Development and application of ligand-based cheminformatics tools for drug discovery from natural products

Entwicklung und Anwendung von ligandenbasierten Cheminformatik-Programmen für die Identifizierung von Arzneimitteln aus Naturstoffen

INAUGURALDISSERTATION

zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Albert-Ludwigs-Universität Freiburg im Breisgau

vorgelegt von

Kiran Kumar Telukunta

aus Hyderabad, Indien

Juli 2018

Dekan:

Prof. Dr. Manfred Jung
Universität Freiburg
Institut für Pharmazeutische Wissenschaften
Chemische Epigenetik
Albertstraße 25
79104 Freiburg

Vorsitzender:
des Promotionsausschusses

Prof. Dr. Stefan Weber
Universität Freiburg
Institut für Physikalische Chemie
Physikalishe Chemie
Albertstraße 21
79104 Freiburg

Referent:

Prof. Dr. Stefan Günther
Universität Freiburg
Institut für Pharmazeutische Wissenschaften
Pharmazeutische Bioinformatik
Hermann-Herder-Straße 9
79104 Freiburg

Korreferent:

Prof. Dr. Rolf Backofen
Universität Freiburg
Institut für Informatik
Lehrstuhl für Bioinformatik
Georges-Köhler-Allee 106
79110 Freiburg

Drittprüfer:

Prof. Dr. Andreas Bechthold
Universität Freiburg
Institut für Pharmazeutische Wissenschaften
Pharmazeutische Biologie und Biotechnologie
Stefan-Meier-Straße 19
79104 Freiburg

Prüfungsdatum: 30 Aug 2018

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit allein und ausschließlich unter Nutzung der direkt oder sinngemäß gekennzeichneten Zitate geschrieben habe. Weiterhin versichere ich, dass diese Arbeit in keinem anderen Prüfungsverfahren eingereicht wurde.

Kiran Kumar Telukunta
Juli 2018

Acknowledgements

“Learning gives creativity; Creativity leads to thinking; Thinking provides knowledge; Knowledge makes you great."

-A. P. J. Abdul Kalam

- Primarily I thank athletic Prof. Dr. Stefan Günther for giving me the opportunity to be part of Pharmaceutical Bioinformatics group, where I enjoyed my journey of knowledge and challenges of science expedition. I am happy to see his journey from Junior Professor to a Professor which gives me enormous motivation both personally and professionally. His diverse group of knowledge and peer PhDs from various countries making it a unique international research group provides a set platform with great exposure. I am fortunate to have Prof. Dr. Rolf Backofen again as supervisor to my Ph.D. and my special thanks to him. My gratitude to Prof. Dr. Andreas Bechthold for accepting to be my third supervisor.

- I sincerely thank my funding institutions German National Research Foundation (DFG) and University of Freiburg.

- One of the high motivation for me pursuing Ph.D. and completing it is the philosophy of IndiaYouth.info which I have quoted in my master thesis, and I am blissful that this envision still sought and it evolved in me in every Planck length (l_p) of the moment. This philosophy gave me the opportunity to thank Rama Rao Madarapu who joined me in this philosophy and gradually became part of my life. His substantial support is an essential asset during this period.

- I'm happy that Dr. Stephan Flemming and Dr. Xavier Lucas born in different countries because now I can express this easily that if somebody asks me if I were to born in Germany or Spain then probably I would have been like these guys respectively. They were everything to me in my new country of living. Both of them were always accessible and available for my ideas, queries and all topics of discussion. More importantly, several exciting activities with these guys will always be great memories in my life.

- Dr. Björn Grünning and Dr. Fidele Ntie-Kang were great inspiring colleagues during my Ph.D. Their maximum utilization of their resources in contributing to scientific progress is admirable.

- It was enormous pleasant feeling working and contributing to science along with Dr. Kersten Döring, Dr. Dennis Klementz, and Paul Zierep. I had a lovely environment working with Dr. Anika Erxleben, Pankaj Mishra, Dr. Gwang-Jin Kim, Mehrosh Pervaiz, Ammar Qaseem, Jianyu Li, Stefan Bleher, Dr. Martin Gotthardt, Dr. Nan Liu, Daniel Eckhardt, Dr. Maria Hörnke, and Lars Rösch.

- I thank Moritz Konrad, Laura van Hazendonk, and Lea Purschke for being part of my work team during my stay at Pharmaceutical Bioinformatics.

- My respect to the source of my life is my parents for their love and support. I gratefully appreciate and thank the unconditional support of my sister, brother-in-law, and especially my Aunt and Uncle during the journey of my Ph.D. Meinen großen Respekt und meine Liebe zu meiner deutschen Mutter Gudrun Lorenz und Familie, die mir ein besonderes Familiengefühl in Deutschland vermittelt hat.

- In the end but the extraordinarily special person in my life is my life partner Sheetal Arya Telukunta who gave me the opportunity of going through all my emotions during this period gives me immense pleasure in my joyride, and I am remarkably thankful to her for all sensations she has given to me.

- I thank my second home Germany for all its support during my Masters and Doctorate.

Abstract

In the drug-discovery identification of small molecules that selectively bind to a biological target from virtually infinite chemical space is a time-consuming crucial step among many other critical steps in drug development. Identified molecules need to possess adequate residence times of drug-target complexes to modulate the function of the target protein and must affect the desirable phenotype. Furthermore, examination of the pharmacological activity of compounds in *in vivo* studies is required which are characterized by pharmacokinetic and pharmacodynamic properties. Finally, the efficacy of the drug has to be validated in human.

The huge number of natural products being approved drugs indicates the importance of natural compounds for drug discovery. Genome-mining tools can be applied to identify a substantial number of novel natural products and ligand- or structure-based virtual screening methods will further increase the pace of therapeutic compound discovery.

The present doctoral thesis focuses on developing cheminformatic tools which aid basic research for lead identification in drug development. The following applications were co-developed within the scope of this work:

Tools evaluating existing literature by applying text-mining in natural language processing are becoming an essential part of identifying compound-protein, drug-drug, and protein-protein interactions along with their associations to diseases in literature. PubMedPortable is a framework developed for accessing large-scale biomolecule associated data and bridges the gap between natural language processing components and relation extraction methods by providing a local queryable and searchable instance of the literature.

NANPDB annotates thousands of compounds from the Northern African region; Strep-

tomeDB is an updated database of molecules produced by actinobacteria. Both developed libraries contribute significantly to the biologically relevant natural chemical space. Furthermore, provided web services allow for the retrieval of information about a therapeutic application, physicochemical properties, and synthesis routes.
Structural elucidation of biosynthetic substances is a hurdle. The developed web tool SeMPI provides a pipeline to identify encoding gene clusters from genomic data and predicts the basic structure of related natural products.
DVS offers an algorithm that serves to narrow down the chemical space that has to be screened to identify putative drugs. FragPred provides a solution in another direction by predicting the activity of compounds based on the knowledge of contained active substructures.

The cheminformatic tools presented in the thesis are useful for creating hypotheses for the discovery of novel drugs to certain diseases. A case study, diabetes mellitus, illustrates these tools and their operation. Starting with finding literature on diabetes mellitus and the identification of existing drugs for treatment, proposes alternative compounds from the presented natural databases. Finally, for the alternative compounds, putative targets are predicted. The manifested drug-discovery cheminformatic tools demonstrate the importance of *in silico* methods in modern drug discovery.

Zusammenfassung

Die Identifizierung kleiner Moleküle aus dem nahezu unendlichen chemischen Raum, die selektiv an ein biologisches Zielprotein binden, ist neben vielen anderen kritischen Schritten in der Arzneimittelentwicklung ein zeitaufwändiger Prozess. Identifizierte Moleküle müssen über ausreichende Verweilzeiten am Wirkstoff-Zielkomplex verfügen, um auf die Funktion des Zielproteins einzuwirken und den gewünschten Phänotyp zu beeinflussen. Darüber hinaus charakterisieren pharmakokinetische und pharmakodynamische Eigenschaften die pharmakologische Aktivität von Molekülen, die in *in vivo* Studien weiter untersucht werden muss. Schließlich muss die Wirksamkeit des Medikaments im Menschen validiert werden. Die große Zahl der zugelassenen Naturstoffe zeigt deren Bedeutung für die Arzneimittelforschung. *Genome-mining* kann eingesetzt werden, um eine Vielzahl neuartiger Naturprodukte zu identifizieren. Liganden- oder strukturbasierte virtuelle *screening*-Methoden unterstützen die Entdeckung neuer therapeutischer Wirkstoffe.

Die vorliegende Dissertation beschäftigt sich mit der Entwicklung cheminformatischer Werkzeuge, mit deren Hilfe die Grundlagenforschung zur Identifizierung von Leitstrukturen unterstützt werden kann. Die folgenden Anwendungen wurden im Rahmen dieser Arbeit mitentwickelt:

Text-Mining Werkzeuge mit deren Hilfe Literatur durch die Verarbeitung der natürlichen Sprache evaluiert werden kann, sind zu einem wichtigen Bestandteil der Identifizierung von Interaktionen zwischen Metaboliten, Arzneistoffen und Proteinen geworden und ermöglichen auch die Erkennung von Assoziationen zu Krankheiten in Texten. PubMed-Portable ist ein *framework* mit dessen Hilfe die genannten Biomoleküle und Assoziationen aus Texten extrahiert werden können und dafür eine lokal verfügbare Instanz der PubMed-

Datenbank bereitstellt.

NANPDB und StreptomeDB sind biologische Datenbanken, die den verfügbaren chemischen Raum um relevante Wirkstoffe erweitern: NANPDB beschreibt bekannte Naturstoffe aus dem nordafrikanischen Raum und StreptomeDB Moleküle, die in Aktinobakterien produziert werden. Darüber hinaus ermöglichen die bereitgestellten *webservices* das Abrufen von Informationen über therapeutische Anwendungen, physikalisch-chemische Eigenschaften und Syntheserouten.

Die strukturelle Aufklärung biosynthetischer Substanzen ist sehr aufwendig. Die Anwendung SeMPI stellt eine *pipeline* zur Verfügung, mit deren Hilfe Gen-Cluster aus genomischen Daten identifiziert und die Grundstruktur der synthetisierten Naturprodukte vorhergesagt werden kann.

DVS und FragPred sind Algorithmen zur effizienten Filterung des chemischen Raums in Bezug auf die adressierten Zielproteine bzw. zur Vorhersage der Aktivität eines Moleküls basierend auf bekannten darin enthaltenen aktiven Substrukturen.

Die cheminformatischen Werkzeuge, die im Rahmen dieser Dissertation entwickelt wurden, sind nützlich für die Gewinnung von Hypothesen zur Medikamentenentwicklung und veranschaulichen die Bedeutung von *in silico*-Methoden in der modernen Wirkstoffforschung. Anhand einer Fallstudie, der Krankheit Diabetes mellitus, werden die genannten Werkzeuge und deren Bedienung illustriert. Beginnend mit einer Literatursuche und der Bestimmung der verfügbaren Medikamente, folgt die Identifizierung weiterer möglicher Wirkstoffe mithilfe der zur Verfügung gestellten biologischen Datenbanken. Daran anschließend werden für die betreffenden Moleküle Wechselwirkungen mit Proteinen vorhergesagt.

Table of contents

Chapter 1

Introduction

1.1 Chemical space

In the field of cheminformatics, chemical space has fundamental relevance for medicinal chemistry and chemical informatics. The type of bond along with the number of atoms and their topological connections between structural formula majorly defines molecules. Combinatorially arranging compounds with heteroatoms, rings and obeying the laws of chemical valence erupts the chemical space to 10^{400} [1]. Applying *de novo* design (Section 2.1) following Lipinski's rule (Section 2.1.6) of potential drugs, the number of molecules that have drug-like characteristics are estimated to scale down to 10^{60}. However, that would be still more small molecules than the atoms in the Solar System [2], and so it is termed as Chemical Cosmos. Virtually there are an infinite set of possible organic compounds [3]. Nonetheless, with modern screening methods and with the expansion of biomedical knowledge combining with the completion of Human Genome Project (Figures 3.43, 3.44, 3.45) the ability to discover new molecules is inconceivable [4–6]. Discovery of new molecules to a certain extent simplified by natural products. According to the analysis of drugs approved by FDA since 1939 more than one-third of new molecular entities are natural products and their derivatives [7]. Although, the total fraction of natural products diminished on the other side semi-synthetic, and synthetic natural product derivatives have increased. Natural compounds play a vital role in the discovery of lead structures for drug development as revealed in numerous analyses [8–13].

