498, DSM265, MMV048 [31]
TCP2 (retired and combined to TCP1 fast killers of blood schizonticides, and long-acting molecules) [34]			
TCP3 Anti-relapse	Hepatic stage hypnozoites	Primaquine and tafenoquine [34]	KAI407 [22]
TCP4 Chemoprotection	Hepatic stage schizonts	Atovaquone-proguanil [31]	DSM265, KAF156, P218 [34] DDD498, KAF156, MMV048 [31]
TCP5 Transmission blocking	Gametocytes/Gametes	Primaquine [31]	KAF156 [34]
TCP6 Transmission blocking	Insect vector (endectocides)	Ivermectin [34]	

Recent research has focussed on the development of much stronger antimalarial portfolios. For example, new molecules have achieved interesting activity in transmission-blocking *in vitro*, at promising concentrations for future clinical application. These compounds include OZ439, KAE609, KAF156, SJ733, and DDD498 [30,35]. Drugs currently exist for the different identified TCPs (Table 1-1). DDD498, KAF156, DSM265, and MMV048 combine asexual and hepatic schizont stages activities [31]. KAF156 showed good potential as a multi-stages active compound and active against resistant strains including where artemisinin resistance is evident [36]. Still many are areas of improvement remain, especially given the biological threats. Anti-relapse compounds are lacking compared to other TCPs. Currently, only tafenoquine and primaquine are approved but induce hemolysis in glucose-6-phosphate dehydrogenase (G6PD)-deficient patients, a population of 350 million people. This may be particularly challenging to overcome [22,37].

Beyond the traditional drug discovery methods and combinatorial therapies, other strategies that can contribute to antimalarial development have been identified [30]. Drug repurposing is an interesting strategy that has been identified as a particular route in antimalarials development. This strategy uses known drugs to treat different diseases beyond their initial indications [38]. Particularly in the context of rapid emergence of resistance, this approach may accordingly accelerate the development of antimalarials [39,40]. Methylene blue, rosiglitazone, fosmidomycin,

imatinib and sevuparin are examples of molecules being explored for their potential repurposing as antimalarial [30]. Polypharmacology is another alternative strategy in drug discovery [41]. Drug development can be more expensive in the case of a combinatorial therapy than a single compound, which may still have the same effect through multitargeting. Drug-drug negative interaction risks are higher for drug combinations [41]. Multi-target antimalarials can be developed through a single hybrid molecule design or compound multiple activity explorations [42]. Chloroquine analogs with dual activity showed excellent activity against resistant strains from Thailand and Cambodia [43]. Chen *et al.* recently identified FP-2 and PfDHFR dual inhibitors [44]. MMV007571 and MMV020439, from the Malaria Box library were identified as dual inhibitors of the dihydroorotate dehydrogenase (DHODH) and the parasites' new permeability pathways (NPPs) [45]. Through virtual screening, eight compounds from the ChemBridge library were found to have dual activity on falcipain-2 and falcipain-3 [42]. Hence, multitargeting strategy also fits with antimalarial development, especially within the context of drug resistance [40].

1.3.1 Overview of current targets

In this section, we give an overview of *P. falciparum* targets to identify drug discovery opportunities. The process of drug design first identifies a suitable target [46] through methods such as genetic knockout and gene silencing[47]. A good pool of potential targets exists [48,49]. Malaria elimination will certainly benefit from new MoAs for resistance but also multi-stage active compounds [22]. However, parasite cell screens have demonstrated that even inhibitors with distinct scaffolds seem to converge toward the same ~12 targets [48], making it challenging to hunt for compounds achieving a new MoA. Searching the DrugBank website (January 9th,2021) for Food and Drug Administration (FDA) approved compounds targeting plasmodium returned 14 compounds for 19 unique targets. Yet, many inhibitors at different stages of development, focused on different pathways and enzyme targets still exist in the antimalarials pipeline [22,35,50,51]. The parasite genome sequencing [51] has opened the door for target space mining with the identification of chokepoint enzymes [52]. Two notable studies deciphered its target space to uncover drug discovery opportunities. Gomes *et al.* developed a genetic screening approach to find *Plasmodium berghei* essential genes. *P. berghei* relative growth rates were estimated for more than 2,500 genes, confirming the druggability of several kinases [49]. Similarly, *P. falciparum* essential genes were also uncovered, confirming 2680 essential genes, including ~1000 essential genes conserved in the parasite [53]. This has enabled the prioritization of high-value targets, especially for small molecule inhibition. Keeping in mind the low correlation between essential gene and high-value target [48], even if only 10% of this set is suitable for small molecule inhibition, that is still a good pool of about ~268 protein targets.

Yet, the current antimalarials discovery pipeline is marked by the so-called Harlow-Knapp effect, or “searching under the lamppost,”. Few targets are extensively studied while as highlighted in the above paragraph many opportunities exist (Figure 1-3) [54]. PfDHODH [55], PfDXR [56], PfDHFR are well studied with numerous crystal structures in Protein Data Bank (PDB), deposited bioactivity data in ChEMBL. Yet for some of these, including for DXR, this research has not resulted in an approved drug.

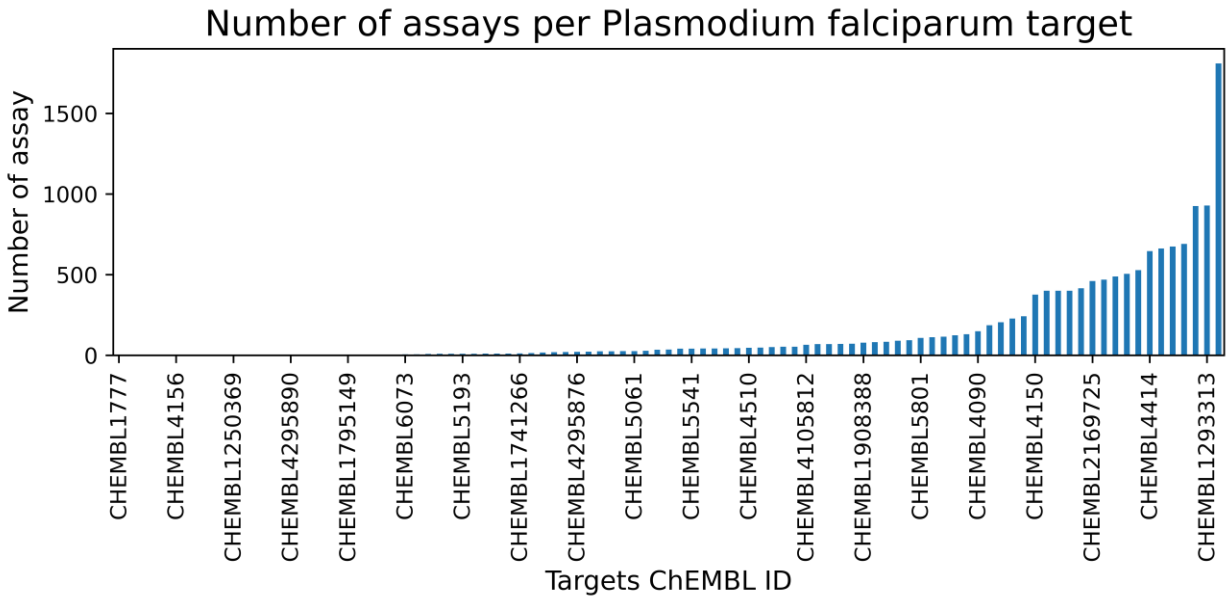


Figure 1-3 Distribution of bioactivity data deposited per *Plasmodium falciparum* protein targets. The x and y-axis represent the number of assay data and their respective targets. The targets are sorted by assay count. A few targets have the highest amount of data with a long tail distribution illustrating the Harlow-Knapp effect. The data was extracted from chembl_webresource_client version 0.10.2.

Aneja *et al.* reviewed targets, pathways, organelles, and their related inhibitors optimizations from a structure-based (SB) design perspective [51]. The apicoplast implied in the fatty acid type II (FAS-II), isoprenoids and heme synthesis is a key organelle for antimalarial discovery [57]. Good selectivity is achievable for targets such as DXR and the apical membrane antigen 1 since these do not have human homologs. Fosmidomycin, one of the most advanced DXR inhibitor candidates failed in monotherapy and is now under investigation in terms of its combination with piperaquine [58,59]. Hemoglobin digestion in the food vacuole is another key pathway with approved drugs (for example halofantrine) targeting aspartic proteases (plasmepsins) [60]. The Medicines for Malaria Venture (MMV) P218 is in clinical trials and acts on DHFR in terms of inhibiting folate metabolism. Proguanil and pyrimethamine are FDA approved antimalarials targeting DHFR. The electron transport system in the mitochondria has key enzymes such as the cytochrome bc1 complexes and PfDHODH, and the NADH ubiquinone oxidoreductase (PfNDH2) [51]. KAF156 is a compound that inhibits PfDHODH, and it is currently in Phase 2b clinical trials [61]. Given the need for new MoA covering the different TCPs to combat resistance, disease control, and elimination [22,31], current antimalarial development focuses on new MoA [61]. Candidates with new MoA include Methylene Blue, MMV048, KAF156 which target glutathione reductase, phosphoinositol 4-kinase and PfCARL respectively [61].

1.3.2 NPs as antimalarials

Another strategy for antimalarial discovery is through natural compounds (NPs). NPs have been and continue to be a major source of drugs, including antimalarials. Avermectin and artemisinin have revolutionized onchocerciasis, lymphatic filariasis, and malaria and their discovery was awarded the 2015 Nobel Prize in Physiology or Medicine [62]. The first antimalarial was the NP

quinine, isolated from Cinchona bark. Its synthetic derivative, chloroquine, has played a key role in malaria chemotherapy[63]. NP databases offer antimalarial compounds [64,65] and specialized databases including antimalarial activity have been created (AfroMalariaDB [66]). Assessing 122 drugs derived from plants, 80% were linked to their initial ethnopharmacological uses [67]. Wells underlines the importance of firstly identifying the clinical activity of a herbal medicinal product used by the community before further pharmacological investigation. This approach was initially suggested by Chen Guofu in 1952 and named dao-xing-ni-shi or 'acting in the reversed order' [68]. This is a good opportunity for malaria-endemic countries to use their ethnopharmacological heritage. The African Network for Drugs and Diagnostics Innovation has emphasized the role of African countries in valorizing their NPs [69]. They will continue to be a key source of new structural leads [70] and are expected to provide innovative chemotypes to develop antimalarials [71].

1.4 *In silico* drug discovery

Bohacek *et al.* estimated at 10^{63} molecules the size of the drug-like chemical space [72]. In the drug discovery landscape, from discovery to registration, *in silico* approaches mostly contribute to the discovery phase by filtering and selecting molecules with interesting activity from chemical libraries [73]. They offer time and cost-efficient solutions for mining that space [74] compared to the time and experimental cost of current drug discovery pipelines further undermined by a high attrition rate [75]. The Centre for Medicines Research data benchmark analysis indicated a new compound in preclinical evaluation only has an 8% chance to be part of a product [31]. The "Holy Grail" in virtual screening is to estimate accurately and precisely the binding free energy for billion of molecules at a practical cost. The cost combines technicalities of the setups, runtime, and required resources. Virtual screening workflows are usually set up on the tradeoff between cost and accuracy. Besides the quest for a practical and accurate affinity estimation, a compound needs to be optimized with respect to absorption, distribution, metabolism, excretion and toxicity (ADMET) properties in order for it to reach its target in enough concentration for activity without toxicity. Hence, drug discovery turns into a multi-objective optimization task with ADMET properties and affinity all embedded in the same structure [76,77]. A structural change for a better property might negatively impact affinity and vice-versa hence requiring a gait on tiptoes in the pharmacokinetic and pharmacodynamics space. In the following section, we describe some virtual screening approaches, especially the ones used in this work. The techniques are described in order of increased accuracy and cost with an overview of their theoretical background, differences, limits, and advantages.

Ligand-Based Virtual Screening (LBVS) approaches make use of active molecules and/or inactive ones and rely on molecular similarity to predict a molecule activity [78,79]. They may use target information that enhances performance [80]. Compounds are described in terms of 1D, 2D, or 3D descriptors, topology, pharmacophore, molecular field, shape and volume [79]. They tend to be less accurate than SBVS ones in general [80–82] even though this may simply be related to their lower usage [81]. On the other hand, they are faster and applicable to big data (billions of molecules) and do not require target structure knowledge [83,84]. The recent NIH Virtual Workshop on Ultra-Large Chemistry Databases highlighted fast search methods for bioactivity (Rapid Isostere Discovery Engine (RIDE), SmallWorld and Arthor [84]), cloud-based architecture, data compression strategies,

and synthesis of the virtual space [85] for practical multi-billion compound libraries mining for bioactivity [86]. Elsewhere, LBVS is also used to build focused libraries and for hit identification [79].

SBVS is the search for a ligand given a biological target structure. The method is based on the knowledge biological target three-dimensional structure. This structure can be obtained through x-ray crystallography, NMR spectroscopy or homology modelling. This knowledge enables the screening for or the de-novo design of compound with structure for optimum interaction with the target. The ultimate goal is to identify compound with a therapeutical effect. The approach uses computational approaches such as docking, MD and free energy calculations[87 88].

Molecular docking predicts a molecule's binding pose and affinity on a target. The first is done through sampling the protein-ligand conformational space. Notable search methods and programs relevant to small molecules are simple rigid docking (DOCK) [89], genetic algorithm (AutoDock) [90], particle swarm optimization (SODOCK) [91], (GOLD), incremental construction (FlexX) [92], Iterated Local Search global optimizer with parallelism using multi-threading (Vina) [93], and hierarchical approaches (Glide) [94 95]. More recently, global optimization (Monte Carlo) combined with essential Local and Location optimizations through BFGS method (QuickVina-W) [96] have been used. Given its acceptable accuracy (Kendall's tau rank correlation with experimental value up to 0.46 [81]) and ability for good pose prediction (predicted pose root-mean-square deviation (RMSD) <= 2 Å compared to crystal structure [97]), speed (order of seconds per "~< 25 torsion molecules" with recent significant speed gain on GPU architectures [98]), docking is widely used in virtual screening for hit identification. A recent large-scale application is OpenEye "GigaDocking" with the docking of REAL Enamine 1.43 Billion molecules on Purine Nucleoside Phosphorylase (PNP) and Heat Shock Protein 90 with a 24-hour runtime on the Orion cloud using ~27,000 CPUs for PNP [99]. A library of 170M compounds was also screened against AmpC β-lactamase (AmpC) and the D4 dopamine [100]. If posing is almost considered a solved problem, scoring has many pitfalls. This is related to docking atoms fixed partial charges, water in binding site treatment, proper H-bonds scoring, and receptor flexible [101]. Strategies to overcome these challenges include polarizable force fields, identifying structurally conserved waters, rigid or semi-flexible docking (only a few residues) receptors [102], targeted instead of blind-docking [96], and also Molecular Dynamics (MD) [103].

1.4.1 Molecular Dynamics (MD)

Given the difficulty of solving the Schrödinger equation for biomolecular systems with thousands of atoms, classical molecular dynamics are most often used. Classical MD in computational simulations is atomic movements driven by Newton's 2nd law (1-1) [103–105]. Beyond docking and its lock-and-key model, MD gives insight into protein dynamics, receptor flexibility and the movement of water molecules [102]. MD and its extensions (such as replica exchange dynamics (REMD), Steered molecular dynamics (SMD) and Umbrella Sampling (US)) have diverse applications in drug discovery especially for protein-ligand binding [106]: *in silico* validation of hits, identification of cryptic pockets, validation of binding poses [107], exploration of the energy landscape [103], and binding affinity estimation [108].

(1-1)

$$F_i(t) = m_i a_i = m_i \frac{d^2 r_i(t)}{dt^2}$$

$F_i(t)$: $Force\ acting\ on\ atom\ i$

m_i: $mass\ of\ atom\ i$

a_i : $acceleration\ of\ atom\ i$

r_i: $position\ vector$

t: $time$

Potential energy functions are also known as force fields (FFs), and these have made MD applicable to biological systems with the first simulation of a small globular protein (bovine pancreatic trypsin inhibitor) over 8ps in 1977. FFs are made of energy terms describing atomic interactions: short-range bonded and non-bonded ones (electrostatics, repulsion, dispersion) [109]. Bonded terms can be divided into torsional (between 4 atoms), bending (between three atoms), and stretching (between two atoms). The non-bonded forces include van der Waals and electrostatic terms. The system's total energy can be expressed using (1-2), (1-3), and (1-4) [110,111].

$$E = E_{bonded} + E_{nonbonded} \quad (1\text{-}2)$$

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral} \quad (1\text{-}3)$$

$$E_{nonbonded} = E_{electrostatic} + E_{vand\ der\ Waals} \quad (1\text{-}4)$$

Many software tools (GROMACS[112], NAMD, LAMMPS, GROMOS, CHARMM and AMBER) and FFs (AMBER, CHARMM, and OPLS) have been developed for biomolecular simulation [103,113]. Equation (1-5) represents the AMBER03 force field [114]. Yet, classical FFs do not capture key quantum effects due to a point charge atomic model. Some current limitations include the absence of polarization, charge transfer, charge penetration [109,115]. Moreover, in the case of transition metals, quantum mechanical ligand-field, spin-state, trans, and Jahn–Teller effects are more significant and not well captured in classical FFs [116,117]. Simulation time remains the order of nanoseconds in most cases or microseconds with coarse-graining and/or greater resources [118]. This may not be enough to explore many biological processes. Structures' high-energy states (transition states and/or rare conformations) are rarely sampled in MD [119].

$$E_{total} = \sum_{bonds} K_b\left(b - b_{eq}\right)^2 + \sum_{angles} k_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{torsions} \frac{V_n}{2}\left[1 + \cos\left(n\emptyset - \gamma\right)\right] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right] \quad (1\text{-}5)$$

In the AMBER03 force field (Equation 1-5), the bonded interactions are described by the first three terms. Harmonic potentials model covalent bonds and angles. K_b and $k_\theta\ are$ bonds and

angles force constants and θ_{eq} , b_{eq} are the equilibrium angles and bond length respectively. The variables $\emptyset$, V_n, γ represent the dihedral angle, the force constant and the phase angle respectively for dihedrals angles. The last two terms model the non-bonded interactions. R_{ij} is the distance between two particles. (A_{ij}) and (B_{ij}) are Van der Waals interactions and the London dispersion terms respectively. $q_i \, and \, q_j$ are the partial charges to model the Coulombic interactions. Finally, ϵ is the dielectric constant [114].

Recent advances in MD implementations on GPU architecture have reduced the overheard from polarizable force fields and enhanced sampling techniques [109]. In the Drude force field, instead of a single point charges model, electronic degrees of freedom are modelled with particles negative charges attached to their parents' atoms via harmonic springs with added computational cost but with some promises on GPU [120]. An approximation strategy through increasing hydrogen mass while reducing that one of their connected heavy atom enables a 4 fs timestep for better sampling. Hydrogens' high-frequency vibrations are reduced by increasing their masses [118]. There have been improvements in the packing of hydrophobic residues [121]. Machine Learning (ML) based force fields are promising with converged simulations and accuracy attaining quantum-chemical CCSD(T) for a few dozen atoms [122].

1.4.2 Free energy calculation

In the context of protein-ligand drug discovery, the binding free energy is the energy difference between the protein-ligand unbound and bound states. The unbound states is the ligand and protein free in solvent, while the bound one refers to the fully formed complex between the protein and the ligand. The binding energy can be calculated sampling many configurations between these two states. Free energy calculation methods can be divided into endpoint and alchemical methods. The first class only considers the bound and unbound states of the protein and the ligand. The latter samples the full reaction coordinate, ligand and protein-bound and unbound states, and the intermediate states [123].

1.4.2.1 MM-PBSA

MM-PBSA is an endpoint method with binding affinity and is calculated using Equation (1-6) [108]. Beyond the MM bonded terms, van der Waals and electrostatic ones, MM-PBSA includes polar, non-polar, and entropic terms. The polar term is obtained through the Poisson-Boltzmann (PB) equation (PBE) or the generalized Born (GB) one in the MM-GBSA variant. The non-polar term is estimated from the solvent-accessible surface area (SASA). It is an implicit solvation method considering protein and ligand desolvation energies. The entropic term S is evaluated through normal-mode analysis of the vibrational frequencies [108,124].

$$G = E_{\mathrm{bnd}} + E_{\mathrm{el}} + E_{\mathrm{vdW}} + G_{\mathrm{pol}} + G_{\mathrm{np}} - TS \qquad (1\text{-}6)$$

The method is imprecise with an inaccurate entropy evaluation and uses a uniform dielectric constant. It depends on the system and the solvent model. It has been noted that ligand polarity increases MM-PBSA uncertainty. This is further exacerbated when the ligand binds in charged binding pockets [124]. Current accuracy thus varies widely across systems [123,125]. However, in general, it is more accurate than docking scoring functions but less than alchemical methods

[123,124]. It has poor ranking power especially for compounds with a difference in affinity less than 12 kJ/mol [123]. Hence, it is used for *in silico* hit validation but less so for hit optimization.

Current strategies to overcome these limitations and other advances include the use of many independent simulations to achieve convergence, residue-specific dielectric constants [123,124], GPU acceleration for faster PBE solvers [126], and machine learning [127].

Alternative to MM-PBSA such as linear Interaction Energy, free energy perturbation and quantum mechanics (QM) approaches exist. Yet, they have higher computational cost but often have higher accuracy. Umbrella sampling approach for example used in this work is alternative method to calculate binding free energy but with higher computational cost [108,124,128].

1.4.2.2 Umbrella Sampling (US)

Alchemical methods can be further divided into four types: perturbation theory, histogram approaches, non-equilibrium work simulation, and thermodynamic Integration from constrained and unconstrained dynamics [129]. US is an enhanced sampling method combining non-equilibrium simulation and histogram approaches.

Because of limited sampling, the protein-ligand binding process is rarely sampled in MD. Approaches such as umbrella sampling use a biased potential along a certain reaction coordinate (RC or ξ); here the unbinding process of a complex can overcome the limitations of poor sampling and the free energy of the binding process can be evaluated [119]. Free energy calculation is commonly done through the Weighted Histogram Analysis Method (WHAM) [103]. It constructs the potential of mean force (PMF) or free energy along ξ. From the biased simulations, system configurations (windows) are extracted and a histogram, h(ξ), describing the probabilities of finding the system at individual locations along ξ is constructed. Each window is weighted by a factor dependent on the applied bias potential, giving information on the free energy in that window. It estimates the uncertainty in an unbiased P(ξ) to calculate the PMF with a minimal statistical error. The resulting PMF constructed is the free energy along ξ [119,130].

Windows for US can be generated using SMD in which a constant harmonic force pulls some chosen atoms along an RC [131]. In a protein-ligand system, the ligands are often pulled to form the unbound states [119]. SMD can also generate intermediate states useful for bond-forming and breaking reactions in QM/MM systems[131].

In the D3R Grand Challenge 2, in the binding free energy methods, a combined Jarzynski non-equilibrium pulling and umbrella sampling achieved a centered root-mean-square error (RMSEc) of experimental and predicted binding free energy difference of 0.94 kcal/mol and Kendall tau of 0.62. The method was the top performer in free energy set 2 of the challenge [81]. Another advantage of US is the ability to decompose the free energy into its Van der Waals and electrostatic contributions [129]. On the other hand, the main disadvantage of US is its more complex computational setup requiring configurations from a biased simulation. These configurations require adapted force and velocity in SMD for pulling in protein-ligand systems, the careful adjustment of windows size and force constraint, and there is an associated cost with sampling each window [103,119]. Jagdish Suresh Patel *et al.* have combined US with coarse-grained MD in a model that may be useful in reducing this computational cost [132].

1.4.3 *In silico* antimalarial discovery

Many studies have characterized the malarial parasite targets [49,52,53,55,133,134]. Recently, combining stage specificity and metabolomic profiling Murithi *et al.* have also identified the good potential of unexplored druggable pathways [135]. Computational approaches have contributed extensively to antimalarial discovery. In a systematic review on *in silico* approaches in antimalarial drug discovery between 2008 and May 2015, Anurak *et al.* identified 17 articles covering the topic. These studies used both ligand and structure-based approaches including molecular docking, homology modelling, 2D- or 3D-QSAR, and pharmacophore modeling. All of these make use of common virtual screening approaches. However, these studies focused on one target [136], where current antimalarial discovery strategies would benefit from a holistic approach in the elimination era [32]. More recently, Kushwaha *et al.* used molecular docking against *Plasmodium* orotidine 5-decarboxylase, plasmepsin 2, HSP90, PfATPase to find hits with better docking scores than their respective standard inhibitors [137]. From a library of thiazole-1,3,5-triazine derivatives docked on *Pf*-DHFR eight compounds with IC_{50} from 11.29 to 40.92 µg/ml against a chloroquine-resistant strain were identified [138]. Arshadi *et al.* built the DeepMalaria system using Graph Convolutional Neural Networks and evaluated compounds with respect to *P. falciparum* growth inhibition and mammalian HepG2 cell cytotoxicity. The authors made use of transfer learning, pretraining the model with weights transferred from a model trained on a large unrelated dataset to overcome the small sample size limitation. From this, DC-9237 was identified as a fast-acting compound inhibiting asexual stages [139].

1.5 Research problem statement and justification

Malaria is a major health concern with its parasite continuously developing drug resistance. Chemotherapy plays a key role in the fight against the disease, yet the current WHO recommended ACT is threatened. The recent COVID-19 crisis also has impacted some programs resulting in treatment and diagnosis being disrupted in 37 of the 64 endemic countries. Progress in the eradication of malaria has stalled in recent years and the 2020 GTS milestones were not met [7]. Despite many potential antimalarials in development, a complaisant attitude may significantly hamper previous efforts. In general, the approval rate is low in drug discovery [75] including antimalarials [31]. Additionally, drug resistance is occurring faster than drug approval [40]. Moreover, the quest for new MoA may be challenging due to the intrinsically greater attrition risk for new chemotypes [31]. With no effective vaccine yet, the biological threats remain and there is a need for the continuous development of antimalarials. *Plasmodium falciparum* is particular among all malaria-causing species being the most prevalent and deadly. A significant reduction was observed in recent years with a 97% incidence reduction in the Mékong region. However, resistance to artemisinin has been observed, here artemisinin is the current best antimalarial [7]. The further existence of a deletion in pfhrp2/3 genes has rendered this malarial diagnostic difficult [19]. Moreover, it is the most prevalent species (about 100%) in the WHO African region which held 82% and 94% of malaria cases and deaths worldwide respectively in 2019 [7]. Hence, the current work focuses on *P. falciparum*.

1.6 Aims

This project aims to contribute to antimalarial development using *in silico* approaches to find hit compounds for *P. falciparum*. These strategies have been adapted for the current trends of antimalarial development for elimination and control on the one hand. Besides, changes have been made to the virtual screening approach to overcome known pitfalls. The first chapter describes an integrated *P. falciparum* proteome-scale drugs repurposing pipeline to find hits. The second chapter focuses on a single target (PfDXR) screening but uses more extensive virtual screening methods through a consensus hierarchical LBVS-SBVS approach. Hits are further assessed using MD, steered MD, and free energy calculation through MM-PBSA and US to find potential PfDXR hits. Finally, the third chapter covers the SANCDB NPs library update, as a resource in the search for antimalarials and more generally in drug discovery.

1.7 Research objectives

To achieve the above aims, the following general objectives were defined:

1 Explore holistic approach to *in silico* antimalarials discovery through screening on a set of *Plasmodium falciparum* targets.

2 Explore drug repurposing strategies through screening FDA approved drugs on these targets.

3 Application LBVS and SBVS screening for identification of hits for PfDXR

4 Explore Advanced MD sampling and free energy calculations.

More specifically the approach was to:

1. Setup and assess a screening pipeline on *Plasmodium falciparum* targets (Chapter 2)
2. Identify hits from FDA approved drugs for potential repurposing using the pipeline (Chapter 2)
3. Conduct hit *in-vitro* plasmodial activity and human toxicity assessment (Chapter 2)
4. Perform consensus LBVS on a ZINC lead-like subset and a consensus query of DXR inhibitors (Chapter 3)
5. Perform docking and consensus scoring on LBVS hits (Chapter 3)
6. Conduct MD, SMD and US on docking hits (Chapter 3)
7. Update the SANCDB database with new compounds and their commercially available analogs (Chapter 4)

Chapter 2: Potential Repurposing of Four FDA Approved Compounds with Antiplasmodial Activity Identified through Proteome Scale Computational Drug Discovery and in Vitro Assay

2.1 Introduction

The number of high-resolution structures of drug targets allows for proteome scale screening. Currently, the PDB has more than 155,000 DNA, RNA and protein structures. About 73% of these structures are co-crystalized with one or more ligands. This vast experimental data offers an excellent mining opportunity for novel drugs. Indeed, it has already contributed to ~90% of the 210 new drugs FDA-approved between 2010 and 2016 [140]. These structures can also contribute to antimalarial drug discovery.

Despite the availability of many structures, only a few targets are really studied. Despite the identification of 2680 *P. falciparum* essential genes[53], many potential target structures remain unsolved. Indeed, the PDB data has only about 600 *P. falciparum* structures with high redundancy. For example, PfDXR and PfDHFR count up to 19 and 26 structures of these 600, respectively. This same pattern continues in the bioactivity data. Assessment of *P. falciparum* bioactivity data in ChEMBL showed that Hexose transporter 1, Dihydroorotate dehydrogenase and Plasmepsin 2 have significantly more bioactivity data than other targets. This may contribute to understanding a ligand series structure-activity relationship (SAR) on the same target, but this also impairs target diversity. Hence, the number of approved drugs may be restrained to few targets while the target space is broader and remains unexplored. Exploring new targets is certain to be beneficial in the case of malaria.

The malaria parasite has a complex life cycle. It has three different phases which have a long history of molecular co-evolution with its hosts (human and mosquitoes). This long co-evolution has made this parasite highly adapted to humans with several survival mechanisms [23,27]. Moreover, the high plasticity of the parasite genome together with its permissive nature contributes to its ability to develop resistance. Further, the parasite has shown its ability for *de novo* resistance, occurring without meiotic recombination. Also, drug action may be impaired through transporters in the parasite. For instance, chloroquine resistance occurs through the drug active H^+-dependent efflux out of the digestive vacuole [141]. Given these mechanisms, *P.*

falciparum has shown resistance to all used drug treatments including Artemisinin-based Combination Therapies (ATCs) [40]. Target Product Profiles (TPP) and Target Candidate Profiles (TCP) for malaria elimination have emphasized the need for compounds with new mechanisms of action and with multi-stage activity [22,37].

Screening against an array of proteins helps in the discovery of novel drug-target interactions [142]. The virtual screening pipeline usually focuses on single target screening. An exception may be for kinases, in which protein arrays have been developed since their inhibitors act through multitargeting [142,143]. Given the complex biology of the malaria parasite and its ability to develop drug resistance, targeting more pathways and proteins may be beneficial. This approach toward system biology fits complex disease models [144]. Further, a proteome-based approach can help identify multitarget compounds that will be less susceptible to resistance - drugs with pleiotropic modes of action may well be resistance-proof [141]. This could be the "holy grail" for malaria elimination, indeed, multitarget drugs already showed to have the longest lifespan in terms of clinical efficacy in the case of malaria [40]. There are desired characteristics for new antimalarial compounds, including transmission-blocking, and activity on blood and liver stages of the parasite lifecycle [50]. Only tafenoquine and primaquine are currently approved as liver-stage active compounds [145]. More liver-stage active drugs would require an exploration of the parasite targetome. Additionally, cross-docking helps to model drug cocktail activity to optimize their synergy. Combinatorial chemotherapy plays a key role in the fight against malaria, and specifically ACT has been one of the most effective drug regimens known [146].

In the specific case of *plasmodium*, large-scale virtual screening has been done but on a limited set of targets [144,147]. WISDOM-I and II are two notable projects aiming at exploring plasmodium targetome for virtual screening. WISDOM-I aimed at PfGST, PfDHFR, PvDHFR targets while WISDOM-II extended this set to the *P. vivax* orthologs[144,147]. A later study used a larger target set but in a target fishing exercise [148]. To our knowledge, the potential of structure-based drug discovery at proteome-scale has not yet been identified or explored in malaria especially in the case of *P. falciparum*.

Drug repurposing is a cost and time-effective strategy to face the problems of resistance and attrition. This strategy has been successfully applied and is promising in the case of malaria [40]. Doxycycline and clindamycin are antibiotics that have been successfully repurposed for malaria [149], and even heparin is being investigated as a potential antimalarial [150]. Through a comparative structural and sequence analysis, Ramakrishnan *et al.* identified a further potential 71 FDA-approved drugs for *P. falciparum* [151]. Many other drugs are being investigated for repurposing as antimalarial [39]. In addition to repurposing, efficiency indices may help with respect to the attrition problem. Efficiency indices may be used for better hit selection but also to guide their optimization [76,152]. They also fit the holistic approach philosophy by combining drug potency and pharmacokinetic properties. Efficiency metrics have been increasingly used in publications since their introduction [153].

The motivation behind this study is to set some basis for repurposing current FDA approved drugs for malaria treatment. The current pipeline approach is in line with malaria elimination requisites by using proteome scale virtual screening, target diversity and cost-effective screening strategies such as drug repurposing.

2.2 Methods

Our approach hence combines proteome scale screening, ligand efficiency metrics, and standardization and ranking strategies. The below-described approach is the final version of the screening pipeline; the iterative improvements that resulted in the described final approach are detailed in the discussion related to the methods section.

2.2.1 Data retrieval and structures preparation: starting from known drugs and clean targets.

Raw PDB data requires pre-processing for virtual screening applications, and this has driven the design of cleaner PDB subsets such as sc-PDB [154,155] (this database also provides additional information such as binding site similarity scores through the Shaper scores [156]). In this study, protein (MOL2 format) and ligand (SMILES) structures were retrieved from the sc-PDB (Screening Protein Data Bank) (release v.2017 frozen PDB data on 2016-11) [156] and DrugBank (version 5.1.2, released 2018-12-20) [157]. The protein structures with fewer missing residues were chosen as representative for protein structures having the same UniProt [158] IDs. Fewer missing residues is advantageous in modeling the full structure more accurately for MD. These structures were first modelled using Prime version 5.4 (r012) (Schrodinger2018-4) [159]. A final set of 36 proteins was used. DrugBank ligands were further filtered in the following way. Compounds not affecting and not targeting *Plasmodium spp* were selected using the DrugBank search menu. New molecules without known antimalarial activity were ideal for use in this prospective study. Next, only orally active, and rule of five compliant compounds were selected. These filters fit some of the criteria for new antimalarials such as the one of oral administration[50]. Finally, the compounds with the greater Quantitative estimate of druggability (QED) [160] were selected for pairs of highly similar compounds (Tc >= 0.8). This was to reduce the computational cost while maintaining good quality compounds and diversity, thus maintaining ideal conditions in the virtual screening process particularly with respect to searching for new scaffolds. Also, similar ligands are likely to have similar properties. Finally, 796 ligands were used. Structures' pdbqt format were generated using AutoDock Tools [161]. RDKit (version 2018.09.1) was used to calculate molecular properties [162] and Crippen's method for cLogP [163].

2.2.2 Optimized cross-docking, rescoring, standardization, and complex ranking pipeline assessed by the MRE.

The docking was first assessed with a minimalistic setup using blind redocking of co-crystalized ligands to all proteins (all-vs-all). This all-vs-all approach evaluates the pipeline ability to retrieve the true co-crystalized for a specific protein. QuickVina-W [96] was used as adapted for blind docking. A good quality pose has RMSD <= 2.00 Å when compared to co-crystallized one. This threshold is commonly used in docking pose evaluation studies [164]. RMSD values were calculated using GROMACS 2016 [112].

The above-mentioned setup runtime will scale approximately linearly with the number of docking experiments (number of protein times the number of ligands). Therefore, it is important to optimize running costs for a practical runtime, hence the use of Quick-Vina-W [96]. This enhanced version of Autodock Vina [165] has a mean and maximum normalized overall time acceleration of 3.60 and 34.33 fold respectively compared to Vina [96]. We evaluated the runtime of QuickVina-W

[96] vs Vina [165] and confirmed this improved speed on a test dataset (Figure 2-4d). Concerning pose and affinity predictions, the tool is reported to maintain similar or greater accuracy [96]. The method applies an enhanced search algorithm by spatiotemporal integration suitable for blind docking while using AutoDock Vina scoring function (SF)[165] [96]. In the current study, the exhaustiveness was adjusted with respect to the dimensions of every target box. The scaling factor used a reference value of 24 for a box dimension of 30^3 Å (3 X the default Autodock Vina [165] exhaustiveness value). Ten poses were predicted for each docking. Three CPUs per docking computation and eight jobs per computer node (24 cores per node) were used for internal parallelization and external parallelization for optimum computational efficiency. These parameters yielded the maximum efficiency with Autodock Vina [166] on a computer cluster.

GRIM (Grscore) and RF-Score SFs were used for rescoring. Their accuracy has been assessed using the MRE as illustrated in a comparable study[167]. In our all-vs-all 1296 (36X36) docking experiments, we obtain a matrix of scores $S_{[i,\ j]}$, where i and j are row and column indices, respectively. Proteins are in columns, while ligands are in the rows. A ranking error (Err_j) is computed for every row (set i of ligands on protein j) using equation (2-1). The diagonal of this matrix includes all the protein-co-crystallized ligand pairs. S_{jj} is the score for the co-crystallized ligand j on protein j. S_{jbest} and S_{jworst} are the scores for the best and the most ligands respectively on protein j. The range $S_{jbest} - S_{jworst}$ is the one of all scores on protein j. $S_{jbest} - S_{jj}$ is the score difference between the best ligand and the co-crystallized ligand. This must be or close to zero in an ideal scoring, as the co-crystallized should be or close to being the best ligand. The MRE is the mean of all Err_j (across all proteins). 1.0 is the worst MRE, 0.5 is for a random ranking, and ~0 for an ideal SF. Another alternative is to use the number of correctly identified co-crystalized ligands. Indeed, such an approach was also used in a similar study[167].

$$Errj = \frac{S_{jbest} - S_{jj}}{(S_{jbest} - S_{jworst})} \qquad (2\text{-}1)$$

From the resulting binding energy scores, Lipophilic efficiency (LipE) (equation (2-4)), Surface efficiency index (SEI) (equation (2-3)), and Binding efficiency index (BEI) (equation (2-2)) were calculated. These metrics have gained interest in identifying quality hits, since they rank better than sole potency [168]. Selection of hits based on high-scoring ligand efficiency metrics can lead to good quality leads [169–172]. Further, they can guide the compounds' optimization path through the efficiency plane [170]. Here we combined SEI and BEI which can be derived from the 2D efficiency plane properties [170]. Given the orthogonal nature of the two dimensions, we simply used the radial coordinate which corresponds to the square root of the sum (SEI^2 + BEI^2), i.e. ($\sqrt{SEI^2 + BEI^2}$)[173,174]. Efficiency metrics were also assessed with the MRE.

$$BEI = \frac{pIC50}{MW(kDa)} \qquad (2\text{-}2)$$

$$SEI = \frac{pIC50}{(PSA/100Å^2)} \qquad (2\text{-}3)$$

$$LipE = pIC50 - logP \qquad (2\text{-}4)$$

MW: Molecular Weight.

PSA: Polar surface area.

LogP was used for simplicity, since logD requires a pKa (dissociation constant) calculation and was not available in RDKit[162]. For the distribution coefficients of charged compounds, LogD is more accurate than the calculated partition coefficients (log P) [175]. The potency metrics IC_{50}, Kd, and Ki are interchangeable. The binding affinity was converted to Ki (dissociation constant of the enzyme-inhibitor complex) using equations (2-5) and (2-6) where ΔG is the binding affinity (kcal/mol), $R = 1.98$, and $T = 298.15$ K. pKi is obtained from equation (2-6).

$$K_i \text{ (unit in Molar)} = \frac{e^{\frac{1{,}000 \times \Delta G}{RT}}}{1{,}000} \tag{2-5}$$

$$\text{pKi} = \log 10(\text{Ki}_{\text{Molar}}) \tag{2-6}$$

Binding site characteristics (depth, size, hydrophobicity may cause significant variation in binding affinities. This makes it challenging to compare a ligand affinity on two targets. Indeed, this inter-protein scoring noise has been shown in various works. Score standardization techniques were suggested to minimize it [167,176–178]. Here, scores were transformed to their z-score by deducting the mean and then dividing by the standard deviation. This was applied, per column (all ligands' scores on every protein) and afterward to every row (a ligand's scores on all proteins) to obtain the z-score (Figure 2-2). The z-score was computed using SciPy[179] (equation (2-7)). The standardization strategy was applied on scores of every protein, centering them around a mean of zero with a standard deviation of one. This minimizes the inter-protein scoring noise. Hence ligand scores on two proteins may be compared. A comparable phenomenon was noted for ligands. These tend to have a greater affinity associated with their increase in molecular weight, causing false positives in docking. This was noted with Vina. Normalizing this bias was shown to enhance ligand affinity ranking in VS [177,179,180].

$$z - score = \frac{x - \mu}{\sigma} \tag{2-7}$$

On the resulting standardized scores, the complex ranking was applied to reveal protein-ligand pairs having a high reciprocal affinity in the dataset[142]. Each score was transformed to its rank. Where scores were equal, the mean of the score ranks was used. Hence ranks may have non-integer values. A complex rank is defined as the total of the protein rank plus ligand rank as in equation (2-8). The $\text{ligand}_{\text{rank}}$ respective to a protein is its rank compared to all other ligands. Similarly, the $\text{protein}_{\text{rank}}$ relative to the ligand is the rank of the protein compared to all other proteins. Hence, a $\text{complex}_{\text{rank}}$ is simply the total of the ligand and protein ranks. This enables uncovering protein-ligand systems having high mutual specificity, with a low rank (~2) being used to filter out false positives.

$$\text{Complex}_{\text{rank}} = \text{ligand}_{\text{rank}} + \text{protein}_{\text{rank}} \tag{2-8}$$

Efficiency indices have been used solely or in combination (SEI/BEI, LLE-LELP efficiency planes) in drug discovery pipelines. This rank transformation allowed us to integrate LipE, SEI, BEI as well as the Grscore. This addition provides a more holistic approach. Hits selected using efficiency metrics only may not maintain key interactions as the ones of the co-crystalized. Integration of the Grscore helps to take into account these interactions as GRIM scores molecular interactions Grscore [181].

2.2.3 MD simulation

All MDs were done as described here to assess protein-ligand complex stability, and thus to remove false positives. Indeed, MD is often used in the last stage of virtual screening pipelines [103]. Due to force field limitations, metal ions (MG in 1D5C, MG 1P9B, MN in 2PML and MG in 3FI8) were removed in MD. Cofactors were retained in the structures in both MD and docking. Hydrogen Mass Repartitioning (HMR) was applied to the structures. Masses of hydrogens connected to heavy atoms were repartitioned enabling an increased 4-fs time step. HMR consists in raising hydrogens' masses by a factor of four and deducting the added mass from the connected heavy atom as explained in GROMACS documentation [182]. The system total mass is conserved. HMR has been shown to accurately speed up MD[118,183–185]. Ligands' charges were obtained from Discovery Studio Visualizer V1.7 and their topologies generated with ACPYPE[186]. A dodecahedron box with a 1.0 Å distance between the box and solute and the tip3p water model were used with a concentration of 0.15 M (Na^+ (sodium) and Cl^- (chloride) ions). Steepest descent was used for energy minimizing using a max force of < 1000.0 kJ/mol/nm and a maximum of 50000 steps. Systems were equilibrated at 300 K and 1 atm with 50 ps MD in the isothermal-isobaric ensemble and later in the canonical ensemble. The Lennard-Jones and the short-range electrostatic interaction cut-offs were set at 10 Å. The smooth particle mesh Ewald and a fourth-order interpolation were used for long-range electrostatic interactions. Leap-frog algorithm was used for integration during the twenty nanoseconds MD. Simulations were performed on a remote computer at Center for High-Performance Computing (CHPC) using GROMACS [112] version 2018.2, with the Amber ff99SB-ILDN [187] force field. After the simulations, the GROMACS[112] module trjconv was used to adjust for periodicity. Protein rotation as well as translation were removed by fitting it to the initial structure. Nglview [188] and Pytraj [189] were used for analysis and visualization in a Jupyter Notebook[190]. The analysis metrics were clustered into geometry related (radius of gyration) Rg, RMSD, protein-ligand center of mass (COM) distance), interactions (hydrogen bonds) and finally energy-related (protein-ligand interaction energy) ones. Analysis involved initial evaluating proteins' structural stability via their RMSD and the Rg calculated using the corresponding GROMACS [112] modules. Rg is related to the overall compactness of the protein, which can thus assess structure instability, especially when unfolding[191]. It can also be linked to the different protein folds [191]. An increasing Rg indicates a less compact structure. The ligand heavy atoms RMSD was fitted to the backbone of the protein. This metric has been shown to better capture ligand stability [192]. Additionally, its interaction energy, COM distance to the protein COM ("COM" is used for simplicity), and hydrogen bonds were used.

The current pipeline combines two concepts. First, we use a holistic approach through proteome-scale docking, and through the use of multiple metrics (energy and molecular properties (ligand efficiency) and interaction scoring (GRIM)) in scoring the ligands. Secondly, we explored the drug

repurposing aspect. Further MD simulations including HMR schemes were done. Figure 2-1 represents the overall screening workflow.

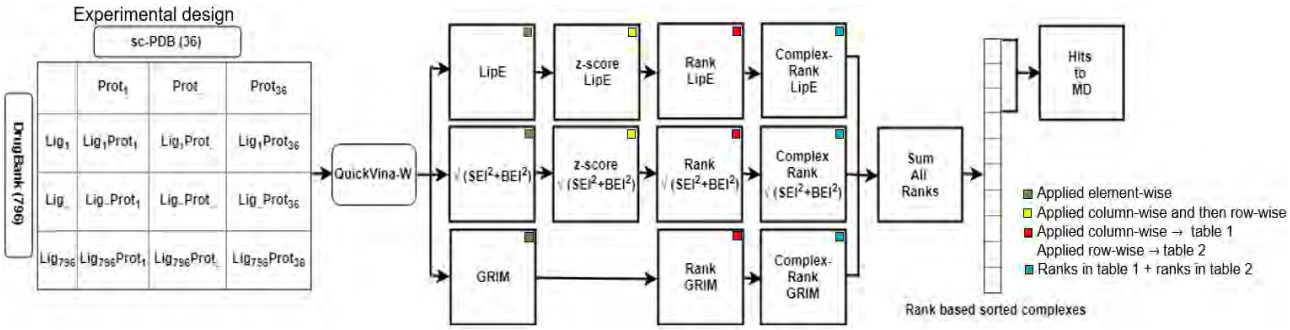


Figure 2-1 Overall screening pipeline from left to right. The table in the experimental design carries on throughput the workflow. Square boxes represent tables of protein-ligand complexes as described in the experimental design. The metrics in the boxes represent the values in the corresponding table. The transformations in each table are shown in the color code.

2.2.4 Antiplasmodial and human cytotoxicity assays

Compound antiplasmodial activity evaluation was done against the *Plasmodium falciparum* 3D7 strain. The method has been fully described previously[193]. As a pre-screen, the cultured parasites were incubated with each compound at 20 µM for 48 hours. A control of untreated parasites was also used. The plasmodium lactate dehydrogenase (pLDH) assay[194] determined the parasite viability percentage relative to the control. The assay and the 48-hour incubation were repeated in 3-fold serial dilutions. IC_{50} evaluation was performed for active compounds, the ones that decreased parasite viability below 50%. Their values were calculated through non-linear regression analysis of parasite viability % vs. log[compound].

Active compounds' human cytotoxicity was evaluated on HeLa cells (human cervix adenocarcinoma). Compounds were incubated at 20 µM in three-fold serial dilutions (100 to 0.0457 µM) in a 96-well plate. An untreated control well was also used. HeLa cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) with 5 mM L-glutamine (Lonza), supplemented with antibiotics (amphotericin B/streptomycin/penicillin) and 10% fetal bovine serum (FBS) at 37 ºC in a 5% CO_2 incubator for 24 h. The surviving cells to drug exposure were counted using the resazurin based reagent and resorufin fluorescence quantified (Excitation560/Emission590) in a SpectraMax M3 plate reader (Molecular Devices) [195,196]. The wells fluorescence readings were converted to cell viability percentage relative to the control average readings, after deducting background readings from wells without cells. Cell viability % vs. log[compound] plots were used to determine IC_{50} using GraphPad Prism (v. 5.02) through non-linear regression.

2.3 Result

2.3.1 Pipeline accuracy assessment: 77% correct poses and an MRE of 0.08

The pipeline accuracy was first assessed in the all-vs-all experiment. Redocked co-crystallized ligands' poses RMSD, binding energies, their standardized values and ranks are presented in Figure 2-2. Diagonal cells on the heatmaps show protein and co-crystallized ligands pairs. 77% of ligands were docked accurately considering that poses with the lowest RMSD have RMSD <= 2 Å (Figure 2-2a). Comparable percentages of accurate pose were found in pose accuracy studies[164,197]. Quick-Vina-W showed similar accuracy in its original paper [96].

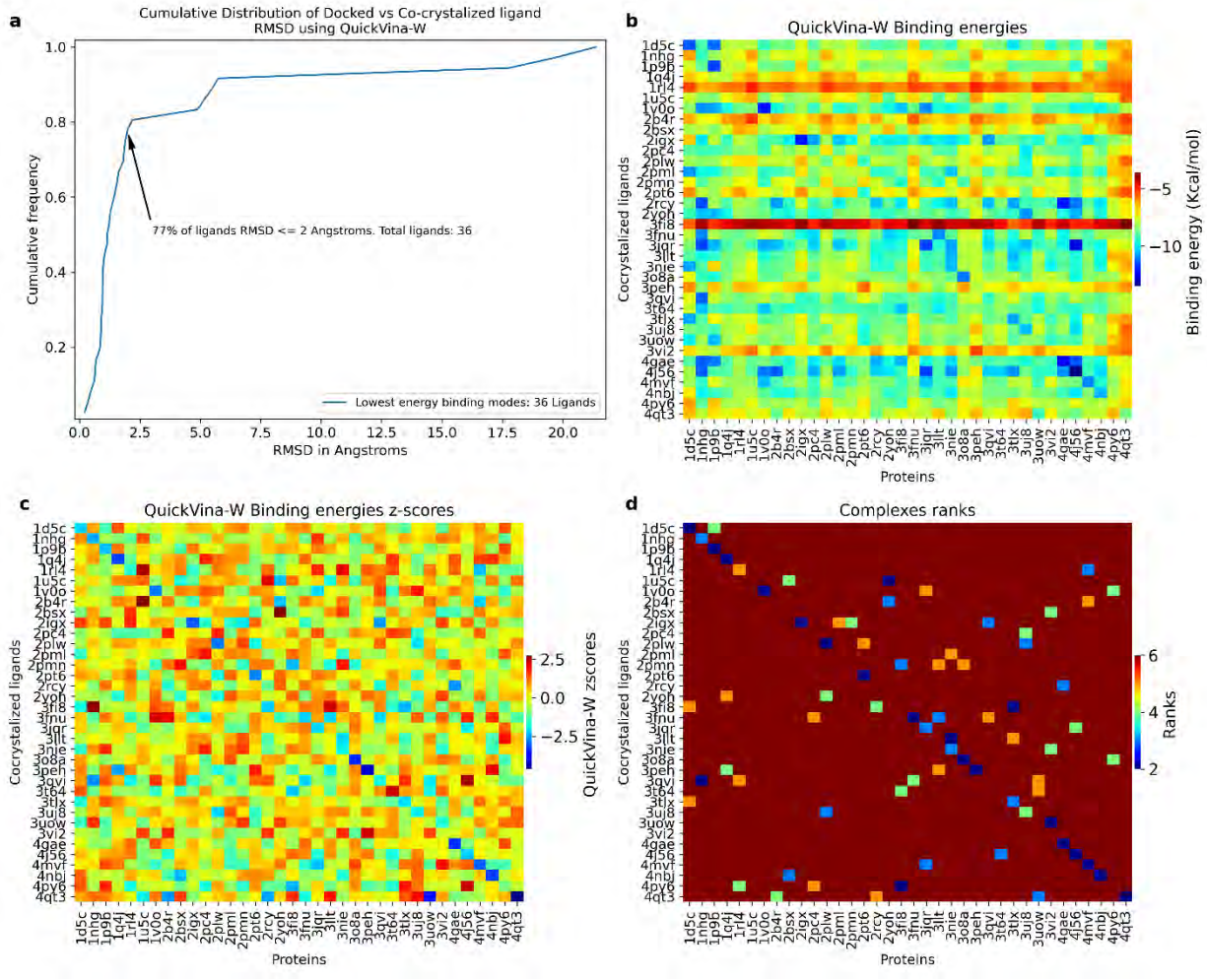


Figure 2-2 Workflow assessment validation of docking poses and scores transformations. a. Docked vs co-crystallized poses RMSD cumulative distribution. b. QuickVina-W[96] binding affinities. c. Standardized values; d Complex ranks (for clarity, only complexes ranks ≤ 6 are shown. Rows (ligands) and columns (proteins) are alphabetically ordered on the heatmaps. The figure was produced using Seaborn version 0.9[198].

Co-crystallized ligands in these proteins docked with RMSD above 2 Å: 1U5C (4.8 Å), 2b4r (21.3 Å), 2pc4 (5.7 Å), 2rcy (2.1 Å), 3JQR (5.4 Å), 3QVI (19.7 Å), 3uow (17.7 Å) and 3vi2 (5.1 Å). Although

in some of these cases the ligand-bound in the active site area, the co-crystallized ligands in 3QVI, 3uow, and 2b4r had RMSD deviations above 6 Å: 21 Å, 19 Å, and 17 Å, respectively after docking. In investigating the cause for these RMSD deviations, the probability of docking program error was first ruled out by trying other programs with different search algorithms: Vina[93], Smina[199], and VinaXB[200]. AES has a halogen atom, and for this case, VinaXB[200] which takes into account halogen bonds could be helpful. However, all of these programs gave similar results to Quick-vina-w[96] with the RMSD attaining (0.01 Å) when comparing poses generated by the four programs. Following this, structural analysis was performed to examine alternative receptor conformations, water-mediated interactions, resolution, flips of asparagine, glycine or histidine, missing residues, particularly in the active site and protonation states at the protein working pH. As such, careful inspection of input files, their combinations, and other aspects influencing docking were conducted on these structures. In 2b4r, MolProbity [201] indicated a flip of ASN185, an active site residue (Figure 2-3). The co-crystallized ligand (AES) was redocked with an RMSD of 1.23 Å with ASN185 flipped. This could be the residue correct conformation. It is additionally notable that AES in 2b4r was assigned to an unexpected electron density[202]. On the other hand, the histo-aspartic protease (3QVI) works at low pH (5.5)[203]. The correct pose was not reproduced using that protonation state pH using Schrodinger[204]. Inspection of the structure showed that the co-crystalized ligand is bound into an unusual δ-turn conformation. A tight domain-swapping makes the flap pocket (enzyme active site) inaccessible[205]. For 3uow, MolProbity [201] showed a flip of GLN476 (chain A). The residue in that conformation and the above-mentioned changes did not allow reproduction of the correct pose. An explanation might be the rigid nature of the receptor. A considerable conformational change happens upon binding of XMP[206,207]. This latter may be binding to its initial binding site before the induced conformational change leading to the co-crystallized one. Flexible residues or an induced-fit docking might be a better choice. These three cases might be a limitation. As 3uow, 2b4r, and 3QVI conformations did not reconstruct the respective co-crystallized ligand poses, only considering successfully redocked targets would have been a better strategy for the rest of the experiments.

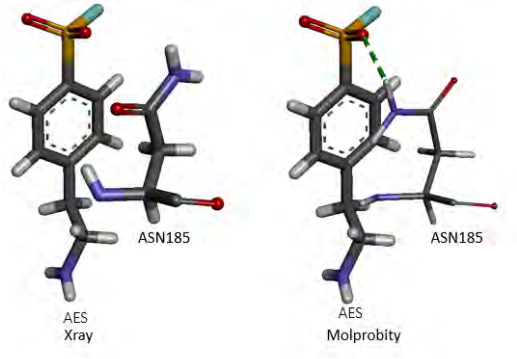


Figure 2-3 2b4r-ASN185 conformations in the crystal structure (left) and flipped from the Molprobity (right). AES redocked successfully with the flipped conformation with ASN185 AND AES forming an hydrogen bond.

These docking failures were not expected, as sc-PDB is a dataset designed for molecular modelling. In the structure preparation for this dataset, hydrogens were added to structures

taking into account the ionization state of titratable groups[208]. For other residues, hydrogens were added according to ionized templates built from HET group dictionary which contains information on hydrogen atom connectivity and bonding type[208]. The intermolecular hydrogen bonds were optimized using the BioSolveIT Hydescorer program. No missing residue was noted in the active site areas. The worst resolution among the selected structures was 2.8 Å. Sc-PDB is specifically designed for docking methods[156]. Thus, we assumed that the set of proteins and co-crystallized ligands suitable for this study. Yet, as indicated, we observed that some redockings were challenging due to residue flips, or due to receptor conformations. The quality of the sc-PDB structures could be further improved with residue flip analysis. Validating structures through redocking might improve the quality of the dataset for virtual screening.

In the initial stages of this project, all DrugBank compounds and all *P. falciparum* structures (~600) were retrieved from PDB to construct the target set for screening. This set was associated with many challenges including structure and binding site redundancy, target validity, binding site druggability, docking runtimes, target size for blind docking, and all the above-required structure preparations. This lack of preparation caused poor redocking accuracy. Only 61% of ligands had RMSD <=2 with Qvina-w while this percentage was 51% for Vina. On a test set of 560240 dockings, q-vina and vina had an average runtime per compound of 148.77s and 399.68s respectively (Figure 2-4d). Despite its faster speed of q-vina, the total runtime required to complete the docking for the entire set, using the available 240 CPUs, was 48 days and was impractical for the scope of the project. This required reduction of the ligand and/or target set sizes. Filtering redundant structures (90% sequence identity) gave a set of ~235 structures. These structures were then classified to estimate *P. falciparum* target space coverage (Figure 2-4b). Many of them were not suitable for small molecule drug discovery (especially antigens, transporters, immunoglobulins etc.). Fpocket [209] filtered for druggable pockets (Figure 2-4a). Considering only pockets having a druggability score equal to or above 0.4 (the Fpocket threshold for druggable pockets), a final set of 61 structures for a total number of 367 binding sites was selected. Cofactor sites and sites other than the active site were considered. Also, homodimers and homotetramers for example have multiple copies of the same binding site. This multiplicity in the binding site allowed for additional strategic analysis in the aggregation of results from multiple targets.

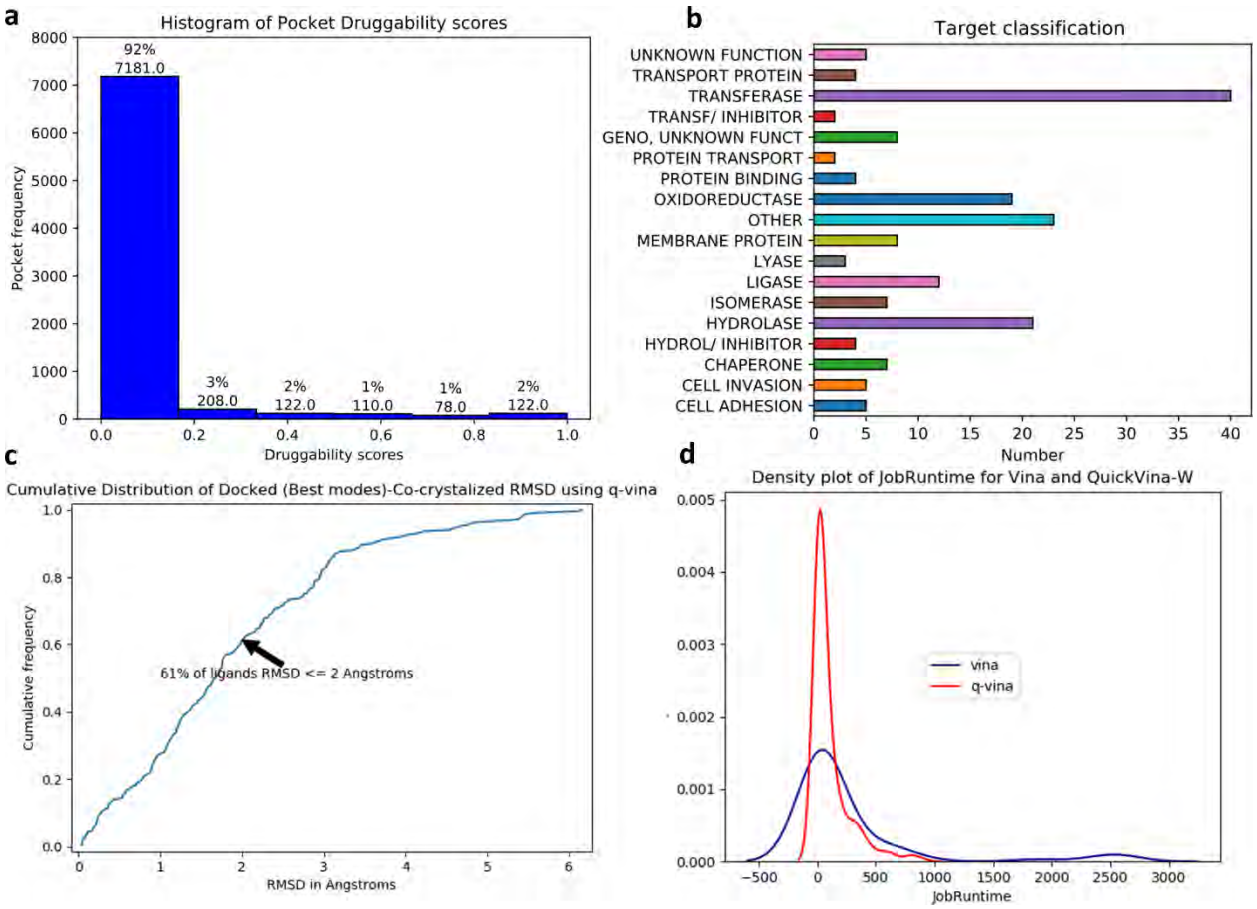


Figure 2-4 Challenges associated with the initial PDB set preparation for screening. **a**. Histogram of Fpocket druggability scores. **b**. *P. falciparum* Structures classification. **c**. Cumulative distribution of QuickVina-W[96] generated poses RMSDs on a test set. **d**. Runtimes (in seconds) density pots for Vina (blue) and QuickVina-W[96] (red).

