College of Medicine

STRUCTURE-BASED *IN SILICO* IDENTIFICATION OF NOVEL ANTICANCER NATURAL COMPOUNDS FOR ENHANCED BREAST CANCER CHEMOTHERAPY

By

Jonathan Tatenda Bvunzawabaya

(Bachelor of Science (Chemistry))

A Thesis Submitted to the Department of Biomedical Sciences in Fulfilment of the Requirements of the Master of Philosophy in Biomedical Sciences Degree

January 2022

CERTIFICATE OF APPROVAL

The thesis of **Jonathan T. Bvunzawabaya** is approved by the Thesis Examination Committee:

……..

(Chairman, Post Graduate Committee)

…….

Prof. Fanuel Lampiou (Primary Supervisor)

……

(Head of Department)

DECLARATION

I **Jonathan T. Bvunzawabaya**, hereby declare that this thesis, "Structure-based *in silico* identification of novel anticancer natural compounds for enhanced breast cancer chemotherapy" is my own work that was performed under the supervision of Professor F. Lampiao and Professor G. Mugumbate and it has not been submitted before for a degree or examination at the University of Malawi or any other University. All resources I have used have been acknowledged or referenced.

Name of Candidate: Jonathan T. Bvunzawabaya

Signature: 

Date: January 2022.

DEDICATION

This work is dedicated to my son, Atidaishe, and the Almighty.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to the following, without whom this journey would have been hard:

My Primary Supervisor, Professor F Lampiao, for sharing with me his wisdom and experience in research throughout the course of this study.

My secondary Supervisor, Professor G. C Mugumbate, for providing me motherly support, guidance in research and for always having faith in me from the beginning of work.

Dr. A. G. Mtewa, for offering opportunities and training in academic research writing, collaboration, and presentation.

Biomedical Sciences Department (Kamuzu University of Health Sciences), for allowing me to study for my degree and for providing a friendly environment that is conducive for research and learning

My family and friends, who were always available to encourage and support me all the way,

Finally, yet importantly, **The Almighty God**, for His grace, wisdom, knowledge, and skill, and for providing all resources needed the work to be carried out.

Financial Support

I am grateful to the Africa Centre of Excellence in Public Health and Herbal Medicine (ACEPHEM) for the academic scholarship.

OUTPUTS FROM THE STUDY

1. **Oral presentation** titled "Structure-based virtual screening and Pharmacokinetic analysis of potential anti breast cancer natural compounds from Africa" at Biodiversity for Malawi 2063: Research Dissemination Conference, Mulanje, Malawi. 20th-22th July 2021. Awarded best oral presentation for the conference.
2. **Poster presentation** titled" Chemical space and Chemical diversity profiling of drug-like compounds from African natural product databases for Fragment based-drug discovery" at Biodiversity for Malawi 2063: Research Dissemination Conference, Mulanje, Malawi. 20th-22th July 2021. Awarded best poster presentation for the conference. (**Appendix 2**)
3. **Oral presentation** titled "The Designing of synthetic drug scaffolds from existing libraries and natural products towards acceptable African pharmaceutical response to pandemics" at The Africa Research and Impact Network International Conference on Africa in the Post Covid-19 World: Lessons for Research and Policy 18th – 20th November 2020, Nairobi, Kenya.
4. **Research manuscript (drafted)** Identification of natural inhibitors for γ-secretase enzyme from African natural product databases. An in silico approach in combating multi-drug resistance in breast cancer chemotherapy. (**Appendix 3**)
5. **Research manuscript (drafted)** Chemoinformatic profiling of drug-like compounds from African natural product databases for lead prioritization and discovery. (**Appendix 4**)
6. **Book chapter (published)** Molecular optimization of phytochemicals into antidotes doi.org/10.1016/B978-0-12-821556-2.00006-2. (**Appendix 5**)
7. **Book chapter (published)** Applications of phytochemicals against nerve agents in counterterrorism. doi.org/10.1016/B978-0-12-821556-2.00020-7. (**Appendix 6**)
8. **Book chapter (accepted)** The Designing of synthetic drug scaffolds from existing libraries and natural products towards acceptable African pharmaceutical response to pandemics.
9. **Research manuscript (published)** Ligand-Protein interactions of Plant-isolated (9z,12z)-Octadeca 9,12-dienoic Acid with B-Ketoacyl-Acp Synthase (Kasa) in Potential Anti-Tubercular Drug Designing. doi.org/10.1016/j.sciaf.2021.e00824. (**Appendix 7**)
10. **Research manuscript (submitted)** Flexible Docking-Based Screening of African Natural Products for New Drug Leads against Enoyl-acyl Carrier Protein Reductase of C.

trachomatis (CtFabI): An In-silico Approach to Neglected Tropical Disease Drug Discovery. (**Appendix 8**)

ABSTRACT

Breast cancer remains a serious public health concern all over the world with heavier burdens on developing countries. In Sub-Saharan Africa, the mortality rate for breast cancer is currently on the rise because of late diagnosis and poor treatment. Currently available chemotherapeutic drugs for breast cancer are associated with severe side effects and are now facing multi-drug resistance. Therefore, identifying new anti-cancer drugs by using structural information of potential drug targets such as the gamma secretase will enhance the fight against breast cancer. The objective of this study was to identify and characterize drug-like natural anti-breast cancer compounds from selected African natural products databases using structure-based virtual screening and chemoinformatic approaches. Exactly11304 compounds from four databases (Afrodb, NANPDB, Afrocancer, and ConmedNP) were curated and filtered to remove structural alerts and compounds that violate drug-like rules according to Lipinski and Veber, using KNIME analytics. The resulting druglikeNP dataset (437 compounds) had its, chemical space, scaffold diversity, and complexity analyzed using scaled Shannon entropy and cyclic system retrieval curves (CSR). The 437 compounds were docked into the binding site of gamma-secretase enzyme and the pharmacokinetic properties of the hit compounds were profiled using the pkCSM server. 60% of the compounds in the druglikeNP dataset contained lead-like physicochemical properties and occupied the same space as FDA-approved drugs. The scaffold diversity of druglikeNP was observed to be higher than that of FDA drugs based on Shannon entropy and CSR. Docking studies identified 12 compounds as potential inhibitors of the gamma-secretase enzyme (with binding energies ranging between -7.6 and -8.8 KCal/mol). *In silico* ADMET predictions revealed a majority of the 12 hit compounds have good pharmacokinetic and toxicity profiles. In this study

12 drug-like natural compounds of African origin with potential inhibitory properties against a validated breast cancer target; gamma secretase enzymes were computationally identified. This work could be valuable to the ongoing efforts to discovery novel drugs for enhanced breast chemotherapy.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND ACRONYMS

COMREC	College of Medicine Research and Ethics Committee
TNBC	Triple negative breast cancer
GSIs	*γ-secretase* inhibitors
PS1	Presenilin 1
NP	Natural product
CADD	Computer-aided drug design
SBDD	Structure-based drug design
LBDD	Ligand-based drug design
ADMET	Absorption distribution, metabolism and Excretion
SSE	Shannon scaled entropy
KNIME	Konstanz information miner
t-SNE	t-distributed stochastic neighbor embedding
PUMA	Platform of unified molecular analysis
CSR	Cyclic System Retrieval
CDF	Cumulative distribution frequency
PAINS	Pan Assay Interference Compounds
FDA	Food and Drug administration
TPSA	Topological polar surface area
MW	Molecular weight
MR	Molar refractivity

INTRODUCTION

1.1 Background

Breast cancer (BC) is the most common invasive cancer in women, and continues to be a global health burden in both transitioned and transitioning countries. In 2020 it was classified as the most commonly diagnosed cancer with approximately 2.3 million cases and 684 996 deaths in 185 countries (1,2). Although in Africa incident cases are relatively low, breast cancer is the second leading cause of death from cancer in women in sub-Saharan Africa because of low survival rates in patients (3). This disease is characterized by the uncontrolled growth of epithelial cells originating in the ducts or breast lobule in women (4). Chemotherapy is currently one of the most commonly used methods in BC treatment and has been associated with increased chance of cure in addition to other treatments (5). However current chemotherapeutic drugs are linked to two primary setbacks; severe side effects and lately multi-drug resistance, warranting the search for more options (5). In addition to these setbacks, metastatic or advanced BC is not considered curable using currently available chemotherapeutic options. Rapid discovery of new novel drugs which target unique cellular pathways that are responsible of multi-drug resistance and proliferation of tumorous cells is a promising approach in breast cancer management (6). The notch signaling pathway, regulated by γ-secretase enzyme is one of the cellular pathways that has recently emerged as a propitious breast cancer drug target (7). On the other hand, natural products have been a bedrock of drug discovery for many decades. It is estimated that more than half of all cancer drugs and antibiotics are natural product-inspired (8). Today more than 400 000 natural compounds have been elucidated from various organisms; plants, fungi, bacteria, and animals, and their chemical information is readily available to the drug research community (9).

Chemoinformatics and computer-aided drug discovery techniques have formerly proved to be useful in prioritizing bioactive compounds for clinical screening saving time and money (10) are providing alternative ways of exploring the natural product databases for new anti-cancer drugs. With the rising interest in fragment-based drug discovery, chemoinformatics approaches have assisted in characterizing the drug-likeness, fragment-likeness and chemical diversity of natural products (11). While other computer aided methods such as structure-based and ligand-based virtual screening are assisting in identifying potentially active drug hits for clinical studies *in silico* pharmacokinetic screening methods have also complimented other computational methods as they have been employed to predict the pharmacokinetic properties of drug hits prior to clinical studies (12).

1.2 Problem statement

Breast cancer is the most common cancer in women, with the highest mortality rate in women in sub-Saharan African and worldwide (13). Five-year survival estimates for breast cancer patients in Africa are estimated to be very low on average (less than 50%) chiefly because of poor treatment and diagnosis. Current chemotherapeutic approaches in breast cancer management involve the use of drugs that exhibit non-selectivity between cancerous cells and normal DNA tissue cells which is results in severe toxicity, and adverse side effects in cancer patients (5). The Notch signaling pathway regulated by γ-secretase enzyme promotes the development of drug-resistant breast cancer cells which results in possible regeneration or recurrence of the disease thus lessening the effectiveness of the cancer drugs (7). To date, there has not been any clinically approved drug that regulates the notch signaling pathway and preventing cancer stem cell proliferation (14). Advanced stages of breast cancer still lack treatment to date, as all available treatments are

rendered ineffective in stopping metastasis. African natural products databases provide a promising source of drug-like compounds that have not been screened and characterized for breast cancer drug development purposes using chemoinformatic approaches. On the other hand, the process of drug discovery takes about 10-15 years, costing between 500-800 million dollars for one drug to be on the market and hence the computational discovery of novel anti-cancer drugs (15).

2. LITERATURE REVIEW

2.1 The global burden of breast cancer

Breast cancer is the most prevalent and devastating malignancy among women worldwide (16). At a molecular level, the disease is characterized by rapid, uncontrollable, irreversible, independent, autonomous, uncoordinated, and relatively unlimited, and abnormal growth of breast tissues. Rapid creation of abnormal cells occurs in the milk ducts and glands, these grow beyond their boundaries leading to the formation of a tumor that invades adjacent normal tissues (metastasis) (4). Breast cancer is also categorized as a heterogeneous disease and currently is split into four subtypes clinically based on immunohistochemical analysis: luminal A (HR+/HER2-) and luminal B (HR+/HER2+) (expressing the estrogen receptor (ER)), human epidermal growth factor receptor 2 (HER2)-enriched (without ER expression) and basal-like also known as triple-negative breast cancer (TNBC) (17,18). Breast cancer progression is usually classified according to stages, from early-stage breast cancer (stage I) to advanced or metastatic (stage IV). The risk of breast cancer has been consistently associated with age, a family or personal history of breast cancer, reproductive and hormonal factors that include early menarche, late age at first pregnancy, small number of pregnancies, short or no periods of breastfeeding, and later menopause, hormone replacement therapy (HRT), obesity for postmenopausal breast cancer only, alcohol consumption, physical inactivity, exposure to ionizing radiation, and genetic predisposition (17). Genetic predisposition has been reported to play a significant role as a causative factor in breast cancer pathogenesis. High-risk predisposition alleles conferring a 40 -85 % lifetime risk of developing breast cancer include BReast CAncer genes (BRCA1, BRCA2), HER2 neu mutation, and elevated ER expression. Several studies also report that women with BRCA1 and BRCA2 deleterious

mutations have a lifetime breast cancer risk ranging from 65 % to 85% for BRCA1 carriers and 45 - 85% for BRCA2 carriers (19).

In 2020, Global cancer statistics (GLOBOCAN) reported that female breast cancer has surpassed lung cancer as the most commonly diagnosed cancer with an estimated 2.3 million new cases (2). It was reported to be the fifth leading cause of cancer mortality worldwide, with 685,000 deaths (1). Among women, breast cancer accounts for 1 in 4 cancer cases and 1 in 6 cancer deaths. Although incident rates in developed countries are 88% higher than in developing countries, the reported mortality rates in developing countries are higher by 17% (1). A recent study on 3-year survival estimates of breast cancer patients in five countries in sub-Saharan Africa reported that less than 50% of cancer patients survive for at most 3 years after diagnosis (3). The low survival rates and high mortality rates in African have been attributed to poor diagnosis and lack of effective treatment (13). Currently used breast cancer treatment approaches are mostly a combination of; surgery, radiotherapy, chemotherapy, endocrine therapy, and targeted therapy depending on the tumor stages (5). Despite the past developments in breast cancer therapy, surviving breast cancer is still a big challenge in oncology. Triple-negative breast cancer (TNBC) and metastatic (stage IV) cancer have limited options of treatment. most therapeutic approaches for these cancers aim to prolong survival, control symptoms with low treatment-associated toxicity and maintain or improve quality of life (that is, improved quality-adjusted life expectancy (20).

2.1.1 Breast cancer chemotherapy and drug discovery

Chemotherapy is defined as the use of drugs to destroy cancer cells. It usually works by keeping the cancer cells from growing, dividing, and multiplying. Chemotherapy is one of the most popular

and competent treatment options in breast cancer management (6). During chemotherapy, drugs are administered as the only treatment or complementing other methods of treatment such as radiotherapy, surgery, hormonal, and immunotherapy (21). Currently, the Food and Drug Administration (FDA) in the United States of America approve 83 drugs for breast cancer treatment. These drugs can fundamentally be classified into two categories; adjuvant and neoadjuvant drugs and drugs for advanced or metastatic breast cancer. Adjuvant and neoadjuvant drugs are drugs given before and after another treatment respectively. Figure 1 shows some of the drugs used in advanced breast cancer management.

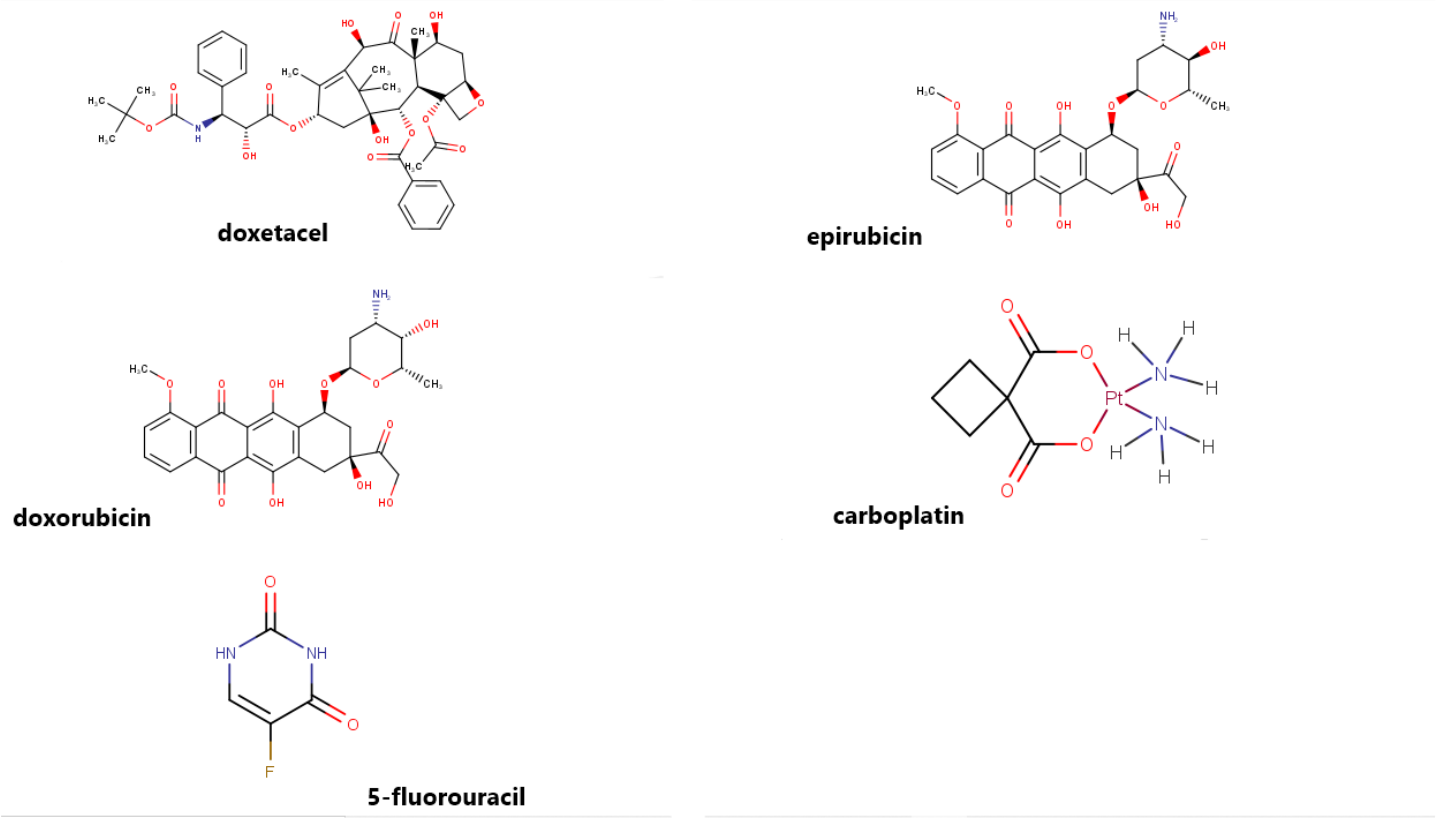


Figure 1: Examples of approved drugs used in breast cancer chemotherapy

Although the application of chemotherapeutic drugs is still a dependable approach in breast cancer management, adverse side effects have contributed to their unpopularity. Some of the most

common possible side effects include hair loss, nail changes, mouth sores, loss of appetite or weight loss, Nausea, vomiting, and diarrhea (5). Such toxicity issues have warranted research and development effort to discover safer drugs. In addition to toxicity, emergence of drug resistant tumors has also reduced the efficacy of most breast cancer drugs. Several breast cancer drug candidates are still under clinical evaluation. In 2020 FDA approved four breast cancer drugs to be used in management of triple negative breast cancer and metastatic breast cancer (22). As the complexity and the heterogeneity of breast cancer cells is unveiled timely leading to the discovery of more cancer molecular pathways and drug targets, it is vital for medicinal chemists to identify novel anti-cancer drugs that exploit new biological pathways.

2.1.2 Role of notch signaling in breast cancer pathogenesis and drug discovery

The role of normal cell signaling pathways in regulating cancer cell proliferation and division, cell death, cell differentiation and fate, and cell motility has been reported in several studies. Commonly known pathways such as; estrogen receptor (ER) signaling, HER2 signaling, and canonical Wnt signaling are currently targeted by most chemotherapeutic drugs (18). Gamma (γ)-secretase -regulated notch signaling is one of the signaling pathways that has drawn a lot of attention in oncology for the past decade. Notch signaling pathway is highly conserved cell signal mechanism that regulates a vital role in proliferation, stem cell maintenance, cell fate specification, differentiation, and homeostasis of multicellular organism and implicates angiogenesis (23). During notch signaling notch receptors (notch1-4) interacts with notch ligands (Jagged1, Jagged2, Delta1, Delta3, and Delta-like 4) expressed on adjacent cells leading to proteolytic cleavages of Notch receptor. The cleavage step catalyzed by the γ-secretase complex results in the release of the Notch intracellular domain (NICD). The NICD then moves to the nucleus, where it interacts with CSL (RBP-Jk/CBF1) and Mastermind to activate transcription of downstream target genes.

The Notch signaling pathway has been reported to be highly deregulated in several human cancers including lung, breast, cervical, colon, pancreatic, and renal carcinoma. In several clinical studies, Notch 1 was associated with an enhanced progression of tumor cells and chemotherapy resistance. In 2016, Ying et al reported that Notch 1 levels were significantly associated with metastasis and triple-negative breast cancer (TNBC) in 115 tumor tissues from primary lessions (24). Another study found that Notch4 was expressed in 55.6% of TNBC samples compared to 25.5% of ER+ samples (25). The inhibition of the notch pathway has been viewed as a promising therapeutic strategy to cure breast cancer and completely eradicate chemotherapy resistance. One popular approach that has been employed do is to inhibit γ-secretase, an enzyme responsible for proteolytic cleavage of the receptors that release the active intracellular fragment which is one of the most crucial steps (7).

γ-secretase is a transmembrane aspartyl protease that comprises four subunits, the catalytic subunit Presenilin 1/2 (PS1/2), nicastrin, Aph-1, and Presenilin enhancer (PEN). The aspartyl protease is known to catalyze over 90 protein substrates including notch receptors using its catalytic subunit Presenilin (26). Over the past decades, γ-secretase inhibitors (GSIs) have been actively investigated for their potential to block the generation of Ah peptide that is associated with Alzheimer's disease. Some of the GSIs are still being tried in clinical studies for various carcinomas since inhibition of γ-secretase has been shown to reduce tumor size and cancer cell proliferation (27). GSIs such as MK-0752, RO4929097, DAPT, PF-03084014, BMS-986115, and LY90009 (Figure 2) have been used in phase 1 and phase 2 clinical trials in combination with other cancer drugs to treat breast, ovarian, colorectal, melanoma, pancreatic, and non-small cell lung cancer (27,28).

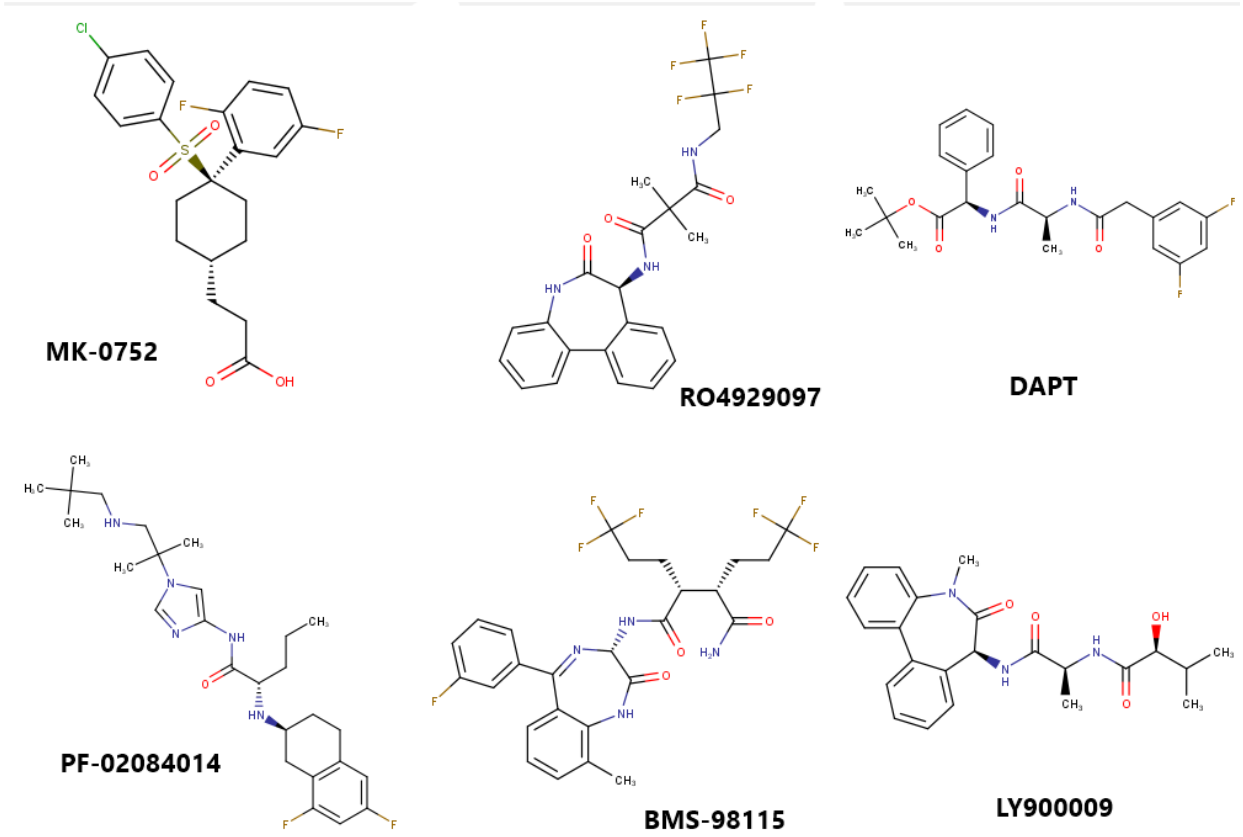


Figure 2: Some of the GSIs that have been clinically investigated as anti-cancer drugs

Although more than 100 GSI have been synthesized to date, none have been approved for clinical use because of reported toxicity and efficacy issues (29). Therefore, the discovery of structurally diverse GSIs from other synthetic compounds libraries and natural products offers a promising alternative in cancer drug discovery. Furthermore, the introduction of crystalized gamma γ-secretase complexes in protein database such as the RCSB PDB could enhance the discovery of enzyme specific γ-secretase inhibitors.