1.2 Natural compounds

Naturally occurring organisms observed in nature which have metabolites produce natural compounds. Chemical transformations occurring in the cells of life-entities produce metabolites in the process of metabolism. The metabolites which are essential for the growth and maintenance of cellular function are primary metabolites. Vitamins, amino acids, nucleosides and organic acids constitute primary metabolites. Whereas, the products like alkaloids, flavonoids, steroids, antibiotics, gibberellins, and toxins are the secondary metabolite compound produced during the stationary phase of the cell growth and are not essential for growth and maintenance of cellular functions. Secondary metabolites are the end products of the primary metabolism. They have various features such as structure, signaling, stimulatory and inhibitory effects on enzymes, catalytic activity, defense, and interactions with other organisms. These functions of metabolites are responsible for many therapeutic actions of drugs. These propitious functions make Natural products potential, or novel drug leads in drug discovery [14]. On applying principal component analysis on chemical space, significant overlap between natural products and FDA-approved drugs of the chemical space indicates the importance of NPs to be potential lead compounds [15]. On average between 1981 and 2014 natural products contributed to drugs up to 42% and the share of natural products increases up to 75% in the area of cancer [12]. Scaffolds (Section 2.1.8) in natural products provide vital information as a starting point for hit-lead in drug discovery (Section 1.5). Collecting information about such natural products into a library is an extensive process. StreptomeDB 2.0 (Figure 1.3, Section 3.2) and NANPDB (Figure 1.3, Section 3.3) are such database (Section 2.5) collection combined with several beneficial features to browse and query compound information. Nature's library of chemicals are the consequences of biosynthetic pathways and are with evolutionary benefits atop human-made chemicals. The possibility of arranging natural products in number of ways with natural product biosynthetic logic [16] described by the variety of biosynthetic paradigms [17–21] makes it interesting to have insight into the diversity and distributions of natural product biosynthetic gene clusters.

1.3 Gene clusters

A *gene* is the basic physical and functional unit of heredity and are made up of a sequence of nucleotides, i.e., *DNA* encoding molecules having specific functions to form *proteins*. In humans, genes vary in size ranging from few hundred bases to more than 2 million bases. The Human Genome Project (HGP) initiative upon finishing the human genome sequence estimated human protein-coding genes to be 20,000-25,000 [5]. Genes during the evolution due to duplication of a single gene form a set of similar nucleotides share important characteristics. When a set of such similar genes originated by duplication with similar biochemical functions consisting of a similar sequence of DNA nucleotides forms a *gene family*. A gene family can also be from different genes if the set of genes participate in the same processes. *Phylogenetic tree* is formed by identifying similar characteristics and aligning inferred evolutionary relationships among various entities of biological species. The phylogenetic tree is a dendrogram or a tree structure representing the relationship between biological species (interspecific taxonomic level) or strains (intraspecific taxonomic level). Phylogenetic tree of nucleotide sequences is a rich resource for drug discovery process [22–25]. In phylogenetic tree (Figure 1.1) of *Streptomyces glaucescens* a source of acarbose compound forms the sibling of *S. flaveolus* forms the sibling and their common ancestor is *S. tenebrarius*. Classifying genes into families helps to predict where and when a specific gene is active or expressed, subsequently gives indications for identifying genes that are involved in a specific disease [26]. A *gene cluster* is a portion of a gene family and is an assortment of two or more genes available within an organism's DNA that encode for similar polypeptides, proteins which as a group share a generic function. Recognizing homologous genomic regions is a vital step for genomic analyses. Identifying homologous regions is a vital step when two genomic regions located within a few thousand base pairs. The detection of homologous regions which are scrambled due to evolutionary events is a difficult task; this requires more effort when gene clusters are in diverged genomes, i.e., gene and the order of gene not preserved in those locations merely content of gene is similar [27]. The gene clusters can be from some few genes to several hundreds of genes.

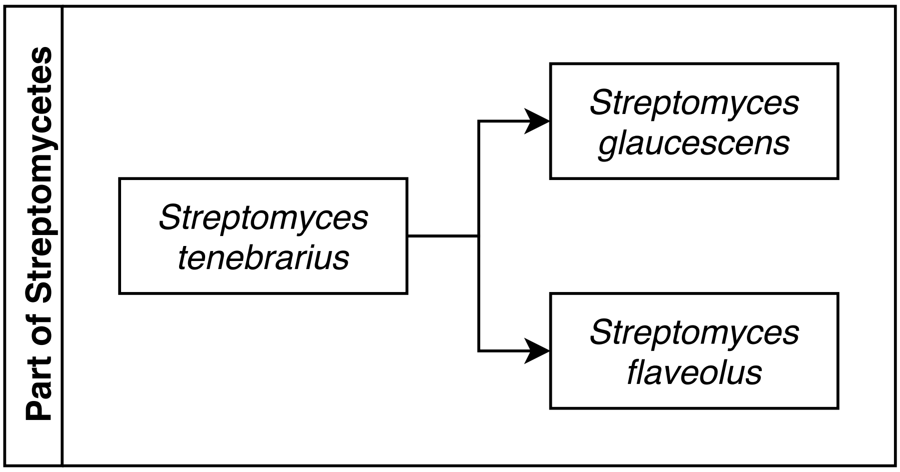


Fig. 1.1 An example of phylogentic tree representation taken from StreptomeDB 2.0 [22] of *Streptomyces glaucescens*. A source organism of *Acarbose*

Metabolic gene clusters are gene clusters which are genes participating in a typical, discrete metabolic pathway producing either primary or secondary metabolites. Secondary metabolites are the prolific basis for most antibiotic pharmaceutical compounds, chemical machinery used for chemical communication and forming a synthetic engagement between organisms and their environment. Secondary metabolic gene clusters are a common feature and are a rich source of metabolic diversity in bacteria such as prokaryotic actinomycetes (Section 3.2), filamentous fungi, and plants. Secondary metabolites are likely to bestow critical selective advantages for the producing organisms in nature as antibiotics. However, identifying the structures of secondary metabolite is a complex process to accomplish this we have devised a web-based tool SeMPI (Section 3.4). Human and veterinary medicines use these compounds which are the source of bioactive compounds. Clustering facilitates regulation of a set of genes in controlling steps in the biosynthetic pathway [28]. With rapidly decreasing costs, genomic screening has become a vastly used method in the search for new natural drugs [29]. Deciphering the biological activity presented by NP libraries and biosynthetic pathways of bioactive compounds is exciting for both therapeutic and diagnostic uses [30, 31].

1.4 Biological activity

In drug discovery, *biological activity* or *pharmacological activity* is one of the important factors which is defined as the capability of a compound, to alter or to create an effect or to elicit a response of one or more chemical or physiological functions of living tissue. Small compounds or the structure of the compound which exerts an activity on the living matter with the help of *in vitro* studies [32]. These studies determine if the compound exhibits positive or the negative effect and are accordingly stated to be a drug or a toxic substance. An example (Figure 1.2) of such bioactivity exhibited by StreptomeDB 2.0 compound (acarbose[1]) which is a potential diabetes drug with *alpha-glucosidase* bioactivity at target maltase-glucoamylase, intestinal. Thus acarbose upon clinical trials forms drug Glucobay.

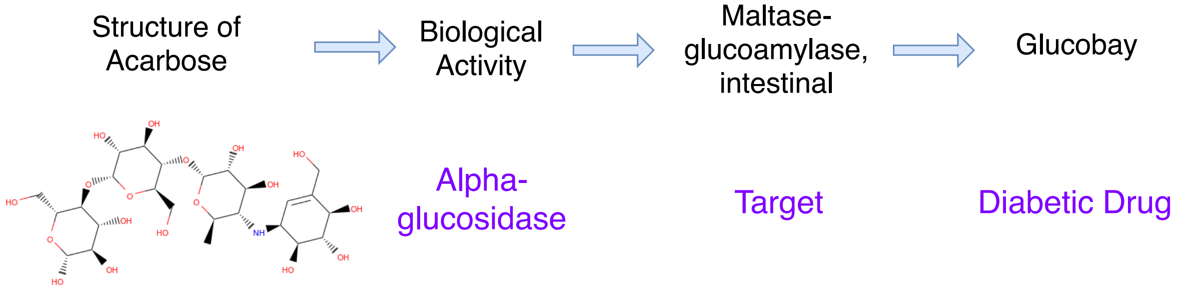


Fig. 1.2 An example of diabetes mellitus bioactivity by the compound in NANPDB

[1] http://phabi.de/streptomedb2/get_drugcard/9895/

1.5 Drug discovery

The motivation for the drug discovery arises with the condition of unavailability of the medical prescription for a disease or a clinical state or with the precautionary principle [33]. In certain cases, researchers majorly from academia or industry participates in the basic research (Figure 1.3) and engender data to foster a hypothesis that the activation of protein or pathway or inhibition will result for curing diseases by showing a therapeutic effect from the *in silico* study. The engendered data is an ever-increasing resource especially PubMed [34] is a huge resource and to search for such humongous resource we have worked up with a tool PubMedPortable (Figure 1.3, Section 3.1). According to Owens [35], if a molecule can modulate a protein or the biological target with high affinity while it is binding to it and if the binding ability modifies the function of the target producing a therapeutic effect to the patient, then the target is said to be *druggable*. The primary aim of this stage is to select a target which should be *druggable*, safe to use, i.e., not toxic, adapt to clinical needs, commercially viable. At this stage, the target can be any biological entity such as proteins, genes, and RNA. A possible target can efficiently bind to the putative small molecule. Drug molecules evoke a bio-activity response (Section 1.4) when they attach to the target. Right target identification reduces the time to drug discovery, as the target will have relevance to the disease. Data mining tools such as PubMedPortable (Section 3.1) induce the process of target identification in the available ever expanding biomedical data. Validation of selected target involves several possible techniques which range from *in vitro* tools, where animal models are used to modulate the desired target in patients. Other validation techniques available are disease associated with their genetics and expression data, over-expression of transgenics, comparative genetics, analysis of molecular signaling pathways, interactions with immunoprecipitation yeast two-hybrid, tools using bioactive molecules and literature survey and their competitor information.

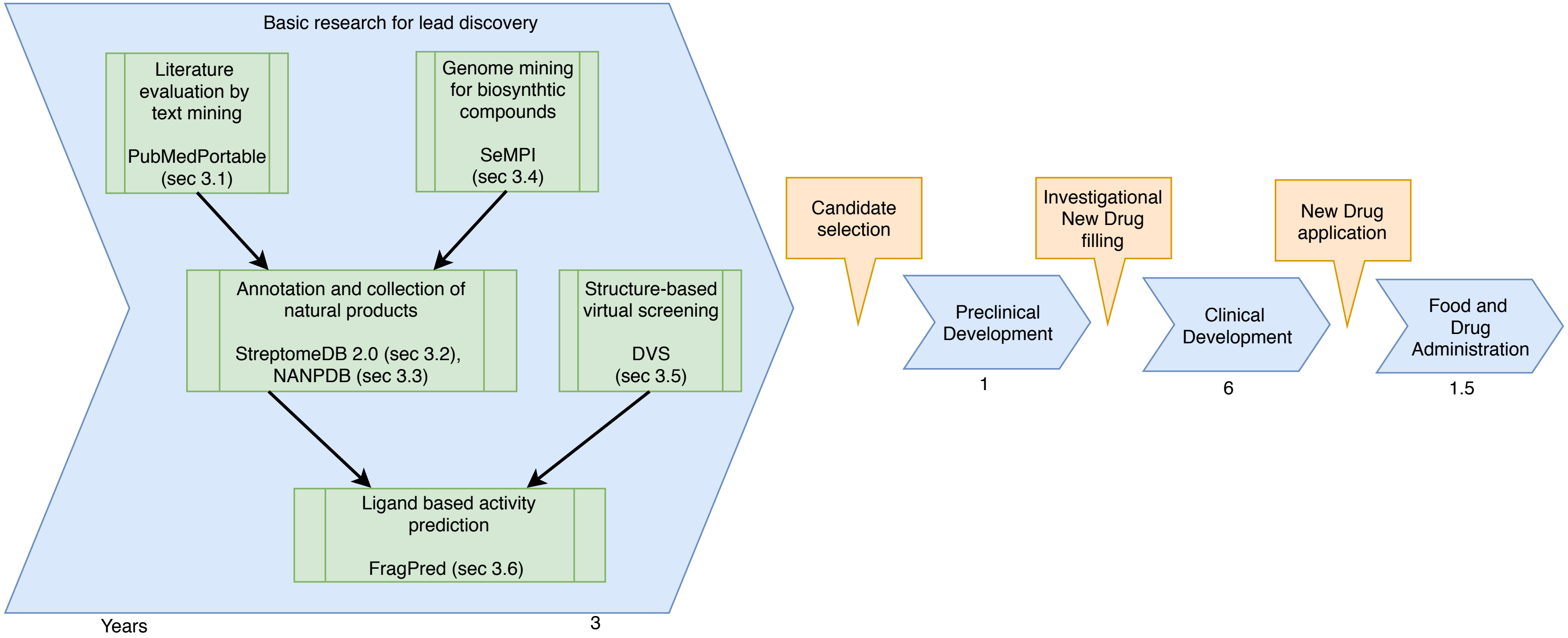

Basic research for lead discovery
Literature evaluation by text mining
PubMedPortable (sec 3.1)
Genome mining for biosynthtic compounds
SeMPI (sec 3.4)
Annotation and collection of natural products
StreptomeDB 2.0 (sec 3.2), NANPDB (sec 3.3)
Structure-based virtual screening
DVS (sec 3.5)
Ligand based activity prediction
FragPred (sec 3.6)
Years
3
Candidate selection
Preclinical Development
1
Investigational New Drug filling
Clinical Development
6
New Drug application
Food and Drug Administration
1.5

Fig. 1.3 Drug discovery process, adopted from Hughes et al. [36]. The work written in this thesis is concentrated in the intial basic research phase. Section number of the projects written in parentheses

Further, in the lead-discovery phase or the hit-to-lead (H2L) phase compound screening assays are developed where an activity of the compound is affirmed. Identification of hit molecules in this stage can use any of the screening methods available such as high throughput screening, where screening of the collected library against the drug target or an assay system whose bioactivity is dependent on the target and gets confirmed with the secondary assays to determine their efficacy [37]. Biophysical testing, such as nuclear magnetic resonance (NMR) section 2.1.2. The screening is left out with the option of utilizing laboratories if there is no advance knowledge of the activity of the compound at the target. However, this is a time-consuming process.