One of the initial objectives was to investigate for multitarget binders for which binding site similarity between structures is key. The availability of that information, co-crystalized ligands that allow use of the GRIM SF, and all the above-mentioned challenges associated with the initial PDB set motivated the choice of using the sc-PDB dataset.

Binding energy standardization minimized protein and ligand-related biases. 1NHG and 4qt3 had the highest and lowest average binding energies (-9.11 kcal/mol and -6.50 kcal/mol respectively) (Figure 2-2b and c), which corresponds to a difference of 2.61 kcal/mol. This might be explained by the buried active site in 1NHG compared to the greater solvent-exposed active site of 4qt3. The standardization procedure lowered this inter-protein noise by centering the mean of binding energies on each protein at zero. Similarly, with ligands, 2-aminoethyl dihydrogen phosphate and (2R)-2- [(hydroxy-- amino)methyl] hexanoic acid co-crystallized ligands in 3FI8 and 1RL4, have low binding energies across all proteins (Figure 2-2b). Their low molecular weights 141 Da and 189 Da respectively may explain their reduced scores. Vina SF has a ligand size-related bias[180]. By comparison, the co-crystallized ligand from 4J56, flavin-adenine dinucleotide, has a high molecular weight (785 Da) which might explain its high promiscuity. Moreover, it presents

multiple centers for H-bonding. Overall, the workflow assessment showed that 77% of the redocking was successful as evidenced from their RMSDs (<= 2 Å) (Figure 2-2). Score standardization removed ligand and protein-related biases. Finally, the complex ranking revealed mutually selective protein-ligand pairs.

In, addition to QVina-W SF, GRIM and RF-Score were also evaluated for their ability to retrieve original pairs of protein-ligands. Figure 2-5 shows MRE values for the different scoring schemes.

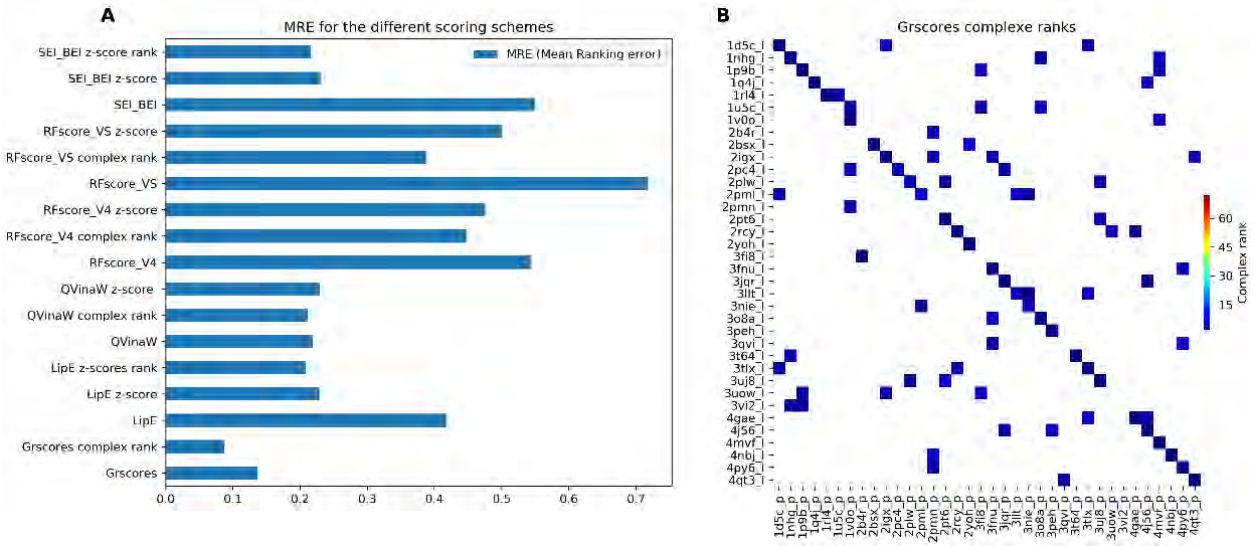


Figure 2-5 MRE values and Complexes ranks from Grscores. **A.** Bar chart of the MRE values for the different scoring schemes. **B.** Heatmap of the Grscores complex ranks described in this chapter Methods section (only complexes with a rank value ≤ 6 are shown for clarity. On the heatmaps, rows (ligands) and columns (proteins) are alphabetically ordered. Being similarity scores, Grscores were not standardized. SEI_BEI is the radial coordinate ($\sqrt{SEI^2 + BEI^2}$) . LipE z-score, LipE z-score complex rank is the standardized value and complex ranks derived from LipE. A similar naming pattern is used for QVina-W, RF-Score-v1 and v4, and SEI_BEI.

The best MRE (0.08) was obtained with Grscore after complex ranking. LipE and SEI_BEI had MREs of 0.20 and 0.21 respectively in their complexes ranking. This MRE value is comparable to the QuickVina-W[96] one (0.21). Remarkably, the machine learning SF, RF-Score, yielded the highest MRE values. Indeed, both RF-Score VS and also RF-Score-v4 were found to rank poorly the co-crystallized ligands by having an MRE of 0.71 and 0.54 respectively. In both versions, RF-Score VS and v4 had an MRE greater than 0.5, worse than random, and for that reason, they were not used in subsequent experiments or in the final version of the pipeline.

The best MRE with the Grscore may simply be explained by the accuracy of docked poses. These docked poses would give accurate molecular interactions which are used in GRIM scoring approach. Also, the GRIM approach may be more advantageous as the co-crystalized ligand is used as a reference and the MRE evaluates the tools' ability to retrieve the correct protein-co-crystalized ligand. Simple logic would drive us to solely choose this metric for further screening. However, we considered the importance of the consensus approach [142,210,211]. Further, the DrugBank ligands used in screening are not the co-crystallized ligands. Finally, analysis of drug

attrition has highlighted the importance of including molecular properties in the early stages of drug discovery [171,212], and so we decided to integrate the efficiency metrics.

MRE enhancement through scores' standardization plus complexing ranking is well-observed as the various scoring strategies are implemented (Figure 2-5A). For instance, one can note a reduction in the MRE starting with SEI_BEI to SEI_BEI z-score and finally to SEI_BEI z-score rank. Likewise, a similar trend is noted in RF-Score VS and RF-Score-v4, and LipE. Complex ranking produced the lowest MREs whereas the highest values are observed for the scores before standardization and complex ranking (Figure 2-5A).

Ligands in 3QVI, 3uow, 3vi2, 1U5C had a Grscore of 0.57, 0.56, 0.57 and 0.58 respectively. The value 0.594 is the threshold to distinguish similar from dissimilar co-crystallized ligand interaction patterns [181]. This may be caused by the absence of good binding poses as shown above in their pose assessment. Indeed, Grscore scores the molecular interaction similarity which inevitably depends on pose quality.

Despite having good ligand poses (RMSD < 2) 3FI8 had a complex rank of 16. The system had a protein rank and ligand rank of 11 and 5, respectively. Hence, the protein rank in this case is the main contributor to the poor complex rank. The co-crystallized ligand has a simple interaction consisting of hydrogen bonds with GLN290, ASP288, and a water molecule. The simplicity of the interaction pattern may make its reproducibility easier. Hence, many other ligands may bind similarly, leading to the target having a high average Grscore. Indeed, comparing its Grscore to the other systems, 3FI8 target had the highest average Grscore. This observation highlights the dependence of the Grscore on the reference ligand molecular interaction complexity.

Complexes OPE603 (ligand from original protein:3FI8)-2b4r and ANP (ligand from original protein:3LLT)-3nie had a complex rank of 2 while not being original complexes. These two false positives may explain the limit in Grscore success. The ligand in 3FI8 (OPE603) binds around AES602 (in 2b4r chain P) and interacts with AES602 (Figure 2-6a). It is also noteworthy that a residue in AES602 binding site in 2b4r did not have the correct conformation. MolProbity [201] showed flip of ASN185 AES602 binding site. OPE603 binds with a high Grscore (0.74) which may be explained by its interaction with the reference ligand.

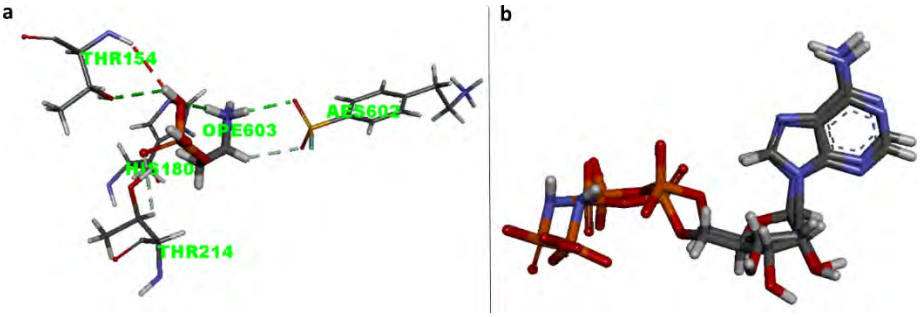


Figure 2-6 **a.** Interactions OPE603-AES602. b. Redocked ANP in 3LLT superimposed with co-crystallized ANP in 3FI8. The image is rendered using Discovery Studio Visualizer 2017 R2.

In the second case, phosphoaminophosphonic acid-adenylate ester (HET CODE: ANP) was co-crystallized with two kinases (3LLT and 3nie). In the redocking, ANP in 3LLT was found binding in

3nie in a pose close to the native ANP in 3nie (Figure 2-6b). This may explain their low complex rank of 2.

2.3.2 Top predicted complexes

GRIM and ligand efficiency were added to the workflow (Figure 2-1). 796 drugs from DrugBank [213] were docked on 36 Plasmodium falciparum targets. LipE, BEI, and SEI were computed for every ligand's best binding energy. All poses were rescored using the GRIM tool to get the Grscore and the one having the best Grscore was used for ranking. LipE, SEI, BEI scores were standardized via the z-score. A complex rank was computed from scores rank transformation as described in the method section. Complexes ranked list was obtained by summing all metrics relative ranks and ranked accordingly. One ligand was selected for each target and assessed in MD. 25 stable complexes were kept (Table 2-1).

Table 2-1. Top predicted ligand (names) with their predicted targets PDB IDs, and compounds names, DrugBank IDs, binding energies, GRIM Grscores, and ligand efficiency (LipE, SEI, BEI) values. Ligands are sorted by their mean of PLIE.

Protein name (PDB ID)	Compound name (DrugBank ID)	ΔG Q-vina (Kcal/mol)	Grscore	LipE	BEI	SEI	ΔG PLIE[1] (Kcal/mol)	Plot Labels
Thioredoxin reductase 2 (4J56)	Prazosin (DB00457)	-11.4	0.71	9.6	30	11	-331.42	22
Phosphoethanolamine N-methyltransferase (3UJ8)	Abacavir (DB01048)	-9.2	0.75	8.7	34	10	-261.05	20
Protein serine/threonine kinase-1 (3LLT)	Sitaxentan (DB06268)	-10.5	0.59	6.8	24	10	-231.08	17
Spermidine synthase (2PT6)	Sotalol (DB00489)	-7.9	0.62	7.7	32	11	-224.17	12
Dihydroorotate dehydrogenase (3O8A)	Nadolol (DB01203)	-9.0	0.78	9.0	31	12	-214.64	18
L-lactate dehydrogenase (1U5C)	Gemifloxacin (DB01155)	-8.2	0.67	8.1	23	7	-213.04	6
D-aminoacyl-tRNA deacylase (4NBJ)	Triamcinolone (DB00620)	-9.4	0.63	9.3	25	9	-205.09	24
Protein kinase 7 (2PMN)	Lamotrigine (DB00555)	-7.8	0.69	6.7	34	10	-201.50	11
Protein kinase domain-containing protein (2PML)	Terazosin (DB01162)	-9.0	0.62	8.6	25	9	-190.83	10
Bromodomain protein putative (4PY6)	Pirbuterol (DB01291)	-6.7	0.69	7.2	33	9	-190.13	25
Glutathione S-transferase (1Q4J)	Saxagliptin (DB06335)	-9.4	0.63	8.8	31	11	-187.68	4
Plasmepsin 2 (2IGX)	Fingolimod (DB08868)	-7.9	0.79	5.6	29	13	-177.18	9
Purine nucleoside phosphorylase (2BSX)	Temozolomide (DB00853)	-7.5	0.74	10.6	44	8	-175.49	8
Choline kinase (3FI8)	Tenoxicam (DB00469)	-8.8	0.74	7.8	28	10	-173.48	14
GTPase (Rab6) (1D5C)	Dianhydrosorbitol 2,5-dinitrate (DB00883)	-7.7	0.73	9.7	37	7	-171.46	2
Calcium-dependent protein kinase 2 (4MVF)	Abiraterone (DB05812)	-11.0	0.69	5.7	32	33	-166.17	23
Ferredoxin–NADP reductase apicoplast (3JQR)	Grepafloxacin (DB00365)	-9.5	0.71	7.7	28	13	-165.17	16
Cell division control protein 2 homolog (1V0O)	Anastrozole (DB01217)	-10.1	0.70	7.5	36	13	-163.73	7
Plasmepsin III (3FNU)	Darifenacin (DB00496)	-9.5	0.73	6.0	23	18	-152.76	15
Adenylosuccinate synthetase (1P9B)	Tafamidis (DB11644)	-10.3	0.71	6.1	34	17	-151.97	3
Thymidylate kinase (2YOH)	Salbutamol (DB01001)	-7.7	0.70	7.4	36	12	-151.25	13
Peptide deformylase (1RL4)	Ruxolitinib (DB08877)	-9.0	0.69	6.2	31	12	-150.06	5
Enoyl-acyl carrier reductase (1NHG)	Moclobemide (DB01171)	-9.0	0.73	8.2	36	23	-146.70	1
Histo-aspartic protease (3QVI)	Stavudin (DB00649)	-7.6	0.60	9.3	38	10	-142.12	19
1-deoxy-D-xylulose 5-phosphate reductoisomerase (4GAE)	Dihydromorfinon (DB00327)	-9.6	0.67	8.4	35	20	-140.28	21

1: PLIE: Mean of Protein-ligand interaction energy in MD.

2: ΔG Protein Ligand Q-vina Binding energy

3: Plot labels: Drugbank compound labels on the plot in Figure 2-7.

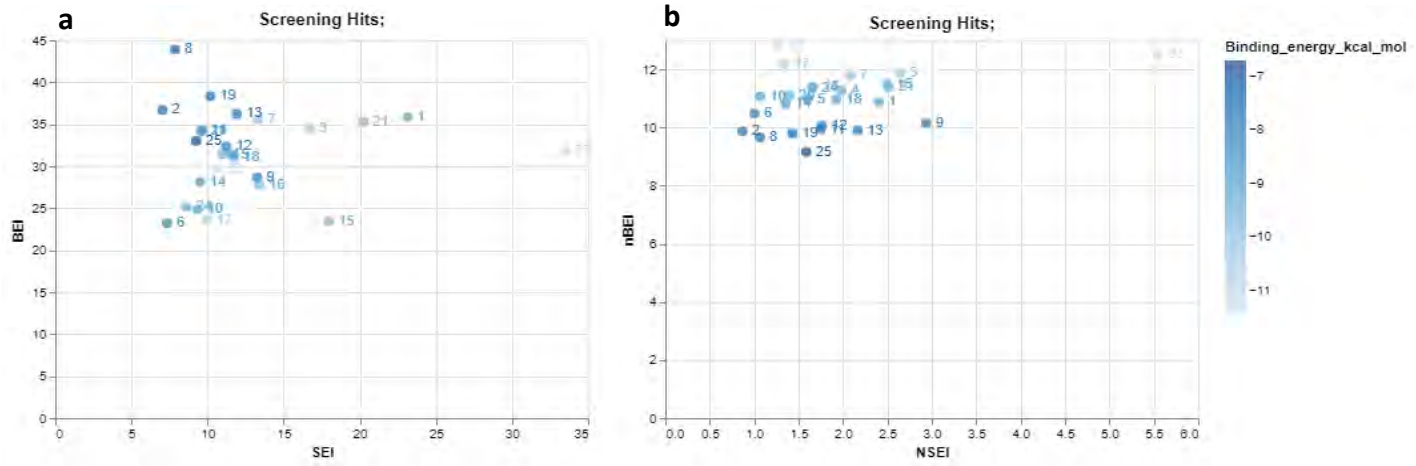


Figure 2-7 Scatter plot of the screening hits on the efficiency planes subplot **a**: (SEI/BEI), subplot **b**: (NSEI/nBEI). Points' labels represent the DrugBank IDs of the hits. The colors bar represents the binding energies on their respective best predicted targets. Plots labels mapping are in Table 2-1. Formula for $nBEI = log10(Ki_{\mathrm{Molar}}/NHA)$ and $NSEI = log10(Ki_{\mathrm{Molar}}/NPOL)$. With NHA and NPOL being the number of heavy atoms and the number of polar atoms respectively.

Efficiency metrics were found to be in an acceptable range for the different efficiency metrics thresholds. Minimum BEI and LipE values were 23 and 3, and their lower accepted thresholds are 3 and 27[171] respectively. Kumar *et al.*[214] defined the value 15 as a lower threshold for SEI. In the set of identified ligands, the average SEI value (12) was less than this SEI lower limit. Indeed, the compounds dianhydrosorbitol 2,5-dinitrate (Total Polar Surface Area (TPSA) of 123.2 Å^2), gemifloxacin (TPSA: 127 Å^2), temozolomide (TPSA: 105.94 Å^2), triamcinolone (TPSA: 115.06 Å^2), pirbuterol (TPSA: 85.61 Å^2), terazosin (103.04 Å^2), tenoxicam (99.6 Å^2), abacavir (101.88 Å^2), and Sitaxentan (TPSA: 107.73 Å^2) were found to have SEI values lower than 10. These molecules tended to have high polarity values. 7 (gemifloxacin) and 33 (for abiraterone) were the lowest and highest SEI values, respectively. For abiraterone, the compound's high hydrophobicity explains its high SEI value. Indeed, it has a PSA of 33 Å^2 and a cLogP of 5.3.

Figure 2-7 represents the scatter plot on SEI/BEI and NSEI/nBEI for the 25 DrugBank hits compounds. Most hit compounds are in the favorable region around ideal values of nBEI (10.5) and NSEI (1.5)[215]. This was expected about these are marketed drugs and thus likely to already have good molecular properties. Only compound labeled 23 (DB05812), Abiraterone is in the fast east region of the graph. This may be linked to its hydrophobicity. Indeed, it is highly hydrophobic with a PSA of 33.12Å^2 and a logP of 5.39. Indeed it was one of the four active compounds from the *in vitro* testing.

Compounds with favorable every do not necessarily score high on the SEI/BEI plane. Indeed these metric counterbalance potency with molecular property.

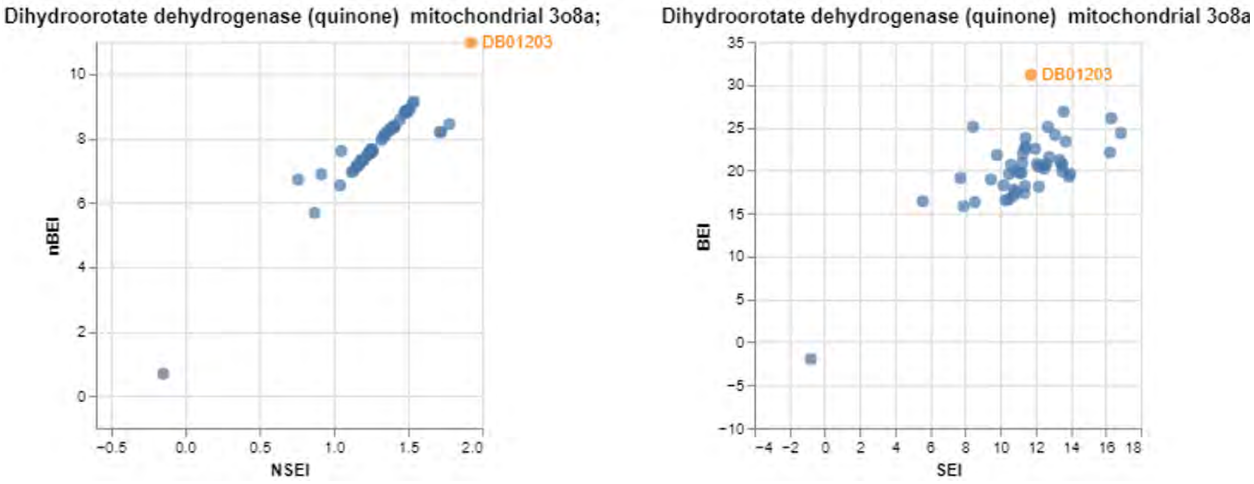


Figure 2-8 Scatter plot of LEIs of Nalodol and available Dihydroorotate dehydrogenase inhibitors in PubChem. **Left subplot**: x-axis NSEI y-axis nBEI, **Left subplot**: x-axis SEI y-axis BEI. Nalodol is in orange dot while other Dihydroorotate dehydrogenase inhibitors are in blue dots.

The plots illustrates nalodol optimized LEIs by its localization in North-east region of the plot. Compared to known Dihydroorotate dehydrogenase, it has better LEIs values. Interestingly, the compounds was not found active in the assays. Yet Dihydroorotate dehydrogenase is a validated target.

Despite having comparable molecular interactions, hits and co-crystallized ligands had different chemical scaffolds. Indeed, identified ligands had a Grscore above 0.58 indicating similar interactions to co-crystallized ligands. However, compared to their respective targets known inhibitors (found in ChEMBL[216]) none of the compounds showed a Tanimoto Coefficient (Tc) score greater than 0.6. Hence, they also present different scaffolds from known target inhibitors. This indicates GRIM ability for scaffold hopping despite searching for similar interaction patterns as previously indicated in the related publication[181]. Additionally, hits were compared to FDA-approved antimalarials. The maximum similarity (0.52) was between primaquine and terazosin. Even still, this low value does not imply structural similarity. This supports the likelihood of hits presenting a new mechanism of action from current antimalarials, an ideal scenario in the resistance context.

Furthermore, identified hits have appropriate pharmacological properties for further optimization. Indeed, these compounds have acceptable molecular weight and logP for further growing or modification. Hence, using efficiency metrics combined with GRIM, normalization, and complex ranking may have guided hit selection toward a more druglike chemical space and have avoided molecular weight-related bias. Additionally, some of the hits binding poses show the possibility for further extension within the binding site.

2.3.3 Twenty-five stable complexes in MD

2.3.3.1 MD: Ligand binding did not induce a conformational change

The 25 complexes were simulated for 20 ns to assess their stability and ligand binding affinities. MD is an effective approach to assess ligand binding mode conservation. For the protein structures, their stability in the *apo* and complex form was assessed using RMSD and Rg. The *apo* systems were compared to their respective complex forms.

For the *apo* proteins, the mean of RMSD per system ranged from 0.11 nm to 0.47 nm while ranging from 0.20 nm to 0.50 nm for the complexes (Figure 2-9). For the standard deviations, the ranges were [0.10 nm - 0.15 nm] and [0.01 nm - 0.06 nm] for *apo* and complexes proteins, respectively. Both maximum RMSD for *apo* and complexes were observed in 2YOH with 0.38 nm and 0.5 nm, respectively. These values are greater than 3 Å and are therefore above structural similarity thresholds. Nevertheless, they maintained low standard deviations 0.035 nm (apo) and 0.066 (complex) during simulation. This indicated that despite an initial deviation from the initial conformation, the new conformation is maintained throughout the simulation.

To assess ligand binding effects on protein structures, the complexes and their respective *apo* systems, the absolute difference of the protein RMSDs were analysed. The highest difference was 1.1 Å for 2YOH_DB01001. Considering protein structural similarity threshold of 3 Å [217], this difference is not significant for conclusions on structural change. Hence ligand binding did not induce any significant conformational change in any system.

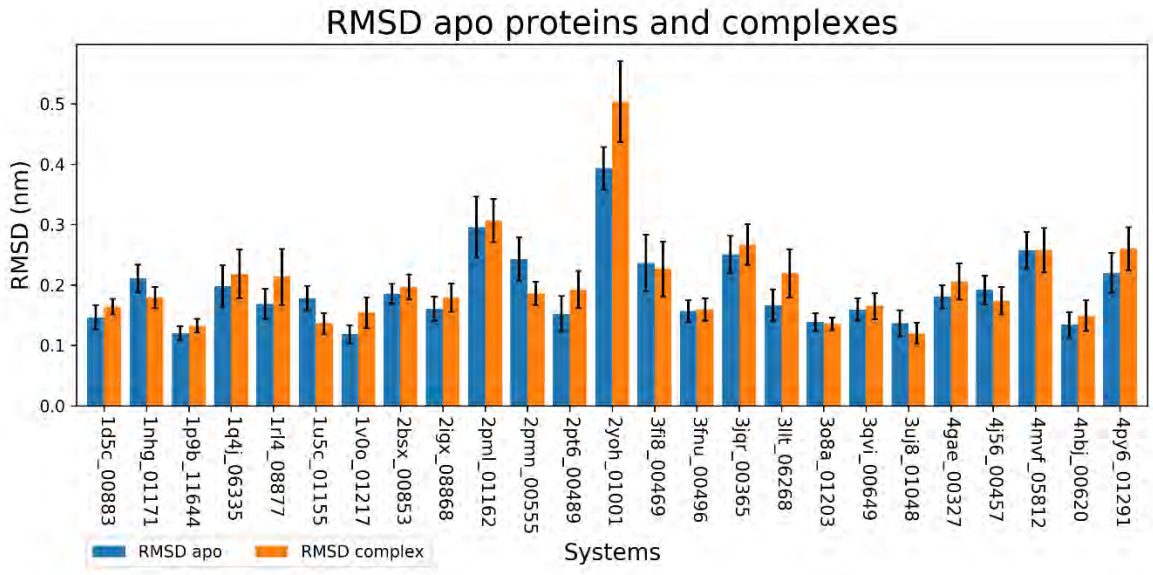


Figure 2-9**.** Mean of RMSD of backbone atoms for *apo* proteins and complexes. Complexes are represented by their DrugBank IDs (last five digits) and PDB IDs. Error bars are the standard deviations of the means. The complexes and *apo* proteins are in orange and blue respectively.

Ideally, this analysis may be done in a pair-wise comparison of all frames in the *apo* vs all frames in the complex system. This will allow comparing the entire ligand-bound conformational space throughout the simulation to the *apo* system. The abovementioned RMSDs are computed respective to the initial MD run frame. This may be different from the initial structure before the

minimization and equilibration steps of the system. Indeed, the output of these processes may result in different structures. Ideally, the reference system should be a set of structures, representing conformational space rather than an individual conformation.

Protein stability was also assessed through the Rg. There is an overlap between the two metrics (Rg, and RMSD) as all changes in Rg are normally captured by the RMSD.

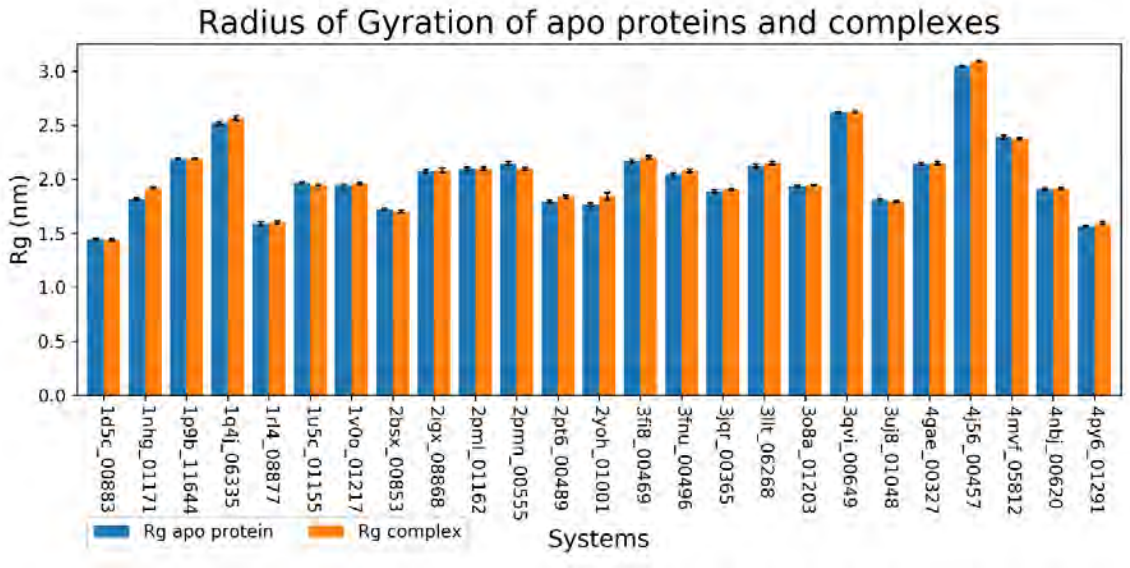


Figure 2-10**.** Means of proteins backbone atoms Rg for *apo* and complexes. Complexes are represented by their DrugBank IDs (last five digits) and PDB IDs. Error bars are the standard deviation of the means. The complexes and *apo* proteins are in orange and blue respectively.

The highest Rg standard deviations were 0.02 nm (4MVF) in *apo* proteins and 0.05 nm (2YOH_DB01001) for complexes. Hence, given these low values for both *apo* and complexes, these respective levels of compactness support their stability. Minimum and maximum Rg values were 1.45 nm (1D5C) and 3.04 nm (4J56) respectively for the *apo* proteins while they were 1.44 nm (1D5C_DB00883) and 3.09 nm (4J56_DB00457) for complexes. The inter-system difference is linked to the respective protein sizes. Indeed, protein structures can be grouped into different classes depending on their radius of gyration values [191].

Both proteins Rg and RMSD in their *apo* vs ligand-bound indicated that ligand binding does not induce significant conformational change. Ligand stability in the different systems was also assessed. Geometric metrics related to ligand stability were its RMSD and COM distance to the protein. Regarding COM distances, the standard deviation ranged from 1 Å (1RL4_DB08877) to 0.2 Å (3O8A_DB01203). These low standard deviations indicate no ligand dissociation. The actual COM distance ranged from 0.61 nm (2PT6_DB00489) to 2.33 nm (4J56_DB00457). Intersystem COM distances variation is more likely linked to binding sites proximity to protein COM, hence its absolute value is less likely to be linked to ligand stability, but its variation may be so. It is also noteworthy that change in protein COM will affect COM distance while ligand stability is not necessarily affected. Hence, this metric is to be used cautiously. Changes in protein COM may be tracked with its RMSD.

The RMSD for the ligand alone was low, below 2.5 Å (the highest observed value) in all systems, indicating stability. However, the ligand RMSD fitted to the protein backbone was characterized by a much greater fluctuation, reaching up to 8 Å with high standard deviations; as previously shown RMSD is more sensitive to ligand movements relative to the protein[192]. However, the consistency in hydrogen bonding and protein-ligand binding energy supports the stability of the ligands.

Molecular interactions play an important role in ligand stability in proteins, and hydrogen bonds are the strongest [218]. Figure 2-11 shows the hydrogen bond count time evolution in each complex. The maximum hydrogen bond count was observed with abacavir (DB01048). Its initial docked pose formed five hydrogen bonds, which may explain its high interaction energy, the 2nd most favourable (-261.05 kcal/mol) (Figure 2-12). Darifenacin (DB00496) and fingolimod (DB08868) had the lowest mean of hydrogen bond with 0.22 and 0.39 respectively. They mainly have hydrophobic contacts with their proteins. For instance, fingolimod formed only one hydrogen bond with ASP121 in its docked pose. Overall, ligands maintained the initial number of hydrogen bonds in the majority of the complexes (Figure 2-11).

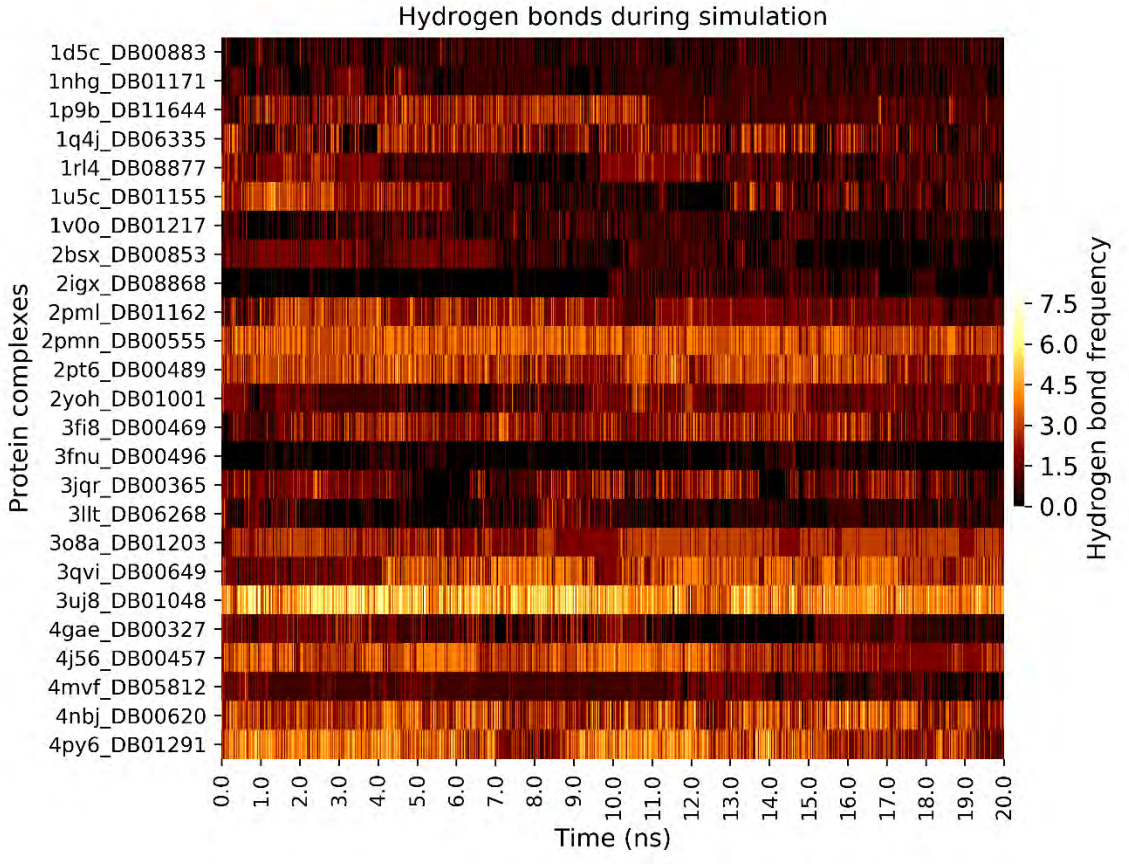


Figure 2-11**.** Time evolution of hydrogen bonds between the protein and the ligand. The y-axis represents the PDB ID and DrugBank IDs. The heatmap was produced with Seaborn version 0.9[198].

In terms of system interaction energies, prazosin had the best interaction energy with -331.42 kcal/mol as a mean value. It is difficult to compare the different energy values between systems

due the difference in binding sites. Pocket hydrophobicities may cause significant differences in binding energies. Ideally, interaction energies may be compared to the co-crystallized ligand as a reference. The ratio of two affinities can give a good estimate of the sample ligand affinity. Overall, all complexes had a negative energy value ranging from -140.28 kcal/mol to -331.42 kcal/mol indicating a favorable protein-ligand interaction (Table 2-1 and Figure 2-12).

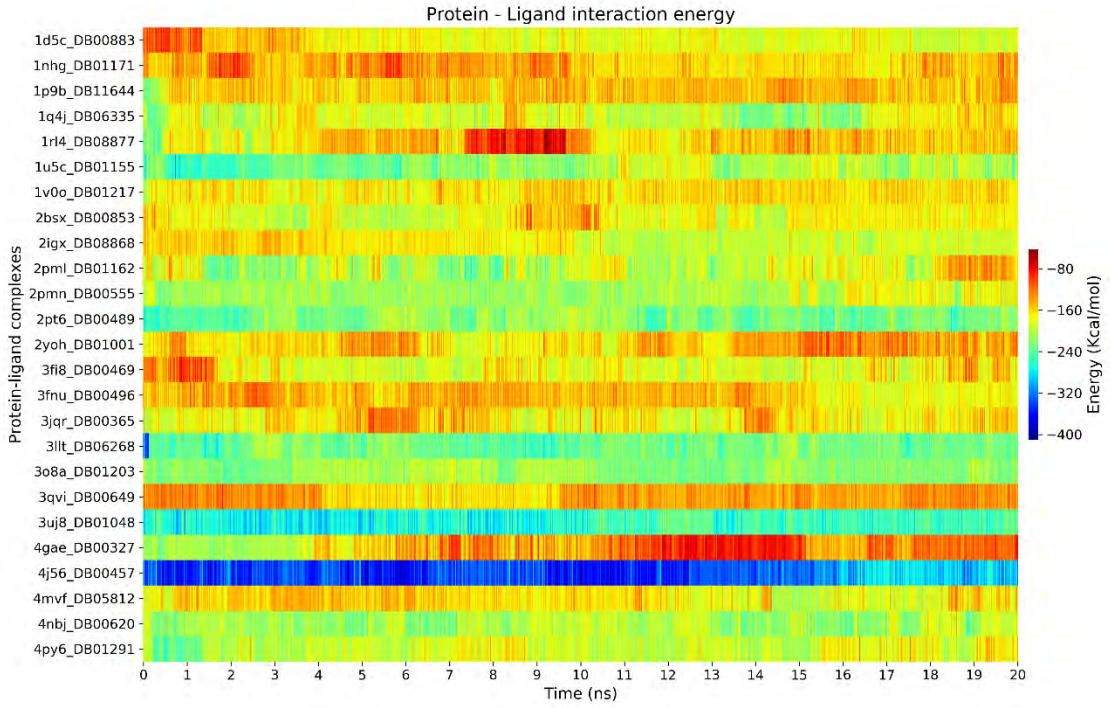


Figure 2-12 Time evolution of protein-ligand interaction energies. The heatmap was generated using Seaborn version 0.9 [198].

Compared to other metrics, the interaction energy, beyond assessing ligand stability and dissociation through its variation, also gives a measure of its affinity for the protein.

In summary, MD showed stable targets in both their *apo* and complex forms indicated by their Rg and RMSD. The different complexes were not different in stability from their *apo* forms. Ligand RMSD and COM distances to their respective proteins indicated their stability, while their hydrogen bonds and interaction energies indicated favorable protein-ligand interactions.

2.3.4 *In vitro* assays: four active compounds

Antiplasmodial and human cytotoxicity assays were performed for sixteen commercially available compounds. Four compounds were active against *P. falciparum* 3D7 (S37). Fingolimod and abiraterone had IC_{50} values of 2.21 µM and 3.37 µM respectively (Figure 2-13a), as the two most active compounds. Given that their activity values are in the single-digit µM range, these compounds are promising for further optimization. *P. falciparum* Plasmepsin 2, and *P. falciparum* Calcium-dependent protein were the two predicted targets for fingolimod and abiraterone, respectively. In the human cytotoxicity assays, fingolimod showed toxicity despite being an approved drug. It reduced HeLa cells viability to below 50% (1.98%) and later had an IC_{50} of 1.63

µM**.** Its immunosuppressant properties may explain its HeLa cells cytotoxicity. Indeed, the drug targets the sphingosine-1-phosphate receptor on T cell membranes[219]. Interestingly, it is currently studied as a treatment for COVID-19 [220]. In all active compounds, only fingolimod showed significant toxicity against human cells (Figure 2-13b).

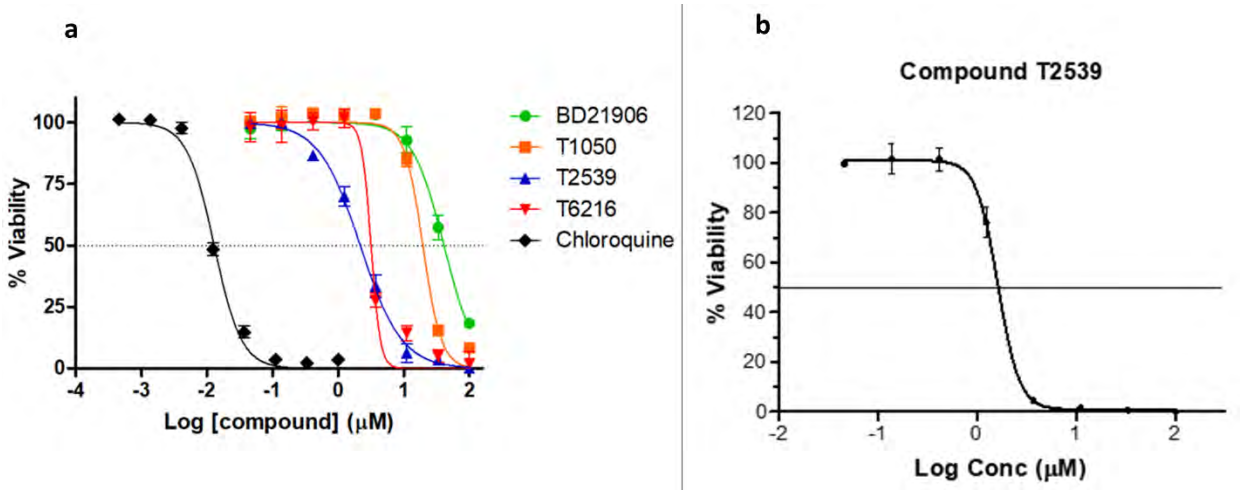


Figure 2-13 **a.** Antiplasmodial dose-response plots. *P. falciparum* viability percentage is plotted against the Log (compound concentration). Chloroquine, the positive control is the black curve. **b**. Dose-response plots for human cells. The viability percentage is plotted against the Log (compound concentration). In both plots, IC_{50} values were obtained by non-linear regression. The error bars are the standard deviation from the triplicate test. BD21906, T1050, T2539, T6216 correspond to terazosin (DB01162), prazosin (DB00457), fingolimod (DB08868) and abiraterone (DB05812) respectively.

Prazosin and terazosin also had active IC_{50} values of 16.67 µM and 34.72 µM respectively. A similar activity difference in cell viability assay was also observed. Interestingly, despite a two-fold difference in their activities, they are analogs (Tc of 0.7). Indeed, both compounds share a n-arylpiperazine scaffold with a piperazine ring having an aryl group substituent on the nitrogen ring atom. Their structures only differed by a tetrahydrofuran ring on terazosin, while prazosin has a furan one (Figure 2-14b and c). Moreover, the two compounds were predicted on two different targets: thioredoxin reductase 2 for prazosin and PfPK7 for terazosin. Hence, they may have a dual-action. Given their structural analogy, one would expect the two to bind the same targets. This difference may be linked to the hit selection procedure which enforced selecting a single compound for each target. Thioredoxin reductase 2 is putatively a good target for liver-stage active compounds[221]. Prazosin also had the most favorable protein-ligand interaction energy.

These active compounds could be tested in combination with chloroquine or artemisinin to identify potential synergistic activities. The combinatorial approach is currently the main one in malaria chemotherapy. The compounds could also be tested against parasite-resistant strains.

In a cell viability assay used as pre-screen, three other compounds (lamotrigine, salbutamol, and moclobemide) decreased cell viability to 72.23%, 71.83%, 61.24% respectively. These activities

were not considered enough for IC_{50} evaluation. Previously, salbutamol[222] and Moclobemide[223,224] and were shown to have activity against *P. falciparum* [222–224].

2.3.4.1 Binding modes of the active compounds

Here we describe the binding modes of the active compounds in their predicted selected target. Some may bind to other targets, as discussed, but this will not be explored here.

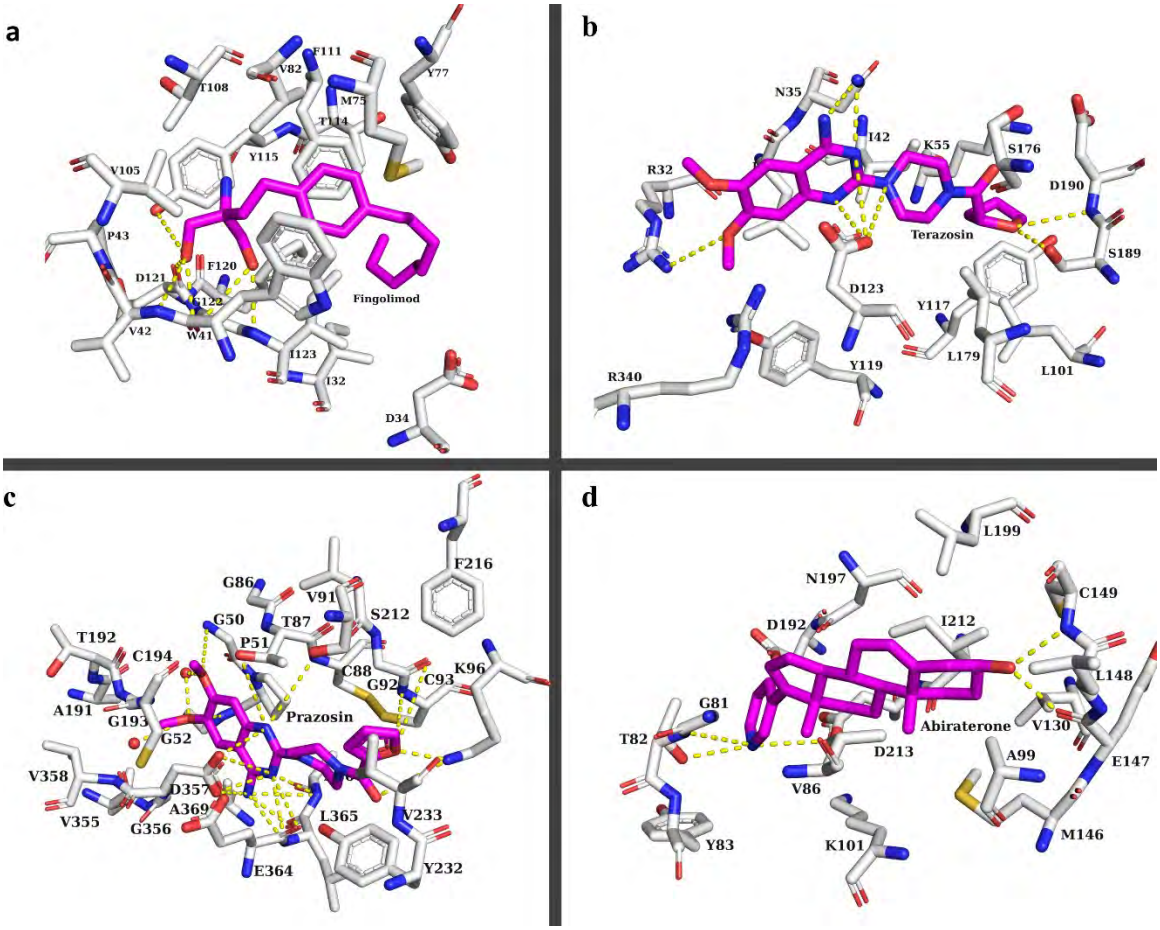


Figure 2-14 Active compounds binding modes in their predicted targets. **a.** fingolimod, **b.** terazosin, **c.** prazosin, **d.** Abiraterone. Active compounds are in magenta and residues in a radius of 3.5 Å are in white. Residues are labeled with their one-letter code and residue numbers. Dashed yellow lines are polar contacts. The figure was prepared using Pymol [225] and the show_contacts script [226].

2.3.4.1.1 Fingolimod (DB08868) and plasmepsin 2 (PDB ID: 2IGX):

Plasmepsin 2 is an aspartic proteinase A1A associated with hemoglobin degradation. Halofantrine, a known antimalarial targets this proteinase [227]. The plasmepsin 2 structure used 2IGX has been co-crystallized with A1T. The drug fingolimod is used to treat relapsing-remitting multiple sclerosis by regulating the sphingosine 1-phosphate receptor. It binds 2IGX in a buried hydrophobic pocket, forming a hydrogen bond with ASP121 in the pocket depth (Figure 2-14a), a crucial binding pattern for potent inhibitors [228]. Yet, the polar group in the hydrophobic pocket may come with an unfavorable energetic cost[229]. It forms Pi-Pi interactions with TYR77 and PHE111 and makes hydrophobic contacts with VAL82, PHE111, TYR77, and ILE123 via its aromatic

ring. Finally, the carbon chain has alkyl interactions with PHE111 and PHE120. The co-crystallized inhibitor A1T and fingolimod only have a Tc of 0.22. They share only a common long aliphatic chain as is often observed in plasmepsin 2 inhibitors. Additionally, fingolimod expands into the trench area outside the pocket. This region may provide an excellent opportunity to extend its scaffold for increased potency. Optimization studies have been carried out on fingolimod for better selectivity for sphingosine-1-phosphate receptor 1 (S1P1) and against S1P3, responsible of bradycardia. This optimization resulted in a compound with over 1250-fold selectivity for S1P1 compared to S1P3 and no bradycardia [230]. The same compound may be worth investigating as antimalarial, but also to overcome the observed fingolimod human cytotoxicity.

2.3.4.1.2 Abiraterone (DB05812) and calcium-dependent protein kinase 2 (PfCDPK2, PDB ID: 4MVF)

PfCDPK2 has no human homolog [231]. Abiraterone binds to a trench-like hydrophobic pocket making contacts with VAL86, VAL130, MET146, LYS101, ALA99, LEU199, ILE212, ASP213, CYS149, and VAL86 (Figure 2-14d). The compound had the highest SEI (33) which may be linked to its low PSA (33.12 Å2) and high logP (5.39). Its only polar contact is with THR82 in a more exposed area of the binding site. Its most similar compound among calcium-dependent protein kinase 2 inhibitors was CHEMBL602580 with a similarity of 0.5. The 4MVF binding site has a Shaper score above 0.44 with respect to 3gie and 3LLT, indicating the compound is a potential binder to these targets too.

2.3.4.1.3 Terazosin (DB01162) and protein kinase 7 (PfPK7) (PDB ID: 2PML)

PfPK7 has features different from its mammalian homologs[232] making selectivity possible [232,233]. Terazosin interacts mainly through hydrophobic contacts (TYR117, SER189, LYS55, LEU34, LEU179, ASN35, LEU101, ASP123, ASP190, and ILE42) (Figure 2-14b) while forming a hydrogen bond with ARG32. This pattern resembles that of known inhibitors [232]. A difference with terazosin is that it extends to a superficial area of the pocket, also having a moiety fitting more deeply in the binding site[232], while the ATP analog in 2PML binds more superficially. The terazosin and the ATP analog structures are also quite different with a Tc of only 0.23. Compared to 2PML's (Target ID: CHEMBL6169) known inhibitors, terazosin had the highest similarity (Tc=0.5) with CHEMBL602580 sharing a common long chain connected to a benzene ring.

2.3.4.1.4 Prazosin (DB00457) and thioredoxin reductase (PfTrxR, PDB ID: 4J56)

PfTrxR is essential for *Plasmodium falciparum* [234] and is a putative liver-stage target[221]. Prazosin binds to a buried pocket, the FAD binding site. Here it interacts with CYS88 and CYS93 which form the protein redox centers [234] and forms hydrogen bonds with ASP357, and LYS96. Additionally, it makes contacts with VAL233, THR87, SER212, PRO51, GLY52 and ALA191, and some water molecules (Figure 2-14c). An aromatic quinazoline similar to the quinoxaline found in the PfTrxR (CHEMBL4547) inhibitor (CHEMBL380953) is also present on prazosin. The two rings present many hydrophobic interactions. Prazosin is selective for the parasite and inactive on mammalian thioredoxin reductase (PubChem [235] BioAssay IDs 588453, 488773, and 488772). Additionally, its selective profile is confirmed by the human cytotoxicity assay presented here. 4J56 is an isolated target: not having a similar binding site above 0.44 (Shaper score) to any of the current targets. Moreover, prazosin (CHEMBL2) is not predicted for PfTrxR in ChEMBL target predictions [236].

Hence the current predicted target requires further validation, and this should be treated with care.

2.3.4.2 A 25% hit rate

The comparison between the *in silico* predictions and the observed experimental values from the assays indicated the hit rate to be 25%. Four of the 16 tested hits were active. Virtual screening pipelines are reported to have hit rates in the range of between 1% and 40% [237] with a 5% hit rate often considered as successful [142]. A larger screening library may help improve the current hit rate as this strategy is shown to be effective [100]. However, this may oppose the repurposing strategy as the set of approved drugs does not make a large library. The effectiveness of MD for screening may require improvement. Eight of the 16 tested compounds did not have any activity in the cell viability assay. These compounds were stable in MD on validated targets. Advanced binding free energy methods such as umbrella sampling [192] and using the co-crystalized ligand as a reference may be helpful in future studies. In addition to the all-vs-all assessment, the pipeline can be evaluated using a set of actives vs decoys molecule libraries. The final ligand ranking would then be assessed through the enrichment factor and/or the area under the curve in the receiver operating curve as commonly used for screening pipelines [211].

2.3.4.3 Promiscuous active compounds

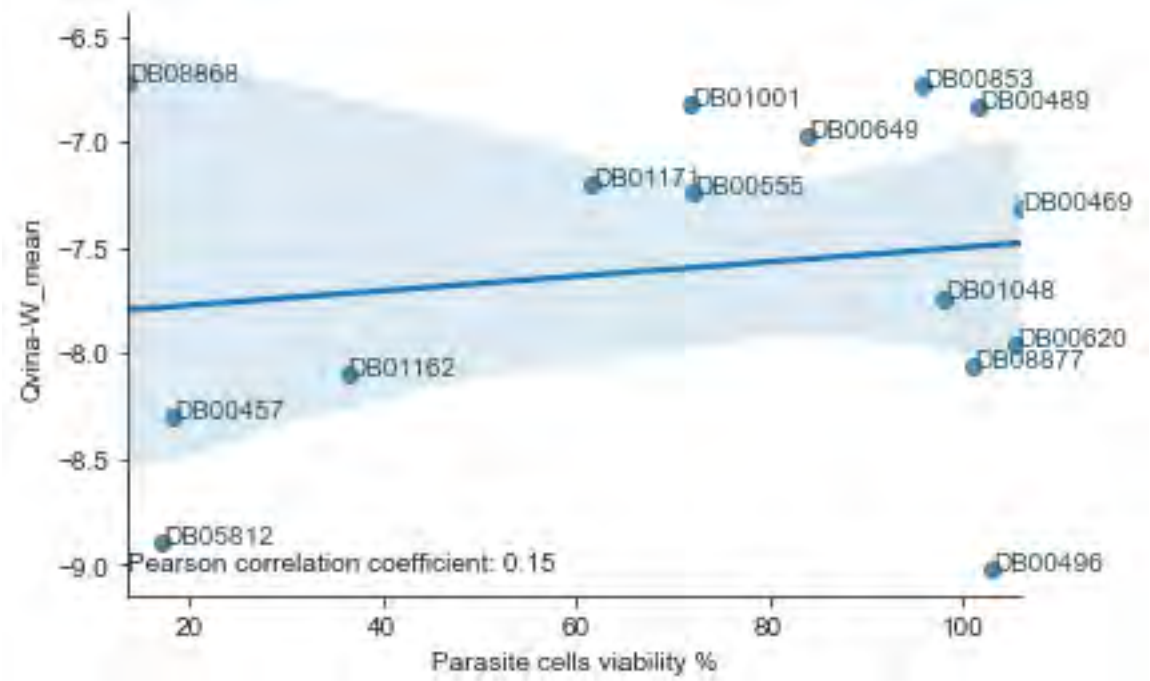


Figure 2-15 Parasite cell Viability % vs the average of Qvina-W binding energies scatter plot. Compounds are labeled with their DrugBank ids. The blue line and light area are the regression line and confidence interval at 95 % respectively.

Figure 2-15 shows the scatterplot of parasite cell viability % and their corresponding Qvina-W binding energies averaged across the different targets. The predictions are poorly correlated to the experimental values (Pearson correlation coefficient: 0.15, and Kendall tau of -0.03). Two outlier compounds and seem to be the reason DB08868 and DB00496. Discarding the outliers

from the dataset improved the correlation to 0.25 (Kendall) and 0.58 (Pearson). Fingolimod (DB08868) is active but predicted poorly, while darifenacin (DB00496) has good predictions but was not active. The darifenacin inactivity remains unclear. The fingolimod case might be explained in terms of its specificity to a crucial target in the parasite. However, it also had a highly toxic effect on human HeLa cell cytotoxicity assays. Based solely on binding affinity, fingolimod would not have been selected as a hit. Yet it is active. The efficiency metrics-based and complex ranking may have contributed to its selection.

Interestingly, applying a -8 kcal/mol threshold filter on the average Qvina-W binding energies in the pipeline results in a ~50% hit rate. Only six of the tested compounds pass that threshold and three of them are active: Abiraterone (DB05812), terazosin (DB01162), and prazosin (DB00457). In this case, active compounds show a promiscuous nature, having good binding energies across multiple targets. This unexpected finding contradicts the selectivity-based approach implemented through the complex ranking used here. The paired ranking was used to select selective compounds and avoid promiscuous ones. Selecting compounds simply based on their average binding energy returns a high number of compounds. It is the combination of applying a -8 kcal/mol threshold on the average binding energy of hits selected using the paired ranking scheme which significantly improved the pipeline hit rate. Nevertheless, these compounds' promiscuity profile can reduce drug resistance probability, which is ideal in the context of malaria elimination [40].

One distinction of the current method from most virtual screening pipelines is the integration of proteome-scale, consensus scoring (GRIM and QVina-W), repurposing and paired ranking strategies. Ligand and target choice strategies were to meet the different TCP/TPP profiles for malaria elimination [22,37]. The hit selection strategy enforced a ligand for each target. This is in line with the aim to identify multiple targets for target diversity (Figure 2-4b) and new mechanisms of action. Moreover, this comprehensive approach still uses cost-effective strategies such as the HMR scheme in MD together with on-the-fly rescoring with GRIM.

Drug target array screening together with the complex ranking model the behavior of drug-target interactions *in vivo*. In single target screening, not all library compounds will exclusively bind to the target of interest. However, interactions with other targets occur. Screening on target arrays model those interactions. In addition, the complex ranking [142] helps select a specific target for a drug. Yet, this process is limited as the drug now competes with other drugs in the library for that target. Hence a more selective drug may be chosen for that target. Moreover, this approach allows modelling drug combination activity. Drug combination has been a key strategy in the fight against malaria. Further, one could integrate a “target essentiality score”. Despite being validated, targets may have varied essentiality in similar parasites. The PlasmoGEM project evaluated the relative growth rates of the parasite *Plasmodium berghei* for more than 2,500 genes [49]. A similar experiment was later done on *P. falciparum* [53]. Each target could be mapped to its essentiality score. Hence each multitarget compound could be scored by a multitarget index: summing the Grscore or binding affinity for each target weighted by its gene essentiality score. This approach could well help in prioritizing high-value drug targets. Similarly, Loza-Mejia *et al.* used a weighted multitarget index in which targets were weighted with a desirability

coefficient of binding [238]. This desirability coefficient can also be extended to include human key proteins for improved selectivity [239].

2.4 Conclusion

This work identified antiplasmodial property in four orally available and known drugs (fingolimod, abiraterone, prazosin, and terazosin). To the best of our knowledge, none of the compounds had previously known antiplasmodial activity. Abiraterone and fingolimod had IC_{50} values in the single-digit range. Abiraterone is not only an orally approved drug, safe on human cells as shown in assay results but also predicted on a putative liver-stage essential target. It hence fulfills many of the requirements for a new antimalarial [50].

The pipeline incorporates multiple metrics: molecular properties (ligand efficiency), energy (QVina-W), and molecular interactions (GRIM) scorings. It further uses normalization and ranking strategy for selectivity ideal in protein array screening and improving scoring bias [154]. It shows a promising 25% hit rate given the proteome-scale screening and cost-effective approaches context. Further analysis shows that this hit rate can be significantly increased by combining compound complex ranking with their average binding energy across the targets.

In the malaria drug resistance context, its elimination will benefit from cost and time-effective approaches for chemotherapy. This repurposing pipeline contributes to identifying antiplasmodial drugs from sets of known drugs for future accelerated development. Given the parasite complex biology, disease elimination will benefit from holistic approaches toward system biology. This workflow sets the stage for a multi-objective, proteome-scale virtual screening pipeline.