2.1.3 Natural products-based drug discovery of breast cancer drugs

Natural products (NPs) are broadly defined as compounds or chemicals derived from natural sources that possess biological activities (30). These compounds are mostly metabolites and/or by-products from biological sources, e.g. plants, microorganisms, or animals. Natural products (NPs) have been the center of attention of the research community in the last decades and the interest around them continues to grow incessantly, especially in drug discovery (9). This is because NPs have proved to be a reliable source of drugs or drug-like scaffolds for several diseases, that contributed about 3.8% unaltered NPs, and 18.9% NP derivatives of the total approved drugs between 1981 and 2019, (31). These include the drugs (Figure 3) Taxol from *Taxus brevifolia*, and Vinblastine from *Catharanthus roseus*, effective in treating cancer (32). NPs have enormous scaffold diversity and structural complexity. They also have a higher molecular mass, a larger number of sp^3 hybridized carbon atoms, and more oxygen atoms but fewer nitrogen and halogen atoms, compared with synthetic compounds (33,34).

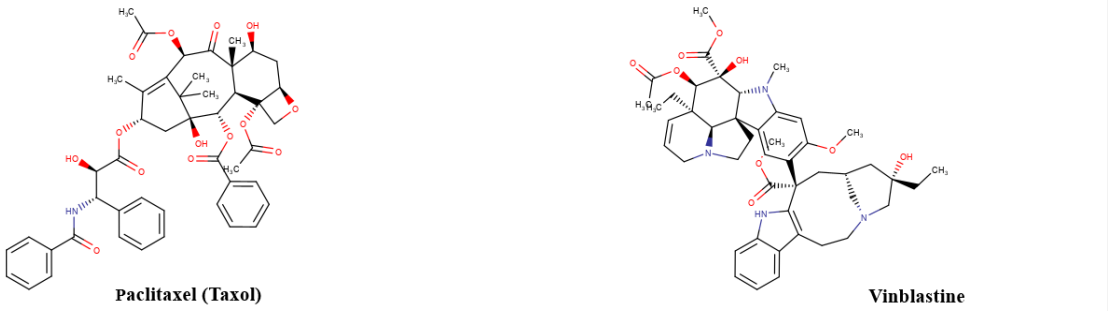


Figure 3: Some of the natural products approved for clinical use; Paclitaxel (Taxol) and Vinblastine as anti-cancer drugs

For the past two decades, there has been a rapid multiplication of NP information resulting from dozens of studies in natural products research. This information encompasses biological and geographical origins, ethnobotanical, chemical and pharmacological information of natural

products. These developments have been followed by a compilation of physical and virtual libraries of NPs. A most recent study in 2020 reported that there are 123 databases of natural products (NPs) that have been published and cited in the scientific literature from 2000 to 2019 (9). The study was followed by assembling of the most complete up-to-date COlleCtion of Open Natural ProdUcTs (COCONUT) which is a compendium of 50 open-access databases collecting over 400,000 compounds (35). The motive behind this development was to have an online resource regrouping all known NPs in just one place, to simplify NPs research, advance virtual screening projects, and other *in silico* applications that improve lead discovery in drug design pipelines (30). In Africa, remarkable efforts have been done to characterize and document many medicinal plants alongside the constituted phytochemicals for the past ten years. An attempt to improve African NPs accessibility to the scientific community has led to the development of small online natural product databases such as SANCDB, ETM-DB, EANPDB, and NANPDB. Several datasets (Table 1) are also available in literature and provide a good starting point for virtual screening campaigns in cancer drug discovery and development. A majority of the compounds in these datasets have never been explored further to identify bioactive drug-like compounds for breast cancer chemotherapy using *in silico* approaches.

Table 1: List and description of currently available databases of natural product compounds of African origin in 2021

NP Database/Library	Data source and description	No. of NP compounds present	Web link	References
NANPDB	Mostly plants and some endophytes, animals (e.g., corals), fungi, and Bacteria from Northern African countries.	>6000 compounds	http://african-compounds.org/nanpdb	(36)
EANPDB	NPs from about 300 source species from the Eastern African region and literature data of the period from 1962 to 2019	1870	http://african-compounds.org/nanpdb	(37)
SANCDB	Mostly from plants and some from marine life from South Africa	>1000 compounds	https://sancdb.rubi.ru.ac.za/	(38,39)
p-ANAPL	From medicinal plants across Africa	≈500 compounds	Not available	(40)
AfroDB	From medicinal plants across Africa	>1000 compounds	Not available	(41)
AfroCancer	From medicinal plants that have reported anti-cancer activity across Africa	≈400 compounds	Not available	(42)
ConMedNP	From 376 distinct medicinal plant species from central African Flora (Congo basin)	≈3200 compounds	Not available	(33)
CamMedNP	From 224 distinct medicinal plants from Cameroon flora	≈2500 compounds	Not available	(43)
AfroMalariaDB	From pure compounds derived from African flora which have exhibited anti-malarial properties	≈300 compounds	Not available	(44)
Afrotryp	Potential trypanocidal compounds derived from medicinal plants in Africa	≈ 300 compounds	Not available	(45)
ETM-DB	From 1054 Ethiopian medicinal herbs with reported 1465 traditional therapeutic uses	≈ 4000 compounds	http://biosoft.kaist.ac.kr/etm/home.php	(46)

2.2 Computer-aided drug designing (CADD)

Computer-aided drug discovery (CADD) or *in silico* drug discovery refers to the use of computer technology and software to support drug discovery efforts. It also considers the application of computational methods to support, guide, and streamline the processes of drug discovery, design, development, and optimization. CADD methods have emerged as a powerful tool in the development and discovery of drug-like molecules for over three decades, allowing higher hit rates than experimental high-throughput screening (HTS) approaches in a very short space of time and at low costs(47). Most if not all CADD approaches can be classified under Ligand-based drug design (LBDD) and structure-based drug design (SBDD).

SBDD approaches depend on information from the 3D structure of the drug target to search for novel compounds active against the particular target. The 3D structural information is usually derived from the experimental crystal structure of the drug target (48). The structure is used to guide the discovery of new ligands either by informing the design of novel compounds or through the identification of new ligands from the screening of virtual compound libraries. Experimental 3D structures of proteins are stored in virtual protein data banks and can be downloaded free when needed. If the experimental 3D structure of a particular target is absent, a theoretical structure can be designed or modeled using a technique termed homology modeling (15). Docking is one of the most popular SBDD techniques used in a virtual simulation of molecular interactions, it can predict the conformation and binding affinity of ligands within a target active site with high accuracy (15,49). Other methods such as molecular dynamics and structure-based virtual screening have assisted in provided relevant insights into ligand-receptor interactions. LBDD methods depend on the information of known molecules which bind to the biological target active site of interest (50).

These molecules are used to design a suitable model, which is useful in identifying other compounds that are potential binders of the same biological target. Approaches available for Ligand-based drug designing include quantitative structure-activity relationship (QSAR) pharmacophore modeling, reverse docking, and Target Fishing (51). Recently, there has been a trend towards integrating both SBDD and LBDD approaches in identification of novel compounds. The purpose is to enhance the reliability of CADD approaches by combining structural and physicochemical properties of known ligands and protein information. Structure-based pharmacophore modelling is one example of integrated techniques CADD.

2.2.1 Chemoinformatics

Chemoinformatics involves the organization, integration, curation, standardization, simulation, mining, and transformation of pharmacology data (compounds and bioactivity) into knowledge that can drive rational and viable drug development decisions (52). Chemoinformatic approaches have gained considerable popularity in CADD. The field has helped handle chemical data for *in silico* studies. Chemoinformatic approaches in CADD are popularly known for curating, characterization, or profiling large libraries and databases of compounds for virtual screening campaigns n (52). In these endeavors, pharmacologically relevant chemical descriptors used in estimating the drug-like and lead like character of compound libraries are computed. Some of the descriptors used include; molecular weight (MW), the octanol/water partition coefficient (SlogP), topological polar surface area (TPSA), hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), and the number of rotatable bonds (RB). These descriptors are also used to assess compliance of molecules to Lipinski (53) and Veber (54) rules of five which is used predict the oral bioavailability of drug candidates. A recent study on 400 potential anticancer, cytotoxic, and

ant proliferative agents derived from the African flora showed that 165 out of the 390 compounds (42.3%) could be regarded as "lead-like", having physicochemical parameters falling within the intervals ($150 \leq MW \leq 350$; $\log P \leq 4$; $HBD \leq 3$; $HBA \leq 6$) (42). Physicochemical properties of NPs have formally been used to study the differences between NPs, synthetic chemicals and approved drugs, through analysis of chemoinformatic parameters such as chemical space, scaffold diversity and molecular complexity (31).

2.2.2 Application of Chemoinformatics in pharmacokinetic and toxicity evaluations of drug hits

Attrition is a major concern in anticancer drug development with up to 95% of drugs examined in phase I trials not getting FDA approval and marketing authorization (55). Most preclinical and clinical candidates are rejected due to toxicity or lag of optimal pharmacokinetics properties, resulting in high costs and increased timelines for the drug discovery process(56). Pharmacokinetic evaluations in drug development involve evaluations of absorption, distribution, metabolism, excretion and toxicity (ADMET) properties of drug candidates. Computational or *in silico* assessment of ADMET properties of drug leads prior to expensive clinical trials has been considered as a cost and time-saving approach (57). *In silico* methods employed as applied in this area, makes use of computational chemoinformatic predictive models and algorithms based on the QSAR principle in medicinal chemistry (12,58). Common ADMET parameters assessed by *in silico* models include water solubility, Caco-2 intestinal permeability, human intestinal absorption, human skin permeability, activity as P-glycoprotein substrate-glycoprotein inhibition; volume of steady-state distribution blood-brain barrier permeability, CNS permeability, metabolic interactions with cytochromes CYP2D6, CYP3A4, CYP1A2, CYP2C19, CYP2C9, CYP2D6, and

CYP3A4, total renal (log mL/min/kg); OCT2 substrate; AMES toxicity; human ERG inhibition and hepatotoxicity (56). In 2020 BIOFACQUIM, a Mexican NP database was analyzed for ADMET by Durán-Iturbide et al. and they reported that the absorption and distribution profiles of NPs in BIOFACQUIM are similar to those of FDA approved drugs, while the metabolism profile is comparable to that in the other natural product database (59). While for the past decade Fidele Ntie-Kang et al have comparatively screened ADMET properties in several NP databases of African origin with special emphasis on toxicity screening (33,37,41,43,60).

.

2.2.3 Tools and software used in computer aided drug designing and Chemoinformatics

CADD and Chemoinformatics has developed rapidly from the time it was first introduced in the 1980s. Currently, several tool/toolkits or software are available, that can utilize structural data of NP compounds to advance drug discovery endeavors (61). Although the most efficient are commercialized, many chemoinformatic tool developers continue releasing tools that are open source to benefit research in academia. Some of the popularly used CADD tools are listed in Table 2.

Table 2: List of CADD and chemoinformatic tools popularly used in drug discovery

Name	Description	Availability (accessed as of January 2022)	References
Chemical descriptor calculation			
RDKit	A standalone toolkit that can carry out 2D and 3D molecular operations such as descriptor and Fingerprint generation for machine learning	https://www.rdkit.org/ (open source)	(62)
CDK	A standalone toolkit that can carry out Substructure and SMARTS pattern searching and figure print-based similarity searching	https://cdk.github.io/ (open source)	(63)
QuaSAR-Descriptor (MOE)	A module in MOE that can calculate over 100 2D and 3D molecular descriptors for QSAR purposes	https://www.chemcomp.com/ (commercial)	(64)
Scaffold diversity analysis and chemical space visualization			
Scaffold Hunter	Java-based tool for the visual analysis of chemical data sets. Can also generate, scaffold trees and clusters for compound databases	http://scaffoldhunter.sourceforge.net/index.html (open source)	(65)
Data warrior	Standalone software that can generate chemical Scatter plots, box plots, bar charts, and pie charts. Can also compute multiple scaffolds or compound substitution patterns	https://openmolecules.org/datawarrior/ (open source)	(66)
ChemGPS-NP	Web-based tool for computing the eight principal components (dimensions)	https://chemgps.bmc.uu.se/batchelor/ (open source)	(67)

	describing physical-chemical properties for chemical libraries even NPs		
WebMolCS	A Web-Based Interface for Visualizing Molecules in Three-Dimensional Chemical Spaces	http://www.gdbtools.unibe.ch:8080/webMolCS/	(68)
CDPs-V.2	A web-based tool used to compare and classify data sets using diversity metrics (i.e. scaffold counts, fingerprints similarity, molecular properties)	https://consensusdiversityplots-difacquim-unam.shinyapps.io/RscriptsCDPlots/ (open source)	(69)
ADMET profiling			
PkCSM	Web-based ADMET for predicting small-molecule pharmacokinetic properties using graph-based signatures	http://biosig.unimelb.edu.au/pkcsm/ (open source)	(70)
SwissADME	Web-based tool for computing physicochemical descriptors as well as predicting ADME parameters and pharmacokinetic properties	http://www.swissadme.ch/ (open source)	(71)
QikProp/ Schrödinger	Standalone software that can rapidly predict pharmacokinetic properties of compound libraries at once.	https://www.schrodinger.com/products/qikprop (Commercial)	(72)
ADMETlab	A web interface for systematic ADMET evaluation of chemical compounds based on a comprehensive database comprising 288,967 entries	http://admet.scbdd.com/ (open source)	(73)
Docking			
Autock4	Locally installed docking program; grid-based; grids have to be calculated before docking	http://autodock.scripps.edu (open source)	(74)

Autodock vina	Standalone docking program; no grids have to be calculated before docking, usually used in virtual screening projects	http://autodock.scripps.edu (open source)	(75)
Glide/ Schrödinger	Standalone docking program that performs high throughput virtual screening jobs with high accuracy.	https://www.schrodinger.com/products/glide (commercial)	(76)
GOLD	Standalone docking program; based on a genetic Algorithm	https://www.ccdc.cam.ac.uk/solutions/csd-discovery/Components/Gold/ (Commercial)	(77)
Pharmacophore modeling			
MOE	Structure- and ligand-based modeling and virtual screening tool implemented in a large modeling suite	https://www.chemcomp.com/ (commercial)	(78)
Phase/ Schrödinger	Ligand-based modeling and virtual screening tool implemented in Schrodinger suite	https://www.schrodinger.com/products/phase (commercial)	(79)
Pharmer	A tool for Structure- and ligand-based modeling and virtual screening tool	https://sourceforge.net/projects/pharmer/ (open source)	(80)
ZincPharma.	Software that uses the Pharmer open-source pharmacophore search technology	http://zincpharmer.csb.pitt.edu/ (open source)	(81)
Ligand Scout	Software that can create three-dimensional (3D) pharmacophore models from structural data of macromolecule–ligand complexes	http://www.inteligand.com/ligandscout/ (commercial)	(82)

The ever-growing studies in Computer sciences and data science now enable software developers to design software tools that can incorporate many chemoinformatic tasks on a single interface. One such software is Konstanz Information Miner (KNIME), an open-source analytics platform,

which is the leading tool for wide-ranging data processing, integration, analysis, and exploration. The software allows chemical data scientist to execute several important tasks for chemoinformatics analysis by accessing open-source chemoinformatics toolkit such as CDK, Indigo, RDKit, Vernalis, CACTVS, Enalos, Lhasa, OpenBabel, OCHEM, Chemical Identifier Resolver, ErlWood, EMBL-EBI Nodes, and CheSMapp in form of nodes. The platform user-friendly workflows to include numerous tasks. In a study to investigate the chemical space and diversity of NuBBE Database Saldívar-González et al used KNIME to facilitate the visual representation of the seven molecular descriptors and generate a visual representation of the chemical space, a principal components analysis (PCA) (83). The tool continues to gain popularity in designing machine learning protocols of computer-aided drug design projects.

In this study, chemoinformatic methods and tools were used to characterize the chemical diversity of drug-like natural products from four African natural products databases relative to FDA bioavailable drugs. Structure-based virtual screening approaches were used to identify novel γ-secretase inhibitors from natural sources for breast cancer chemotherapy.

2.3 Justification of study

Bioactive natural products have proved to be selective towards cancer proliferating cells and are nontoxic towards normal body cells as such they would be more preferable in chemotherapy to synthetic drugs (84). Africa natural product databases contain a large number of diverse compounds that provide a promising starting point for breast cancer drug discovery (60). Chemoinformatic characterization of drug-like natural products in databases can assist in the prioritization of bioactive drug-like and lead-like compounds for screening and synthetic purposes

(52). Since the traditional drug development pipelines are time-consuming and take years for target identification, validation, and subsequent design optimization of the lead candidate compounds, CADD and chemoinformatic approaches can be used to rapidly identify drug hits by virtually screening them against validated drug targets (50). In addition, *in silico* approaches are capable of profiling the pharmacokinetic properties of natural product hits prior to clinical ADMET analysis of bioactive compounds, thus saving time and cost (12). Identification of natural inhibitors γ-secretase is a positive step towards regulating the notch signaling pathway, combatting multi-drug resistance in breast cancer chemotherapy and thus reducing the health burden of breast cancer in Africa and globally (27).

2.4 Objectives of the study

2.4.1 Main objective

To identify and characterize drug-like natural product inhibitors for gamma-secretase complex from selected African natural products databases using structure-based virtual screening and chemoinformatic approaches.

2.4.2 Specific objectives

1. To characterize the physicochemical distribution and chemical diversity of druglike compounds from selected African natural products databases.
2. To identify novel anti-cancer compounds from drug-like natural products dataset using structure-based virtual screening.
3. To profile the pharmacokinetic (absorption, distribution, metabolism, excretion, and toxicity) properties of novel γ-secretase inhibitors.

3. METHODOLOGY

3.1 Study design

This was an *in silico* study conducted at Kamuzu University of Health sciences. The study, applied concepts of computational chemistry, pharmacology, and bioinformatics. Figure 4 shows a schematic diagram of the study design.

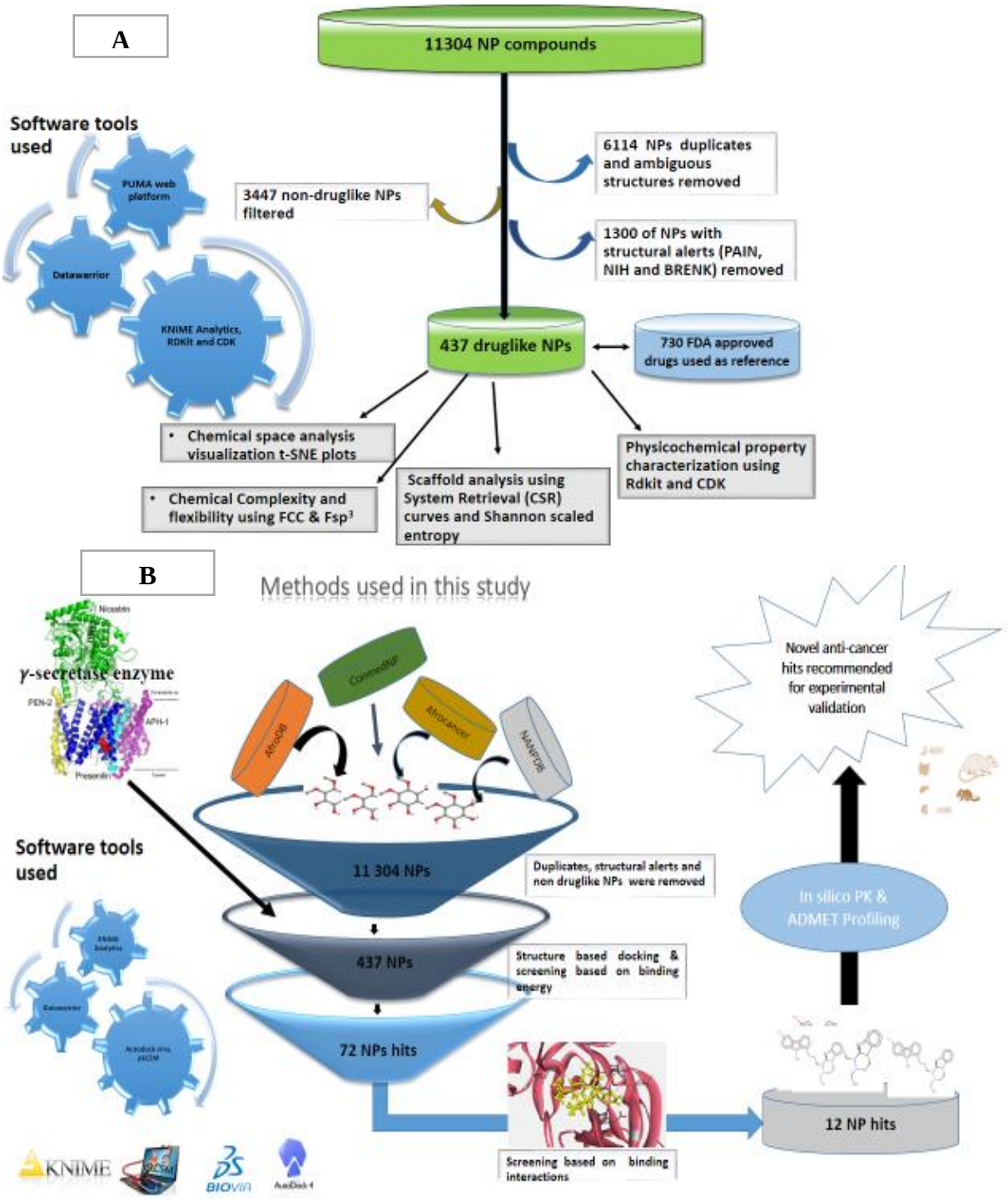


Figure 4: (**A**) Summary and workflow of methods using chemoinformatic analysis of compounds in African natural products databases and (**B**) used in structure-based virtual screening of drug-like natural products and ADMET analysis of the final compounds.

3.2 Cleaning and curation of the NP databases

A total of 11 304 compounds were retrieved from 4 virtual African NPs databases listed in **Table 2.1** were used; namely NANPDB (36), ConMedNP (33), Afrocancer (60) and AfroDB (41). The NP libraries were loaded in KNIME analytics software (85) and structurally inspected for compounds duplicates, incomplete structures, and they were also normalized and geometrically optimized using the MMFF94 force field.

All compounds that violate Veber (54) and Lipinski (86) drug-likeness rules and compounds that contain structural alerts sub-structures were also removed using filtering nodes in KNIME analytics. The nodes contained PAINS, BRENK, NIH, and ZINC filters that were designed to screen chemical substructures that are possibly toxic and are capable of giving false positives in assays(87,88). The final NP dataset was be referred to as the "DruglikeNP" set in this research. Another dataset with FDA approved drugs from drugbank (https://go.drugbank.com/) was curated to create a drug-like FDA drugs dataset, which was used in scaffold diversity analysis of NPs. The drug-like FDA dataset was named "druglikeFDA"

3.3 Chemoinformatic profiling of the DruglikeNP dataset

3.3.1 Physicochemical descriptor chemical analysis

Pharmaceutically relevant molecular descriptors of the dataset such as hydrogen-bond donors (HBD), hydrogen-bond acceptors (HBA), partition coefficient octanol/water (XlogP), calculated Aqueous solubility(LogS), molecular weight (MW), number of rotatable bonds (RB), topological polar surface area (TPSA), were computed and statistically analyzed using CDK RDKit and Java nodes in KNIME (31,83).

3.3.2 Chemical space visualization

To explore property similarity between natural products in the druglikeNP dataset and 750 drug-like drugs in FDA approved drugs dataset (89), the t-SNE method was used to generate a 3D plot. t-SNE is a nonlinear dimension reduction in which Gaussian probability distributions over high-dimensional space are constructed and used to optimize a Student t-distribution in low-dimensional space. The low-dimensional space maintains the pairwise similarity to the high-dimensional space, leading to a clustering on the embedding space without any significant loss of structural information (90,91). Eight descriptors computed in section 3.1.1 were used to generate a t-SNE plot.