1.5.1 Final phases of drug discovery

The earnings of compound screening is a hit compound which has an activity at the target protein. Further in the hit and lead discovery phase assay development is involved. In this test-systems are devised which quantify and determine the cellular levels of a specific protein, levels of the metabolite in serum or urine, examine catalytic activity of an enzyme in normal and abnormal tissue. Assays can be biochemical or cell-based. At last in this stage a potential compound having ADME properties, physicochemical and pharmacokinetic measurements adequate to examine their efficacy in *in vivo* models. Subsequently, in the lead optimization phase, agreeable properties molecules are kept intact while improving on flaws. Once a candidate compound is selected, it undergoes preclinical steps, which usually takes unspecified or the longer duration. In the end, the compound chosen heads towards clinical phase, once a compound entering into this phase has a high probability of reaching to the market after clearing FDA approval.

Chapter 2

Methods

2.1 Virtual screening

Briefly passage to VS can be divided into two major methods one based on the structure of the biological target where its three-dimensional description is known, structure-based drug discovery (SBDD) commences with 3D structural information with the knowledge on the target of interest. Mostly, the identification of structure is by the experimental methods of X-Ray, NMR, homology modeling [38]. Molecular dynamics (MD) simulation can play a key role in the screening process by providing qualitative information about the potential of the target activity by calculating the binding stabilities of plausible hits with their target [39]. Further, a scoring function is utilized to approximate free energy binding score between the protein and the ligand in docking pose. Subsequently, selection of compound is made using filtering tools and empirical rules [40]. DVS (Figure 1.3, Section 3.5) is such method which gives attention to SBDD. In pharmaceutical chemistry a *ligand* is a molecule or an ion binds to a chemical entity with non-covalent bonds, interactions governing non-covalent bonds can be of electrostatic, π-effects, van der Waals forces, and hydrophobic effects. Thus a ligand forming a relationship with a binding partner is a function of charge, hydrophobicity, and molecular structure. In protein-ligand binding, the ligand is usually a molecule that binds to a site on target protein [41]. SBDD based on the ligand can be a *de novo* design [42] in which the ligand molecules are built up with the understanding of the binding pockets near the targets. Assembling of ligand molecules involves individual atoms or small fragments and with the help of computer-assisted tools. The major advantage of *de novo* method is the structures are completely novel and synthetic could well be not in any of the chemical databases. Another category of VS includes ligand-based techniques such as similarity and substructure searching, quantitative structure-activity relationships (QSAR) [43], pharmacophore [44] and three-dimensional shape matching [45] made up the ligand-based virtual screening (LBVS). Properties of known ligands inspire to unearth new binding compounds called as ligand-based virtual screening. Here 2D molecular similarity (Section 2.2) is applied to identify new similar molecules [46, 38, 47]. The similarity technique is also used to cluster data sets to recognize similar chemotypes [48]. Screening of vast libraries can also be narrowed down by partitioning compounds binding to specific target classes, knowledge of similar structures, compounds classified by their stereogenicity or hybridization all forming into smaller focused libraries [6]. Another type of screening involves *in vivo* techniques. In this physiological screening effects of the drug at tissue are built into a model relevant to the patient condition and toxic effects of drugs [49]. In this method complexity of tissue is better understood which yields drugs pertinent to disease.

2.1.1 Fragment based drug discovery

Inspired by focused or the knowledge-based screening, in fragment screening initially a fragment is acquired, positioned on data of its crystal structure, molecular weight, DNA-encoded methods recurrently enhanced into robust binding molecule [50]. Furnishing the knowledge of targets of the existing compounds to the fragments improve fragment-based screening technique, this predicts the target of the new compound; adaptation of this method is more explained in FragPred (Figure 1.3, Section 3.6). Arranging the present compounds in fragments is a combinatorial problem which falls into two categories either biased toward a broad range of targets or to a specific target. For any fragment based method, fragmenting the molecule is an important step, and Lewell et al. [51] proposed RECAP (Retrosynthetic Combinatorial Analysis Procedure) computational based combinatorial technique. It starts by collecting active structures based on a target class. Subsequently, fragment space is built based on predefined cleavage rules. Knowledge of standard chemical reactions and affinity towards synthesizable fragments defines RECAP cleavage rules. Molecules having any of the eleven chemical bonds (Figure 2.1) triggers the bond cleavage. RECAP does not cleave

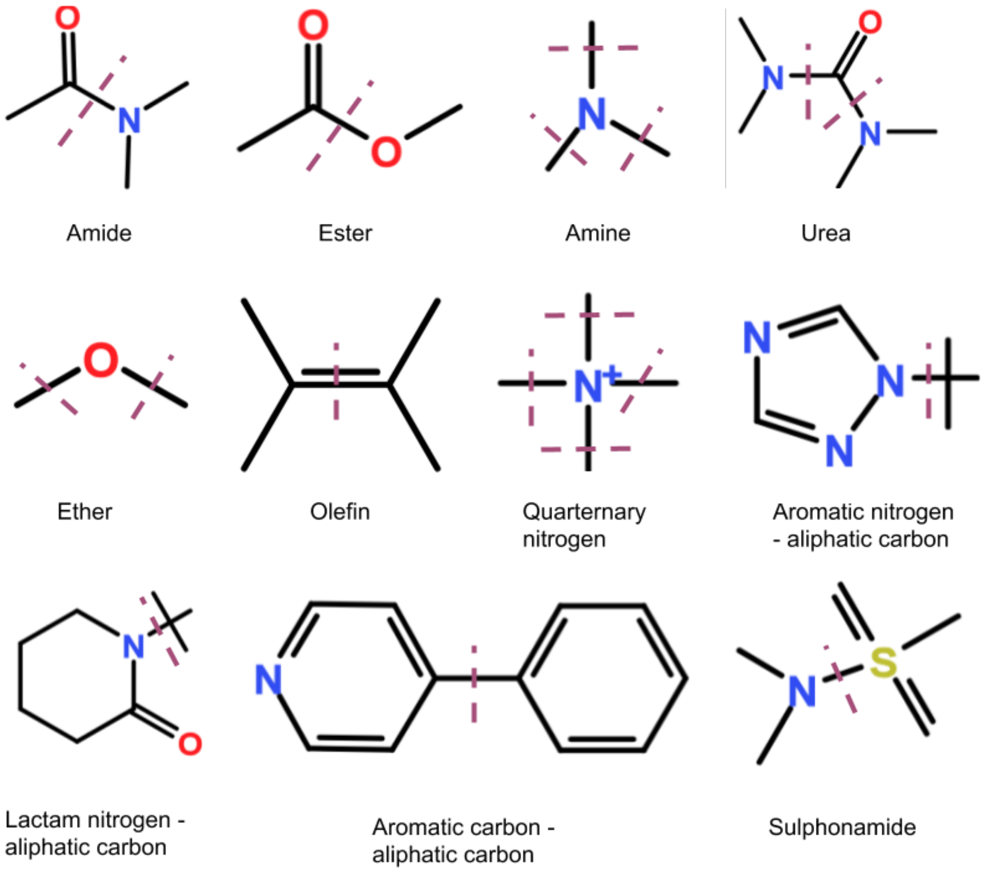

Amide
Ester
Amine
Urea
Ether
Olefin
Quarternary nitrogen
Aromatic nitrogen - aliphatic carbon
Lactam nitrogen - aliphatic carbon
Aromatic carbon - aliphatic carbon
Sulphonamide

Fig. 2.1 RECAP bonds prescribed by Lewell et al. [51]

terminal bonds if they include any of the predefined functional groups (hydrogen, methyl, ethyl, propyl, and butyl).

2.1.2 NMR

Nuclear magnetic resonance (NMR) spectroscopy is a physical phenomenon, the initial phase of drug discovery (Section 1.5) takes advantage of this technique. Isotopes with an odd number of protons have angular and intrinsic magnetic momentum. In the presence of this static magnetic momentum, the resonance frequency of a particular substance is directly proportional to the strength of the applied magnetic field. These properties get used in the physical experiment. The absorption and re-emission of electromagnetic radiation by nuclei in the arranged magnetic field of the required substance emits a measurable amount of radio frequency signal. This signal can give information about the molecular structure and their interactions at the atomic level. The NMR chemical shift being sensitive can give vital information of the small molecule if it binds to the biological target, i.e., protein or nucleic acid. It can also give the information about which unit of the small molecule bound to the macromolecular target [52]. NMR is one of the simple methods for ligand binding studies about hit identification and validation [53]. StreptomeDB (Section 3.2) and some of the NANPDB (Section 3.3) compounds are provided by its NMR information.

2.1.3 Mass spectrometry

According to Boggess [54] mass spectrometry (MS) is an analytical metric which measures the mass of different molecules within the observed sample. It is an experimental technique which converts the chemical molecule into an atom, molecule, or substance into an ion or ions, typically by removing one or more electrons. The ions then sorted based on their mass-to-charge (m/z) ratio. The process of MS involves vaporizing a sample of molecules, and the electron beam bombards the vapors, which converts the gases to ions. Since mass spectra measure the mass of the charged ions or broken fragments, neutral molecules are invisible in this process subsequently positive and negative ions are noticeable. Formation of positive and negative ions are by taking electrons away and giving electrons to molecule respectively. The acceleration and the deflection of ions produce the m/z values. Lighter particles move faster towards negative plates at speed dependent on their mass compared to the heavier ones, and the deviation of atoms is dependent on the mass of the particles. MS functions with a constraint only if vaporization of a chemical substance does not decompose the compound. In drug discovery mass spectrometry can be a vital tool in determining the structures of drugs and metabolites and screening for metabolites in biological systems.

2.1.4 QikProp

QikProp [55] is a proprietary software from *Schrödinger* Company. QikProp computes physical relevant descriptors and uses them to perform absorption, distribution, metabolism, and excretion (ADME) prediction designed by Prof. W. L. Jorgensen. It's an easy to use software, the prediction of pharmaceutically relevant properties of small molecules makes QikProp vital tool. QikProp extracts properties of small molecules individually, or in batch mode, which makes the software serviceable as usually in chemical spaces (Section 1.1) there is a large number of molecules. *Pharmacokinetic profiling* uses the properties produced by QikProp. The complete list of QikProp descriptors and properties are a list in [56].

2.1.5 Pharmacokinetic profiling

Pharmacokinetics is a fundamental discipline in the applied therapeutics and pharmacy. Proper medication provides patients assistance in health who are in need of medicines for a clinical condition. The dose is chosen by an evidence-based approach so that it can be compatible with other drugs or even alternative therapies if the patient is taking or suitable for the patient's metabolism [57]. Pharmacists follow drug use process (DUP) and take into consideration of need, choice, the frequency of the drugs, goals of the therapy, dosage, monitoring, and counseling. By the patient's drug handling parameters, which depend on the processes absorption, distribution, metabolism, and excretion. All these processes of pharmacokinetics provide the mathematical basis to assess the ability to react for medicine of the patient against the disease and to determine the drug concentration in the body; these processes are popularly referred as ADME. Mathematical parameters of ADME provide the fundamental understanding of them. QikProp generates all these properties.

2.1.6 Lipinski's rule

Lipinski's rule of five [58, 59] is a rule of thumb which is not accurate and also not reliable for every situation. The rule with the ADME descriptors estimates the drug-likeness of a chemical compound with a biological activity if it has chemical and physical properties that would make compound orally active drug for humans. Compounds satisfying the indicators are just highly probable to be a drug. However, the rule does not predict if the compound is pharmacologically active. The rule is advantageous in screening large chemical space. Molecules satisfying the rule of five (RO5) have lower rejection rates during the clinical trials. Lipinski's rule prescribes that an orally active drug has no more than one violation of the following criteria:

- The total number of nitrogen-hydrogen and oxygen-hydrogen bonds forms the hydrogen bond donors, and they should not be more than 5.
- Either all nitrogen or oxygen atoms forms the hydrogen bond acceptors which have to be no more than 10.
- Molecular mass of the molecule should be less than 500 Da
- The octanol-water partition coefficient (log P), i.e., the measure of solubility or the extent of hydrophilic nature of the chemical substance should not be great than 5.

Lipinski [59] states that 90% compounds satisfying this rule have achieved phase II clinical status. RO5 is useful in evaluating compound libraries (Sections 3.2, 3.3).