In the future, the pipeline can be extended with the set of docking SFs used in Chapter 3: . Both target and ligand sets can be extended. Particularly, target human analogs, other human proteins sensitive to toxicity can serve to improve selectivity and avoid toxicity as shown with fingolimod. Ligands can be extended to the set of all experimental drugs in line with repurposing or a special *P. falciparum* custom-made library. Not only co-crystallized but active and inactive compounds in *P. falciparum* could be used as a baseline for a first thorough evaluation of the pipeline. The approach here may also be applied in other disease areas such as tuberculosis.

Chapter 3: Consensus Ligand and Structure-Based Screening for Identification of PfDXR Inhibitors

3.1 Introduction

The methyl-d-erythritol 4-phosphate (MEP) pathway is present in all intra-erythrocytic stages of the parasite and is essential for hepatic stage optimal development [240]. Its intermediates are present in gametocytes and its products are required for gametocytogenesis, showing thus potentiality for transmission-blocking compounds [241]. *Plasmodium falciparum* 1-deoxy-d-xylulose 5-phosphate reductoisomerase (PfDXR) catalyzes MEP second step, a rate-limiting one. It converts 1-Deoxy-D-xylulose 5-phosphate to 2-C-methyl- D-erythritol-4-phosphate (MEP) by isomerization and reduction [242]. It has no human homologs and hence ideal as a target for drugs matching the new antimalarials criteria [243]. The enzyme is a homodimer using NADPH as a cofactor and a metal ion (Mg^{2+}, Co^{2+} or Mn^{2+}) both required for the enzyme catalytic activity [244].

Fosmidomycin attracted much attention as a PfDXR inhibitor [245]. Yet, its pharmacokinetic properties need improvement for it to be effective as a drug [246]. Hence, studies on finding potent and drug-like PfDXR inhibitors are still needed [243].

Virtual screening approaches, both ligand-based and structure-based, can be used in the early stages of drug discovery. While the former is in general, faster, the latter uses target information giving insight into drugs' binding modes. Their combination can allow for fast screening of large libraries using ligand-based virtual screening (LBVS) followed by SBVS for a more thorough screening. The D3R grand challenge results showed that current scoring functions do not perform significantly better than logP or Molecular Weight (MW) to rank compound affinities [164]. Consensus approaches in virtual screening have proven to be more effective than a single approach [210,211,247,248]. Advanced and more accurate methods require high computational resources and are difficult to set up for high-throughput experiments. Simple rescoring of protein-ligand complexes requires much less in terms of computational resources [164].

Previous virtual screening studies contributed to identifying PfDXR hits using LBVS and SBVS [249–253]. In an earlier similar virtual screening study on DXR, Wadood *et al.* identified inhibitors from ChemBridge [254] using a pharmacophore model [250]. Recently, natural product hits were identified using a joint pharmacophore and MD approach [255]. Open Eye software FRED was used to finding hits from the ZINC dataset, which were later docked. Potential identified hits had better-predicted potency than fosmidomycin [251]. ZINC12 was screened in ArgusLab using a shape-based method to search for hits with comparable functional groups to fosmidomycin.

In this chapter, a hierarchical virtual screening pipeline combining LBVS and SBVS was used to find hits for PfDXR. Compounds are further assessed using MD, steered MD, and free energy calculation through MM-PBSA, and US LC5, a 280 nM [256] PfDXR inhibitor was used as a baseline for hit identification.

In the following section, a brief overview of the different LBVS tools and docking scoring functions (SFs) is presented.

3.1.1 Ligand-based virtual screening

LBVS is based on the principle that in general similar ligands have similar properties [78,257]. In some rare cases, they may also have very different properties (e.g. activity cliffs) [258]. Methods related to ligand properties such as shape, topology, physicochemical properties, etc. have been used to measure ligand similarity. These encode chemical compounds into a set of numerical values, a critical step for accurate molecular representation in numeric form. Methods such as distance and similarity coefficients then compare compounds [257,259] through their numerical encoding.

That compound similarity is related to their relative biological activity is relevant in drug discovery. The compounds' absorption, distribution, metabolism, and excretion (ADME) properties also remain important. Molecular properties have been linked to their structures [260] leading to SAR and quantitative structure-activity relationship (QSAR) studies. The latter assumes compounds' properties are encoded in their structures, resulting in similar structures having similar properties. It thus remains to find the structural encoding that best explains activity, given that any of them apart from the electron density result in loss of information [257]. For example, a specific conformer SMILES encoding loses the structure conformation information and subsequently the electron density. Given a compound of interest's structure and a library of molecules with known structures, LBVS aims to find compounds with similar properties using structural searches. There are 2D and 3D similarity methods depending on compounds' structural representation. 3D approaches can also be divided into shape-based, pharmacophore modeling, molecular field, and 3D fingerprint approaches [261].

In this study, six LBVS methods especially shape-based approaches are used to screen 3 M Zinc lead-like compounds on DXR.

3.1.1.1 USR

Ultrafast shape recognition (USR) is a method for molecular shape comparison based on the relative spatial positions of atoms [262]. This cuts the need for alignment or translation of molecules. However, the full set of compound interatomic distances have redundant information, more than is required for its shape description. As result, only atomic distances from four molecular locations are considered: the molecular centroid (ctd), the closest atom to ctd (cst), the farthest atom to ctd (fct), and the farthest atom to fct (ftf). Given a compound 3D structure, the distribution of distances from every atom to each centroid is generated. Finally, the descriptor vector consists of the normalized first three moments of each distribution, giving a vector of twelve numbers encoding the shape information. The similarity of two compounds is the inverse of the translated and scaled Manhattan distance between their descriptor vectors. One and zero correspond to the maximum and minimum similarity respectively [262]. This approach is independent of the atom number in each compound and is suitable for finding new scaffolds.

It is also fast, 2038 times faster than that of Shape Signatures [263] and at least 5 059 times faster than ROCS (rapid overlay of chemical structure, OpenEye Scientific Software) [262]. The speed improvement is a result of the absence of molecular surface or volume calculations. Using only one processor, USR can screen a multibillion database to find similar compounds to a single query in a matter of minutes [262].

3.1.1.2 ElectroShape

ElectroShape (ES) adds a fourth one to the three dimensions (x, y, z) used in USR. This dimension considers atoms' charge thus together they unify both compound shape and electronic information. To find the exact position of a point in the n-dimensional space Rn, n+1 centroids are needed. Hence, five centroids are used to take into account ElectroShape four dimensions. ES maintains the speed of USR while almost doubling the average enrichment ratio at 1% in virtual screening. It also differentiates between enantiomers through the chiral shape recognition (CSR) method and shows good enrichment in terms of scaffold novelty. To ensure unit consistency between the fourth dimension (charge) and the x, y, z coordinates (in angstroms), a scaling factor (μ) which counts for the number of angstroms per electron is used. This makes the approach flexible, balancing between a pure shape-based (using a small μ) and partial charge based using a large μ. Optimal μ values are determined by choosing the value that gave the best average enrichment using the directory of the useful decoy (DUD) dataset. Each molecule is encoded in a fifteen-number vector of the first three moments of the distances to centroids distributions. ElectroShape similarity score is the inverse Manhattan distance between two compounds' vectors [264]. In an exemplary case, the rank receiver operating characteristic (ROC) curve showed better early enrichment than the similarity ROC curve. This latter gave the best area under the curve for a similarity threshold set at 0.8 between active and inactive compounds [264].

3.1.1.3 USRCAT

Ultrafast Shape Recognition with CREDO Atom Types (USRCAT) is a fast shape-based method, extending USR by adding pharmacophoric constraints with atom typing from the CREDO database [265]. Moments for specific subsets of a molecule's atoms (hydrophobic, aromatic, hydrogen bond donor or acceptor atoms) are added to USR moments. The three moments resulting from each distribution from the 4 centroids (see USR 3.1.1.1) for all atoms and the four subsets result in a USRCAT descriptor vector of 60 elements (5x12 (4 centroids * 3 moments for each)). The first 12 are identical to USR moments. For empty subsets, the corresponding elements in the vector are set to zero (if no hydrogen bond donor is found, for example). From two USRCAT vectors, the similarity score is as in USR the inverse of the translated and scaled Manhattan distance between their descriptors vectors [266].

Descriptors and similarity scores for USRCAT, ES, and USR were calculated from ODDT toolkit [267].

3.1.1.4 RDKit 3D pharmacophores

The RDKit[162] 3D pharmacophore is calculated based on pharmacophore feature points identified on a compound followed by inter-feature topological distance calculations. The pair feature-distance combinations are assigned bit ids which are stored as bits or counts. Features are customizable and RDKit[162] uses features defined by Gobbi *et al.* [268]. The 3D pharmacophore fingerprint was computed using RDKit[162] by feeding a 3D distance matrix to the 2D-

pharmacophore machinery as described in the documentation [269]. The similarity score is then calculated from the pharmacophore fingerprint using the Tanimoto similarity.

3.1.1.5 Obspectrophore

From compounds' 3D structures, obspectrophore computes the following atomic properties: partial charges, lipophilicity indices, shape deviations and softness properties which encode the molecular field generated by a structure's topology. Each molecule, or more specifically each conformation, is inserted into an artificial cage of points (an artificial receptor) followed by calculating the interaction energy between each of the atom and the surrounding points using equation (3-1).

$$V(c,p) = -100 \sum_{i} \sum_{j} \frac{A(j,p)P(c,i)}{r_{ij}} \qquad (3\text{-}1)$$

Given a structure with j atoms and p atomic properties (in the current implementation p = 4), and a cage *c* with *i* cage points and the cage values P(*c*, *i*), the total interaction value V(*c*, *p*) of property p and is calculated according to a standard interaction energy equation (3-1). r_{ij} is the Euclidean distance between atom *j* and cage point *i*. The structure is rotated along all its axes within the cage and the most favorable interaction values are kept as the final result. The approach is hence independent of the compound orientation. Default values are 12 different cages (c) and 4 different atomic properties (p) generating a final spectrophore of 48 values per molecule. The similarity score between two compounds is obtained with the Euclidian distance between their spectrophores with a threshold <= 50 used to infer two compounds' similarity [270]. Some parameters control the approach application: accuracy controls the angular step sizes for the compound rotation in the cage which gives faster but less accurate computation with larger step sizes. Resolution controls the distance between the molecule and the cage (default value = 3 angstroms). The 12 cages described above have a symmetrical distribution of points, making the approach insensitive to molecules' enantiomeric configuration. The stereospecificity parameter makes possible an asymmetric distribution of points with 18 asymmetric cages. Finally, a normalization parameter which is done on a per-property basis allows focusing on the relative differences in the spectrophore values rather than on the absolute numbers. Among the default parameters, only the normalization was changed from none to normalization with zero mean as recommended for virtual screening. A Euclidian distance of 50 or below can be used to identify similar compounds [270]. The accuracy value was 20°, the resolution 3 Å, and the stereospecificity sets to 'none'.

3.1.1.6 MHFP6

MHFP6 encodes structures using the extended-connectivity principle as in extended-connectivity fingerprint (ECFP) up to a diameter of six bonds and combines it with molecular shingling and MinHash encoding methods. From a starting structure, the shingling writes circular substructures around each atom as SMILES. These are then assigned to bit values using a local sensitive hashing (LSH) scheme, the MinHash hashing. The generated hash values sets can be indexed by an LSH algorithm for approximate nearest neighbor search (ANN). This approach solves the problem of dimensionality and allows for structures indexing in extremely sparse Tanimoto space. From two

structures shingling (A and B), the Jaccard similarity coefficient of the molecules is calculated according to equation (3-2).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3\text{-}2)$$

MHFP6 uses the same fingerprint as the ECFP6 but a different algorithm for hashing (minhash). This accelerates the approach and suits it to large databases. The python implementation of the method was used from https://github.com/reymond-group/mhfp. Molecules are sanitized to ensure that they are "reasonable" and kekulized (converting aromatic rings to their Kekulé form) using RDKit [271]. The default radius of three encodes SMILES to MHFP6. Structures' rings in the molecules are included in the fingerprint [83].

Except for the MHFP methods, all the approaches used here use require the 3D structures of molecules of interest.

3.1.2 Docking SFs

In SBVS, SFs predict protein-ligand binding affinity using mathematical functions. They are generally divided into 4 classes: force field-based, knowledge-based, empirical, and machine learning ones [272,273].

3.1.2.1 Rf-score

RF-Score is a machine learning SF using the random forest algorithm to predict affinity. In its fourth version, it uses 47 features, of which 36 are RF-Score features and 11 are AutoDock Vina features. RF-Score versions 1 to 4 were used [267]. RF-Score had a higher hit rate when compared to Vina with its top 1% providing a 55.6% hit rate, while that of Vina was of 16.2%. Recently, ML SFs showed to perform better than classical SF. Primary versions of RF-Score-VS v1-3 outperformed state-of-the-art classical SFs [274]. However, ML SFs are often qualified as black boxes without interpretability [275].

3.1.2.2 Cyscore

Cyscore is an empirical SF focusing on improving the prediction of hydrophobic free energy. Contrary to many SFs which treat the term as surface tension, proportional to the interfacial surface area, Cyscore uses a curvature-dependent surface-area model. This approach can distinguish convex, planar, and concave surface in hydrophobic free energy calculation while the former approach ignores the role of molecular shape. The curvature-dependent surface-area model was shown to be superior to the conventional surface area model [276]. This is particularly important for ligands binding in narrow pockets where the curvature factor has a higher contribution to the hydrophobic free energy. The hydrophobic effect was estimated to contribute to perhaps 75% of the free energy of most binding. Charge–charge interaction, water-mediated protein-ligand interaction, and the π–system interactions are not considered by Cyscore. Entropy is estimated by the number of rotatable bonds on the ligand [276]. Cyscore predicts affinity in a negative range of value, the lower the value, the better the affinity.

3.1.2.3 DSX (DrugScore eXtended)

DSX is an improved version of DrugScore, extended with a more specialized set of atom types. It is a knowledge-based SF that includes different statistical potentials: distance-dependent pair, torsion angle, and solvent accessible surface-dependent one. In an assessment of its scoring power, DSX ranked second after Xscore. It is recommended to adjust the weight for the different terms in DSX functional to produce a target-tailored SF even though the authors did not do so in the related paper [277].

3.1.2.4 Xscore

Xscore average binding affinities predictions from three empirical SFs: HPScore, HMScore, and HSScore. All three SFs were calibrated through a multivariate regression analysis of a set of 200 protein-ligand complexes. They reproduced the binding free energies of the entire training set with a standard deviation of 2.2 kcal/mol (HPScore), 2.1 kcal/mol (HMScore), and 2.0 kcal/mol (HSScore). They differ in their modeling of the hydrophobic effect. This latter is estimated by the buried solvent-accessible molecular surface, or by the hydrophobic matching of the ligand with the binding site, or by the number of hydrophobic contacts between the protein and the ligand. Xscore accounts for metal interactions of interaction in the hydrogen bonding term, given that it has a Lewis acid-base pairing nature. The functional form of the free energy in Xscore is given by equation (3-3).

$$\Delta G_{\text{bind}} = \Delta G_{\text{vdw}} + \Delta G_{\text{H-bond}} + \Delta G_{\text{deformation}} + \Delta G_{\text{hydrophobic}} + \Delta G_0 \quad (3\text{-}3)$$

ΔG_{vdw}: protein-ligand van der Waals interactions.
$\Delta G_{\text{H- bond}}$: hydrogen bonding term.
$\Delta G_{deformation}$: deformation effect (only accounting for the ligand, the protein one being neglected).
$\Delta G_{hydrophobic}$: hydrophobic effect.
ΔG_0: regression constant.

Metals can contribute to binding affinity by forming a bond with lone pairs in the ligand. Xscore accounts for such type of interaction in the hydrogen bonding term, given that it has a Lewis acid-base pairing nature. This is of importance since in this study the Mn atom in the binding site and the surrounding water molecules were kept in the protein.

3.1.2.5 NNScore

NNScore 2.0 is a machine learning SF [278] trained on a set of receptor-ligand complexes with features derived from BINANA descriptors [279]. Structures were extracted from MOAD [280] and PDBbind-CN [281] databases. The SF was compared to its first version, AutoDock and AutoDock Vina ones. Interestingly, the first version NNScore 1.0 outperformed all of them and while NNScore 2.0 ranked second.

3.1.2.6 AutoDock

AutoDock is a popular semi-empirical free energy force field trained using a set of 30 structurally known protein-ligand complexes. The docking approach uses the Lamarckian Genetic Algorithm combined with semiempirical force field SF for free energy estimation. Calibrating the method

on a diverse set of 188 protein-ligand complexes, the SF produced a standard error of about 2-3 kcal/mol [161]. In this study, the AD4 SF implemented in Smina which uses the same terms as AutoDock4 is referred to as AutoDock.

3.1.2.7 AutoDock Vina

AutoDock Vina speeds up AutoDock 4 by two orders of magnitude using multithreading on multi-core machines in terms of molecular docking. It is more accurate in binding mode prediction with a standard error of 2.85 kcal/mol in energy prediction. Vina was inspired by Xscore functional form and is trained on the PDBbind dataset [282]. Beyond the intermolecular contributions to the free energy, Vina also accounts for intra-molecular contributions. It, however, does not account for hydrogen atoms explicitly. There is no directionality in the hydrogen bonding term [165]. Vina continuously was found among the top-scoring in different SF assessment studies [81,197,283]. Smina and Vinardo SFs were inspired by Vina [199,284].

3.1.2.8 Vinardo

Vinardo (Vina RaDii Optimized) is derived from Vina with fewer parameters and with a more physics-based character than the ML approach used in Vina. The terms in the SF can be related to some terms used in current SFs. In Vinardo's development, a set of 72 functions were derived from Vina by changing parameters weights, terms, and atom radii and trained with a reduced set of PDBBIND 2013. Vinardo presents two differences compared to Vina: the absence of the second Gaussian term in Vinardo which produces a second minimum in the steric interactions and the change in atomic radii. Pre-minimized complexes were used for the fitting of the SF. The authors indicate that the real Scoring power of a function is better measured by predicting binding energy on energy minimized structures. Testing its ranking power and virtual screening capabilities, Vinardo was found to be more successful compared to Vina [284].

3.1.2.9 Smina

Smina is an empirical SF derived from AutoDock Vina. It extends the Vina default functional terms with simple property counts, an electrostatic , an AutoDock 4 desolvation term, (45) a non-hydrophobic contact term, and a Lennard–Jones 4–8 van der Waals term. Hence, it aims to identify the most useful linear combination of these terms. The function was fitted using the Community Structure-Activity Resource (CSAR) 2010 data set [199].

3.1.2.10 Protein-ligand extended connectivity (PLEC)

PLEC presents the particularity to be developed from fingerprint. Protein-ligand interactions are encoded using the ECFP and used as features to train different machine learning models. A simple linear model could achieve a predictive power of 0.817 on the Protein Databank (PDB) bind v2016 'core set'. The affinity value is predicted in p*Ki*/d unit [285].

Overall, different SFs from four main classes: force field, machine learning, knowledge-based and empirical ones were used in this study.

3.2 Methods

3.2.1 Data Retrieval

3.2.1.1 ZINC lead-like set

About 3.8 M compounds from the "lead-like" subset of ZINC15 were used. The subset was filtered for compounds with MW in the range <400 Da, fitting in the PfDXR active site. Only anodyne (no pan assay interference compounds (PAINS), no michael acceptor) compounds and commercially available were selected. A final subset of ~3 M compounds was used. Their structures were retrieved from ZINC in MOL2 and pdbqt formats.

3.2.1.2 DXR inhibitors

DXR possesses a good wealth of inhibitors crystallized in their bound conformations. Ligands tend to bind to proteins in their lowest energy conformation [286]. A shape-based approach may present the advantage of finding hits with shape already matching the bound conformation of PfDXR inhibitors. Seventeen protonated PfDXR inhibitors in their bound conformation were retrieved from the PDB [287]. They have a MW between 250 and 400 and logP between 0.5 and 3. Their structures (Figure 3-1) present a phosphonate and hydroxamate (for metal chelation) groups, and a backbone spacer ideally of 3 carbons in length. Although essential for activity, these groups present poor pharmacokinetic properties [246]. Extensive research has been done to develop fosmidomycin/FR900098 analogs [49,51,288]. They show some key features on PfDXR inhibition structure-activity relationships [56]. For example, introducing an aromatic ring on the backbone chain resulted in more potent PfDXR inhibitors[289].

.

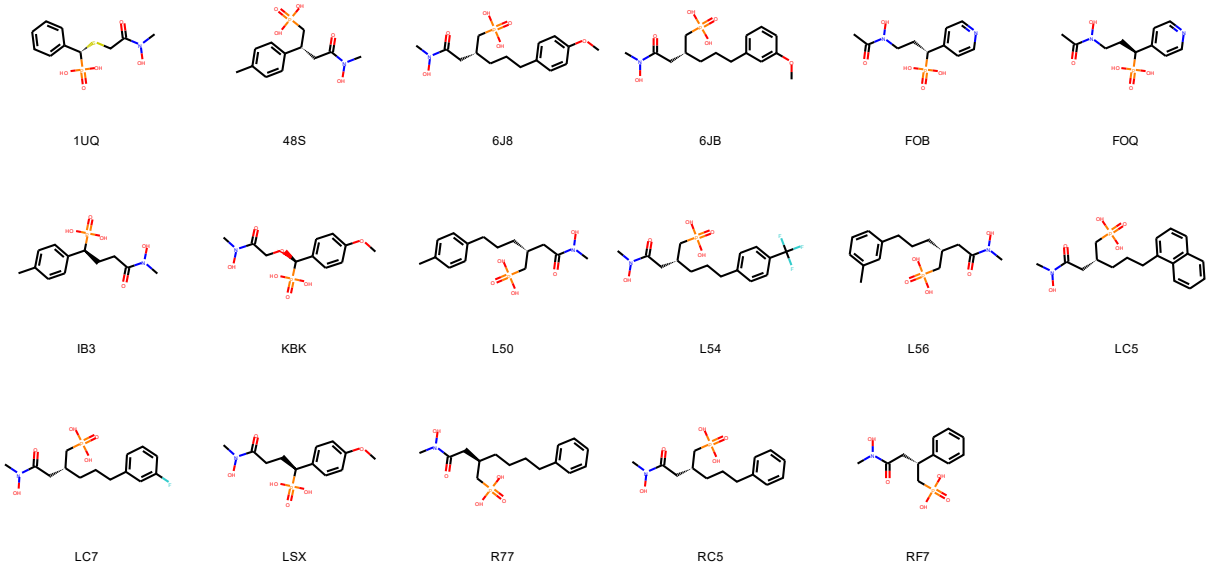


Figure 3-1 PfDXR inhibitors structures. IDs represents ligand IDs in the PDB structures. Structures were drawn using RDKit [162].

3.2.1.3 Target structure

The crystal structure PDB ID: 5JAZ from PDB [290], PfDXR its closed conformation co-crystallized with LC5 ([(2R)-2—5-(naphthalen-1-yl)pentyl]phosphonic acid) was used for docking and MD. Only a monomer (Chain B) was used. To mimic the same assay conditions as LC5, the structure was used without NADPH. The manganese atom parameters for AMBER03 [114] force field were retrieved from previous study [291]. LC5, the target co-crystalized binds similarly to fosmidomycin with its hydroxamate coordinating Mn^{2+}, the phosphonate interacts ASN311, SER306, SER269, Ser270) in a polar region (Appendix H). In addition, the compound aromatic ring extends toward the loop covering the binding site [292].

3.2.2 LBVS and SBVS

A total 3,078,845 ZINC compounds and 17 PfDXR inhibitors descriptors of the different similarity search approaches were generated using the respective tools (Table 3-1) and GNU Parallel (version 20160422) [293]. Any compound having a run-time error with any method was removed. Each ZINC compound was compared to each known inhibitor using every listed method here. For each method, a consensus query approach was used [247]. The similarity scores across the 17 inhibitors were averaged and rank transformed. The ranks across the different methods were summed, and this latter was rank transformed and sorted, giving the final list of ranked ZINC compounds, thus ranking using the rank by rank approach of consensus scoring [211]. For convenience, the negative of the Euclidian distance was used in the case of obspectrophore in ranking the compound.

1 Table 3-1 LBVS methods and the used parameters

LBVS	Tool	Parameters	Similarity metric	Methods	References
Obspectrhphore	Open Babel / obspectrophore	ZeroMean, accuracy: 20°, resolution: 3 Å, stereospecificity: 'none'	Euclidian distance	Molecule environment properties (partial charges, lipophilicities, shape deviations and electrophilicities.)	294,295
USR		ODDT / shape.usr	usr_similarity	Distance atoms - centroids distributions	267,296
ES		ODDT / shape.electroshape	usr_similarity	Usr + atoms charges	267,297
USRCAT		ODDT / shape.usr_cat	usr_similarity	Usr + CREDO atom types.	266,297
MHFP	mhfp_encoder.encode	mhfp_encoder.encode(mol.smiles, radius = 3, rings = True, kekulize = True, sanitize = True)	1. - MHFPEncoder.distance	ECFP + molecular shingling + MinHash encoding	83
RDkit_3d_pharm	rdkit.Chem.Pharm2D import Gobbi_Pharm2D, Generate	Generate.Gen2DFingerprint(mol.Mol, factory, dMat = Chem.Get3DDistanceMatrix(mol.Mol))	Tanimoto similarity	Pharmacophore features points + inter-feature topological distances	298 269

2

3

The top 50000 compounds were selected for molecular docking. The set of PfDXR inhibitors and the receptor (chain B of 5JAZ) were prepared using AutoDockTools (ADT) [161]. Docking was performed using Q-vina-w [96], which uses the same SF as Vina [96]. Water molecules in the active site (at a distance <=4 Å to the co-crystallized ligand) were included after charge calculation using ADT graphical user interface. PfDXR active site was targetted by first validating the docking parameters (spacing of 20 Å and exhaustiveness of 64) with the co-crystallized ligand LC5. This latter was redocked to the receptor obtaining an RMSD value between the docked and co-crystallized pose of 0.58 Å, less than 1 Å, indicating good reproduction of the pose. Indeed, a threshold of 2 Å in RMSD is often used to qualify a docked pose as good [81]. The RMSD was computed using Obrms in Open Babel [299].

After docking, the top pose for each compound was rescored with the different SFs. Scores were transformed to their absolute values and integrated using the exponential consensus ranking scheme of a set of scoring function [211]. These were done using Dask [300]. The top 20 (Table 3-3) compounds were selected for MD simulations. The different tools and used parameters are summarized in Table 3-2.

Table 3-2 Scoring functions

Program/SF	Version	Classification	Affinity Unit	Reference
ODDT /Vina	1.1.2 (May 11, 2011)	Hybrid (Empirical+ knowledge-based)	kcal/mol	[301]
Cyscore	V 2.0.3	Empirical	kcal/mol	[302]
Xscore	V 1.2	Empirical	kcal/mol	[303]
Smina / Vinardo	Smina Feb 12 2019. Based on AutoDock Vina 1.1.2	Empirical	kcal/mol	[199,284]
Smina / Smina	Smina Feb 12 2019. Based on AutoDock Vina 1.1.2	Empirical	kcal/mol	[199]
Smina / ad4	Smina Feb 12 2019. Based on AutoDock Vina 1.1.2	Empirical	kcal/mol	[199]
Smina / dk_scoring	Smina Feb 12 2019. Based on AutoDock Vina 1.1.2	Empirical	kcalmol	[199]
DSX (DrugScore eXtended)	V 0.9 (17.04.2015)	knowledge-based	kcal/mol	[277]
ODDT /RF-Score	V 4	Machine Learning (Random Forest)	pKd	[274,301]
ODDT / RF-Score	V 1	Machine Learning (Random Forest)	(pKi/d)	[274,301]
ODDT / nnscore	V 2	Machine Learning (Neural Network)	(pKi/d)	[274,301]
ODDT / RF-Score	V 2	Machine Learning (Random Forest)	(pKi/d)	[274,301]
ODDT /RF-Score	V 3	Machine Learning (Random Forest)	(pKi/d)	[274,301]
ODDT / pleclinear	PLEClinear_p5_l1_s65536	Machine Learning (Linear regression)	(pKi/d)	[285,301]

3.2.2.1 Dask a python parallel computing library for large scale drug discovery

Unlike pandas [304] data frame commonly used in python data analysis pipelines, Dask data frame scales to computer clusters. Cells in Dask [300] can contain python objects like an RDKit [269] mol object. For instance, in our use case, a Dask data frame was used to rescore docked poses. Every row was a ligand (in the form of an RDKit mol object). The different SFs were wrapped into python functions to only return affinity values. Hence, columns were generated by applying the functions to the pdbqt files using a parallel scheme. From the RDKit[162] mol objects, input files are generated for scoring functions with paths saved in the same Dask dataframe or as RDKit mol object attribute, thus creating links for better organization. The dataframe can be saved in pickle format or in a more compressed formats: feather or parquet. This eases experimental setups, generation of inputs, parsing of output files, collection of data. More importantly, Dask Application Programming Interface (API) is similar and built upon Pandas API [305]. This offers extensive analytical functions on data frames, rows, and columns. These were used for LBVS and SBVS scores distributions analysis. Customized python methods can be applied to RDKit mol objects. For instance, this can be used to generate molecular properties for large data. Indeed, Dask is used for big data (billions of rows) analytics. This fits well the purpose of mining the large chemical space for molecular properties. Here, Dask was used with a small dataset of about 50000 ligands. The rationale behind this was to anticipate any time-consuming rescoring but also to make use of the associated pandas API for analysis. Moreover, Dask was found to be easily scalable, flexible as it adapts to different cluster configurations (SLURM, PBS, etc.), is user-friendly and can be used from a remote Jupyter notebook [190] running on the computer cluster.

3.2.3 Molecular dynamics

Eighteen of the top 20 compounds selected from docking were assessed in a 20 nanosecond MD run using GROMACS (version 2018.6) [182] and the AMBER03 [114] force field to assess their stability. Acpype [186] was used to generate ligands' topologies. MDs were run in a cubic box with a distance between the solute and the box of 1.0 nm and using the Simple Point Charge (spc216) solvent model with a concentration of 0.15 M (Na^+ and Cl^- ions). Systems' energies were minimized using the steepest descent method with a maximum force set at <1000.0 kJ/mol/nm and a maximum number of steps of 50000. The temperature was set to 300K and the pressure at 1 atmosphere for a 50 ps equilibration in the isothermal-isobaric ensemble and later in the canonical one. The particle-mesh Ewald algorithm was used for long-range electrostatic interactions and the short-range non-bonded interaction cut-off distance was 1.2 nm. The equation of motion was integrated using a time step of 2 fs. After MDs, GROMACS[112] modules and PYTRAJ [189] were used for analysis. Protein-ligand interaction energy (PLIE) [106] was used to assess stability in the protein: fluctuations in the PLIE can be informative for ligand stability. Also, the metric can be useful for ranking ligand by affinity. Simulations were carried out on a remote machine at the CHPC and visualized using NGLview [188] in a Jupyter Notebook [190]. From the MD simulations, the top nine ligands were selected for MM-PBSA, and US.

3.2.4 MM-PBSA Binding Free Energy

MM-PBSA is a popular approach in computational drug discovery, often used as an in-silico validation approach of hits and to reproduce experimental findings [108]. Binding free energy (BFE or ΔG_{US}) calculations were performed for the last five ns of 20ns dynamics where frames are

saved every 100 ps (50 frames) using the g_mmpbsa package (version 1.6) [306,307]. G_MM-PBSA calculates relative binding free energy using the MM-PBSA method, using Molecular mechanics (MM), Poisson Boltzmann (PB), and Solvent Accessible (SA) energy values. The BFE of the protein-ligand complexes was calculated using equations (3-4) to (3-7).

$$\Delta G_{binding} = G_{complex} - (G_{protein} + G_{ligand}) \quad (3\text{-}4)$$

$G_{complex}$, $G_{protein}$ and ΔG_{ligand} : isolated free energies of the complex (protein-ligand), the protein and the ligand, respectively.

$\Delta G_{binding}$: binding free energy of the protein-ligand complex in the solvent.

$$G_x = \langle E_{MM} \rangle - TS + \langle G_{solvation} \rangle \quad (3\text{-}5)$$

G_X : free energy for each entity: ligand, protein, or protein-ligand complex.

$\langle E_{MM} \rangle$: average mechanical potential in a vacuum.

TS : entropic contribution (T is temperature and S is entropy).

$\langle G_{solvation} \rangle$: free energy of solvation.

$$E_{MM} = E_{bonded} + E_{nonbonded} = E_{bonded} + (E_{vdW} + E_{elec}) \quad (3\text{-}6)$$

E_{MM} : vacuum molecular mechanics potential energy.

E_{bonded}: bonded interactions such as bonds, dihedrals, angles, and improper interactions.

$E_{nonbonded}$: non-bonded interactions: electrostatic and van der Waals interactions modelled using Coulomb and Lennard-Jones (LJ) potential functions.

$$G_{solvation} = G_{polar} + G_{nonpolar} \quad (3\text{-}7)$$

$G_{solvation}$: energy required to transfer the protein-ligand solute from a vacuum into a solvent.

G_{polar} : electrostatic energy contributions.

$G_{nonpolar}$: non-electrostatic energy contributions.

3.2.5 Steered molecular dynamics (SMD) and umbrella sampling (US)

The top eighteen compounds together with LC5 were also simulated in SMD and later nine were selected for US. SMD mimics the Atomic Force Microscopy (AFM) experiment and has been successfully used to study protein-ligand binding affinity [103,308–310]. A constant velocity, constant force (cv-cf) SMD was applied to the selected protein-ligand systems. A monomer (PDB ID:5JAZ chain B,) was used for the docking, MD, SMD, and US simulations. The Caver 3.0.1 Pymol [225] plugin was used to find an optimal unbinding path for the co-crystallized ligand (LC5). The tool finds possible unbinding pathways (Appendix I) from the binding pocket and ranks them based on

their characteristics which include the tunnel bottleneck radius, length, curvature, cost, and throughput to find an ideal unbinding path and was used in earlier SMD studies [308,311–313]. The C-alpha atoms of four residues (residues number: 329, 264, 330, 265, Figure 3-2) located at the opposite extreme of the unbinding path were restrained to avoid dragging the protein during pulling. The last conformations of the MD simulations were used as starting structures for the pulling simulations after transformation to fit the Caver unbinding path/pulling direction to the z-axis. Pull groups were the ligand and the protein, the reaction coordinate (ξ) was defined as the COM of the ligand and pulling direction was given by Caver (see red arrow in Figure 3-2). A single dimension free energy evolution along ξ was constructed. A harmonic biasing potential defined by a force constant of 1,000 kJ mol^{-1} nm^{-2} (~ 1,700 pN/nm, the upper limit of k in AFM experiments) with a pull rate of 0.005 nm/ps was applied on the ligand COM. The geometry of the reaction coordinate was direction-periodic. The SPC water model was used as a solvent, and 100 mM NaCl was present in the simulation cell, which was a dodecahedron box. The systems' energies were minimized using the steepest descent method with a maximum force set at <1,000 kJ/mol/nm and a maximum number of steps of 50000. A 100 ps equilibration was done under an Isothermal–isobaric ensemble (NPT) ensemble. For each system, ten replicate simulations of one ns each (including minimization and equilibration phases) each using a random seed were performed. The results (pulling work (Wpull), and the rupture force (Fmax)) from the 10 different simulations were aggregated for each metric by averaging them. In umbrella sampling, the different configurations (umbrellas) were extracted from the ξ of the first SMD simulation using a window spacing of 0.05 nm. Configurations were equilibrated under an NPT ensemble for 100 ps and finally put into a ten ns simulation with a pull rate of zero. Coordinates were saved every picosecond in the corresponding SMD simulation.

All simulations were conducted with GROMACS, version 2018.6 using the AMBER03 [114] force field. The analysis was done on the pulling work (Wpull) (equation (3-8)), the rupture force (Fmax), and the $\Delta G_{TIE.}$ This latter refers to the total interaction energy difference between the unbound and bound state of the ligand. The protein-ligand interaction energy is the total non-bonded interaction energy (short-range Coulombic and Lennard-Jones) giving an estimation of the strength of the interaction [106].

The PMF was obtained from the histogram analysis using the WHAM algorithm via the g_wham package in GROMACS[112], with 200 bootstraps to estimate statistical uncertainty [314]. The binding free energy ΔG is the difference between the maximum and the minimum values of energy on the PMF curve [315].

$$Wpull = v \int_0^t F(t) d\, t \qquad (3\text{-}8)$$

$Wpull$: work of the external force.

v : pulling speed.

F: external force.

t : time.

The screening workflow is summarized in Figure 3-2 below.

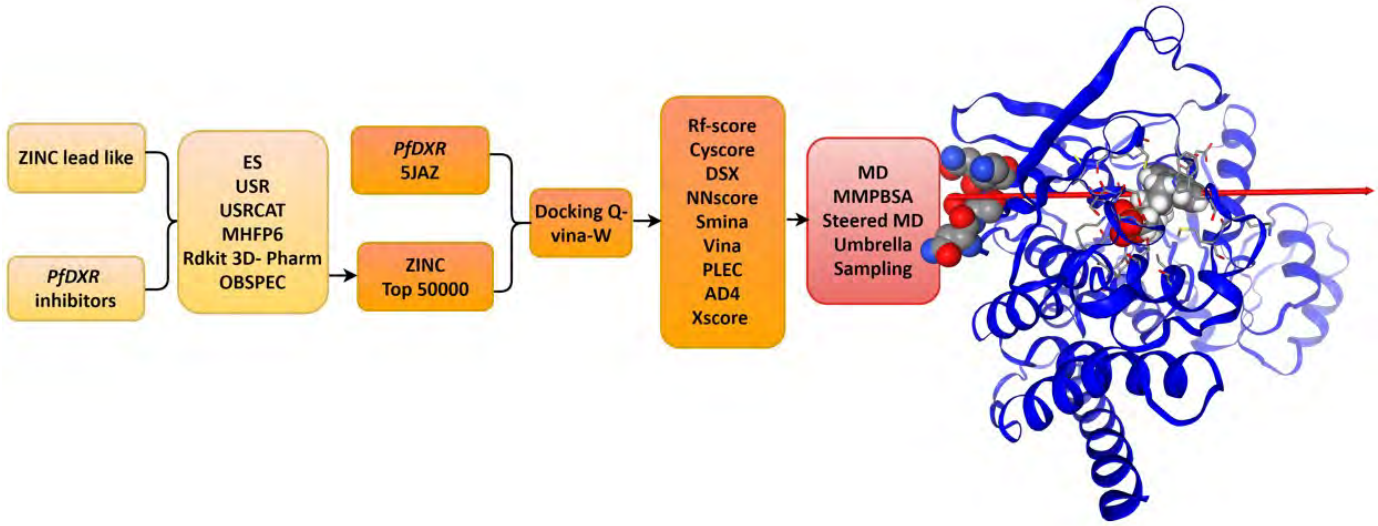


Figure 3-2 Screening workflow summary. PfDXR's structure is represented in blue ribbon. The red arrow represents the pulling direction with restraint residues at the back (extreme left) and LC5 in the middle in ball and stick representation.

3.3 Results - Discussions

3.3.1 Ligand-Based Virtual Screening

3.3.1.1 LBVS hits

About 3 M ZINC compounds were compared to 17 DXR inhibitors using six LBVS methods.

Figure 3-3 shows the structures of the top 16 compounds identified after using a consensus ranking.

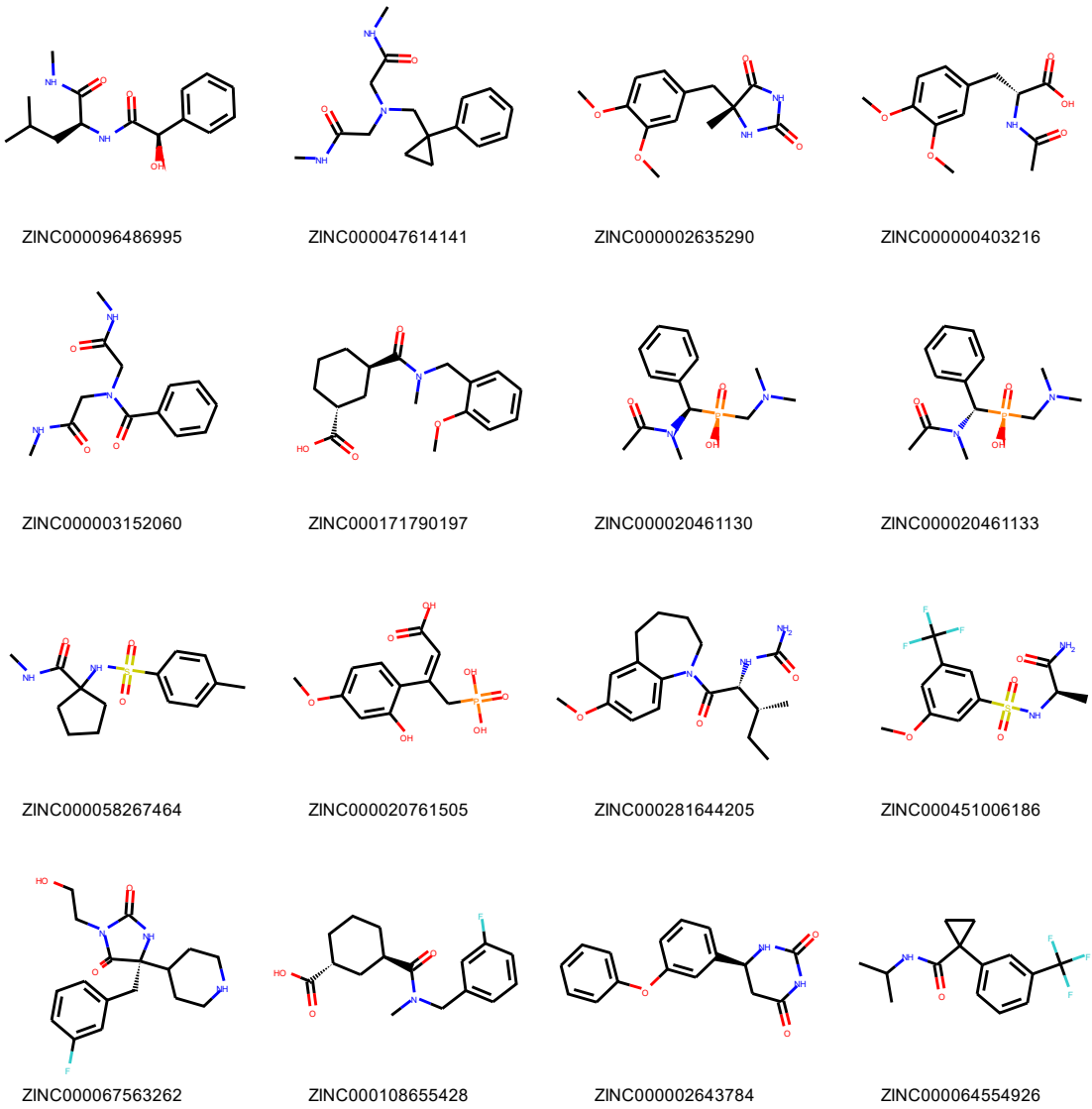


Figure 3-3 Top 16 LBVS hits structures. IDs represent ZINC [316] database IDs of the structures. Structures were drawn using RDKit[162]

The selected hits structures from LBVS lacked the hydroxamate group, a key feature, known for metal chelation in the known inhibitors [317]. Indeed, no hydroxamate group was found in these top compounds, a moiety conserved among PfDXR inhibitors' structures in their bound conformations and used as queries. However, both normal and reverse orientations of the group can be effective for inhibitory activity [318]. Amide groups were present in ZINC000281644205 and ZINC000451006186. ZINC000003152060, ZINC000096486995, and ZINC000000403216 present donor groups which may coordinate, making the compound bidentate ligands serving for metal chelation. Of these hits, only ZINC000020761505 presented a terminal group matching known inhibitors, which bind in a region of the active site forming a network of hydrogen bond with the residues SER269, SER270, SER306, ASN311, LYS312 and HIS293 [319]. New ring structures: propyl, hexane and heptane were present while all rings in the known inhibitors were benzene rings.

Also, some compounds present have many rings (e.g., ZINC000067563262, ZINC000002643784, ZINC000058267464) while the inhibitors, except LC5, only present a single benzene ring. ZINC0000020462230 and ZINC0000020461133 are isomers.

3.3.1.2 Similarity scores distribution

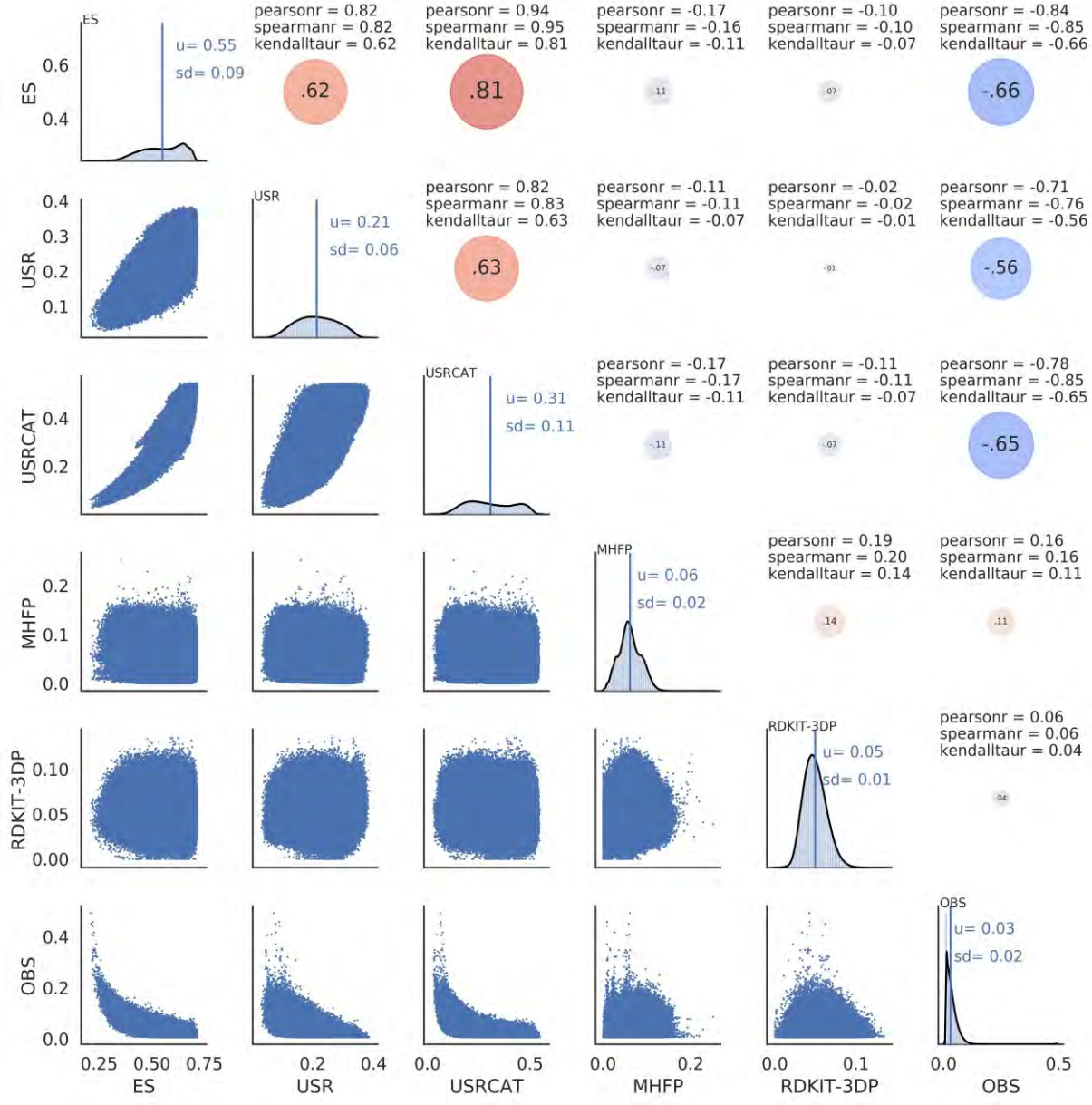


Figure 3-4 LBVS scores distributions and correlations. The sizes of the dots in the upper triangle of the grid represent are proportional to the Kendall τ correlation coefficients. The red color indicates positive correlation while the blue one indicates the negative one. The vertical bars in the distribution plots on the diagonal indicate the means. SD values are also annotated on the diagonal. The grid plot was generated using Seaborn [198].

Figure 3-4 presents histograms of the similarity scores distributions for the ZINC compounds compared to PfDXR inhibitors. Similarity scores had different ranges (from the lowest to the best

scores): ES ([0.02 – 0.71]), MHFP [0 – 0.25], OBSPEC [49231-573], RDKit_3dpharm [0 – 0.13], USR [0.03 – 0.38] and USRCAT [0.03 – 0.54]. The standard deviations were 0.09, 0.06, 0.1, 0.02, 0.01 and 2238 for ES, USR, USRCAT, MHFP, RDKit_3dpharm and OBSPEC respectively. OBSPEC was the most dispersed distribution, followed by USRCAT and ES. Thus, these may have the highest screening/differentiation power for compounds. However, the large SD in OBSPEC could be related to the large left-skewness of the distribution which may result in many outliers influencing the SD. On the other hand, MHFP and RDKit_3dpharm have the most compact distributions with an SD of 0.02 and 0.01 respectively. They have maximum similarity scores as low as 0.25 and 0.13 respectively.

Except for ES, all methods have a low percentage of compounds passing the similarity threshold. Euclidean distance values from spectrophores showed a significant deviation from the low values. Indeed, the observed values here [49231-573] are much higher than those observed in the original paper ([0-700] and [0-1000]) [270]. Also, the threshold to infer compound similarity is 50 [270]. OBSPEC best score was 573 where a Euclidian distance of 50 a good threshold for similarity in virtual screening, which implies the absence of similar PfDXR inhibitors in the screened set. Comparing inhibitors to each other, high values were found with the lowest one being 588 for 48S. These high values may be caused by the absence of normalization which can result in a shift in the spectrophore values [295]. Indeed, the default normalization “none” was used while zero mean is recommended for virtual screening [270]. This may not impact ligand rank ordering.

About scores distribution, only RDKit_3dpharm and USR have bell-shaped curves or a normal distribution. OBSPEC scores distribution presents a long tail to the left. The bimodal nature of ES and USRCAT distributions may reflect their combination of the shape and the electronic nature of the compounds: each mode corresponding to one of the two characteristics. This bimodal nature may be related to the different weights given to the shape or electronic nature when merging. For ES, a scaling factor (μ) counts for the number of electron(s) per angstroms and makes the approach flexible between a purely shape-based and purely partial charge. RDKit_3dpharm also accounts for compound conformation (with the relative distance of atoms) and the pharmacophoric points but does not show a bimodal distribution. MHFP had an interesting symmetric distribution with four inflection points which may imply a mixture distribution. These may be linked to MHFP fingerprint design, as Tanimoto similarity scores are usually normally distributed [320]. While USR, ES, USRCAT, RDKit_3dpharm, and OBSPEC are shape-dependent MHFP is conformation independent.

3.3.1.3 Similarity scores correlation

The agreement between the different methods was analyzed using the Kendall tau correlation, a non-parametric statistical test ranging from 1 to -1. When compounds have a similar rank across different methods, the Kendall correlation is high (toward 1) and vice versa [321]. The different correlations values are present in the upper triangle of the grid (Figure 3-4). The different methods are correlated (Kendall tau correlation > 0.5) (Figure 3-4) except RDKit_3pharm and MHFP, which are the two fingerprint-based methods. The latter is particularly conformer independent. USR, ES and USRCAT use a similar method to compare compound shapes which may explain their strong correlation. Interestingly, USR does not encode atom typing, nor electronic properties, yet it shows a high rank correlation with ES and USRCAT. USR, even though

it only accounts for the shape of compounds, has a significant correlation with OBSPEC, ES, USRCAT. On the other hand, MHFP and RDKit_3pharm do not show a correlation to any of these methods, nor between themselves (Kendall tau correlation 0.14). RDKit_3pharm had the lowest Kendall tau correlation with USR (0.01). This may be linked to the absence of pharmacophore information in USR which is solely shape-based while RDKit_3pharm is specifically designed for pharmacophores.

Four shape-based methods (OBSPEC, ES, USRCAT and USR) showed high correlations between themselves. There were only two different methods. The final ranks may be biased toward the first four methods. A ranking scheme with associated weight to each method or using a set of decorrelated methods could have provided a better consensus ranking in this scenario. For example, principal component analysis of the different scores will provide the component with the highest variance which can later be combined using the exponential consensus ranking scheme [211]. Optionally, an approach with a weighting scheme and higher weights for these two fingerprint methods (RDKit and MHFP) could have been used when combining the different rankings.

3.3.2 Structure-Base Virtual Screening

This section presents the scoring function correlations.

3.3.2.1 Docking score distribution and correlations

The top 50000 ligands identified in LBVS were docked and rescored. A total of 48972 were successfully scored by all used SFs. The top 20 ligands were selected using exponential consensus ranking [211] for MD simulations. In the section below, an analysis of the docking scores is presented. As SFs estimate ligands' affinities for DXR, a correlation between the different scoring functions is expected.

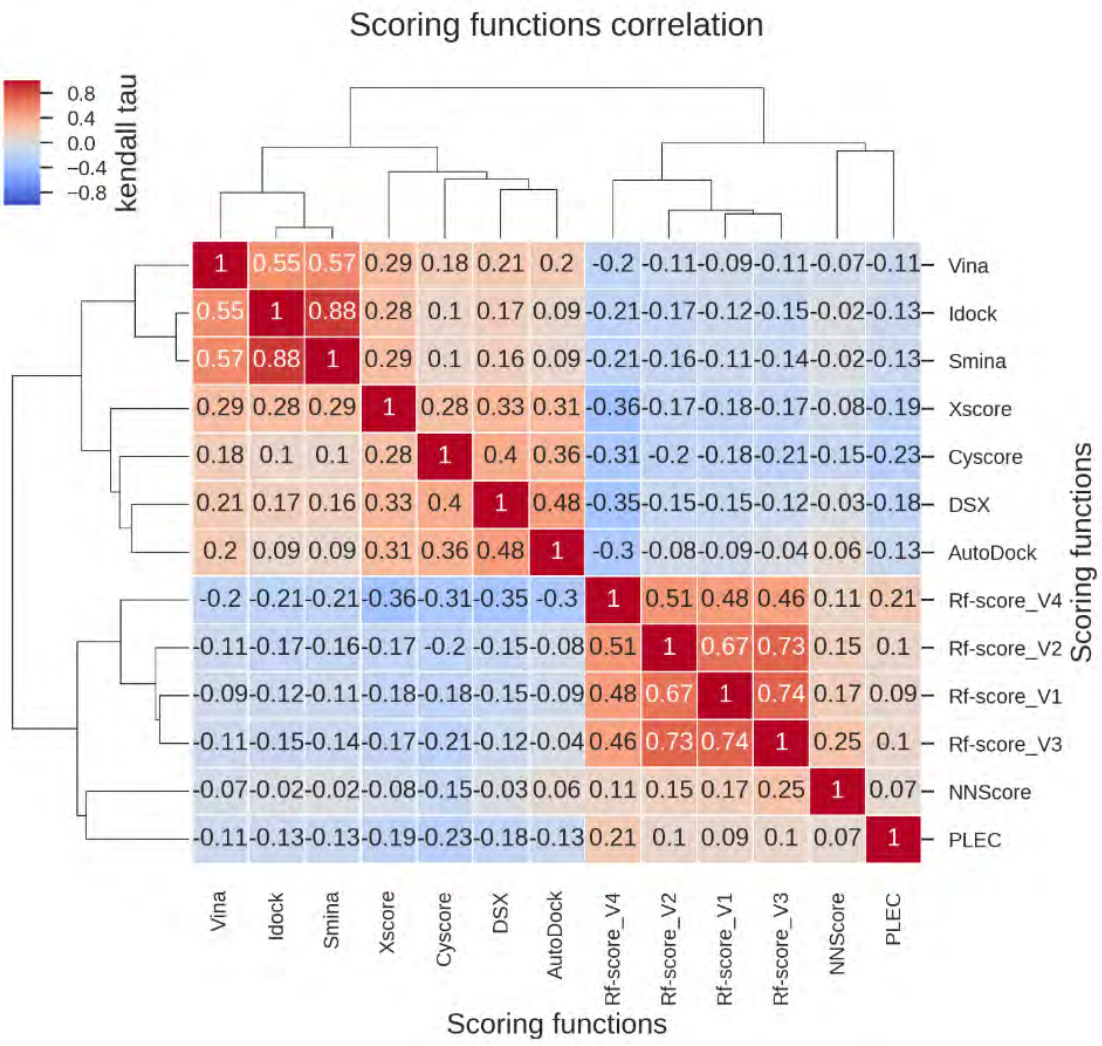


Figure 3-5 Clustermap of Kendall τ correlation coefficients between the different scoring functions. The color key is scaled from -1 to 1. The clustering is done based on the Euclidian distance between the different Kendall τ. NNScore, Rf-score SFs, PLEC predict affinity in positive pKd values and thus are negatively correlated with the other SFs predicting it in negative kcal/mol. Descriptive statistics for each SF distribution are in Appendix B. The figure was produced with Seaborn [198].

Figure 3-5 presents correlation coefficients (Kendall τ) between the distributions of the scores from the different SFs. Three main clusters are formed: (Vina, Idock, and Smina), (AutoDock, DSX, Cyscore, Xscore) and the Rf-score group (Rf-score_V1 to V4). PLEC and NNScore form a cluster but have negligible correlation (0.07) and showed the lowest level of correlation with the other SFs. Indeed, the highest correlation for NNScore was 0.25 with Rf_score_v3 and for PLEC the highest correlation was 0.23 with Cyscore. Among all SFs, the lowest correlation was observed between AutoDock and NNScore (-0.06) while Smina and Idock had the highest one (0.88). Interestingly, this latter was developed from Vina but here shows a better correlation with Smina.

An observed trend was the poor correlation between ML SFs (NNScore, Rf-score SFs group, PLEC) with the empirical and knowledge-based ones. The first three versions of Rf-score also showed a

high level of correlation (>0.7). Interestingly, its fourth version was the least correlated with the others but was closer to AutoDock, DSX, Cyscore and Xscore. Indeed, the SF had a high Kendall τ of -0.36 with Xscore for instance. Hence, its later versions tended to agree with classical SFs.

Clustering of the SFs reflected the distinct classes. Rf-score group uses the random forest algorithm. NNScore uses neural network while and PLEC uses a simple linear regression model. Idock and Smina were developed from Vina and hence clustered together. Cyscore, DSX, AutoDock, and Xscore formed a heterogeneous ensemble considering their different classes: empirical, knowledge-based force-field, and empirical, respectively. The lower correlation coefficient was lower inside that cluster than in the other ones. The highest was 0.48 between AutoDock and DSX.

Elsewhere, the features and datasets used to train these models also may affect this clustering. However, as shown above, the underlying class of SFs supports the clustering pattern observed.

These findings are difficult to generalize to other drug targets as PfDXR presents two peculiarities: substrate binding happens through an induced-fit mechanism and the protein presents a metal ion in its active site [56]. The rigid nature of these docking experiments may not fully model the induced-fit mechanism of binding and thus limit an accurate scoring. Also, many SFs may not accurately score metalloproteins. A previous study showed a variation between SFs accuracy in ranking compounds for zinc-dependent endopeptidases, where, in terms of the Spearman correlation the order of accuracy was DSX, X-Score, Vina followed by AutoDock [322].

Descriptive statistics for each SF are provided in Appendix B. Except for NNScore, Rf-score_v1, Rf-score_v2 and Rf-score_v3 which all had a bimodal distribution, all the SF distributions had a bell-shaped curve characteristic of a normal distribution with different ranges. The bimodal distribution may indicate the mixture of two normal distributions [323] indicating an SF with two key parameters contributing to the affinity estimation.

The above analysis showed a lack of correlation between all SFs, justifying the need for a consensus approach. However, some SFs may simply be predicting the affinity inaccurately, while it is unlikely that all methods are inaccurate. An approach to mediate this could be to use known PfDXR active compounds to assess each SF accuracy in rank ordering.

The following SFs were selected for compound ranking to avoid bias toward a specific SF. For instance, given the high correlation between Rf-score SFs they might bias the final ranking if they were all used. The scores were combined using the exponential consensus ranking scheme [211]. Consensus scoring is similar to multiple sampling in which the mean value has a higher probability to be close to the truth than any of the single SFs given their current limitation. It rectifies or minimizes the errors in the result when compared to individual scoring, thus reducing false-positive and improved ranking and hit rate. Hence, it is advantageous over using a single scoring [324]. A task was to define a combination rule. Recently, the exponential consensus ranking scheme has been shown to outperform previous combination schemes [211]. The top 20 hits (Table 3-3) selected were further assessed in MD simulations.

Table 3-3 Top 20 ligands selected from the exponential consensus ranking. Ligands are ranked from top (1st) to bottom (20th). RFS denotes RF score.