3.3.3 Scaffold complexity and flexibility

To quantify molecular complexity two descriptors were computed in the study of fraction of sp3-hybridized atoms (Fsp3), calculated as the number of sp3-hybridized atoms over the total atom count, and the fraction of chiral carbons (FCC) calculated as the number of chiral centers over the carbon count (92,93). Molecular flexibility was also measured using the molecular flexibility index in Datawarrior (66).

3.3.4 Scaffold diversity analysis

Scaffolds were generated through the rcdk function embedded in the Platform of Unified Molecular Analysis version 1 (PUMA)(94), which removes all side chains and converts the compound into a ring system joined by linkers. The distribution and diversity of the molecular scaffolds in the two datasets (druglikeNP and druglikeFDA) were calculated and analyzed using Cyclic System Retrieval (CSR) curves in PUMA (69). These curves were plotted with the fraction

of scaffolds in the x-axis and the fraction of compounds in the y-axis. While the Shannon entropy and scaled Shannon entropy was used to identify unique scaffolds as it is known to give a more accurate measure of the most common chemotypes populated in a dataset. To compare the structural diversity of the entire molecules based on fingerprints between druglikeNP and druglikeFDA compounds, extended connectivity molecular fingerprints with a diameter of 4 (ECFP_4) were used. Other fingerprints were also computed, including MACCS (166-bits) and PubChem (881-bits). The distribution of pairwise similarities of the compounds was analyzed by means of cumulative distribution function (CDF) curves using the Tanimoto index in PUMA (95).

3.3 Structure-based virtual screening of γ-secretase inhibitors (molecular docking)

3.3.1 Target identification and Protein preparation

Seven 3D structures of γ-secretase were downloaded from the PBD database (http://www.rcsb.org/pdb/home/home.do). The structures were inspected for quality through literature and visual methods. Discovery Studios Visualizer v 17.2.0.16349. distributed by Dassault systems Biovia Crop and Pymol distributed by Schrodinger, Inc were used to visualize the structure and binding pockets of γ-secretase enzyme units. From the inspection, one structure (PDB ID: 6IYC) with atomic resolution, 2.6 Å was selected for structure-based docking studies.

The Cryo structure of γ-secretase enzyme (PDB ID; 6IYC) was prepared using Discovery Studio Visualizer v17.2.0.16349 Presenilin (PS1) unit was isolated from the γ-secretase complex, the unit was protonated and all heteroatoms not related to the activity of the unit were deleted. PS1 was saved in PDB format and was further converted to PDBQT format using Autodock Tools.

3.3.2 Ligands preparation

SDF files containing 437 drug-like NPs (23 NPs from Afrocancer, 56 NPs from Afrodb, 136 NPs from ComMedDB, and 222 NPs from NANPDB) were prepared for docking purposes. Open Babel version 2.4.1 ref to protonate pH 7.4 (physiological pH for the biological activity of gamma secretes enzyme) and convert the ligands to PDB format. The software was also used to give each NP (ligand) an identifier name related to its source NP library, for example, 9-O-beta-glucopyranosyl trans-cinnamyl alcohol from the NANPDB dataset was named nanpdb1.

3.3.3 Setting up the Docking protocol

PyMol Molecular Graphics System, distributed by Schrodinger LLC, Autodock plugin was used to set up the virtual screening workflow. Pymol was able to automate the docking process by accessing necessary Autodock suite tools; Autodock Vina, Autodock Tools, and Auto ligands. Grid coordinates for the binding site were configured to; 56, 52, and 58 for X, Y, and Z respectively. Grid points of the center were set to 158.062, 182.457, and 157.725 for X, Y, and Z planes respectively. Autodock Vina was configured to produce nine poses for each ligand. 13 Known γ-secretase inhibitors (GSIs) with documented activity (IC_{50}) were used in validating the docking protocol. A positive correlation between IC_{50}s and Vina binding energies of 70% of GSIs was taken to validate the docking protocol since the Cryo structure of PS1 did not have a co-crystalized ligand.

3.3.1 Post docking analysis based on binding energy

OSIRIS Data warrior v5.2.1 distributed by Idorsia Pharmaceuticals was used to analyze the binding scores and physicochemical properties of drug-like NPs and GSIs.

3.3.2 Post docking analysis based on molecular interactions

NPs with lower binding energy than GSIs were further visually inspected their binding modes to PS1 binding site using Discovery studio visualizer. NPs that interacted with more than four residues common to GSIs binding modes were taken as potential candidate for biological evaluations.

3.4 ADMET and analysis of promising NP compounds

The structures of the most promising compounds were converted to SMILE structures using DS visualizer and used for *in silico* ADMET analysis. The absorption, distribution, metabolism, elimination, and toxicity (ADMET) parameters were predicted using pkCSM (70). The pkCSM software provided information on the following parameters: water solubility (log mol/L); Caco-2 intestinal permeability (cm/s); human intestinal absorption (%); human skin permeability (log Kp); activity as P-glycoprotein substrate; activity as P-glycoprotein I substrate and II inhibitor; volume of steady-state distribution (log L/kg); blood-brain barrier permeability (log BB); CNS permeability (log PS); metabolic interactions with cytochromes CYP2D6, CYP3A4, CYP1A2, CYP2C19, CYP2C9, CYP2D6, and CYP3A4; total renal (log mL/min/kg); OCT2 substrate; AMES toxicity; human ERG I and II inhibition and hepatotoxicity.

3.5 Ethical approval and consent

Ethical approval for the study was sort from - College of Medicine Research and Ethics Committee (COMREC; P.05/20/3064, Appendix 1)

4. RESULTS

4.1 Cleaning the databases and creation of the druglikeNP dataset

In this study, a KNIME workflow (Figure 5) and four structural filters packaged in the form of RDKit and CDK nodes were used to remove duplicate structures, salts and sugars, and disconnected, ambiguous structures from the generated dataset of 11304 compounds. The dataset didn't contain salts but contained 6 sugars, 2031 duplicate structures, and 3837 disconnected and ambiguous structures.

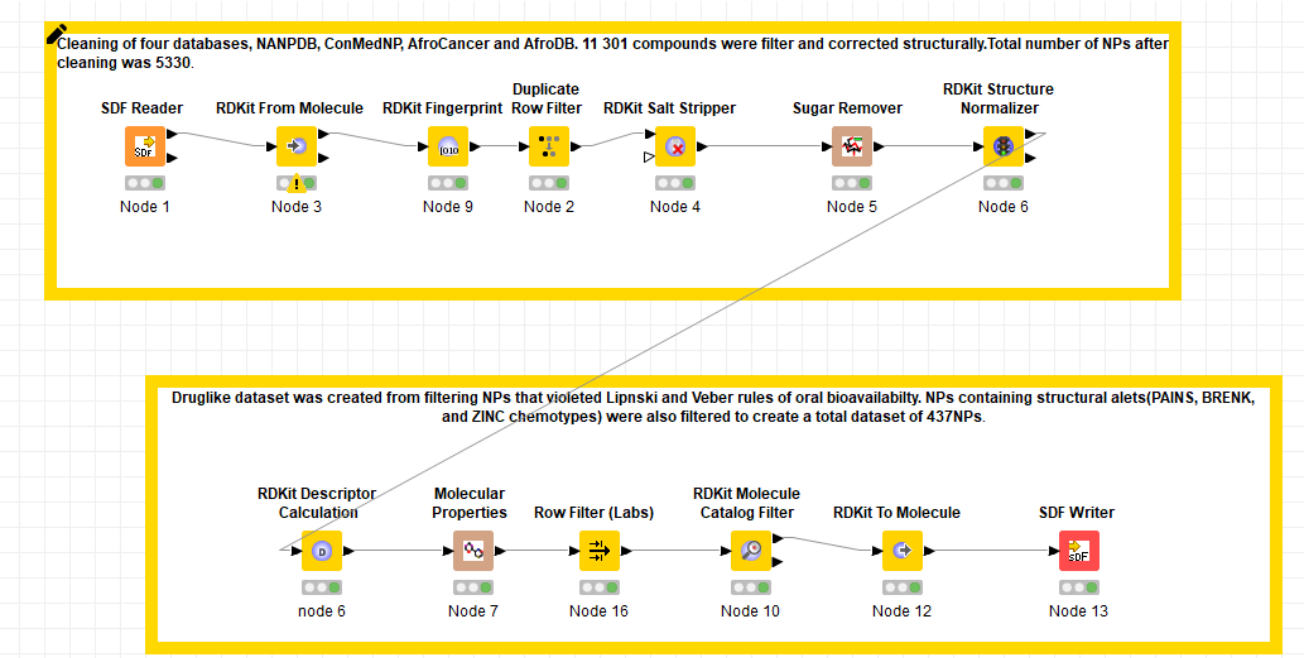


Figure 5: KNIME workflow used to curate and filter natural products database to create 437 drug-like compounds

The total number of NPs that had drug-like properties was 437 after applying the filters listed in **Table 3**. The final dataset was named druglikeNP.

Table 3: Types of structural features used to filter undesirable natural products compounds.

Types of filter used	Number of compounds filtered out of the 11304 NPs
Disconnected and ambiguous structure filter	3637
Duplicated structures	2477
Sugars remover	6
Lipinski and Veber drug-like filter	3447
PAINS, Brenk, ZINC, and NIH filters	1300

4.2 Chemoinformatic characterization of the DruglikeNP dataset

Pharmacologically and Chemoinformatics relevant molecular properties of the druglikeNP dataset were computed and statistically analyzed to understand the property landscape of the natural product dataset.

3.3.4 Physicochemical descriptor chemical analysis

Physicochemical properties of compounds such as hydrogen bond acceptors (HBAs), hydrogen bond donors (HBDs), number of rotatable bonds (NRBs), the octanol/water partition coefficient (XlogP), Solubility (LogS), topological polar surface area (TPSA), molecular weight (MW) and molar refractivity (SMR) are pharmacologically relevant in drug discovery. These descriptors are used to predict bioactivity in quantitative structural activity relationship studies as well as oral bioavailability (96). In this study, all these properties were computed using RDKit nodes in KNIME. Statistical analysis of the descriptors was done using the Data explorer node in KNIME (Table 4).

Table 4: Statistical values of eight physicochemical descriptors used to describe the druglikeNP dataset

Molecular descriptor	Min	Mean	Median	Max	SD
HBA count	1	2.88	3.0	8	1.67
HBD count	1	2.18	2.0	5	1.17
TPSA	12.0	75.06	73.8	149.8	28.8
MW	150.1	291.1	295.1	512.1	76.6
LogS	-6.43	-2.76	-2.91	1.25	1.35
Molar refractivity(MR)	3.01	7.62	7.63	13.8	2.04
XlogP	1.93	1.85	1.87	4.89	1.36
RB count	0	2.55	2	8	1.91

Figure 6: gives distribution of eight physicochemical properties of drug-like natural products

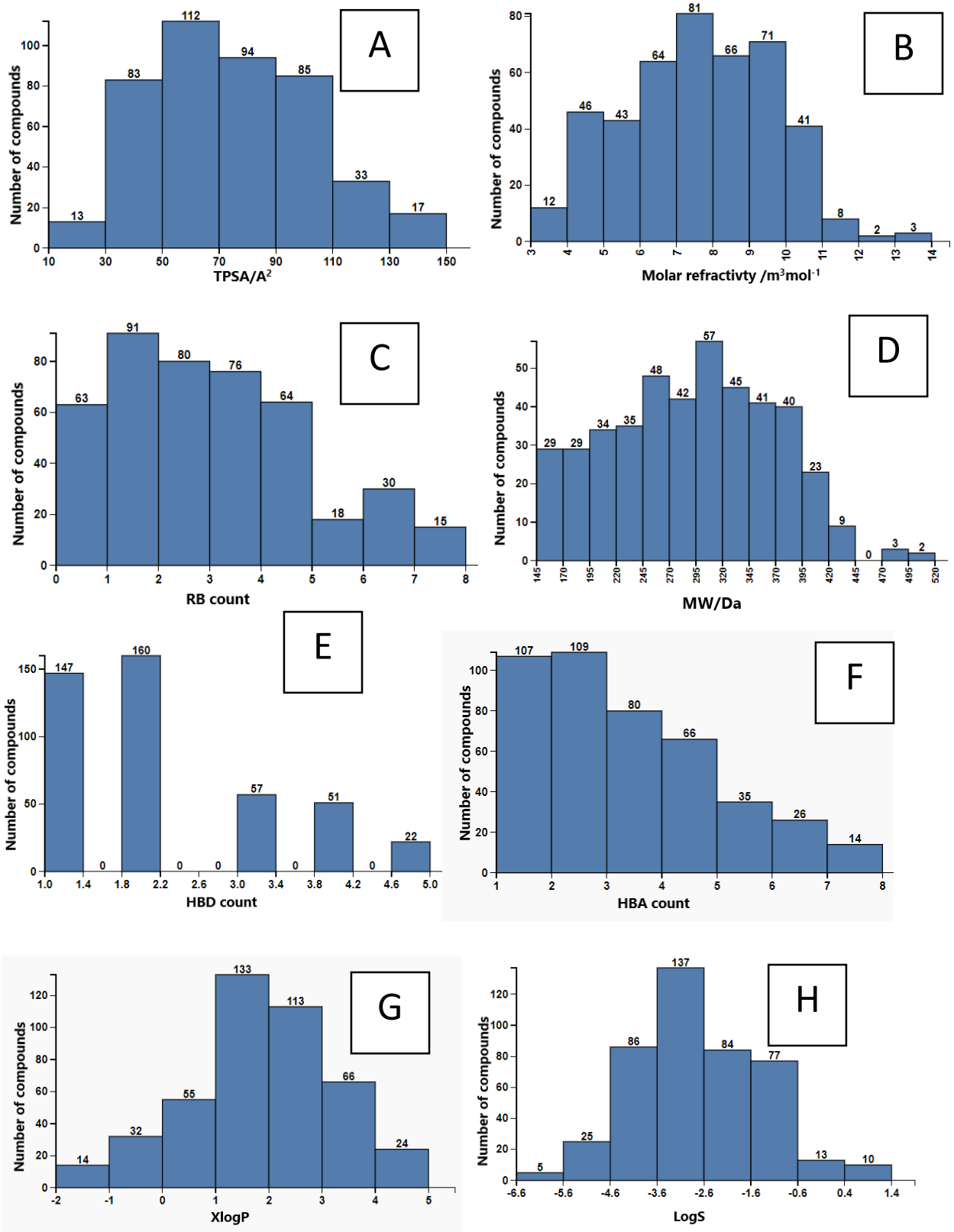

A
Number of compounds
TPSA/A²
13
83
112
94
85
33
17
B
Molar refractivty /m³mol⁻¹
12
46
43
64
81
66
71
41
8
2
3
C
RB count
63
91
80
76
64
18
30
15
D
MW/Da
29
29
34
35
48
42
57
45
41
40
23
9
0
3
2
E
HBD count
147
0
160
0
0
57
0
51
0
22
F
HBA count
107
109
80
66
35
26
14
G
XlogP
14
32
55
133
113
66
24
H
LogS
5
25
86
137
84
77
13
10

Figure 7: Distribution of eight physicochemical properties of drug-like natural dataset: A) Topological polar surface area (TPSA), B) Molar refractivity(MR), C) Rotatable bond count (RB), D) Molecular weight(MW), E) Hydrogen bond donor count, F) hydrogen bond acceptor count, G) log Octanol/water partition coefficient (XlogP) and H) Water solubility LogS.

4.2.1 Chemical space visualization

In this study, the approach was used to check the similarity of natural products from the druglikeNP dataset with the FDA approved drugs that adhere to Lipinski and Veber rules of drug-likeness. T-SNE coordinates were calculated from eight pharmacological chemical descriptors in Table 4. The chemical space map of the two datasets is shown in Figure 8.

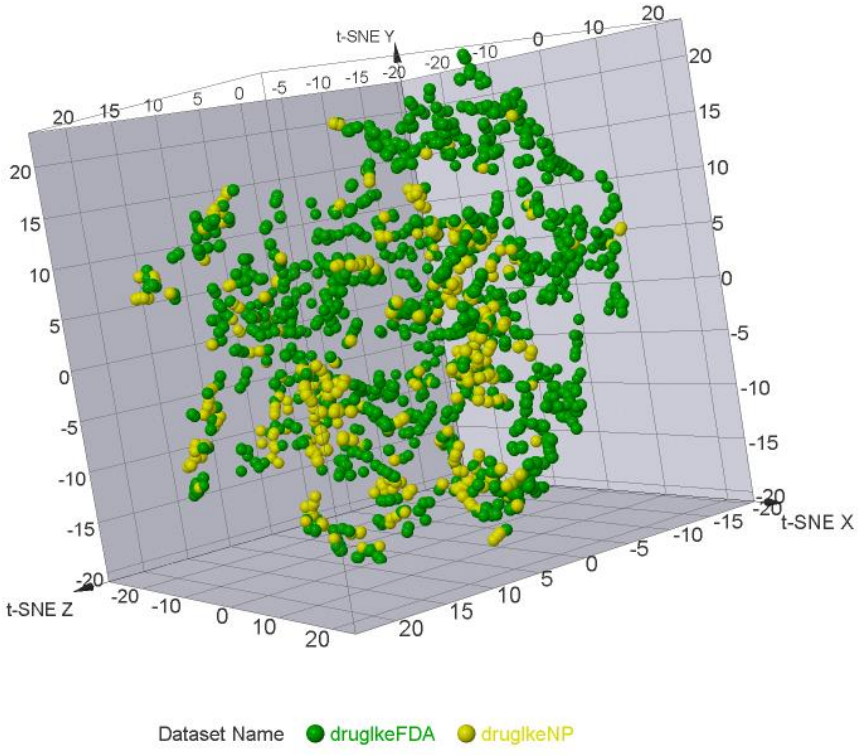


Figure 8: Chemical space of druglikeNP (**yellow**) and druglikeFDA (**green**)

4.2.2 Molecular complexity and flexibility of the druglikeNP dataset

Molecular complexity and flexibility are two important parameters in drug discovery and have been associated with the success of drug candidates in progressing into clinical development, target selectivity, and compound safety (92,97). In this study, two parameters were used to estimate complexity; fraction of sp^3 carbons in RDKit and fraction of chiral centers (FCC). Statistical distribution of molecular complexity in druglikeNP dataset and druglikeFDA is shown in Table 5 and Figure 9.

Table 5: Statistical values of Molecular complexity measurements (Fraction of sp^3 (Fsp3) carbons and fraction of chiral) and molecular flexibility in data of druglikeNP dataset

Dataset	Complexity parameter	Min	mean	Median	Max	SD
druglikeNP	Fsp^3 carbons	0	0.43	0.39	1	0.25
	FCC	0	0. 16	0.16	0.86	0.16
	Molecular flexibility	0	0.29	0.29	0.85	0.14
druglikeFDA	Fsp^3 carbons	0	0.41	0.39	1	0.23
	FCC	0	0.093	0.056	0.83	0.12
	Molecular flexibility	0	0.39	0.42	0.84	0.15

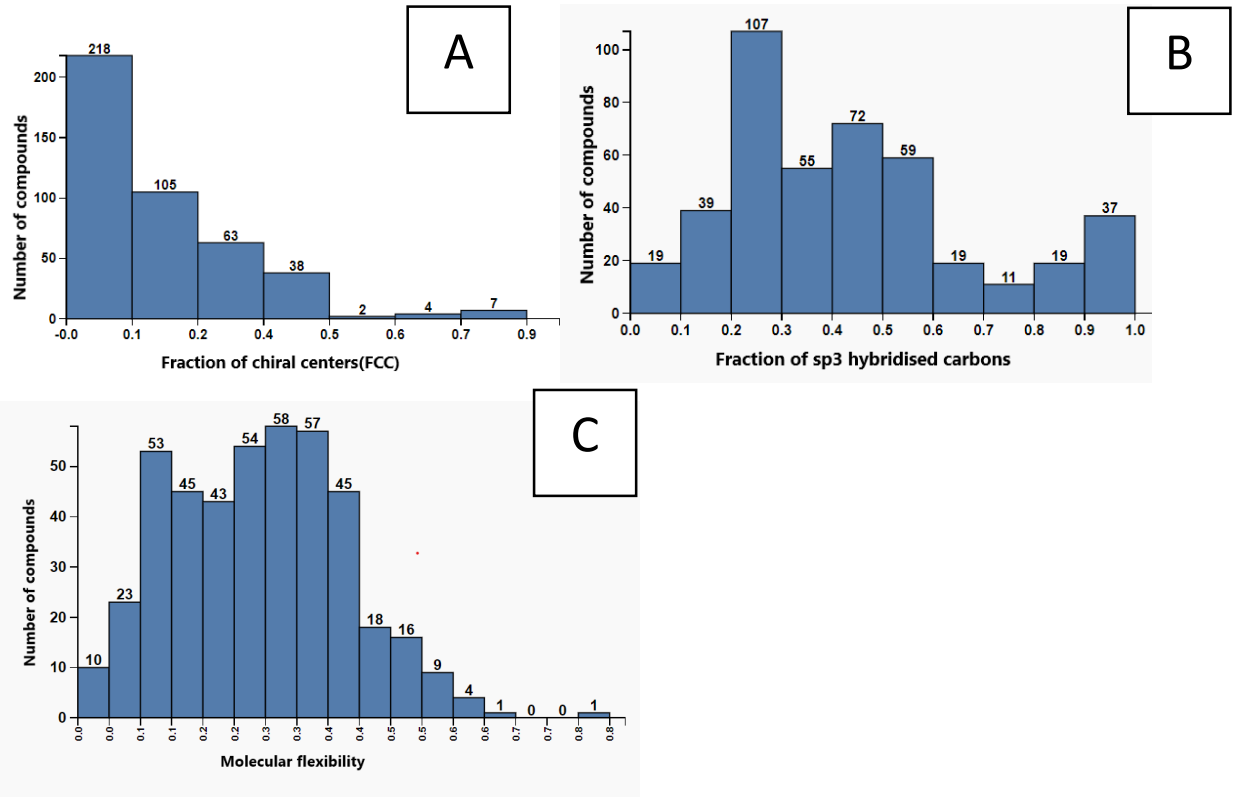


Figure 9: Distribution of molecular complexity and flexibility measurements in druglikeNP dataset; A) FCC, B) Fraction of sp^3 carbons (Fsp^3) and C) molecular flexibility.

4.2.3 Scaffold analysis

The term "scaffold' is commonly used to describe the core structure of a compound to which functional groups are attached (98,99). In this research, Bemis and Murko's definition of a molecular scaffold was used; which is the molecular core of an organic structure without the side chains. Scaffold diversity was analyzed using two approaches; Cyclic System Retrieval (CSR) curves and Shannon entropy (SE) or scaled Shannon entropy (SSE).

Cyclic System Retrieval (CSR) curves Results

CSR curves analyze the distribution and diversity of molecular scaffolds by plots the fraction of chemotypes (x-axis) versus the fraction of compounds that contain those chemotypes (y-axis) (98). The curves can be further characterized by calculating the area under the curve (AUC), and the fraction of chemotypes required to retrieve 50% of the molecules (F50). Figure 10 and table six summaries scaffold diversity of the two datasets evaluated in this study.

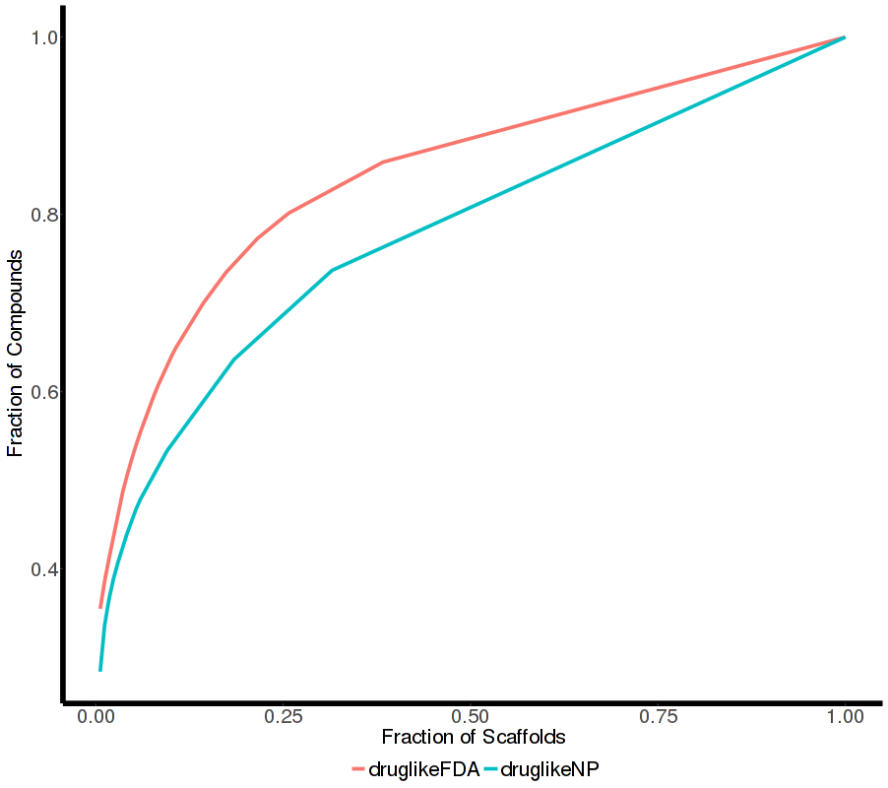


Figure 10: Cyclic system retrieval curves for two datasets evaluated in this study

Table 6: Scaffold diversity summary of the two datasets

Dataset	M	N	FN/M	NSING	FNSING/M	FNSING/N	AUC	F50
druglikeFDA	730	167	0.23	103	0.14	0.62	0.84	0.042
druglikeNP	437	168	0.38	115	0.26	0.68	0.78	0.071

M = number of molecules in the dataset, N = number of chemotypes or scaffolds, FN/M = chemotype diversity fraction, NSING = singleton number, FNSING/M = singleton fraction between total molecules, FNSING/N = fraction of singleton among total chemotypes, AUC = area under the curve, F50 = fraction of chemotype required to recover 50% of the molecules.