2.1.7 SP docking

According to Friesner et al. [60] glide HTVS (High throughput virtual screening) and SP docking use a series of hierarchical filters to search for possible locations of the ligand in the binding-site region of a receptor. The shape and properties of the receptor are represented on a grid by different sets of fields that provide progressively more accurate scoring of the ligand pose. During docking process, exhaustive enumeration of ligand torsions generates a collection of ligand conformations. Deterministically initial screening done over the entire phase of chemical space available to the ligand to locate promising ligand poses. The docking calculations were performed with Glide 5.6 [61]. Glide 5.6 has two modes one Standard-Precision (SP Docking) mode is a milder approach and minimizes false negatives. The second mode is Extra-Precision (XP Docking) mode is a more laborious function that applies penalties for poses that violate established physical chemistry principles such as that charged and strongly polar groups be adequately exposed to solvent [62].

2.1.8 Scaffolds

A *scaffold* is a single unit of a compound and usually defined according to the focus of the problem which is getting addressed. The unit left after dissecting rings, linkers or the single bonds connecting the ring structures, side chains or the leaves of the graph considering the molecule as a graph is a scaffold. However, it is not that often that this unit becomes the critical component or components of the molecule. It is sometimes possible that there can be multiple deserted units. The definition of a scaffold is better expressed when it is subjective to the objective presuming that structures are sharing a scaffold also share common synthetic pathway. Once the scaffold is defined is applied to the screening library, and the number of units or the diversity represents the character of the compound space. Either the compact representation or the sparse description of scaffolds can elevate problem of redundancy or rapid generation of structure-activity relationships respectively [63]. A suitable depiction of the scaffold is invariant, objective and not dependent on the dataset that enhances the scaffold diversity.

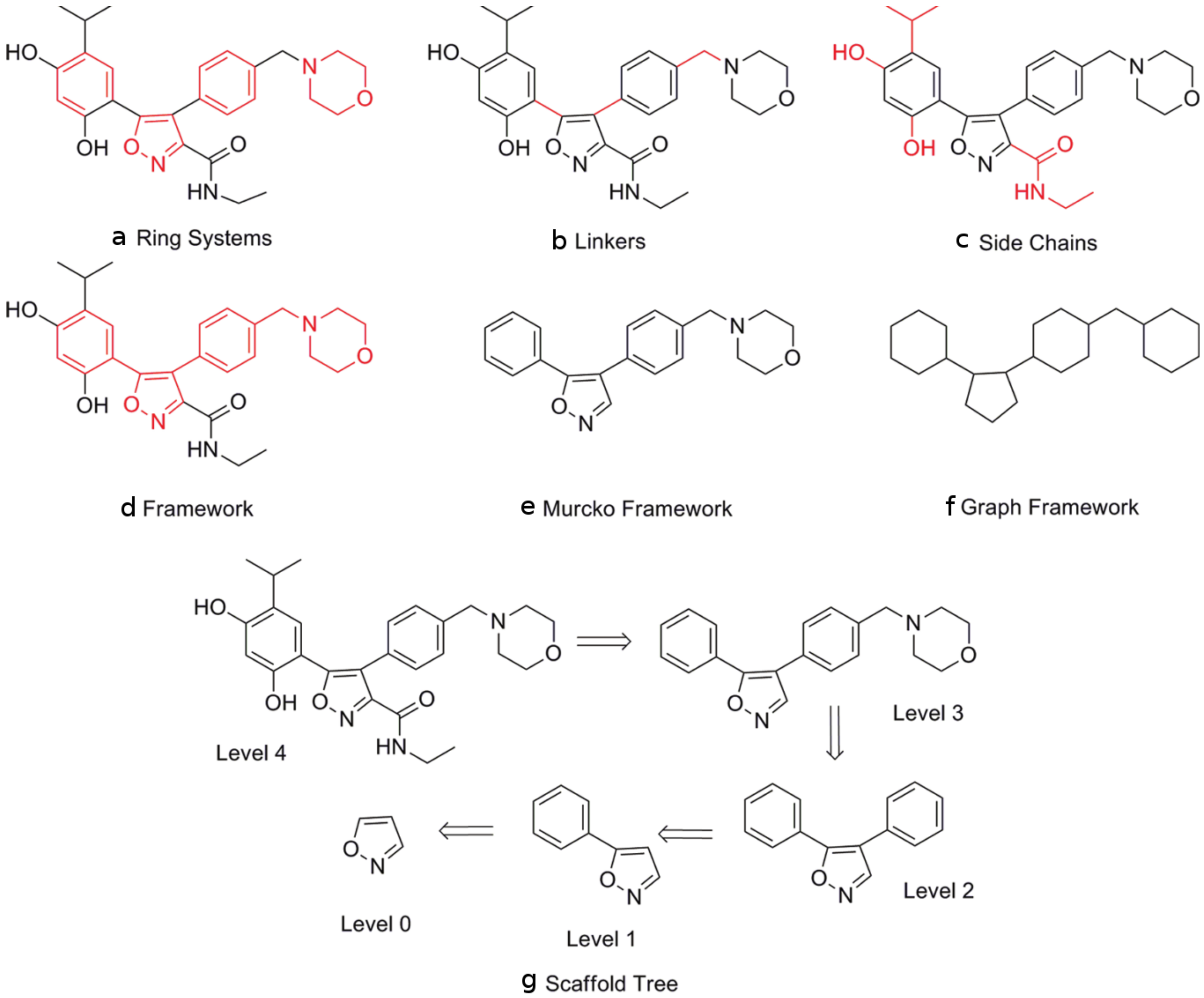


Fig. 2.2 Scaffold representation following Murcko Framework of HSP90 inhibitor taken from Langdon et al. [63]

And and Murcko [64] proposed a framework to analyze the structures of the discovered drugs. It suggests forming a structure into rings (Figure 2.2-a), edges connecting the rings acting as linkers (Figure 2.2-b), leaf nodes or the unsaturated valency bonds (Figure 2.2-c) working as side chains of molecules. The aggregate of all the above elements in the molecule forms the framework (Figure 2.2-d). Holding atom type information finally structure encompasses into Murcko framework (Figure 2.2-e), and the graph representation completes the graph framework (Figure 2.2-f). Translation of Murcko framework into scaffold tree facilitates the computational objectives. Attaining the scaffold tree can obey several schemes [65] of representation according to the requirement securing specific properties such as biological activity. The procedure to achieve scaffold tree cleaves each molecule in the library iteratively by pruning rings based on the predefined rules until a ring (Level 0 in Figure 2.2-g) arrives traversing n levels. Here n^{th} Level is the whole molecule, and the Level $n-1$ is the Murcko framework (Figure 2.2-d).

2.2 Molecule similarity

The existence of bioisosteres [66, 67], i.e., chemical substituents with common physicochemical properties which produce broadly similar biological properties in a molecule and knowledge of formulating them conceives the idea of molecular similarity. A measure of similarity or the dissimilarity indicates how close two entities are or group of things are to one another. Measurement takes into account a confined space within which the closeness or the similarity is. Among several methods available for distance d measurement Euclidean distance is one which measures the distance between two points x and y within the Euclidean space with dimension S.

$$d = \sqrt{\sum_{i=1}^{S} (x_i - y_i)^2} \tag{2.1}$$

The Euclidean distance relates to the molecular similarity. However, the distance measurements cannot be same for molecule distance. To measure the distance between two molecules points have to be defined, and these points can be chemotypes [48], cliff-forming compound pairs [68], i.e., compounds which are structurally similar where bioactivity is recognized. Molecules can also be described based on the criteria followed to select a compound and sometimes along with its target. Describing the molecule in such a way that it can be measured can ease the effort of comparing fragments. These multiple descriptors expressed as bits of binary string or instead fingerprint completes the description of a compound. The primary division of family of descriptors is into 2D and 3D topological information.

2.2.1 Fingerprints

The idea of molecule similarity is the way of eliminating molecules from the chosen chemical space. *Screening* or the substructure search is the method applied to eliminate unrequired candidates. The elimination process makes use of *fingerprints*. Fingerprints are an abstract representation of molecules which are suitable for computational process consequently can be applied to a big dataset. A pattern of target or substructure, a molecule structure forms the basic requirements. Fingerprints differ on their length, pattern size, size of count vectors, active bits, and the cost or the order of algorithms applied to them. Substructure search is considered to be a *non-poliynomial-complete* (*NP-complete*) class of computational problem [69]. It is similar to subgraph isomorphism problem and its worst-case time can be of the order $\mathcal{O}(\mathcal{K}^{k\mathcal{N}})$ where $\mathcal{N}$ is the number of atoms or bonds, which means every addition of new entry the time complexity doubles. The range of the fingerprint bits length can be 32 to 16,384 in general. However, the length can be more depending on the programming language capabilities. In C and C++ the typical datatypes found are *unsigned short int* (16-bit word), *float* (32 bit floating point precision), and *double* (64 bit double precision). The bits in the fingerprint represent if the encoded feature is available or not in the object or the molecule in the consideration.

2.2.2 Types of fingerprints

Classification of fingerprints based on the algorithms can be dictionary-based or structural-based, topological or path-based, circular and pharmacophores. List of current fingerprints used for similarity metrics is in table 2.1. Several types of fingerprints were developed for similarity metrics as no one similarity measure will be the best in every case. In molecule similarity measure there is no canonical definition of molecule similarity. There have been several studies [47, 70] to explain about performance and benchmarking of fingerprints. Further studies confirm that fingerprints performance majorly dependent on the type of data-set composition or the decoys chosen and the ability of the fingerprint to describe the molecules regarding bits for computing for a specific activity against a specific target. The diversity of the molecules in consideration, if they are very similar or very nonsimilar can make fingerprints uncomfortable as the active and the inactive nature of the compound cannot be differentiated. Studies also suggested which fingerprints are best for medicines and which kind of fingerprints to avoid. These studies majorly have used ChEMBL datasets which have information about their activity.

It is often important to have a performance overview of the fingerprints. O'Boyle and Sayle [70] illustrated a tree of fingerprint performance (Figure 2.3); it illustrates relative fingerprint review summarized in a directed graph format. Performance evaluation for benchmarking in O'Boyle and Sayle [70] adopted single-assay benchmark, multi-assay benchmarks and also illustrated directed graph for the benchmarking performed by Riniker and Landrum [47]. According to O'Boyle and Sayle [70] *single-assay* benchmark analyses the ability to rank very similar structures relative to a reference chosen. The molecules chosen were differing by 0.4 log units and are structurally similar. Starting from the reference being a most bioactive unit others test units are decreasingly active units, assuming that more similar activity to the reference, the more similar structure. On the other hand *multi-assay* benchmark test focused on different structures relative to the reference. It has taken four data set of molecules from the different articles such that from one group to another one connecting molecule is similar and within the group, they are similar with increasing distance of similarity between them. Based on the performance of the single and multi assays for all the fingerprints (Table 2.1) a directed graph is generated (Figure 2.3) fingerprints lower in the graph are worse than the fingerprints which are higher. Moreover, the distance or the numbers of directed nodes (Figure 2.3) indicate the net difference, i.e., number of times better minus number of times worse. Comparing the single-assay benchmarks (Figure 2.3-a) a notable change observed for atom pair fingerprints [71], AP and HashAP are on top of the directed graph and were expected to be poor as they do not encode paths in the structure compared to Daylight-type fingerprints. With the results from multi-assay fingerprints, LECPF4 and LECFP6 are consistently among the best fingerprints. And in general extended-connectivity and feature-class fingerprints performed better than others. Multi-assay benchmarks also reveal that there is not much difference between LECFP6 and ECFC6 fingerprints. And LFCFP4/FCFC4 are lower in the ladder compared to ECFC6. In case of ECFP fingerprints longer version LECFP performed better indicating an increase in the length of the fingerprint, increases the performance.