ZINC IDS	RFS_V1	VINA	NNSCORE	RFS_V2	RFS_V3	PLEC	RFS_V4	IDOCK	CYSCORE	DSX	SMINA	AD4	XSCORE
ZINC000002969522	7.83	-8.52	7.61	7.52	7.40	6.74	7.59	-8.00	-4.58	-111.47	-8.17	-38.48	-8.70
ZINC000000202238	8.05	-8.32	5.68	7.52	7.40	6.05	7.28	-8.77	-4.82	-114.24	-9.03	-38.44	-9.04
ZINC000173601880	7.75	-9.27	6.73	7.38	7.39	5.56	7.22	-9.73	-4.34	-116.11	-9.70	-39.32	-9.18
ZINC000008735333	7.83	-8.46	6.49	7.44	7.56	6.96	7.35	-8.91	-4.61	-101.52	-9.02	-37.57	-8.87
ZINC000225472873	8.14	-8.16	6.42	7.54	7.28	5.43	7.28	-8.28	-4.51	-124.19	-8.51	-35.51	-8.75
ZINC000230215778	8.11	-8.49	7.09	7.24	7.43	6.68	7.04	-8.15	-4.16	-120.67	-8.48	-37.42	-8.92
ZINC000072302893	7.85	-8.64	6.51	7.35	7.16	5.37	7.04	-8.90	-3.65	-115.66	-9.24	-36.64	-8.88
ZINC000010271232	7.78	-8.42	6.87	7.19	7.30	5.83	6.96	-8.72	-4.27	-107.14	-8.69	-31.91	-8.76
ZINC000057348471	8.15	-8.12	6.77	7.48	7.62	5.14	7.29	-8.73	-3.98	-116.10	-8.91	-32.96	-8.76
ZINC000193973285	7.93	-8.32	6.89	7.59	7.47	5.96	6.92	-8.67	-3.95	-106.07	-8.70	-31.02	-8.56
ZINC000409241945	8.31	-8.16	5.72	7.54	7.37	6.03	7.74	-8.92	-4.18	-125.93	-9.13	-33.10	-8.90
ZINC000028943558	7.78	-8.47	6.71	7.37	7.30	5.18	7.00	-8.49	-4.36	-109.26	-8.74	-32.93	-8.73
ZINC000000182272	8.14	-8.87	6.20	7.25	7.08	5.73	7.16	-8.85	-4.24	-105.58	-9.04	-32.02	-8.82
ZINC000013940913	8.18	-8.33	6.01	7.22	7.15	5.29	7.31	-9.22	-3.91	-107.26	-9.19	-35.12	-9.04
ZINC000091845778	7.73	-8.39	6.15	7.21	7.21	5.90	7.30	-8.75	-4.12	-102.52	-8.82	-32.38	-8.84
ZINC000058430530	8.23	-8.37	6.21	7.53	7.14	4.75	7.35	-8.84	-4.13	-109.67	-8.79	-34.89	-8.87
ZINC000023128752	7.49	-8.18	6.62	7.46	7.35	6.47	7.44	-8.05	-4.85	-126.51	-8.17	-43.32	-9.02
ZINC000065625934	7.72	-7.73	6.92	7.42	7.13	5.31	7.18	-8.29	-4.27	-111.29	-8.31	-36.72	-9.08
ZINC000050633276	8.25	-8.44	6.27	7.22	7.12	4.93	7.29	-9.06	-3.75	-119.66	-9.12	-32.15	-8.73
ZINC000065625931	7.62	-8.01	6.28	7.26	7.10	6.23	7.31	-9.76	-3.35	-115.01	-9.86	-35.67	-8.95

3.3.3 Molecular dynamics

In the following MD sections, results from conventional MD, SMD, MM-PBSA, and US are presented. SMD (Wpull and Fmax) was used to select compounds for MM-PBSA and US. Final hits were selected from the US simulations. The diagram below (Figure 3-6) gives a flowchart of the simulations and the ID and rank compound ranking ordering in each simulation type. The top 20 compounds were selected from rescoring. Eighteen ligands and LC5 were simulated in MD and SMD. Further eight ligands and LC5 were selected for US and MM-PBSA. SMD Wpull and Fmax are aggregated from ten independent replicates.

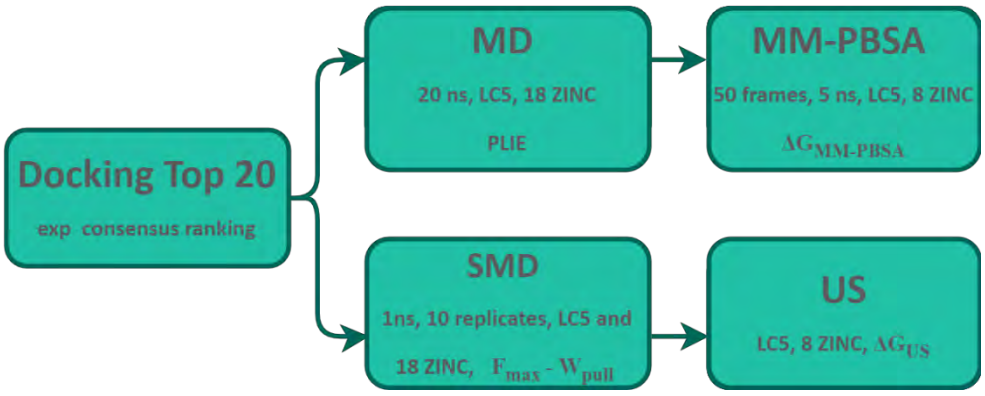


Figure 3-6 Simulations workflow and hits selection.

A range of metrics can be derived from MD simulations, RMSD, radius of gyration, protein-ligand COM distance to assess protein-ligand complexes stability. They are derived from relative atomic coordinates. As structural variations are driven by the energy gradient, we decided to focus on the analysis of energy-related metrics to reduce data dimensions for simplicity. More energy-based metrics also give an estimation of not only stability but affinity between the protein and the ligand. The protein-ligand interactions are also analyzed through the residues' contributions to binding energy in MM-PBSA and the broken interactions in SMD.

PLIE (Appendix C) was derived and analyzed from conventional MD. It gives a relative measure of affinity which can be used to rank compounds. Ligands outperforming LC5 in PLIE may have good experimental PfDXR affinity. Only two compounds ZINC000050633276 (-273.36 kcal/mol) and ZINC000230215778 (-272.91 kcal/mol) outperformed LC5 according to the average PLIE (Appendix C). Indeed, LC5 showed an average PLIE of -272.91 kcal/mol. All systems showed negative PLIE showing favorable binding with an average ranging from -273.36 kcal/mol (ZINC000230215778) to -166.09 kcal/mol (ZINC000013940913). Contrary to MM-PBSA, LC5 was ranked well according to PLIE.

3.3.4 Steered molecular dynamics.

SMD generates protein-ligand unbinding process configurations for sampling in US. Three metrics (the rupture force, the pull work, and the protein-ligand interaction energy) can also be derived from SMD and rank protein-ligand systems in terms of affinity. We analyzed the ranking of the

18 ligands selected from MD according to these three metrics. Simulations were not successfully run for two protein-ligand systems (ZINC000225472873, ZINC000008735333) due to challenge with water molecules settling during systems solvation causing failure in their energy minimization step.

3.3.4.1 Rupture forces

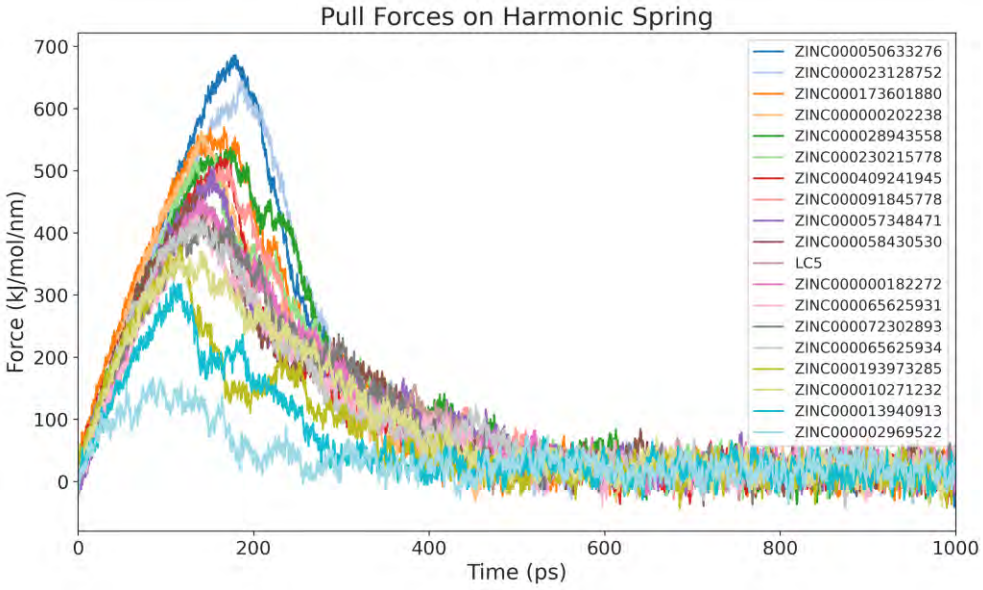


Figure 3-7 Pull forces on the harmonic spring. The y-axis represents the forces in kJ/mol/nm and the x one is time in picoseconds. The co-crystallized ligand is used as a reference for comparison. Ligands' names in the legend are sorted according to their rupture force. The graph was produced with pandas [304] and matplotlib [325].

Figure 3-7 shows the time-dependent evolution of the pull forces and the pull works (averaged from the ten independent replicated SMD). The curves followed the forms of observed pull force evolution in earlier pulling simulations [103,308,326,327]. There was a steady increase of the force to reach the rupture force (Fmax) followed by a decrease and finally and stable phase with a force of around zero continuing until the simulation end. ZINC000050633276, ZINC000023128752 and ZINC000173601880 were the top three compounds based rupture forces with 685.38 kJ/mol/nm, 644.47 kJ/mol/nm and 570.21 kJ/mol/nm, respectively. The co-crystallized ligand LC5 had a rupture force of 453.71 kJ/mol/nm. The top ZINC compounds thus showed better affinities for the protein than the co-crystallized ligand. This latter is a potent inhibitor in the nanomolar range (280 nM). Hence, the ZINC compounds could potentially be good inhibitors. A set of ten other ZINC compounds had a better Fmax than LC5. All compounds attained their Fmax at around 100 to 200 ps. The corresponding time-point to the rupture force is called Tmax and tends to be proportional to Fmax [308]. This is similar to ligand resilience time in the binding site. This metric could also be used to rank the compounds. For example, ZINC000002969522 which has the lowest Fmax also showed has the lowest Tmax. On the other hand, ZINC000050633276 had the highest Fmax, had a lower Tmax than ZINC000023128752, the second compound with the highest

Fmax (Figure 3-7). Fmax was strongly correlated to Tmax with a Pearson correlation coefficient of 0.91 (Appendix D).

All compounds were fully solvated after the first 600 ps of simulation, with rupture forces attained before 250 ps. This supports SMD computationally cost-effectiveness compared to advanced free energy approaches which are more expensive without a significant gain in accuracy [312].

Rupture forces presented different peaks. Higher ones are associated with sharper peaks than lower ones. For example, ZINC000050633276 presented a sharp peak at Fmax while ZINC000002969522 shows a flatter profile. This flatter profile could correspond to successively broken interactions. While the sharper peak results from the simultaneous breakage of multiple interactions. The different orientations of the bound ligands and a unique pulling direction may contribute to this difference in the peak profiles. Ligands can bind in an orientation in which all or majority of protein-ligand interactions' directions are parallel to the pulling direction. Interactions are thus likely to have a simultaneous contributing effect against the rupture force. While when the ligands bound orientation presents interactions orthogonal to the pulling direction, we may rather observe a detachment effect during unbinding. In that process, interactions are successively broken, resulting in a lower force peak. This cooperativity of molecular interactions, increasing the rupture force, was also observed in a previous study [328].

3.3.4.2 Pulling work

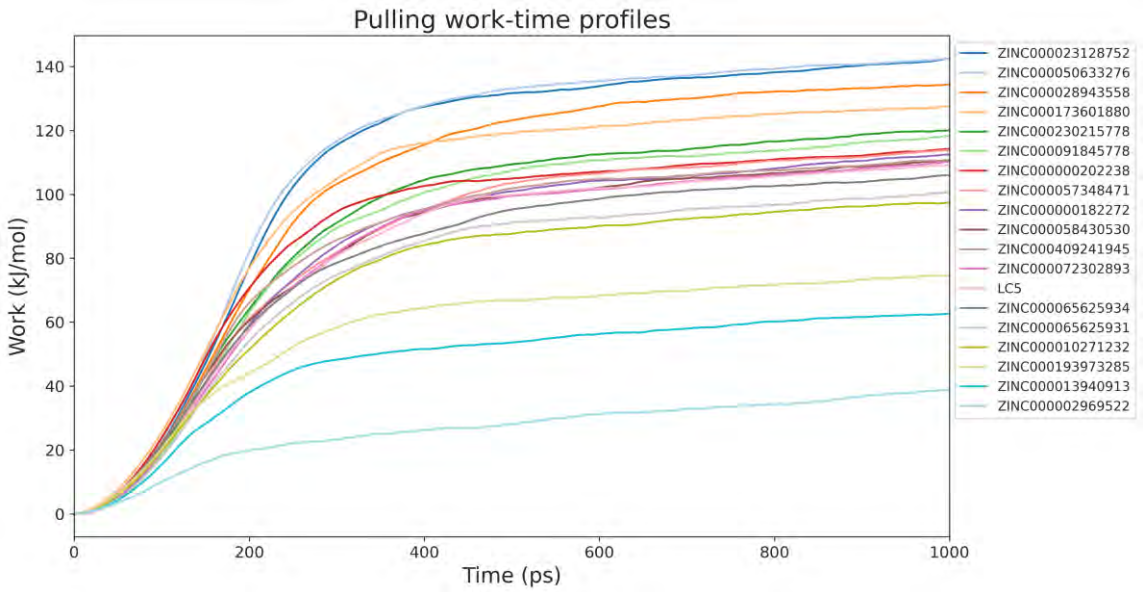


Figure 3-8 Pulling work-time profiles. The y-axis represents the work in kJ/mol unit and the x one is time in picoseconds. The co-crystallized ligand is used as a reference for comparison. Ligands' names in the legend are ranked according to the pulling works. The figure was prepared with matplotlib [325] and pandas [304].

Figure 3-8 shows the time evolution of the pulling works for the different systems. This latter has been shown to have better agreement with experiment than Fmax [308]. The top compounds were ZINC000173601880 (127.56 kJ/mol) ZINC000000202238 (114.25 kJ/mol) and ZINC000050633276

(142.25 kJ/mol) according to the pulling work (Wpull). LC5 showed a Wpull of 109.11 kJ/mol. Hence, with the pulling work, these compounds show better affinity than the co-crystallized ligand. Ten compounds showed better Fmax and Wpull than the co-crystallized ligand in total. The pulling work plateaus at around 500 ps with no significant variation afterward.

SMD results are dependent on the ligand pathway during unbinding from the protein [308]. Ideally, the ligand should be free to move in all dimensions during unbinding as in a realistic biological system. However, such a setup requires a larger simulation box in all dimensions, increasing computational cost [106,312]. This could be practical in SMD, as it is in the order of picoseconds. However, it can be expensive in US as independent simulations are carried out on different windows from the reaction coordinate. A current limitation of the current approach was the unidirectional pulling determined using Caver. Even though this direction may represent the lowest energetically cost for unbinding of the co-crystallized ligand, it may not be the optimum for other ligands especially if these are binding in different orientations in the active site. To overcome such limitations, SMD approaches with adaptative direction during simulation have been proposed [326]. A minimal steric hindrance showed better agreement with experiment than direction obtained from Caver [326,329]. However, these approaches are not available in the GROMACS simulation package [112]. Yang *et al.* and Gu *et al.* proposed an SMD method with adaptive direction adjustments where the optimum path of ligand is navigated by minimizing the pulling force automatically during the simulation [326,329].

As described in the methods sections, ten independent SMD were run and the metrics (Fmax, Wpull) were averaged across the different simulations. Another approach could have been to simply choose the system having the lowest energy during the unbinding.

Ligand binding to PfDXR happens through an induced-fit into a rather confined binding site with a loop over it acting as a lid [330]. These observations may make the unidirectional pulling not suitable for such a system. A slow velocity is important in such a scenario to allow enough time for the complex to relax and adopt more energetically favorable conformation, thus avoiding abrupt unbinding distorting binding site residues. Indeed, during experimental setup, different values of pulling force and velocities were tested. Further, the protein was not fully restrained during the SMD simulation. To avoid dragging protein along the reaction coordinate, restraints can be applied to the full receptor or part of it. For example, restrains can be applied to the protein C-alpha while keeping the side chain flexible [312,331]. In this study, only C-alpha atoms of four residues (residue numbers: 329, 264, 330, 265) located at the opposite extreme of the unbinding path were restrained. Hence the protein residues were flexible and able to adopt energetically favorable conformation during the unbinding process. This was to help prevent the entire protein structure from following the ligand during pulling and also prevent protein rotations which may affect the pulling direction. The protein remained flexible with possible conformational change for the active site residues.

It is important to note that the unbinding path for all ligands used was the same as for the co-crystallized ligand. Given that all ligands bind in the active site, the same direction should be closed to the optimum for each case. However, the unbinding path may be different for a ligand in a different pose.

3.3.4.3 Total protein-ligand interaction

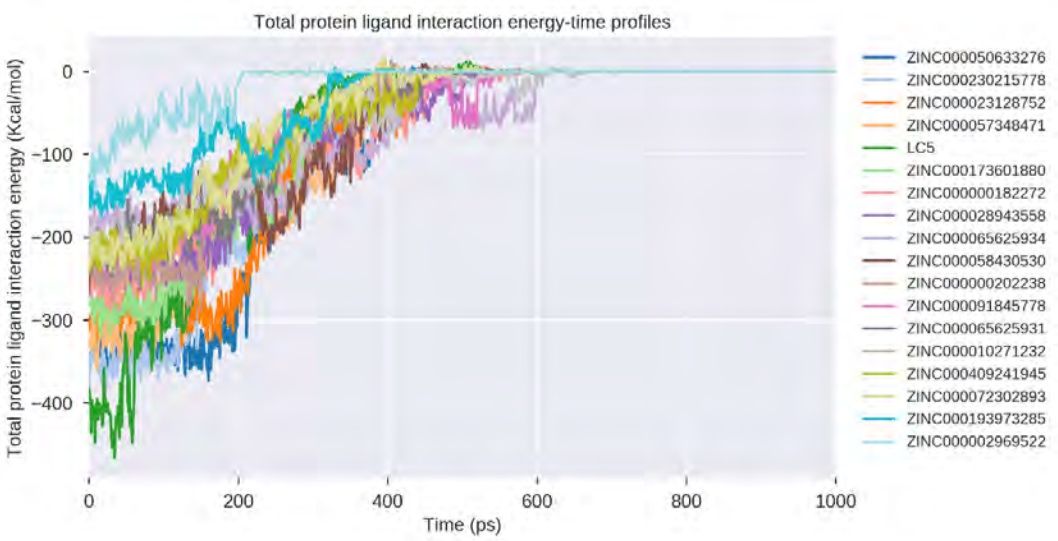


Figure 3-9 Total protein-ligand interaction energy-time profiles. The y-axis represents the PLIE in kcal/mol unit and the x one is time in picoseconds. The co-crystallized ligand is used as a reference for comparison. The figure was produced with matplotlib [325] and pandas [304].

Figure 3-9 shows the time evolution of the total PLIE between the protein and ligand during the SMD simulation. The total interaction energy (sum of short-range Coulombic and Lennard-Jones interaction energy) is a simple decomposition of the system potential energy, to only take into account the non-bonded terms between the selected atom groups (here the protein and the ligand). It does not correspond to a binding free energy [106]. All systems' G_{TIE} were negative, indicating favorable interactions between the ligands and the protein. Energy values increase toward zero where ligands are fully solvated. The profile was similar to the Fmax and Wpull profiles. The ΔG_{TIE} (total interaction energy difference between the unbound and bound state of the ligand) ranged from -154.47 Kcal/mol (ZINC000002969522) to -477.74 Kcal/mol (LC5). Hence, the co-crystallized ligand showed the best affinity. ZlNC000230215778, ZINC000050633276 and ZINC000057348471 were the top three ZINC compounds.

When considering G_{TIE} individual contributions, the vdW interaction energy shows higher contributions than the electrostatic one (Appendix E and Appendix F). However, in the top ligands, the electrostatic contribution tends to be higher. Indeed, LC5 and the top ZINC compounds (ZlNC000230215778, ZINC000050633276 and ZINC000057348471) showed more favorable electrostatic contributions relative to the vdW contribution. Hence, potentializing electrostatic contribute may be a good strategy for potent PfDXR inhibitors.

3.3.4.4 Events on the unbinding path

Force peaks are often associated with strong interactions in SMD [309,328]. We assume residues implied in these interactions to be strong anchoring points in the binding site for potent inhibitor design. Fmax is the value for the highest peak on the force profile and Tmax is the corresponding time-point. Figure 3-10 shows broken interactions at Tmax. A broken interaction has a relative frequency >= 0.5 before and <= 0.1 after Tmax. We assume those interactions to contribute to Fmax.

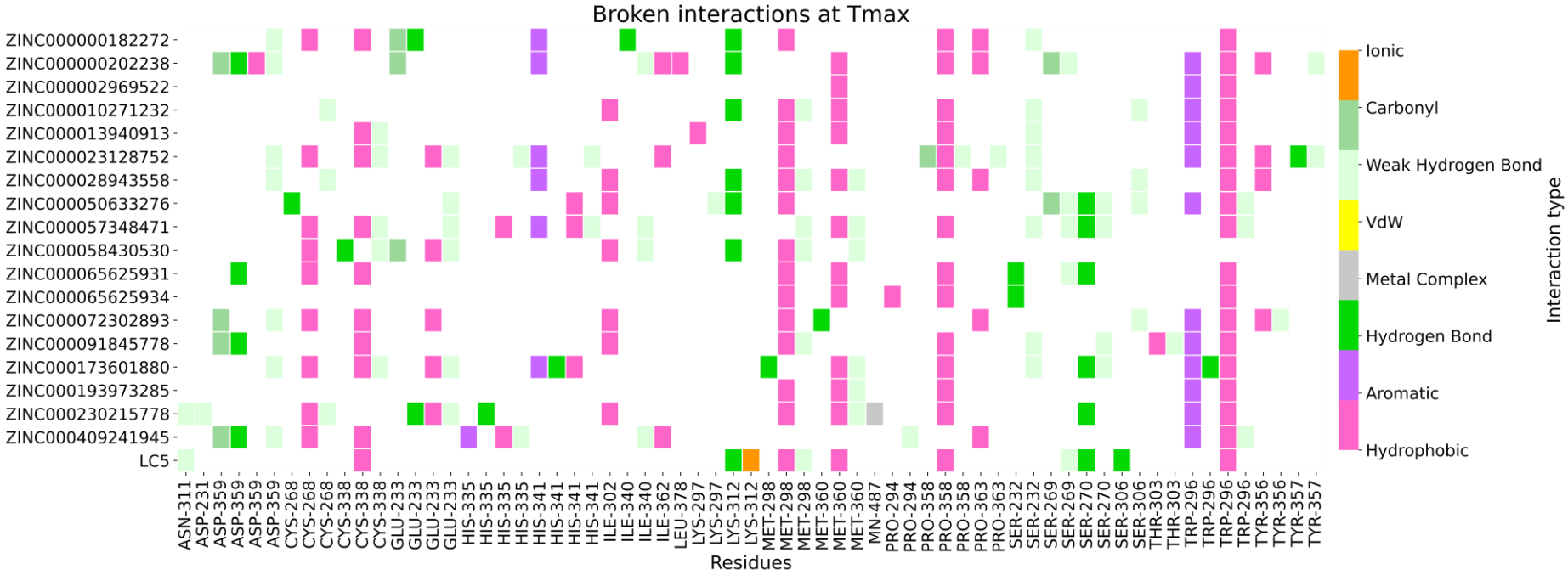


Figure 3-10 Broken interactions at Tmax. Ligands are on the y-axis and residues on the x one in their three letter code and residue number. Interactions and their types are represented by a colored box if present at Tmax. White areas represent the absence of interaction. Duplicate residues on the x-axis have different types of interactions. The heatmap was produced using Seaborn [198]. The broken interactions were analyzed on the first SMD simulation of the 10 replicates. Interactions were determined using Arpeggio [332].

Globally weak hydrogen bonds were the most common broken interaction, representing 35.4% in the type of interaction proportions. Hydrophobic and hydrogen bonds were also found in high numbers with respectively 27% and 23%. Rare carbonyl, aromatic, ionic, and metallic interactions were also found.

Regarding the residues, ASP359, HIS335, TRP296, HIS341, and GLU233 were the most frequently involved. MET360, MET298, PRO358, and CYS338, TRP296 were the most contributing residues to hydrophobic interactions. SER270, SER312, SER232, LYS312, GLU233, ASP359 were the most common residues where hydrogen bonds were broken at Tmax.

Some particular residues showed notable patterns. TRP296 is located on the loop covering the active site and had a hydrophobic contact identified as a broken interaction with all compounds except ZINC000058430530. By contrast, GLU233 was the only residue found to interact with the ligands among the metal coordinating residues. This residue could be interesting in optimizing inhibitor potency in that region. It is the most exposed among the residues implied in metal coordination, making it more accessible. LYS312, a buried residue in the phosphonate binding region made charged ionic interaction with LC5 only. This may explain its high protein-ligand interaction energy, especially in the electrostatic contribution (Appendix F). Hence, the possibility of forming strong charged, ionic interaction with this residue could be an exploitable anchoring point for optimizing inhibitors. MET298, while mainly implied in hydrophobic contacts, forms a hydrogen bond only with ZINC000173601880.

Comparing the frequency of interaction to Fmax intensity, ZINC000002969522 which has the lowest number of interactions broken (only three) also has the lowest Fmax (247.93 kcal/mol). On the other hand, ZINC000023128752 showed the highest number of broken interactions at Tmax with a rupture force of (610.96 kcal/mol). Interaction frequency is not necessarily linked to Fmax, as individual contributions may significantly vary. Hydrogen bonds, for example, may have a stronger contribution to Fmax than other interaction types. The heatmap indicated a higher frequency of hydrophobic contacts at Tmax than hydrogen bonds. However, hydrogen bonds most likely have a higher contribution to Fmax due to their stronger nature [218].

Broken interactions were only analyzed for the first SMD simulations. This set of simulations had a higher resolution (time steps between saving frames), where coordinates were written on disk every 500 steps contrary to the rest of SMD simulations in which they were written every 50000 steps. This was done to save disk space. Hence, frames corresponding to Tmax could be found with higher precision for the first replicate than the remaining ones. Yet, broken interactions could be computed on the other replicates considering the immediate frames before and after Tmax. The different broken interactions could then be aggregated using a probabilistic approach.

Broken interactions at Tmax are likely to be binding pocket gatekeepers. For instance, MET298 and TRP296 interacted with most ligands probably due to their location on the binding site loop. These may not necessarily be the only hotspots. Indeed, strongly binding residues, even buried in the active site may contribute to Fmax due to the slow SMD process allowing ligands to rearrange and maintain interactions. Another analytics approach could map all peaks on the force profile to interactions. Hence residues contributing to any peak could be identified.

To investigate residues association with Fmax intensity, Tmax was mapped to the corresponding broken interactions. This gives a QSAR-like model using Fmax as a continuous variable and the presence or absence of a specific interaction as a dichotomous variable (0 or 1). This results into a point-biserial correlation identical to the Pearson one [333].

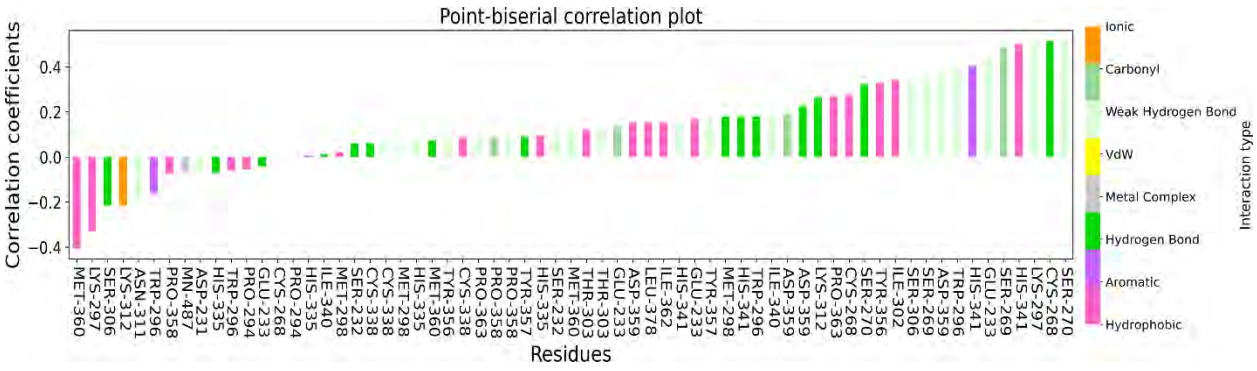


Figure 3-11 Point biserial coefficient plot. Coefficients of the point-biserial correlation between each specific residue interaction (residues and interaction type) and the intensity of Fmax.

Sixty-five residue interactions were identified with correlations ranging from -0.4 to 0.51. Most residue interactions did not show significant association with Fmax intensity, with a correlation coefficient in the range -0.2 to 0.2. CYS-268, SER-270, HIS-341 and LYS-297 showed the highest correlation (> 0.5) with high Fmax intensity. HIS-341 and CYS-268 were not previously noted as key residues in PfDXR binding site or to contribute to known inhibitor potencies. For instance, fosmidomycin beta-arylpropyl analogs preferably pointed toward TRP296 indole ring instead of HIS341 [289]. LYS-297 loop position may explain its correlation Fmax high values as shown in Figure. Other notable residues with a Pearson correlation above 0.3 were the well-known PfDXR residues GLU233, ASP-359, ILE302, SER269, SER306, TRP296. GLU233, TRP296, SER269 (Appendix A).

Some residue interactions showed a negative association with Fmax intensity. MET360 hydrophobic interaction had the lowest correlation (-0.4). It is located at the pocket periphery, toward the NADPH binding site. This peripheral location may explain its association with lower values of Fmax. This contrasts with its favorable contribution to MM-PBSA binding energy. A similar observation can be made for LYS312.

Residues correlated to high Fmax and contributing favorably to MM-PBSA binding energy are GLU233, CYS268, SER270, TRP296 and HIS341. Therefore, not only are they binding residues but they maintain this binding up to Tmax, hence their contribution to Fmax. Therefore, they may be stronger anchoring points. Broken interactions not present in MM-PBSA are likely implied as only guiding the unbinding process.

A limitation here is the unidirectional pulling SMD as previously noted [119]. This direction was determined using LC5 and applied to all ligands. This has already been discussed in section 3.3.4.2. In a future study, aggregating all data from the 10 SMD simulations and correlating Fmax to broken interactions in each replicate may provide a better sampling for better predictions. However, the current data comes from a set of 19 ligands, having different scaffolds and binding

differently from DXR. Hence, we can estimate a good sampling of the different possibilities in the binding pocket. As yet another alternative approach, PLIE could have been used instead of Fmax. Indeed, MD PLIE time series vs interactions information may also be used to build a QSAR-like model from the MD trajectories.

In the future, the identified interactions here may be developed into a pharmacophore hypothesis. De-novo compound generation biased toward these interactions may provide potent compounds. Combined with the knowledge of the binding pocket metal-binding regions and phosphonate binding region, these strategies may result in ligands with greater potency.

In some systems, the coulombic component of the interaction energy showed some peaks while the Lennard Jones component had already decreased to zero. When visualized, the ligands had electrostatic interactions with some residues on the active site loop (TRP296, LYS297, and LYS295) previously associated with high Fmax. For instance, ZINC000173601880 formed a weak hydrogen bond with LYS295 (Figure 3-12). This ligand is almost fully solvated, hence the total absence of LJ energy. Lysine residues may play a role in ligand binding/unbinding as charged side chains are expected to attract polar ligands[334]. Further, mutations K295N and K297S are reported to have a 24-fold decrease on PfDXR catalytic efficiency [335]. Hence, LYS297 and LYS295 residues may be key in substrate recognition.

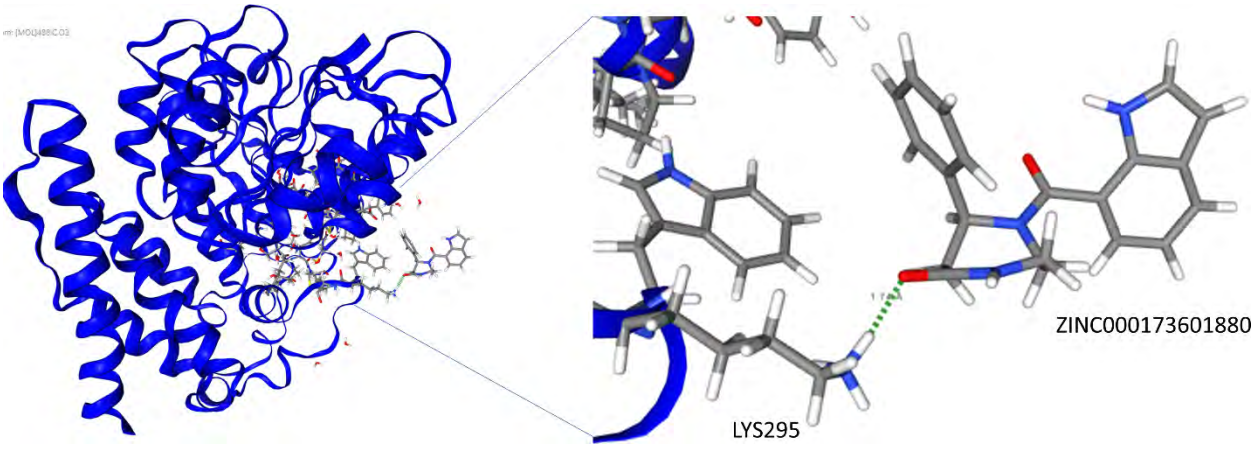


Figure 3-12 ZINC000173601880 last interaction in SMD. PfDXR in blue ribbon on the left. The active site area in zoomed in on the right. ZINC000173601880 and interacting residues are in licorice representation and atom types coloring. ZINC000173601880 formed a weak hydrogen bond showed in green dashed line with LYS295. The illustration was generated using NGLview [188].

SMD gives QSAR insights and is cost-effective considering the simulation length (about 1-2 ps). In addition, Wpull showed a good correlation with experimental affinity [331]. Hence, the approach is not only useful for a more accurate affinity prediction but also may be used in rational inhibitor design. It might be a more valuable approach compared to the classical conventional MD scheme combined with MM-PBSA. Beyond the analysis of residues associated with Fmax, SMD also provides a diversity of protein conformations especially in the binding site area. This diversity of

conformation may be valuable in the investigation into the potential energy landscape of binding, given that the DXR native ligand binds through an induced-fit mechanism [242].

3.3.5 Molecular Mechanics Poisson Boltzmann: MM-PBSA

3.3.5.1 Protein-ligand binding free energy.

MM-PBSA predicts protein-ligand complexes binding free energy (ΔG). The different energy components' contributions and the residue contributions provide further insights for rational inhibitor design. Positive contributions are unfavorable to the binding process while negative ones are favorable. Figure 3-13 shows the binding free energies for the ten systems simulated (LC5 and nine ZINC compounds).

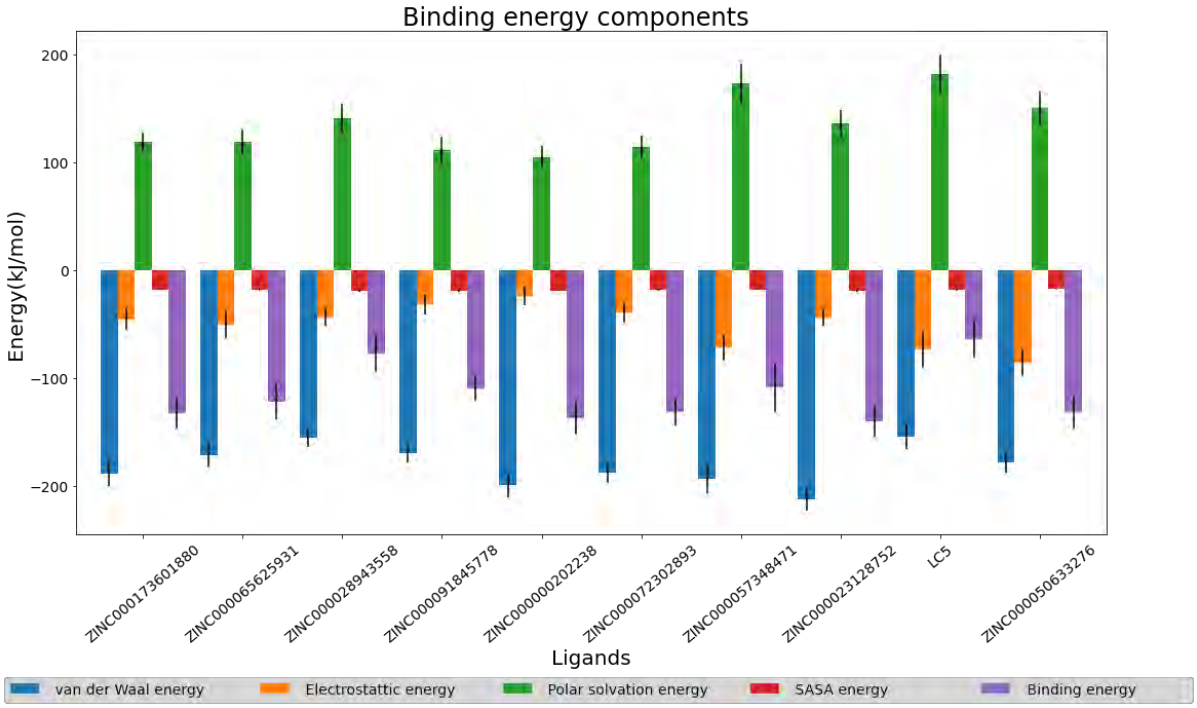


Figure 3-13 Binding free energies and their components for LC5 and the hits. Van der Waal, electrostatic, polar solvation, SASA contributions are presented. Standard deviations are indicated by error bars. The co-crystallized ligand is used as a reference for comparison. The bar plot was generated with matplotlib [325] and pandas [304].

All systems showed negative binding energy, thus showing their favorable binding. All ligand hits showed better binding energy than LC5. Indeed, LC5 had a ΔG of -64 kJ/mol while the most unfavorable of the hits was correspondingly -77.1 kJ/mol. Contrary to the results of SMD metrics (Wpull, Fmax, and PLIE) and US, LC5 was poorly ranked in MM-PBSA. The top three compounds identified in MM-PBSA were ZINC000023128752, ZINC000000202238 and ZINC000173601880 with binding energies of -140 kJ/mol, -137.14 kJ/mol and -132.63 kJ/mol. These compounds showed a twofold better affinity than LC5, an inhibitor potent in the nanomolar range. Hence these predictions indicate promising hits.

Interestingly, these top three ligands are ranked 5th, 6th, 7th according to the PLIE in SMD while LC5 was the top ligand. Hence compound rankings differed across the methods.

Regarding the different components of the free binding energy. Similar contribution patterns can be noted across the ligands. There was unfavorable polar solvation energy and negative vdW and electrostatic energies consistent across all systems. The vdW contributions were the most favorable while the polar solvation energy was the most unfavorable. vdW energy was also a significant contributor to the PLIE. Interestingly, LC5, ZINC000050633276 and ZINC000057348471 have the strongest electrostatic contributions in both MM-PBSA and PLIE. SASA contributions showed the least variance across the ligands. Hence, the binding pocket size and its accessibility to water molecules remained independent of the bound ligand. As a result, given the small, confined pocket, optimization strategies may benefit from prioritizing substitution on a lead or hit scaffold rather than further their expansion in the pocket.

The polar solvation energy was the most correlated with binding energy (Pearson correlation of 0.89). Hence, its optimization might an excellent strategy for potent PfDXR inhibitors. On the other hand, the electrostatic energy was poorly correlated (0.09).

MM-PBSA is known for its poor precision [336]. Here the standard error of the mean ranged 4.24 to 1.62 kJ/mol for ZINC000244774073 and ZINC000091845778, respectively. It is the ratio of the standard deviation to the square root of the number of frames [336]. Hence, this seems to give better precision in estimating the binding free energy compared to 11 and 14 kJ/mol reported by Weis *et al.*[336,337]. This may be linked to the default 10 ps time step extended to 100 for better sampling and to save computational cost. However, replicating simulations (20 -50 replicates of 100-200 ps simulations) is known to provide better precision [108].

3.3.5.2 Residues contributions to the binding energy

G_mmpbsa also estimates the residue's energetic contribution which is obtained through the decomposition of the total binding energy. This helps to gain insight into the residue's interactions. Figure 3-14 shows residues energetic contributions to binding free energy for the nine ZINC compounds and LC5.

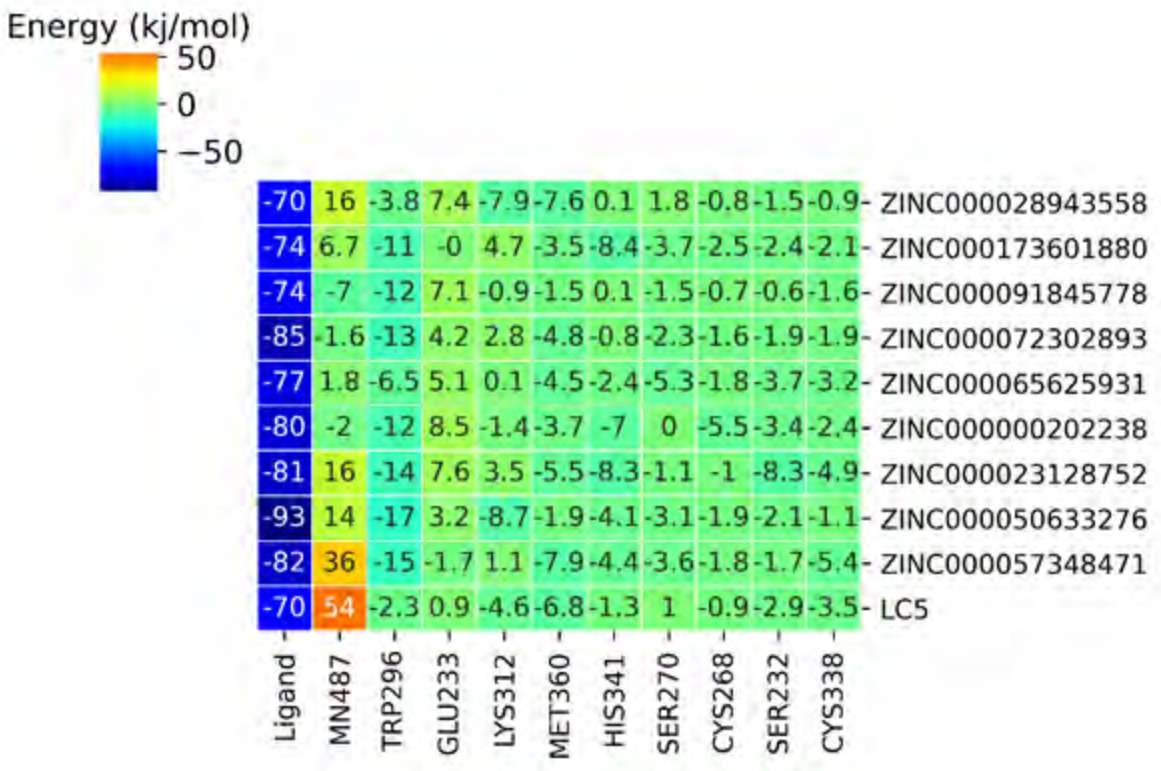


Figure 3-14 Residues energetic contributions (kilojoules per mole) to the total binding free energy. Residues are on the x-axis while the ligands are on the y-axis. From left to right residues are ordered from the highest variations to the least in their contributions. The figure was prepared using Seaborn [198].

SER232, GLU233, CYS268, SER270, TRP296, LYS312, CYS338, HIS341, MET360, and the MN metal were the greatest contributors to binding energy either favorably or unfavorably (Figure 3-14). Except for CYS268 and MET360, these residues were known to be implied in inhibitor interactions with PfDXR [338]. SER232, SER270, and LYS312 are implied in hydrogen bonding while SER270 and LYS312 are known to bind to the fosmidomycin phosphonate moiety [56,244].

GLU233 is a coordinating residue of MN and is involved in hydroxamate binding [242,339]. It is the only residue involved in metal chelating that was present among the most contributing residues. The absence of ASP231 and GLU315 might be explained by their buried nature. Further, to the best of our knowledge, no known PfDXR ligand-bound crystal structures interact with these residues.

TRP296 consistently showed the most favorable contribution to the binding energy. As shown in the interactions heatmap, it was also implied in many aromatic and hydrophobic interactions. Interactions with it are known to potentialize PfDXR inhibitors as it is known to form an aromatic hotspot in the loop region covering the active site [289].

In general, the MN had an unfavorable, positive contribution to the binding energy; only in three cases, it was favorable to it. ZINC000091845778 was the only ligand hit showing interaction with the metal. The MN contributed favorably to the binding energy in this specific case. In a similar MM-PBSA calculation, Anu *et al.* prioritized ligands showing interaction with the metal when

selecting hits [255]. Given the MN favorable contribution to binding energy in ZINC000091845778, this might be a valuable option for optimizing the ligand. Furthermore, it is interesting that LC5 is known to chelate MN, yet it has the poorest contribution from MN in this set with a value of +54 kJ/mol. Visualization of the simulation showed the hydroxamate moiety keeping only one oxygen coordinating to the MN coordination center. This may explain its poor ranking here.

Previous characterization of the PfDXR binding site showed three main regions: a polar one (SER269, SER270, SER306, ASN311), a hydrophobic one (HIS293, TRP296, MET298, CYS338, and PRO358) and the metal coordination region (ASP231, GLU233, and GLU315) [250]. The intersection of these residues with the above-mentioned ones (SER270, TRP296, and CYS338) might be a set of strategic residues with which to optimize interactions with for potent inhibitors. Despite the polarity of two core regions (metal binding and phosphonate binding), our results showed vdW interactions were the most contributing to the binding energy. Interestingly, CYS338, even though lodged in a hydrophobic region of the pocket, showed the ability to participate in hydrophilic interaction. This has also been noted in a previous study with interaction with CYS338 significantly potentialized pyridine containing inhibitors [338]. Additionally, SER232, CYS268, LYS312, HIS341, and MET360 may be residues to optimize interactions with as well.

Anu *et al.* performed MM-PBSA on PfDXR using the same crystal structure (PDBID: 5jaz) in complex with some natural products. A consistent, positively polar solvation energy was found. Raw binding energy values observed on the natural products show weaker binding than in this study. However, their simulation was done using a different force field (Amber99SB force-field) [255]. So, these comparisons should be taken with caution due to the different force field and GROMACS[112] versions. Further, in their study a ligand-dependent solute dielectric constant (pdie) was used [255].

Overall, as key findings, all hits showed more favorable binding than LC5. The residues energetic contributions combined with types of interactions provide insight into PfDXR rational-based drug design: a potent chelating agent is required to bind the metal ion, hydrogen acceptor binding must be possible in the phosphonate binding region, and finally it is necessary to optimize hydrophobic interaction in the loop. More specifically, optimizing interactions with residues such as SER270, TRP296 and CYS338 seem to be indicated for the design of potent PfDXR inhibitors.

In the set of experiments, the residues contribution to $\Delta G_{MM\text{-}PBSA}$ and the Fmax mapping to interactions gave information on the potential residues' contribution to the binding energy. Considering their overlap, we can note the following key points. GLU233 and the manganese atom positive (unfavorable) contribution should be minimized. TRP292 was the most frequent and the most contributing to $\Delta G_{MM\text{-}PBSA}$ favorably. LYS312, SER270, and SER232 contribute favorably through hydrogen bonds. HIS341 showed a strong contribution to aromatic and hydrophobic interactions. Finally, CYS338 and CYS268 MET360 contribute favorably through hydrophobic contacts.

3.3.6 Umbrella sampling

Top ligands for the US simulation were selected based on their Fmax and Wpull. Figure 3-15 presents their PMF profiles.

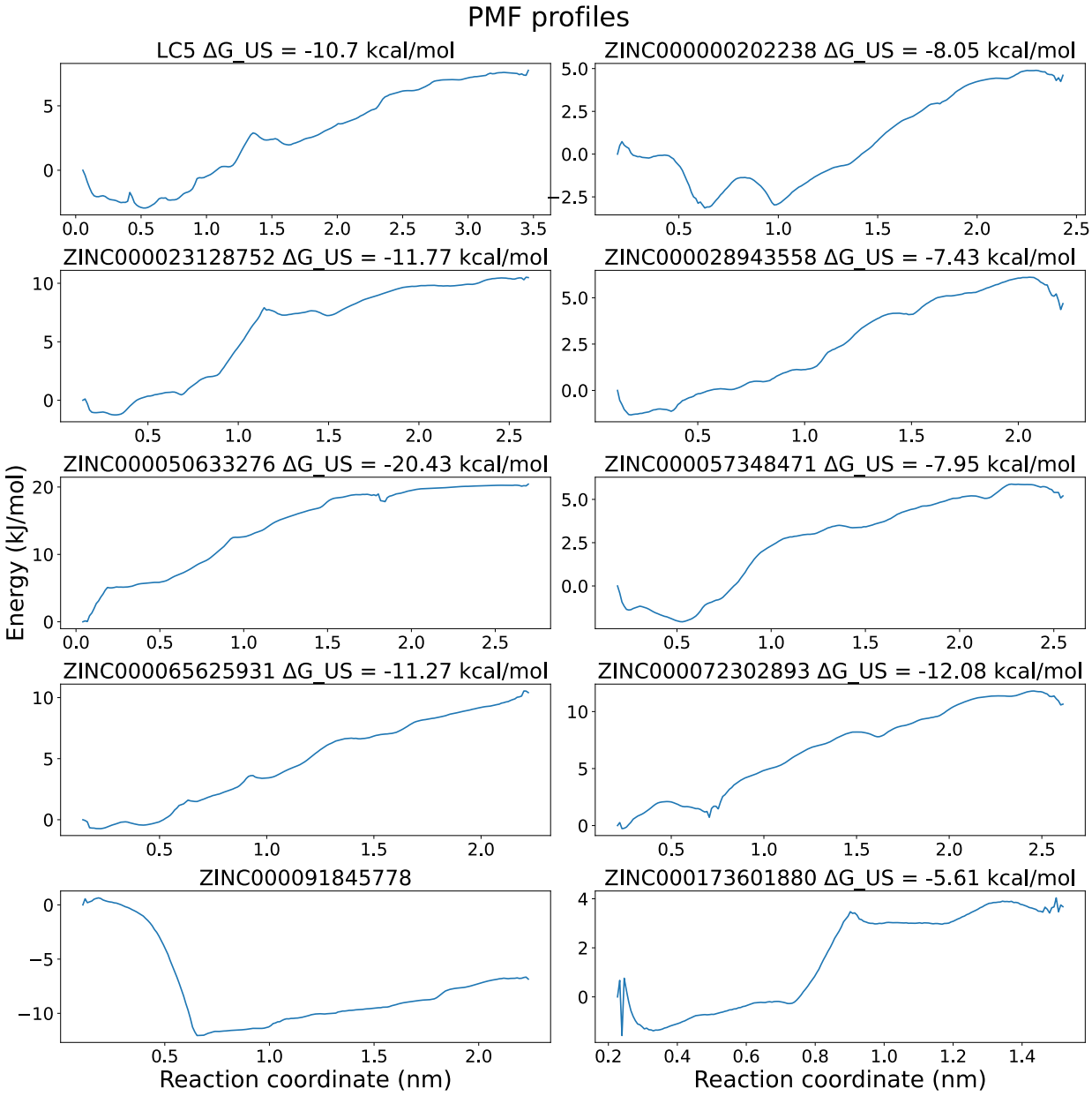


Figure 3-15 PMF curves obtained from WHAM analysis for the different systems. The related histograms for the different systems are presented in Appendix G. The x-axis is the reaction coordinate (protein-ligand COM displacement) while the Y one represents the potential energy. The figure was generated with matplotlib [325] and pandas [304].

The calculated binding free energy from the US simulation (ΔG_{US}) can be determined as the difference between the largest and smallest values of PMF. The free energy (y-axis) varied across the reaction coordinate (x-axis) which characterizes the protein-ligand dissociation. Calculated Binding free energy values obtained using the US method showed to be in good correlation with the experimental data [313,340].

The PMF curves had similar trends: an early decrease to a minimum value after starting at zero for the free energy. This is an unexcepted event as all systems were first minimized and equilibrated. It finally increases to reach a stable value at around 2nm along ξ, where the ligand was fully solvated. This early decrease in the energy value has also been observed in a similar study applying US to protein-ligand systems [313]. In general, systems showed a smooth increase in energy which can be linked to the choice of a slow pulling velocity. However, ZINC000173601880 and ZINC000023128752 presented a steady increase.

LC5 ΔG_{US} was -10.7 kcal/mol. The compound has an inhibitory constant of 280 nM [256] which corresponds after conversion using (1-5) to a binding free energy of -9 Kcal/mol. Hence the predicted value here showed an overestimation of 1.7 kcal/mol. In MM-PBSA, the binding free energy was - 63.99 kJ/mol (-15.29 kcal/mol) likewise much lower than the experimental value. The exact binding free energy was overestimated in both approaches. However, since interest is in the relative activities of compounds, it is noted that the overestimation may be systematic across the different systems.

$$\Delta G = RT \ln(K_i) \quad (3\text{-}9)$$

Ki (Inhibitory constant)

R (Gaz constant) = 1.98 and

T (temperature) = 298.15 Kelvin

Four other compounds: ZINC000050633276 (-20.43 kcal/mol), ZINC000072302893 (-12.07 kcal/mol), ZINC000023128752 (-11.77 kcal/mol), ZINC000065625931 (-11.27 kcal/mol) (Figure 3-16) showed better binding free energy than LC5. As this latter has an inhibitory constant in the nanomolar range (280nM), these compounds may have similar potency against DXR. ZINC000050633276 showed a notable ΔG_{US} of -20 kcal/mol corresponding to Ki of 1.934 fM. Hence the compound may bind strongly. Indeed, its binding pose showed numerous polar contacts (Figure 3-17). However, the compound was not the top performer in MM-PBSA. Further, in this case of ZINC000050633276, MN had an unfavorable contribution of 14 kJ/mol. Optimizing its interaction with the metal or merging it with PfDXR based inhibitors scaffolds may yield potent inhibitors. These four hits present scaffolds very different from fosmidomycin. None of the compounds, for instance, contain the hydroxamate or the phosphonate groups.

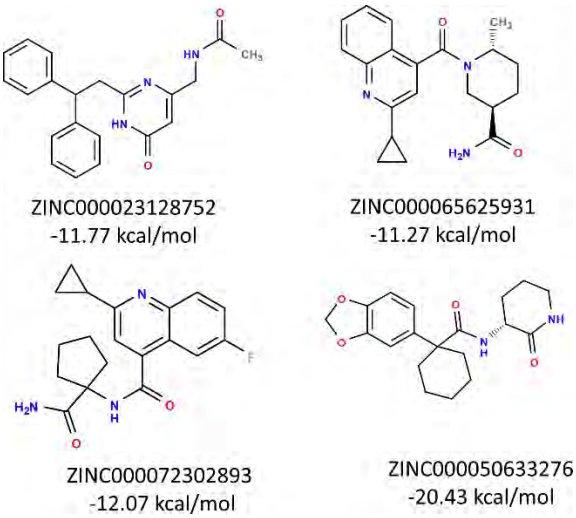


Figure 3-16 2D depictions of the four hits. Structures were depicted using Open Babel [299]

ZINC000091845778 shows an unusual trend of PMF for protein-ligand systems. Indeed, the curve shows a sharp decrease of free energy in the early part of the simulation. This was not expected, as the first conformation was supposed to be the most energetically favorable one. The potential energy decreased sharply to -12 kcal/mol. The next phase follows a classic PMF profile of protein-ligand unbinding. ZINC000091845778 umbrella histogram analysis showed a lack of sampling in that region of the reaction coordinate (Appendix G). Hence, the compound binding energy was not included in the hit selection.

With ZINC000065625931, the potential energy does not seem to have plateaued. The end of the reaction coordinate was determined based on ligand complete solvation. This latter was determined based on visualization and using the SMD curves. This was done to save computational cost. A significant increase in the potential energy is not expected and the ligand current binding energy is better than the co-crystallized one.

Ligand binding to PfDXR has an induced-fit mechanism to accommodate the ligand [56]. As ligand unbinding may require conformational rearrangement, a slow pulling rate was, therefore, ideal to allow active residues time to adopt favorable conformations.

The free energy of binding can also be obtained using Jarzynski's Equality from multiple independent SMD simulations. The method showed comparable efficiency to US [341].

As conclusion, a set of uncorrelated features may be a better approach to consensus scoring. One can drop one of two SFs showing a correlation of 0.5 or greater. Each SF was expected to contribute to a more accurate ligand rank ordering. However, ideally, one would test against a background truth the accuracy of different combinations of SFs. Using an experimental reference would have been ideal for evaluating the different methods but the most suited approach for integration. In this study, for example, each LBVS method could have been evaluated based on its ability to rank accurately known inhibitors according to their potency and later integration of the different methods could have been evaluated the same way. At the SBVS level, the inhibitors receptor conformation can be used as well, although this will significantly increase the

dimensionality of the data and the computational cost. Alternative approaches are flexible residues docking or induced-fit docking with higher computation cost though. An induced-fit docking approach may have provided a better approach to modelling.

About the cost-accuracy trade-off, the SMD approach is attractive. It has the lowest runtime (2 ns) compared to MD, MM-PBSA and US. It provides Fmax and Wpull as estimation of affinity. Moreover, it provides analytical insight into residues behavior along binding/unbinding path. About 20 windows were produced for the different ligands for US. The ten nanoseconds of sampling in each window result in a 100 times higher computational cost than the 2 ns in SMD. The potential accuracy gain may not be worth this cost. A limitation in the current chapter is the lack of experimental data for predicted affinities. Besides LC5, known inhibitors used in LBVS could have guided on the accuracy of the current methods.

Some umbrella histograms showed a lack of sampling in some regions (Appendix G). For example, this was the case with the two first windows in ZINC000000202238. In those cases, an attempt to concentrate more windows in that region did improve the profile as windows were shifting away from those regions. These regions may present a high-kinetic barrier or high-energy state. LC5 umbrella histograms had a good overlap of the reaction coordinate. This may be explained by the choice of the pulling direction based on its bound conformation. That direction determined according to CAVER [331] provides the most favorable unbinding path, thus avoiding high-kinetic or energetic barriers.

Overall, four compounds (ZINC000050633276, ZINC000072302893, ZINC000065625931, ZINC000023128752) had a higher affinity than LC5 in US.

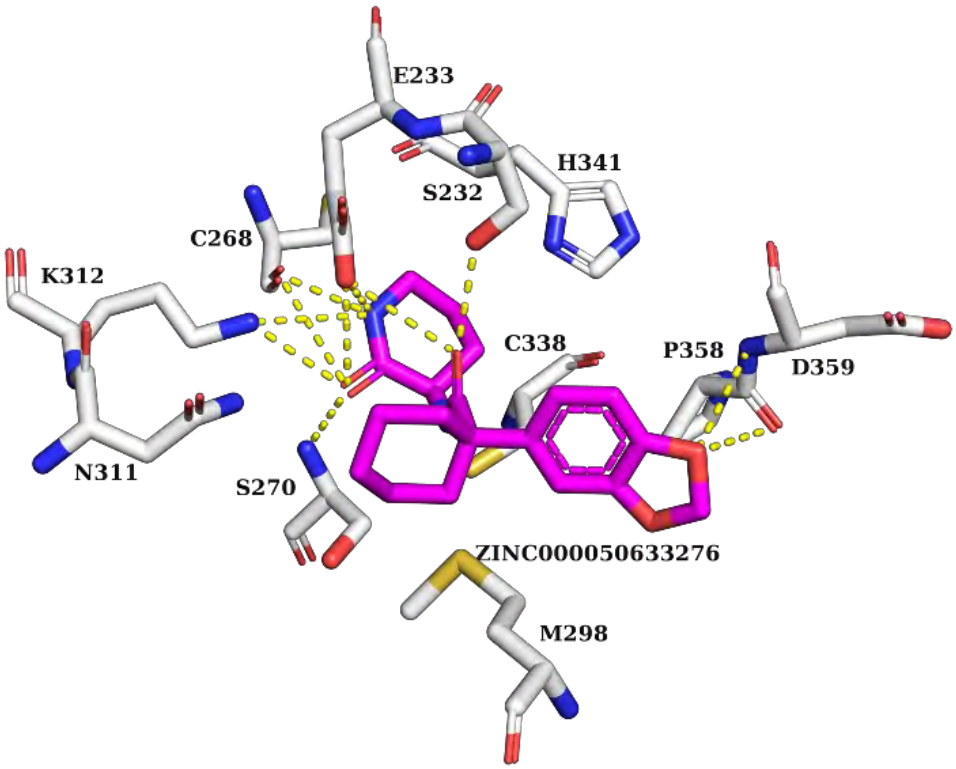


Figure 3-17 ZINC000050633276 (magenta) docked pose in PfDXR active site. ZINC000050633276 interacting residues are indicated in light grey. Protein residues in a radius of 3.5 Ångströms of the ligand are labelled with their one-letter code and their residue numbers, displayed in stick and colored atom types (other elements) and white (carbon). Polar contacts with the ligand are displayed in dashed lines in yellow. The figure was generated using Pymol [225] and the show_contacts script [226].