Scaled Shannon entropy (SSE) Results

SSE is a metric used to quantify the relationship between the entropy and information content It is useful to measure the scaffold diversity based on the n number of most frequent scaffolds in a dataset (in contrast, the CSR curves provide information of the scaffold diversity of all the data set). A SSE value closer to one indicates maximum scaffold diversity, while SSE closer to zero (0) is an indicative of low scaffold diversity. In this work, SSE for n ranging from 10 to 60 (**Table 7**) was calculated and plotted (**Figure 11**) (100,101).

Table 7: SSE of the 10−60 Most Populated Scaffolds in druglikeFDA and druglikeNP

Datasets	SSE10	SSE20	SSE30	SSE40	SSE50	SSE60
druglikeFDA	0.80	0.83	0.84	0.85	0.85	0.85
DruglikeNP	0.88	0.89	0.90	0.91	0.91	0.92

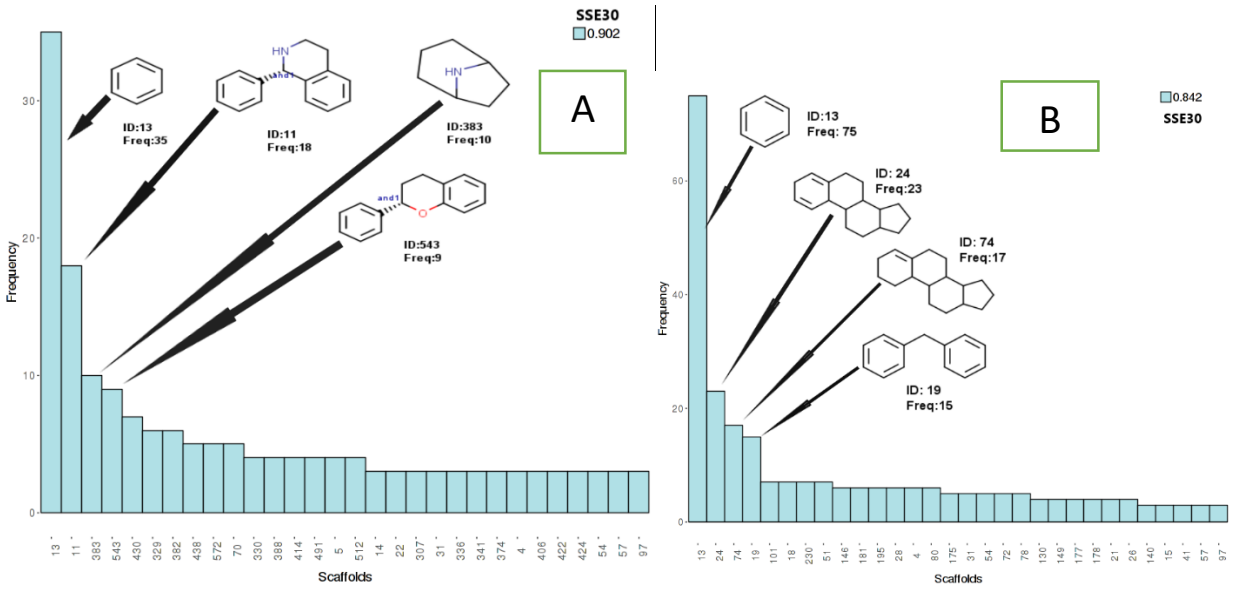


Figure 11: 30 most frequent scaffolds of (A) druglikeNP and (B) druglikeFDA datasets. It indicates the value of SSE for the 30 most frequent scaffolds (SSE30) and the structures of the four most frequent scaffolds in each dataset. The number underneath each bar is an ID assigned to each scaffold.

Figure 12 below show twenty-five scaffolds common in druglikeNP and druglikeFDA datasets

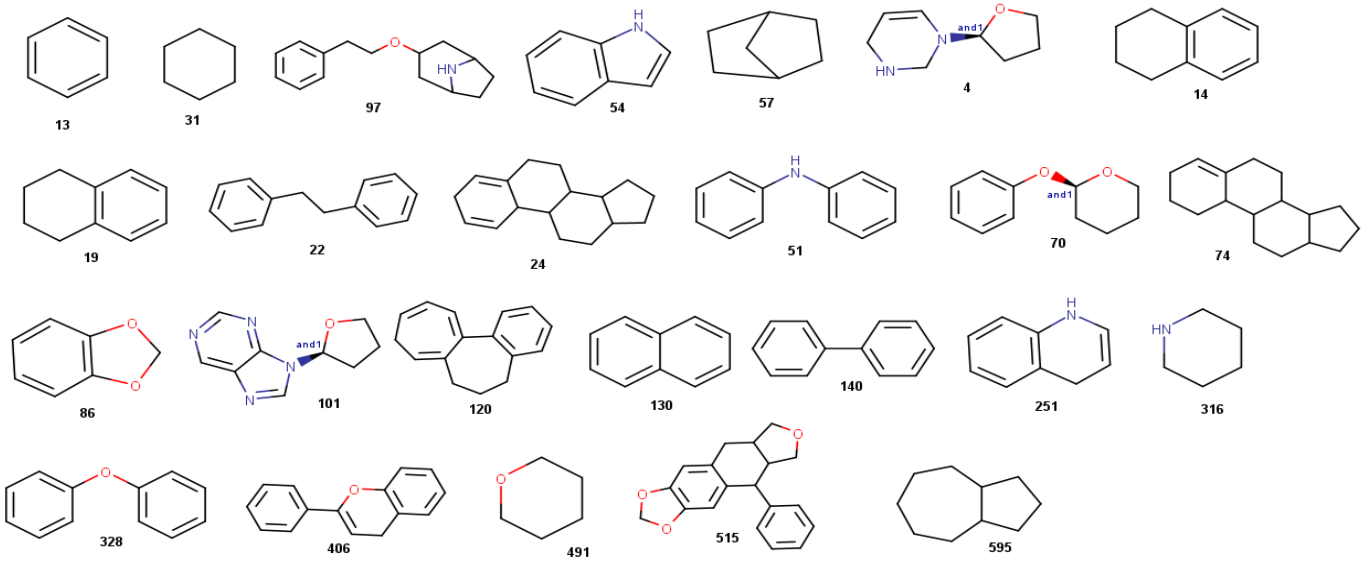
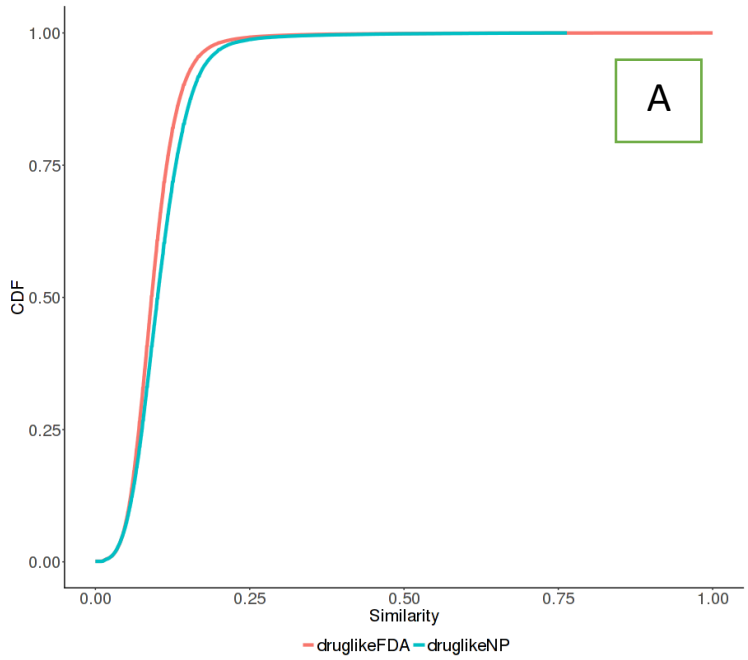


Figure 12: Structures and ID numbers of twenty-five scaffolds common in both druglikeNP and druglikeFDA.

Structural Fingerprints Results

The distribution of the pairwise similarities of the compounds within each dataset (druglikeNP and druglikeFDA) was analyzed by means of cumulative distribution function (CDF) curves (**Figure 13**) for three structural binary fingerprints, including ECFP_4, MACCS keys (166-Bits), and PubChem (881-Bits).



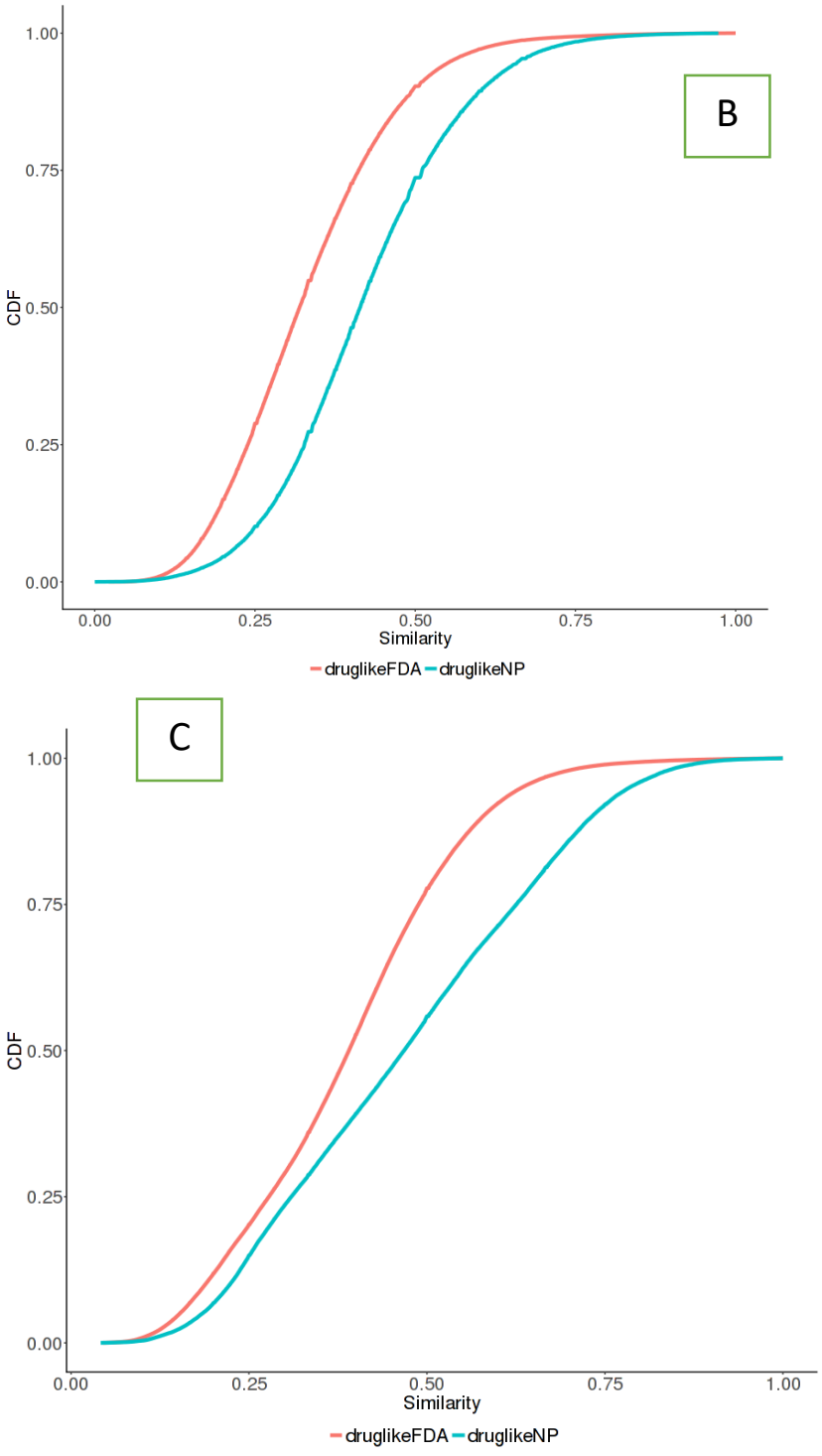


Figure 13: Cumulative distribution functions (CDFs) of the similarity regarding pairwise values computed using extended connectivity fingerprints with a diameter of 4 (A), MACCS keys (B), and PubChem (C).

Table 8: The statistical values of the similarity of the Tanimoto coefficient with ECFP-4.

Fingerprint	Dataset	Min	1st Q	Median	Mean	3rdQ	Max	SD
ECFP_4	DruglikeFDA	0	0.07	0.091	0.097	0.12	1	0.045
	DruglikeNP	0	0.075	0.10	0.11	0.13	0.85	0.051
MACCS(166-bits)	druglikeFDA	0.02	0.24	0.33	0.33	0.41	1	0.13
	DruglikeNP	0	0.33	0.42	0.42	0.51	0.98	0.14
PUBCHEM(881-bits)	druglikeFDA	0.04	0.28	0.39	0.39	0.49	1	0.15
	DruglikeNP	0.02	0.31	0.47	0.47	0.63	1	0.19

Max maximum, Q quartile, SD standard deviation

4.3 Structure-based virtual screening of γ-secretase inhibitors (molecular docking)

Molecular docking is one of the most popular structure-based approaches used to examine the affinity compounds on a particular disease target in which the 3D structure is available (49). In this study, 437 drug-like natural products were virtually screened on a breast cancer drug target, γ-secretase catalytic unit presenilin (PS1) using Autodock vina.13 experimentally validated γ-secretase inhibitors (GSIs) with documented IC_{50} activity were used to validate the docking protocol and were used as reference ligands throughout the docking experiments and analysis.

4.3.1 Validation of docking protocol

Validation of a docking protocol is a crucial step in most structure-based virtual screening campaigns. The purpose of this stage is assess the docking protocol accuracy in pose prediction and affinity-based ranking of the docking software. Docking accuracy is commonly assessed by root mean square deviation (RMSD) calculations comparing predicted pose and experimental pose given than the receptor used as the drug target has a co-crystalized ligand. In this study the PS1 unit was not crystallised with any ligand hence to validate the docking protocol, GSIs were docked into the reported binding site of PS1 (**Table 9**). A graph was plotted to estimate the accuracy of Autodock Vina (**Figure 14**).

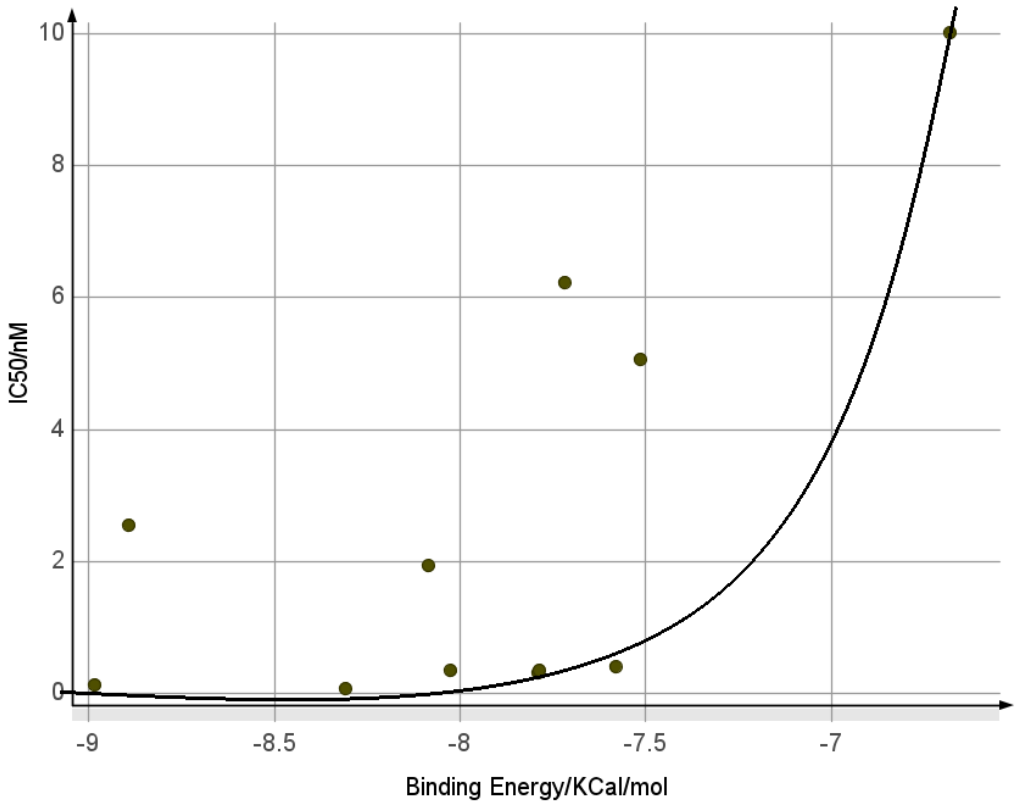


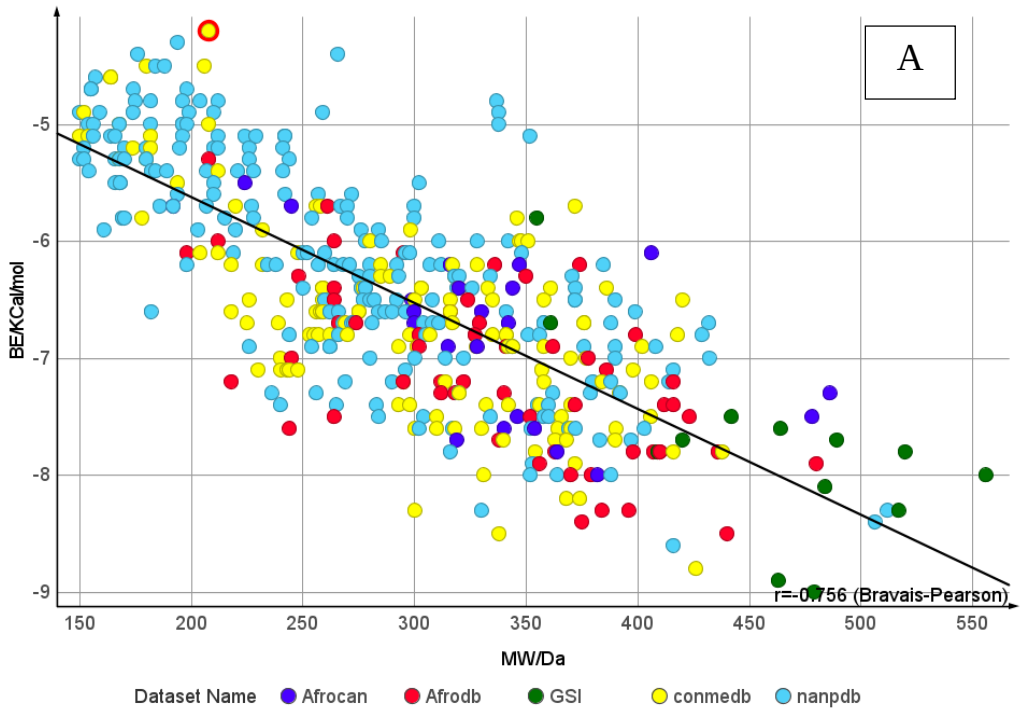
Figure 14: Correlation curve of binding energy and IC_{50} of 13 GSIs from molecular docking results

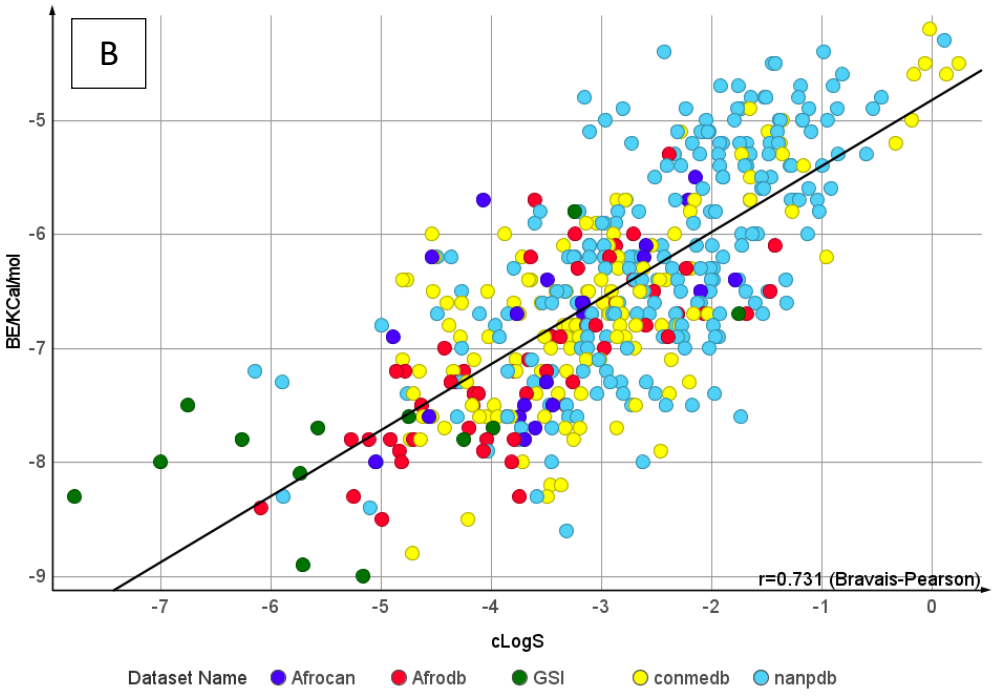
Table 9: GSIs and their reported IC_{50} activity on PS1 unit of γ- secretase

GSIs	Reported IC_{50} activity on PS1 (nM)	Binding Energy/KCalmol^{-1}
LY411575	0.078 (102)	-9.0
YO01027	2.60 (103)	8.9
MRK560	0.14 (104)	-8.3
BMS708163	0.33 (105)	-7.8
BMS906024	0.29 (106)	-8.0
RO4929097	1.88 (107)	-8.1
LY900009	0.27 (108)	-7.8
ADAPT	115 (109)	-7.7
PF3084014	6.2 (110)	-7.7
LY3039478	0.40 (111)	-7.6
MRK0752	5.00 (112)	-7.5
LY450139	10.0 (113)	-6.7
IMR1	2600(114)	-5.8

4.3.2 Ranking of docked hits according to the binding energy

The binding energies of 437 docked NPs were plotted against molecular weight, calculated lipophilicity (Xlog P), and calculated aqueous solubility (clogS) of the 437 NPs as shown in (**Figure 15**). This was done to understand the correlation between binding energy and pharmacologically relevant drug properties.