Fingerprint	Family	Algorithm / Class	Radius (Circular)/ Key length (Sub-structure)/ Path-based (length of path)	Notes
ECFP (0, 2, 4, 6) [72]	Morgan FP	Circular	0, 1, 2, 3	Extended-connectivity fingerprints
ECFC (0, 2, 4, 6) [73]			0, 1, 2, 3	Count vector form of ECFP
FCFP (2, 4, 6) [74]			1, 2, 2	Functional-class fingerprints. Properties that relate to ligand binding.
FCFC (2, 4, 6) [47]			1, 2, 3	Feature-connectivity count vector
LECFP (4, 6) [70]			2, 3	Longer version of ECFP (instead of 1024 bits here 16384 bits)
LFCFP (4, 6) [70]			2, 3	Longer version of FCFP (instead of 1024 bits here 16384 bits)
MACCS [75]		Substructure -based (or) dictionary -based	166	Molecular ACCess System structural keys
FP3 [75]			55	Uses SMARTS pattern
FP4 [75]			307	Uses SMARTS pattern
Avalon [43]	Flag based			Enumeration of feature classes of molecular graph
AP [71]	Atom Pair FP	Path-based or topological		Shortest path separations between all pairs of atoms in topology.
hashAP				Hashed forms of atom pairs
TT [76]	Topological torsion FP			Inspired by the basic conformational element, torsion angle
hastTT				Binary vector with torsion angle
RDK (5, 6, 7) [77]	RDKit FP		5, 6, 7	Hashed code of linear subgraphs of specified length of path within minPath and maxPath
FP2				Indexes small molecule fragments based on linear segments

Table 2.1 Various fingerprints utilized for ligand similarity

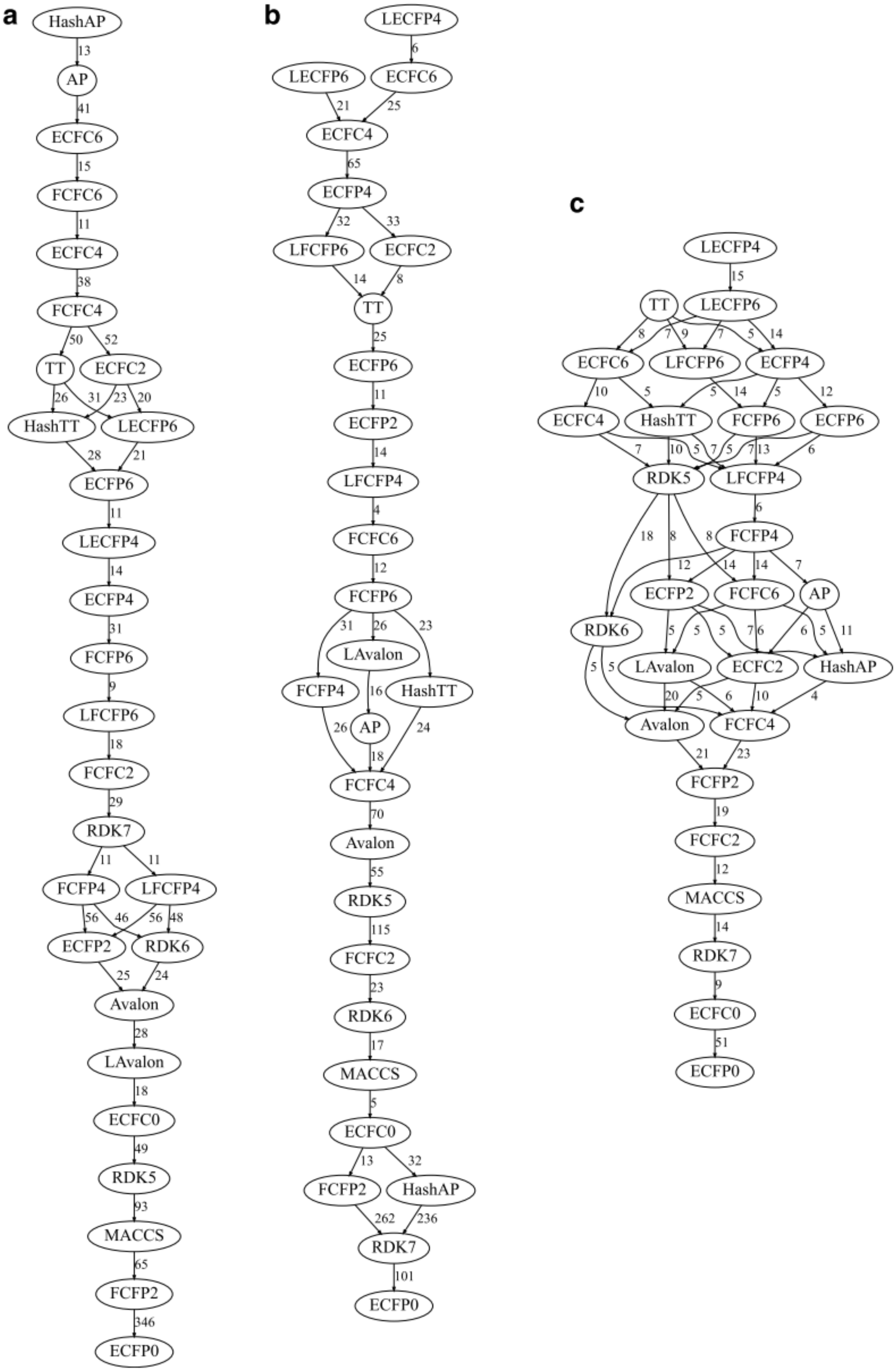


Fig. 2.3 Fingerprint performance taken from O'Boyle and Sayle [70] a) performance with single-assay, b) performance with multiple-assay, c) Riniker and Landrum [47] benchmark

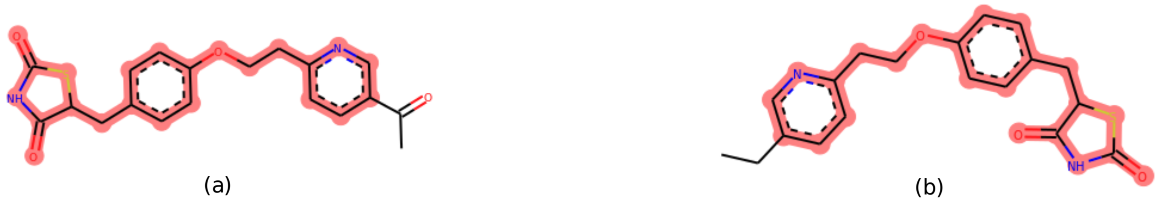


Fig. 2.4 Similarity comparison of Diabetic DrugBank compound with StreptomedDB 2.0 compound and has a Tanimoto coefficient 0.76 while using the ECFP4 fingerprint. (a) StreptomeDB 2.0 compound ChEMBL1268. (b) DrugBank diabetic compound Pioglitazone

2.2.3 Tanimoto coefficient

After the choice of a fingerprint, the next step is to measure the similarity of the molecules, i.e., as the Euclidean distance. It is important to understand that Tanimoto measure is independent of the molecular descriptors or the bits which are encoded. Fingerprints have to satisfy the requirement of having size with a number raised by 2 and is a multiple of 8; this eases the requirement of computational encoding of bits. Tanimoto coefficient also referred as Jaccard coefficient is the simple measure of comparisons among the presence of bits in two fingerprint strings. Considering two fingerprint bit strings *A* and *B*, the formulation can be *a* count of bits in fingerprint A but not in B. Let *b* be a count of bits in fingerprint B but not in A and lastly *c* count of bits which are present in both the fingerprints. Tanimoto does not consider the count of bits which are present in both A and B. Lastly it can be computed with equation (Equation 2.2). In figure 2.4 a similarity comparison given it is between diabetic DrugBank compound with StreptomedDB 2.0 compound and has a Tanimoto coefficient of 0.76 while using the ECFP4 fingerprint. Given a *N* bits string of fingerprint, the number of possible Tanimoto similarities is its number of *Farey numbers*. And the asymptotic expected number of coefficients, i.e., Tanimoto proximities is given by Euler equation 2.3 [78]. The equation approximates the number of proximities as the length of *N* increases.

$$\text{tanimoto coefficient} = \frac{c}{a+b+c} \tag{2.2}$$

$$\Phi(N) = \frac{3}{\pi^2}N^2 + \mathcal{O}(N\log N) \tag{2.3}$$

2.3 Softwares

2.3.1 RDKit

RDKit [79] is an Open-source toolkit Cheminformatics Software which has an extensive cheminformatics Python[1] and C++[2] API. It is based on BSD license, and its basic library is based on data structures and algorithms in C++ which makes it better regarding performance. An interface Python wrapper is built over it using Boost libraries for both 2.x, and 3.x version of Python and the latest 2018.03.1 version completely supports Python 3.x. It has a vibrant library for calculating all the fingerprints. RDKit operates in Linux along with Mac and Windows.

RDKit provides an extensive list of cheminformatics libraries in its API. However, some of the notable libraries which are used in this thesis are listed here:

- Package Chem[3]: Module for molecules, several kinds of fingerprints generation. Accessing all atom pairs of molecules. *RDKFingerprint*[4] part of Chem package gives options to control minimum path size(1), maximum path size(7), fingerprint size(2048), number of bits per hash(2), minimum fingerprint size(64 bits), and target on bit densitiy(0.3). The numbers in the parentheses are the default values observed by RDKFingerprint. Changing these paremeter most of the fingerprints can be acheived.
- Package Draw[5]: Useful to generate 2D structure images of compounds
- Package pyAvalonTools[6]: Module containing the functionality from the Avalon toolkit.
- Package DataStructs[7]: Module containing an assortment of functionality for basic data structures.

[1] http://www.rdkit.org/docs/api/index.html
[2] http://www.rdkit.org/docs/cppapi/index.html
[3] http://rdkit.org/Python_Docs/rdkit.Chem-module.html
[4] http://www.rdkit.org/docs/api/rdkit.Chem.rdmolops-module.html#RDKFingerprint
[5] http://www.rdkit.org/Python_Docs/rdkit.Chem.Draw-module.html
[6] http://www.rdkit.org/Python_Docs/rdkit.Avalon.pyAvalonTools-module.html
[7] http://www.rdkit.org/Python_Docs/rdkit.DataStructs-module.html

2.3.2 Python

Python programming language is a easy, expressive, free and open-source, high-level, portable object-oriented language. Projects PubMedPortable (Section 3.1) StreptomeDB 2.0 (Section 3.2), NANPDB (Section 3.3), FragPred (Section 3.6) were implemented with python.

2.3.3 Galaxy

The evaluation of complex biological datasets is often associated with the incorporation into program environments and libraries such as R [80] and BioConductor [81], BioPython [82] or also BioPerl [83]. Galaxy allows the use of these partially complex projects without knowledge of the underlying programming and represents the functionalities of them with the help of a minimalistic surface. The individual steps of analysis are transferred to different modules and can be linked to a sequence. The reproducibility of the results of a study depends on the availability of the corresponding data sets and the complete documentation of the individual processing steps. Galaxy allows the creation of analysis schemes for the processing and evaluation of data from various Bio and cheminformatic fields. The functional scope of Galaxy as delivered can be considerably improved with the help of packages called tool sheds. Tools sheds offer through a decentralized organization added to the official galaxy tool shed [84]. Galaxy uses a client-server model. Accordingly, the program on the server is readily accessible via the client to the user and application does not necessarily run on the same computer but can be used within the same network or also openly via the Internet. Access is via a web browser, and the central computing load lies with the server computer, which means that even in computationally intensive processes does not use the local machine. Complete programs or individual program methods can be divided into modules and combined with each other. Each module is assigned one or more input and output modules. Files with a specific data type along with their format specifications are taken into account. Sequences of workflows is a useful method to work with the more substantial pipeline.

The ChemicalToolBoX [85] is a collection of cheminformatics tools integrated into Galaxy-workflow-management enables applications for similarity and substructure searches, clustering of compounds, prediction of properties and molecular descriptors. It also has fragmentation and fragment merging tools. Combination of existing tools more customized tools can be designed. ChemicalToolBox, an open-source software, can quickly be deployed locally or on a large scale cluster. Galaxy also has several easy to use statistical tools.

2.4 Statistical tools

2.4.1 Enrichment analysis

Enrichment analysis is a method to identify whether specific event appears in a set of another group of entity more frequently than expected by chance when examined with a background set of events related to the entity. And the basic underlying assumption that all events have an equal probability of being selected under the null hypothesis (H_0), by default null hypothesis is considered to be valid. The attempt is made to gather the evidence to reject the null hypothesis. However, there is a chance of dismissing null hypothesis incorrectly, and it is a *Type I Error*, which is a false positive study result. The level of probability of making a Type I error is called α, and the counter likelihood is *Type II Error* (β). To quantify the statistical significance of H_0, *p-value* is evaluated as an evidence. *p-value* is the probability of obtaining a result at least as extreme as the current one assuming H_0 is true. When the *p-value* is very small there is a disagreement between the data considered with the null hypothesis conversely higher the *p-value* fail to reject the null hypothesis. The evaluation of *p-value* is the enrichment analysis. Computing *p-value* value there are several methods *Fisher's exact test* is one of the method.

2.4.1.1 Fischer's exact test

Fisher's exact test proposed by Fisher [86], Agresti and Agresti [87] is a statistical significance test of independence when there are two nominal variables or the categorical data classifying objects in two different ways, Fisher test examines the significance of the association between two kinds of classification. Most common representation of Fisher test is a $2x2$ contingency table (Table 2.2). To represent Fisher's test mathematically, we take rows and columns marginal totals and the total represented by N. Thus, the frequency of the event that we are interested in is a. To evaluate the hypergeometric distribution gives the empirical probability of a random variable x which is represented by the frequency of a, such that b is the frequency of event A in the second population, c is the frequency of counter event A

in the first population, and d is the frequency of counter event A in the second population. Then the hypergeometric equation is given by equation 2.4.