3.3.7 Additional discussions

3.3.7.1 On finding SFs, documentation, running parameters

In this work, we searched for freely available SF. A key remark for some SF was their lack of proper documentation. Some SFs related publications had broken links to the respective tool. Hence, efforts in developing these tools may be lost and the tools may disappear. USRCAT [266] and NNscore [342] original publications had broken links (http://hg.adrianschreyer.eu/usrcat and http://www.nbcr.net/software/nnscore/) to the tools. Moreover, we found a lack of documentation on parameters in some tools. In future work, a fully documented repository of SFs with examples can be constructed. ODDT toolkit was helpful as it already collected a set of SFs and offer rescoring functionality. We unsuccessfully attempted to integrate more SFs to the pipeline due to the lack of documentation or compatibility. SFs used in the current study can be integrated in ODDT [301]. DockBox [343] is a similar initiative focusing on molecular docking tools. Each SF may have a particular contribution to an accurate binding affinity estimation. Hence, as emphasized here with the consensus approach or the wisdom of the crowd, affinity estimation will benefit from every SF. Moreover, as discussed earlier a strategic combination of SFs can contribute to building a customized SFs depending on the target. Hence a future approach could automatically build a customized SF depending on available experimental data, ligand molecular properties and the strategic integration of docked scores generated from ligand having experimental values. Tailor-made SF can be helpful as SFs accuracy also tends to be inconsistent across different targets [81].

3.4 Conclusion

In this chapter, the ZINC lead-like subset was screened for potential PfDXR inhibitors using a consensus LVBS and SBVS approach including MD, MM-PBSA, SMD, and US. Four hits$ki_molar = ki_nM/1000000000$ outperformed LC5 in US, a 280 nM potent PfDXR inhibitor. US is the most advanced free energy prediction method [103] used here. ZINC000050633276 showed a promising -20.43 kcal/mol as binding free energy corresponding to Ki of 1.934 fM. The top identified hits were associated with higher electrostatic interaction contributions than vdW interaction ones in the PLIE.

GLU233, CYS268, SER270, TRP296, and HIS341 had a significant contribution to binding free energy in MM-PBSA and their breaking in SMD was also associated with higher values of Fmax.

Comparing the methods ranking correlation, some remained uncorrelated in both LVBS and SBVS, while others agreed. In LBVS, two main clusters (ES, USR, USRCAT, OBSPEC) and (MHFP, RDKit_3dpharm) are noted while in SBVS (Vina, Idock, and Smina), (AutoDock, DSX, Cyscore, Xscore) and the Rf-score group (Rf-score_V1 to V4) formed distinct clusters.

Comparing the current approaches to US, Wpull had the highest correlation. This is supported by a previous finding and can be explained by its link to the binding free energy through the isobaric-isothermal Jarzynski's equality [308]. We thus recommend SMD to validate docking hits especially in computational resources limited settings. Indeed, the approach was less expensive than conventional MD and MM-PBSA.

As future work, the activity of these hits identified by *in silico* experiments may be confirmed *in vitro*. Elsewhere, combining HMR [118], and SMD can result in a considerable gain in speed while still maintaining accuracy. A 1-2 ns SMD joint with HMR (using a timestep 4 fs) on a structure of about 500 residues on a 24 cores (Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz) machine for about 2 hours. Moreover, the methods used here can be combined with the *P. falciparum* targets screening in Chapter 2: . The table of protein ligands can be rescored using the different scoring functions used here. This can be associated with a preselection of a suited screening library using the LBVS methods. A custom library can be made of the union of targets' known inhibitors analogs.

Facing the limits of the current state of art approach SF [164] and the high computational cost associated with more advanced approaches [278], the wisdom of the crowd can be a better alternative in high throughput virtual screening [344]. These current SFs can be integrated into tools such as ODDT [301] and/or VirtualFlow [345]. Finally, different SFs can be used to generate features for ML models.

Chapter 4: SANCDB: An Update On South African Natural Compounds And Their Readily Available Analogs

4.1 Introduction

Throughout history, mankind used nature as a source of food, cosmetics, pesticides, and medicines [70]. Plant usage as medicines may date back up to Neanderthals [346]. Even now humans still harness many benefits from natural sources [70]. A substance produced by a living organism: animals, plants, or microorganisms (bacteria, algae, fungi) are defined as natural products (NPs) [347]. Currently, these NPs represent up to 35% of medicines [348]. In the decade 2010-2019, Newman *et al.* estimate that 25% to 33% of approved small molecules are from NPs [349].

This importance of NPs in drug discovery is rooted in their chemical structure. NP scaffolds have been optimally shaped for living systems, and selectivity for biological targets has been driven by natural selection over millions of years of evolution [350]. In this survival of fittest, organisms have developed defense mechanisms, including chemical compounds as antibiotics. For instance, penicillin is one of the most successful antibiotics used by humans, but it comes from fungi, where its purpose is for its defense [350]. Moreover, NPs cover a larger area of chemical space and are more structurally diverse than synthetic compounds [351].

Given NPs' contribution to drug discovery and the justification of this in their chemistry, many NPs' data repositories have emerged [347]. These repositories often cover different geographic areas. The South African National Compound Database (SANCDB) contains chemical structures of NPs isolated in South Africa. Other information such as their sources, bioactivities, structures' properties classification, literature references are also available [65]. Beyond storing and archiving information, it has been used in diverse works, counting up to 50 citations in Google scholar. Figure 4-1 shows SANCDB yearly citations counts and other similar NP databases. These studies are related to cheminformatics, machine learning [352–355], and virtual screening [356–362].

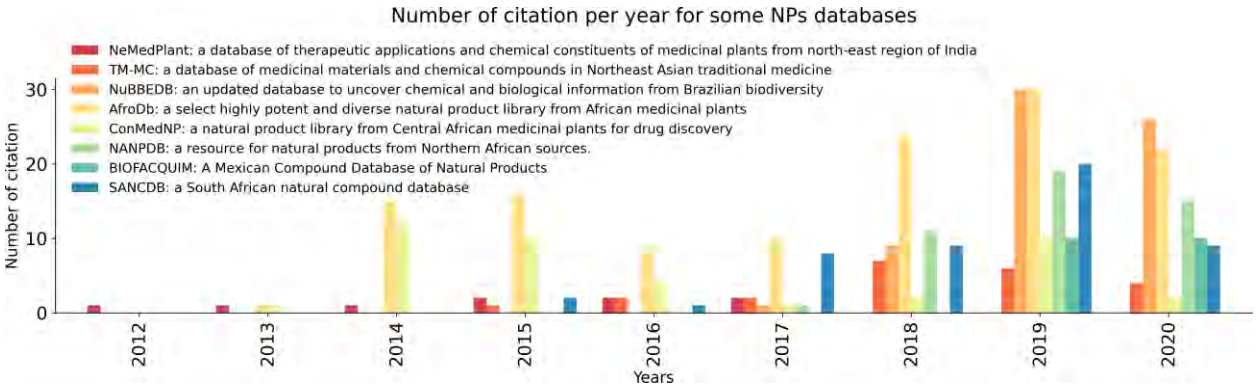


Figure 4-1 Yearly citation distributions for SANCDB and some regional NP databases articles similar to SANCDB. Databases may have been introduced at different times.

Hits identified in virtual screening are often confirmed *in vitro*. The physical unavailability of SANCDB compounds poses a challenge. In a past study [363], SANCDB hits were not commercially available for further *in vitro* tests. Five authors corresponding to five compounds, were contacted for its availability. Out of five, two authors replied that the compounds are not readily available. Similarly, 20(29)-Lupene-3β-isoferulate (SANC00518) a potential allosteric modulator human Hsp90α [362], Gordonoside A (SANC00456), for *Plasmodium falciparum* Prolyl tRNA synthetase modulation [360], and discorhabdin N (SANC00132), for Hsp72 and Hsc70 [364] have also been identified from SANCDB. Yet none of these compounds are currently commercially available. Hence, it remains challenging to confirm these predicted activities *in vitro*. NPs' unavailability is not specific to SANCDB. For example, only 10% of the ZINC NP subset are readily purchasable compounds [351]. Besides the unavailability problem for compounds, they have costly isolation methods [348,365] and or have complex synthetic routes because of their scaffold complexity [366]. Hence, NPs' scaffold complexity is an attractive chemistry on one side but is difficult of access on the other. This reverse of the medal [366] may hamper their full exploitation.

The initial 600 NPs in SANCDB can be augmented. By comparison, Brazil counts species around 170,000 to 210,000 species [367] and the NuBBEDB database counts 2147 compounds [368]. South Africa is the third most biodiverse nation with over 100,000 known organisms [369–371]. This biodiversity contributes to compound diversity [372] and one aim in virtual screening is to identify new scaffolds [373]. More recent years have been prolific in NP research in the region. For example, oligosaccharides, flavonoids, proanthocyanidins, quinic acid derivatives, and ellagitannins were isolated from *Myrothamnus flabellifolia Welw* in 2016 [374]. A further eleven compounds were isolated from *Aspalathus linearis* by Fantoukh *et al.* [375]. In 2019, Awolola *et al.* reported four compounds from the genus *Ficus* [376]. These isolations happened after SANCDB establishment in 2015 [65]. Hence, the database update should follow the same trend as compound isolations. SANCDB website has an automatic deposition pipeline that has been underutilized by natural product chemists. So far, only a few compounds have been deposited by such researchers. Given that the initial set of 600 compounds likely under-represents the country's potential in NPs, a database update is required.

Further, large libraries are ideal for virtual screening. The larger the library, the more likely it is that virtual screening will find more potent and diverse scaffolds [100].

A common problem in NPs' resources is the lack of update maintenance. Although some databases are regularly maintained and or are recently updated [368,377], a broad problem across many other NP databases is the lack of updates, maintenance and accessibility. For example, some published NP repositories have broken website links, which may be due to loss of data, and will certainly cause a loss of accessibility of data [347].

As an alternative solution to the compound availability problem, we propose exploring compound analogs, since similar compounds will have similar properties [78,257]. Analogs, more than being an alternative to the availability of NPs, may in themselves be more potent in the context of drug targets. Hence, more than simply maintaining the NP compound activity, analogs can further optimize this. For instance, quinine and artemisinin have been used as starting points for more potent antimalarials [378,379].

Motivated by the above-mentioned problems, the current research aims to further add more compounds to the database, aiming to reach a thousand compounds in SANCDB and to further include commercially available analogs for all compounds. The current work can therefore be separated into two parts; the first is the database update with new NPs, the second is to make available readily available commercial analogs associated with each of the NPs. The term "new NPs" is used relative to the already present NPs in the database. An additional cheminformatic analysis related to the compound scaffolds and different drug discovery relevant subsets is performed in the context of the whole updated SANCDB.

4.2 Methods

4.2.1 Compounds update

In this section, the methods used in this work are explained in more detail.

Most regional compound databases were established through a literature search, for example for NuBBE [380], Database@Taiwan [381] and BIOFACQUIM [377]. Isolated compounds are often published in the literature, and in this context similar methods were used. The process is described in the original SANCDB paper [65] searched literature using keywords ("isolate", "South Africa", "natural product"). The classic search through keywords in search engines has limits. A general remark was that about half of the search results through these keywords did not return the expected relevant results or produced results that were redundant. To overcome this, we additionally focused on the current authors' list of references in the database. Most compounds isolations are done by the same research groups and investigators. For example, Davies-Coleman is an authority in South African NPs research and is the author of 115 NPs in the current set of compounds. Hence besides the keyword search, all publications associated with all authors and co-authors using the reference Digital Object Identifier (DOI) were retrieved through the Scopus API [382]. This allowed for programmatic access to all scholarly databases indexed by Scopus [382] for collections, parsing, and extraction of organized literature references. Redundant publications and articles in which any author affiliation was not from South Africa were removed. Further, the current references in the database were excluded. Through this strategy, we found a set of references with a high probability for NPs isolation in South African.

One major drawback of the current and also with the previously used method was the required but time-consuming step of reading and accurately identifying compound isolation in the text. The manual and the most intense stage was then to examine and validate compounds' South African origin. More structure retrieval from publication figures and ensuring structure accuracy were also time-consuming and error-prone. Yet, this is a critical step as the structure is the key information in the database and its accuracy is of utmost importance. For instance, identified hit structures in a virtual screening experiment must agree with the one in the database.

From the above search, each compound's Chemical Abstracts Service (CAS) [383] number was obtained using SciFinder [383]. Every compound was identified using this unique CAS number. All the above was done through a semi-automated process using Selenium [384]. Selenium automates some browser actions such as filling forms with given information from a table. Selenium can map a spreadsheet column to specific fields in web form for content transfer [384]. This saves the time-consuming and error-prone steps of copy-pasting or the manual forms filling. Another alternative would be to link the MySQL table to the Excel sheets or to the CSV, which once filled may then directly update the database.

From the compound's CAS IDs and source species information, the remaining information for the database could be automatically obtained. From the CAS identifier, PubChemPy [385], PubChem API [235], Chemical Identifier Resolver (CIR) [386] solved the IDs for different databases (ChEMBL [387], DrugBank [213], ZINC [388], PubChem [235]) and compounds molecular properties.

New compounds' structures were prepared with OpenBabel [299] and minimized using GAMESS at RM1 level of theory [389]. ClassyFire [390] classified compounds and Pygbif [391], a python client for the Global Biodiversity Information Center (GBIF) [392] API linked sources organisms to their kingdoms, families, and genera. AutoDock pdbqt formats were prepared with the Autodock Tools prepare_ligand4.py script, and Schrödinger Maestro formats [393] were added to the database besides the already available MOL2, PDB, SMILES, and SDF formats. A single SMILES was made available for each compound. The entire set in SDF file containing 3D structures and related compounds information, IDs, source, PubChem ID... was made available for download. Additionally, structures' depictions were updated: adding stereochemistry to those lacking and all aromatic rings were depicted in their "Kekulé" forms.

4.2.2 Commercially available analogs

MolPort (October 2019) [394], Mcule (October 2019) [395] and SciFinder [383] were searched for SANCDB compounds analogs. Different methods exist to quantify the similarity between two chemical structures. Some related methods have been explained Chapter 3: . Here we used the Open Babel FP2 fingerprint together with the Tanimoto similarity. Open Babel is a popular cheminformatic tool with over 4000 citations in Google Scholar [299]. This eases reproducibility and is a more sustainable solution, as no dependencies are required. This is also an ideal solution with respect to the context of API integration. Hence, future updates of the database will be able to make use of these tools to reproduce the approach. FP2 is a path-based fingerprint that indexes linear fragments in molecules up to 7 atoms. With the set of indexed fragments, a hash number from 0 to 1020 is used to set a bit in a 1024-bit vector. From the set of fingerprint bits for two compounds

A and B, the similarity score was computed using equation (4-1) [259]. A similarity threshold of > 0.6 classified compounds as analogs.

$$\text{Tanimoto coefficient}_{A,B} = \frac{AB}{A + B - AB} \qquad (4\text{-}1)$$

Analogs are thus now available for download on the SANCDB website [396] and a link to each analog in Mcule and Molport is provided. The web interface displays the analogs and their similarity scores to their respective SANCDB molecules. An automated periodic update of analogs through Mcule and MolPort APIs was setup. This automation and inclusion within the web interface were facilitated by Michael Glenister [397]. Analogs' structures were prepared with OpenBabel [299] and resulting geometries minimized in RDKit[162] under the Merck Molecular Force Field (MMFF94) [398]. This particular aspect of work was done by Thommas Musyoka [399].

4.2.3 Cheminformatic analysis

This analysis assesses the database potential for drug discovery. Compound drug-likeness and scaffolds were calculated. In addition, an analysis of the updated SANCDB compound chemical space coverage by commercially available analogs was conducted.

SANCDB chemical space coverage may help identify patterns particularly with regard to structures without analogs. These structures may exist in a hard to synthesize region of SANCDB chemical space. To evaluate the SANCDB chemical space coverage by analogs, Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) were conducted. While PCA has been commonly used for dimension reduction and chemical space analysis, t-SNE has been recently proposed as a more efficient approach for clustering chemical compounds [400–402]. The implementation of this in scikit-learn was used [403,404]. T-SNE measures the similarity of two data points in the low-dimensional space through a variant of Stochastic Neighbor Embedding using a Student-t distribution [405]. In t-SNE, a perplexity of 50 and a learning rate of 100 were used while keeping the rest of the parameters to default. For this, the compounds' non-normalized Molecular Quantum Numbers (MQN) descriptors [406] were computed using RDKit [407].

Virtual screening aims to find new scaffolds [373], making scaffold diversity an ideal characteristic for screening databases. The Bemis-Murcko scaffold decomposition is commonly used to assess database scaffold diversity [351,377,408–412]. This approach is used here using the Scopy package [413]. Molecule clouds have been recently proposed for visualization of compound database scaffolds [413,414], and these are used here.

Compound subsets are defined by thresholds on molecular properties. Screening libraries such as ZINC [415] are often subdivided into subsets, which may fit different drug discovery project scenarios. For instance, screening for protein-protein interaction inhibitors may benefit from a library rich in PPI-like inhibitors. Fragments may fit early-stage drug discovery in the identification of potent chemotypes for later optimization. They may also be useful for merging strategies to more potent leads [416]. Frequent hitters and toxicophores may be avoided by using a PAINS-free

library [417]. Hence subsets can be helpful to set up good quality libraries. Here, we use the following definitions of subsets (Table 4-1): Drug-like, Extended drug-like, Lead-like, Fragment-like and PPI-like [418]. The used molecular properties are the MW, logP, Number of hydrogen bond acceptor (nHA), nHD (Number of hydrogen bond donor), TPSA and the number of rings (nRing).

Table 4-1 Molecular properties conditions for subsets

Subsets	**Conditions**
Lead-like	MW ≥ 250 & MW ≤ 350 & nRot ≤ 7 & logP ≤ 3.5
Extended drug-like	Druglike & nRot ≤7 & TPSA < 150
Drug-like	MW ≤ 500 & MW ≥ 150 & logP ≤ 5 & nHD ≤ 5 & nHA ≤10
PPI-like	nRing ≥ 4 & MW > 400 & nHA > 4 & logP > 4
Fragment-like	nHA ≥ 3 & MW ≤ 300 & nHD ≤ 3 & logP ≤ 3

All data analysis and plots were done in a Jupyter notebook environment [419] using python packages Pandas [420], Pandas-profiling [421], Matplotlib [325] and Seaborn [198].

4.3 Results – Discussions

This section outlines the results of the current set of 1012 compounds and their related attributes (sources, classes, biological activities). A second part analyses analogs and explores drug discovery related metrics such as compound scaffold subsets. During this work, 288 new compounds were added, for this total 1012. All the current analyses are done on the full set of 1012 compounds. SANCDB had been continuously updated since the initial set of 600 compounds

[65] and contained 716 before this work. Here we present an analysis of the database content (sources, compound classes, and activities).

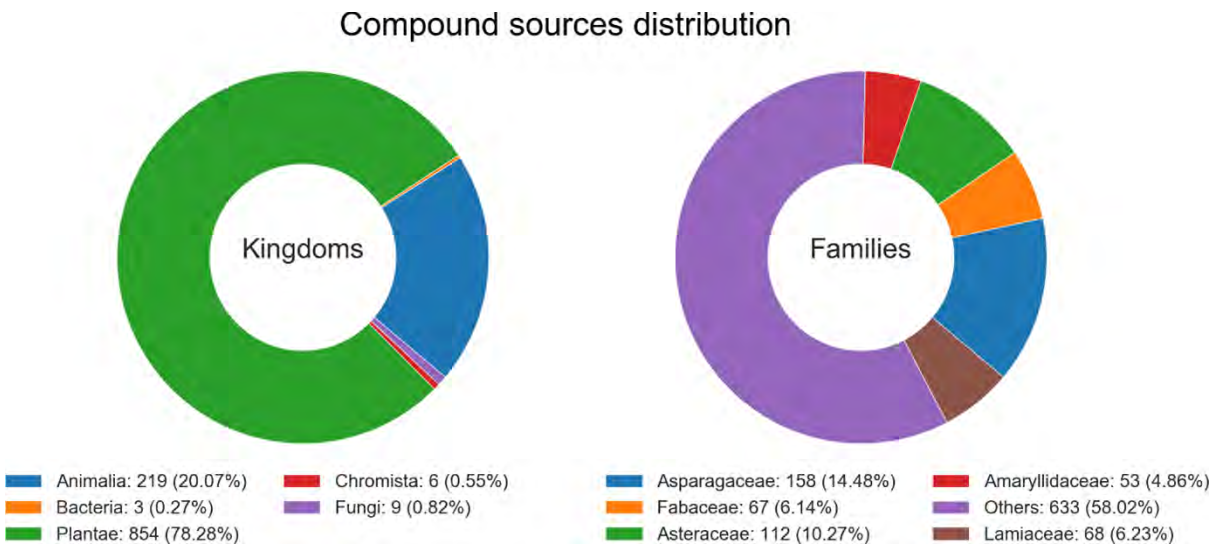


Figure 4-2 Compound sources distribution. Sources' species were mapped to their kingdom, families and genera using pygbif [391]

The database content was derived from 321 distinct sources. As a compound can be derived from multiple sources, there was a higher number of sources than the number of compounds. The distribution of compound sources with respect to biological kingdoms and families is presented in Figure 4-2. Plants were the main source in that 78.3% (854 of compounds) were isolated from plants. They are the major sources of compounds in many NPs databases [368] and some databases only focus on plants [373,422,423] probably linked to their predominance as NPs sources. Animals, fungi, chromista, and bacteria followed with 219 (20.1%), 9 (0.8%), 6 (0.5%), and 3 compounds (0.3%) respectively. Bacteria had the lowest proportion. All three compounds isolated from bacteria were isolated from *Streptomyces sp.* However, generally, bacteria are major sources of potent antimicrobials and NPs [351]. Similarly, in SANCDB only four fungi sources were recorded: *Clathrina aff reticulum*, *Eurotium rubrum*, *Termitomyces microcarpus* and *Fusarium proliferatum*. The low proportion of microbial sources may show an under-exploration of their potential in NPs in South Africa. The NuBBE database also showed a comparable source distribution to that presented in this study [368]. This may also be explained by plants' larger and more documented uses in traditional medicines and easier accessibility. Concerning compound source families, Asparagaceae (158 – 14.48%), Asteraceae (112 – 10.27%), Lamiaceae (68 – 6.23%), Fabaceae (67 – 6.14%) and Amaryllidaceae (53 – 4.86%) were the most frequent families. Top genera were Ornithogalum Senecio, Eucomis, Salvia, and Plocamium with the following number of compounds and proportions (62 – 5.7%), (58 – 5.3%), (39 – 3.6%), (39 – 3.6%) and (38 – 3.5%) respectively.

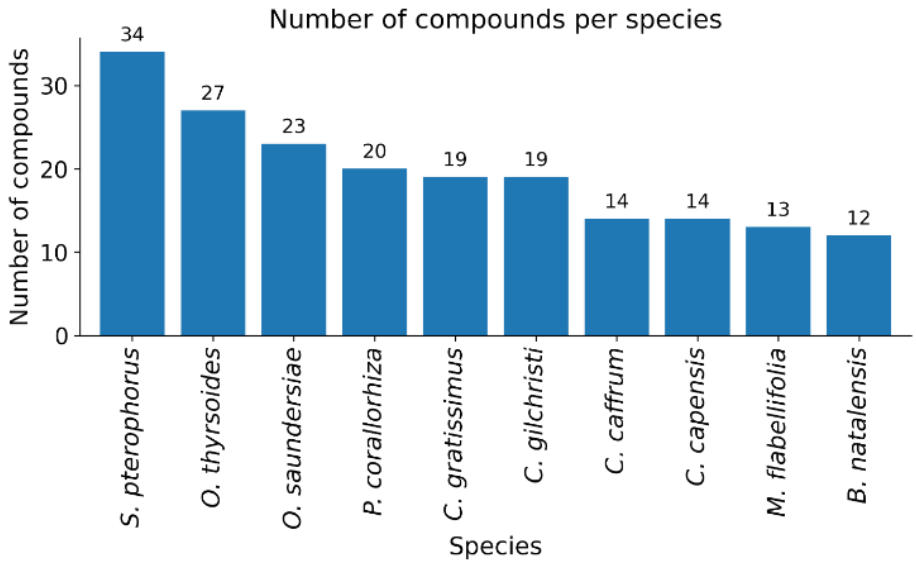


Figure 4-3 Top 10 species, producing the highest numbers of NPs in SANCDB.

The frequency of the source species was also analyzed to identify most prolific species. Figure 4-3 shows the top ten most frequent source species found in SANCDB. The Asteraceae, *Senecio pterophorus* was the most productive with 34 compounds (3.1%). It produces macrocyclic diester pyrrolizidine alkaloids which showed teratogenic, genotoxic, hepatotoxic, and carcinogenic activity [424]. Next, 26 (2.4%) and 23 (2.1%) compounds were isolated from two Asparagaceae: *Ornithogalum thyrsoides* and *Ornithogalum saundersiae,* respectively. The thyrsoides are cytotoxic against HL-60 human promyelocytic leukemia cells. *Ornithogalum thyrsoides* is abundant in the Western Cape region of South Africa [425,426]. *Ornithogalum saundersiae* is an ornamental flower, toxic for cattle, found in Swaziland, Mpumalanga, and KwaZulu-Natal regions [427,428] . Twenty compounds (1.8%) were isolated from *Plocamium corallorhiza*, a red algae of the Plocamiaceae family yielding halogenated monoterpenes [429,430] . The fifth most prolific source was the tubeworm [431], *Cephalodiscus gilchristi* with 19 compounds (1.7%). It produced cephalostatin 1, a potent cell growth inhibitor, and alkaloids active against lymphocytic leukemia [432].

A common characteristic among these sources was that they are naturally widespread in South Africa. This eases their accessibility for research and explains the high numbers of the compounds that have been identified from these sources. Source productivity information might guide conservation strategies toward prioritizing these main producers and their underlying dependencies. Currently, these prolific sources are not on the list of endangered species of the South African National Biodiversity Institute (SANBI) [433].

4.3.1 Compounds classification

Eleven superclasses, 79 classes, and 124 subclasses of compound were found using ClassyFire. SANCDB covered about 50% of all superclasses (26) available in the tool. Hence, we noted compounds' diversity across the different levels of classification. One can thus expect a range of diversified biological activities. However, SANCDB compounds only covered 77 (10%) of the 764

classes available in ClassyFire. This may be linked to the vastness of the chemical space and/or the crowdedness of the SANCDB one. One should note though that the tool also covers inorganic compounds.

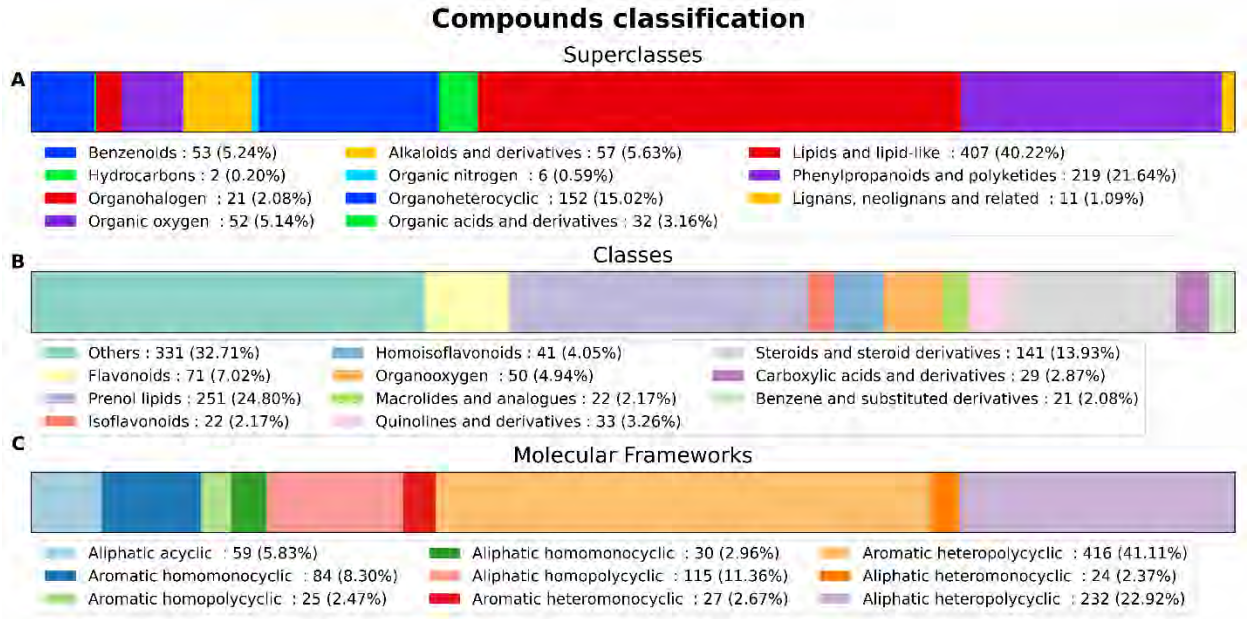


Figure 4-4 Stacked bar charts of the compound classifications. **A)** SANCDB compounds superclasses. **B)** SANCDB compounds classes. **C)** SANCDB molecular frameworks. Classifications were obtained from ClassyFire.

The most frequent compound classes were the phenol lipids, the steroids and steroid derivatives, the flavonoids, the organooxygen compounds, and the homoisoflavonoids counting for 24.8%, 13.9%, 7%, 4.9%, and 4.1% of the content respectively (Figure 4-4b). Regarding the compounds' molecular framework, polycyclic compounds were the most frequent (Figure 4-4c). SANCDB is rich in cyclic compounds. The aromatic heteropolycyclic count 41.11% of the database (416 compounds). The molecular framework distribution showed that only 59 (5.8%) of the compounds were acyclic. Nine distinct molecular frameworks were found (Figure 4-4c). The most common frameworks were the aromatic heteropolycyclic, the aliphatic heteropolycyclic, and the aliphatic homopolycyclic, counting for 41.1%, 22.9%, and 11.4% respectively. The molecular cloud showed many large compounds, including macrocycles (Figure 4-12).

NPs classes distribution in SANCDB was similar to other databases in the literature [368,423,434]. For instance, SANCDB and the Integrated Ethiopian Traditional Herbal Medicine and Phytochemicals Database (ETM-DB) databases have the same top three superclasses (the lipids and lipid-like, the phenylpropanoids and polyketides, and the organoheterocyclic). However, ETM-DB showed greater diversity with 22 superclasses and 200 classes for 3,930 compounds. Its compound classification was also done using ClassyFire [390]. Bioassays, Ecophysiology, and Biosynthesis of Natural Products Database (NuBBE) and the 500 Pan-African Natural Products Library (p-ANAPL) have 14 and 30 classes of compounds respectively [368,435]. It is noteworthy that these classifications were done using a different scheme.

Compound classification helps cluster compounds to illustrate their diversity. This may guide molecular optimization by deriving libraries from a specific class with known biological activity. Compound classification is a domain expert task requiring human intervention. Different classification systems exist and help assess diversity [368,436]. For instance, Dewick's biosynthesis theory classifies according to the compound's synthetical origin. There are also structure-based classifications and those based on biological activities [368,390]. Djoumbou Feunang *et al.* introduced ClasssyFire, an automated classification tool [390]. This allows a faster and standardized classification without human intervention. The approach is also reproducible for future updates. The tool classifies into subclasses superclasses, classes, and kingdoms using structural patterns [390]. Other classification levels are available: compounds molecular frameworks, parents, and substituents. A compound's ring count, aliphatic or aromatic nature, and atom types characterize its molecular framework [390]. The concept resembles the scaffold one [411] and evaluates compound databases diversity for screening [437].

More NPs may be available, especially beyond published literature. A limit in the current study and the earlier one is the lack of search in theses. For example, MSc and Ph.D. theses were used for building the CamMedNP database [373]. These resources may in the future extensively extend SANCDB content. Another important consideration is that SANCDB is a single group research effort. This may not be sufficient, especially when considering the volume and the potential additional curation required for non-published literature. Collaboration with an institution focusing on NPs research such as SANBI [369] may be fruitful. For instance, NuBBe is a collaborative effort with other institutions such as the National Council for Scientific and Technological Development, which significantly contributed to the literature search [368].

4.3.1.1 A more efficient text mining strategy

Several strategies can speed up database construction and improve their maintenance. Most current NP databases are built through literature parsing. Automated text mining tools can ease human effort in collecting chemical data from the literature. Examples of such tools are Molminer, ChemEx, ChemicalTagger, ChemDataExtractor [438]. SureChEMBL is a successful example of using an automated parsing pipeline without human curation to collect bioactivity data from patents [439]. Advances in natural language processing can be used to link compounds to their sources. More e-alerts systems with NP research-related journals can help in terms of a continuous update. Authors isolating NPs should also be aware of the already available deposition system. Indeed, we noticed a lack of data deposition from the authors themselves. Authors referenced in the database and others involved in NP research can be invited to deposit unpublished data. This unpublished nature can be indicated on the related records.

4.3.2 Compounds classes and sources relationships

The relationships between sources and type of produced molecules were analysed. This can be done at different levels using classes or superclass of the produced molecules and the source families. The heatmap in Figure 4-5 shows the relationship between source family and compounds superclass. The relations between source family and chemical classes were also analysed.

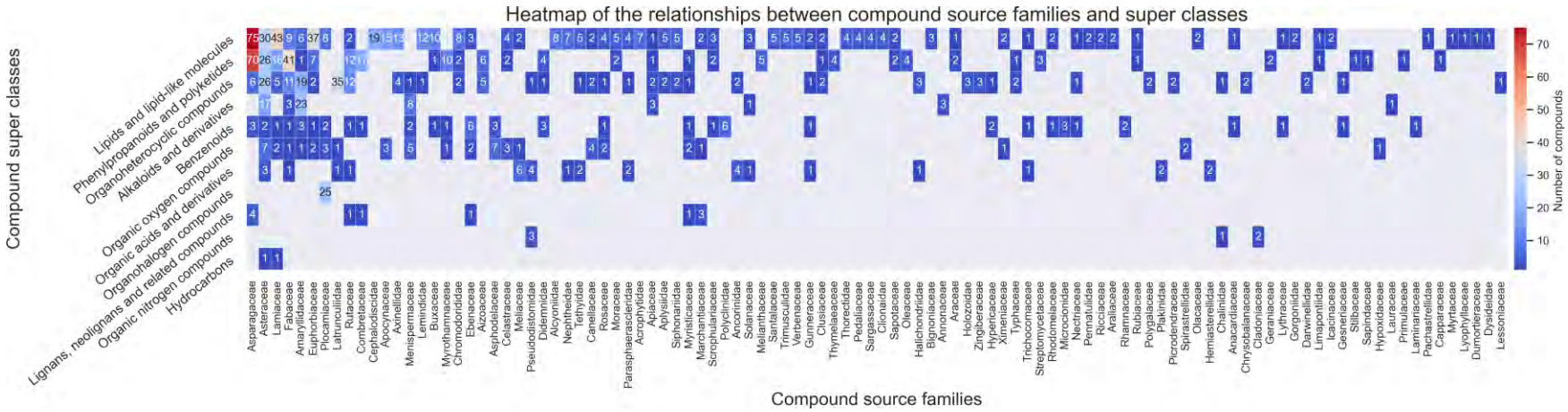


Figure 4-5 Heatmap of the occurrence of Classyfire superclasses (y-axis) and the source family (x-axis). For visualization, only superclasses are displayed. There were over 70 classes. A Fisher's exact test was performed to test compounds classes distribution uniformity in the sources. P-values were computed by Monte Carlo simulation as the table was larger than 2×2 [440].

SANCDB compound classes were distributed unequally among the sources (p-value 0.0004, confidence level = 0.95). The different superclasses had a variety of sources. For instance, the phenol lipids were produced by most sources (Figure 4-5). By contrast, organohalogen (mainly vinyl halides and organochlorides) were only produced by Plocamiaceae - all 25 halogenated compounds were isolated from the Plocamiaceae. Four species of Plocamiaceae were found in the database: *P. corallorhiza*, *P. cornutum*, *P. maxillosum*, and *P. suhrii Kützing*. These algae produce halogenated monoterpenes, both acyclic and cyclic, and these compounds do exhibit levels of cytotoxicity, as well as having known anticancer properties, particularly for anti-esophageal cancer [429,441], and antiplasmodial [442] activities. Some natural cyclic polyhalogenated monoterpenes from Chilean red alga *Plocamium cartilagineum* had insecticidal activity against *Macrosteles fascifrons* and the *Aster leafhopper* [443]. Similarly, alkaloids were mainly produced by Asteraceae and Amaryllidaceae. There is an overlap in some of the Classyfire superclasses. For instance, “alkaloids and derivatives” are a more specific case of "organic nitrogen compounds". Similarly, a compound could be an organoheterocyclic and also an "organic nitrogen” compound. Yet an organoheterocyclic compound might not be an organic nitrogen one. Compounds are hence classified in the most specific category.

Concerning the relationship between compound superclass level and source family, the associations with the highest number were lipids and lipid-like molecules with the Asparagaceae, Asteraceae, Lamiaceae and Euphorbiaceae. The phenylpropanoids and polyketides were associated with the Asparagaceae, Asteraceae, and Fabaceae. Asteraceae and Latrunculiidae were the main producers of organoheterocyclic compounds. Plocamiaceae were the main source of organohalogens. All these associations included at least 25 compounds (Figure 4-5). With respect to chemical class and source family, Asparagaceae produced the highest number of steroids (68) and homoisoflavonoids (58). Lamiaceae were the source of the highest number of phenol lipids (43). Other top associations were between phenol lipids and Euphorbiaceae (35), and phenol lipids and Asteraceae (29). Asteraceae, Lamiaceae, Fabaceae, and Amaryllidaceae as shown in Figure 4-2 produced many compounds of diverse classes including phenol lipids, steroids, flavonoids, homoisoflavonoids and quinolines.

These relationships are most likely rooted in the sources' inherent biosynthetic pathways [444] and may follow established chemotaxonomy. They may help refine taxonomy or identifying biochemical markers for some sources. For instance, quinoline alkaloids are known markers for the Rutaceae [368,444]. Similarly, Asparagaceae are major producers of homoisoflavonoids [445,446]. Quinolines and derivatives were mainly produced by Latrunculiidae and Amaryllidaceae. Latrunculiidae are known sources of pyrroloiminoquinone alkaloid and discorhabdins [447]. Also, some sources may be of particular interest as they are the only producer of a specific class of compounds. For instance, Plocamiaceae was the only producer of halogenated molecules. The associations can also guide compound discoveries by focusing on sources producing compounds of interest.

4.3.3 Compound activities

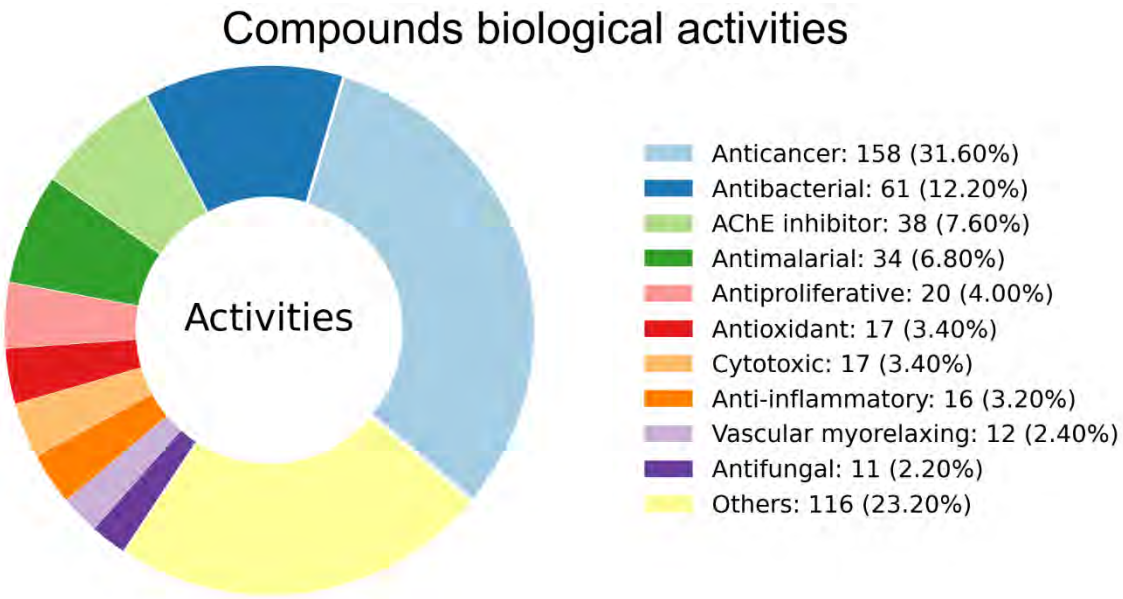


Figure 4-6 Biological activities of SANCDB compounds represented as a donut chart. 318 compounds activities were recorded in SANCDB. The 10 most reported activity classes are represented. All other activities are grouped in the "Others" category.

Fifty-nine distinct types of activities were found to be reported for these compounds in the literature. Anticancer was the most common biological activity. Indeed, 158 (31.6%) compounds had anticancer activity. Figure 4-6 shows the compounds' activities distribution. Antibacterials, acetylcholinesterase inhibitors, antimalarials and antiproliferative agents followed with these counts and percentages: (61 - 12.2%), (38 - 7.6%), (34 - 6.8%) and (20 - 4.0%) respectively. An interesting observation is the lack of anti-tubercular and anti-HIV compounds given that they are key health priorities in South Africa [448]. However, some compounds with anti-tubercular properties may fall into the antibacterial group. It is also important to note that the record of biological activities for the database was restricted to only compounds having a significant level of activity. Moreover, these records may still need to be standardized. Assays for biological activity can be done at different biological levels: disease, cellular, or molecular. Some compounds were found to be associated with multiple biological activities. Quercetin, Isoorientin, Ouabain, Combretastatin A-1 and Acovenoside A were associated with at least 6 different bioactivities. More interestingly, between them they had at least 15 predicted targets with 90% confidence in ChEMBL [216]. Hence, these are particularly good starting points for multi-target drugs.

These proportions of compound families and types of biological activities did not significantly differ from the previous content of the database.

4.3.4 Commercially available analogs

The non-availability of the previously identified hits in SANCDB mentioned in the introduction could simply have been that these were rare cases. Hence, the availability of the entire dataset was first assessed before searching for their analogs. Seventy percent of SANCDB was not available commercially. Only 316 and 327 of the compounds were found in MolPort [394] and Mcule [395] respectively. This may be explained by NPs' poor coverage in commercial libraries, as previously found [351]. For example, as previously mentioned, only ten percent of the ZINC NPs subset is commercially available [351]. Further, NP synthesis is often difficult [366], adding a greater challenge to their lack of availability. The probability densities distribution of SANCDB compounds synthetic accessibility score is shown in Figure 4-7. A synthetic accessibility [366] score greater than 6 was found for 118 of SANCDB compounds. Hence, most of the database may be accessible synthetically as the majority of compounds had a score below 6. However, for the ready availability of compounds, available analogs may be a good alternative.

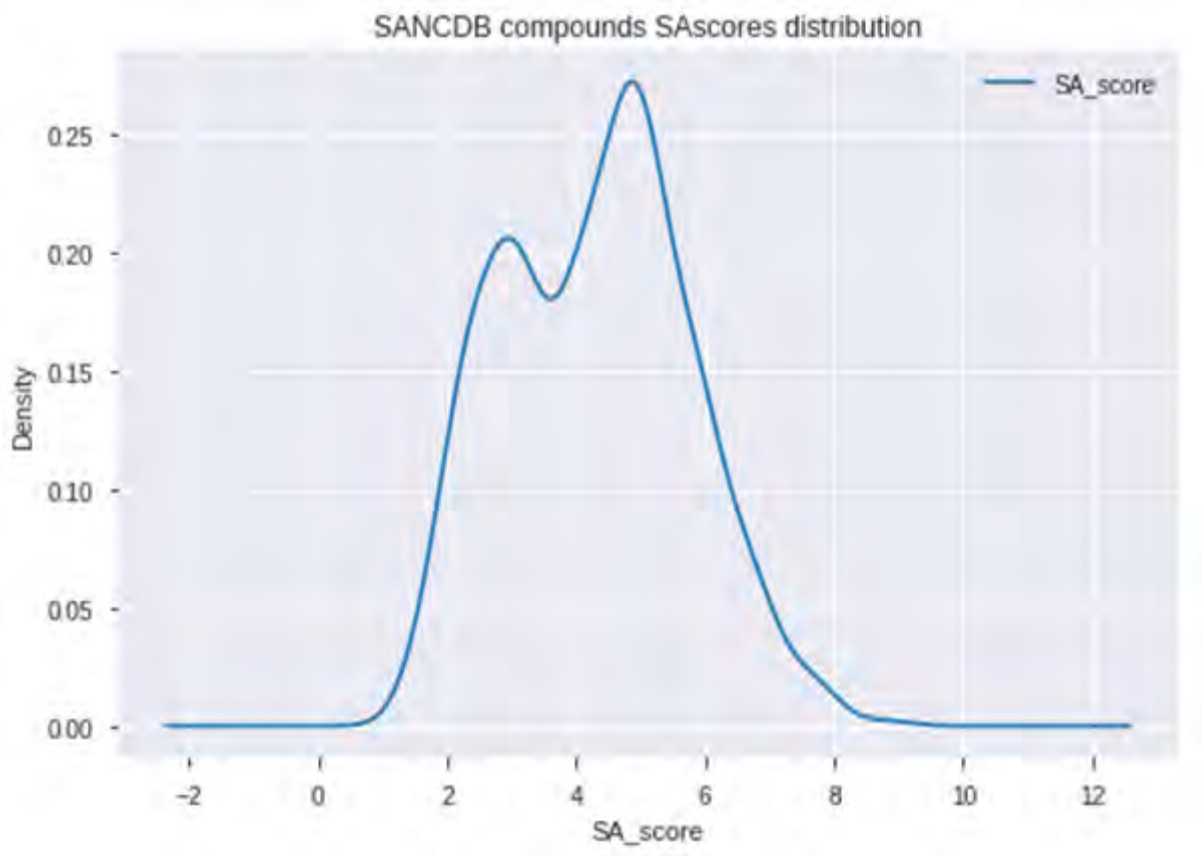


Figure 4-7 SANCDB compounds SAscore (synthetic accessibility score) distribution. Probability densities and SA_scores are on y and x-axis, respectively.

The number of compounds in Molport and Mcule was 7,597,214 and 9,884,200, respectively, at the date of download (October 2019). 1,487 analogs were found on average for each SANCDB compound. The circular bar plots in Figure 4-8 show the distribution of the number of analogs per compound in the updated SANCDB database.

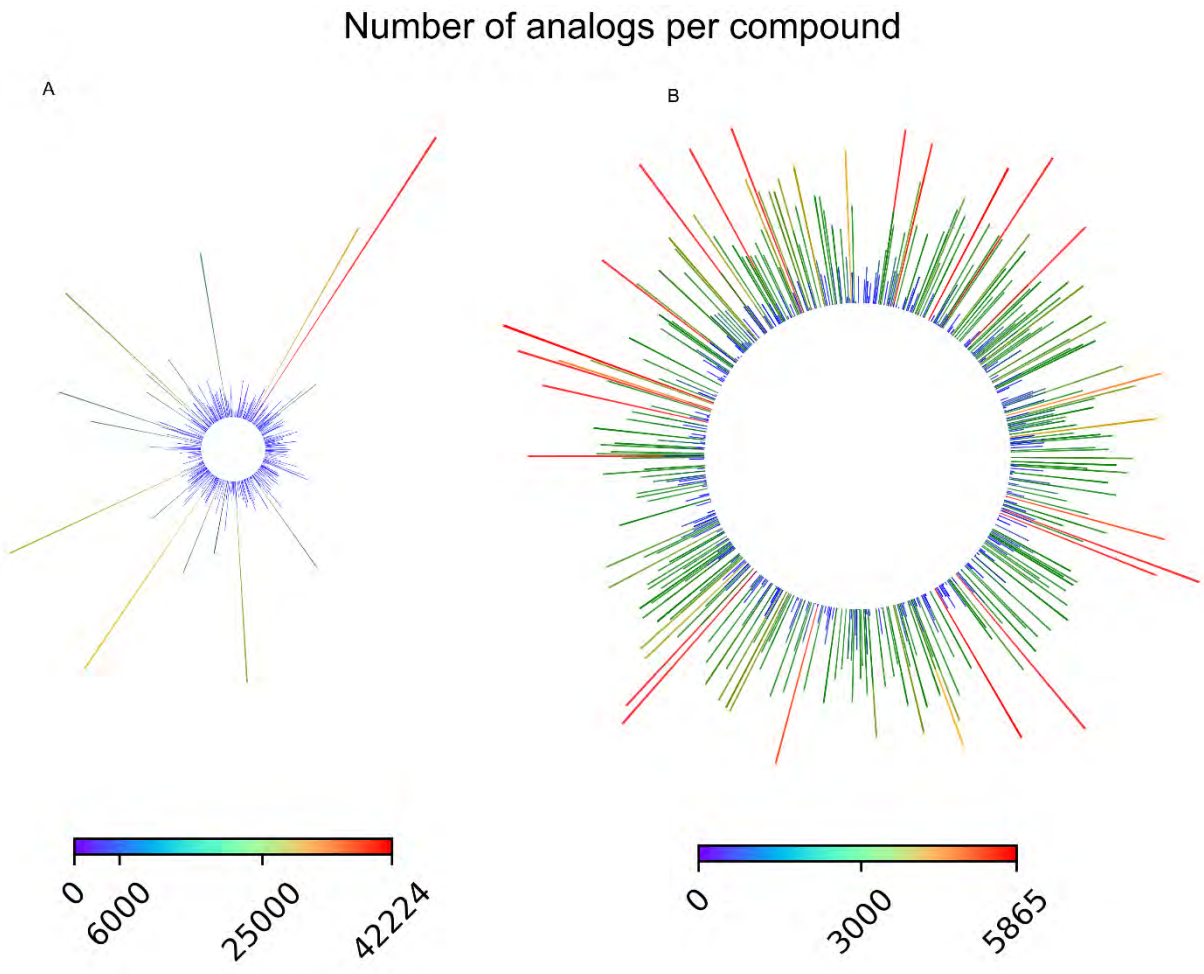


Figure 4-8 Circular bar plot of analogs count per compound. A) All content (1,012 compounds) B) Compounds having fewer than 6000 analogs. The analogs count is depicted in the color key.

Frequencies ranged from 42,224 to zero analogs. All analogs have linked from the SANCDB web interface to their respective page on either Mcule or Molport. The Tanimoto similarity scores and the SMILES were made available on a specific page for each SANCDB compound. Hence, users interested in hit optimizations through analogs can easily download structures associated with a hit. The total number of unique analogs added to the database was 374,067. With the latter automated update through the API, this number may increase. 141,320 were found in Molport, while 232,747 compounds were from Mcule. Analogs count per compound varied from 42,224 to zero. The highest number of analogs were found with SANC00428 (42,224) SANC00815 (29,045), SANC00656 (27,823), SANC00967 (26,638) and SANC00425 (24,993). Seventy compounds had zero analogs. These compounds had a low MW (<300 Da). All compounds with over 10,000 analogs had an MW bellow 300 Da. However, comparing the analogs per compound and MW only had a negligible Pearson correlation of -0.09 (Figure 4-9).

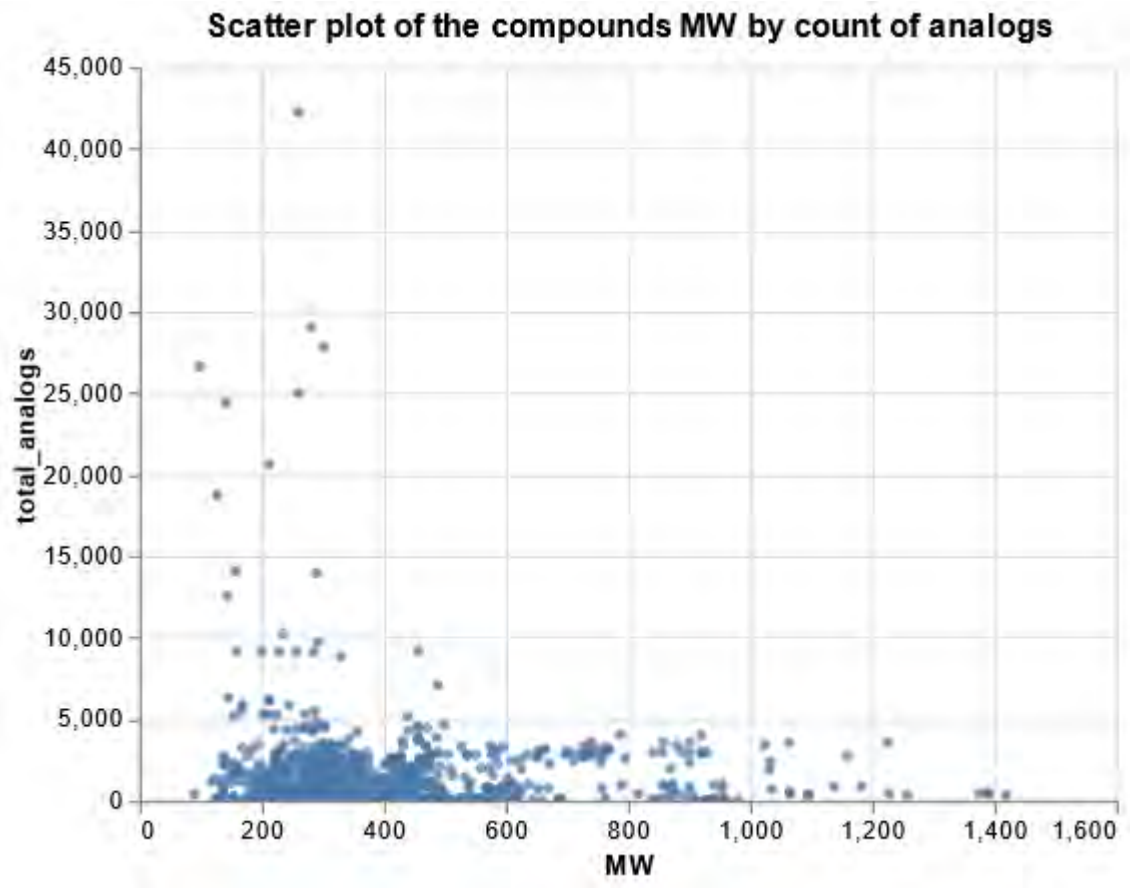


Figure 4-9 Scatter plot of compounds MW versus analogs count. X-axis and y-axis correspond to MW (Dalton) and the number of analogs respectively

Analogs were searched in SciFinder [383] for the 70 compounds without analogs in Mcule and Molport. One to 29 analogs were identified for these compounds. SciFinder seems to be a more comprehensive database than Mcule and Molport. Yet, it does not offer a batch search functionality or API access. Hence, the search was manual for a similarity interval at a time. For example, a user can only search an interval (e.g. [0.75-0.79[). Hence this requires six independent searches to cover the [0.6-1] interval (0.6/0.05 similarity). To save time, only the first interval having analogs from all searched ones are shown here.

No more than 1,000 analogs were found for up to 570 SANCDB compounds. Given the large size of these two databases (7,597,214 and 9,884,200 molecules for Molport and Mcule respectively) and the low similarity threshold (0.6) used, this shows the lack of SANCDB chemical space coverage in these datasets.

This work expands SANCDB chemical space and provides the commercially available analogs to help further *in vitro* testing following modelling. These analogs can also be used for screening hits optimization. This may also be applied to the previously identified hits.

4.3.5 SANCDB and chemical space

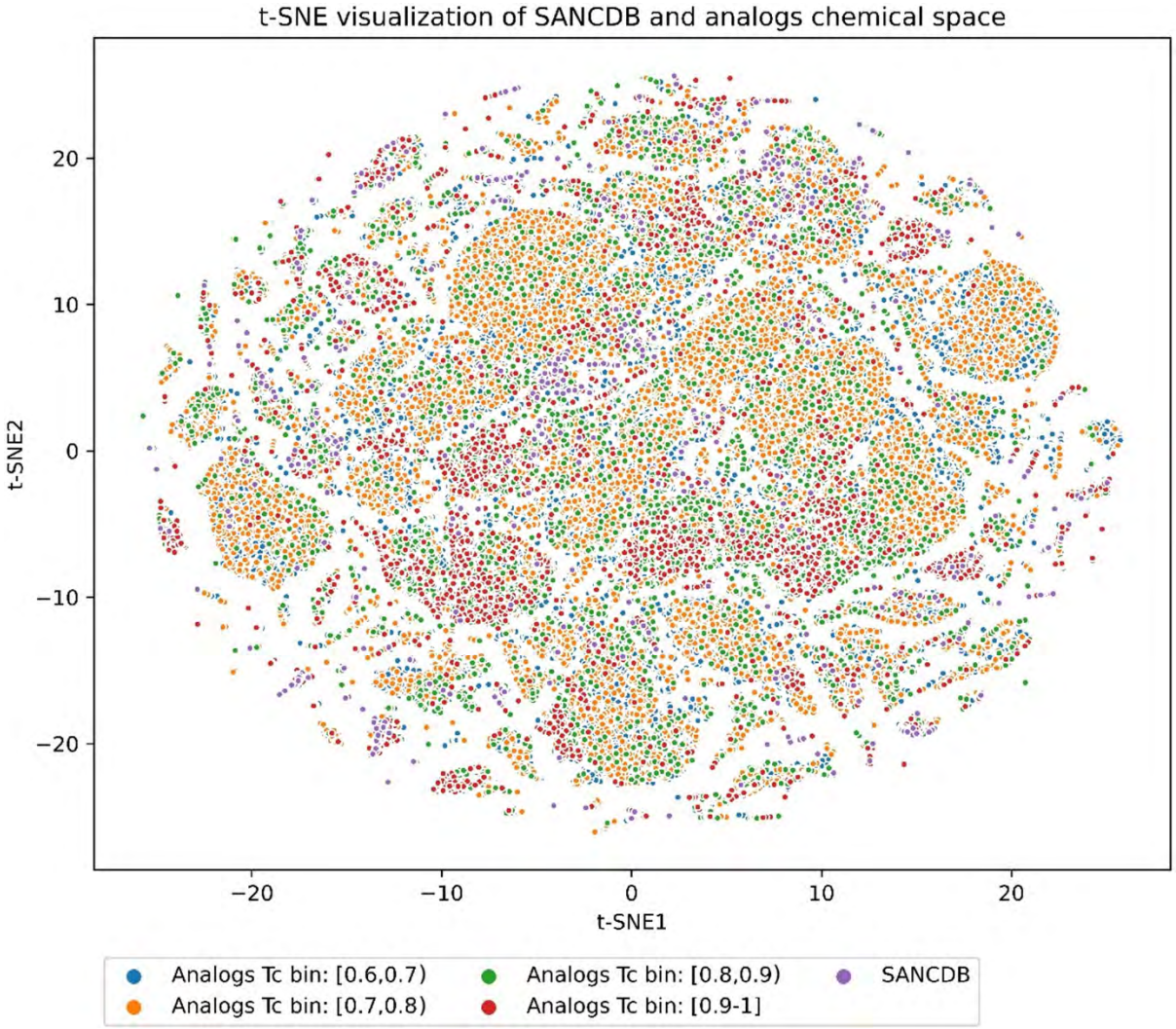


Figure 4-10**:** visualization of SANCDB and analogs chemical space. Compounds (n = 375061) are represented in dots. SANCDB (violet, n = 1012). Analogs are in bins of similarity values [0.6,0.7) (blue, n = 266147), [0.7,0.8) (orange, n = 69336), [0.8,0.9) (green, n =24679), [0.9-1] (red, n = 13887). As an analog may have different similarity scores with different SANCDB compounds, the maximum similarity score was chosen for each analog.

In terms of t-SNE visualization, compounds clustered into a single ball-like (Figure 4-10). This may indicate that all analogs and SANCDB compounds fitting in the same region despite the possibility of high diversity in that region. Isolated SANCDB compounds (isolated violet points) may correspond to compounds without analogs. The ball-like shape may also indicate a too high learning rate according to t-SNE documentation [404]. However, gradually decreasing the learning rate from 200 to 50 did not change the shape. The plot above was obtained using perplexity 50 and a learning rate of 100. Different values of learning rate and perplexity and combination of both were tried, all showing a similar pattern. These values are in the ranges of recommended values for t-SNE [405]. Despite the general ball-like shape, some internal clusters can be

distinguished. Indeed, some compound clusters can be noticed around the following coordinates (x= 20, y = 10), (x =20, y = -5), and (x =-15, y = 10). These compounds may share common properties. Compounds without analogs did not form a distinct cluster, but rather scattered inside the ball. This may be linked to their inner diversity. Indeed, structural analysis of the compounds without analogs showed many had different chemical scaffolds to other compounds without analogs. Although the discorhabdin scaffold was common in 17 of them.