A
BE/KCal/mol
-5
-6
-7
-8
-9
150
200
250
300
350
400
450
500
550
MW/Da
r=-0.756 (Bravais-Pearson)
Dataset Name
Afrocan
Afrodb
GSI
conmedb
nanpdb


B
BE/KCal/mol
-5
-6
-7
-8
-9
-7
-6
-5
-4
-3
-2
-1
0
cLogS
r=0.731 (Bravais-Pearson)
Dataset Name
Afrocan
Afrodb
GSI
conmedb
nanpdb

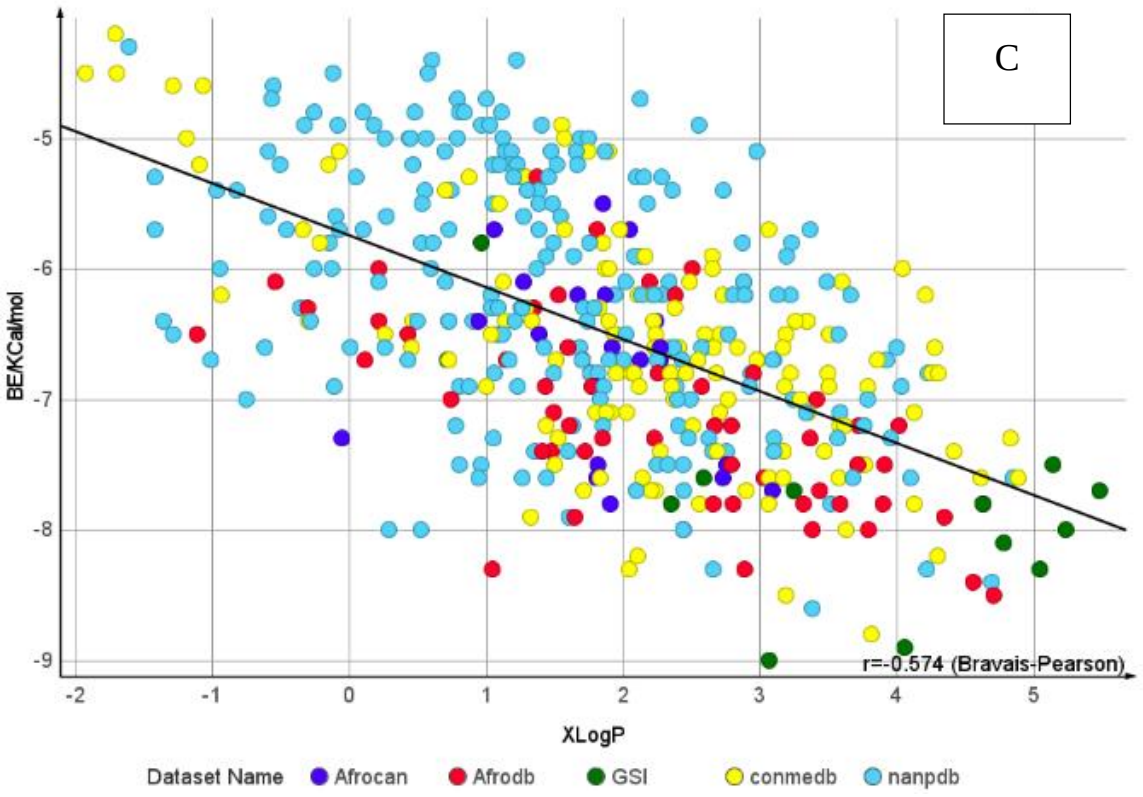


Figure 15: A scatter plot of binding energy of DruglikeNP compounds against different physicochemical properties; A) molecular weight (MW), B) Aqueous solubility and C) Lipophilicity (XlogP). The compounds are also colored according to database origins of natural products; Afrocan (red dots) for NPs sourced from Afrocancer database.

Docked NPs were clustered based on their molecular properties using t-SNE. **Figure 16** shows a chemical space plot of docked compound clusters (A, B, C), colored according to their binding energies.

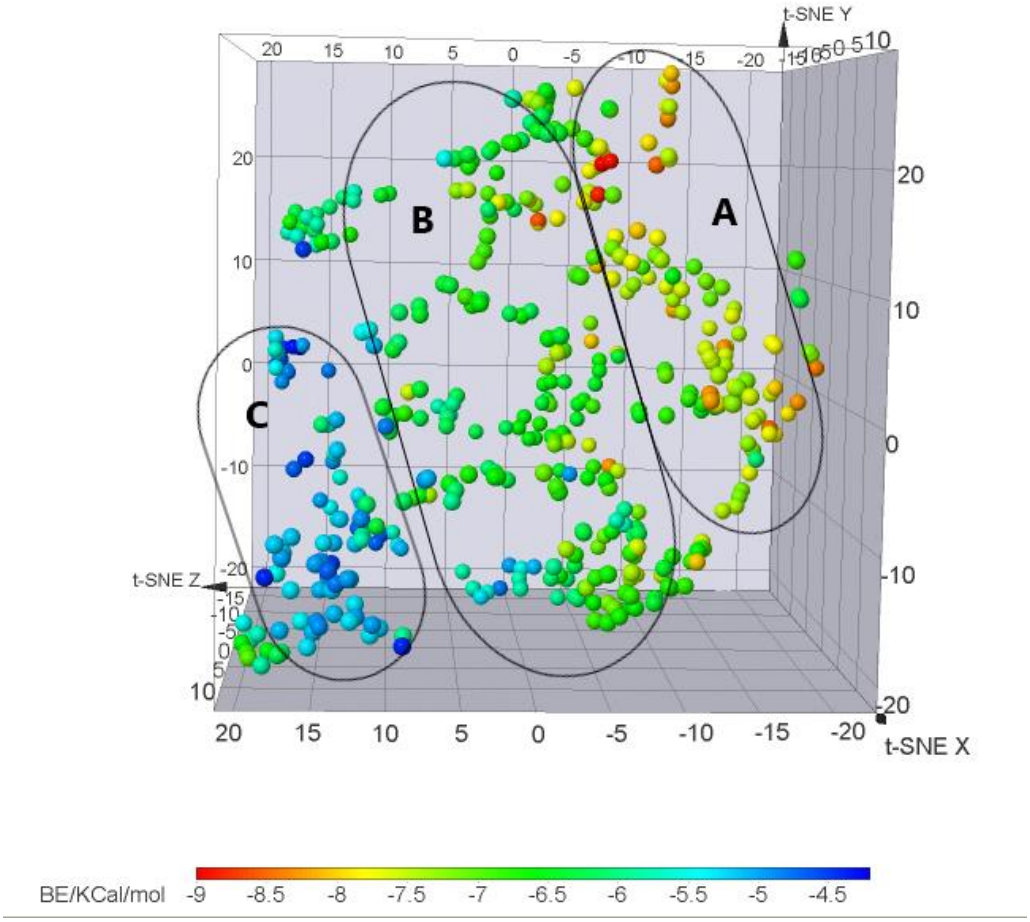


Figure 16: t-SNE chemical space showing molecular property-based similarity of NPs docked compounds colored according to binding energies. Circles show visible clusters of similar compounds A, B, and C.

4.3.3 Ranking of NP drug hits according to protein-ligand interaction of PSI-NPs complexes

72 compounds with binding energy less than -7.5 Kcal/mol (**Figure 17**) were screened further based on protein-ligand interactions. The purpose of this stage was to identify compounds with intermolecular interactions similar to the ones present in GSIs-PS1 complexes as well as identify novel compounds with unique interactions.

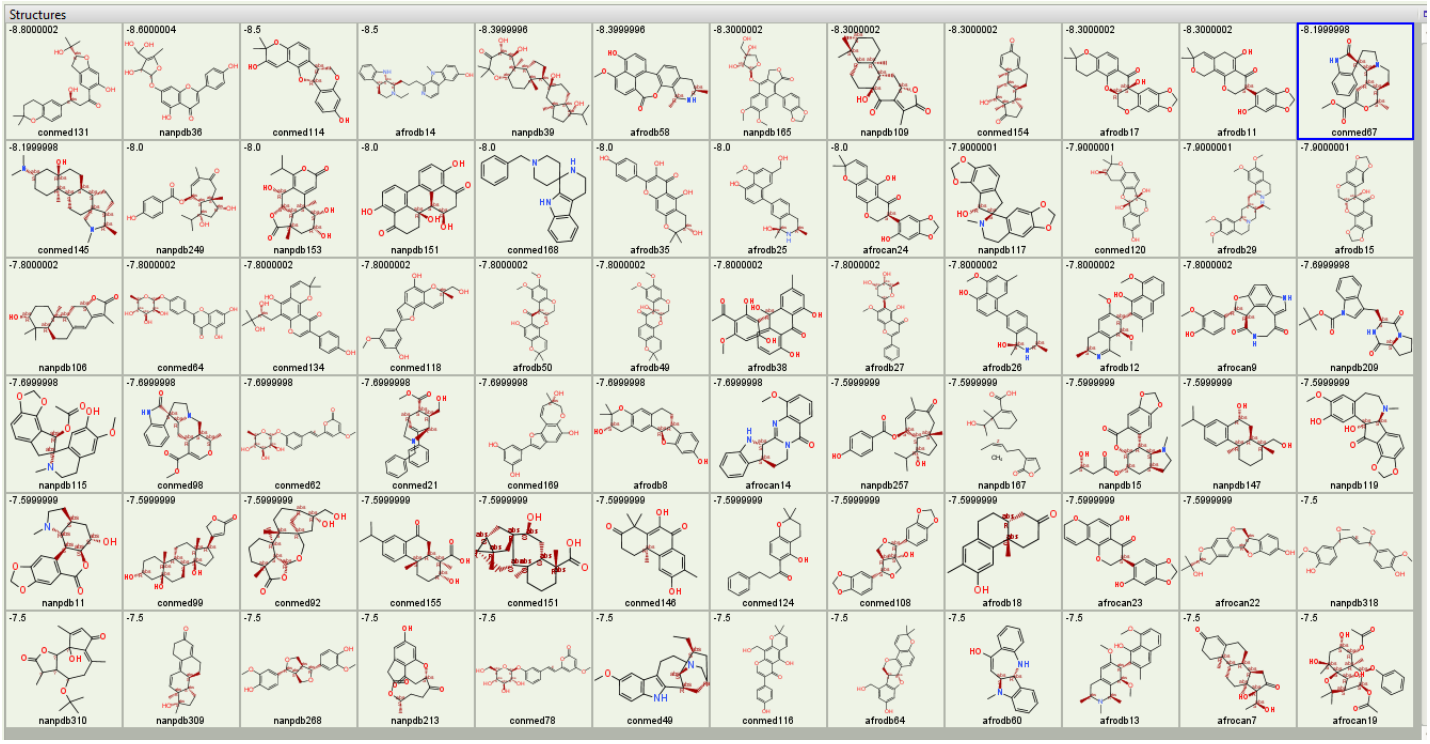


Figure 17: Structures of 72 NPs with binding energy less than -7.5Kcal/mol selected for the targeted protein

To accurately select promising compounds based on interactions, 13 GSI-PS1 complexes were visually inspected to identify amino acid residues common to GSIs on the PS1 binding site. **Table 10** list the binding properties of GSI-PS1 complexes generated in this study. Common amino acids residues (**Table 11**) identified at this stage were used as a reference in screening NP compounds.

Table 10: List of 13 GSIs with associated amino acid residues and types of intermolecular interactions in GSIs-PS1 complexes. **Bolded** are residues that showed to interact with 4 or more GSIs, * residues are the ones that were experimentally reported to interact with ADAPT

GSIs	Binding Energy/KCalmol^{-1}	Interacting Residues	Types of Interactions
LY411575	-9.0	**LEU286, LEU268, GLY384*, ILE143**	• H-bond • C-H bond • Pi-Alkyl
YO01027	8.9	**LEU150, LEU166,** MET146***, TRP165***, **GLY384*, ASP257, ILE143**	• C-H bond • Halogen(Fluorine) bond • Pi-Pi stacked • Alkyl • Pi-Alkyl
MRK560	-8.3	**LEU166, LEU286, TRP165*, ILE143, ASP257, LEU150**, THR147, PHE283*	• C-H bond • Halogen(Fluorine) bond • Pi-sigma • Pi-sulfur • Alkyl • Pi-Alkyl
BMS708163	-7.8	LEU283, **LEU150, LEU286, LEU166, TRP165*, GLY384*,** ALA285, ILE229, ILE213	• H-bond • C-H bond • Halogen (Fluorine) bond • Pi-Donor hydrogen bond • Pi-Sigma • Pi-Pi stacked • Alkyl • Pi-Alkyl
BMS906024	-8.0	ASP385, **ASP257,** ALA434, ILE287, **LEU268, LEU268**, LEU282	• H-bond • Alkyl • Pi-Alkyl

RO4929097	-8.1	**ASP257**, THR147, **GLY384*, ILE143, LEU268**, ALA434	• H-bond • C-H bond • Halogen(Fluorine) bond • Pi-Alkyl
LY900009	-7.8	MET146***, LEU166,** GLY384*****, TRP165***, LEU150**	• C-H bond • Pi-Pi stacked • Alkyl • Pi-Alkyl
ADAPT	-7.7	**ASP257**, ILE287, **GLY384*,** GLY382, LEU435, LEU282, LEU296, ASP385	• H-bond • C-H bond • Halogen(Fluorine) bond • Pi-Anion • Alkyl • Pi-Alkyl
PF3084014	-7.7	**LEU268, LEU286**, ASP385, LEU271	• H-bond • Halogen(Fluorine) bond • Alkyl • Pi-Alkyl
LY3039478	-7.6	**LEU268,** LEU381, ILE287, ASP285, GLY287	• H-bond • C-H bond • Halogen(fluorine) bond • Alkyl • Pi-Alkyl
MRK0752	-7.5	GLY382, LEU432, **LUE150**, LEU433,**LEU268**	• H-bond • C-H bond • Alkyl • Pi-Alkyl
LY450139	-6.7	SER169, SER170, **LEU286, LEU166, LEU150**	• C-H bond • Alkyl • Pi-Alkyl

IMR1	-5.8	**ILE143,** GLY282, LEU432 **ASP257**, ALA434,	• H-bond • C-H bond • Pi-Alkyl

Table 11: List PS1 interacting residues that were present in four or more GSI-PS1 complexes and were used to screen 72 NP compounds

Interacting Residues	Frequency in GSI-PS1 complexes
TRP165	4
GLY384	5
ILE143	5
LEU166	5
LEU150	6
LEU286	8
ASP257	5
LEU268	5

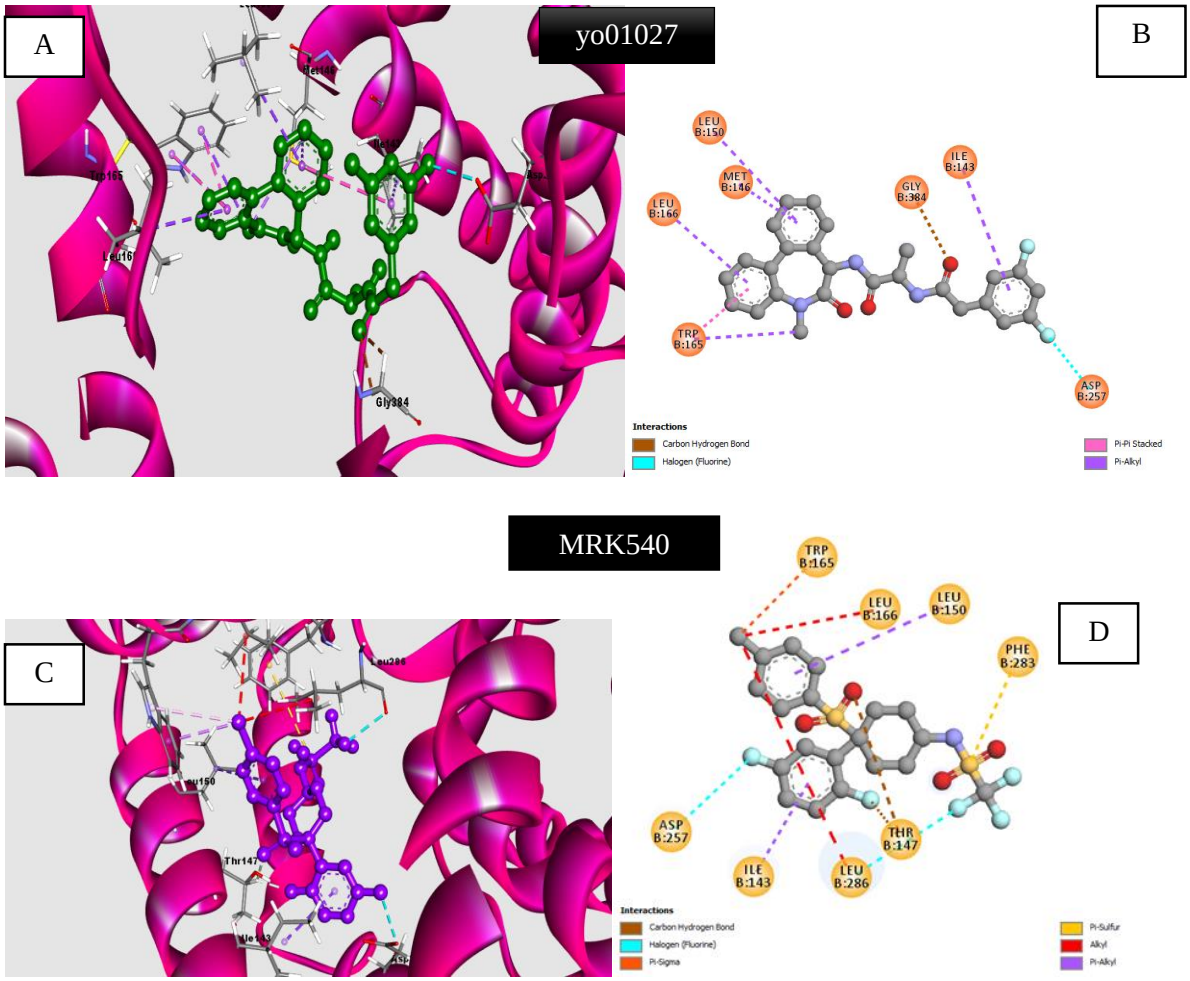
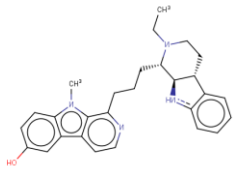
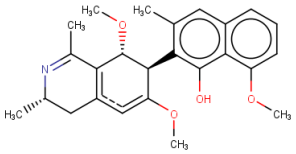
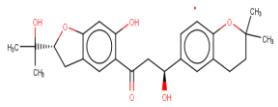
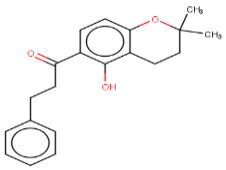
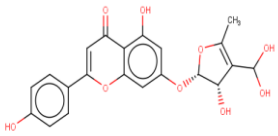
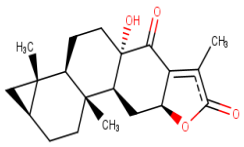
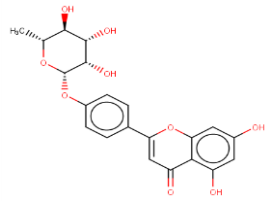
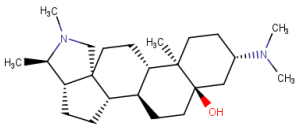
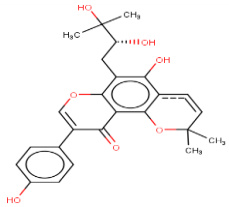
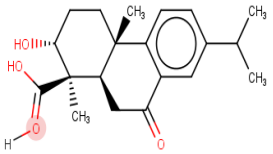
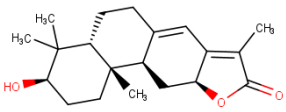
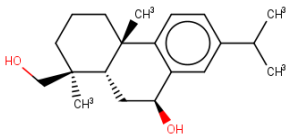


Figure 18: A-D 2D (right) and 3D (left) images of some of GSI-PS1 complexes; yo01027. Non-bonding intermolecular interactions within the complexes include; conventional hydrogen bond(**green**), Carbon-hydrogen bond **(brown**), Pi-Pi stacked (**pink**), Pi-sigma (**orange**), Pi-alkyl **(purple**), Pi-Sulphur **(yellow**) and halogen bond (**Fluorine)**.

Table 12: Table showing the Binding energy, Interacting residues and associated interactions of 12 best NP compounds. The residues in **bold** are residues that had a greater frequency of occurrence in GSI-PS1 complexes as shown in **Table 11** and the * are residues experimentally reported to interact with ADAPT

Name of NP	Binding Energy	Interacting Residues	Types of interactions
Afrodb14 	-8.5	LEU383,**LEU286,LEU166, TRP165**,MET146*,**LEU268, LEU150**	• Alkyl • Pi-Alkyl • Pi-Sigma • Pi-Pi T shaped
Afrodb12 	-7.8	**ASP257,LEU166,TRP165, LEU286,LEU150,ILE143**, MET146,* SER169	• Alky • Pi-Alkyl • Pi-Sigma • Carbon-H bond
Conmed131 	-8.8	**LEU268,LEU286,LEU150,** LEU271,LEU434,LEU282, LEU257,PHE283*,ILE287	• Alkyl • Pi-Alkyl • H-bond • C-H bond
Conmed124 	-7.6	LEU271,LEU282,**LEU268, LEU286**,PHE283*,**ASP257**, ASP385,**ILE143**, **GLY384***	• Alkyl • Pi-Alkyl • Pi-Anion • H-bond
Nanpdb36 	-8.6	**LEU150,LEU286,LEU166, LEU268**,LEU435,**TRP165***, THR147,SER169,PHE388	• Alkyl • Pi-Alkyl • Pi-Pi Stacked • H-bond • C-H bond

Nanpdb109 	-8.3	**LEU268,LEU150,LEU286, GLY384***,LEU271,LEU282, PHE388,PHE283*,LEU383	• Alkyl • Pi-Alkyl • H-bond
Conmed64 	-7.8	**ASP257,ILE143,LEU286,** SER169,ALA285,ILE387, ILE229,LUE383	• Alkyl • Pi-Alkyl • Pi-Sigma • H-bond • C-H bond
Conmed145 	-8.2	**ILE143,LEU268,ASP257**, LEU435,GLY382,PHE388, LYS380,ALA434	• Alkyl • Pi-Alkyl • H-bond • C-H bond
Conmed134 	-7.8	**LEU268,LEU286,ILE143**, LEU381,LEU271,LUE282, LEU435,PHE388,ILE253	• Alkyl • Pi-Alkyl • Pi-Sigma • C-H bond
Conmed155 	-7.6	**LEU268,LEU286,LEU150**, LEU432,LEU282,LEU271, PHE283*,ILE287,VAL272, VAL261	• Alkyl • Pi-Alkyl • H-bond
Nanpdb106 	-7.8	**LEU166,LEU286**,MET233*, LEU383,LEU226,SER170, ALA285,ILE387,ILE229	• Alkyl • H-bond
Nanpdb147 	-7.6	**LEU150,LEU286,LEU268**, LEU271,LEU282,VAL271, VAL261,ALA431,PHE283*	• Alkyl • Pi-Alkyl • Pi-Sigma

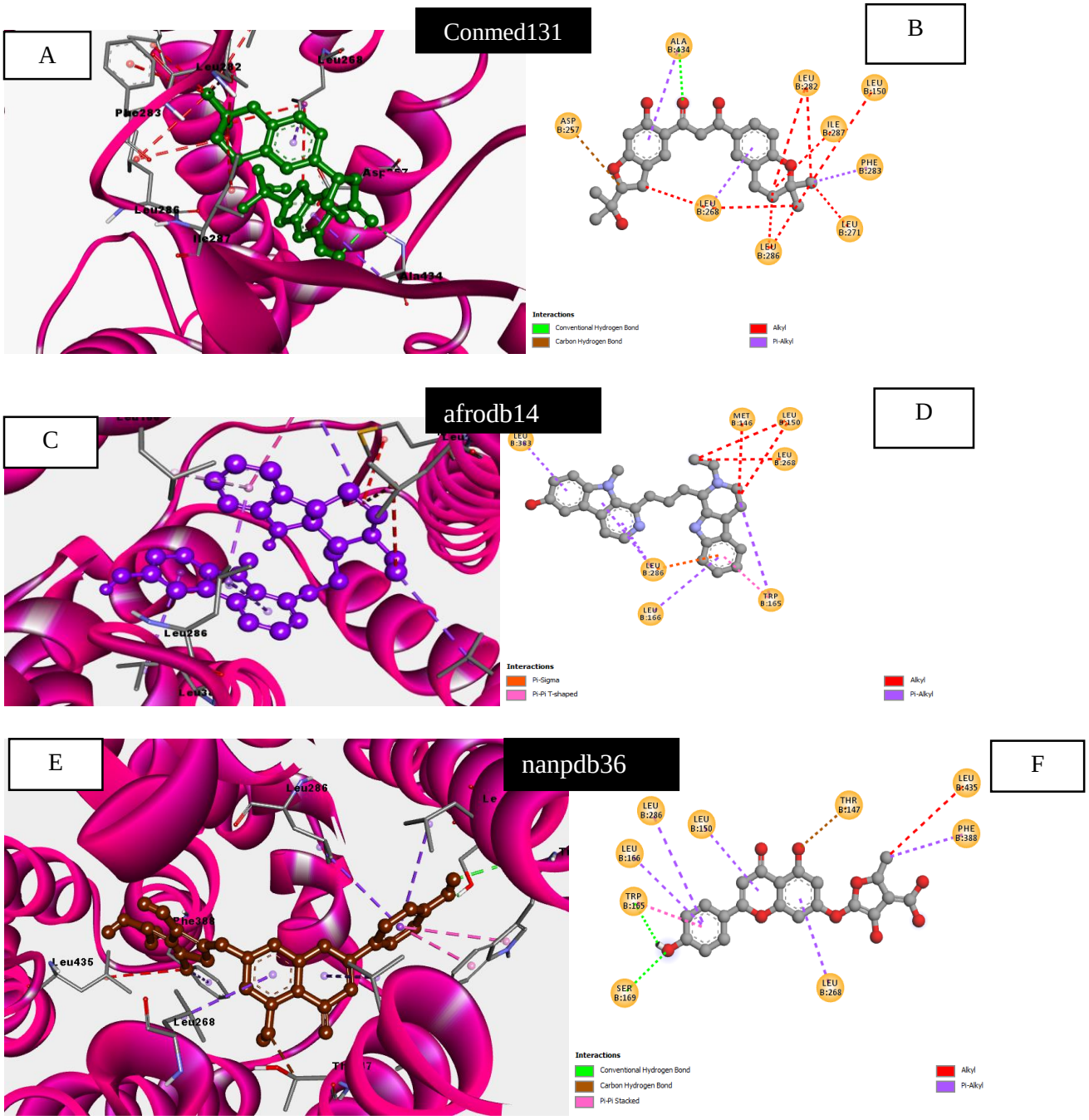


Figure 19: Showing 2D (right) and 3D (left) images of NP-PS1s complexes of three NPs; conmed131, afrodb14 and nanpdb36. Non-bonding intermolecular interactions within the complexes include; conventional hydrogen bond (**green**), Carbon-hydrogen bond **(brown**), Pi-Pi stacked and Pi-T shaped (**pink**), Pi-sigma (**orange**), Pi-alkyl **(purple**) and Alkyl (**red**).