	B	B'	$\sum$
A	f_a	f_b	f_A
A'	f_c	f_d	$f_{A'}$
$\sum$	f_B	$f_{B'}$	$f_a + f_b + f_c + f_d = N$

Table 2.2 Fisher's absolute frequency contingency 2x2 table

- f_a is the absolute frequency of event A in the first population represented in figure 2.5-ii
- f_b is the absolute frequency of event A in the second population represented in figure 2.5-iii
- f_c is the absolute frequency of counter event A in the first population represented in figure 2.5-iv
- f_d is the absolute frequency of counter event A in the second population represented in figure 2.5-v

$$P(x = a) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \tag{2.4}$$

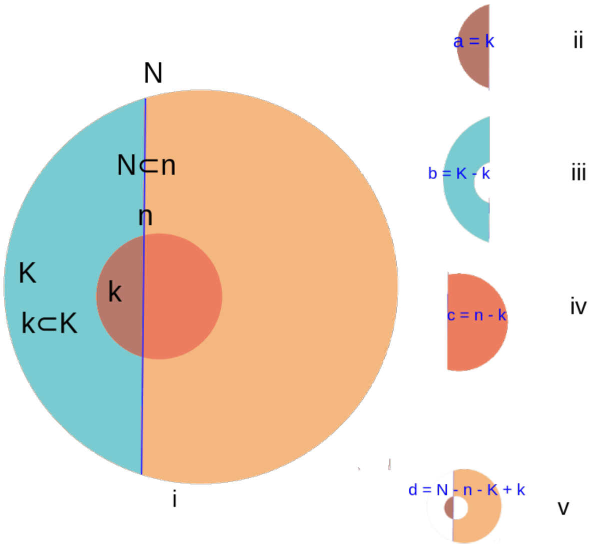


Fig. 2.5 Venn diagram representation of Fisher's exact test of the table 2.2

- N is the total population size
- K is the number of success states in the total population
- n is the number of samples that we check for success states
- k is the number of success states found in n

2.4.1.2 Multiple comparisons problem

As seen earlier in section 2.4.1 α is the level of significance an arbitrary value which determines whether *p-value* is low or high. Further, α implicates the probability of committing *Type I error*. With a selection of a hypothesis including a significant value without conclusion or proof, gives rise to the probability of false positives. The degree of false positives depends on the type of dataset, and the number of false positives could increase when considering simultaneous multiple statistical inferences giving way to *multiple comparisons problem* or *multiple testing problem*. The number of false positives is directly proportional to the number of inferences made. To control this group of false positive or the familywise error rate (FWER) several statistical methods are in use.

Bonferroni correction

As described in Goeman and Solari [88] the seriousness of the loss attained is inversely related to the number of hypotheses rejected. It is apparent to have a desirable error rate control which it termed as false discovery rate (FDR). Considering $H_1, H_2, \dots, H_m$ being family of hypotheses based on *p-values* $p_1, p_2, \dots, p_m$ arranged in increasing order of values. m being the total number of null hypotheses. Then Bonferroni correction rejects the null hypothesis (H_i) for each evaluated by equation 2.5. The procedure controls the FWER at $\leq \alpha$. Bonferroni procedure does not take into account dependency among *p-values* also how many null hypotheses are true, making it be a strict procedure for discarding hypotheses.

$$p_i \leq \frac{\alpha}{m} \tag{2.5}$$

Benjamini–Hochberg procedure

Benjamini and Hochberg [89] tries to address the problem in Bonferroni correction (Section 2.4.1.2) by introducing a parameter k and finding the most substantial value of it satisfying the equation 2.6 and then rejecting the null hypothesis for all $H_{(i)}$ where i is in $1, \dots, k$

$$p_{(k)} \leq \frac{k}{m}\alpha \tag{2.6}$$

2.4.2 Shannon entropy

Entropy or the information entropy is the amount of information content of the stochastic source of data. It is the expected bits of information contained in each message by covering all the areas of the information [90]. For example, to report drug-likeness information (Section 2.1.6) of five compounds set. The entropy of the message would be bits or *shannons* (named after Shannon, and it originates from Pauli and von Neumann) weighted average of total bits of information supplied among the possible messages (2 compounds satisfying or 3 compounds satisfying). To express the *shannons* of information, Shannon entropy ($\mathcal{H}$) proposes logarithmic formula (Equation 2.7). To formalise, let $\mathcal{A} = (A, p)$ be a discrete probability space such that $A = \{a_1, a_2, \ldots, a_n\}$ is a finite set each element with probability p_i then Shannon entropy ($\mathcal{H}(A)$) is given by equation 2.7.

$$\mathcal{H}(A) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2.7}$$

2.5 Databases in bioinformatics

The fast increasing data and information in the field of bio and cheminformatics needed a system which can efficiently manage the humongous data with every passing day generated by the domain. The system must also be smart and intelligent enough to access the information of gene sequences, amino acid sequences in proteins, motifs, domains in proteins, structural data from XRD, NMR, metabolic pathways, PPIs, CPIs, gene expression data from DNA microarrays and many new entities that are getting discovered. The database is a system in which data is an organized, structured, searchable, updated periodically. Biological databases serve the purpose of availability of systemized biochemical data and analysis of computed information is made easy. Features of databases are data heterogeneity, high volume of data, accepts curated scientific data, integration of large-scale data and sharing of it, and it is dynamic to requirements. The database can be classified based on accessibility, i.e., publicly available, available with copyright, browsable, downloadable, academic but not free, proprietary or with restricted access. Based on data sources databases can be a primary database which is an archival database where data is derived experimentally and are directly submitted such as nucleotide sequences and three-dimensional structures. Examples of primary database are Gen bank, UniProt, DDBJ. Secondary databases are also known as curated databases consisting of data derived from analysis of primary data sources such as sequences, secondary structures, etc. Technically databases are available in flat-files namely fasta files, relational database (Section 2.5.1) such as PostgreSQL (Section 2.5.2), object-oriented databases, eg XML database, ontology-based databases.

2.5.1 Relational databases

A relational database (RDB) is a set of multiple relations, where each relation is a dataset composed of *tables*, *records*, and *columns*. RDBs establish a clear relationship between database tables. Codd and F. [91] first proposed the relational data model in 1970. The significant advantage of relational model over classical model is its simple representation of data, and with ease, complex queries can be expressed. Structured query language (*SQL*) is the query language extensively used for creating, manipulating and querying RDBMS. Data definition language (*DDL*) is the subset of SQL supports to create, modify, delete tables and define their integrity constraints. DDL provides access rights to the tables and views. Data manipulation language (*DML*) is another subset of SQL which allows users to pose queries, insert, delete and modify rows. Execution of triggers which are tiny instructions or commands on databases gets executed when given preconditions met. DML consists of set-manipulation constructs which are useful when a set of rows are the result of the select query. *UNION*, *INTERSECT*, and *EXCEPT* are the operations supported under set-manipulation constructs. RDBMS can perform nested queries, correlated queries with the help of *set-comparison operators* (<, <=, =, <>, >=, >). Databases in the extension are used to do computation and summarization with the help of aggregate operators. Aggregate operators include *COUNT, SUM, AVG, MAX, MIN*. Construction of database is always a preplanned procedure and this preplan in this context is termed as schema. Any schema usually has multiple tables of definitions, *GROUP BY* clause extracts the data from multiple tables. Boolean operations in queries are performed with the help of *logical connectives* (AND, OR, and NOT).

Relational algebra is a formal query language for expressing relational model in RDB. Queries of algebra consists of collections of operators which are guided by the property that every operator in the algebra accepts one or two relation instances as arguments and returns a relation instance as the result. This rule makes it easy to compose complex queries in *relational algebra*. Basic operators of relational algebra are *selection* (σ), *projection* (π), *union* ($\cup$), *intersection* ($\cap$), *set-difference* (—), and *cross-product* (×).

msrid	kingdom	country	coll_date
7517	Plantae	Morocco	1998-01-01
3898	Plantae	Egypt	2010-04-01
3366	Animalia	Egypt	2003-06-01
5161	Plantae	Algeria	2005-04-01
1730	Plantae	Morocco	1990-07-01

Table 2.3 Sample sources instance S1 from NANPDB

msrid	kingdom	country	coll_date
3898	Plantae	Egypt	2010-04-01
5161	Plantae	Algeria	2005-04-01
3366	Animalia	Egypt	2003-06-01

Table 2.4 Result instance S2 for the expression 2.8

country	coll_date
Egypt	2010-04-01
Egypt	2003-06-01
Algeria	2005-04-01

Table 2.5 Result instance S3 for the expression 2.9

With the help of relational algebra considering the instance in table 2.3 we can retrive rows corresponding to rows with collection date after year 2000 with expression:

$$\sigma_{coll\_date>2000}(S1) \tag{2.8}$$

We acheive table 2.4 with the expression 2.8 and to project only few colomns, the expression 2.9 yields table 2.5

$$\pi_{country,coll\_date}(S2) \tag{2.9}$$

2.5.2 PostgreSQL - an open source database

Among available open-source databases PostgreSQL is one of them, and it is free. It is an object-relational database management system. PostgreSQL is also referred as *postgres* and the SQL used in PostgreSQL is *psql*. Management of concurrent transactions is efficiently implemented by multi-version concurrency control technique, which manages each transaction independently allowing changes to be made for each transaction are invisible to other transaction until complete changes are committed. With this feature deadlock, read locks are avoided and enable efficient atomicity, consistency, isolation, and duration (ACID) principles. PostgreSQL presents multi-level transaction isolation, which makes it avoid incomplete reads, upon a request for an uncommitted read transaction due to the isolation level it provides read committed transaction, and it obeys serializability. One of the important feature required for a database to be stable and reliable, PostgreSQL highly reliable and stable and can run without crashing for a longer duration. PostgreSQL is extensible its customizable features are completely available due to its open-source nature. PostgreSQL design can handle high volumes of the data environment. PostgreSQL incorporates built-in support for B-tree and hash indexes and four index access methods generalized search trees (GiST), generalized inverted indexes (GIN), Space-Partitioned GiST (SP-GiST) and Block Range Indexes (BRIN). In PostgreSQL, a schema holds all objects (except roles and tablespaces). Schemas efficiently act like namespaces, allowing objects of the same name to co-exist in the same database. By default, newly created databases have a schema called *public*, and additional schemas can be added if required however it is not mandatory. PostgreSQL assists a wide variety of data types. PostgreSQL is efficiently administered by pgAdmin a graphical user interface. In all the works in this thesis, we have used PostgreSQL version from 9.4 to 9.6. Version 9.6 is the first version which supports parallel querying. Incorporating PostgreSQL with Django framework (Section 2.5.3) is an easy process. We have used Django as the web framework for developing web resources (Sections 3.2, 3.3).

In PubMedPortable (Section 3.1), a combination of relational database and the full-text index is utilized to solve the problem of data accessibility. Full-text index builds complex boolean text queries, whereas a relational database stores all meta information of PubMed articles and their statistics. StreptomeDB 2.0 (Section 3.2) and NANPDB (Section 3.3) stores compound information and their relationship based on compound meta information. StreptomeDB 2.0 also stores prepared fingerprints with the help of OpenBabel [92] database wrapper. FragPred (Section 3.6) adopts the schema presented by ChEMBL and extends it to the requirement of fragments generated. The main goal of the design is to access the information with minimum redundancy in the data readily.

2.5.2.1 Psycopg

To the access the data of PostgreSQL database with Python an adapter is required and *Psycopg*[8] provides this bridge. Psycopg implements complete Python DB API 2.0 specification and Python high-level threading interface. It is desinged to perform for mult-threaded applications which can do several inserts and delete operations simultaneously. It completely supports Python 3.x environment. Using SQL syntax in Python is made easy by Psycopg. Django also supports Psycopg library.

2.5.3 Django framework

Django [93] is a python based free and open-source high-level framework. It follows model-view-template (MVT) architectural pattern and has a Model-View-Controller (MVC) design. The model is an abstract layer and is not the actual data. Its function is to access actual data from the database. During access of data Django does not need to know the complete architecture of the database. The models are reusable on different databases which obey its properties. The view represents the presentation layer of the model created during the development phase. The view can collect the user input, and serve it on to the web app presented via the browser. The controller controls the information flow between model and view. The logic defined in the program determines what kind of information has to be taken from the database and which data has to be queried to the user. The controller implements logic by either changing view or modifying the model or by both. In the presentation, layer data is displayed via templates using the data presented by view via the model. The Django core framework includes:

- Standalone web server for development and testing
- The unique nature of differentiating server environment and development environment is most useful. However, it is sometimes cumbersome.
- Caching framework, useful in serving several images

[8] http://initd.org/psycopg/

- Fast environment as it uses python completely

The QuerySet API is an efficient use of accessing the data from databases building very complex queries with very few lines of code. QuerySet is another datatype which has iteration, slicing, pickling as evaluating features. Methods of QuerySet are of two types one which returns a new QuerySet and another does not return. It has field lookups such as exact, startswith, year, etc. Aggregate functions such as expression, Min, Max, etc. Lastly, query related tools which collect objects for QuerySet.

- QuerySet can iterate over data presented by the database.
- Slicing is possible with QuerySet, in which when an unevaluated QuerySet gets another unevaluated QuerySet, then Django executes a database query and return a limited list according to the slicing parameter.
- Pickling is possible in Django it is a precursor to caching. Pickling forces data to be loaded into the memory before it gets used.

Complete implementation of projects StreptomeDB, NANPDB are in Django Framework. NANPDB additionally uses Mezzanine [94] CMS on top of Django framework. Django 1.5 to 1.11 versions were used in all projects of this thesis.