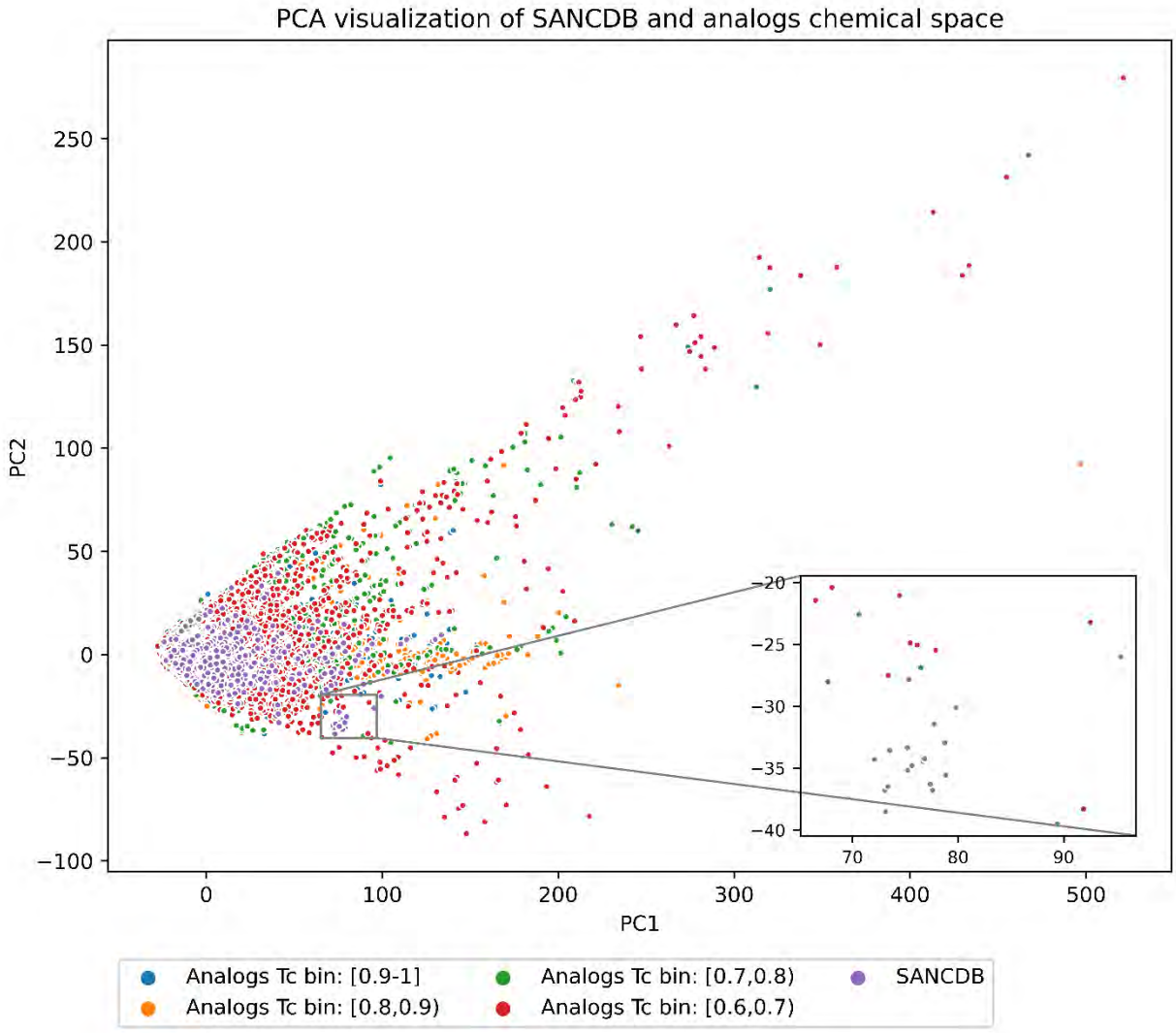


Figure 4-11 PCA visualization of SANCDB and analogs chemical space. Compounds (n = 375061) are represented in dots. SANCDB (violet, n = 1,012). Analogs are in bins of similarity values: [0.6,0.7) blue, n = 266147, [0.7,0.8) orange, n = 69336, [0.8,0.9) green, n =24679, [0.9-1] red, n = 13887. As an analog may have different similarity scores with different SANCDB compounds, the maximum similarity score was chosen for each analog. The first two components explain 81% of the variance (PC1 (66%), PC2 (15%)).

The PCA analysis showed SANCDB chemical space regions coverage with analogs (Figure 4-11). Indeed, there was an overlap with most SANCDB compounds and their analogs. The spread and diversity from SANCDB compounds is due to decreasing similarity scores. As a result, those in the

interval [0.6,0.7) of similarity scores were the most isolated. Some SANCDB compounds without analogs occupied a small, isolated cluster (zoomed region of the plot). They were some of the compounds without analogs and share the discorhabdin scaffold. The t-SNE visualization showed a less dense cloud, having an inner clustering of the compounds compared to PCA. Yet this discorhabdin cluster was not captured in it. The two visualizations seem complementary.

To further contribute to solving the availability problem, analogs can be searched in the synthetically accessible space. Indeed, the current work was limited to readily available compounds. The Enamine database currently contains 1.36 billion synthesizable molecules. By comparison, Mcule and Molport have less than ten million available compounds with the potentiality of some overlap between the two datasets. Hence, an analog search in the Enamine dataset could improve accessibility, especially for compounds without analogs and which are difficult to synthesize.

In the future, analogs can be evaluated as a replacement for hits previously identified. For instance, analogs can be used for *in vitro* testing of hits and allosteric modulators identified previously. These potential future studies may be preceded by *in silico* modeling to compare hits and their analogs predicted activities. Moreover, this work might inspire other NPs' data sources to make their commercial analogs available. For that purpose, the API integration used here offers a regularly updated and sustainable solution.

4.3.6 Scaffolds and compounds subsets

The analysis of compound scaffolds and druggability subsets was carried out to assess database potentiality for drug discovery. sp^3-configured centers found in NPs make them attractive in virtual screening [435,449] but also as a starting point for further optimization [423]. Also, diverse scaffolds in libraries is ideal for virtual screening [373].

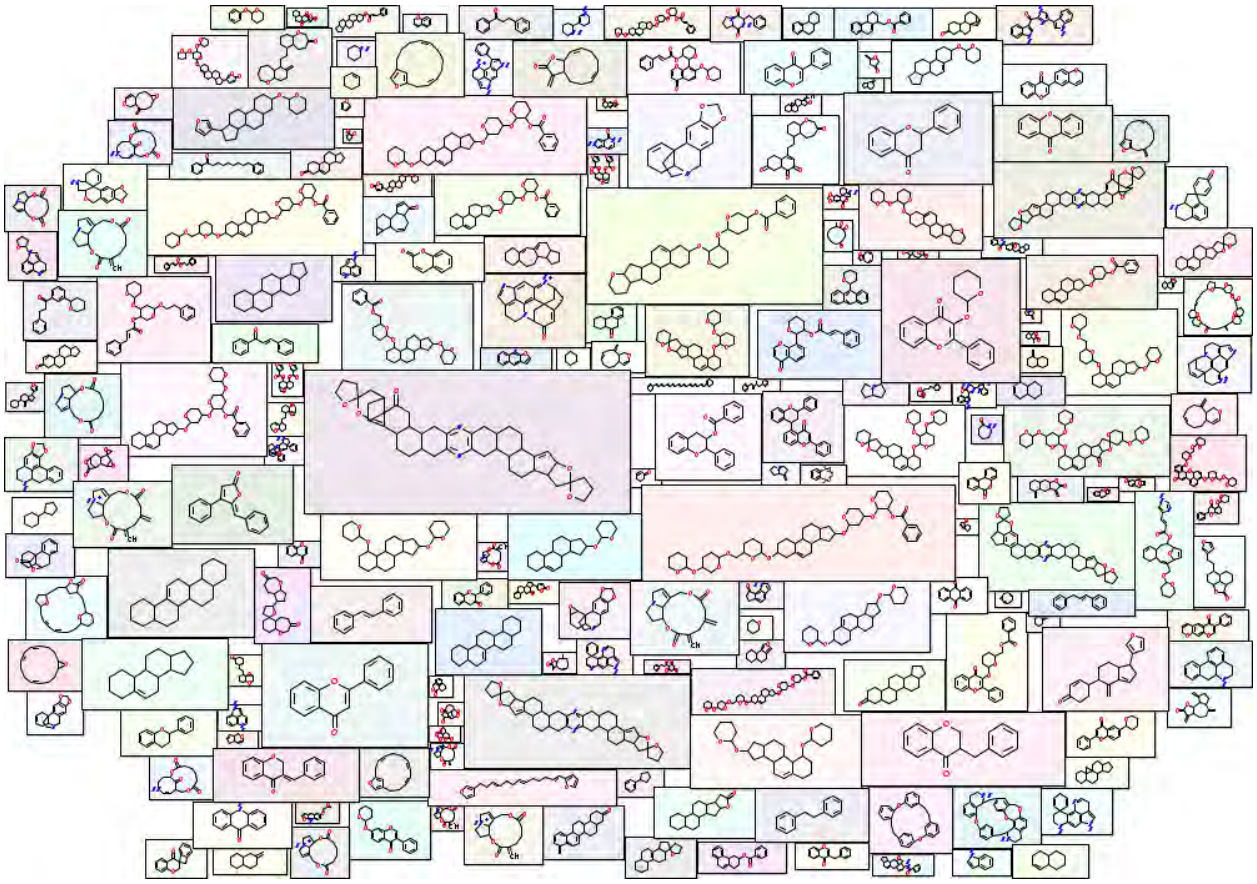


Figure 4-12 Molecule cloud of SANCDB scaffolds. Structure sizes indicate scaffold frequencies. The benzene ring is a special case, being the most frequent scaffold in all large data sets [414]. Therefore, it is not displayed.

The molecule cloud visualization gives an overview of the database diversity, the most frequent scaffolds, and their structural features [414]. However, less common scaffolds with interesting properties might not be highlighted.

501 unique scaffolds were identified. Hence, about half of the database presents a unique scaffold showing the diversity of the SANCDB database. Figure 4-12 shows all scaffolds. Further, the scaffold counts had a “long tail” distribution common to chemical libraries [414] (Figure 4-13) and indicating the high number of singletons. Each of these singletons has a unique scaffold in the database supporting its diversity. 59 compounds were noncyclic.

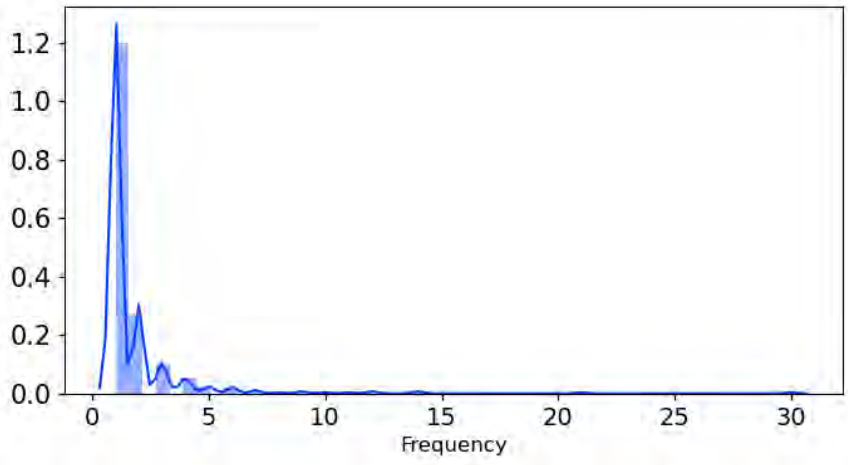


Figure 4-13 Histogram and kernel density distribution of the scaffolds count.

Flavonoids are frequently encountered scaffolds, common in NP libraries [411]. Figure 4-14 presents the top ten scaffolds. Interestingly, the flavonoid class was only the third in class distribution as shown in Figure 4-4. This may be linked to a greater structural diversity encompassed by the first two categories (phenol lipids and steroids and derivatives) compared to the flavonoid class. The chromane 3-Benzylchroman-4-one had the scaffold the highest number of compounds (30). Chromane scaffolds are also frequent in NPs [450]. Its structure consists of a bicyclic 3,4-dihydro-1-benzopyran. Chromanes are known to inhibit human monoamine oxidase B [451] and also for anticancer activity [452]. This may explain the database abundance in terms of compounds with anticancer activity (Figure 4-6). Flavone was the second most frequent scaffold with 21 compounds. The prodrug aminoflavone reached phase 2 clinical trials to treat breast cancer [451]. Flavanone with 14 compounds was the third most common scaffold. All these top scaffolds present a ketone group, a structural alert, which may cause toxicity due to its high reactivity [453]. The fourth scaffold was a delta-5-steroid. Estrona, a compound based on this scaffold, reached phase 3 clinical trials for homeopathic treatment of premenstrual syndrome [454]. The fifth was the chromenone scaffold, these are coumarins, whose derived drug warfarin is known for its anticoagulant properties [455] and these also present a keto group at 2-position. All top five scaffolds were linked to biological activities. Hence, related compounds may be used as a good starting point for further exploration.

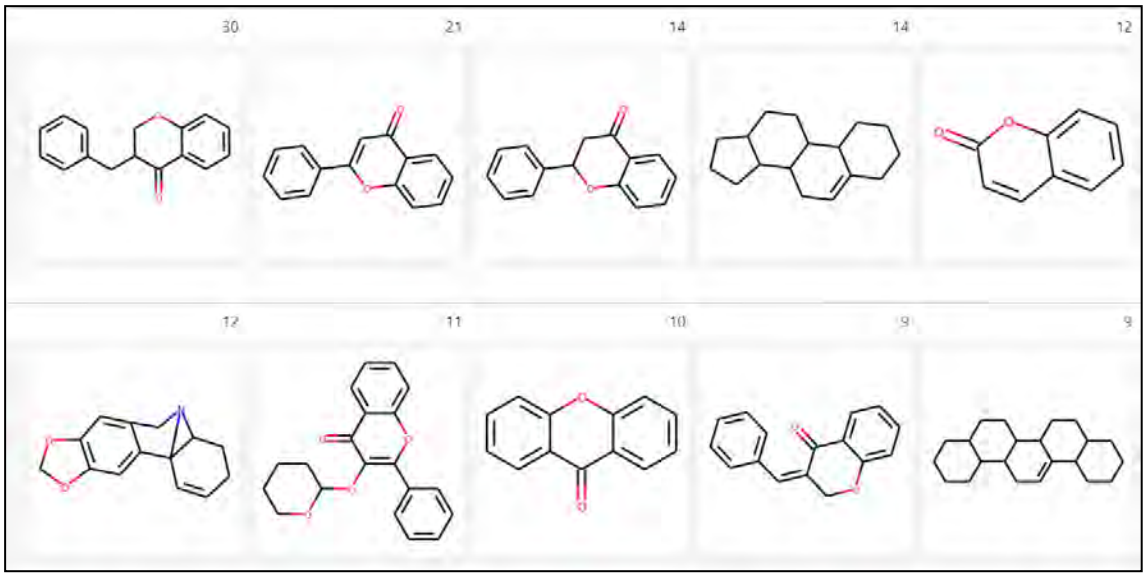


Figure 4-14 Structures of the ten most common SANCDB scaffolds and their counts. Structures were drawn using RDKit [162]

Compounds were further categorized into subsets relevant to drug discovery. Screening datasets such as ZINC [415] are often subdivided into subsets. PAINS patterns are used to filter out frequent hitters in screening [417]. Figure 4-15 shows their repartitioning into drug-like, extended drug-like, PPI-like, fragment-like, and lead-like subsets.

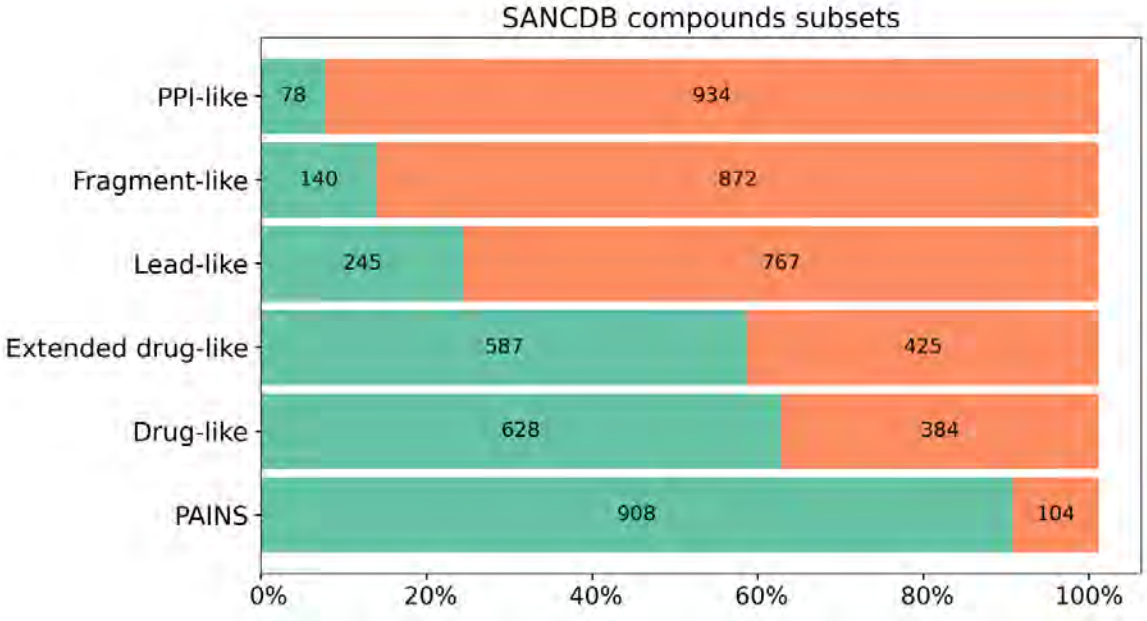


Figure 4-15 SANCDB compounds repartitioning in drug-like, extended drug-like, fragment-like, lead-like, PPI-like subsets on the y-axis. The x-axis represents the number of compounds in each subset with their

related percentages. The green area indicates compounds complying with rules specific to that subset. For PAINS, it corresponds to the absence of PAINS pattern.

About 90% of the database was free of PAINS pattern and more than half of it was extended drug-like or drug-like. These subsets are the most important for drug discovery and hence this indicates the good potential of the database for this purpose. Extended drug-like and drug-like only differed by 41 compounds. The former adds two more conditions to the drug-like (nRot <=7) & (TPSA < 150). Fragment-like and PPI-like had the lowest compound counts with 140 and 78, respectively. The low count in fragments-like compounds can be linked to many SANCDB compounds being polycyclic nature. As shown in the compound classification (Figure 4-4C), polycyclic compounds were common in SANCDB (~75%). PPI-like compounds had the lowest count with 78 compounds. This contrasts with the database containing 75% polycyclic compounds. In general, SANCDB is not biased toward fragments nor very large compounds (PPI-like) but mostly represented by drug-like ones.

Other NPs databases had similar distributions. Extended drug-like and drug-like represent more than 50% while fragment-like and PPI-like, have low proportions [412]. The subsets may fit different drug discovery scenarios, hence guiding virtual screening. For instance, PPI-like and fragment-like compounds can be useful for protein-protein inhibition or fragment-based drug discovery. Also, initial potent chemotypes can be identified from fragments for future optimization.

4.4 Conclusion

This work aimed mainly at updating SANCDB to at least a thousand compounds with new isolated NPs to harness more of the South African biodiversity with the continuously growing isolated NPs in the region [374–376]. Through a literature search, 412 new NPs were added to the database for a total of 1012 compounds. The interest in NPs especially for drug discovery and chemoinformatics [347,351] contrasts with their insufficient commercial availability [351]. Hence, a secondary aim was to provide readily available analogs for all SANCDB NPs. A total of 374,067 analogs were added to the database. SANCDB compounds were also made available in formats relevant in drug discovery AutoDock [456] pdbqt and Schrodinger Maestro [393]. Analogs were linked to their sources on Mcule [395] and Molport [394] and their update was automated through API integration [395] and Molport [394].

This update can benefit chemoinformatics and drug discovery by providing a larger chemical library, especially for virtual screening. More automated API updates of analogs in the database will contribute to database maintenance. Compound subset analysis showed that SANCDB may be promising as a source for drug-like compounds. The entire dataset available in ready-to-dock formats can accelerate virtual screening pipelines setup and foster more prospective screening studies. The classification and scaffold analysis showed content diversity with 501 unique NP scaffolds. The availability of analogs is a unique feature of an NP repository. Analogs can contribute to screening hit optimization, as an alternative, for their *in vitro* testing and may inspire other NPs' resources to integrate NPs analogs given their low commercial availability.

In the future, the database may be extended to cover the southern African region or to serve as a good starting point to cover the entire African continent. Many NPs repositories cover the African continent [64,65,422,434,457]. They usually cover specific countries or regions. There is no

database currently covering the west African region or covering the entire African region. This could be done through the unification of the different country-specific databases. A similar approach is currently being envisaged for the Latin America compound database [458]. Considering the limit of human effort and the growing literature, future updates should consider using automated text mining tools. Elsewhere, a web-based virtual screening pipeline could be integrated into the database for drug discovery. The analogs' space could also be extended to the synthetically feasible compounds.

Chapter 5: Side project, Thymol as a potential antagonist of serotonin 5-HT_{3A} receptor for IBS treatment

5.1 Introduction

Irritable bowel syndrome (IBS) is a complex disease involving multiple symptoms such as visceral discomfort, diarrhea, constipation and disturbance of the gastrointestinal (GI) transit [459]. Its global prevalence ranges from 15 to 45%, causing 20 billion dollars healthcare cost per year in the USA, and is the second most frequent reason for work absenteeism and the most common one for gastroenterologist visits. Its etiology is not fully understood. Some family history, genetic predisposition, female sex, unbalanced or gastrointestinal tract-aggressive diet and stress have been identified as a predisposing risk for IBS [460,461].

Modulating serotonin receptors can alleviate IBS symptoms**.** This signal transducer is implied in anxiety, mood, sleep, and gastrointestinal motility [462]. Earlier work confirmed serotonin's role in IBS pathogenesis [463]. The serotonin (5-HT, 5-hydroxy-tryptamine), type 3 receptors (5-HT_3R) can regulate autonomic functions, such as motility and peristalsis, secretion and visceral perception, and can thus contribute to functional GI disorders, such as IBS [464]. Setrons (5-HT_3R antagonists) are clinically used for IBS treatment [465].

Thymol, a monoterpenoid used in digestion and bowel-related problems [466] reduces the severity of IBS syndrome [467]. Terpenes are known 5-HT_3A modulators [465]. In a stress-induced IBS rat model, thymol enhanced the GI transit, decreased the fecal count and lowered visceral pain. Further immunohistochemical analysis of colon and intestine tissues showed an increase in serotonin receptor after thymol treatment, supporting possible thymol antagonizing effect on 5-HT_{3A}R. Indeed, 5-HT_{3A} antagonist compounds are capable of managing stress-driven IBS defecation [467].

Previous studies showed that thymol is an activator of human 5-HT_{3A} [468] through an allosteric transmembrane site [469]. Investigating several terpenes as human 5-HT_{3A} modulators, carvacrol activated it [465]. However, carvacrol and thymol were found to have an interesting species selectivity on 5-HT_3Rs, where they were observed to be agonists in human but not in mouse [469]. In the same line, colchicine acted as a human 5-HT_3Rs positive allosteric modulator but inhibited the mouse one [470]. More interestingly, site-directed mutagenesis identified transmembrane amino acids either abolishing carvacrol and thymol agonist activity on human 5-HT_{3A}Rs or activating them on mouse 5-HT_{3A}Rs [469]. Additionally, thymol activates human 5-HT_{3A}R

but has a different level of potentialization on the receptors on other subunits, greater at 5-HT_{3A} receptors than 5-HT_{3AB}, 5-HT_{3AC}, 5-HT_{3AD}, or 5-HT_{3AE} receptors. There was no activity data in ChEMBL between thymol (CHEMBL29411) on the mouse 5-HT_{3A} (CHEMBL2111333) nor on the one of the rat (CHEMBL2411).

Table 5-1 Activities of some investigated molecules here on different organisms 5-HT_{3A}Rs

Molecules	Human	Rat	Mouse
Thymol	Allosteric agonist [469,471] weak partial agonists and positive modulators [468]		No agonist or potentiating effect [469]
Serotonin	Orthosteric agonist	Orthosteric agonist	Orthosteric agonist
Tropisetron	Serotonin binding site competitive antagonist [462]		
Carvacrol	weak partial agonists and positive modulators [468]		No agonist or potentiating activity [469]

Understanding the underlying molecular mechanism in IBS is important given its health-associated burden. Also, provided the above *in vivo* and *in vitro* observations on thymol effect, and the previous finding associating thymol and the 5-HT_{3A}R receptor modulation, we intended to further determine whether thymol can competitively antagonize the serotonin receptor. For that purpose, a comparative analysis of serotonin, thymol and tropisetron (a serotonin antagonist) in terms of binding modes, interactions and energies was performed using molecular docking. Further, MD was used to investigate their stability and analyze their agonist-antagonist behaviors.

5.2 Methods

5-HT_{3A}R is a Cys-loop receptor, and a ligand-gated ion channel with a pentameric structure. Its extracellular domain (ECD) has an orthosteric site at the interface of two adjacent subunits (Appendix P). Currently, there is no crystal structure of the rat 5-HT_{3A}R. Recently, Lucie Polovinkin *et al.* solved four cryo-electron microscopy structures of the mouse shedding much light on the protein functional cycle [462]. These mouse structures were used, and are:

- 6HIQ: 5-HT_3R + serotonin (agonist) + TMPPAA (positive allosteric modulator): 6HIQ I2 conformation
- 6HIO: 5-HT_3R + serotonin (agonist): 6HIO I1 conformation
- 6HIN: 5-HT_3R + serotonin (agonist): 6HIN F (Full conformation) open state, activated ECD.
- 6HIS: 5-HT_3R + tropisetron (antagonist): 6HIS T conformation, inhibited state.

The receptor structures were first selected based on their quality. 6HIQ has the best resolution (3.2 Å) among mouse 5-HT_{3A} receptors (Uniprot ID: P23979). We included the 3 other receptors as they presented different conformations. This helped to assess thymol binding to these other conformations. Also, 6HIS is bound to the antagonist tropisetron was used as a positive control to compare its effect to that of thymol in molecular dynamics.

These structures were prepared for docking using AutoDock tools [472]. They were protonated at physiological pH (7) as has been previously done with this receptor [473,474]. Serotonin was redocked through blind docking to 6HIQ using QuickVina-W [96] to validate the docking protocol. The used parameters are provided in Appendix L. Docked vs crystal pose RMSD value was computed without least-squares fitting using GROMACS (Version 5.1.2) [182]. Thymol (PubChem CID: 6989) and serotonin (from the crystal structure) were then docked to all four conformations. Tropisetron was only docked to 6HIS. In each structure exhaustiveness for docking was scaled to the protein size using a reference value of 24 for a 30 $Å^3$ (24 is 3 times the Autodock Vina (Version 1.1.2) default value (8) [165]). Each docking generated ten poses and the lowest energy poses were selected for MD. Nine 50 ns MD with GROMACS[112] using the Amber ff99SB-ILDN [187] force field were done to assess ligands' stability. The nine simulated systems were: serotonin in the four conformations, thymol in the four conformations and tropisetron in 6HIS. Ligand topologies were generated using ACPYPE [186]. TMPPAA (a positive allosteric modulator in 6HIQ) was not included in the simulations. The co-crystalized serotonin and tropisetron, binding on the docked thymol binding site were used. Given large structures' sizes, only the ECD (residue 1 to 219) was simulated. Missing residues in 6HIS ECD were first modelled using Prime version 5.4 (r012) (schrodinger2018-4) [159]. Simulations were done in a dodecahedron 10 Å between the solute and the box set to and using the tip3p water model with 0.15 M [Na^+Cl^-]. Systems were minimized using steepest descent with a maximum force and the number of steps set at < 1000.0 kJ/mol/nm and 50000, respectively. They were equilibrated at 1 atm and 300 Kelvin during 50 ps in the isothermal-isobaric ensemble and canonical one. The leap-frog algorithm was used for integration. Short-range electrostatic and Lennard-Jones thresholds were set at 10 Å. For the long-range electrostatic interactions, a fourth-order interpolation and the smooth particle mesh Ewald were used. Simulations were conducted at CHPC. The trajectories were visualized in a Jupyter Notebook[190] using Nglview [188] and the Pytraj package[475]. GROMACS[112] modules root-mean-square deviation (RMSD), radius of gyration (Rg) and root-mean-square fluctuations (RMSF), protein-ligand interaction energy and the number of hydrogen bond were used to assess systems stability.

5.3 Results and Discussions

5.3.1 Thymol binds serotonin binding site with comparable energies to serotonin in all four conformations

In docking validation, the serotonin best pose had an RMSD of 2.2 Å (Appendix O). Docked on all four conformations (6HIQ, 6HIN, 6HIO, 6HIS) in blind docking, thymol binds in the serotonin binding site and the binding affinities are similar to that of serotonin. Indeed, the 10 docked serotonin poses binding affinities range from -7.6 kcal/mol to -8 kcal/mol while those of thymol range from -7.3 kcal/mol to -7.9 kcal/mol (Appendix J). Given the similar range of binding energy, a partial antagonist mechanism may be envisaged for thymol. On the other hand, tropisetron

consistently had binding affinities lower than - 8 kcal/mol. Thus, the evidence supports that it can competitively antagonize serotonin on the orthosteric site, with a full antagonist mechanism for this site.

Elsewhere, simply rescoring the co-crystalized ligands resulted in lower affinities than the redocked one, even though the co-crystalized and redocked ligand had RMSD <= 2.5 Å (Appendix J).

Thymol binds in a fully buried binding site, mainly stabilized with Pi-Pi interactions with aromatic residues rings (TYR207, PHE199, TRP156, TYR126 and TRP63) (Figure 5-1, Appendix N, Appendix M). The compound does not make any hydrogen bond in its docked conformation.

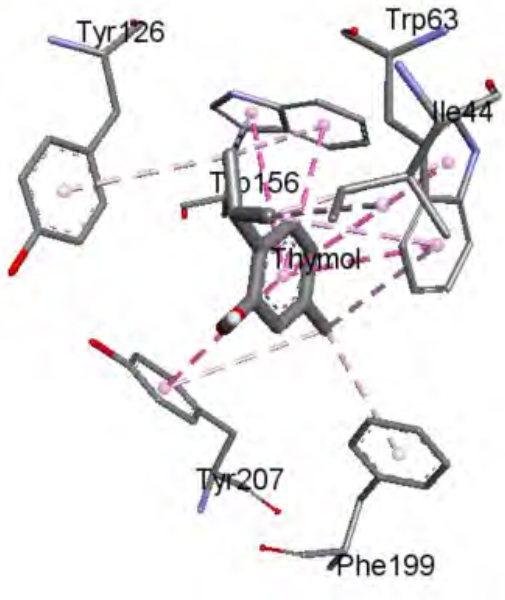


Figure 5-1 Thymol docked in 6HIQ. Interacting residues in stick and their three letter codes and residues numbers are shown. Dashed pink lines represented hydrophobic contact. The plot was obtained from Discovery Studio Visualizer V1.7.2.

In this study, docking predicted thymol binding on the orthosteric site in four ECD conformations as well on the full protein including the transmembrane region (Figure 5-2, A, B). This site is also the binding site for tropisetron, a competitive antagonist for 5-HT_3 receptor. All lowest energy conformations (LECs) bound the same orthosteric site in the ECD in all four conformations and interacted with all obligatory binding residues PHE199 and TYR207 TYR126 and TRP63, except for PHE199 in 6HIS [462]. Further, with the exception of 6HIS, all other poses bound the site. In 6HIS, the 8th and 7th poses bound the membrane domains, in a region near the extracellular domain. Still, the region was distinct from the proposed one by Lansdell *et al.* in which was predicted to bind human 5-HT_{3A}R transmembrane [469]. Hence, according to the current docking results, a possible allosteric mechanism through the transmembrane region may be excluded.

The mouse structure (4PIR) was also used (Appendix P). The structure was used by Lansdell *et al.* [469] for homology modelling of the human one. Thymol, serotonin, and NAG (N-acetyl-D-Glucosamine, the co-crystal ligand in 4PIR) were blindly docked in the 5 five structures. An exhaustiveness of 6000 was used to efficiently explore the large search space. The redocked serotonin in 6HIQ showed an acceptable RMSD with the co-crystallized serotonin. Concerning thymol, all the best poses (the most energetically favourable) are bound to the extracellular domains of the structures. Moreover, all the other poses also bind the same domain in all structures except in 6HIS. In that last case, the 7th and 8th poses bind in the intracellular domains in an extreme region close to the extracellular domain. However, this region is still different from the one proposed by Lansdell *et al.* This was expected given 4PIR structural similarity to the other structures. Comparing 4PIR to other structures, 6HIN, 6HIQ, 6HIO all had RMSD below 2.7 Å using both only c-alpha and all backbone atoms. On 6HIN, on the other its RMSD was 5.1 Å for both the c-alpha and all backbone atoms.

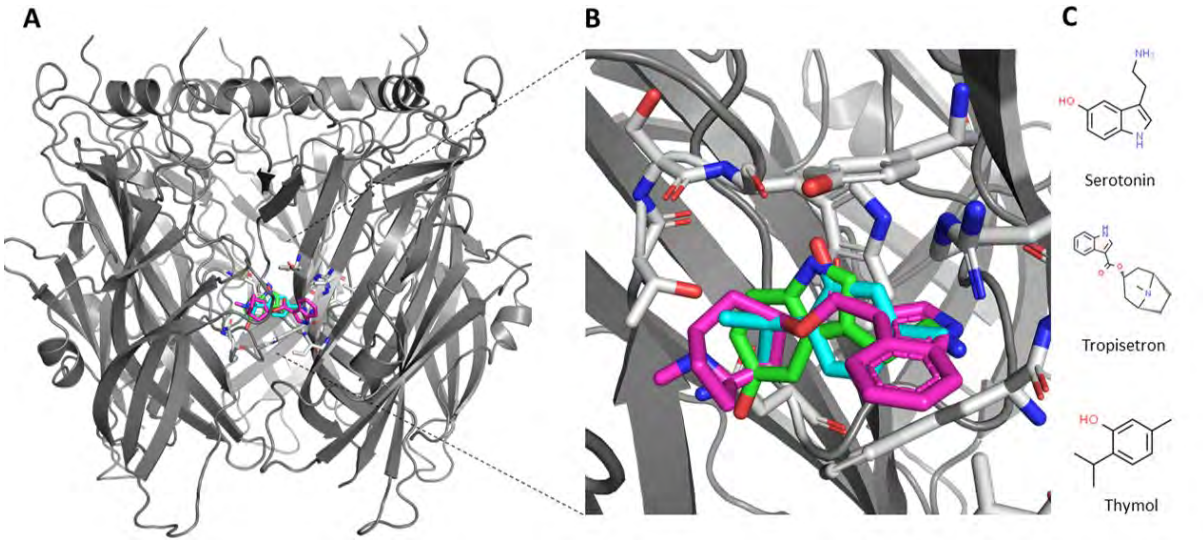


Figure 5-2 Tropisetron (magenta), thymol (cyan) and serotonin (green) docked in 6HIS. (**A**) ECD in cartoon representation. Docked ligands and crystalized tropisetron superimposed in the active site. (**B**) Active site zoomed-in view. Interacting residues (light grey). (**C**) 2D depiction for thymol, serotonin and tropisetron structures.

Thymol and serotonin have similar binding energies (Appendix J, Appendix K). Further, it also showed similar interacting residues to serotonin in the three protein conformations (6HIQ, 6HIO, 6HIN). Some variations are noticeable in 6HIS. Thymol and tropisetron bound the same site in 6HIS (Figure 5-2) and interacted with identical residues in TYR207, TRP156 in chain A, and TRP63, ARG65, ILE44 in chain E (Appendix N, Appendix M). Thymol interacts with all residues that tropisetron binds to in the crystal structure. The sole common residue to all three compounds is TRP156 polar contact in chain A. Hence serotonin does not share common residues with thymol and tropisetron in 6HIS. This may be explained by 6HIS being the inhibited state of the protein and that thymol is acting as an antagonist like tropisetron. However, thymol and serotonin had similar binding energies in that particular protein conformation: -6.5 kcal/mol and -6.7 kcal/mol

respectively. On the other hand, tropisetron had a significantly better binding energy of -9.3 kcal/mol to this conformation. The energy difference may be explained by an interaction between the tropisetron indole ring and 6HIS ARG65 (Appendix N). The difference in interaction residues might explain thymol's *in vitro* effect. However, the binding energies do not seem to support this.

In all conformations, thymol interactions are mainly driven by hydrophobic contact especially between its phenol ring and TRP and TYR residues while serotonin consistently formed 1-2 hydrogen bonds (Appendix N).

Comparing ligands' structures, thymol and serotonin share a phenol ring but serotonin and tropisetron present an indole absent in thymol. This indole ring is a familiar moiety in 5-HT_{3A} bioactive compounds (CHEMBL4972). Moreover, most 5-HT_3R antagonists follow the pharmacophore constraint of an aromatic ring, hydrogen bond acceptor and a ring-embedded nitrogen [476] as in tropisetron structure (Figure 5-2). Even though thymol possesses an aromatic ring and HBA it lacks a ring-embedded nitrogen, and the constraint between the aromatic ring and the HBA is not fulfilled. Hence its potential antagonism may take place through a different molecular mechanism, hence it is a new class of 5-HT_{3A} modulator, worthy of further structure–activity relation investigation.

5.3.2 Molecular dynamics simulations

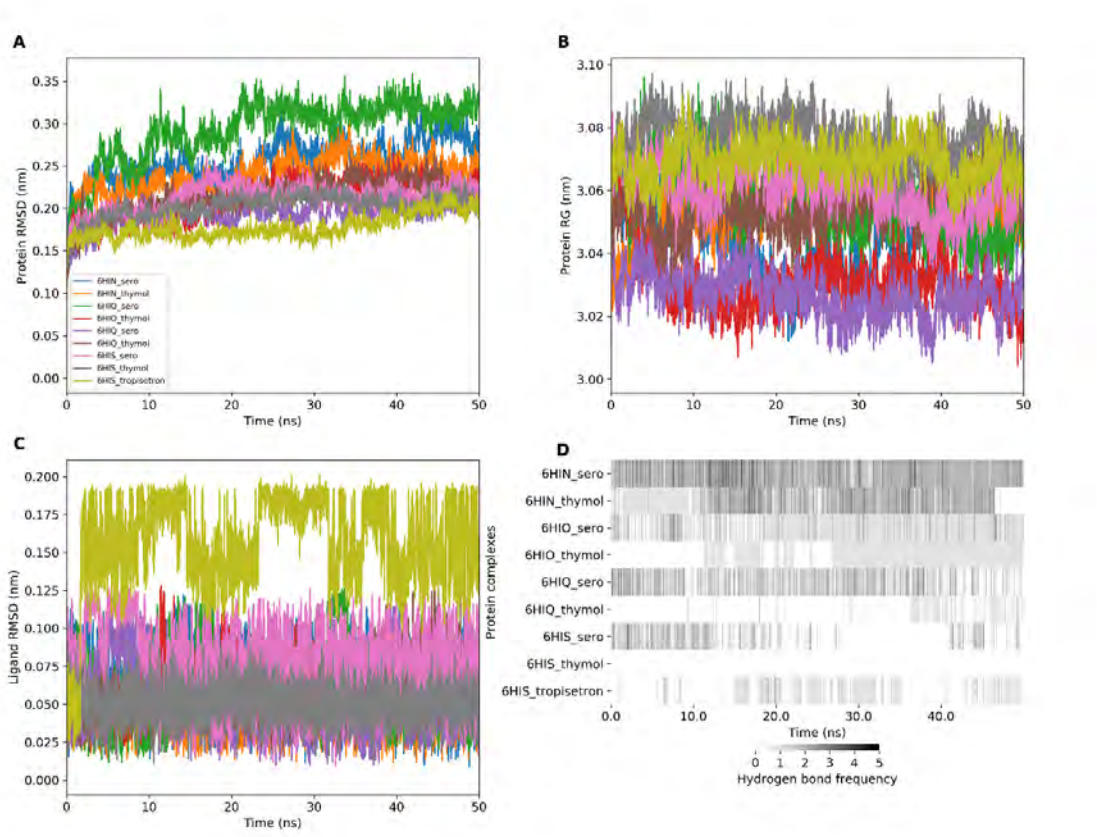


Figure 5-3 Molecular Dynamic simulations. (A) Protein RMSD, (B) Protein radius of gyration (Rg), (C) Ligand RMSD, (D) Hydrogen bond frequency between protein and ligand. Color code for subplots (A), (B), (C) is given in subplot (A). RMSD and Rg values are presented in nanometer (nm) and time in nanosecond (ns).

Dynamics simulation results showed stable complexes with proteins' RMSDs lower or equal to 3.5 Å (Figure 5-3A). Structures are in acceptable range of RMSD to assume no significant global change [477]. Additionally, the different radius of gyration showed no variation maintaining a value of ~3.05 nm (Figure 5-3B). There was no ligand dissociation, and the highest ligand RMSD considering all ligands was ~ 1.25 Å. Only tropisetron in 6HIS had an unexpected increase of RMSD to ~ 2 Å given it was crystalized with the structure. In terms of interactions, serotonin consistently had more hydrogen bonds than thymol and tropisetron (Figure 5-3D). As shown in the interactions in its binding pose, thymol in 6HIS mostly interacted with hydrophobic contacts, hence the absence of hydrogen bonds in MD. Thymol makes more hydrogen bonds in the 6HIN conformation than in all other conformations, even though we observed a decrease toward the end of the simulation. Interestingly, in both 6HIO and 6HIQ, thymol seems to rearrange and adopt a more stable pose, having a more consistent hydrogen bonding toward the end of the simulation (Figure 5-3). This is in accord with the docked pose which mainly includes hydrophobic interactions (Pi-Pi stacking on thymol benzene ring). Thymol in 6HIS does not make any hydrogen bonds, but it is mainly stabilized by hydrophobic contacts, in contrast to tropisetron which showed

~1-2 hydrogen bonds. Hence the underlying molecular mechanism resulting in their similarly induced residues fluctuation might be different. Yet the two compounds interacted with similar residues in their docked poses (Appendix M, Appendix N).

5.3.3 Thymol inducing a different protein behavior, rigidifying the structure.

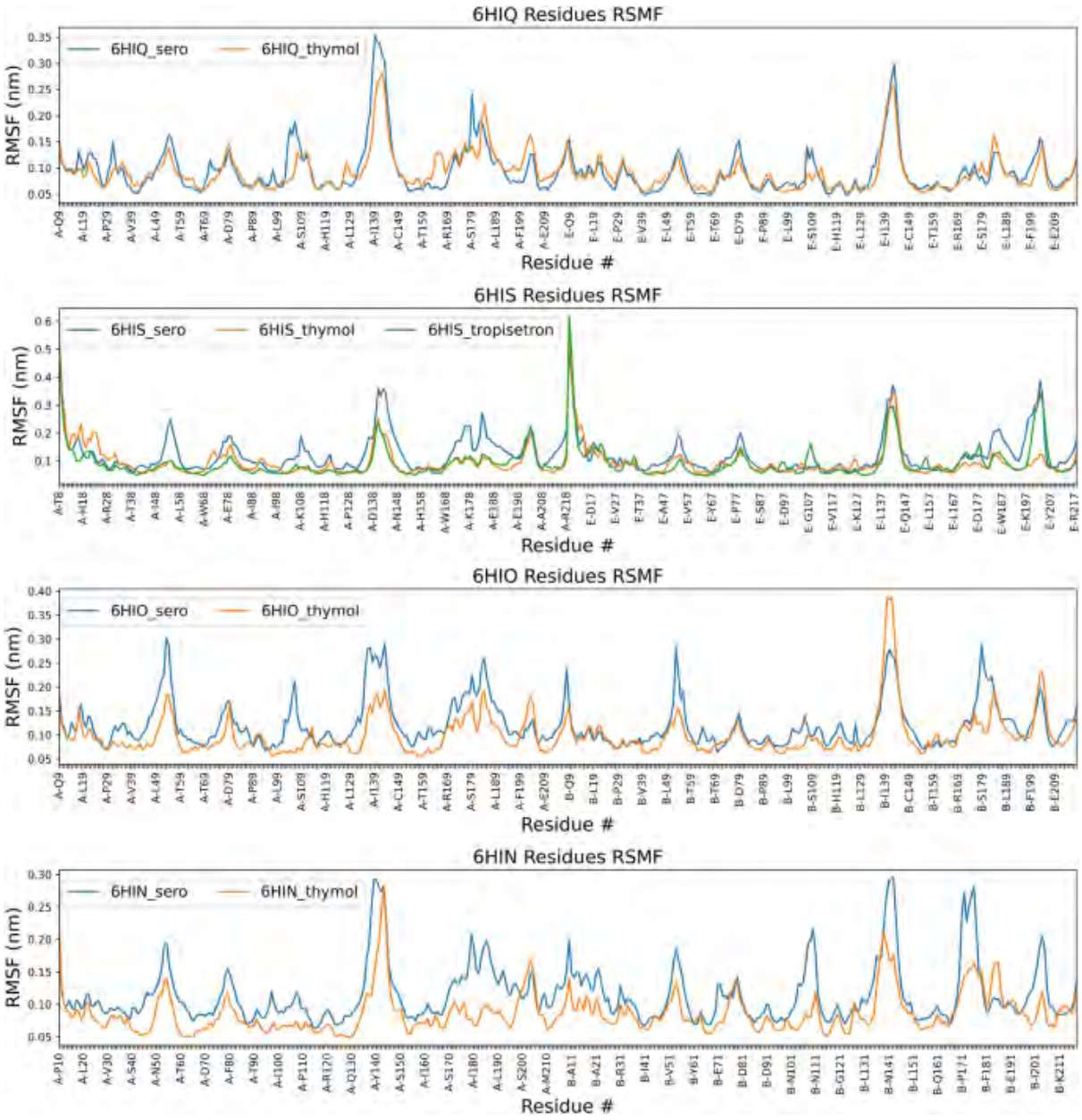


Figure 5-4 5-HT_{3A} ECD RMSF in ligand-bound in the four conformations. Only RMSF values of residues in chains forming the bound-compound binding site are plotted. 5-HT_{3A} is a pentamer with five equivalent binding sites formed at the subunits interfaces in its ECD.

Residues RMSF in MDs revealed different structure behavior depending on bound serotonin or thymol (Figure 5-4). Residues are more flexible with serotonin bound to the structures than thymol and tropisetron. This pattern is especially more accented in 6HIO and 6HIN. Also, in the

inhibited state (6HIS) tropisetron and thymol bound structure have more similar residues fluctuation pattern compared to serotonin. Only in the terminal region around Y207 (CHAIN E) tropisetron and serotonin bound structures have similar behavior that is different from thymol. The main binding residues PHE199, TYR207, TYR126 and TRP63 [462] do not differ in their fluctuations. This higher fluctuation with serotonin may indicate protein reactivating from the inhibited state by serotonin, while this is not the case for thymol and tropisetron.

This work suffers from a number of limitations – notably, only one ligand was used on the entire structure. Yet there are four binding sites on the ECD. Future studies might consider simulating with a ligand-bound on each site on the ECD. This analysis only focused on RMSF of chains forming the binding site. Considering all the chains, with ligand-bound in all four binding sites may provide more insight. More simulation of the protein with the membrane domain may help investigate ion channel activation or inhibition mechanism depending on compound binding. Moreover, in the absence of a rat structure, the mouse one which has was used. However, the mouse sequence (Uniprot ID P23979) had 93.0% sequence identity to the one of the rat one (Uniprot ID P35563). Nevertheless, given 5-HT_{3A}R modulators species specificity [469] it is worth investing a model of the rat sequence.

5.3.4 Thymol had a lower affinity than serotonin and tropisetron.

To further investigate thymol antagonist behavior, its affinity through the PLIE was compared to that one serotonin and tropisetron. Figure 5-5 shows the PLIE for thymol and serotonin in different conformations during the 50 ns.

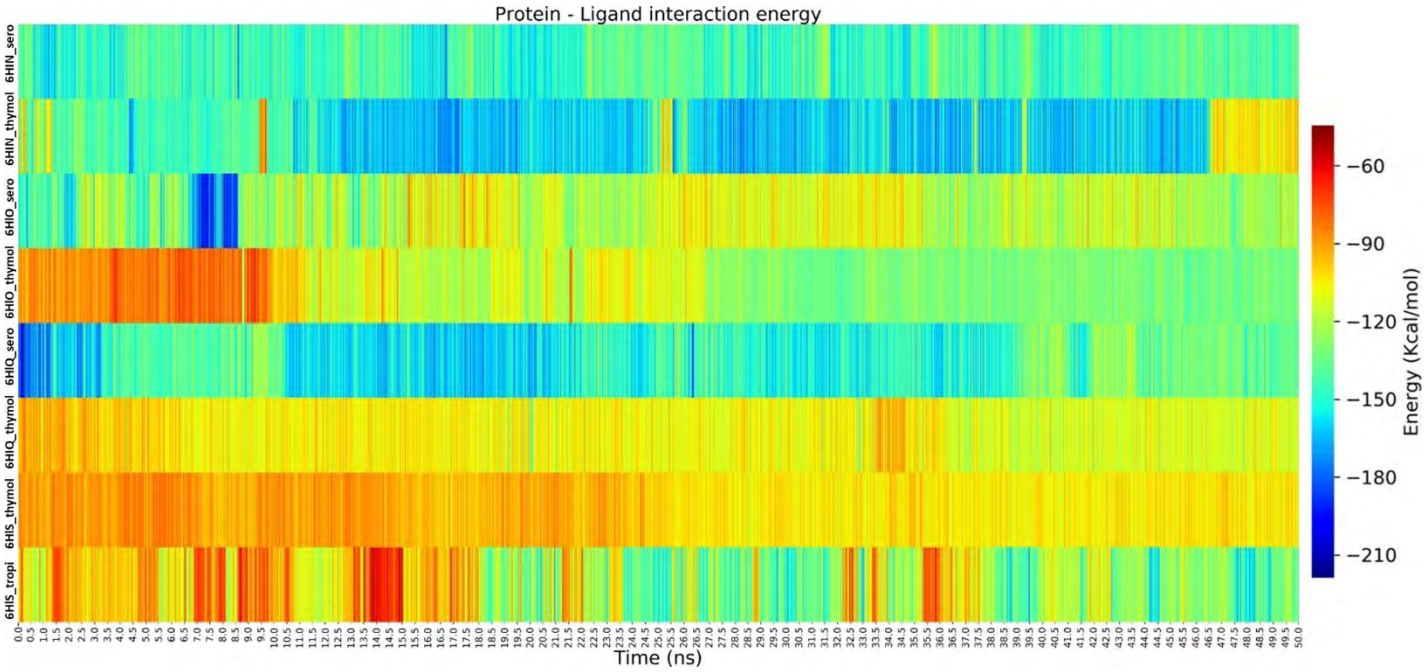


Figure 5-5 PLIE for the following complexes 6HlS_tropi, 6HlS_thymol, 6HIQ_thymol, 6HlQ_sero 6HlO_thymol, 6HlO_sero, 6HIN_thymol, 6HIN_sero during the 50 ns simulation.

The total protein-ligand interaction energy is different from a binding free energy. It quantifies the nonbonded (sum of Coulombic and Lennard-Jones) interaction energies [192]. The total protein-ligand interaction energy is also an indicator of ligand dissociation from the protein which will result in the absence of any energy value. All systems had negative total protein-ligand interaction energy with the averages per system in a range of -97.10 kcal/mol (6HIS_thymol) to -151.39 kcal/mol (6HIN_thymol). The highest energy difference is observed between thymol and serotonin in 6HIQ (-106.43 kcal/mol and -147.44 kcal/mol respectively). This could be related to thymol's movement in that conformation as seen before. The energy difference in other conformations remains lower than 15 kcal/mol. Comparing the ligands' average interaction energies in their respective proteins, serotonin and tropisetron have better energies than thymol in all proteins except in 6HIN in which thymol has a more favorable averaged PLIE. Also, thymol had more favorable PLIE in the last 25 ns in the 6HIO structure.

This thymol lower affinity may be explained by the higher number of hydrogen bonds made by serotonin in docked posed and also in the hydrogen bonding analysis during MD while thymol binding was mostly driven by hydrophobic contacts (Figure 5-3, Appendix N). Actually, in the 6HIN conformation in which thymol makes significant hydrogen bonding, a more favorable affinity is observed (Figure 5-5).

From this energy analysis, thymol may be partially competing with serotonin as this latter had in general a more favorable PLIE. Given it lower PLIE compared to serotonin but similar induced structural behavior to tropisetron, thymol may be acting as a partial antagonist to serotonin. In 6HIS, tropisetron had more favorable PLIE than thymol. This may be expected as tropisetron is co-crystallized ligand, a known antagonist and 6HIS is the inhibited state. A limitation in the current analysis is the absence of PLIE for serotonin in 6HIS. This could provide more insight in the serotonin and thymol comparative behavior in the inhibited state.

5.4 Conclusion

The current data *in vitro* and modelling can support that thymol antagonizes serotonin through 5-HT_3 orthosteric site. Docking results indicate that thymol binds the orthosteric, serotonin binding site, with similar energies to serotonin in all four mouse 5-HT_{3A}R conformations, and with further conservation of that binding mode in MD. These findings make its transmembrane region binding and allosteric mechanism unlikely. MD showed that the two compounds induce different residue fluctuations with thymol inducing a structure behavior similar to tropisetron, a known antagonist. The higher RMSF observed with serotonin in contrast to thymol and tropisetron might result from its agonist mechanism, different from tropisetron and potentially thymol antagonist ones. However, given that thymol has similar binding energies to serotonin but lower compared to tropisetron it may act through a partial antagonist mechanism supported by their significant structural difference.

Yet more evidence is needed in light of the previous literature findings especially compounds species selectivity in 5-HT_3. Further investigation especially on a rat model including the membrane region could help understand the different behavior. Also, a 5-HT_3 enzymatic assay with thymol could be done to fully confirm thymol antagonizing effect.

Chapter 6: Conclusion and future perspectives

In the era of malaria eradication, its chemotherapy remains hampered by drug resistance and hypnozoite reservoirs. Additionally, drug approval rate hampered by the general attrition rate [75] is slower than resistance development from the parasite. *P. falciparum* counts up to 90% of the current disease burden [7]. Its genome high plasticity, diverse resistance mechanisms, potential difficult diagnostic with pfhrp2/3 genes deletion [19] and adaptation to the human host rooted in thousands of years of co-evolution [23,27] turn its eradication into a complex problem. It already showed resistance to all antimalarials classes including the current WHO-recommended ACTs treatment [40]. Given this current situation, stakeholders in antimalarials discovery have emphasized the need for new MoAs defined in the TCPs and also cost-effective strategies to stay ahead of resistance [22]. Despite the highlight of TCPs for antimalarial discovery, the current drug development is marked by the Harlow-Knapp effect, with emphasis on few targets despite other available opportunities [54]. Pipelines are also marked by a high attrition rate despite higher investments, one of the reasons being the lack of proper physicochemical properties in the hunt for potency [76]. This extends to *In silico* approaches which lack a cost-efficient and accurate method to mine the vast chemical space especially illustrated in docking SFs limitation [81]. LEIs may guide the choice of lead compounds and their optimization strategies[478].

Given this context, this work aimed to contribute to computational antimalarial discovery by identifying potential *P. falciparum* inhibitors. Strategies are adapted to the current threats to malaria elimination. Hence, a cost and time effective through drug repurposing combined with a proteome scale screening to find new MoAs. Besides, the work also explores computational screening approaches through the use of efficiency metrics for an accurate pipeline given the limitations of current computational approaches. Hence, an extensive screening pipeline emphasizing consensus scoring is used against PfDXR. This target inhibition can offer a new MoA. Additionally, we present a contribution to the SANCDB library and propose analogs as an alternative to NPs availability.

Using the cost-effective repurposing strategy, four orally available FDA approved drugs (fingolimod, abiraterone, prazosin, and terazosin) with antiplasmodial properties and predicted activities on four different targets are identified. Abiraterone is predicted on a putative liver-stage essential target contributing to covering the different antimalarials TCPs. This part also proposes a proteome scale screening pipeline using a consensus of multiple metrics. A set of 36 *P. falciparum* targets was used. The screening metrics include ligand efficiency and on-the-fly rescoring of docked poses through GRIM. The first metric integrates logP, PSA and MW to avoid drug attrition and the second one contributes to using a consensus of docking scoring SFs. The pipeline also includes normalization and ranking strategy to face scoring biases and reveal as shown in the pipeline evaluation mutually selective protein-ligand complexes. These transformations are beneficial in the context of screening an array of proteins. Yet the predicted targets should be experimentally confirmed.

Given their approved nature, their further exploration (including that of derivatives) may be open to fast-track approval as antimalarials. The integrated pipeline introduced can further be extended with the scoring tools used in Chapter 3: . Additionally, a thorough assessment of the pipeline is possible using a larger set of active/inactive molecules different from the co-crystalized ligand solely. Both ligands and targets sets may be significantly increased for a higher probability of finding more potent hits and alternative MoAs. More target-resistant variants may be included. The identified hits may be further optimized. The pipeline is applicable in other disease areas and may fit complex diseases. It can be a foundation for modelling an *in silico* cell for virtual phenotypic screening.

In the second chapter, a hierarchical virtual screening pipeline combining LBVS and SBVS was used to find hits for PfDXR. This enzyme inhibition will offer an antimalarial with a new MoA. The ZINC lead-like of 3M is used to identify hits with better physicochemical properties for PfDXR inhibition. A more extensive collection of SFs with the philosophy of the "wisdom of the crowd" are used to identify with better-predicted potency than LC5 a nanomolar inhibitor, our baseline. Compounds are further assessed using MD, steered MD, and BFE calculation through MM-PBSA and US. In the end we have identified four lead-like compounds with better-predicted affinity than LC5 a potent nanomolar inhibitor. These hits' scaffolds are different from that of fosmidomycin. To date, most active compounds deposited in ChEMBL are based on that latter which already proved poor physicochemical properties which hampered its usage.

A detailed analysis of interacting residues and the types of interactions revealed GLU233, CYS268, SER270, TRP296, and HIS341 high contributions to Fmax intensity in PfDXR irrespective of the ligand. The same residues were also associated with high contributions to BFE in MM-PBSA.

Analysis of the different LBVS and SFs used revealed that these methods give correlated or distinct rank-ordering of compounds. About the SFs, three groups: the Rf-score group (Rf-score_V1 to V4), (Vina, Idock, and Smina) and (AutoDock, DSX, Cyscore, Xscore) formed different groups. Two main groups: (ES, USR, USRCAT, OBSPEC) showed agreement in LBVS, while RDKit_3pharm and USR had a Kendall tau correlation as low as 0.01. The same trend extends to MD approaches in which affinity evaluation through SMD, MM-PBSA and US did not provide a consistent rank-ordering. We recommend the SMD method given its lower runtime.

As future work, the identified ZINC hits can be validated *in vitro* and their commercially available analogs and/or a derived library may be pursued for better potency. Their lead-like character may predispose them for good optimization. Identified high contributing residue to BFE in MM-PBSA and SMD Fmax may guide these optimizations. SMD and/or US combination with HMR may reduce their cost by two-fold. Further, including PfDXR known inhibitors in SMD, MM-PBSA and US will guide on selecting an accurate BFE method, especially in the context of induced-fit binding and metalloprotein.

Finally, an emphasis was put on NPs given their importance as antimalarials. Literature data were searched to identify NPs isolated in South Africa. The SANCDB library was extended to 1012 NPs, and these were linked to 374,067 commercially readily available analogs in Mcule [395] and Molport [394]. The APIs integration will ensure the links and the constant update with more compounds available. The addition of file formats (AutoDock [456] pdbqt and Schrodinger Maestro [393]) for drug

discovery combined with the availability of the full library for download in these formats can foster drug discovery can ease and speed up pipelines setups. This extended library can contribute to antimalarials given NP already known importance in antimalarials [63]. The analogs may optimize screening hits and may be used for *in vitro* hit validation. The compound classification was automated and standardized. The scaffolds analysis showed database diversity with 501 scaffolds for 1012 compounds and the good potential for drug discovery with a good portion of the database being drug-like or lead-like.

The identification of relevant literature data from scholarly APIs combined with a more exhaustive and thorough text mining for chemical information can help in a larger context such the African one. This can contribute to highlight the continent's ethnobotany heritage and contribute to antimalarials, given that the continent holds about 90% of the disease burden. A docking server can be integrated to the SANCDB website. This server can also integrate SFs used in Chapter 3: for more efficient drug discovery.

Finally, Chapter 5 which departs from the general theme of in silico antimalarial discovery was done as a side project in collaboration with the Department of Pharmaceutical Science and Chinese Traditional Medicine at Southwest University. The chapter investigates the potential mechanism of thymol on serotonin receptor 5-HT_{3A}R after observations of its alleviating effects on IBS symptoms. Thymol binding in 5-HT_{3A}R orthosteric site and similar behavior to tropisetron in RMSF from MD but lower affinity in the protein-ligand interaction energy supports a potential partial antagonist. However, more evidence is needed to support this mechanism, especially in the context of previous literature findings.