4.4 Pharmacokinetic (PK) property analysis of 12 most promising compounds

Pharmacokinetic properties of 12 most promising compounds were predicted using an online web server pkCSM, which uses graph-based signatures. The parameters investigated were categorized into five fields popularly used to describe PK properties of compounds. Absorption, distribution, metabolism, excretion, and toxicity.

4.4.1 Absorption

Table 13: Absorption profiles of 12 most promising compounds

Compound	Solubility(LogS)	Caco-2 Permeability log(Papp)/10^{-6} cm/s)	Intestinal absorption(human)/% Absorbed
Afrodb14	-3.8	1.07	88
Afrodb12	-5.1	1.06	94
Conmedb131	-4.1	1.05	87
Conmedb124	-4.5	1.37	95
Conmed64	-3.22	0.219	53
Conmed145	-3.0	1.19	88
Conmed134	-3.4	0.006	83
Conmed155	-3.9	0.94	93
Nanpdb106	-4.1	1.31	96
Nanpdb147	-4.6	1.61	93
Nanpdb36	-3.9	0.467	64
Nanpdb109	-3.9	1.06	99

Caco-2 cell permeability (log Papp in 10−6 cm/s >0.09); intestinal absorption (human), % absorbed (>30).

4.4.2 Distribution

Table 14: Distribution profiles of 12 most promising compounds

Compound	Volume of Distribution VDss /Log(VDss)	Fraction unbound	BBB permeability (LogBB)	CNS permeability Log(PS)	P-gp substrate	P-gp I inhibitor	P-gp II inhibitor
Afrodb14	1.4	0.07	0.12	-1.76	+	+	+
Afrodb12	0.7	0.04	0.28	-1.72	+	+	+
Conmedb131	0.59	0.13	-0.89	-2.92	+	+	-
Conmedb124	0.51	0	0.21	-1.45	-	+	-
Conmed64	-0.30	0.09	-1.6	-3.76	+	-	-
Conmed145	0.28	0.45	0.7	-1.82	+	+	-
Conmed134	-0.17	0.09	0.09	-2.04	+	+	+
Conmed155	-0.29	0.11	-1.4	-2.96	-	-	-
Nanpdb106	0.13	0.16	0,06	-2.36	-	+	-
Nanpdb147	0.86	0.01	-0.03	-1.86	-	-	-
Nanpdb36	-0.33	0.03	-1.7	-4.14	+	-	+
Nanpdb109	0.09	0.13	-0,02	-1.83	-	+	-

BBB: Blood-brain barrier, CNS: central nervous system, PS: permeability-surface area product, VDss (human) (log L/kg) (low if<−0.15 and high if >0.45), logBB (low if <-1 and high if>0.3, logPS (low if<-3 and high if >-2), P-gp: P glycoprotein

4.4.3 Metabolism

Table 15: Metabolism profiles of 12 most promising compounds

Compound	CYP1A2 inhibitor	CYP2C19 inhibitor	CYP2C9 inhibitor	CYP2D6 inhibitor	CYP3A4 inhibitor
Afrodb14	+	+	+	+	-
Afrodb12	-	+	-	+	-
Conmedb131	-	+	-	+	-
Conmedb124	+	+	-	+	+
Conmed64	-	-	-	-	-
Conmed145	-	-	-	-	-
Conmed134	-	-	-	-	-
Conmed155	+	+	-	+	-
Nanpdb106	-	-	-	-	+
Nanpdb147	+	+	-	-	-
Nanpdb36	-	-	-	-	-
Nanpdb109	-	-	-	-	+

4.4.4 Elimination/Excretion

Table 16: Excretion profiles of 12 most promising compounds

Compound	Total Clearance Log(CLtot)	Renal OCT2 substrate
Afrodb14	0.97	-
Afrodb12	0.99	-
Conmedb131	0.24	-
Conmedb124	0.99	+
Conmed64	0.58	-
Conmed145	0.31	-
Conmed134	0.03	-
Conmed155	0.76	-
Nanpdb106	0.61	+
Nanpdb147	0.84	-
Nanpdb36	0.76	-
Nanpdb109	0.45	+

4.4.5 Toxicity

Table 17: Toxicity profiles of 12 most promising compounds

Compound	AMES toxicity	hERG I inhibitor	hERG II inhibitor	Hepatotoxicity	Skin sensitization
Afrodb14		-	+	+	-
Afrodb12	-	-	+	-	-
Conmedb131	-	-	+	-	-
Conmedb124	-	-	+	-	-
Conmed64	-	-	+	-	-
Conmed145	-	-	+	-	-
Conmed134	-	-	+	-	-
Conmed155	-	-	-	-	-
Nanpdb106	-	-	-	-	-
Nanpdb147	-	-	+	-	-
Nanpdb36	-	-	+	+	-
Nanpdb109	-	-	-	-	-

5. DISCUSSION

5.1 Curation and isolation of drug-like natural products

In this study four compound natural products databases of African origin (NANPDB (36), Afrodb (41), Afrocancer (42), ConMedNP (33) were compiled, curated to come up with one dataset which constituted of natural products that didn't contain structural alerts PAINS and in whose physicochemical properties were compliant to Lipinski and Veber criteria of drug-likeness. Out of the 11304 NP compounds processed in this study, 2477 (22%) were filtered as structural duplicates, 3637 (32%) were filtered as structures that were disconnected or ambiguous and six (0.05%) were removed as sugars. This stage revealed that the four compound libraries contained a considerable number of structures that were originally captured with improper chemical structures, while the presence of duplicates revealed the four libraries had a considerable number of identical structures. The final step in creating a small drug-like dataset of natural products involved filtering all structures that disobeyed all Veber and Lipinski rules of drug-likeness as well as the removal of compounds containing PAINS substructural alerts and other alerts defined by Brenk (87), NIH (115) and the ZINC databases. 3447 (30%) of the compounds were non-Lipinski and non-Veber compliant while only 1300 (11%) NPs contained structural alerts and PAINS. Lipinski and Veber drug-like rules are common drug rules used to predict the oral bioavailability of drug candidates (116). The percentage of compounds with zero violations of drug-like rules and without PAINS and other toxiphores made only 4%. Pan-assay interference structures (PAINS) are substructures which were reported to give false-positive results in high-throughput screens as react nonspecifically with many biological targets rather than specifically affecting one desired target (88). Other small compound filters such as Brenk, ZINC, and NIH identify and remove

structural alerts that were reported responsible for toxicity in multiple drug assays. During the compilation of the natural product databases used in this study, the identification of toxiphores was not reported; therefore, this study is the first to identify pharmacologically unwanted substructures in NANPDB, ConmedNP, Afrocancer, and Afrodb. The resulting dataset named druglikeNP consisted of 437 compounds (4% of 11304 compounds). This study therefore established that African natural products contain a very small percentage of compounds that do not violet any Lipinski and Veber rules of drug-likeness and unwanted substructures.

5.2 Chemoinformatic characterization of DruglikeNP dataset

Chemoinformatic analysis of the natural product drug-like dataset in this study was done mainly through statistical analysis of physicochemical properties and scaffold diversity. This exercise revealed that the natural product dataset contains lead-like properties and were more diverse and complex that drug-like FDA drugs. Each parameter computed is discussed comprehensively in the following sections.

5.2.1 Molecular properties

Topological polar surface *(TPSA*): A majority of compounds in druglikeNP dataset (**figure 7A** and **Table 4)** (86%) had TPSA ranging between 30-110 $Å^2$. 87% of these compounds were found with TPSA less than 90 $Å^2$ with mean of 75 $Å^2$, meaning than most of the good drug could be used as psychotic drug candidates given that there are bioactive (117). A small portion of the compounds (4%) could be classified as too polar since they had TPSA ranging between 150 and 130 $Å^2$.

Molar refractivity (MR): In this study it was observed that 99% (**Figure 7B**) of the compounds fell within MR range of 40-130 and the mean molar refractivity was 76 (**Table 4**). According to Ghose(1999) compounds with such MR range are considered drug-like as they have good drug polarizability(86,89,118)(119).

Rotatable bond count (RB): (**Figure 7C**) 86% of compounds in druglikeNP dataset had less than or equal to five RB and 55% of the natural products had three or less rotatable as shown by average of 2.55 (**Table 4**). While the relevance of rotatable bonds in drug-likeness measurement was initially proposed by Veber in extension to Lipinski(54), Congreve et al (120) used the parameter in classification of fragment-like compounds called the rule of three. Therefore, more than half of the compounds in druglikeNP dataset fit in the lead-like criteria of rotatable bonds.

Molecular weight (MW): molecular weight is one the most important descriptor in drug design, it has been associated with drug complexity, Lipophilicity, and bioavailability (89). In druglikeNP, 99% of the compounds had MW between 150 and 145 Daltons (**Figure 7D**), the average MW calculated to 291, and the median was observed to 295(**Table 4**). According to Ghose (119), bioavailable drugs should have MW ranges between 180 and 480, 93% of the dataset fulfilled that criteria. 63% of the compounds had lead-like mass ranges (MW<300).

Number of HBA & HBD: These two property descriptors define a drug's capability to bind, which in turn determines the polarity and permeability. High hydrogen bonding capacity reflects low permeability that is a direct measure of its low solubility(54,121).70% of the druglikeNP had HBD count<2(**Figure 7E**), while 83% had HBA count<5(**Figure 7F**). From these results it is safe to

infer that the dataset was constituted of compounds with few very electronegative atoms (mostly oxygen and nitrogen) bonded or non-bonded to hydrogen atoms. The average number (**Table 4**) of HBD and HBA (2.81 and 2.88 respectively) testified that the majority of compounds fit into the HBD and HBA count criteria of fragments or lead-like compounds.

Aqueous solubility (LogS) and Lipophilicity (XlogP)*:* **Figure 7H** show that 88% of compounds in the druglikeNP dataset have logS ranges between -4.6 and -0.6. 20% of the compounds had logS values ranging between -4.6 and 3.6. Lipophilicity descriptor, logP (log Octanol/water partition coefficient) is used to evaluate the compounds 'probability of transportation through lipid membrane. Increased LogP of a drug is associated with greater lipophilicity. A compound with LogP value range between -2 and 5 is consider good lead candidate. 84% of the dataset was within the range of 0 and 4 (**Figure 7G**).

5.2.2 Chemical space visualization

In this study, the chemical space druglikeNP was mapped comparatively to 730 FDA drugs with no violation of drug-like rules using t-distributed stochastic neighbor embedding (t-SNE) method (91). Eight chemical descriptors in **Table 4** were used. The space occupied druglikeNP (**Figure 8**) was similar to the one occupied with druglikeFDA, although there are areas of in which approved drugs occupied only. These results showed how similar druglikeNP natural products are to FDA approved drugs as most natural products were packed near approved drug clusters.

5.2.3 Molecular complexity and flexibility analysis.

97% (**Figure 9A**) of the investigated NPs had FCC less than 0.5, while 33% (**Figure 9B**) of the NP dataset had Fsp^3 above 0.5. The reference dataset used druglikeFDA also contained few compounds with Fsp^3 and FCC more than 0.5 as the average values (**Table 5**) for FCC and Fsp^3 were 0.093 and 0.41 respectively. Several studies have reported that 84% of marketed drugs have average values of $Fsp^3 \geq 0.42$ (122).The natural products used in this study reported an average Fsp^3 of 0.43 (**Table 5**). From these results it can be concluded that fewer natural compounds in druglikeNP are more likely to have a 3D structure and a less complex (100). The stereochemical complexity of both dataset was comparable lower on average as reported by average values of FCC for both datasets. The molecular flexibility of compounds has been known to influence the stability of drug-protein interaction. You et al , reported in 2016 that keeping the molecules flexible, as opposed to making them rigid, both reduced the entropy penalty and created a stronger binding (123). This study compounds in druglikeNP (0.29) were reported to have a lower average of molecular complexity index than approved drugs in druglikeFDA (0.39).

5.2.4 Scaffold diversity analysis

Two methods of measurements were employed in this study to explore scaffold diversity of druglikeNP; cyclic system retrieval (CSR) curves and Scaled Shannon entropy (SSE). Diversity based on fingerprints analysis was done using cumulative distribution function (CDF) curves and Tanimoto coefficient.

Diversity based on cyclic system retrieval (CSR) curves: These curves are used in diversity analysis to capture the distribution of compounds in the cyclic systems of a compound collection.

The lower the area CSR curve (AUC), the larger the scaffold diversity(101). **Figure 10** shows that druglikeNP is more diverse as its curve is closer to the diagonal compared to the one for druglikeFDA. Calculated AUC for druglikeNP was lower (0.78) than that of druglikeFDA (0.84). The difference in the number of scaffolds or chemotypes (N) is very small, as druglikeNP has 168 scaffolds (**Table 6**) compared to druglikeFDA with 167 yet the fraction of scaffold in druglikeNP (N/M) is greater than that of druglikeFDA. The fraction of singletons (fraction of scaffolds having one compound) (FNSING/N and FNSING/M) of druglikeNP shows that drug-like natural products contains more unique scaffolds relative to the total number of molecules or scaffolds than druglikeFDA. Scaffold diversity-based on CSR curves showed that the druglikeNP dataset was more diverse and contained more unique scaffolds than the drug-like FDA approved dataset. This analysis confirmed what was reported by Saldívar-González et al., 2019 (83) that most natural product libraries are chemically diverse than drug libraries.

***Diversity based on Scaled Shannon entropy (SSE)*:** The scaled Shannon entropy (SSE) is a normalized value that measures the most common chemotypes present in a databases (101). Unlike CSR curves, SSE is more comprehensive as it measures the scaffold diversity of the most populated scaffolds. It takes values ranging from 0 up to 1; 0 indicates that the majority of compounds share the same scaffold (minimum diversity) while a value closer to 1.0 indicate that the compounds are evenly distributed between the n acyclic and/or cyclic systems (maximum diversity) (97,124). SSE values for 10-60 most frequent scaffolds was measured for druglikeNP and druglikeFDA in this study. The scaffold diversity of druglikeNP is higher than that of druglikeFDA (**Table 7**) (SSE values ranging from 0.88 to 0.92 and 0.80 to 0.85, respectively). Scaffold with ID: 13, 11, 383 and 543, had the highest frequency in druglikeNP (**Figure 11A**)

whilst scaffold ID: 13, 24, 74 and 19 dominated in druglikeFDA (**Figure 11B**). The two datasets only shared one scaffold which was more frequent in each dataset; the benzene ring. It is reported that this scaffold makes up most of the databases which have been recently studied (11). Only 25 scaffolds (**Figure 12**) were identified to be common between druglikeNP and druglikeFDA. This suggests that the two datasets are very diverse from each other.

Diversity based on fingerprint similarity*:* In this study, three structural binary fingerprints, including ECFP_4, MACCS keys (166-Bits), and PubChem (881-Bits) were computed using PUMA (94) for the two datasets; druglikeNP and druglikeFDA. To understand the fingerprint diversity of the two datasets, the Tanimoto index as a similarity coefficient was used and represented in cumulative distribution functions (CDF) (**Table 8** and **Figure 13**). The curves obtained with ECFP-4 (**Figure 13A**) did not prove to be a suitable fingerprint representation for these datasets as it showed that all datasets are occupying the same mathematical path without an observed significant difference. Previous similar studies on several databases have also came to the same conclusion (101,124,125). All the pairwise intra-library similarity values in **Table 8** show that druglikeNP is the least diverse set with median MACCS keys and PubChem similarity values of 0.33 and 0.39, respectively. Contrastingly, druglikeFDA was the most diverse with MACC keys and PubChem similarity values of 0.24 and 0.28, respectively. The druglikeFDA was previously shown to be the least diverse (**Table 7** and **Figure 11**) when studied based solely on the scaffolds, which means that acyclic systems and side chains are the contributing factors to the diversity of its compounds.

5.3 Structure-based virtual Screening (molecular docking) of drug-like natural products (druglikeNP)

Structure-based virtual screening (SBVS) of small compound datasets is now the fastest approach used to identify novel leads in most drug design campaigns(49). The method is fast and relatively cheaper than its experimental counterpart is, high throughput screening (HTS) is. SBVS was used for the first time to identify novel γ-secretase inhibitors from 437 natural product compounds, the "druglikeNP" dataset, in effort to discover novel cancer drugs. The campaign protocol used in this study can be broken down into four major steps; protocol validation and docking, ranking hits based on binding energy, and ranking hits based on protein-ligand interactions.

5.3.1 Protocol validation and molecular docking

Autodock Vina is a widely used academic, free access molecular docking program that has been regarded as accurate in hit identification and ranking. In a recent study to comprehensively evaluate the prediction accuracy of sampling power and scoring power of the 10 most popularly used docking programs (free and commercial), Autodock Vina was recorded to have the best scoring power (126,127). In this study, the same program was used to screen natural products on the binding site of the catalytic unit of the γ-secretase enzyme, presenilin (PS1). To validate the docking protocol validate 13 γ-secretase inhibitors (GSIs) (**Table 9**) docked around PS1 reported binding site residues Met146, Met233, Trp165, Phe283 and Gly384 (128). The correlation curve (**Figure 14**) showed that Auto dock Vina could rank GSIs in the same order of their experimentally reported IC_{50} activities because 8/11 γ-secretase inhibitors were ranked the same order as their IC_{50} values. Docking protocol assessment is a crucial step in almost every structure-based virtual screening campaign. The most popularly used method to validate docking methods is redocking

the co-crystalized ligand into the binding site of the receptor under study. If the best root mean square deviation between the docked conformation and the co-crystalized is less than 2Å it means the protocol and the program are reliable enough (129). Enrichment factor EF measurements are also popular in validating molecular docking programs as the docking program is tested to check if it can differentiate active compounds (actives) and inactive (decoys) ones(130). In this study, a compromise as the cryo-structures of γ-secretase enzymes available in protein data banks did not have a co-crystalized during the time they were accessed June 2020 (131).

5.3.2 Binding energy-based ranking of 437 docked natural products

Estimation of binding affinities of small molecules in the binding region of a protein is one of the most fundamental steps in docking simulation. Mathematical functions called scoring functions are used to do these estimations(49). Vina uses an empirical scoring function, which calculates the affinity or fitness of protein-ligand binding by summing up the contributions of number of terms, which represents an important energetic factor in protein-ligand binding. Vina scores the affinity of docked compound in $KCalmol^{-1}$, the smaller the value the greater the affinity of compounds on the receptor (75). **Figure 15** shows a scatter plots of binding energy (BE) of 437 docked natural products and docked GSIs during validation stage. A majority of NPs (90%) had BE ranges between -7.8 and -4.8 $KCalmol^{-1}$. Average BE for NPs was calculated to be -6.5$KCalmol^{-1}$, with NP with ligand ID: conmed131 having the lowest energy; -8.8$KCalmol^{-1}$ and conmed22 having the highest binding energy; -4.2$KCalmol^{-1}$. The average BE of GSIs (-7.7$KCalmol^{-1}$) is lower than that of NPs as expected since GSIs are compounds that were designed and optimized to maximize inhibition of γ-secretase. Figure 4.9A shows the relationship of BE and molecular weight MW. A visible linear correlation (r = 0.76 using Bravais-Pearson) could be observed as BE decreased with

the increase in MW. This bias could also be used to justify why most GSIs had very low binding energies (<-7.5KCalmol^{-1}) since the majority of GSIs have MW>400. The possible reasons for this BE-MW bias is that bigger ligands tend to fill the binding pocket and they interact closely with binding site residues whereas small molecules in this case (MW<300) do not have the ability to contact all residue binding sites or fill up the binding pocket. With the rising interest of fragment-based drug design, there has been a shift to focus on low molecular weight compounds for early lead discovery and optimization as large weight molecules have be considered to be rigid to optimization (132). Calculation of ligand efficiency, is one method used to minimize molecular weight-binding energy bias and prioritizing good binding fragments (133). In this study NPs such as conmed154, nanpdb109 and conmed114 could be classified as promising fragments as they relatively have very low BE (less than -8.0KCalmol^{-1}) and low MW (less than 350). **Figure 15B** and **C** shows scatter plots of binding energy of NPs correlations with solubility (cLogS) and lipophilicity (XlogP) respectively. As shown by the plots BE of NPs is increasing as cLogS increases and decreasing as XlogP is increasing. These results show that lipophilic compounds favored strong interactions, thus implying that the binding site used is very hydrophobic. The docked NPs were also clustered using t-SNE method using eight physicochemical properties computed earlier in this study. Based on physicochemical properties three clustered were generated and the distribution of the BE on clusters was also inspected (**Figure 16**). From the visualization, the three clusters could be categorized by their BE ranges cluster A with BE ranges -9.0 to -7.5 KCalmol^{-1}, cluster B with BE ranges -7.4 to -5.5 KCalmol^{-1} and cluster C with BE ranges -5.4 to -4.5 KCalmol^{-1}. From these results 72 (**Figure 17**) compounds with BE<-7.5 KCalmol^{-1} (average BE of GSIs), in Cluster A were selected from the druglikeNP of docked NPs for further pose and interaction studies with PS1.

5.3.3 Ranking and screening of 72 NPs compounds based on binding interactions in PSI complexes

The orientation of the ligand relative to the receptor and the intermolecular interactions in protein-ligand complexes are of great importance in molecular docking studies (49). From these, mechanisms of action of enzymes or proteins are predicted as well as activation or inhibition of the enzyme. In this studies 72 NPs with BE<-7.5KCalmol^{-1} were further screened based on their interactions with PS1. The purpose of this stage was to identify NPs that interacted with essential amino acids within the binding pocket of PS1. To achieve this 13 GSI-PS1 docked complexes were analyzed to identify key amino acid residues common to most GSI inhibition of γ-secretase. The common amino acids were then used as reference in screening NPs. **Table 11** shows the interacting residues in 13 GSI-PS1 complexes and the types of intermolecular interactions associated with them. 8 amino acids (Trp165, Gly384, Ile143, Leu166, Leu150, Leu286, Asp257 and Leu268) in PS1 were found to interact with at least 4 or more GSIs. Although the grid parameter for the docking experiment was configured, around four hydrophobic residues reported to interact with DAPT inhibitor (Met146, Met233, Trp165, Phe283 and Gly384) it is worth noting that only two of these residues, Trp165 and Gly384 actually interacted with at least four GSIs. Asp257, one of two catalytic aspartate considered crucial in catalytic activity of PSI, interacted with five GSIs. Hydrophobic interactions such as Pi-alkyl, Pi-Pi stacking, Alkyl and Pi-Sigma dominated in 80% of GSI-PS1 complexes and hydrogen, halogen and C-H bond were the only polar bond observable in these complexes. Attempts to demystify binding modes of GSIs in PS1 have been done for the past few years using different experimental methods (128,134,135). These studies have revealed that GSIs bind differently to the PS1(106). Svedružić et al (131), recently conducted a molecular docking and molecular dynamic studies on PS1 using three GSIs; DAPT,

LY-45130, LY411575, and BMS-708164 which revealed other 3 more binding sites of PS1 which could be used in further studies.