Chapter 3

Results

The section introduces and explains the specific works in detail. The cheminformatic tools presented in the thesis are useful for creating hypotheses for the discovery of novel drugs to certain diseases, and metabolic disorder *diabetes mellitus* (DM) illustrates such hypothesis as a case study. Diabetes mellitus, commonly referred to as diabetes is the sixth most affected disease [95]. DM is a chronic disease which occurs when the pancreas does not produce enough insulin, or it can even occur when the body cannot effectively use the insulin which it produces, i.e., when the cells of the body are not responding properly to the insulin produced. *Hyperglycaemia* is a disorder in which consistently high blood sugar levels are present for a sustained period. When diabetes exists for over a time, it leads to serious damage and causes many complications to the body's systems, especially the nerves and blood vessels. Acute complications such as *diabetic ketoacidosis*, *hyperosmolar hyperglycemic state* are major determinants for the hospitalization of diabetic patients [96]. Hyperglycaemia and diabetes are the major causes of mortality which gives rise to cardiovascular disease in patients. The risk of CVD almost doubles [97] in patients with diabetes compared to patients who are not affected by diabetes [98]. Other complications include *ischemic stroke* to the brain; chronic kidney disease is a state in which kidneys functions are abnormal. In diabetic foot ulcer, wound healing is an innate ability of the human body gets disturbed. Diabetic retinopathy is a multifactorial microvascular disease that is also caused by chronic hyperglycemia and subsequent adverse metabolic sequelae [99]. DM can be mainly of three types:

- **Type 1 DM:** This occurs due to deficient in insulin. In this state, the pancreas fails to produce sufficient insulin thereby it is called as insulin-dependent diabetes mellitus (IDDM) or juvenile diabetes. Recent studies show there is a rise in Type 1 DM [100].
- **Type 2 DM:** It occurs due to excess body weight and physical inactivity. In this state, cells fail to respond to insulin adequately. Subsequently, non-insulin-dependent diabetes mellitus (NIDDM); a form of diabetes in which insulin production is inadequate, or the body becomes resistant to insulin.
- **Gestational diabetes:** It occurs in the woman who is in pregnancy period. She develops hyperglycemia, i.e., similar to Type 2 DM. During this state the risk of pre-eclampsia, depression and requiring Caesarean section increases. It is the degree of glucose intolerance during pregnancy first recognition [101].

There were several works and statistics are done in the domain of DM. There is an increased mortality rate due to direct and clinical sequelae of CVD and diabetes. Fasting and nonfasting blood glucose data were collected and examined in COX models, a correlation between risks of CVD and normal fasting glucose is detected. These associations are independent of age, sex, and region of the population. In conclusion, significant benefits observed by lowering blood glucose levels of at least to 4.9 mmol/l [98]. A collaborative analysis of Diagnostic Criteria in Europe (DECODE) study has confirmed that people with impaired fasting glucose (IFG) and impaired glucose tolerance (IGT) has a high risk of death from CVD which varies between an average person and diabetic groups [102]. The DECODE data also interpreted that 2-h plasma glucose concentration after a 75-g OGTT is an independent predictor of CVD mortality, while prediction of CVD death by FPG is mostly due to correlation with 2-h PG concentrations [97] and these concentration levels were superior in Asian populations. Higher-than-optimum blood glucose concentration attributes to IHD and stroke diseases; the work used the population distribution of FPG to measure blood glucose. The data was collected extensively from 52 countries from individual-level records from surveys, systematic reviews, and data provided by investigators. 84% CVD deaths were in low and middle-income countries among which 55% were IHD and 28% were the stroke, and 17% were coded to diabetes directly. Having hemoglobin A_{1c} less than 5% is the normal range and are a lower risk of CVD. However, an increase in hemoglobin A_{1c} by 1% point increases the average relative risk of death by any cause is by 1.24 with 95% CI in men and goes up to 5.01 when more than 2% point rises. And in woman average relative risk increases by 1.28 with 95% CI and climbs up to 6.91 when hemoglobin A_{1c} increases by more than 2% death due to any cause of reason. These relative risks are independent of age, BMI, waist-to-hip ratio, systolic blood pressure, serum cholesterol concentration, cigarette smoking and history of CVD [103].

In the recent research related there is a rise in search of synthetic drugs for anti-diabetic agents [104]; further, it discusses interactions between synthetic compounds targets associated with DM. The association of significant adverse effects of existing medications triggered for therapeutic strategies to adopt synthetic medications as inhibitors to reduce blood glucose levels [105–107] and also indicates the significance of natural sources. A *sulfonylurea* class of organic and antidiabetic compound [108] exhibited its ability to alleviate DM associated vascular complications. StreptomeDB (Section 3.2) and NANPDB (Section 3.3) being natural compound databases in sections 3.2.4, 3.3.4 example pipelines are shown how these natural sources can be helpful for discovery of drugs against this disease.

3.1 PubMedPortable: Text mining framework

Information science in the field of the biopharmaceutical is growing exponentially, leading and crucial information retrieval from ever-increasing literature is becoming increasingly entangled. This complexity is because of the scientific results which are produced as a means of textual publication and requires text mining techniques to search for needed findings. There have been several enhancement and advances using computational methods for increased sharing and redistribution of experimental setups and their results. However, there are suggestions that these enhancements will not cease scientific publication growth. Equivalently, digitization of preliminary findings and clinical experiences data with structured data models happening, still the wealth of medical records unstructured using natural methods, conventional methods [109]. Natural language processing and text mining applications can expedite in the domain of biomedical science.

The significant barriers in NLP and text mining applications are broadly divided into two categories. One, the problem of interoperability between NLP components, i.e., identifying specific words or the entities and classifying them in groups, for example, groups such as identification of protein-protein interactions [110], drug-drug interactions [111], compound-protein interactions [112], and their aid to treat diseases [113]. An NER task is to map named entities to concepts in vocabulary. Performance of later experiments depends on the mapping of objects, but due to unavailability of standards, this step is complicated. Secondly, research literature data should be reachable, this metamorphoses into the difficult task as there is an accelerated growth of publications articles in biomedical science.

In the PubMedPortable publication [114] which is majorly developed by Dr. Kersten Döring[1] we tried to address the mentioned problems, *PubMedPortable* became useful in accessing the subject required data. Tools employed to use this data are the relational database (Section 2.5.1), full-text index of all the required text from the PubMed. The tools can be utilized to acquire the results of complex boolean queries. Combination of both relational database and full-text index gets used for collecting statistics of the scientific data.

[1] http://phabi.de/main/members/

PubMed as of mid of 2017 consists more than 27.3 million records stretching back to 1966, also including selective records to the year 1809. Each year approximately 500,000 new records are added. Among the listed records 48% are listed with their abstracts, and 52% articles have links to full-text. About 2.5% of the records in PubMed correspond to diabetes. Published, and unpublished approaches are available to deal with such a significant amount of data, and quite a few also provide a BioC interface. BioC is a simple XML format for sharing biomedical text, their annotations, libraries which can read and write any format presented to it [115, 116], this enhances interoperable tools for NLP and makes the tool to be modular which further can be used with least investment on new systems and languages. Similar web service tools which provide support for the BioC interface are OntoGene [113]. BioC approach is called as the *minimalist* approach.

3.1.1 Processing of PubMed data

The basic idea of PubMedPortable is to be able to work in a local machine conveniently, which leads to the prerequisite of garnering PubMed data into the local machine. In PubMedPortable various methods are developed to download data from NLM FTP server as XML files of the specific subject related term [117]. The XML dataset is also possible to download as a single file and can later be broken chunks of files as per the requirement of the hardware availability. For the low and medium range of hardware, the downloaded XML file is split into smaller pieces of data which gets processed quickly.

The downloaded XML files data is stored into the database which will be more beneficial to perform various queries as compared to processing with XML files. There are various tools available which can be used to parse XML files such as LingPipe project, and to build Lucene based full-text index [118]. Argo [119] is a web-based, text-mining workbench. Argo also provides a repository of configurable analytics which accepts and produce interchangeable formats of data. Text mining works are the main application of Argo. Within the workbench, there is a possibility of workflow construction. It used an old and widely used architecture called, Unstructured Information Management Architecture (UIMA) [120–122] approach in which it uses the benefits of its assets such which are organizational structure, skills alignment, development inefficiencies, and speed of the product. This architecture was introduced by IBM research to have a common software architecture for developing. Mainly two classes are observed in the analysis of unstructured information, namely document-level and collection-level analysis. Text Analysis Engines (TAEs) does the document-level analysis. These are analogous to another widely used architecture called General Architecture for Text Engineering (GATE) [123, 113]. A TAE is a recursive structure which can either consist of sub or component engines each performing a different stage of analysis in both GATE or the UIMA architecture. UIMA gets messy quickly with its complex capability. GATE evolved from mere text processing tool to an extensive software system for developers, language engineers from various fields. It has the strong footprint and wide range of lifecycle in the field of text analysis systems. Its major characteristics included scalability, flexibility, and robustness [109]. For efficient indexing, search and access it possess augmenting full text, annotating graphs, and structuring data based on knowledge-based ontology. However, due to all these capabilities, complexity supersedes. BioC uses a minimalistic approach, as only the data format of XML files is defined in a document type definition (DTD) file, an additional file describing the semantics of data and annotations serves as user-specific properties. The interoperability is ensured within the BioC workflow, defining an Input Connector to read and an output Connector to write BioC XML data. The interface to these BioC classes is implemented in several programming languages [116, 124]. Producing a BioC-compatible XML document is much simpler than a UIMA-compatible module [124].

3.1.1.1 BioC interface

The premier advantage of using BioC interoperability is that it can facilitate a wide range of available BioC tools. Some of those tools are NER tools from PubTator [125, 122]. These tools can identify genes/proteins, chemicals/species, mutations/variations, taxonomy, diseases. The minimalistic approach of BioC facilitates crowdsourcing projects an example is mark2cure crowdsourcing project another example can be gene tagger [126]. The BioC workflow as shown in figure 3.1 shows how any tool supporting input and output format can add annotations to a document, supporting the idea of interoperability. In PubMedPortable MeSH terms are added to BioC XML document from the local relational database. The BioC

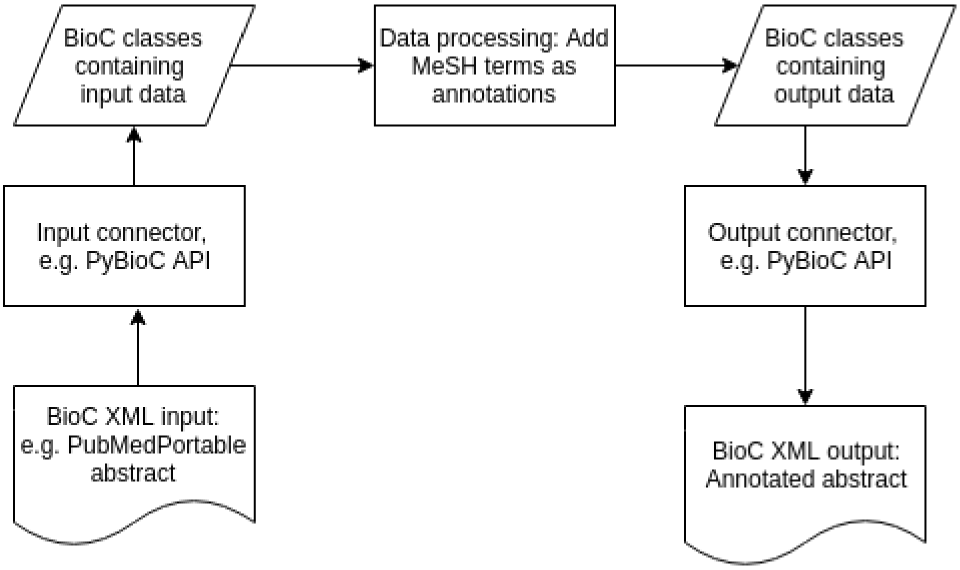

BioC classes containing input data
Data processing: Add MeSH terms as annotations
BioC classes containing output data
Input connector, e.g. PyBioC API
Output connector, e.g. PyBioC API
BioC XML input: e.g. PubMedPortable abstract
BioC XML output: Annotated abstract

Fig. 3.1 **General BioC workflow:** Minimalistic approach [116] combining with example from PubMedPortable [114]. The example shows how to add MeSH terms to BioC PubMed titles and abstracts from the PubMedPortable PostgreSQL database.

data model describes a data model using an XML DTD, and it avoids dependency on used language features providing a standardized file format, familiar to the biomedical domain.

3.1.1.2 Employing tools and accessing the data

The requirements of PubMedPortable to be employable are the installation of Python, Xapian, and PostgreSQL (Section 2.5.2) and recommended in Linux based Operating systems. Tested on Ubuntu and Fedora operating systems and their installation methods are described on GitHub project page[2] respectively. A disk space of around 400 GB which will be useful to process complete PubMed XML files. The basic workflow of PubMedPortable represented in figure 3.2. PubMedPortable operates with step by step code snippets execution and is usable via a command-line interface which requires PubMed XML files as input. A database needs to be created and configured at first. The schema subsequently is built on it.