REFERENCES

1. Winzeler, E. A. Malaria research in the post-genomic era. *Nature* **455**, 751–756 (2008).
2. Crutcher, J. M. & Hoffman, S. L. Malaria, chapter 83, p 997. *Med. Microbiol. 4th ed. Univ. Texas Med. Branch Galveston, Galveston, TX* (1996).
3. Control, I. of M. (US) C. for the S. on M. P. and, Stanley C. Oaks, J., Mitchell, V. S., Pearson, G. W. & Carpenter, C. C. J. *Parasite Biology*. (National Academies Press (US), 1991).
4. Barber, B. E., Rajahram, G. S., Grigg, M. J., William, T. & Anstey, N. M. World Malaria Report: time to acknowledge Plasmodium knowlesi malaria. *Malar. J.* **16**, (2017).
5. Kyu, H. H. *et al.* Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1859–1922 (2018).
6. Head, M. G. *et al. The allocation of US$105 billion in global funding from G20 countries for infectious disease research between 2000 and 2017: a content analysis of investments*. www.thelancet.com/lancetgh (2020) doi:10.1016/S2214-109X(20)30357-0.
7. World Health Organization. World Malaria Report: 20 years of global progress and challenges. *World Health* **WHO/HTM/GM**, 238 (2020).
8. Sarma, N., Patouillard, † Edith, Cibulskis, R. E. & Arcand, J.-L. The Economic Burden of Malaria: Revisiting the Evidence. *Am. J. Trop. Med. Hyg* **101**, 1405–1415 (2019).
9. Weekly epidemiological record Relevé épidémiologique hebdomadaire.
10. World Malaria Day 2020 - EDCTP. http://www.edctp.org/news/world-malaria-day-2020/.
11. World Health Organisation. Malaria Threat Map. https://apps.who.int/malaria/maps/threats (2020).
12. Cui, L., Mharakurwa, S., Ndiaye, D., Rathod, P. K. & Rosenthal, P. J. Antimalarial Drug Resistance: Literature Review and Activities and Findings of the ICEMR Network. *Am. J. Trop. Med. Hyg.* **93**, 57–68 (2015).
13. Menard, D. & Dondorp, A. Antimalarial drug resistance: a threat to malaria elimination. *Cold Spring Harb. Perspect. Med.* **7**, 1–24 (2017).
14. Cowell, A. N. & Winzeler, E. A. The genomic architecture of antimalarial drug resistance. *Brief. Funct. Genomics* **18**, 314–328 (2019).
15. Hemingway, J. *et al.* Tools and Strategies for Malaria Control and Elimination: What Do We Need to Achieve a Grand Convergence in Malaria? *PLoS Biol.* **14**, (2016).

16. *Report on antimalarial drug efficacy, resistance and response*. (2019).

17. Thu, A. M., Phyo, A. P., Landier, J., Parker, D. M. & Nosten, F. H. Combating multidrug-resistant Plasmodium falciparum malaria. *FEBS J.* **284**, 2569–2578 (2017).

18. Seyfarth, M., Khaireh, B. A., Abdi, A. A., Bouh, S. M. & Faulde, M. K. Five years following first detection of Anopheles stephensi (Diptera: Culicidae) in Djibouti, Horn of Africa: populations established—malaria emerging. *Parasitol. Res.* **118**, 725–732 (2019).

19. Thomson, R. *et al.* Prevalence of plasmodium falciparum lacking histidine-rich proteins 2 and 3: A systematic review. *Bulletin of the World Health Organization* vol. 98 558–568F (2020).

20. Wirth, D. F. The parasite genome: Biological revelations. *Nature* **419**, 495–496 (2002).

21. Morrissette, N. S. & Sibley, L. D. Cytoskeleton of Apicomplexan Parasites. *Microbiol. Mol. Biol. Rev.* **66**, 21–38 (2002).

22. Yahiya, S., Rueda-Zubiaurre, A., Delves, M. J., Fuchter, M. J. & Baum, J. The antimalarial screening landscape—looking beyond the asexual blood stage. *Curr. Opin. Chem. Biol.* **50**, 1–9 (2019).

23. Cowman, A. F., Healer, J., Marapana, D. & Marsh, K. Malaria: Biology and Disease. *Cell* **167**, 610–624 (2016).

24. Soulard, V. *et al.* Plasmodium falciparum full life cycle and Plasmodium ovale liver stages in humanized mice. *Nat. Commun.* **6**, 7690 (2015).

25. Campo, B., Vandal, O., Wesche, D. L. & Burrows, J. N. Killing the hypnozoite – drug discovery approaches to prevent relapse in Plasmodium vivax. *Pathog. Glob. Health* **109**, 107–122 (2015).

26. Biamonte, M. A., Wanner, J. & Le Roch, K. G. Recent advances in malaria drug discovery. *Bioorg. Med. Chem. Lett.* **23**, 2829–2843 (2013).

27. Antinori, S., Galimberti, L., Milazzo, L. & Corbellino, M. Biology of Human Malaria Plasmodia Including Plasmodium Knowlesi. *Mediterr. J. Hematol. Infect. Dis.* **4**, (2012).

28. Bennink, S., Kiesow, M. J. & Pradel, G. The development of malaria parasites in the mosquito midgut. *Cell. Microbiol.* **18**, 905–918 (2016).

29. White, N. Antimalarial drug resistance and combination chemotherapy. *Philos. Trans. R. Soc. B Biol. Sci.* **354**, 739–749 (1999).

30. Tse, E. G., Korsik, M. & Todd, M. H. The past, present and future of anti-malarial medicines. *Malaria Journal* vol. 18 1–21 (2019).

31. Burrows, J. N. *et al.* New developments in anti-malarial target candidate and product profiles. *Malar J* **16**, 26 (2017).

32. Carolino, K. & Winzeler, E. A. The antimalarial resistome – finding new drug targets and

their modes of action. *Curr. Opin. Microbiol.* **57**, 49–55 (2020).

33. Gaur, A. H. *et al.* Safety, tolerability, pharmacokinetics, and antimalarial efficacy of a novel Plasmodium falciparum ATP4 inhibitor SJ733: a first-in-human and induced blood-stage malaria phase 1a/b trial. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(19)30611-5.

34. Hooft van Huijsduijnen, R. & Wells, T. N. The antimalarial pipeline. *Current Opinion in Pharmacology* vol. 42 1–6 (2018).

35. Ashley, E. A. & Phyo, A. P. Drugs in Development for Malaria. *Drugs* **78**, 861–879 (2018).

36. Kublin, J. G. *et al.* Safety, Pharmacokinetics, and Causal Prophylactic Efficacy of KAF156 in a Plasmodium falciparum Human Infection Study. *Clin. Infect. Dis.* (2020) doi:10.1093/cid/ciaa952.

37. Technologies, T. malERA R. C. P. on B. S. and E. malERA: An updated research agenda for basic science and enabling technologies in malaria elimination and eradication. *PLOS Med.* **14**, e1002451 (2017).

38. Vora, P., Somani, R. & Jain, M. Drug Repositioning: An Approach for Drug Discovery. *Mini. Rev. Org. Chem.* **13**, 363–376 (2016).

39. Fontinha, D., Moules, I. & Prudêncio, M. Repurposing drugs to fight hepatic malaria parasites. *Molecules* **25**, 3409 (2020).

40. Verlinden, B. K., Louw, A. & Birkholtz, L.-M. Resisting resistance: is there a solution for malaria? *Expert Opin. Drug Discov.* **11**, 395–406 (2016).

41. Hopkins, A. L., Mason, J. S. & Overington, J. P. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **16**, 127–136 (2006).

42. Tibon, N. S., Ng, C. H. & Cheong, S. L. Current progress in antimalarial pharmacotherapy and multi-target drug discovery. *European Journal of Medicinal Chemistry* vol. 188 111983 (2020).

43. Mushtaque, M. & Shahjahan. Reemergence of chloroquine (CQ) analogs as multi-targeting antimalarial agents: A review. *European Journal of Medicinal Chemistry* vol. 90 280–295 (2015).

44. Chen, W. *et al.* Novel dual inhibitors against FP-2 and PfDHFR as potential antimalarial agents: Design, synthesis and biological evaluation. *Chinese Chem. Lett.* **30**, 250–254 (2019).

45. Dickerman, B. K. *et al.* Identification of inhibitors that dually target the new permeability pathway and dihydroorotate dehydrogenase in the blood stage of Plasmodium falciparum. *Sci. Rep.* **6**, 37502–37502 (2016).

46. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. *Diagnosing the decline in pharmaceutical R&D efficiency*. www.nature.com/reviews/drugdisc (2012) doi:10.1038/nrd3681.

47. Schenone, M., Dančík, V., Wagner, B. K. & Clemons, P. A. Target identification and mechanism of action in chemical biology and drug discovery. *Nature Chemical Biology* vol. 9 232–240 (2013).

48. White, J. & Rathod, P. K. Indispensable malaria genes. *Science (80-.).* **360**, 490–491 (2018).

49. Gomes, A. R. *et al.* A genome-scale vector resource enables high-throughput reverse genetic screening in a malaria parasite. *Cell Host Microbe* **17**, 404–413 (2015).

50. Burrows, J. N. *et al.* Antimalarial drug discovery - the path towards eradication. *Parasitology* **141**, 128–39 (2014).

51. Aneja, B., Kumar, B., Jairajpuri, M. A. & Abid, M. A structure guided drug-discovery approach towards identification of Plasmodium inhibitors. *RSC Adv.* **6**, 18364–18406 (2016).

52. Yeh, I., Hanekamp, T., Tsoka, S., Karp, P. D. & Altman, R. B. Computational analysis of Plasmodium falciparum metabolism: Organizing genomic information to facilitate drug discovery. *Genome Res.* **14**, 917–924 (2004).

53. Zhang, M. *et al.* Uncovering the essential genes of the human malaria parasite Plasmodium falciparum by saturation mutagenesis. *Science (80-.).* **360**, (2018).

54. Lunev, S., Batista, F. A., Bosch, S. S., Wrenger, C. & Groves, M. R. *Identification and Validation of Novel Drug Targets for the Treatment of Plasmodium falciparum Malaria: New Insights*. *Current Topics in Malaria* (InTech, 2016). doi:10.5772/65659.

55. Lunev, S., Batista, F. A., Bosch, S. S., Wrenger, C. & Groves, M. R. Identification and Validation of Novel Drug Targets for the Treatment of Plasmodium falciparum Malaria: New Insights. *Curr. Top. Malar.* (2016) doi:10.5772/65659.

56. Murkin, A. S., Manning, K. A. & Kholodar, S. A. Mechanism and inhibition of 1-deoxy-D-xylulose-5-phosphate reductoisomerase. *Bioorg. Chem.* **57**, 171–185 (2014).

57. Uddin, T. Drug targets in the apicoplast of malaria parasites. (2017).

58. Armstrong, C. M., Meyers, D. J., Imlay, L. S., Meyers, C. F. & Odom, A. R. Resistance to the antimicrobial agent fosmidomycin and an FR900098 prodrug through mutations in the deoxyxylulose phosphate reductoisomerase gene (dxr). *Antimicrob. Agents Chemother.* **59**, 5511–5519 (2015).

59. Mombo-Ngoma, G. *et al.* Efficacy and Safety of Fosmidomycin-Piperaquine as Nonartemisinin-Based Combination Therapy for Uncomplicated Falciparum Malaria: A Single-Arm, Age De-escalation Proof-of-Concept Study in Gabon. *Clin. Infect. Dis. Fosmidomycin-Piperaquine as Malar. NACT • CID* **2018**, 1823.

60. Friedman, R. & Caflisch, A. Discovery of Plasmepsin Inhibitors by Fragment-Based Docking and Consensus Scoring. *ChemMedChem* **4**, 1317–1326 (2009).

61. Belete, T. M. Recent progress in the development of new antimalarial drugs with novel

targets. *Drug Des. Devel. Ther.* (2020) doi:10.2147/DDDT.S265602.

62. Tambo, E., Khater, E. I. M., Chen, J. H., Bergquist, R. & Zhou, X. N. Nobel prize for the artemisinin and ivermectin discoveries: A great boost towards elimination of the global infectious diseases of poverty. *Infectious Diseases of Poverty* vol. 4 58 (2015).

63. Tajuddeen, N. & Van Heerden, F. R. Antiplasmodial natural products: An update. *Malar. J.* **18**, (2019).

64. Ntie-Kang, F. *et al.* CamMedNP: Building the Cameroonian 3D structural natural products database for virtual screening. *BMC Complement. Altern. Med.* **13**, 88 (2013).

65. Hatherley, R. *et al.* SANCDB: A South African natural compound database. *J. Cheminform.* **7**, (2015).

66. Amoa Onguéné, P. *et al.* The potential of anti-malarial compounds derived from African medicinal plants, part I: a pharmacological evaluation of alkaloids and terpenoids. *Malar. J.* **12**, 449 (2013).

67. Fabricant, D. S. & Farnsworth, N. R. The value of plants used in traditional medicine for drug discovery. *Environ. Health Perspect.* **109 Suppl**, 69–75 (2001).

68. Wells, T. N. C. Natural products as starting points for future anti-malarial therapies: going back to our roots? *Malar. J.* **10 Suppl 1**, S3 (2011).

69. Mboya-Okeyo, T., Ridley, R. G., Nwaka, S. & ANDI Task Force. The African Network for Drugs and Diagnostics Innovation. *Lancet* **373**, 1507–1508 (2009).

70. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **79**, 629–661 (2016).

71. Wells, T. N. C., Van Huijsduijnen, R. H. & Van Voorhis, W. C. Malaria medicines: A glass half full? *Nat. Rev. Drug Discov.* **14**, 424–442 (2015).

72. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).

73. Prieto-Martínez, F. D., López-López, E., Eurídice Juárez-Mercado, K. & Medina-Franco, J. L. Computational Drug Design Methods—Current and Future Perspectives. *Silico Drug Des.* 19–44 (2019) doi:10.1016/b978-0-12-816125-8.00002-x.

74. Salim, N. O., Azian, N., Yusuf, M., Adyani, F. & Fuad, A. *Plasmodial enzymes in metabolic pathways as therapeutic targets and contemporary strategies to discover new antimalarial drugs: a review*. *AsPac J. Mol. Biol. Biotechnol* vol. 27 (2019).

75. Waring, M. J. *et al.* An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discov.* **14**, 475–486 (2015).

76. Hann, M. M. Molecular obesity, potency and other addictions in drug discovery. doi:10.1039/c1md00017a.

77. Nicolaou, C. A., Brown, N. & Pattichis, C. S. Molecular optimization using computational multi-objective methods. *Curr. Opin. Drug Discov. Devel.* **10**, 316–324 (2007).

78. Johnson, M. A., Maggiora, G. M. & American Chemical Society. Meeting (196th : 1988 : Los Angeles, C. . *Concepts and applications of molecular similarity*. (Wiley, 1990).

79. Vázquez, J. *et al.* molecules Merging Ligand-Based and Structure-Based Methods in Drug Discovery: An Overview of Combined Virtual Screening Approaches. doi:10.3390/molecules25204723.

80. Cleves, A. E. & Jain, A. N. Structure- And ligand-based virtual screening on DUD-E+: Performance dependence on approximations to the binding pocket. *J. Chem. Inf. Model.* **60**, 4296–4310 (2020).

81. Gaieb, Z. *et al.* D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided. Mol. Des.* **32**, 1–20 (2018).

82. Baumgartner, M. P. & Evans, D. A. Lessons learned in induced fit docking and metadynamics in the Drug Design Data Resource Grand Challenge 2. *J. Comput. Aided. Mol. Des.* **32**, 45–58 (2018).

83. Probst, D. & Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *J. Cheminform.* **10**, 66 (2018).

84. Irwin, J. J. *et al.* ZINC20A Free Ultralarge-Scale Chemical Database for Ligand Discovery. doi:10.1021/acs.jcim.0c00675.

85. Grygorenko, O. O. *et al.* Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **23**, 101681 (2020).

86. NIH Virtual Workshop on Ultra-Large Chemistry Databases, Dec 1-3, 2020. https://cactus.nci.nih.gov/presentations/NIHBigDB_2020-12/NIHBigDB.html.

87. Lionta, E., Spyrou, G., Vassilatis, D. K. & Cournia, Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.* **14**, 1923–38 (2014).

88. Maia, E. H. B., Assis, L. C., de Oliveira, T. A., da Silva, A. M. & Taranto, A. G. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Frontiers in Chemistry* vol. 8 (2020).

89. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269–288 (1982).

90. Morris, G. M. *et al.* Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **19**, 1639–1662 (1998).

91. Chen, H.-M., Liu, B.-F., Huang, H.-L., Hwang, S.-F. & Ho, S.-Y. SODOCK: Swarm optimization for highly flexible protein–ligand docking. *J. Comput. Chem.* **28**, 612–623 (2007).

92. Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A fast flexible docking method using an

incremental construction algorithm. *J. Mol. Biol.* **261**, 470–489 (1996).

93. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–61 (2010).

94. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).

95. Sethi, A., Joshi, K., Sasikala, K. & Alvala, M. Molecular Docking in Modern Drug Discovery: Principles and Recent Applications. in *Drug Discovery and Development - New Advances* (IntechOpen, 2020). doi:10.5772/intechopen.85991.

96. Hassan, N. M., Alhossary, A. A., Mu, Y. & Kwoh, C. K. Protein-Ligand Blind Docking Using QuickVina-W with Inter-Process Spatio-Temporal Integration. *Sci. Rep.* **7**, 15451 (2017).

97. Gaieb, Z. *et al.* D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings. *J. Comput. Aided. Mol. Des.* (2019) doi:10.1007/s10822-018-0180-4.

98. LeGrand, S. *et al.* GPU-Accelerated Drug Discovery with Docking on the Summit Supercomputer: Porting, Opti-mization, and Application to COVID-19 Research. 10 (2020) doi:10.1145/3388440.3412472.

99. GigaDocking™ - Structure Based Virtual Screening of Over 1 Billion Molecules Webinar. https://www.eyesopen.com/webinars/giga-docking-structure-based-virtual-screening.

100. Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).

101. Pantsar, T. & Poso, A. Binding affinity via docking: Fact and fiction. *Molecules* vol. 23 1DUMMY (2018).

102. Śledź, P. & Caflisch, A. Protein structure-based drug design: from docking to molecular dynamics. *Curr. Opin. Struct. Biol.* **48**, 93–102 (2018).

103. De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **59**, 4035–4061 (2016).

104. Petrenko, R. & Meller, J. Molecular Dynamics. in *eLS* (John Wiley & Sons, Ltd, 2001).

105. Allen, M. P. & others. Introduction to molecular dynamics simulation. *Comput. soft matter from Synth. Polym. to proteins* **23**, 1–28 (2004).

106. Lemkul, J. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]. *Living J. Comput. Mol. Sci.* **1**, 0–53 (2019).

107. Fusani, L., Palmer, D. S., Somers, D. O. & Wall, I. D. Exploring Ligand Stability in Protein Crystal Structures Using Binding Pose Metadynamics. *J. Chem. Inf. Model.* **60**, 1528–1539 (2020).

108. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding

affinities. *Expert Opin. Drug Discov.* **10**, 449–461 (2015).

109. Jing, Z. *et al.* Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. (2019) doi:10.1146/annurev-biophys-070317.

110. Vanommeslaeghe, K., Guvench, O. & MacKerell, A. D. Molecular Mechanics. *Curr. Pharm. Des.* **20**, 3281–3292 (2014).

111. MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).

112. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**–**2**, 19–25 (2015).

113. González, M. A. Force fields and molecular dynamics simulations. *École thématique la Société Française la Neutron.* **12**, 169–200 (2011).

114. Duan, Y. *et al.* A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24**, 1999–2012 (2003).

115. Ganesan, A., Coote, M. L. & Barakat, K. Molecular dynamics-driven drug discovery: leaping forward with confidence. *Drug Discov. Today* **22**, 249–269 (2017).

116. Neves, R. P. P., Sousa, S. F., Fernandes, P. A. & Ramos, M. J. Parameters for Molecular Dynamics Simulations of Manganese-Containing Metalloproteins. *J. Chem. Theory Comput.* **9**, 2718–2732 (2013).

117. Hu, L. & Ryde, U. Comparison of Methods to Obtain Force-Field Parameters for Metal Sites. **7**, 2452–2463.

118. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comput.* **11**, 1864–1874 (2015).

119. Lemkul, J. A. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package, v1.0. 1–52 (2018).

120. Huang, J., Lemkul, J. A., Eastman, P. K. & MacKerell, A. D. Molecular dynamics simulations using the drude polarizable force field on GPUs with OpenMM: Implementation, validation, and benchmarks. *J. Comput. Chem.* **39**, 1682–1689 (2018).

121. Mei, Z. *et al. Current MD forcefields fail to capture key features of protein structure and fluctuations: A case study of cyclophilin A and T4 lysozyme*.

122. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. doi:10.1038/s41467-018-06169-2.

123. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10**, 449–61 (2015).

124. Wang, E. *et al.* End-Point Binding Free Energy Calculation with MM/PBSA and MM/ GBSA: Strategies and Applications in Drug Design. (2019) doi:10.1021/acs.chemrev.9b00055.

125. Poli, G., Granchi, C., Rizzolio, F. & Tuccinardi, T. Application of MM-PBSA Methods in Virtual Screening. *Molecules* **25**, 1971 (2020).

126. Qi, R., Botello-Smith, W. M. & Luo, R. Acceleration of Linear Finite-Difference Poisson-Boltzmann Methods on Graphics Processing Units. *J. Chem. Theory Comput.* **13**, 3378–3387 (2017).

127. Terayama, K., Iwata, H., Araki, M., Okuno, Y. & Tsuda, K. Machine learning accelerates MD-based binding pose prediction between ligands and proteins. *Bioinformatics* **34**, 770–778 (2018).

128. Cournia, Z., Allen, B. & Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* (2017) doi:10.1021/acs.jcim.7b00564.

129. Chipot, C., Shell, M. S. & Pohorille, A. Springer Series in Chemical Physics: Introduction. *Springer Series in Chemical Physics* vol. 86 1–31 (2007).

130. Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13**, 1011–1021 (1992).

131. Yang, Y., Pan, L., Lightstone, F. C. & Merz, K. M. The Role of Molecular Dynamics Potential of Mean Force Calculations in the Investigation of Enzyme Catalysis. in *Methods in Enzymology* vol. 577 1–29 (Academic Press Inc., 2016).

132. Patel, J. S. & Ytreberg, F. M. Fast Calculation of Protein-Protein Binding Free Energies Using Umbrella Sampling with a Coarse-Grained Model. *J. Chem. Theory Comput.* **14**, 991–997 (2018).

133. Yahiya, S., Rueda-Zubiaurre, A., Delves, M. J., Fuchter, M. J. & Baum, J. The antimalarial screening landscape—looking beyond the asexual blood stage. *Curr. Opin. Chem. Biol.* **50**, 1–9 (2019).

134. Mathews, E. S. & Odom John, A. R. Tackling resistance: emerging antimalarials and new parasite targets in the era of elimination. *F1000Research* **7**, 1170 (2018).

135. Murithi, J. M. *et al.* Combining Stage Specificity and Metabolomic Profiling to Advance Antimalarial Drug Discovery. *Cell Chem. Biol.* **27**, 158–171.e3 (2020).

136. Anurak, C. & Kesara, N.-B. A systematic review: Application of in silico models for antimalarial drug discovery. *African J. Pharm. Pharmacol.* **12**, 159–167 (2018).

137. Kushwaha, P. P., Vardhan, P. S., Kumari, P., Mtewa, A. G. & Kumar, S. Bioactive lead compounds and targets for the development of antimalarial drugs. in *Phytochemicals as Lead Compounds for New Drug Discovery* 305–316 (Elsevier, 2019). doi:10.1016/B978-0-12-817890-4.00020-2.

138. Sahu, S. *et al.* In silico ADMET study, docking, synthesis and antimalarial evaluation of thiazole-1,3,5-triazine derivatives as Pf-DHFR inhibitor. *Pharmacol. Reports* **71**, 762–767 (2019).

139. Arshadi, A. K., Salem, M., Collins, J., Yuan, J. S. & Chakrabarti, D. Deepmalaria: Artificial intelligence driven discovery of potent antiplasmodials. *Front. Pharmacol.* **10**, (2020).

140. Goodsell, D. S. *et al.* RCSB Protein Data Bank: Enabling biomedical research and drug discovery. *Protein Sci.* **29**, 52–65 (2020).

141. Blasco, B., Leroy, Di. & Fidock, D. A. Antimalarial drug resistance: Linking Plasmodium falciparum parasite biology to the clinic. *Nature Medicine* vol. 23 917–928 (2017).

142. Li, Y. Y., An, J. & Jones, S. J. M. A computational approach to finding novel targets for existing drugs. *PLoS Comput. Biol.* **7**, e1002139 (2011).

143. Karaman, M. W. *et al.* A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **26**, 127–132 (2008).

144. Kasam, V. *et al.* WISDOM-II: Screening against multiple targets implicated in malaria using computational grid infrastructures. *Malar. J.* **8**, 88 (2009).

145. Frampton, J. E. Tafenoquine: first global approval. *Drugs* **78**, 1517–1523 (2018).

146. World Health Organization. *WHO Malaria report 2019*. *Malaria report 2019* https://www.who.int/publications-detail/world-malaria-report-2019 (2019).

147. Jacq, N. *et al.* Grid-enabled virtual screening against malaria. *J. Grid Comput.* **6**, 29–43 (2008).

148. Negi, A., Bhandari, N., Shyamlal, B. R. K. & Chaudhary, S. Inverse docking based screening and identification of protein targets for Cassiarin alkaloids against Plasmodium falciparum. *Saudi Pharm. J.* **26**, 546–567 (2018).

149. Andrews, K. T., Fisher, G. & Skinner-Adams, T. S. Drug repurposing and human parasitic protozoan diseases. *Int. J. Parasitol. Drugs Drug Resist.* **4**, 95–111 (2014).

150. Lantero, E., Aláez-Versón, C. R., Romero, P., Sierra, T. & Fernàndez-Busquets, X. Repurposing Heparin as Antimalarial: Evaluation of Multiple Modifications Toward In Vivo Application. *Pharmaceutics* **12**, 825 (2020).

151. Ramakrishnan, G., Chandra, N. & Srinivasan, N. Exploring anti-malarial potential of FDA approved drugs: An in silico approach. *Malar. J.* **16**, 290 (2017).

152. Álvarez-Carretero, S., Pavlopoulou, N., Adams, J., Gilsenan, J. & Tabernero, L. VSpipe, an Integrated Resource for Virtual Screening and Hit Selection: Applications to Protein Tyrosine Phospahatase Inhibition. *Molecules* **23**, 353 (2018).

153. Meirson, T., Samson, A. O. & Gil-Henn, H. An in silico high-throughput screen identifies potential selective inhibitors for the non-receptor tyrosine kinase Pyk2. *Drug Des. Devel. Ther.* **11**, 1535–1557 (2017).

154. Rognan, D. Proteome-scale docking: myth and reality. *Drug Discov. Today Technol.* **10**, e403–e409 (2013).

155. Kellenberger, E. *et al.* sc-PDB: An annotated database of druggable binding sites from the Protein Data Bank. *J. Chem. Inf. Model.* **46**, 717–727 (2006).

156. Desaphy, J., Bret, G., Rognan, D. & Kellenberger, E. Sc-PDB: A 3D-database of ligandable binding sites-10 years on. *Nucleic Acids Res.* **43**, D399–D404 (2015).

157. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672 (2006).

158. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).

159. Jacobson, M. P. *et al.* A Hierarchical Approach to All-Atom Protein Loop Prediction. *Proteins Struct. Funct. Genet.* **55**, 351–367 (2004).

160. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).

161. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).

162. Landrum, G. RDKit: open-source cheminformatics software. (2016).

163. Wildman, S. A. & Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999).

164. Gathiaka, S. *et al.* D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions. *J. Comput. Aided. Mol. Des.* **30**, 651–668 (2016).

165. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, NA-NA (2009).

166. Jaghoori, M. M., Bleijlevens, B. & Olabarriaga, S. D. 1001 Ways to run AutoDock Vina for virtual screening. *J. Comput. Aided. Mol. Des.* **30**, 237–249 (2016).

167. Vigers, G. P. A. & Rizzi, J. P. Multiple Active Site Corrections for Docking and Virtual Screening. *J. Med. Chem.* **47**, 80–89 (2004).

168. García-Sosa, A. T., Hetényi, C. & Maran, U. K. O. Drug efficiency indices for improvement of molecular docking scoring functions. *J. Comput. Chem.* **31**, 174–184 (2010).

169. Leeson, P. D. & Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* **6**, 881–890 (2007).

170. Abad-Zapatero, C. Ligand efficiency indices for effective drug discovery. *Expert Opin. Drug Discov.* **2**, 469–488 (2007).

171. Mignani, S. *et al.* Present drug-likeness filters in medicinal chemistry during the hit and lead

optimization process: how far can they be simplified? *Drug Discov. Today* **23**, 605–615 (2018).

172. Freeman-Cook, K. D., Hoffman, R. L. & Johnson, T. W. Lipophilic efficiency: the most important efficiency metric in medicinal chemistry. *Future Med. Chem.* **5**, 113–115 (2013).

173. Cortes-Cabrera, A., Morreale, A., Gago, F. & Abad-Zapatero, C. AtlasCBS: A web server to map and explore chemico-biological space. *J. Comput. Aided. Mol. Des.* **26**, 995–1003 (2012).

174. Abad-Zapatero, C., Champness, E. J. & Segall, M. D. Alternative variables in drug discovery: promises and challenges. *Future Med. Chem.* **6**, 577–593 (2014).

175. Arnott, J. A., Kumar, R. & Planey, S. L. *Lipophilicity Indices for Drug Development*. *Journal of Applied Biopharmaceutics and Pharmacokinetics* (2013).

176. Luo, Q. *et al.* The scoring bias in reverse docking and the score normalization strategy to improve success rate of target fishing. *PLoS One* **12**, e0171433 (2017).

177. Jacobsson, M. & Karlén, A. Ligand bias of scoring functions in structure-based virtual screening. *J. Chem. Inf. Model.* **46**, 1334–1343 (2006).

178. Fukunishi, Y., Kubota, S. & Nakamura, H. Noise reduction method for molecular interaction energy: Application to in silico drug screening and in silico target protein screening. *J. Chem. Inf. Model.* **46**, 2071–2084 (2006).

179. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

180. Shityakov, S. & Förster, C. In silico predictive model to determine vector-mediated transport properties for the blood-brain barrier choline transporter. *Adv. Appl. Bioinforma. Chem.* **7**, 23–36 (2014).

181. Desaphy, J., Raimbaud, E., Ducrot, P. & Rognan, D. Encoding protein-ligand interaction patterns in fingerprints and graphs. *J. Chem. Inf. Model.* **53**, 623–637 (2013).

182. Pronk, S. *et al.* GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).

183. Pines, G. *et al.* Genomic Deoxyxylulose Phosphate Reductoisomerase (DXR) Mutations Conferring Resistance to the Antimalarial Drug Fosmidomycin in E. coli. *ACS Synth. Biol.* **7**, 2824–2832 (2018).

184. Maláč, K. & Barvík, I. Complex between Human RNase HI and the phosphonate-DNA/RNA duplex: Molecular dynamics study. *J. Mol. Graph. Model.* **44**, 81–90 (2013).

185. Gu, S. *et al.* Phosphoantigen-induced conformational change of butyrophilin 3A1 (BTN3A1) and its implication on Vγ9Vδ2 T cell activation. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E7311–E7320 (2017).

186. Sousa da Silva, A. W. & Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE.

BMC Res. Notes **5**, 367 (2012).

187. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–8 (2010).

188. Nguyen, H., Case, D. A. & Rose, A. S. NGLview–interactive molecular graphics for Jupyter notebooks. *Bioinformatics* **34**, 1241–1242 (2018).

189. Nguyen, H., Roe, D. R., Swails, J. & Case, D. A. PYTRAJ: Interactive data analysis for molecular dynamics simulations. *New Brunswick, NJ Rutgers Univ.* (2016).

190. Kluyver, T. *et al.* Jupyter Notebooks-a publishing format for reproducible computational workflows. in *ELPUB* 87–90 (2016).

191. Lobanov, M. Y., Bogatyreva, N. S. & Galzitskaya, O. V. Radius of gyration as an indicator of protein structure compactness. *Mol. Biol.* **42**, 623–628 (2008).

192. Lemkul, J. A. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]. *Living J. Comput. Mol. Sci.* **1**, 0–53 (2019).

193. Lunga, M. J. *et al.* Expanding the SAR of Nontoxic Antiplasmodial Indolyl-3-ethanone Ethers and Thioethers. *ChemMedChem* **13**, 1353–1362 (2018).

194. Makler, M. T. & Hinrichs, D. J. Measurement of the lactate dehydrogenase activity of Plasmodium falciparum as an assessment of parasitemia. *Am. J. Trop. Med. Hyg.* **48**, 205–210 (1993).

195. Borra, R. C., Lotufo, M. A., Gagioti, S. M., Barros, F. de M. & Andrade, P. M. A simple method to measure cell viability in proliferation and cytotoxicity assays. *Braz. Oral Res.* **23**, 255–262 (2009).

196. Riss, T. L. *et al.* Cell viability assays. in *Assay Guidance Manual [Internet]* (Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2016).

197. Li, Y., Han, L., Liu, Z. & Wang, R. Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *J. Chem. Inf. Model.* **54**, 1717–1736 (2014).

198. seaborn: statistical data visualization — seaborn 0.11.0 documentation. https://seaborn.pydata.org/.

199. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).

200. Koebel, M. R., Schmadeke, G., Posner, R. G. & Sirimulla, S. AutoDock VinaXB: Implementation of XBSF, new empirical halogen bond scoring function, into AutoDock Vina. *J. Cheminform.* **8**, 27 (2016).

201. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom

structure validation. *Protein Sci.* **27**, 293–315 (2018).

202. Robien, M. A. *et al.* Crystal structure of glyceraldehyde-3-phosphate dehydrogenase from Plasmodium falciparum at 2.25 Å resolution reveals intriguing extra electron density in the active site. *Proteins Struct. Funct. Bioinforma.* **62**, 570–577 (2005).

203. Nasamu, A. S., Polino, A. J., Istvan, E. S. & Goldberg, D. E. Malaria parasite plasmepsins: More than just plain old degradative pepsins. *J. Biol. Chem.* **295**, 8425–8441 (2020).

204. Ash, J. & Fourches, D. Characterizing the Chemical Space of ERK2 Kinase Inhibitors Using Descriptors Computed from Molecular Dynamics Trajectories. *J. Chem. Inf. Model.* **57**, 1286–1299 (2017).

205. Bhaumik, P. *et al.* Structural insights into the activation and inhibition of histo-aspartic protease from Plasmodium falciparum. *Biochemistry* **50**, 8862–79 (2011).

206. Oliver, J. C., Linger, R. S., Chittur, S. V. & Davisson, V. J. Substrate activation and conformational dynamics of guanosine 5'-monophosphate synthetase. *Biochemistry* **52**, 5225–5235 (2013).

207. Ballut, L. *et al.* Active site coupling in Plasmodium falciparum GMP synthetase is triggered by domain rotation. *Nat. Commun.* **6**, 1–13 (2015).

208. Bhat, T. N. *et al.* The PDB data uniformity project. *Nucleic Acids Res.* **29**, 214–218 (2001).

209. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).

210. Perez-Castillo, Y. *et al.* CompScore: boosting structure-based virtual screening performance by incorporating docking scoring functions components into consensus scoring. doi:10.1101/550590.

211. Palacio-Rodríguez, K., Lans, I., Cavasotto, C. N. & Cossio, P. Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Sci. Rep.* **9**, 5142 (2019).

212. Abad-Zapatero, C. & Blasi, D. Ligand efficiency indices (LEIs): More than a simple efficiency yardstick. *Mol. Inform.* **30**, 122–132 (2011).

213. Wishart, D. S. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* vol. 34 D668–D672 (2005).

214. Kumar, M., Kaur, T. & Sharma, A. Role of computational efficiency indices and pose clustering in effective decision making: An example of annulated furanones in Pf-DHFR space. *Comput. Biol. Chem.* **67**, 48–61 (2017).

215. Abad-Zapatero, C., Champness, E. J. & Segall, M. D. Alternative variables in drug discovery: promises and challenges. *Future Med. Chem.* **6**, 577–593 (2014).

216. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).

217. Maiorov, V. N. & Crippen, G. M. *Size-independent comparison of protein three-dimensional structures*. *Proteins: Structure, Function, and Bioinformatics* vol. 22 http://doi.wiley.com/10.1002/prot.340220308 (1995).

218. Fabrizio Mancin. The strength of the interaction. 1–32 (2017) doi:10.2495/SDP1100.

219. Brody, T. Clinical Trials 2nd edition. https://www.elsevier.com/books/clinical-trials/brody/978-0-12-804217-5 (2016).

220. Fingolimod in COVID-19 - Full Text View - ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/NCT04280588.

221. Derbyshire, E. R., Prudêncio, M., Mota, M. M. & Clardy, J. Liver-stage malaria parasites vulnerable to diverse chemical scaffolds. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 8511–8516 (2012).

222. Prado-Prado, F. J., García-Mera, X. & González-Díaz, H. Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species. *Bioorg. Med. Chem.* **18**, 2225–2231 (2010).

223. Trager, W. *et al.* Human malaria parasites in continuous culture. *Science (80-.).* **193**, 673–675 (1976).

224. Plouffe, D. *et al.* In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 9059–9064 (2008).

225. Delano, W. L. *The PyMOL Molecular Graphics System*. (2002).

226. Show contacts - PyMOLWiki. https://pymolwiki.org/index.php/Show_contacts.

227. Friedman, R. & Caflisch, A. Discovery of Plasmepsin Inhibitors by Fragment-Based Docking and Consensus Scoring. *ChemMedChem* **4**, 1317–1326 (2009).

228. Boss, C. *et al.* Achiral, cheap, and potent inhibitors of plasmepsins I, II, and IV. *ChemMedChem* **1**, 1341–1345 (2006).

229. Barratt, E. *et al.* Thermodynamic Penalty Arising from Burial of a Ligand Polar Group Within a Hydrophobic Pocket of a Protein Receptor. *J. Mol. Biol.* **362**, 994–1003 (2006).

230. Series, M. C. Methods and Principles in Medicinal Chemistry. **1**, 438–438 (2007).

231. Färber, P. M., Graeser, R., Franklin, R. M. & Kappes, B. Molecular cloning and characterization of a second calcium-dependent protein kinase of Plasmodium falciparum. *Mol. Biochem. Parasitol.* **87**, 211–216 (1997).

232. Merckx, A. *et al.* Structures of P. falciparum Protein Kinase 7 Identify an Activation Motif and Leads for Inhibitor Design. *Structure* **16**, 228–238 (2008).

233. Cabrera, D. G. *et al.* Plasmodial Kinase Inhibitors: License to Cure? *J. Med. Chem.* **61**, 8061–8077 (2018).

234. Fritz-Wolf, K. *et al.* Crystal Structure of the Plasmodium falciparum Thioredoxin Reductase–Thioredoxin Complex. *J. Mol. Biol.* **425**, 3446–3460 (2013).

235. Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).

236. Bosc, N. *et al.* Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *J. Cheminform.* **11**, 4 (2019).

237. Zhu, T. *et al.* Hit Identification and Optimization in Virtual Screening: Practical Recommendations Based Upon a Critical Literature Analysis. *J. Med. Chem.* **56**, 6560–6572 (2013).

238. Loza-Mejía, M. A. *et al.* In Silico Studies on Compounds Derived from Calceolaria: Phenylethanoid Glycosides as Potential Multitarget Inhibitors for the Development of Pesticides. *Biomolecules* **8**, 121 (2018).

239. Protein Arrays for Assessment of Target Selectivity - Drug Discovery World (DDW). https://www.ddw-online.com/protein-arrays-for-assessment-of-target-selectivity-1358-200212/.

240. Guggisberg, A. M., Amthor, R. E. & Odom, A. R. Isoprenoid biosynthesis in Plasmodium falciparum. *Eukaryot. Cell* **13**, 1348–1359 (2014).

241. Wiley, J. D. *et al.* Isoprenoid precursor biosynthesis is the essential metabolic role of the apicoplast during gametocytogenesis in Plasmodium falciparum. *Eukaryot. Cell* **14**, 128–139 (2015).

242. Murkin, A. S., Manning, K. A. & Kholodar, S. A. Mechanism and inhibition of 1-deoxy-d-xylulose-5-phosphate reductoisomerase. *Bioorg. Chem.* **57**, 171–185 (2014).

243. Odom, A. R. Five Questions about Non-Mevalonate Isoprenoid Biosynthesis. *PLOS Pathog.* **7**, e1002323 (2011).

244. Umeda, T. *et al.* Molecular basis of fosmidomycin's action on the human malaria parasite Plasmodium falciparum. *Sci. Rep.* **1**, 9 (2011).

245. Hale, I., M. O'Neill, P., G. Berry, N., Odom, A. & Sharma, R. The MEP pathway and the development of inhibitors as potential anti-infective agents. *Medchemcomm* **3**, 418–433 (2012).

246. Wiesner, J., Borrmann, S. & Jomaa, H. Fosmidomycin for the treatment of malaria. *Parasitol. Res.* **90**, S71–S76 (2003).

247. Berenger, F., Vu, O. & Meiler, J. Consensus queries in ligand-based virtual screening experiments. *J. Cheminform.* **9**, 60 (2017).

248. Temml, V., Voss, C. V., Dirsch, V. M. & Schuster, D. Discovery of new liver X receptor agonists by pharmacophore modeling and shape-based virtual screening. *J. Chem. Inf. Model.* **54**, 367–371 (2014).

249. Tangyuenyongwatana, P. & Gritsanapan, W. Virtual screening for novel 1-deoxy-d-xylulose-5-phosphate reductoisomerase inhibitors: A shape-based search approach. *Thai J. Pharm. Sci.* **41**, (2017).

250. Wadood, A. *et al.* In silico identification of promiscuous scaffolds as potential inhibitors of 1-deoxy-d-xylulose 5-phosphate reductoisomerase for treatment of Falciparum malaria. *Pharm. Biol.* **55**, 19–32 (2017).

251. Chaudhary, K. K. & Prasad, C. V. S. S. Virtual Screening of compounds to 1-deoxy-Dxylulose 5-phosphate reductoisomerase (DXR) from Plasmodium falciparum. *Bioinformation* **10**, 358–364 (2014).

252. Cobb, R. E. *et al.* Structure-guided design and biosynthesis of a novel FR-900098 analogue as a potent Plasmodium falciparum 1-deoxy-D-xylulose-5-phosphate reductoisomerase (Dxr) inhibitor. *Chem. Commun. (Camb).* **51**, 2526–2528 (2015).

253. de Ruyck, J., Brysbaert, G., Blossey, R. & Lensink, M. F. Molecular docking as a popular tool in drug design, an in silico travel. *Adv. Appl. Bioinform. Chem.* **9**, 1–11 (2016).

254. ChemBridge | Home. https://www.chembridge.com/.

255. Manhas, A., Patel, D., Lone, M. Y. & Jha, P. C. Identification of natural compound inhibitors against PfDXR: A hybrid structure-based molecular modeling approach and molecular dynamics simulation studies. *J. Cell. Biochem.* **120**, 14531–14543 (2019).

256. Sooriyaarachchi, S. *et al.* Targeting an Aromatic Hotspot in Plasmodium falciparum 1-Deoxy-d-xylulose-5-phosphate Reductoisomerase with beta-Arylpropyl Analogues of Fosmidomycin. *ChemMedChem* **11**, 2024–2036 (2016).

257. Nikolova, N. & Jaworska, J. Approaches to Measure Chemical Similarity– a Review. *QSAR Comb. Sci.* **22**, 1006–1026 (2003).

258. Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspect. Drug Discov. Des.* **9/11**, 225–252 (1998).

259. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, 20 (2015).

260. Brown, A. C. & Fraser, T. R. On the Connection between Chemical Constitution and Physiological Action; with special reference to the Physiological Action of the Salts of the Ammonium Bases derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia. *J. Anat. Physiol.* **2**, 224–42 (1868).

261. Kumar, A. & Zhang, K. Y. J. Advances in the development of shape similarity methods and their application in drug discovery. *Front. Chem.* **6**, 315 (2018).

262. Ballester, P. J. & Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **28**, 1711–1723 (2007).

263. Zauhar, R. J., Moyna, G., Tian, L. F., Li, Z. J. & Welsh, W. J. Shape Signatures: A New Approach

to Computer-Aided Ligand- and Receptor-Based Drug Design. *J. Med. Chem.* **46**, 5674–5690 (2003).

264. Armstrong, M. S. *et al.* ElectroShape: Fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput. Aided. Mol. Des.* **24**, 789–801 (2010).

265. Schreyer, A. & Blundell, T. CREDO: A Protein-Ligand Interaction Database for Drug Discovery. *Chem. Biol. Drug Des.* **73**, 157–167 (2009).

266. Schreyer, A. M. & Blundell, T. USRCAT: real-time ultrafast shape recognition with pharmacophoric constraints. *J. Cheminform.* **4**, 27 (2012).

267. Wójcikowski, M., Zielenkiewicz, P. & Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A new open-source player in the drug discovery field. *J. Cheminform.* **7**, 1–6 (2015).

268. Gobbi, A. & Poppinger, D. Genetic optimization of combinatorial libraries. *Biotechnol. Bioeng.* **61**, 47–54 (1998).

269. RDKit Cookbook — The RDKit 2019.09.1 documentation. https://rdkit.readthedocs.io/en/latest/Cookbook.html.

270. Gladysz, R. *et al.* Spectrophores as one-dimensional descriptors calculated from three-dimensional atomic properties: applications ranging from scaffold hopping to multi-target virtual screening. *J. Cheminform.* **10**, 9 (2018).

271. Landrum, G. *et al.* rdkit: 2016_03_4 (Q1 2016) Release. *Release 2017.09.1* (2017) doi:10.5281/zenodo.60510.

272. Bentham Science Publisher, B. S. P. Scoring Functions for Protein-Ligand Docking. *Curr. Protein Pept. Sci.* **7**, 407–420 (2006).

273. Liu, J. & Wang, R. Classification of current scoring functions. *J. Chem. Inf. Model.* **55**, 475–482 (2015).

274. Wójcikowski, M., Ballester, P. J. & Siedlecki, P. Performance of machine-learning scoring functions in structure-based virtual screening. *Sci. Rep.* **7**, 46710 (2017).

275. Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. & Koes, D. R. Protein–Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **57**, 942–957 (2017).

276. Cao, Y. & Li, L. Improved protein–ligand binding affinity prediction by using a curvature-dependent surface-area model. *Bioinformatics* **30**, 1674–1680 (2014).

277. Neudert, G. & Klebe, G. DSX: A knowledge-based scoring function for the assessment of protein-ligand complexes. *J. Chem. Inf. Model.* **51**, 2731–2745 (2011).

278. Durrant, J. D. & McCammon, J. A. NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *J. Chem. Inf. Model.* **51**, 2897–2903 (2011).

279. Durrant, J. D. & McCammon, J. A. BINANA: A novel algorithm for ligand-binding characterization. *J. Mol. Graph. Model.* **29**, 888–893 (2011).

280. Hu, L., Benson, M. L., Smith, R. D., Lerner, M. G. & Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins Struct. Funct. Bioinforma.* **60**, 333–340 (2005).

281. Wang, R., Fang, X., Lu, Y., Yang, C. Y. & Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **48**, 4111–4119 (2005).

282. Renxiao Wang, Xueliang Fang, Yipin Lu, and & Wang*, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. (2004) doi:10.1021/JM030580L.

283. Guedes, I. A., Pereira, F. S. S. & Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.* **9**, 1089 (2018).

284. Quiroga, R. & Villarreal, M. A. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLoS One* **11**, e0155183 (2016).

285. Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M. M. & Siedlecki, P. Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* **35**, 1334–1341 (2019).

286. Searcey, M. The Handbook of Medicinal Chemistry-Principles and Practice. Edited by Andrew Davis and Simon E. Ward. *ChemMedChem* **10**, 2111–2112 (2015).

287. Burley, S. K. *et al.* RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* **47**, D464–D474 (2019).

288. Chofor, R. *et al.* Synthesis and Bioactivity of β-Substituted Fosmidomycin Analogues Targeting 1-Deoxy-d-xylulose-5-phosphate Reductoisomerase. **58**, 2988–3001.

289. Sooriyaarachchi, S. *et al.* Targeting an Aromatic Hotspot in Plasmodium falciparum 1-Deoxy-d-xylulose-5-phosphate Reductoisomerase with β-Arylpropyl Analogues of Fosmidomycin. *ChemMedChem* **11**, 2024–2036 (2016).

290. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

291. Diallo, B. N., Tastan Bishop, Ö. & Lobb, K. In silico study of Plasmodium 1-deoxy-d- xylulose 5-phosphate reductoisomerase (DXR) for identification of novel inhibitors from SANCDB Bakary N ’ tji Diallo. (2018).

292. Sooriyaarachchi, S. *et al.* Targeting an Aromatic Hotspot in Plasmodium falciparum 1-Deoxy-d-xylulose-5-phosphate Reductoisomerase with β-Arylpropyl Analogues of Fosmidomycin. *ChemMedChem* **11**, 2024–2036 (2016).

293. Tange, O. GNU Parallel 2018. (2018) doi:10.5281/ZENODO.1146014.

294. Spectrophores™ — Open Babel v2.3.1 documentation. http://openbabel.org/docs/current/Fingerprints/spectrophore.html.

295. Gladysz, R. *et al.* Spectrophores as one-dimensional descriptors calculated from three-

dimensional atomic properties: applications ranging from scaffold hopping to multi-target virtual screening. *J. Cheminform.* **10**, 9 (2018).

296. Ballester, P. J., Finn, P. W. & Richards, W. G. Ultrafast shape recognition: Evaluating a new ligand-based virtual screening technology. *J. Mol. Graph. Model.* **27**, 836–845 (2009).

297. Armstrong, M. S. *et al.* ElectroShape: Fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J. Comput. Aided. Mol. Des.* **24**, 789–801 (2010).

298. Landrum, G. *Fingerprints in the RDKit*. https://www.rdkit.org/UGM/2012/Landrum_RDKit_UGM.Fingerprints.Final.pptx.pdf.

299. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).

300. Rocklin, M. *Dask: Parallel Computation with Blocked algorithms and Task Scheduling*. *PROC. OF THE 14th PYTHON IN SCIENCE CONF* https://www.youtube.com/watch?v=1kkFZ4P-XHg (2015).

301. Wójcikowski, M., Zielenkiewicz, P. & Siedlecki, P. Open Drug Discovery Toolkit (ODDT): A new open-source player in the drug discovery field. *J. Cheminform.* **7**, 1–6 (2015).

302. Cao, Y. & Li, L. Improved protein–ligand binding affinity prediction by using a curvature-dependent surface-area model. *Bioinformatics* **30**, 1674–1680 (2014).

303. Wang, R., Lai, L. & Wang, S. *Further development and validation of empirical scoring functions for structure-based binding affinity prediction*. *Journal of Computer-Aided Molecular Design* vol. 16 https://www.ics.uci.edu/~dock/manuals/xscore1.1_manual/xscore.pdf (2002).

304. McKinney, W. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python High Perform. Sci. Comput.* 1–9 (2011).

305. API reference — pandas 1.1.3 documentation. https://pandas.pydata.org/docs/reference/index.html.

306. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10037–41 (2001).

307. Kumari, R., Kumar, R. & Lynn, A. G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model.* **54**, 1951–1962 (2014).

308. Ngo, S. T., Hung, H. M. & Nguyen, M. T. Fast and accurate determination of the relative binding affinities of small compounds to HIV-1 protease using non-equilibrium work. *J. Comput. Chem.* **37**, 2734–2742 (2016).

309. Zhang, J. L., Zheng, Q. C., Li, Z. Q. & Zhang, H. X. Molecular dynamics simulations suggest Ligand's binding to Nicotinamidase/Pyrazinamidase. *PLoS One* **7**, (2012).

310. Patel, J. S., Berteotti, A., Ronsisvalle, S., Rocchia, W. & Cavalli, A. Steered molecular dynamics simulations for studying protein-ligand interaction in cyclin-dependent kinase 5.

J. Chem. Inf. Model. **54**, 470–480 (2014).

311. Li, M. S. Ligand migration and steered molecular dynamics in drug discovery: Comment on "Ligand diffusion in proteins via enhanced sampling in molecular dynamics" by Jakub Rydzewski and Wieslaw Nowak. *Phys. Life Rev.* **22**–**23**, 79–81 (2017).

312. Thai, N. Q., Nguyen, H. L., Linh, H. Q. & Li, M. S. Protocol for fast screening of multi-target drug candidates: Application to Alzheimer's disease. *J. Mol. Graph. Model.* **77**, 121–129 (2017).

313. Ngo, S. T., Vu, K. B., Bui, L. M. & Vu, V. V. Effective estimation of ligand-binding affinity using biased sampling method. *ACS Omega* **4**, 3887–3893 (2019).

314. Hub, J. S., De Groot, B. L. & Van Der Spoel, D. G-whams-a free Weighted Histogram Analysis implementation including robust error and autocorrelation estimates. *J. Chem. Theory Comput.* **6**, 3713–3720 (2010).

315. Lemkul, J. A. & Bevan, D. R. Assessing the stability of Alzheimer's amyloid protofibrils using molecular dynamics. *J. Phys. Chem. B* **114**, 1652–1660 (2010).

316. Irwin, J. J. & Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **45**, 177 (2005).

317. Kholodar, S. A. & Murkin, A. S. DXP Reductoisomerase: Reaction of the Substrate in Pieces Reveals a Catalytic Role for the Nonreacting Phosphodianion Group. *Biochemistry* **52**, 2302–2308 (2013).

318. R. Jackson, E. & S. Dowd, C. Inhibition of 1-Deoxy-D-Xylulose-5-Phosphate Reductoisomerase (Dxr): A Review of the Synthesis and Biological Evaluation of Recent Inhibitors. *Current Topics in Medicinal Chemistry* vol. 12 706–728 http://www.eurekaselect.com/96365/article (2012).

319. Kunfermann, A. *et al.* IspC as target for antiinfective drug discovery: Synthesis, enantiomeric separation, and structural biology of fosmidomycin thia isosters. *J. Med. Chem.* **56**, 8151–8162 (2013).

320. Baldi, P. & Nasr, R. When is chemical similarity significant? the statistical distribution of chemical similarity scores and its extreme values. *J. Chem. Inf. Model.* **50**, 1205–1222 (2010).

321. Kendall, M. G. A NEW MEASURE OF RANK CORRELATION. *Biometrika* **30**, 81–93 (1938).

322. Shamsara, J. Evaluation of 11 Scoring Functions Performance on Matrix Metalloproteinases. *Int. J. Med. Chem.* **2014**, 1–9 (2014).

323. Ray, S. & Lindsay, B. G. The topography of multivariate normal mixtures. *Ann. Stat.* **33**, 2042–2065 (2005).

324. Madhavilatha, K. N., Rama, G. & Babu, M. Systematic approach for enrichment of docking outcome using consensus scoring functions. doi:10.1088/1742-6596/1228/1/012019.

325. Hunter, J. D. Matplotlib: A 2D Graphics Environment https://doi. org/10.1109/MCSE. 2007.55 Comput. Sci. (2007).

326. Gu, J., Li, H. & Wang, X. A self-adaptive steered molecular dynamics method based on minimization of stretching force reveals the binding affinity of protein-ligand complexes. *Molecules* **20**, 19236–19251 (2015).

327. Do, P. C., Lee, E. H. & Le, L. Steered Molecular Dynamics Simulation in Rational Drug Design. *J. Chem. Inf. Model.* **58**, 1473–1482 (2018).

328. Li, D., Ji, B., Hwang, K.-C. & Huang, Y. Strength of Hydrogen Bond Network Takes Crucial Roles in the Dissociation Process of Inhibitors from the HIV-1 Protease Binding Pocket. *PLoS One* **6**, e19268 (2011).

329. Yang, K., Liu, X., Wang, X. & Jiang, H. A steered molecular dynamics method with adaptive direction adjustments. *Biochem. Biophys. Res. Commun.* **379**, 494–498 (2009).

330. Kholodar, S. A. & Murkin, A. S. DXP reductoisomerase: Reaction of the substrate in pieces reveals a catalytic role for the nonreacting phosphodianion group. *Biochemistry* **52**, 2302–2308 (2013).

331. Vuong, Q. Van *et al.* A New Method for Navigating Optimal Direction for Pulling Ligand from Binding Pocket: Application to Ranking Binding Affinity by Steered Molecular Dynamics. *J. Chem. Inf. Model.* **55**, 2731–2738 (2015).

332. Jubb, H. C. *et al.* Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *J. Mol. Biol.* **429**, 365–371 (2017).

333. scipy.stats.pointbiserialr — SciPy v0.14.0 Reference Guide. https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.pointbiserialr.html.

334. Khazanov, N. A. & Carlson, H. A. Exploring the Composition of Protein-Ligand Binding Sites on a Large Scale. *PLoS Comput. Biol.* **9**, 1003321 (2013).

335. Jessica L. Goble, H. J. & Jessica, L. G.; Hailey, J.; Jaco, D. R.; Linda, L. S.; Abraham, L.; Gregory, L. B. and Aileen, B. The Druggable Antimalarial Target PfDXR: Overproduction Strategies and Kinetic Characterization. *Protein Pept. Lett.* **20**, 0–0 (2013).

336. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **10**, 449–461 (2015).

337. Weis, A., Katebzadeh, K., Söderhjelm, P., Nilsson, I. & Ryde, U. Ligand affinities predicted with the MM/PBSA method: Dependence on the simulation method and the force field. *J. Med. Chem.* **49**, 6596–6606 (2006).

338. Xue, J. *et al.* Antimalarial and Structural Studies of Pyridine-Containing Inhibitors of 1-Deoxyxylulose-5-phosphate Reductoisomerase. *ACS Med. Chem. Lett.* **4**, 278–282 (2012).

339. Umeda, T. *et al.* Molecular basis of fosmidomycin's action on the human malaria parasite

Plasmodium falciparum. *Sci. Rep.* **1**, 9 (2011).

340. Lan, N. T. *et al.* Prediction of AChE-ligand affinity using the umbrella sampling simulation. *J. Mol. Graph. Model.* **93**, 107441 (2019).

341. Park, S., Khalili-Araghi, F., Tajkhorshid, E. & Schulten, K. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *J. Chem. Phys.* **119**, 3559–3566 (2003).

342. Durrant, J. D. & McCammon, J. A. NNScore: A neural-network-based scoring function for the characterization of protein-ligand complexes. *J. Chem. Inf. Model.* **50**, 1865–1871 (2010).

343. Preto, J. & Gentile, F. Assessing and improving the performance of consensus docking strategies using the DockBox package. *J. Comput. Aided. Mol. Des.* **33**, 817–829 (2019).

344. Chaput, L. & Mouawad, L. Efficient conformational sampling and weak scoring in docking programs? Strategy of the wisdom of crowds. *J. Cheminform.* **9**, 37 (2017).

345. Gorgulla, C. *et al.* An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **580**, 663–668 (2020).

346. Hardy, K. *et al.* Neanderthal medics? Evidence for food, cooking, and medicinal plants entrapped in dental calculus. *Naturwissenschaften* **99**, 617–626 (2012).

347. Sorokina, M. & Steinbeck, C. Review on natural products databases: Where to find data in 2020. *J. Cheminform.* **12**, 1–51 (2020).

348. Calixto, J. B. The role of natural products in modern drug discovery. *An. Acad. Bras. Cienc.* **91**, (2019).

349. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **83**, 770–803 (2020).

350. Wright, G. D. Opportunities for natural products in 21st century antibiotic discovery. *Natural Product Reports* vol. 34 694–701 (2017).

351. Chen, Y., De Bruyn Kops, C. & Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J. Chem. Inf. Model.* **57**, 2099–2111 (2017).

352. Sorokina, M. & Steinbeck, C. Naples: A natural products likeness scorer—web application and database. *J. Cheminform.* **11**, 1–7 (2019).

353. Cockroft, N. T., Cheng, X. & Fuchs, J. R. STarFish: A Stacked Ensemble Target Fishing Approach and its Application to Natural Products. *J. Chem. Inf. Model.* **59**, 4906–4920 (2019).

354. Chen, Y., Stork, C., Hirte, S. & Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* **9**, 43 (2019).

355. Zeng, X. *et al.* NPASS: Natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* **46**, (2018).

356. Najjar, A., Olğaç, A., Ntie-Kang, F. & Sippl, W. Fragment-based drug design of nature-inspired compounds. *Phys. Sci. Rev.* **4**, (2019).

357. L, E., H, C., R, S. & A, B. Computational Approach Revealed Potential Affinity of Antiasthmatics Against Receptor Binding Domain of 2019n-Cov Spike Glycoprotein. (2020) doi:10.26434/CHEMRXIV.12115638.V1.

358. Musyoka, T. M., Kanzi, A. M., Lobb, K. A. & Tastan Bishop, Ö. Structure Based Docking and Molecular Dynamic Studies of Plasmodial Cysteine Proteases against a South African Natural Compound and its Analogs. *Sci. Rep.* **6**, 23690 (2016).

359. Kimuda, M. P., Laming, D., Hoppe, H. C., Ozlem, T. B. & Tastan Bishop, Ö. Identification of Novel Potential Inhibitors of Pteridine Reductase 1 in Trypanosoma brucei via Computational Structure-Based Approaches and in Vitro Inhibition Assays. *Molecules* **24**, (2019).

360. Nyamai, D. W. & Tastan Bishop, Ö. Identification of selective novel hits against plasmodium falciparum prolyl tRNA synthetase active site and a predicted allosteric site using in silico approaches. *Int. J. Mol. Sci.* **21**, 3803 (2020).

361. Musyoka, T. M., Kanzi, A. M., Lobb, K. A. & Tastan Bishop, Ö. Structure Based Docking and Molecular Dynamic Studies of Plasmodial Cysteine Proteases against a South African Natural Compound and its Analogs. *Sci. Rep.* **6**, (2016).

362. Penkler, D. L., Atilgan, C. & Tastan Bishop, Ö. Allosteric Modulation of Human Hsp90α Conformational Dynamics. *J. Chem. Inf. Model.* **58**, 383–404 (2018).

363. Diallo, B. N. In silico study of Plasmodium 1-deoxy-dxylulose 5-phosphate reductoisomerase (DXR) for identification of novel inhibitots from SANCDB. (2018).