A comprehensive analysis of 72 NP-PS1 complexes revealed that NPs interacted diversely with the binding pocket of PS1. Some NPs such as conmed131, conmed151, nanpdb106 and nanpdb36 interacted with more than eight binding pocket residues while NPs such as afrodb17, afrodb15, conmed108 and nanpdb309 interacted with a few residues as little as 2 amino acids. A relationship between the number of interacting residues and BE of NP-PS1 complexes could not be established in this study. In all 72, 24 of them interacted with at least four residues common in GS1-PS1 complexes, shown in **Table 12**. Of the 24, 7 NPs interacted with at least seven amino acid residues. 14 NPs were also identified to interact with at least seven unique residues not common in GSIs. **Figure 19** shows that hydrophobic interactions dominated in NP-PS1 complexes as it could be observed that hydrogen bonds were few. Residues such as Leu282, Leu150, Leu286, Leu268 and Met146 were responsible for forming weak Alkyl interactions with aliphatic portions of NPs. Aromatic residues such as Ph283, Phe288 and Trp165 form strong Pi-Pi and Pi-sigma bonds with aromatic rings of natural products. The final 12 compounds (Table) were selected by considering BE, interacting residues as well as types of interactions formed in NP-PS1 complexes of promising NPs.

5.4 Pharmacokinetic profiling of 12 most promising compounds

ADMET properties of the 12 most promising compounds were assessed using pkCSM server. The following section comprehensively discusses the profiled ADMET parameters computed

5.4.1 Absorption

The absorption profiles of the 12 hits are listed **Table 13.** All the compounds had LogS values within the recommended range of -5.0-0 with a little deviation. The Intestinal absorption in humans (HIA) profiles 10 hits was predicted to be above 80% and while 9/12 compounds were predicted to have high Caco-2 permeability with log P_{app}>0.90. Therefore, most of the compounds identified as potential hits for breast cancer, were predicted to have high oral absorption. These results could be attributed to the database curation steps which removed compounds which disobeyed rules of drug bioavailability as proposed by Lipinski and Veber (53,54).

5.4.2 Distribution

The distribution of a drug refers to the distribution of the compound throughout different compartments within the body (136). The distribution profiles of 12 hits are presented in **Table 14.** Four NPs; nanpdb36, conmed155, conmed134 and conmed64 were predicted to have log VDss < -0.15, thus could be distributed poorly while afrodb12, conmed13, conmed12, nanpdb147 and afrodb14 had log VDss > 0.45, thus it could be concluded that they could be distributed efficiently. The predicted faction of the unbound portions of NP leads was relatively low than 50% for all the 12 compounds. Conmed145 was predicted to have the highest fraction unbound of 0.45 while conmed124 was predicted to attach to the blood plasma protein 100%. Blood brain barrier (BBB) and central nervous system CNS) permeability was predicted as LogBB and LogPS respectively. All NP hits except for conmed145 had low LogBB while all none was predicted to have high CNS permeability. While 5 out of 12 NPs; conmed12, conmed155, nanapdb106, nanpdb147 and nanpdb109 were predicted to be P-glycoprotein (P-gp) inhibitors and the rest were identified as substrates. A larger number of NPs (8/12) were also identified as P-gp I inhibitors yet only 4/12

NPs were predicted to be P-gp II inhibitors. Some NPs (afrodb14, afrodb12, conmed13, conmed145, conmed134 and nanpdb36) were predicted to be gp-substrates and inhibitors at the same time. P-gp proteins have been reported to be multidrug resistance proteins in cancer therapy hence further optimization of NP hits is need to maximize their bioavailability as gp-inhibitors and anti-cancer leads.

3.3.5 Metabolism

Drug metabolism normally involves enzymatic modification or degradation of the compound to facilitate excretion via one of the major clearance organs: liver, kidney, spleen, or bile (73). CYP450 inhibition-based metabolism of 12 NP hits shown in **Table 15** was predicted based on models for five different CYP isoforms (CYP1A2, CYP2C19, CYP2C9, CYP2D6 and CYP3A4).8 NP hits were predicted to inhibit at least one CYP. Conmed124, afrodb14 and conmed155 were identified as consistent CYP inhibitors as they inhibited more than three CYPs. However, one of the most essential, CYP3A4 was only inhibited by three NP hits, conmed124, nanpdb106 and nanpdb109. Conmed64 conmed145, conmed134 and nanpdb36 did not exhibit any inhibition in all five CYP isoforms, can possibly be metabolized perfectly in the liver.

5.4.3 Elimination/Excretion

Drug excretion is pharmacokinetic process that involves the removal of administered drugs or their metabolites from the body mostly carried out by the liver and the kidneys and depends on physicochemical properties of drugs such as molecular weight, lipophilicity and polar surface area (136). **Table 16** summaries elimination profiles 12 identified NPs. Predicted clearance of all 12 NPs was between 0.03 and 0.99. Conmed124 and afrodb12 had the highest clearance of 0.99 while

the clearance of conmed134 had the least clearance of 0.03. While, among the 12 NPs three compounds were predicted as OCT2 substrates (conmed124, nanpdb106 and nanpdb109) while the rest are considered to excrete through the kidneys in another mechanism besides OCT2.

5.4.4 Toxicity

In this study, hepatotoxicity, hERG inhibition, and compound mutageneric (AMES toxicity) models were used to predict the toxicity of NP hits. Results for this analysis are presented in **Table 17.** Only 2/12 compounds were predicted to be hepatotoxic thus could induce liver injury (121). None of the NP compounds test positive for Ames toxicity or mutagenicity. In all 12 compounds, no hERG (human ether-a-go-go gene) I inhibition was predicted while all compounds except of conmed155, conmed106 and nanpdb109 were predicted as hERG II inhibitors.

Pharmacokinetic studies of these potential gamma secretase inhibitors warrant further optimization, to increase the distribution profiles of the hits and remove functional groups responsible of predicted potential toxicity such as CYP inhibition, hERG inhibition and hepatoxicity.

6. CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

Breast cancer is now the most occurring cancer and is responsible for more than 600 000 deaths globally as of 2020. The need to design more chemotherapeutic drugs grows daily, as the currently used breast cancer drugs are nonspecific, associated with adverse toxicity and are multi-drug resistance. Proposing natural products that target breast cancer drug targets such as γ-secretase is a promising approach in effort to develop officious medication against breast cancer. In this study, drug-like natural products from four African natural products databases were prioritized and chemoinformatically profiled to examine their diversity and property distribution relative to FDA approved drugs. The natural products were also virtually screened against γ-secretase enzyme and the final hit compounds were pharmacokinetically profiled. Drug-like natural products from African natural products libraries showed that they have exhibit lead-like character based on physicochemical properties and are very diverse in terms of scaffolds as compared to drug-like FDA approved drugs. In structure-based virtual screening studies, 16% of NPs had binding energy comparable to known γ-secretase while only 5% interacted effectively with γ-secretase catalytic unit PS1 active site residues. From this study, it can be concluded that 12 compounds; afrodb12, afrodb14, conmedb131, conmedb124, conmedb64, conmedb145, conmedb134, conmed155, nanpdb109, nanpdb36, nanpdb147 and nanpdb106 can be proposed as potential inhibitors of γ-secretase. *In silico* ADMET studies predicted that the majority of compounds have good human intestinal absorption, low blood-brain permeability and low hepatotoxicity profiles. However, a considerable number of the hit compounds were predicted to inhibit CYP450 isoforms and hERG

channels. This study provides a good starting point for wet biology assay, especially in target based approach drug discovery of breast cancer drugs.

6.2 Recommendation and future studies

This work provides a good starting point in application of chemoinformatic approaches in natural products research in Africa for advanced drug development purposes. Hence, from this research it is recommended that medicinal scientists should further examine the diversity and the pharmacological relevance of natural product compounds characterized locally and abroad using *in silico* approaches to enhance compound prioritization for synthesis and screening. There is also a need to develop natural products libraries for several African countries (Malawi, Zimbabwe, Zambia, Namibia and many more) currently not covered by currently available databases in order to make available all elucidated natural product compounds to the research community and advance drug discovery in Africa. There is also a need to use molecular dynamics and free energy calculation for the presented drug-ligands complexes, in effort to validate the stability of the protein-ligand complexes generated in this study. *In silico* experiments such as the ones used in this study have been proved more reliable when used in combination with *in vivo* and *in vitro* studies, hence it is it is necessary to use breast cancer *in vitro* or *in vivo* models test the hit compounds identified in this research. Although this study identified potential compounds as *γ-secretase* inhibitors. From the pharmacokinetic property predictions of the compounds, more work on structure activity-based optimization is needed to enhance the potency and the efficacy of the cancer drug hits. Future studies for this work can also include the application of ligand-based virtual screening approaches such as QSAR and pharmacophore modelling to screening all available African natural products databases for enhanced breast cancer drug discovery.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. WHO (World Health Organization). Breast cancer [Internet]. 2021 [cited 2021 Apr 26]. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer
3. Joko-Fru WY, Jedy-Agba E, Korir A, Ogunbiyi O, Dzamalala CP, Chokunonga E, et al. The evolving epidemic of breast cancer in sub-Saharan Africa: results from the African Cancer Registry Network. Int J Cancer. 2020;147(8):2131–41.
4. Waks AG, Winer EP. Breast cancer treatment: a review. JAMA. 2019 Jan 22;321(3):288–300.
5. Leclerc A-F, Jerusalem G, Devos M, Crielaard J-M, Maquet D. Multidisciplinary management of breast cancer. Arch Public Health. 2016 Dec 5;74(1):50.
6. Claessens AKM, Ibragimova KIE, Geurts SME, Bos MEMM, Erdkamp FLG, Tjan-Heijnen VCG. The role of chemotherapy in treatment of advanced breast cancer: an overview for clinical practice. Crit Rev Oncol Hematol. 2020 Sep 1;153:102988.
7. Acar A, Simões BM, Clarke RB, Brennan K. A role for Notch signalling in breast cancer and endocrine resistance. Stem Cells Int. 2016;2016:1–6.
8. Mangal M, Sagar P, Singh H, Raghava GPS, Agarwal SM. NPACT: Naturally occurring plant-based anti-cancer compound-activity-target database. Nucleic Acids Res. 2013 Jan 1;41(D1):D1124–9.

9. Sorokina M, Steinbeck C. Review on natural products databases: where to find data in 2020. J Cheminformatics. 2020 Dec;12(1):20.

10. Egieyeh S, Malan SF, Christoffels A. Cheminformatics techniques in antimalarial drug discovery and development from natural products 1: basic concepts. Phys Sci Rev [Internet]. 2019 Jun 26 [cited 2021 Apr 20];4(7). Available from: https://www.degruyter.com/document/doi/10.1515/psr-2018-0130/html

11. Chávez-Hernández AL, Sánchez-Cruz N, Medina-Franco JL. Fragment library of natural products and compound databases for drug discovery. Biomolecules. 2020 Nov 6;10(11):1518.

12. Kar S, Leszczynski J. Open access in silico tools to predict the ADMET profiling of drug candidates. Expert Opin Drug Discov. 2020 Dec 1;15(12):1473–87.

13. McCormack V, McKenzie F, Foerster M, Zietsman A, Galukande M, Adisa C, et al. Breast cancer survival and survival gap apportionment in sub-Saharan Africa (ABC-DO): a prospective cohort study. Lancet Glob Health. 2020 Sep 1;8(9):e1203–12.

14. Pine SR. Rethinking Gamma-secretase Inhibitors for treatment of non–small-cell lung cancer: is notch the target? Clin Cancer Res. 2018 Dec 15;24(24):6136–41.

15. Batool M, Ahmad B, Choi S. A structure-based drug discovery paradigm. Int J Mol Sci. 2019;18.

16. Momenimovahed Z, Salehiniya H. Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer Targets Ther. 2019 Apr 10;11:151–64.

17. Nazmeen A, Chen G, Ghosh TK, Maiti S. Breast cancer pathogenesis is linked to the intra-tumoral estrogen sulfotransferase (hSULT1E1) expressions regulated by cellular redox dependent Nrf-2/NFκβ interplay. Cancer Cell Int. 2020 Mar 4;20(1):70.

18. Feng Y, Spezia M, Huang S, Yuan C, Zeng Z, Zhang L, et al. Breast cancer development and progression: risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. Genes Dis. 2018 Jun 1;5(2):77–106.

19. Shah R, Rosso K, Nathanson SD. Pathogenesis, prevention, diagnosis and treatment of breast cancer. World J Clin Oncol. 2014 Aug 10;5(3):283–98.

20. Harbeck N, Penault-Llorca F, Cortes J, Gnant M, Houssami N, Poortmans P, et al. Breast cancer. Nat Rev Dis Primer. 2019 Sep 23;5(1):66.

21. Abdulkareem IH. Aetio-pathogenesis of breast cancer. Niger Med J J Niger Med Assoc. 2013;54(6):371–5.

22. Bean M. 20 cancer drugs approved in 2020 [Internet]. 2020 [cited 2021 Jul 12]. Available from: https://www.beckershospitalreview.com/oncology/20-cancer-drugs-approved-in-2020.html

23. Venkatesh V, Nataraj R, Thangaraj GS, Karthikeyan M, Gnanasekaran A, Kaginelli SB, et al. Targeting Notch signalling pathway of cancer stem cells. Stem Cell Investig. 2018;5:5.

24. Zhong Y, Shen S, Zhou Y, Mao F, Lin Y, Guan J, et al. NOTCH1 is a poor prognostic factor for breast cancer and is associated with breast cancer stem cells. OncoTargets Ther. 2016 Nov; 9:6865–71.

25. Edwards A, Brennan K. Notch signalling in breast development and cancer. Front Cell Dev Biol [Internet]. 2021 [cited 2021 Jul 14];9. Available from: https://www.frontiersin.org/articles/10.3389/fcell.2021.692173/full

26. Johnson DS, Li Y-M, Pettersson M, St George-Hyslop PH. Structural and chemical biology of presenilin complexes. Cold Spring Harb Perspect Med. 2017 Dec;7(12):a024067.

27. McCaw TR, Inga E, Chen H, Jaskula-Sztul R, Dudeja V, Bibb JA, et al. Gamma secretase inhibitors in cancer: a current perspective on clinical performance. The Oncologist. 2021 Apr;26(4):e608–21.

28. Ghanbari-Movahed M, Ghanbari-Movahed Z, Momtaz S, Kilpatrick KL, Farzaei MH, Bishayee A. Unlocking the secrets of cancer stem cells with γ-secretase inhibitors: a novel anticancer strategy. Molecules. 2021 Feb 12;26(4):972.

29. Han J, Shen Q. Targeting γ-secretase in breast cancer. Breast Cancer Targets Ther. 2012 Jun 21;4:83–90.

30. Xie T, Song S, Li S, Ouyang L, Xia L, Huang J. Review of natural product databases. Cell Prolif. 2015 Aug;48(4):398–404.

31. Medina-Franco JL, Saldívar-González FI. Cheminformatics to characterize pharmacologically active natural products. 2020;14.

32. Thomford N, Senthebane D, Rowe A, Munro D, Seele P, Maroyi A, et al. Natural products for drug discovery in the 21st Century: innovations for novel drug discovery. Int J Mol Sci. 2018 May 25;19(6):1578.

33. Ntie-Kang F, Onguéné PA, Scharfe M, Owono Owono LC, Megnassan E, Mbaze LM, et al. ConMedNP: a natural product library from Central African medicinal plants for drug discovery. RSC Adv. 2014;4(1):409–19.

34. The International Natural Product Sciences Taskforce, Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov. 2021 Mar;20(3):200–16.

35. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: Collection of Open Natural Products database. J Cheminformatics. 2021 Dec;13(1):2.

36. Ntie-Kang F, Telukunta KK, Döring K, Simoben CV, A. Moumbock AF, Malange YI, et al. NANPDB: A Resource for natural products from northern African sources. J Nat Prod. 2017 Jul 28;80(7):2067–76.

37. Simoben CV, Qaseem A, Moumbock AFA, Telukunta KK, Günther S, Sippl W, et al. Pharmacoinformatic investigation of medicinal plants from East Africa. Mol Inform. 2020 Nov;39(11):2000163.

38. Hatherley R, Brown DK, Musyoka TM, Penkler DL, Faya N, Lobb KA, et al. SANCDB: a South African natural compound database. J Cheminformatics. 2015 Dec;7(1):29.

39. Diallo BN, Glenister M, Musyoka TM, Lobb K, Tastan Bishop Ö. SANCDB: an update on South African natural compounds and their readily available analogs. J Cheminformatics. 2021 May 5;13(1):37.

40. Ntie-Kang F, Amoa Onguéné P, Fotso GW, Andrae-Marobela K, Bezabih M, Ndom JC, et al. Virtualizing the p-ANAPL Library: a step towards drug discovery from African medicinal plants. Lu J, editor. PLoS ONE. 2014 Mar 5;9(3):e90655.

41. Ntie-Kang F, Zofou D, Babiaka SB, Meudom R, Scharfe M, Lifongo LL, et al. AfroDb: a select highly potent and diverse natural product library from African medicinal plants. Barchi JJ, editor. PLoS ONE. 2013 Oct 30;8(10):e78085.

42. Ntie-Kang F, Nwodo JN, Ibezim A, Simoben CV, Karaman B, Ngwa VF, et al. Molecular modeling of potential anticancer agents from African medicinal plants. J Chem Inf Model. 2014 Sep 22;54(9):2433–50.

43. Ntie-Kang F, Mbah JA, Mbaze LM, Lifongo LL, Scharfe M, Hanna JN, et al. CamMedNP: building the Cameroonian 3D structural natural products database for virtual screening. BMC Complement Altern Med. 2013 Dec;13(1):88.

44. Onguéné PA, Ntie-Kang F, Mbah JA, Lifongo LL, Ndom JC, Sippl W, et al. The potential of anti-malarial compounds derived from African medicinal plants, part III: an in silico evaluation of drug metabolism and pharmacokinetics profiling. Org Med Chem Lett. 2014 Dec;4(1):6.

45. Ibezim A, Debnath B, Ntie-Kang F, Mbah CJ, Nwodo NJ. Binding of anti-Trypanosoma natural products from African flora against selected drug targets: a docking study. Med Chem Res. 2017 Mar;26(3):562–79.

46. Bultum LE, Woyessa AM, Lee D. ETM-DB: integrated Ethiopian traditional herbal medicine and phytochemicals database. BMC Complement Altern Med. 2019 Dec;19(1):212.

47. Pereira F, Aires-de-Sousa J. Computational methodologies in the exploration of marine natural product leads. Mar Drugs. 2018 Jul 13;16(7):236.

48. Njogu PM, Guantai EM, Pavadai E, Chibale K. Computer-aided drug discovery approaches against the Tropical Infectious Diseases Malaria, Tuberculosis, Trypanosomiasis, and Leishmaniasis. ACS Infect Dis. 2016 Jan 8;2(1):8–31.

49. Meng X-Y, Zhang H-X, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. Current Computer-Aided Drug Design. 2012;7(2):146-152.

50. Yu W, MacKerell AD. Computer-aided drug design methods. Methods Mol Biol. 2018;24(2): 85–106

51. Gajipara J, Georrge JJ. Tools for ligand based drug discovery. In Recent Trends in Science and Technology-2018 (pp. 57–64). Christ Publications.

52. Kaserer T, Schuster D, Rollinger JM. Chemoinformatics in natural product research. In: Engel T, Gasteiger J, editors. Applied chemoinformatics [Internet]. Weinheim, Germany:

Wiley-VCH Verlag GmbH & Co. KGaA; 2018 [cited 2021 Apr 9]. p. 207–36. Available from: http://doi.wiley.com/10.1002/9783527806539.ch6c

53. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001 Mar 1;46(1–3):3–26.

54. Veber D, Johnson S, Cheng H, Smith B, Ward K, Kopple K. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem [Internet]. 2002 Jun 6 [cited 2021 Jun 20];45(12). Available from: https://pubmed.ncbi.nlm.nih.gov/12036371/

55. Moreno L, Pearson ADJ. How can attrition rates be reduced in cancer drug discovery? Expert Opin Drug Discov. 2013 Apr;8(4):363–8.

56. Kadri HS, Minocheherhomji FP. ADMET analysis of phyto-components of Syzygium cumini seeds and Allium cepa peels. Future J Pharm Sci. 2020 Dec 9;6(1):117.

57. Alqahtani S. In silico ADME-Tox modeling: progress and prospects. Expert Opin Drug Metab Toxicol. 2017 Nov;13(11).

58. Kazmi SR, Jun R, Yu M-S, Jung C, Na D. In silico approaches and tools for the prediction of drug metabolism and fate: a review. Comput Biol Med. 2019 Mar 1;106:54–64.

59. Durán-Iturbide NA, Díaz-Eufracio BI, Medina-Franco JL. In Silico ADME/Tox profiling of natural products: a focus on BIOFACQUIM. 2020;12.

60. Onguéné PA, Simoben CV, Fotso GW, Andrae-Marobela K, Khalid SA, Ngadjui BT, et al. In silico toxicity profiling of natural product compound libraries from African flora with anti-malarial and anti-HIV properties. Comput Biol Chem. 2018 Feb;72:136–49.

61. Begam BF, Kumar JS. A study on cheminformatics and its applications on modern drug discovery. Procedia Eng. 2012;38:1264–75.

62. Bento AP. An open source chemical structure curation pipeline using RDKit. J Cheminform. 2020;12(1):51.

63. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen E. Recent developments of the chemistry Adevelopment Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. Curr Pharm Des. 2006 Jun 1;12(17):2111–20.

64. Lakhlili W, Yasri A, Ibrahimi A. Structure-activity relationships study of mTOR kinase inhibition using QSAR and structure-based drug design approaches. Onco Targets Ther. 2016 Dec; 9:7345–53.

65. Tao L, Zhu F, Qin C, Zhang C, Chen S, Zhang P, et al. Clustered distribution of natural product leads of drugs in the chemical space as influenced by the privileged target-sites. Sci Rep. 2015 Aug;5(1):9325.

66. Sander T, Freyss J, von Korff M, Rufener C. DataWarrior: An open-source program for chemistry aware data visualization and analysis. J Chem Inf Model. 2015 Feb 23;55(2):460–73.

67. Alajlani MM, Backlund A. Evaluating antimycobacterial screening schemes using chemical global positioning system-natural product analysis. Molecules. 2020 Feb 20;25(4):945.

68. Awale M, Probst D, Reymond J-L. WebMolCS: A Web-Based Interface for Visualizing Molecules in Three-Dimensional Chemical Spaces. J Chem Inf Model. 2017 Apr 24;57(4):643–9.

69. González-Medina M, Prieto-Martínez FD, Owen JR, Medina-Franco JL. Consensus diversity plots: a global diversity analysis of chemical libraries. J Cheminformatics. 2016 Dec;8(1):63.

70. Pires DEV, Blundell TL, Ascher DB. pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem. 2015 May 14;58(9):4066–72.

71. Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep. 2017 May;7(1):42717.

72. Ioakimidis L, Thoukydidis L, Mirza A, Naeem S, Reynisson J. Benchmarking the reliability of QikProp. correlation between experimental and predicted values. QSAR Comb Sci. 2008 Apr;27(4):445–56.

73. Dong J. ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. 2018;11.

74. Hou X, Du J, Zhang J, Du L, Fang H, Li M. How to improve docking accuracy of AutoDock4.2: a case study Using Different Electrostatic Potentials. J Chem Inf Model. 2013 Jan 28;53(1):188–200.

75. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2009;NA-NA.

76. Singh KD, Muthusamy K. Molecular modeling, quantum polarized ligand docking and structure-based 3D-QSAR analysis of the imidazole series as dual AT1 and ETA receptor antagonists. Acta Pharmacol Sin. 2013 Dec;34(12):1592–606.

77. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins Struct Funct Bioinforma. 2003 Aug 1;52(4):609–23.

78. Zhou Y, Tang S, Chen T, Niu M-M. Structure-based pharmacophore modeling, virtual screening, molecular docking and biological evaluation for Identification of Potential Poly (ADP-Ribose) Polymerase-1 (PARP-1) Inhibitors. Molecules. 2019 Nov 22;24(23):4258.

79. Dixon SL, Smondyrev AM, Rao SN. PHASE: a novel approach to pharmacophore modeling and 3D database searching. Chem Biol Htmlent Glyphamp Asciiamp Drug Des. 2006 May;67(5):370–2.

80. Koes DR, Camacho CJ. Pharmer: efficient and exact pharmacophore search. J Chem Inf Model. 2011 Jun 27;51(6):1307–14.

81. Koes DR, Camacho CJ. ZINCPharmer: pharmacophore search of the ZINC database. Nucleic Acids Res. 2012 May 2;40:6.

82. Riaz N, Shahbaz A, kalsoom S. Ligand based pharmacophore model development for the identification of novel anti-psychotic drugs. Int J Appl Sci - Res Rev [Internet]. 2018 [cited 2021 Apr 14];05(02). Available from: http://www.imedpub.com/articles/ligand-based-pharmacophore-model-development-for-the-identification-of-novel-antipsychotic-drugs.php?aid=22775

83. Saldívar-González FI, Valli M, Andricopulo AD, da Silva Bolzani V, Medina-Franco JL. Chemical space and diversity of the NuBBE database: a chemoinformatic characterization. J Chem Inf Model. 2019 Jan 28;59(1):74–85.

84. Huang M, Lu J-J, Ding J. Natural products in cancer therapy: past, present and future. Nat Prod Bioprospecting. 2021 Feb 1;11(1):5–13.

85. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: The Konstanz information miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, editors. Data

analysis, machine learning and applications. Berlin, Heidelberg: Springer; 2008. p. 319–26. (Studies in Classification, Data Analysis, and Knowledge Organization).

86. Schneider G. Prediction of Drug-Like Properties [Internet]. Madame Curie Bioscience Database [Internet]. Landes Bioscience; 2013 [cited 2021 Jun 20]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK6404/

87. Brenk R, Schipani A, James D, Krasowski A, Gilbert IH, Frearson J, et al. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. ChemMedChem. 2008 Mar;3(3):435–44.

88. Dahlin JL, Nissink JWM, Strasser JM, Francis S, Higgins L, Zhou H, et al. PAINS in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS. J Med Chem. 2015 Mar 12;58(5):2091–113.

89. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem. 2012 Jan 24;4(2):90–8.

90. Wetzel S, Schuffenhauer A, Roggo S, Ertl P, Waldmann H. Cheminformatic analysis of natural products and their chemical space. Chim Int J Chem. 2007 Jun 27;61(6):355–60.

91. Janssen APA, Grimm SH, Wijdeven RHM, Lenselink EB, Neefjes J, van Boeckel CAA, et al. Drug discovery maps, a machine learning model that visualizes and predicts Kinome–inhibitor interaction landscapes. J Chem Inf Model. 2019 Mar 25;59(3):1221–9.

92. Méndez-Lucio O, Medina-Franco JL. The many roles of molecular complexity in drug discovery. Drug Discov Today. 2017 Jan;22(1):120–6.

93. Korff M von, Sander T. Molecular complexity calculated by fractal dimension. Sci Rep. 2019;

94. González-Medina M, Medina-Franco JL. Platform for unified molecular analysis: PUMA. J Chem Inf Model. 2017 Aug 28;57(8):1735–40.

95. González-Medina M, Medina-Franco JL. Chemical diversity of cyanobacterial compounds: A chemoinformatics analysis. ACS Omega. 2019 Apr 30;4(4):6229–37.

96. Ferreira LT, Borba JVB, Moreira-Filho JT, Rimoldi A, Andrade CH, Costa FTM. QSAR-based virtual screening of natural products database for identification of potent antimalarial hits. Biomorecules. 2021;11(3):459.

97. Saldívar-González FI, Medina-Franco JL. Chemoinformatics approaches to assess chemical diversity and complexity of small molecules. In: Small molecule drug discovery [Internet]. Elsevier; 2020 [cited 2021 Apr 9]. p. 83–102. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128183496000030

98. Ertl P, Schuhmann T. Cheminformatics analysis of natural product scaffolds: comparison of scaffolds produced by animals, plants, fungi and bacteria. Mol Inform. 2020 Nov;39(11):2000017.

99. Yongye AB, Waddell J, Medina-Franco JL. Molecular scaffold analysis of natural products databases in the public domain: scaffold analysis of natural products databases. Chem Biol Drug Des. 2012 Nov;80(5):717–24.

100. González-Medina M, Prieto-Martínez FD, Naveja JJ, Méndez-Lucio O, El-Elimat T, Pearce CJ, et al. Chemoinformatic expedition of the chemical space of fungal products. Future Med Chem. 2016 Aug;8(12):1399–412.

101. A. Olmedo D, L. Medina-Franco J. Chemoinformatic Approach: The case of natural products of Panama. In: Stefaniu A, Rasul A, Hussain G, editors. Cheminformatics and its applications [Internet]. IntechOpen; 2020 [cited 2021 Apr 9]. Available from:

https://www.intechopen.com/books/cheminformatics-and-its-applications/chemoinformatic-approach-the-case-of-natural-products-of-panama

102. Basi GS, Hemphill S, Brigham EF, Liao A, Aubele DL, Baker J, et al. Amyloid precursor protein selective gamma-secretase inhibitors for treatment of Alzheimer's disease. Alzheimers Res Ther. 2010 Dec 29;2(6):36.

103. Wang Y, Yu S, Huang D, Cui M, Hu H, Zhang L, et al. Cellular prion protein mediates pancreatic cancer cell survival and invasion through association with and enhanced signaling of Notch1. Am J Pathol. 2016 Nov;186(11):2945–56.

104. Borgegård T, Gustavsson S, Nilsson C, Parpal S, Klintenberg R, Berg A-L, et al. Alzheimer's disease: presenilin 2-sparing γ-secretase inhibition is a tolerable Aβ peptide-lowering strategy. J Neurosci. 2012 Nov 28;32(48):17297–305.

105. Chávez-Gutiérrez L, Bammens L, Benilova I, Vandersteen A, Benurwar M, Borgers M, et al. The mechanism of γ-Secretase dysfunction in familial Alzheimer disease. EMBO J. 2012 May 16;31(10):2261–74.

106. Ran Y, Hossain F, Pannuti A, Lessard CB, Ladd GZ, Jung JI, et al. γ-secretase inhibitors in cancer clinical trials are pharmacologically and functionally distinct. EMBO Mol Med. 2017 Jul;9(7):950–66.

107. Luistro L, He W, Smith M, Packman K, Vilenchik M, Carvajal D, et al. Preclinical profile of a potent gamma-secretase inhibitor targeting notch signaling with in vivo efficacy and pharmacodynamic properties. Cancer Res. 2009 Oct 1;69(19):7672–80.

108. Pant S, Jones SF, Kurkjian CD, Infante JR, Moore KN, Burris HA, et al. A first-in-human phase I study of the oral Notch inhibitor, LY900009, in patients with advanced cancer. Eur J Cancer. 2016 Mar 1;56:1–9.

109. Dovey HF, John V, Anderson JP, Chen LZ, de Saint Andrieu P, Fang LY, et al. Functional gamma-secretase inhibitors reduce beta-amyloid peptide levels in brain. J Neurochem. 2001 Jan;76(1):173–81.

110. Lanz TA, Wood KM, Richter KEG, Nolan CE, Becker SL, Pozdnyakov N, et al. Pharmacodynamics and pharmacokinetics of the gamma-secretase inhibitor PF-3084014. J Pharmacol Exp Ther. 2010 Jul;334(1):269–77.

111. Borthakur G, Martinelli G, Raffoux E, Chevallier P, Chromik J, Lithio A, et al. Phase 1 study to evaluate Crenigacestat (LY3039478) in combination with dexamethasone in patients with T-cell acute lymphoblastic leukemia and lymphoma. Cancer. 2021;127(3):372–80.

112. Cook N. A phase I trial of the γ-secretase inhibitor MK-0752 in combination with gemcitabine in patients with pancreatic ductal adenocarcinoma. Br J Cancer. 2018; 118(6):793-801.

113. Jämsä A, Belda O, Edlund M, Lindström E. BACE-1 inhibition prevents the γ-secretase inhibitor evoked Aβ rise in human neuroblastoma SH-SY5Y cells. J Biomed Sci. 2011 Oct 21;18:76.

114. Nasser F, Moussa N, Helmy MW, Haroun M. Dual targeting of Notch and Wnt/β-catenin pathways: potential approach in triple-negative breast cancer treatment. Naunyn Schmiedebergs Arch Pharmacol. 2021 Mar;394(3):481–90.

115. Doveston RG, Tosatti P, Dow M, Foley DJ, Li HY, Campbell AJ, et al. A unified lead-oriented synthesis of over fifty molecular scaffolds. Org Biomol Chem. 2014 Dec 17;13(3):859–65.

116. Kadam RU, Roy N. Recent trends in drug-likeness prediction: a comprehensive review of In silico methods. Indian J Pharm Sci. 2007;69(5):609.

117. Pajouhesh H, Lenz GR. Medicinal chemical properties of successful entral nervous system drugs. NeuroRx. 2005 Oct;2(4):541–53.

118. Flores-Sumoza M, Alcázar JJ, Márquez E, Mora JR, Lezama J, Puello E. Classical QSAR and docking simulation of 4-Pyridone derivatives for their antimalarial activity. molecules. 2018 Dec;23(12):3166.

119. Ak G, Vn V, Jj W. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. J Comb Chem [Internet]. 1999 Jan [cited 2021 Jun 20];1(1). Available from: https://pubmed.ncbi.nlm.nih.gov/10746014/

120. Congreve M, Carr R, Murray C, Jhoti H. A 'Rule of Three' for fragment-based lead discovery? Drug Discov Today. 2003 Oct 1;8(19):876–7.

121. Fatima S, Gupta P, Sharma S, Sharma A, Agarwal SM. ADMET profiling of geographically diverse phytochemical using chemoinformatic tools. Future Med Chem. 2020;20.

122. Wei W, Cherukupalli S, Jing L, Liu X, Zhan P. Fsp3: a new parameter for drug-likeness. Drug Discov Today. 2020 Jul 24;S1359-6446(20)30297-X.

123. You W, Huang YM, Kizhake S, Natarajan A, Chang CA. Characterization of promiscuous binding of phosphor ligands to Breast-Cancer-Gene 1 (BRCA1) C-Terminal (BRCT): molecular dynamics, free energy, entropy and inhibitor design. PLOS Comput Biol. 2016 Aug 25;12(8):e1005057.

124. Al Sharie AH, El-Elimat T, Al Zu'bi YO, Aleshawi AJ, Medina-Franco JL. Chemical space and diversity of seaweed metabolite database (SWMD): a cheminformatics study. J Mol Graph Model. 2020 Nov;100:107702.

125. Olmedo DA. Cheminformatic characterization of natural products from Panama. Mol Divers. 2017 Jul 8;11.

126. Wang Z, Sun H, Yao X, Li D, Xu L, Li Y, et al. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. Phys Chem Chem Phys. 2016 May 4;18(18):12964–75.

127. Castro-Alvarez A, Costa AM, Vilarrasa J. The performance of several docking programs at reproducing protein-macrolide-like crystal structures. Mol Basel Switz. 2017 Jan 17;22(1):E136.

128. Bai X, Rajendra E, Yang G, Shi Y, Scheres S. Sampling the conformational space of the catalytic subunit of human γ-secretase. eLife. 2015;

129. Hevener KE, Zhao W, Ball DM, Babaoglu K, Qi J, White SW, et al. Validation of molecular docking programs for virtual screening against Dihydropteroate Synthase. J Chem Inf Model. 2009 Feb;49(2):444–60.

130. Mishra N, Basu A. Exploring different virtual screening strategies for Acetylcholinesterase inhibitors. BioMed Res Int. 2013 Nov 4;2013:e236850.

131. Svedruzic ZM, Vrbnjak K, Martinovic M, Miletic V. Structural analysis of the simultaneous activation and inhibition of gamma-Secretase activity in the development of drugs for Alzheimer's disease. Pharmaceutics [Internet]. 2021 Apr [cited 2021 Jul 1];13(4). Available from: https://lirias.kuleuven.be/3449916

132. de Souza Neto LR, Moreira-Filho JT, Neves BJ, Maidana RLBR, Guimarães ACR, Furnham N, et al. In silico strategies to support fragment-to-lead optimization in drug discovery. Front Chem [Internet]. 2020 [cited 2021 Jun 29];8. Available from: https://www.frontiersin.org/articles/10.3389/fchem.2020.00093/full

133. Kenny PW. The nature of ligand efficiency. J Cheminformatics. 2019 Jan 31;11(1):8.

134. Gertsik N, am Ende CW, Geoghegan KF, Nguyen C, Mukherjee P, Mente S, et al. Mapping the binding site of BMS-708163 on γ-secretase with cleavable photoprobes. Cell Chem Biol. 2017 Jan 19;24(1):3–8.

135. Pozdnyakov N, Murrey HE, Crump CJ, Pettersson M, Ballard TE, Ende CW am, et al. γ-Secretase Modulator (GSM) photoaffinity probes reveal distinct allosteric binding sites on presenilin*. J Biol Chem. 2013 Apr 5;288(14):9710–20.

136. Lagorce D, Douguet D, Miteva MA, Villoutreix BO. Computational analysis of calculated physicochemical and ADMET properties of protein-protein interaction inhibitors. Sci Rep. 2017 Apr 11;7(1):46277.

APPENDICES

Appendix 1: COMREC ethical clearance

CERTIFICATE OF ETHICS APPROVAL

This is to certify that the College of Medicine Research and Ethics Committee (COMREC) has reviewed and approved a study entitled:

P.05/20/3064 - Target-based identification and biological evaluations of novel anticancer natural compounds for enhanced breast cancer chemotherapy: a pilot study for the ethical clearance. by Jonathan Bvunzawabaya

On 09-Jul-20

As you proceed with the implementation of your study, we would like you to adhere to international ethical guidelines, national guidelines and all requirements by COMREC some of which are indicated on the next page for your study

09-Jul-20

Prof. I. Umar -Chairperson (COMREC)

Date

Approved by College of Medicine

11-Jul-2020

(COMREC) Research and Ethics Committee

Appendix 2: Poster presentation

Chemical space and Chemical diversity profiling of drug-like compounds from African natural product databases for Fragment based-drug discovery

Jonathan T. Bvanzawabaya[1,2]*, Andrew G. Mtewa[3], Fanuel Lampiao[1,2], Grace C. Mugumbate[4]

[1]Department of Biomedical Sciences, Kamuzu University of Health Sciences, P/Bag 301, Blantyre, Malawi [2]Africa Center of Excellence in Public Health and Herbal Medicine (ACEPHEM), Kamuzu University of Health Sciences, P/Bag 301, Blantyre, Malawi [3]Chemistry Section, Department of Applied Studies, Malawi Institute of Technology, Malawi University of Science and Technology, P. O. Box 5196, Limbe, Malawi [4] Department of Chemical Technology, Midlands State University, P/Bag 9055, Senga Road, Gweru, Zimbabwe. Corresponding author: bvunzanyari@gmail.com

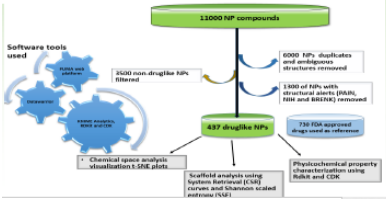
Introduction

- Natural products (NPs) continue to play a key role in development of drugs for tropical diseases and cancer in Africa.
- Over 10 African NP databases with more than 10 000 compounds have been published and re-used since 2010.
- Currently most of the databases have never been chemically characterized for medicinal purposes
- Characterizing the chemical diversity of drug-like NPs in Africa can assist in compound prioritization for early lead screening against tropical diseases and cancer.

Objectives

- The aim of this study was to characterize the chemical diversity, chemical space and physicochemical properties of drug-like NPs in four African natural product databases (NANPDB, Afrocancer, ConmedNP & Afrodb) relative to 730 FDA approved drugs

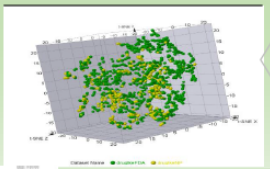
Methodology



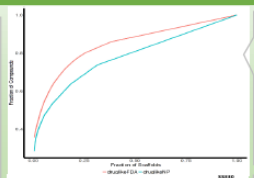
Results & Discussion

descriptor	Min	mean	Median	Max	SD
HBA count	1	2.66	3.0	8	1.67
HBD count	1	2.18	2.0	5	1.17
TPSA	12.0	75.06	73.8	149.8	26.8
MW	193.1	291.1	295.1	512.1	76.6
LogS	-6.42	-2.76	-2.01	1.25	1.35
Molar refractivity(MR)	1.01	7.62	7.61	13.6	2.04
XlogP	1.00	1.85	1.67	4.89	1.36
RB count	0	2.55	2	8	1.91

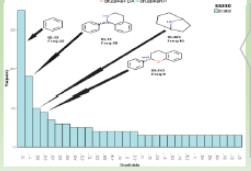
- >50% of druglike NPs had physicochemical properties that satisfied lead-like rules (RO3).
- 88% NPs had aqueous solubility range above the recommend (>-5.7)



- Druglike NPs occupied the same chemical space with FDA drugs.
- Some NPs had high property similarity with FDA drugs



- CSR curves show that AUC for druglikeNP was lower (0.78) than that of druglikeFDA (0.84).
- NPs contained a greater fraction of singletons than FDA drugs



- SSE of 30 most frequent scaffolds in NPs
- Only 25 scaffolds were identified to be common between NPs and FDA drugs datasets.

Conclusion

- Druglike NPs in African natural product databases have similar pharmacological chemical properties with FDA approved drugs(druglike).
- African natural products databases contain many druglike NPs that can be classified as fragments or lead-like.
- Druglike NPs in African NP databases are more diverse in term of molecular scaffolds than druglike FDA approved drugs.
- Therefore African NP databases are a promising source for structurally diverse scaffolds for fragment-based drug discovery.

References

Congreve M, Carr R, Murray C, Jhoti H. A 'Rule of Three' for fragment-based lead discovery? Drug Discov Today. 2003 Oct 1;8(19):876–7
Ntie-Kang F, Telukunta KK, et al. NANPDB: A Resource for Natural Products from Northern African Sources. J Nat Prod. 2017 Jul 28;80(7):2067–76.
Sorokina M, Steinbeck C. Review on natural products databases: where to find data in 2020. J Cheminformatics. 2020 Dec;12(1):20

Appendix 3: manuscript 1

IDENTIFICATION OF NATURAL INHIBITORS FOR γ -SECRETASE ENZYME FROM AFRICAN NATURAL PRODUCT DATABASES. AN *IN SILICO* APPROACH IN COMBATING MULTI-DRUG RESISTANCE IN BREAST CANCER CHEMOTHERAPY

Jonathan T. Bvanzawabaya [1, 2] *, Andrew G. Mtewa[3], Grace C. Mugumbate[4*], Fanuel Lampiao[1, 3]

1Department of Biomedical Sciences, Kamuzu University of Health Sciences, P/Bag 301, Blantyre, Malawi

2Africa Center of Excellence in Public Health and Herbal Medicine (ACEPHEM), Kamuzu University of Health Sciences, P/Bag 301, Blantyre, Malawi

3Chemistry Section, Department of Applied Studies, Malawi Institute of Technology, Malawi University of Science and Technology, P. O. Box 5196, Limbe, Malawi

4 Department of Chemical Sciences, Midlands State University, P/Bag 9055, Senga Road, Gweru, Zimbabwe.

*Corresponding author: bvunzanyari@gmail.com, mugumbateg@staff.msu.ac.zw

ABSTRACT

Selective inhibition of the γ-secretase complex is a promising therapeutic strategy to completely eradicate chemotherapy resistance in breast cancer management. In this work, structure-based virtual screening was used to identify γ-secretase inhibitors from African natural products databases (NANPDB, AfroDB, ConmedNP and Afrocancer). A druglike dataset of 437 compounds was created from cleaning natural product dataset of 11304 compounds. Autodock vina was used to dock 437 compounds against γ-secretase catalytic unit presenilin (PS1). Additionally, comprehensive ADMET analysis for novel inhibitors was done using pkCSM server. 12 compounds (COCONUT ID: CNP0329927, CNP0308825, CNP0419761, CNP0358535, CNP0214749, CNP0207760, CNP0183356, CNP0183161, CNP0358754, CNP0285192 and CNP0193480) which had binding energy ranges between 7.5kcalmol^{-1} and 8.6kcalmol^{-1} were identified. The compounds also formed hydrophobic interactions with catalytic residues of PS1 (TRP165, GLY384, ILE143, LEU166, LEU150, LEU286 and ASP257). ADMET analysis revealed that the majority of novel compounds have good absorption and metabolic profiles but exhibited hERG II toxicity. The compounds identified in this study provide a good foundation for a comprehensive search for natural γ-secretase inhibitors.

Keywords: breast cancer, multi-drug resistance, natural products, docking, pharmacokinetics

Appendix 4: manuscript 2

CHEMOINFORMATIC PROFILING OF DRUG-LIKE COMPOUNDS FROM AFRICAN NATURAL PRODUCT DATABASES FOR LEAD PRIORITIZATION AND DISCOVERY

Jonathan T. Bvanzawabaya [1, 2] *, Andrew G. Mtewa[3], Grace C. Mugumbate[4*] Fanuel Lampiao[1, 2]

1Department of Biomedical Sciences, Kamuzu University of Health Sciences, P/Bag 301, Blantyre, Malawi.

2Africa Center of Excellence in Public Health and Herbal Medicine (ACEPHEM), Kamuzu University of Health Sciences, P/Bag 301, Blantyre, Malawi

3Chemistry Section, Department of Applied Studies, Malawi Institute of Technology, Malawi University of Science and Technology, P. O. Box 5196, Limbe, Malawi

4 Department of Chemical Sciences, Midlands State University, P/Bag 9055, Senga Road, Gweru, Zimbabwe.

*Corresponding author: bvunzanyari@gmail.com, mugumbateg@staff.msu.ac.zw

ABSTRACT

The African continent contains enormous biodiversity resources which, remain a reliable source of pharmacologically active natural products (NPs). For the last ten years, several NP databases containing natural compounds from Africa have been compiled and made available to encourage in silico drug discovery. However, a comprehensive analysis of drug-like compounds in these databases is still missing. Therefore, the aim of this work was to characterize the chemical diversity and chemical space of drug-like NPs from 4 African natural product databases. A total of 11304 compounds from four African natural products libraries were curated and filtered to remove structural alerts and compounds that violate drug-like rules according to Lipinski and Veber. The chemical space and diversity of drug-like NPs was characterized relative to FDA drugs. Drug-like compounds from all 4 libraries demonstrated to be have lead-like physicochemical properties and shared the same chemical space with FDA drugs. However scaffold diversity analysis revealed that drug-like NPs are more unique, diverse and less complex than FDA approved drugs. These findings highlight that the African natural products databases contain drug-like NPs that are chemically diverse and hence provide promising and unique scaffolds for fragment-based drug discovery and development.

Key words: Natural products, druglike, scaffolds, chemical space, physicochemical properties

Appendix 5: manuscript 3

View PDF | Access through your institution | Purchase PDF

Chapter contents | Book contents

Outline

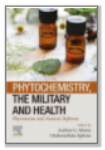
Show full outline

ELSEVIER

Phytochemistry, the Military and Health

Phytotoxins and Natural Defenses

2021, Pages 481-495



24 - Molecular optimization of phytochemicals into antidotes

Andrew G. Mtewa [a, b, c], Jonathan T. Bvunzawabaya [d, e]

Show more

+ Add to Mendeley | Share | Cite

https://doi.org/10.1016/B978-0-12-821556-2.00006-2

Get rights and content

Appendix 6: manuscript 4

Appendix 7: manuscript 5

View PDF Download full issue

ELSEVIER

Volume 12, July 2021, e00824

Ligand-protein interactions of plant-isolated (9z,12z)-octadeca-9,12-dienoic acid with B-ketoacyl-Acp synthase (KasA) in potential anti-tubercular drug designing

Andrew G. Mtewa [a, b], Jonathan T. Bvunzawabaya [c, d], Kennedy J. Ngwira [e], Fanuel Lampiao [c, d], Reuben Maghembe [f, g], Hedmon Okella [b], Anke weisheit [b], Casim U. Tolo [b], Patrick E. Ogwang [b], Duncan C. Sesaazi [b]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.sciaf.2021.e00824

Appendix 8: manuscript 6

Hindawi

Jor

SUBMITTED

Flexible Docking-Based Screening of African Natural Products for New Drug Leads Against Enoyl-acyl Carrier Protein Reductase of C. trachomatis (CtFabI): An In-silico Approach to Neglected Tropical Disease Drug Discovery.

Jonathan Bvunzawabaya[1], andrew Mtewa SA CA [2], Floryn Mtemeli[3], Davies Mweta[4], Fanuel Lampiao[1], Grace Mugumbate[5] + Show Affiliations

Article Type
Research Article

Journal
Journal of Tropical Medicine

Academic Editor Unassigned

Submitted on 2021-09-27 (an hour ago)

Abstract

Neglected Tropical Diseases (NTDs) continue to affect more than one billion people globally and the burden lies heavily on developing countries in Africa due to poverty and lack of proper treatment. NTDs such as trachoma, have very few therapeutic drugs used in treatment of the disease and this warrants rapid discovery of novel drugs. In this study, flexible docking was used to virtually screen 7523 compounds from three African natural products databases (NANPDB, EANPDB and SANCDB) against Chlamydia trachomatis enoyl-ACP reductase (CtFabI). Furthermore, in silico ADMET predictions were done on the promising inhibitors. Seven compounds; R6401, R2074, R2549, R636, R18, R6400 and R164 with binding energies (ΔGbinding) less than that of CtFabI inhibitor, AFN-1252 (-8.249 KCalmol-1) were identified as potential inhibitors of CtFabI. In silico ADMET predictions revealed that the 7 hits compounds have intestinal absorption and Caco2 permeability comparable to antibiotics currently used in trachoma management. Toxicity studies showed that potential Ctfabl inhibitors are non-CYP450 inhibitors, non-heRG channel inhibitors and are not hepatotoxicity. Therefore, this study identifies 7 promising natural compounds of African origin that could be developed as lead compounds for trachoma treatment.