364. Amusengeri, A. & Tastan Bishop, Ö. Discorhabdin N, a South African natural compound, for Hsp72 and Hsc70 allosteric modulation: Combined study of molecular modeling and dynamic residue network analysis. *Molecules* **24**, (2019).

365. Bernardini, S., Tiezzi, A., Laghezza Masci, V. & Ovidi, E. Natural products for human health: an historical overview of the drug discovery approaches. *Nat. Prod. Res.* **32**, 1926–1950 (2018).

366. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, 8 (2009).

367. Lewinsohn, T. M. & Prado, P. I. How many species are there in Brazil? *Conserv. Biol.* **19**, 619–624 (2005).

368. Pilon, A. C. *et al.* NuBBEDB: An updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci. Rep.* **7**, (2017).

369. Home - SANBI. https://www.sanbi.org/.

370. Griffiths, C. L., Robinson, T. B., Lange, L. & Mead, A. Marine Biodiversity in South Africa: An Evaluation of Current States of Knowledge. *PLoS One* **5**, e12008 (2010).

371. SANBI. *National Biodiversity Assessment 2018 - Synthesis Report*. *South African National Biodiversity Institute* http://bgis.sanbi.org/NBA/NBA2011_metadata_formalprotectedareas.pdf%5Cnpapers2://publication/uuid/786A77C5-B11A-4F8D-B139-F3F626EBC802 (2018).

372. Cordell, G. A. Biodiversity and drug discovery - A symbiotic relationship. *Phytochemistry* vol. 55 463–480 (2000).

373. Ntie-Kang, F. *et al.* CamMedNP: Building the Cameroonian 3D structural natural products database for virtual screening. *BMC Complement. Altern. Med.* (2013) doi:10.1186/1472-6882-13-88.

374. Engelhardt, C., Petereit, F., Lechtenberg, M., Liefländer-Wulf, U. & Hensel, A. Qualitative and quantitative phytochemical characterization of Myrothamnus flabellifolia Welw. *Fitoterapia* **114**, 69–80 (2016).

375. Fantoukh, O. I. *et al.* Safety Assessment of Phytochemicals Derived from the Globalized South African Rooibos Tea (Aspalathus linearis) through Interaction with CYP, PXR, and P-gp. *J. Agric. Food Chem.* **67**, 4967–4975 (2019).

376. Awolola, G. V., Sofidiya, M. O., Baijnath, H., Noren, S. S. & Koorbanally, N. A. The phytochemistry and gastroprotective activities of the leaves of Ficus glumosa. *South African J. Bot.* **126**, 190–195 (2019).

377. Pilón-Jiménez, B. A., Saldívar-González, F. I., Díaz-Eufracio, B. I. & Medina-Franco, J. L. BIOFACQUIM: A Mexican compound database of natural products. *Biomolecules* **9**, (2019).

378. Cragg, G. M. & Newman, D. J. Natural products: a continuing source of novel drug leads. *Biochim. Biophys. Acta* **1830**, 3670–95 (2013).

379. Oliveira, A. B. *et al.* Plant-derived antimalarial agents: New leads and efficient phythomedicines. Part I. alkaloids. *An. Acad. Bras. Cienc.* **81**, 715–740 (2009).

380. Valli, M. *et al.* Development of a natural products database from the biodiversity of Brazil. *J. Nat. Prod.* **76**, 439–444 (2013).

381. Chen, C. Y. C. TCM Database@Taiwan: The world's largest traditional Chinese medicine database for drug screening In Silico. *PLoS One* **6**, (2011).

382. Elsevier. Elsevier Developer Portal. *Elsevier.com* https://dev.elsevier.com/tecdoc_text_mining.html (2010).

383. CAS. SciFinder - A CAS Solution. Accessed: 24.10.2015. *Publication* wefw http://www.cas.org/products/scifinder (2015).

384. Selenium with Python — Selenium Python Bindings 2 documentation. https://selenium-

python.readthedocs.io/.

385. Swain, M. PubChemPy: A way to interact with PubChem in Python. (2014).

386. Swain, M. CIRpy-A Python interface for the Chemical Identifier Resolver (CIR). *Matt Swain's Blog* (2012).

387. Gaulton, A. *et al.* ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100-7 (2012).

388. Sterling, T. & Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).

389. Schmidt, M. W. *et al.* General atomic and molecular electronic structure system. *J. Comput. Chem.* **14**, 1347–1363 (1993).

390. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **8**, 1–20 (2016).

391. Chamberlain, S. pygbif 0.4.0 documentation — pygbif 0.4.0 documentation. https://pygbif.readthedocs.io/en/latest/index.html.

392. GBIF. https://www.gbif.org/.

393. Release, S. 1: Maestro. *Schrödinger, LLC, New York, NY* **2017**, (2017).

394. Easy compound ordering service - MolPort. https://www.molport.com/shop/index.

395. Kiss, R., Sandor, M. & Szalai, F. A. http://Mcule.com: a public web service for drug discovery. *J. Cheminform.* **4**, (2012).

396. SANCDB. https://sancdb.rubi.ru.ac.za/.

397. Michael Glenister. *Unpublished pipeline*.

398. Tosco, P., Stiefl, N. & Landrum, G. Bringing the MMFF force field to the RDKit: Implementation and validation. *J. Cheminform.* **6**, 37 (2014).

399. Musyoka, T. M. *Unpublished pipeline*. (2020).

400. A Janssen, A. P. *et al.* Drug Discovery Maps, a Machine Learning Model That Visualizes and Predicts Kinome−Inhibitor Interaction Landscapes. (2018) doi:10.1021/acs.jcim.8b00640.

401. Naveja, J. J. & Medina-Franco, J. L. Finding Constellations in Chemical Space Through Core Analysis. *Front. Chem.* **7**, 510 (2019).

402. Yosipof, A., Guedes, R. C. & García-Sosa, A. T. Data Mining and Machine Learning Models for Predicting Drug Likeness and Their Disease or Organ Category. *Front. Chem.* **6**, 162 (2018).

403. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

404. sklearn.manifold.TSNE — scikit-learn 0.23.1 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.

405. Van Der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2625 (2008).

406. Nguyen, K. T., Blum, L. C., Van Deursen, R. & Reymond, J. L. Classification of organic molecules by molecular quantum numbers. *ChemMedChem* **4**, 1803–1805 (2009).

407. Landrum, G. RDKit Documentation. *Read. Writ.* (2011) doi:10.5281/zenodo.60510.

408. Kearney, S. E. *et al.* Canvass: A Crowd-Sourced, Natural Product Screening Library for Exploring Biological Space. (2018) doi:10.26434/CHEMRXIV.7172369.V2.

409. Sánchez-Cruz, N., Pilón-Jiménez, B. A. & Medina-Franco, J. L. Functional group and diversity analysis of BIOFACQUIM: A Mexican natural product database. *F1000Research* **8**, 2071 (2020).

410. Garcia-Castro, M., Zimmermann, S., Sankar, M. G. & Kumar, K. Scaffold Diversity Synthesis and Its Application in Probe and Drug Discovery. *Angewandte Chemie - International Edition* vol. 55 7586–7605 (2016).

411. Singh, N. *et al.* Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository. *J. Chem. Inf. Model.* **49**, 1010–1024 (2009).

412. Saldívar-Gonzaíez, F. I. *et al.* Chemical Space and Diversity of the NuBBE Database: A Chemoinformatic Characterization. *J. Chem. Inf. Model* **59**, (2019).

413. The Scopy's documentation — Scopy 1.2.3 documentation. https://scopy.iamkotori.com/index.html.

414. Ertl, P. & Rohde, B. The Molecule Cloud - Compact visualization of large collections of molecules. *J. Cheminform.* **4**, 1 (2012).

415. Sterling, T. & Irwin, J. J. {ZINC} 15 – Ligand Discovery for Everyone. **55**, 2324–2337.

416. Scott, D. E., Coyne, A. G., Hudson, S. A. & Abell, C. Fragment-based approaches in drug discovery and chemical biology. *Biochemistry* **51**, 4990–5003 (2012).

417. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**, 2719–2740 (2010).

418. Saldívar-González, F. I., Valli, M., Andricopulo, A. D., Da Silva Bolzani, V. & Medina-Franco, J. L. Chemical Space and Diversity of the NuBBE Database: A Chemoinformatic Characterization. *J. Chem. Inf. Model.* **59**, 74–85 (2019).

419. Kluyver, T. *et al.* Jupyter Notebooks—a publishing format for reproducible computational workflows. in *Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016*

87–90 (2016). doi:10.3233/978-1-61499-649-1-87.

420. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* 56–61 (2010) doi:10.25080/majora-92bf1922-00a.

421. Brugman, S. pandas-profiling: Exploratory Data Analysis for Python. (2019).

422. Ntie-Kang, F. *et al.* AfroDb: a select highly potent and diverse natural product library from African medicinal plants. *PLoS One* (2013) doi:10.1371/journal.pone.0078085.

423. Ntie-Kang, F. *et al.* ConMedNP: A natural product library from Central African medicinal plants for drug discovery. *RSC Adv.* **4**, 409–419 (2014).

424. Castells, E., Mulder, P. P. J. & Pérez-Trujillo, M. Diversity of pyrrolizidine alkaloids in native and invasive Senecio pterophorus (Asteraceae): Implications for toxicity. *Phytochemistry* **108**, 137–146 (2014).

425. Kuroda, M., Ori, K. & Mimaki, Y. Ornithosaponins A-D, four new polyoxygenated steroidal glycosides from the bulbs of Ornithogalum thyrsoides. *Steroids* **71**, 199–205 (2006).

426. Ornithogalum thyrsoides | PlantZAfrica. http://pza.sanbi.org/ornithogalum-thyrsoides.

427. Ornithogalum saundersiae | PlantZAfrica. http://pza.sanbi.org/ornithogalum-saundersiae.

428. Iguchi, T. *et al.* Cholestane glycosides from Ornithogalum saundersiae bulbs and the induction of apoptosis in HL-60 cells by OSW-1 through a mitochondrial-independent signaling pathway. *J. Nat. Med.* **73**, 131–145 (2019).

429. Mann, M. G. A. *et al.* Halogenated monoterpene aldehydes from the South African marine alga Plocamium corallorhiza. *J. Nat. Prod.* **70**, 596–599 (2007).

430. Knott, M. G. *et al.* Plocoralides A-C, polyhalogenated monoterpenes from the marine alga Plocamium corallorhiza. *Phytochemistry* **66**, 1108–1112 (2005).

431. Davies-Coleman, M. & Veale, C. Recent Advances in Drug Discovery from South African Marine Invertebrates. *Mar. Drugs* **13**, 6366–6383 (2015).

432. Pettit, G. R. *et al.* Isolation and structure of the unusual Indian Ocean Cephalodiscus gilchristi components, cephalostatins 5 and 6. *Can. J. Chem.* **67**, 1509–1513 (1989).

433. SANBI. Threatened Species Programme | SANBI Red List of South African Plants. *South African National Biodiversity Institute* http://redlist.sanbi.org/stats.php (2019).

434. Bultum, L. E., Woyessa, A. M. & Lee, D. ETM-DB: Integrated Ethiopian traditional herbal medicine and phytochemicals database. *BMC Complement. Altern. Med.* **19**, (2019).

435. Ntie-Kang, F. *et al.* Virtualizing the p-ANAPL Library: A Step towards Drug Discovery from African Medicinal Plants. *PLoS One* **9**, e90655 (2014).

436. Banerjee, P. *et al.* Super Natural II-a database of natural products. *Nucleic Acids Res.* (2015) doi:10.1093/nar/gku886.

437. Garcia-Castro, M., Zimmermann, S., Sankar, M. G. & Kumar, K. Scaffold Diversity Synthesis and Its Application in Probe and Drug Discovery. *Angew. Chemie - Int. Ed.* **55**, 7586–7605 (2016).

438. Swain, M. C. & Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).

439. Papadatos, G. *et al.* SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res.* **44**, (2016).

440. fisher_test function | R Documentation. https://www.rdocumentation.org/packages/rstatix/versions/0.6.0/topics/fisher_test.

441. Antunes, E. M. *et al.* Identification and in vitro anti-esophageal cancer activity of a series of halogenated monoterpenes isolated from the South African seaweeds Plocamium suhrii and Plocamium cornutum. *Phytochemistry* **72**, 769–772 (2011).

442. Afolayan, A. F. *et al.* Antiplasmodial halogenated monoterpenes from the marine red alga Plocamium cornutum. *Phytochemistry* **70**, 597–600 (2009).

443. Dias, D. A., Urban, S. & Roessner, U. A Historical overview of natural products in drug discovery. *Metabolites* **2**, 303–336 (2012).

444. Wink, M. *Annual plant reviews, biochemistry of plant secondary metabolism*. vol. 40 (John Wiley & Sons, 2011).

445. Castelli, M. V & López, S. N. Homoisoflavonoids: Occurrence, biosynthesis, and biological activity. in *Studies in Natural Products Chemistry* vol. 54 315–354 (Elsevier, 2017).

446. Mottaghipisheh, J. & Iriti, M. Sephadex® LH-20, Isolation, and Purification of Flavonoids from Plant Species: A Comprehensive Review. *Molecules* **25**, 4146 (2020).

447. Li, F., Janussen, D., Peifer, C., Pérez-Victoria, I. & Tasdemir, D. Targeted isolation of tsitsikammamines from the antarctic deep-sea sponge latrunculia biformis by molecular networking and anticancer activity. *Mar. Drugs* **16**, 268 (2018).

448. Health | South African Government. https://www.gov.za/about-sa/health.

449. Meyers, J., Carter, M., Mok, N. Y. & Brown, N. On the origins of three-dimensionality in drug-like molecules. *Future Med. Chem.* **8**, 1753–1767 (2016).

450. Ertl, P. & Schuhmann, T. Cheminformatics Analysis of Natural Product Scaffolds: Comparison of Scaffolds Produced by Animals, Plants, Fungi and Bacteria. *Mol. Inform.* (2020) doi:10.1002/minf.202000017.

451. CHEMBL1766622 Compound Report Card. https://www.ebi.ac.uk/chembldb/index.php/compound/inspect/CHEMBL2079699.

452. Simon, L. *et al.* Synthesis, anticancer, structural, and computational docking studies of 3-benzylchroman-4-one derivatives. *Bioorganic Med. Chem. Lett.* **27**, 5284–5290 (2017).

453. Limban, C. *et al.* The use of structural alerts to avoid the toxicity of pharmaceuticals. *Toxicology Reports* vol. 5 943–953 (2018).

454. Homeopathic Treatment of Premenstrual Syndrome - Full Text View - ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/NCT02402049.

455. Weigt, S., Huebler, N., Strecker, R., Braunbeck, T. & Broschard, T. H. Developmental effects of coumarin and the anticoagulant coumarin derivative warfarin on zebrafish (Danio rerio) embryos. *Reprod. Toxicol.* **33**, 133–141 (2012).

456. Morris, G. M. *et al.* Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **19**, 1639–1662 (1998).

457. Ntie-Kang, F. *et al.* NANPDB: A Resource for Natural Products from Northern African Sources. *J. Nat. Prod.* **80**, 2067–2076 (2017).

458. Medina-Franco, J. L. Towards a unified Latin American Natural Products Database: LANaPD. *Futur. Sci. OA* **6**, FSO597 (2020).

459. Holschneider, D. P., Bradesi, S. & Mayer, E. A. The role of experimental models in developing new treatments for irritable bowel syndrome. *Expert Rev. Gastroenterol. Hepatol.* **5**, 43–57 (2011).

460. Moloney, R. D., O'Mahony, S. M., Dinan, T. G. & Cryan, J. F. Stress-induced visceral pain: Toward animal models of irritable-bowel syndrome and associated comorbidities. *Front. Psychiatry* **6**, 15 (2015).

461. Balmus, I. M. *et al.* Irritable bowel syndrome between molecular approach and clinical expertise—searching for gap fillers in the oxidative stress way of thinking. *Med.* **56**, 38 (2020).

462. Polovinkin, L. *et al.* Conformational transitions of the serotonin 5-HT3 receptor. *Nature* **563**, 275–279 (2018).

463. Moskwa, A. & Boznańska, P. Role of serotonin in the pathophysiology of the irritable bowel syndrome. *Wiad. Lek.* **60**, 371–376 (2007).

464. Cryan, J. F. *et al.* The microbiota-gut-brain axis. *Physiol. Rev.* **99**, 1877–2013 (2019).

465. Juza, R. *et al.* Recent advances with 5 HT3 modulators for neuropsychiatric and gastrointestinal disorders. *Med. Res. Rev.* (2020).

466. Bhalerao, Y. P. & Wagh, S. J. A review on Thymol encapsulation and its controlled release through biodegradable polymer shells. *Int. J. Pharm. Sci. Res.* **9**, 4522–4532 (2018).

467. Subramaniyam, S. *et al.* Oral Phyto-thymol ameliorates the stress induced IBS symptoms. *Sci. Rep.* **10**, 13900 (2020).

468. Ziemba, P. M. *et al.* Activation and modulation of recombinantly expressed serotonin receptor type 3A by terpenes and pungent substances. *Biochem. Biophys. Res. Commun.* **467**, 1090–1096 (2015).

469. Lansdell, S. J., Sathyaprakash, C., Doward, A. & Millar, N. S. Activation of human 5-hydroxytryptamine type 3 receptors via an allosteric transmembrane sites. *Mol. Pharmacol.* **87**, 87–95 (2015).

470. de Oliveira-Pierce, A. N., Zhang, R. & Machu, T. K. Colchicine: a novel positive allosteric modulator of the human 5-hydroxytryptamine3A receptor. *J. Pharmacol. Exp. Ther.* **329**, 838–847 (2009).

471. Price, K. L., Hirayama, Y. & Lummis, S. C. R. Subtle Differences among 5-HT 3 AC, 5-HT 3 AD, and 5-HT 3 AE Receptors Are Revealed by Partial Agonists. *ACS Chem. Neurosci* **8**, (2017).

472. Huey, R. & Morris, G. M. *Using AutoDock with AutoDockTools: A Tutorial*. http://mgltools.scripps.edu/downloads/previous-releases/downloads/tars/releases/DocTars/DOCPACKS/AutoDockTools/doc/UsingAutoDockWithADT.pdf.

473. Reeves, D. C., Sayed, M. F. R., Chau, P.-L., Price, K. L. & Lummis, S. C. R. Prediction of 5-HT3 receptor agonist-binding residues using homology modeling. *Biophys. J.* **84**, 2338–2344 (2003).

474. Verheij, M. H. P. *et al.* Design, synthesis, and structure-activity relationships of highly potent 5-HT3 receptor ligands. *J. Med. Chem.* **55**, 8603–8614 (2012).

475. Hai Nguyen, Daniel R. Roe, Jason Swails, D. A. C. Interactive data analysis for molecular dynamics simulations. Hai Nguyen, Daniel R. Roe, Jason Swails, David A. Case. (2016)[1] PYTRAJ: Interactive data analysis for molecular dynamics simulations. (2016).

476. Thompson, A. J. Recent developments in 5-HT3 receptor pharmacology. *Trends Pharmacol. Sci.* **34**, 100–109 (2013).

477. Maiorov, V. N. & Crippen, G. M. Size-independent comparison of protein three-dimensional structures. *Proteins Struct. Funct. Genet.* **22**, 273–283 (1995).

478. Abad-Zapatero, C. Ligand efficiency indices for effective drug discovery: a unifying vector formulation. *Expert Opinion on Drug Discovery* vol. 16 763–775 (2021).

479. Xue, J. *et al.* Antimalarial and Structural Studies of Pyridine-Containing Inhibitors of 1-Deoxyxylulose-5-phosphate Reductoisomerase. *ACS Med. Chem. Lett.* **4**, 278–282 (2012).

480. Konzuch, S. *et al.* Binding modes of reverse fosmidomycin analogs toward the antimalarial target IspC. *J. Med. Chem.* **57**, 8827–8838 (2014).

481. Deng, L. *et al.* Structures of 1-deoxy-D-xylulose-5-phosphate reductoisomerase/lipophilic phosphonate complexes. *ACS Med. Chem. Lett.* **2**, 165–170 (2011).

482. Saggu, G. S., Pala, Z. R., Garg, S. & Saxena, V. New Insight into Isoprenoids Biosynthesis Process and Future Prospects for Drug Designing in Plasmodium. *Front. Microbiol.* **7**, 1421 (2016).

483. Truong, D. T., Nguyen, M. T., Vu, V. V. & Ngo, S. T. Fast pulling of ligand approach for the

design of B-secretase 1 inhibitors. *Chem. Phys. Lett.* **671**, 142–146 (2017).

484. Masini, T., Kroezen, B. S. & Hirsch, A. K. H. Druggability of the enzymes of the non-mevalonate-pathway. *Drug Discov. Today* **18**, 1256–1262 (2013).

APPENDIX

Appendix A Some PfDXR residues and their identified/suggested roles from literature

PfDXR Residues	Role	References
SER269, ASN311,SER270, SER306, LYS312, HIS293	Bind phosphonate moiety	319,339,479
GLY87, THR86, SER88, LYS116, ILE89, ASN115, , SER117, GLY299, GLU206	Co-factor NADPH binding	339
GLU315, ASP231, GLU233,	Metal coordination and fosmidomycin hydroxamate group binding	250,319,339,479,480
HIS293	May to pre-orientat THE ligand in the binding site and for loop closure.	242
PRO294	Important for maintaining the structure of the flexible loop.	339
GLY299	Suggested to contribute to the flexible loop flexibility	339
MET298	Hydrophobic interactions with NADPH nicotinamide moiety and inhibitors backbone. M298A or M298V mutations may impair substrate (DXP) binding and its turnover.	56
TRP296	DXR inhibitors discrimination, Cover binding site and interact with inhibitor Better interaction with electron-deficient and hydrophobic group.	292,339,481
Linker region	Support the catalytic domain	482

1 **Appendix B SFs rescoring summary statistics.**

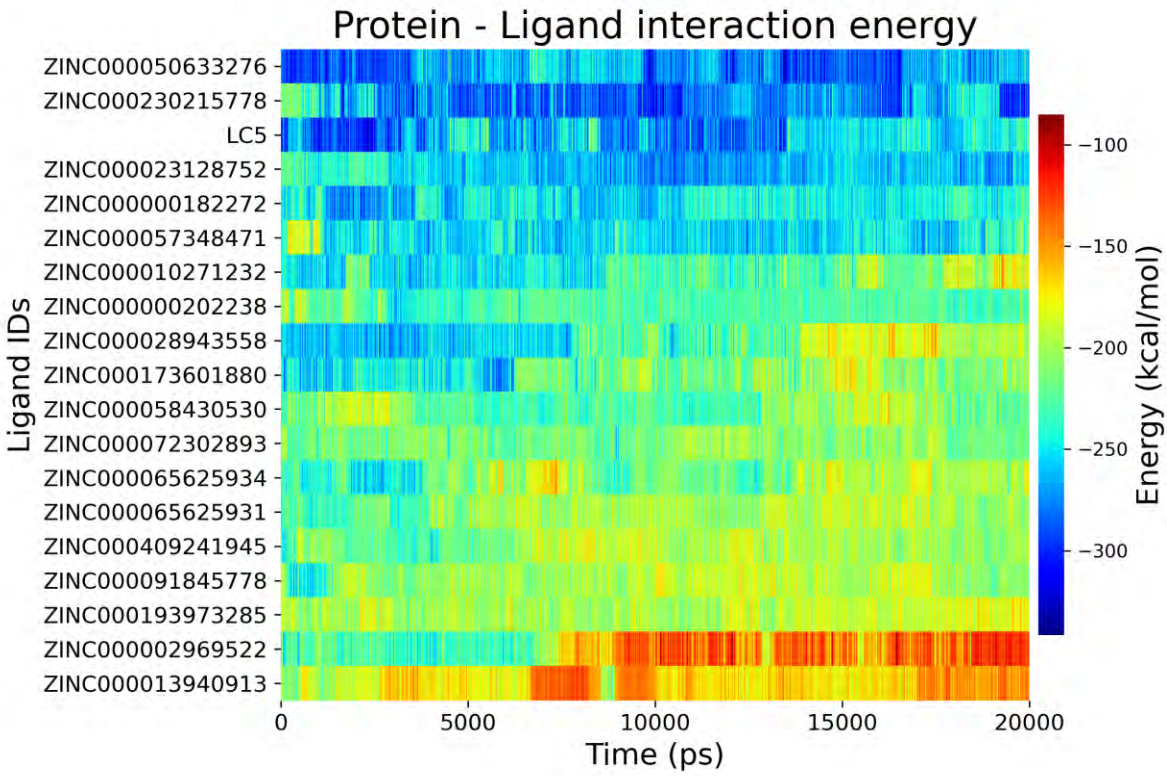
2 Descriptive statistics for all scores for the successfully rescored 48972 ligands.

SFs	mean	std	min	25%	50%	75%	max
Rf-score_V1	6.74	0.68	5.02	6.24	6.82	7.24	8.69
Vina	-7.11	0.64	-10.25	-7.53	-7.08	-6.68	-4.51
NNScore	5.47	0.81	3.02	4.89	5.38	6	7.88
Rf-score_V2	6.52	0.46	4.97	6.1	6.61	6.88	7.75
Rf-score_V3	6.37	0.53	4.68	5.96	6.44	6.8	8.13
PLEC	3.95	1	0.53	3.25	3.91	4.61	8.29

Rf-score_V4	6.32	0.45	4.42	6.06	6.38	6.64	7.74
Idock	-7.39	0.58	-10.09	-7.76	-7.35	-6.98	-5.08
Cyscore	-2.71	0.64	-4.94	-3.15	-2.72	-2.28	0.02
DSX	-83.96	12.83	-135.4	-92.49	-83.79	-75.26	-34.96
Smina	-7.53	0.59	-10.33	-7.91	-7.49	-7.11	-4.91
AutoDock	-24.34	4.92	-43.32	-27.72	-24.66	-21.24	4.15
Xscore	-7.92	0.39	-9.61	-8.18	-6.5	-7.92	-7.66

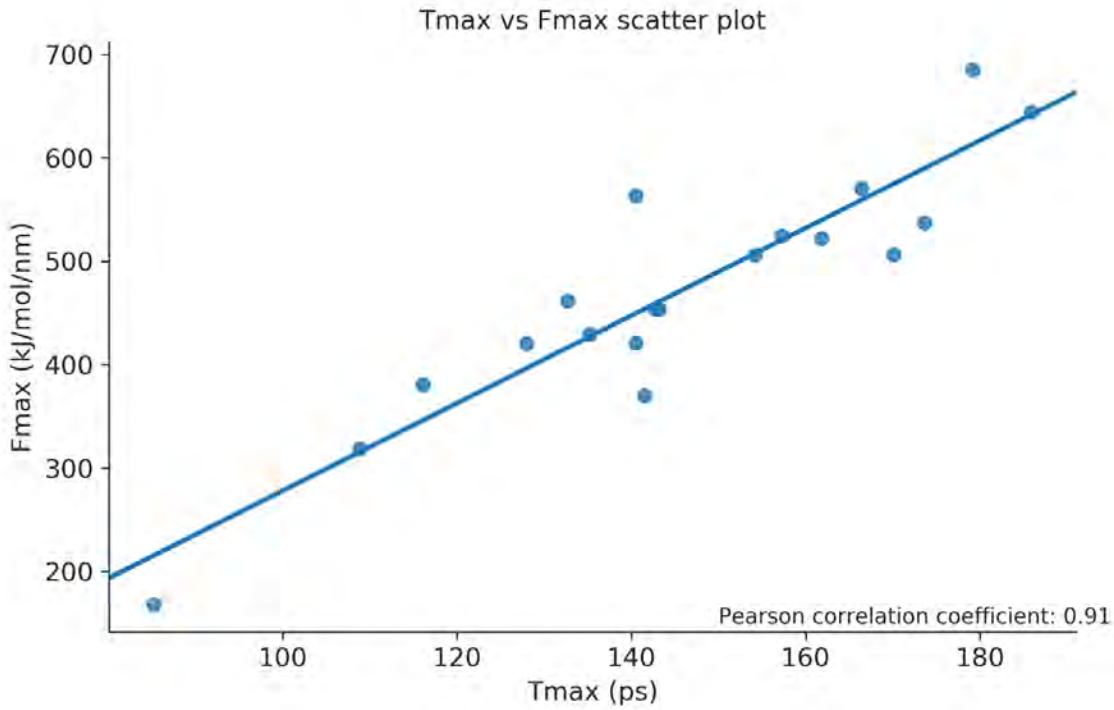
Appendix C Protein-ligand interaction energy during 20 ns MD

Heatmap of the PLIE for the 18 ligands during the 20 ns simulations. Ligand are ranked according to the average PLIE. The colors is scaled to the minimum and maximum of the data. PLIE is given in kcal/mol unit. The figure was generated using Seaborn [198].



Appendix D Correlation between Fmax and Tmax

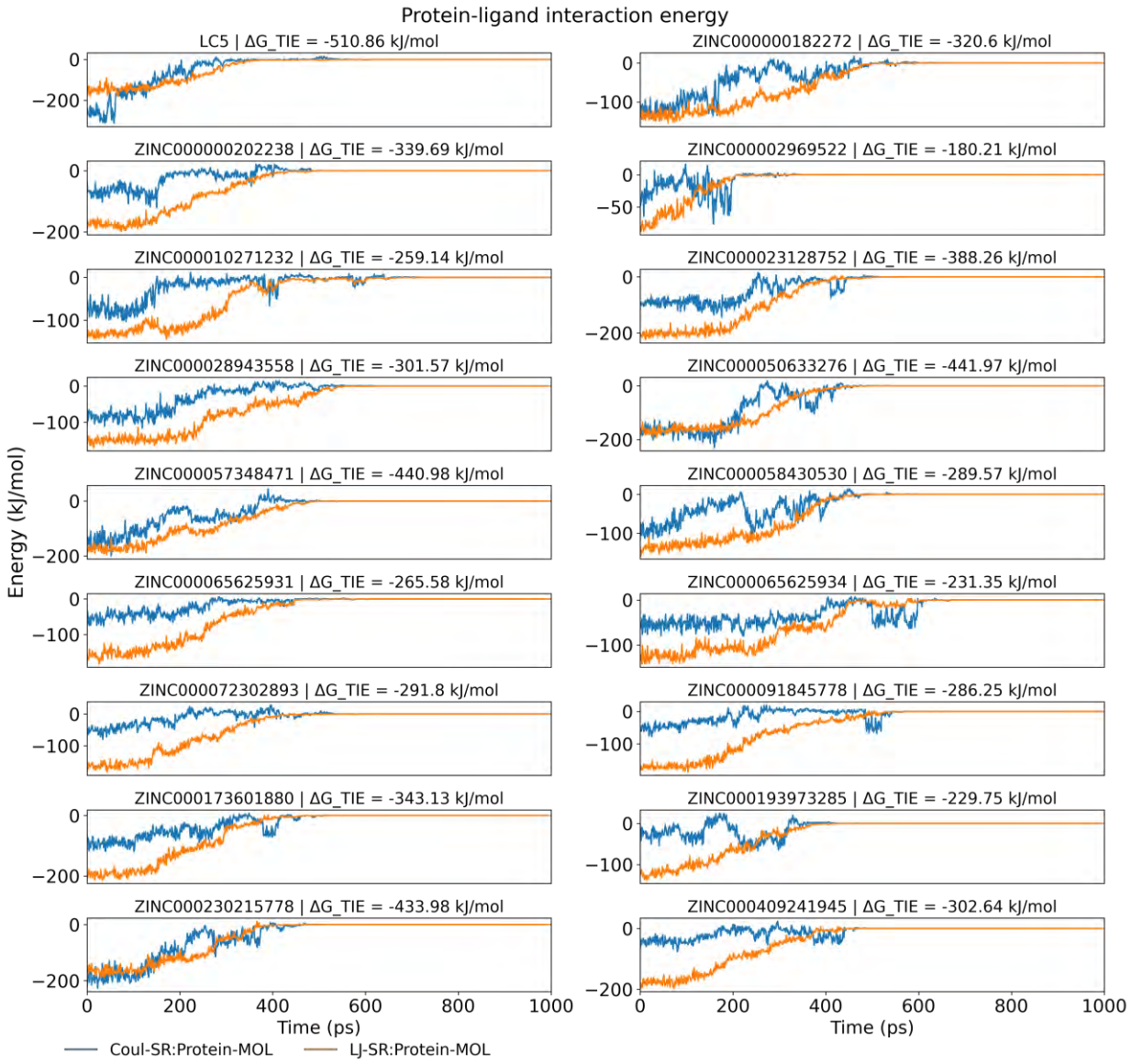
Tmax vs Fmax scatter plot. Every point represents a compound. Every point is a compound. The figure was prepared using Seaborn [198].



A scatter plot showing the correlation between the rupture force (Fmax) and the corresponding time points (Tmax). Each point represents a compound with its Tmax plotted on the x-axis and Fmax on the y-axis. The blue line represents the regression line.

Fmax was strongly correlated to Tmax with a Pearson correlation coefficient of 0.91. This high correlation was only observed on the average of the force-time profiles from the 10 SMD simulations, not on the individual trajectories. The Pearson correlation for the individual trajectories was much lower. Sampling provided convergence of the results, but also this is related to the constant velocity (cv), constant force (cf) nature of the SMD simulation. Fmax-Tmax correlation can hence be used to assess the sampling when combining multiple trajectories in cv-cf SMD. Indeed multiple trajectories have been already often used in SMD in previous studies [310,331,483].

Appendix E Protein-ligand interactions energies (short-range and Lennar Jones) during SMD



Appendix E shows the changes in the interaction energy between PfDXR and the different ligands over the time course of the first SMD simulation. The interaction energy results from the potential energy decomposition to only include nonbonded terms between the protein and the ligand [106]. It is the sum of the electrostatic (short-range) and van der Waals (vdW) interaction energy components extracted from an energy file using the gmx energy module. It is important to note here that these terms decomposition, does not hold physical meaning regardless of the force field used. And only the total interaction energy is useful when the force field has been parametrized in such a way[106].

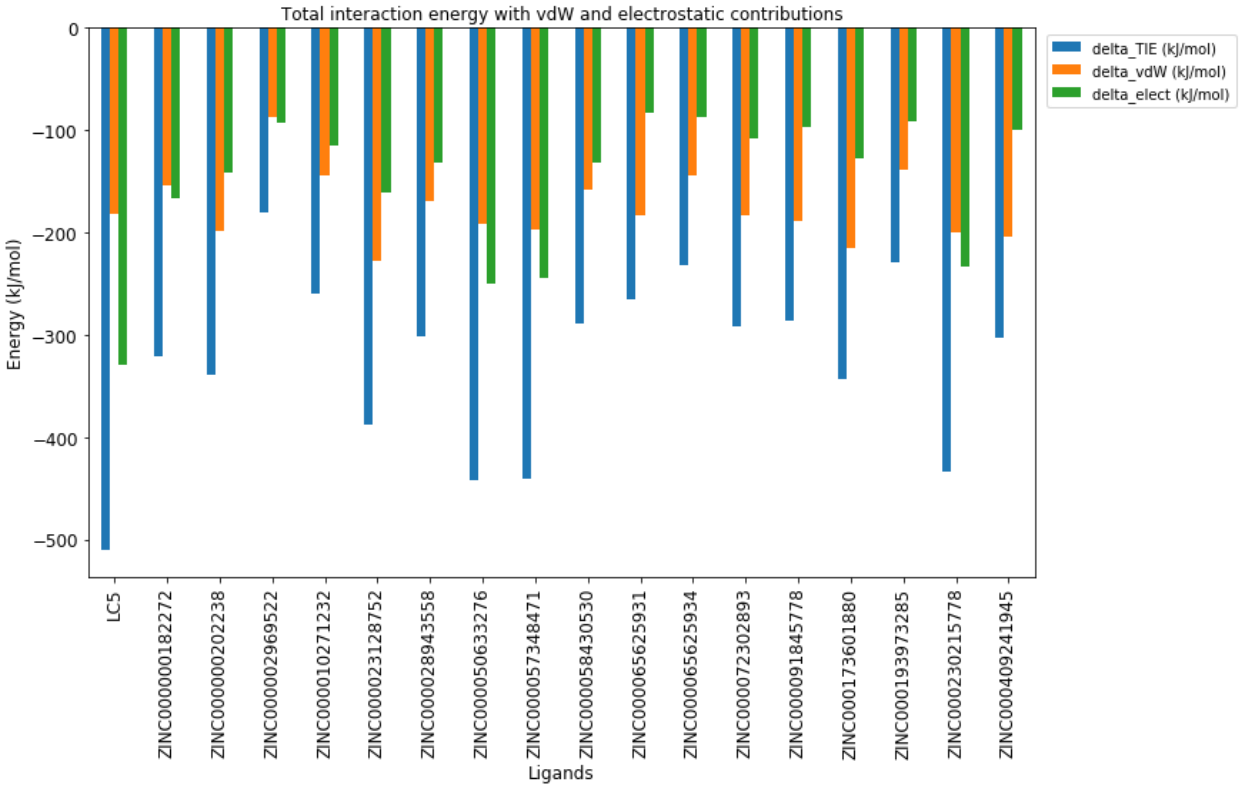
The two time-series are correlated showing similar increasing trend to zero. Both curves are only stable in the early stage of the simulation (about first 100 ps) before starting to increase during the unbinding where interactions are broken. Lowest and maximum values are recorded at 0 ps

and about 500 ps for the different systems. The electrostatic curve spikes after vdW one attained zero in some systems. These two time points correspond to the most stable, with full interaction between the protein and the ligand and the completely solvated state of the ligand respectively.

Truong *et al.* showed that the total interaction energy difference (ΔG_{TIE}) from SMD was the best metric to evaluate the relative binding affinity of the BACE1 inhibitors that comparing it to the rupture force Fmax and to the pulling work (Wpull) [483]

The electrostatic contribution curve was less smooth than the vdW one. Indeed, it shows sharper variations. This can be linked to the stronger nature of the electrostatic interactions than the vdW ones resulting in higher energy variation when formed or broken. In general, vdW interaction energy shows higher contributions than the electrostatic one across the different ligands.

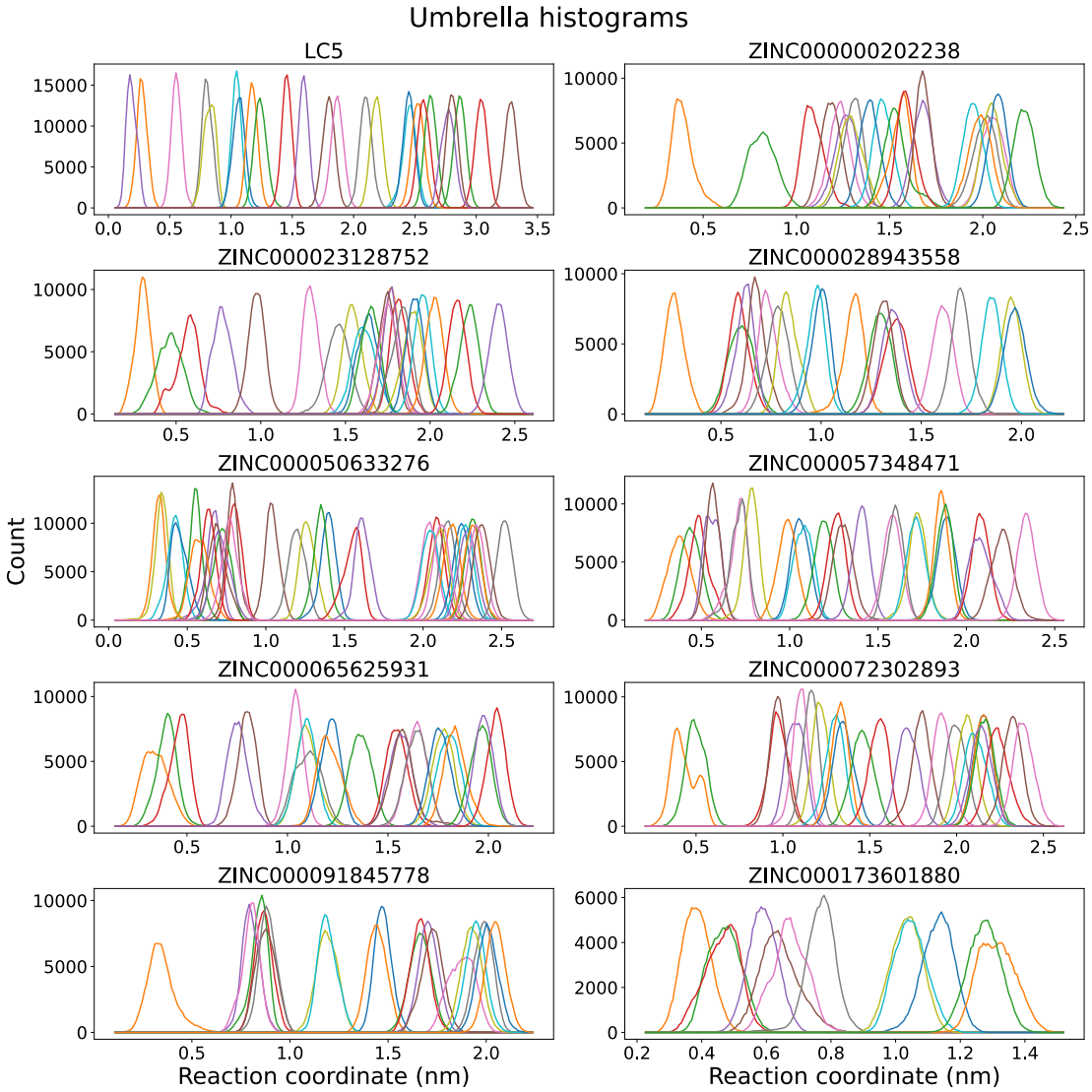
Appendix F Total interaction energy with vdW and electrostatic contributions



Appendix F shows the total interaction energy difference (ΔG_{TIE}) from SMD. $\Delta G_{TIE\ was}$ extracted from the first SMD trajectory only. The $\Delta G_{TIE\ across}$ the different systems range between -180.21 KJ/mol and -510.86 KJ/mol. LC5 showed the best ΔG_{TIE} value with -510.86 KJ/mol, resulting from a significantly higher electrostatic contribution -328.79 KJ/mol which was greater than the double of the average electrostatic contribution of all the ligands. Assuming the concept that the stronger binding inhibitor has stronger non- bonded contact [483], hence the co-crystallized ligand shows the highest affinity here. Interestingly, LC5 showed the best energy value only with the ΔG_{TIE} but not in MMPBSA, US, Fmax or Wpull.

All ΔG_{TIE} are negative, showing favorable interactions between PfDXR and the different ligands. vdW has a higher contribution to ΔG_{TIE} thank the electrostatic interactions an average of -325.40 KJ/mol. Interestingly, ligands with higher electrostatic contribution than vdW seems to perform better. LC5, ZINC000050633276, ZINC000057348471, ZINC000230215778 are the top ligands ranking by the ΔG_{TIE} with respectively -510.86 KJ/mol, -441.97 KJ/mol, -440.98 KJ/mol and -433.98 KJ/mol. The same ligands also have the highest electrostatic contributions to their binding but also have better electrostatic than their vdW energy. This can be related to the rather hydrophilic nature of the binding site. Indeed the binding site has a low ratio of apolar amino acids (0.36) [484] with the phosphate moiety binding in a charged region and the hydroxamate moiety in a hydrophilic area. Electrostatic interactions thus provide room for improvement in order to potentiate ligand binding affinity to DXR.

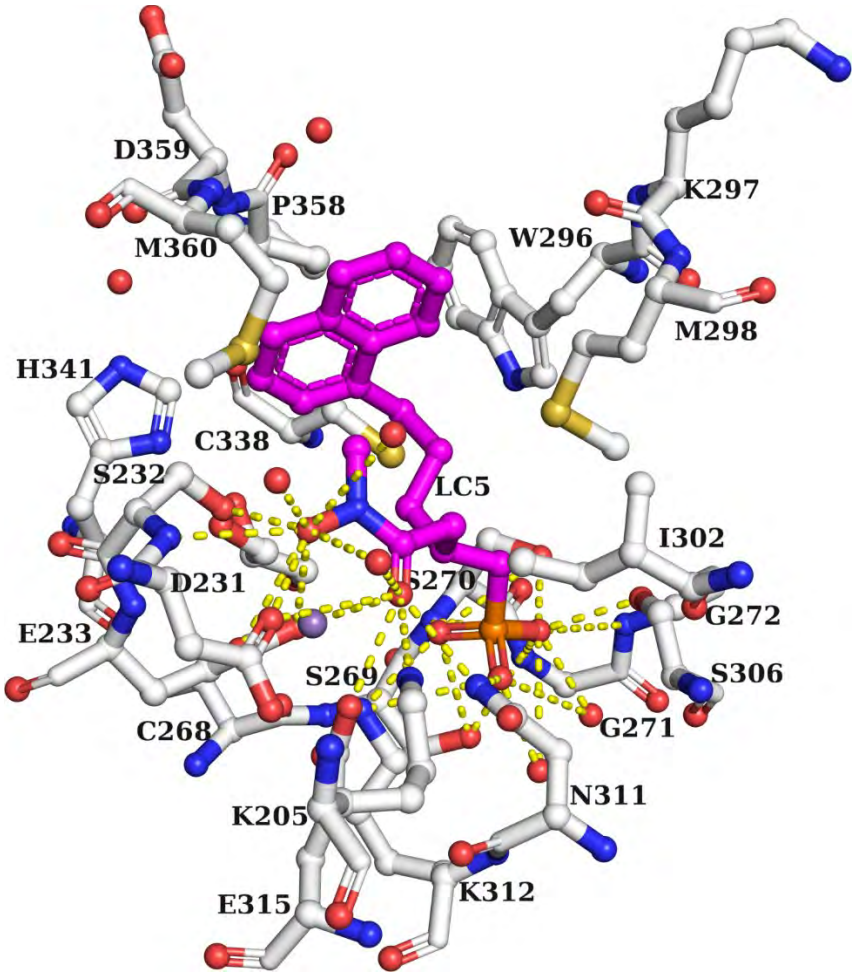
Appendix G Umbrella histograms



Appendix G shows the umbrella histograms for the different ligands. X-axis is the reaction coordinate (Protein-ligand COM distance) the Y one represents the count. Colors do not have meaning.

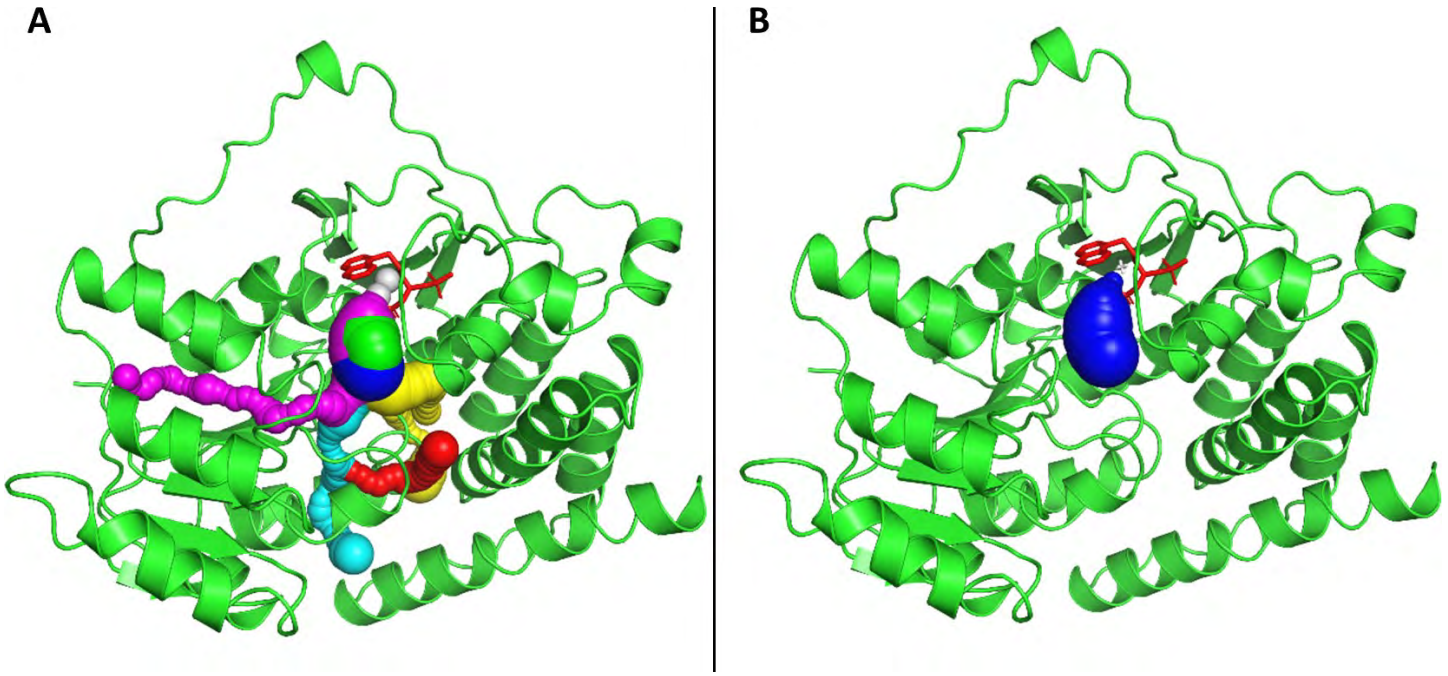
The length of the reaction coordinate was about 3 nm. The histograms provide sufficient overlap for effective sampling of the entire reaction coordinate. We although note some poorly sampled regions for some ligands: ZINC000091845778 (around 0.6nm), ZINC000072302893 (0.6nm), ZINC000173601880 (0.9nm), ZINC000023128752 (1.1nm), ZINC000000202238 (0.6nm and 1nm). However, these lacks sampling in a few points in the reaction coordinate are not likely to influence the final free energy difference.

Appendix H LC5 binding pose in PfDXR active site



LC5 (magenta) crystal structure in PfDXR (5JAZ) active site. Residues in a radius of 3.5 Ångströms light grey of LC5 are labelled with their residue numbers and one-letter code. White indicate carbon atoms and other element are in atom types color. They are drawn in stick. Polar contacts with the ligand are displayed in dashed lines in yellow. LC5 hydroximate group coordinates Mn^{2+} (violet) in addition GLU233, ASP231, GLU315 and two waters (red balls). Its phosphonate moiety bind in a polar region (ASN311, SER306, SER269, SER270). Its aromatic ring extend toward HIS293, TRP296, PRO358 , MET298, CYS338 [289]. The representation was produced using Pymol [225]. Interaction were generated using the show_contacts script [226].

Appendix I LC5 pulling pathways from PfDXR binding site.



The figure in Appendix I shows the pulling pathways determined using Caver 3.0.1 Pymol plugin. PfDXR is in green ribbon with LC5 in red stick.. **A.** All possible unbinding paths (in the different colors) identified by Caver. **B.** The best unbinding path (blue). The path is shorter and is less curved than the other pathways, hence potentially more energetically favourable. The figure was generated using Pymol [225].

Appendix J Best poses binding energies in docking on full structures (Extracellular and Membrane domains)

Proteins	Best pose binding energies (kcal/mol)			
	Serotonin	Tropisetron	Thymol	NAG
6HIQ	-8.0	-8.9	-7.9	-6.7
6HIS	-6.7	-9.3	-6.5	-5.6
6HIN	-7.8	-8.0	-7.9	-6.4
6HIO	-.7.8	-8.6	-8.2	-6.5
4PIR	-6.6	-.8.8	-6.9	-5.6

Appendix K Serotonin and thymol docked extracellular domain binding energies.

Proteins	Best pose binding energies (kcal/mol)		
	Serotonin (Crystal)	Serotonin (redocked)	Thymol
6HIQ	- 6.4	-8.0	-7.9

6HIS Tropisetron (Crystal)	- 4.6 Tropisetron (Crystal)	-9.2	-6.5
6HIN	- 5.9	-7.9	-8.0
6HIO	-5.8	-7.8	-8.2

105

106 **Appendix L Blind docking parameters**

107 Blind docking parameters for full structures (FS) (Extracellular + Membrane domains) and ECD. The docking
108 was first performed on the ECD and later on the FS to investigate potential membrane domain binding.

Receptor	4PIR FS	6HIN FS	6HIO FS	6HIQ FS	6HIS FS	6HIN ECD	6HIO ECD	6HIQ ECD	6HIS ECD
center_x	154.29	124.68	124.68	128.05	124.69	124.68	124.68	128.07	124.69
center_y	203.38	124.68	124.67	128.06	124.69	124.68	124.68	128.05	124.69
center_z	265.85	137.73	129.25	131.77	125.23	156.34	156.67	159.61	153.22
size_x	147.95	87.68	85.02	85.71	85.17	82.68	80.02	80.71	80.17
size_y	136.67	87.48	84.98	85.59	84.97	82.48	79.98	80.59	79.97
size_z	167.21	117.05	155.93	157.93	161.94	67.42	66.99	68.35	69.71
exhaustiveness	6000	6000	6000	6000	6000	529	512	516	513
CPU	24	24	24	24	24	4	4	4	4

109

110 Appendix M Serotonin, tropisetron and thymol interacting residues.

111 Common interacting residues are in color.

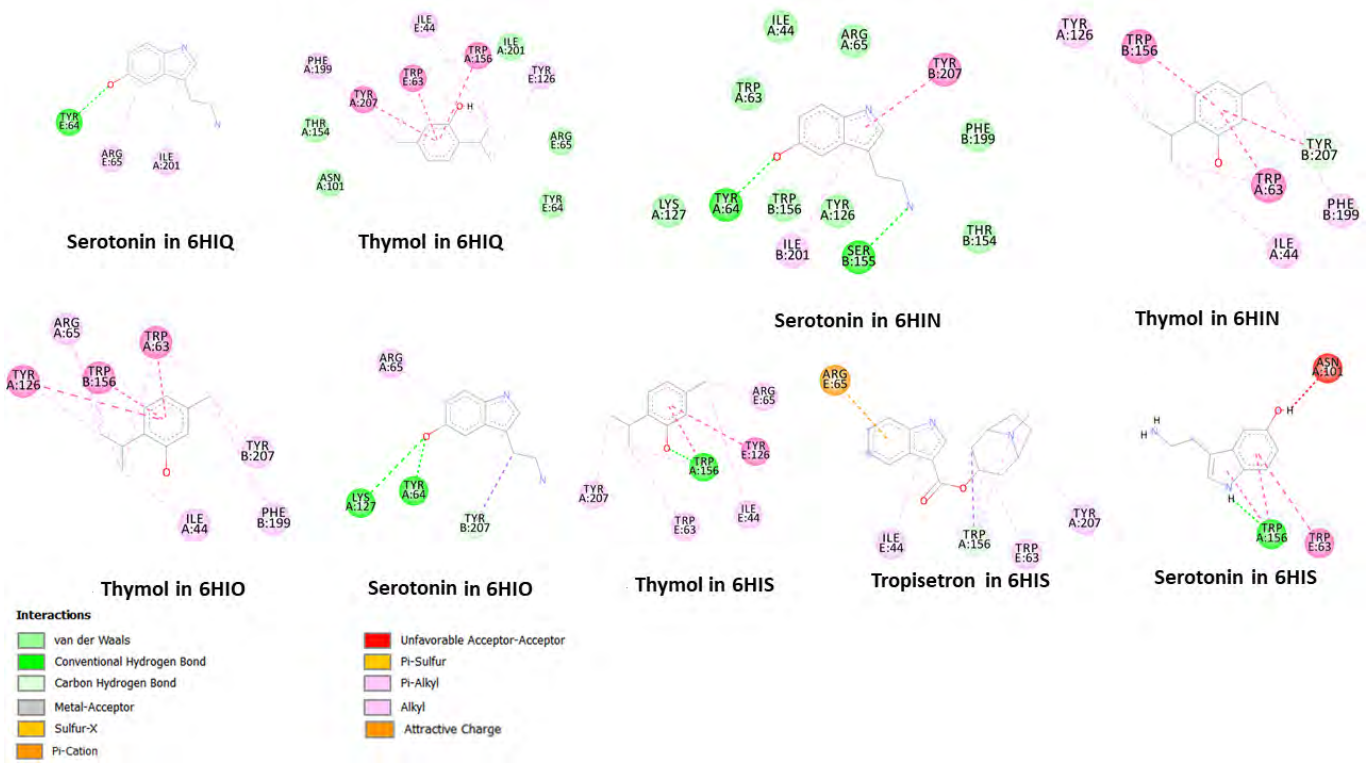
Proteins	Interacting residues	
	Serotonin (Crystal)	Thymol
6HIQ	A-ILE201 E-ARG65 E-TYR64	A-TRP156 A-TYR207 A-PHE199 A-ASN101 A-THR154 A-ILE201 E-TYR126 E-TYR64 E-TRP63 E-ARG65 E-ILE44
6HIS	Serotonin (Docked) A-TRP156	A-TRP156 A-TRP207

	A-ASN101 E-TRP63 Tropisetron (Crystal) A-TRP156 A-TYR207 E-ARG65 E-ILE44 E-TRP63	E-ARG65 E-ILE44 E-TRP63 E-TRP126
6HIN	A-ILE44 A-ARG65 A-TRP63 A-TYR126 A-TYR64 A-LYS127 B-THR154 B-PHE199 B-TYR207 B-ILE201 B-TRP156 B-SER155	A-TRP63 A-TYR126 A-ILE44 B-TRP156 B-TYR207 B-PHE199
6HIO	A-ARG65 A-LYS127 A-TYR64 B-TYR207	A-ARG65 A-TRP63 A-TYR126 A-ILE44 B-TRP156 B-TYR207 B-PHE199

112

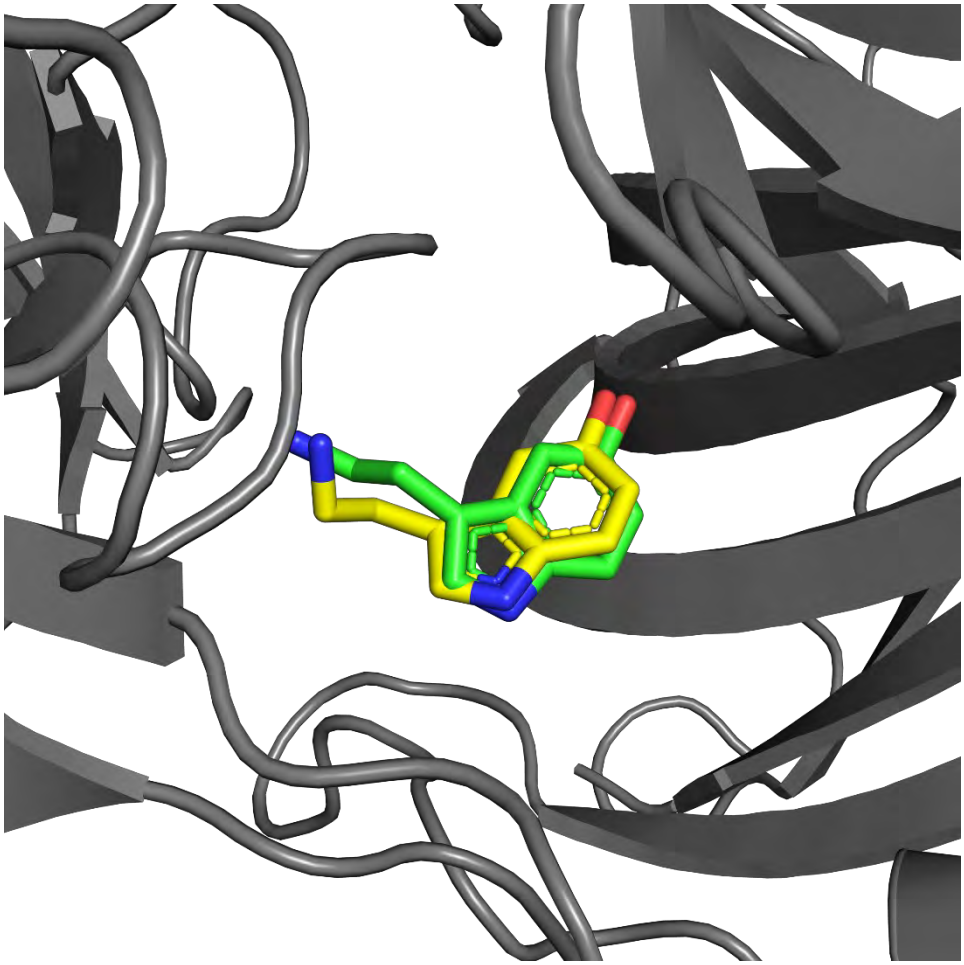
113 **Appendix N 2D interaction plot of receptor-ligand complexes.**

114 Only ligand poses taken to MD simulation are shown. The 2D plots are obtained from Discovery
115 Studio Visualizer V1.7.2.0.16349.



Appendix O Serotonin docking validation.

Crystalized serotonin was redocked to the structure (green) and compared to the blind docking result (yellow). The RMSD value between crystalized and docked serotonin was RMSD 2.2 Å.



Appendix P Thymol, serotonin, and tropisetron docked in 4PIR.

(A) Cartoon representation of serotonin receptor (PDB ID: 4PIR) with docked serotonin (in green), thymol (in cyan), and tropisetron in magenta. B. Thymol, serotonin, and tropisetron docked in 6HIQ. (A) Cartoon representation of serotonin receptor (PDB ID: 6HIQ) with docked serotonin (in green), thymol (in cyan), and tropisetron in magenta. C. Docked thymol (face view) at two subunits interface forming the binding

site. Serotonin in green and thymol in yellow D. Docked Thymol (top view) Serotonin in green and thymol in yellow.