[2] https://github.com/PhaBiFreiburg/PubMedPortable

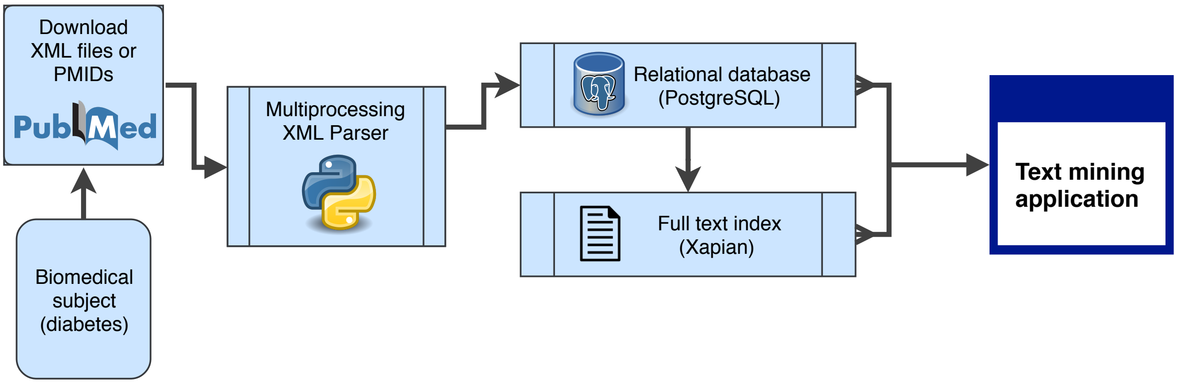

Download XML files or PMIDs
PubMed
Biomedical subject (diabetes)
Multiprocessing XML Parser
Relational database (PostgreSQL)
Full text index (Xapian)
Text mining application

Fig. 3.2 **PubMedPortable workflow:** 1) Download XML files of diabetes from PubMed. 2) Parse the with PubMedPortable parser and with the help of PubMedPortable import tool, import data into local PostgreSQL relational database. 3) Build Xapian full text index. 4) Apply to text mining applications.

In 2004, Oliver et al. [127] compared different approaches in loading PubMed XML files into a relational database, using NLM provided DTD. There are multiple ways by which a schema can be designed, however, while designing schema optimization of speed is considered as a parameter. Priority to minimizing the number of lookups to the database over repetition of information is the choice made to optimize the speed. The work experimented with open-source relational database PostgreSQL (Section 2.5.2), due to its free availability and modifiable property. However, for the final implementation relational databases Oracle 9i and an IBM's DB2 8.1, combined with code available in Java and PERL programming languages. There is an unpublished update version from 2010 using Java 6 and MySQL 5.1 [128]. In PubMedPortable, this SQL schema is wholly adapted and slightly modified, however, combined with object-relational mapping (ORM) in Python to generate a PostgreSQL database from PubMed XML files. Which facilitates, changes to SQL tables or columns can be introduced directly in the parser itself; this accelerates the importing process according to the hardware CPU cores availability. The database schema is available in appendix A.1. Filling of data into tables from PubMed is done by the process of importing XML files using a SAX parser. Subsequently, Xapian full-text index is built by querying abstracts, titles, keywords, MeSH terms, and chemical substances mapped to their data sources from the relational database. Complete setup can also be executed at once using the virtual container

Docker without installing additional software packages. With the finished data setup the subject user-defined terms phrase searches, boolean searches, statistics and their charts are implemented. Explanation of detailed installation, configuration, execution procedures is on GitHub project page[3] [129].

3.1.2 Diabetes mellitus dataset

Downloaded complete DM data as a single 6.5 GB XML data file. Import of records after 1225 splits chunks each consisting of 500 entries into the PubMedPortable database with hardware comprising of Intel Core i5-3570 CPU having four processors at 3.4GHz along with 32GB memory is utilized and tested to store into the PostgreSQL. It took 82 minutes to parse the complete DM dataset. A view of PubMed diabetes corpus is in table 3.1.

Descriptor	#	Descriptor	#
Articles	612507	Distinct authors*	1184462
Abstracts	479473	Grant agencies	844
Databanks	29	Chemical substances	28205
Languages	42	Major mesh names	12718
Gene symbols	587	Not major mesh names*	22541
Countries	113	Years	1788 to 2018
Journals*	11158		

Table 3.1 Properties of DM dataset in PubMedPortable. *Entries which are not clean given number is maximum

In the publication case study it has been shown to extract chemical compound entities names using the tmChem [130] and PubTator web service [125]. Gene and protein synonyms can be excerpted using GeneTUKit [131]. While here in this thesis an attempt is made to find out alternatives to the existing diabetes indications. For this purpose, DrugBank [132] has been selected to get existing compounds related to diabetes. DrugBank yielded 67 such drugs (most commonly occurring top 10 list of drugs is provided in table 3.2). To

[3] https://github.com/PhaBiFreiburg/PubMedPortable

Rank	DrugBank ID	Drug Name	Occurances	Gene	Occurances
1	DB00331	Metformin	11491	PPAR	4434
2	DB01132	Pioglitazone	2723	ACE	4167
3	DB00412	Rosiglitazone	2512	DB	3818
4	DB01124	Tolbutamide	2496	VEGF	3011
5	DB01016	Glyburide	2443	AMPK	2489
6	DB00047	Insulin Glargine	2226	INS	2122
7	DB06655	Liraglutide	1625	OB	1983
8	DB01261	Sitagliptin	1604	GLUT4	1913
9	DB00284	Acarbose	1361	RAGE	1388
10	DB01276	Exenatide	1337	CCL2	1051

Table 3.2 Top 10 DM drugs and genes from DrugBank [132] and UniProt [133] respectively and their number of occurances in PubMed using PubMedPortable [114]

generate a word cloud first the subject is chosen it can be any one of Gene or protein, Disease, Chemical name from any of the databases as per requirement here in this case Genes are taken from UniProt [133], for chemical name entities DrugBank [132] is selected. Subsequently, recognition of entities by their identifiers to summarize the number of entries. Logarithmic values express the frequency of occurrence of each counted entities.

Word cloud (Figure 3.3) of DrugBank DM shows *Metformin* is most frequently occurred drug and in genes (Figure 3.5) *PPAR* occurred most frequently. An exact number of occurrences of all the selected DrugBank drugs is in appendix A.1. Metformin is the drug of choice to treat type 2 diabetes [134], and it prevents type 2 diabetes mellitus [135]. Metformin function is to decrease hepatic glucose production by inhibiting gluconeogenesis(GNG) a metabolic pathway that results in the generation of glucose [136]. On the other hand Peroxisome proliferator-activated receptor gamma (*PPARγ*) is the key glucose regulator, it is a vital pharmacological target against metabolic disorders [137]. And the literature gives the information combining both the data, i.e., Metformin and *PPARγ*. Metformin upregulates *PPARγ* [138, 139] by elevating AMP-activated protein kinase(AMPK) phosphorylation. Pie chart 3.4 shows the number of DM publications country wise, which gives the impression of the amount of research happening related to diabetes in indicated countries. The results of DM obtained with PubMedPortable are used later in the section 3.2.4.

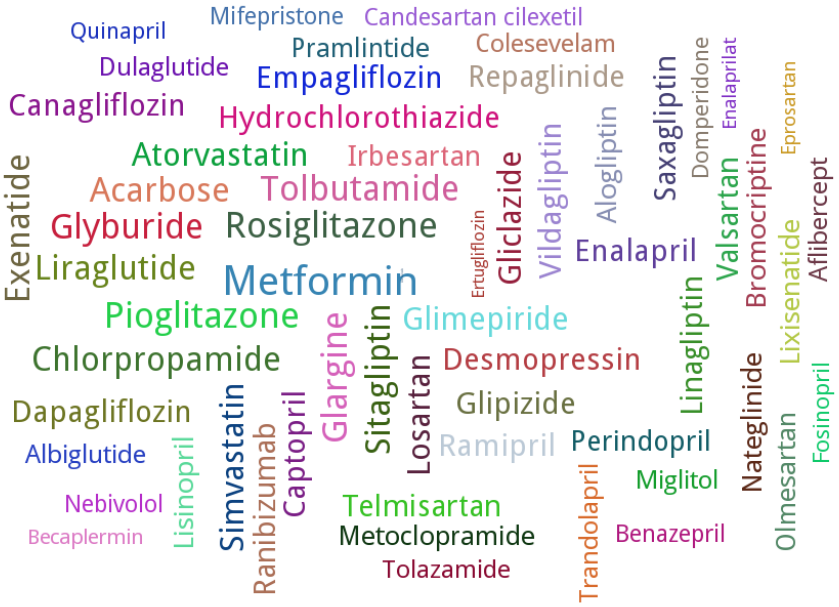


Fig. 3.3 Word cloud of DrugBank DM compound generated using PubMedPortable

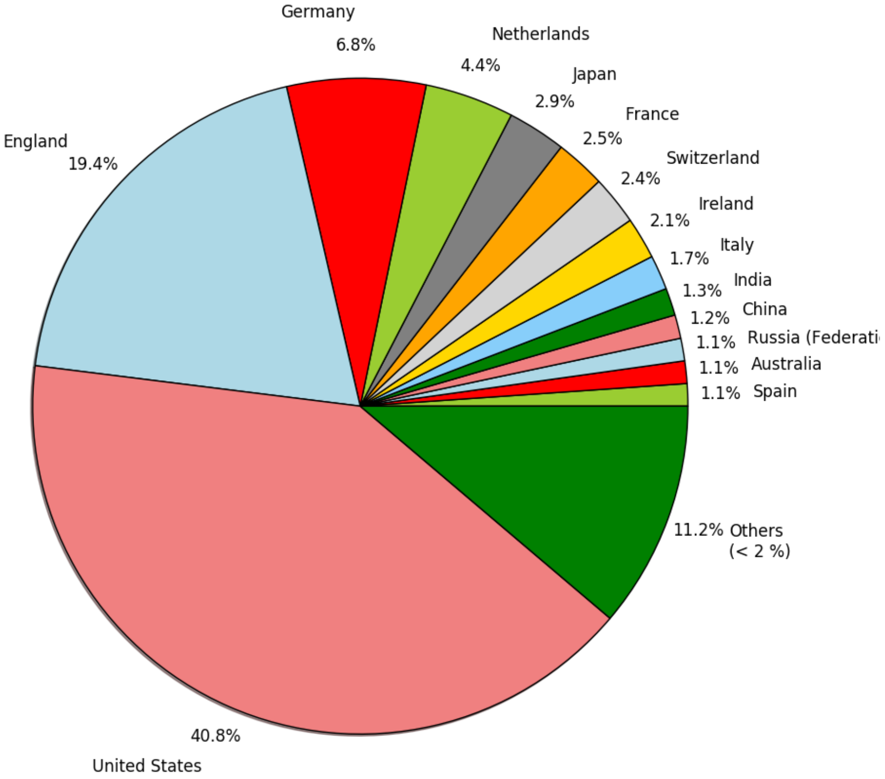


Fig. 3.4 Pie chart indicating number of DM publications country wise

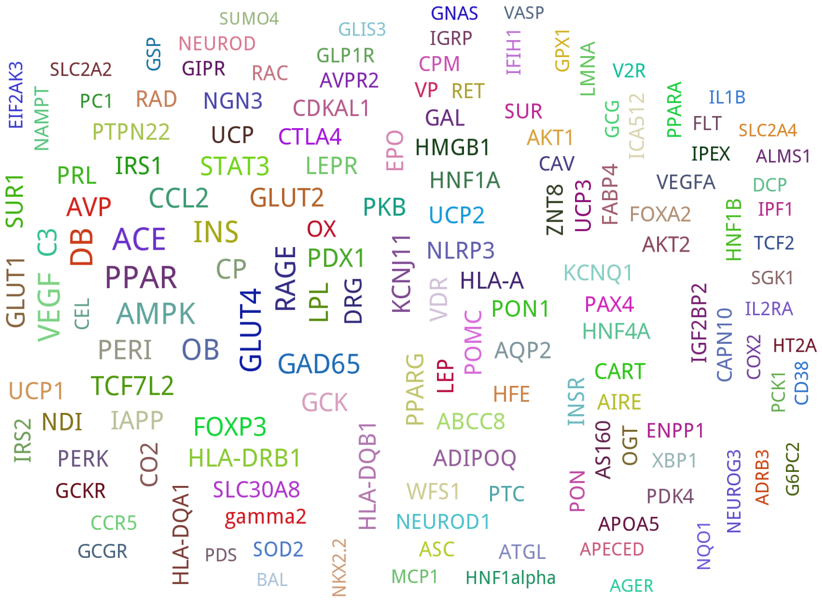


Fig. 3.5 Word cloud of Uniprot DM genes generated using PubMedPortable

The knowledge of genes and proteins related to DM would give an understanding of the function and expression of genes [140], and it is a motivation to create hypotheses for research. Gene terms about DM have been extracted from UniProt, subsequently with PubMedPortable again we can form the word cloud displaying the frequency of occurrences of genes. The word cloud (Figure 3.5) of genes shows most frequently occurred genes in DM literature of PubMed. The selection of gene names for PubMedPortable synonyms is from UniProt [133]. Although occurrences of genes do not mostly vary, the table 3.2 illustrates the top 10 genes and an extensive list of genes is in appendix A.2.

From the table 3.2 it can be seen that drug Metformin mostly occurs. To find most frequently co-occurring words in texts that contain the search term Metformin a word cloud (Figure 3.6) was generated.

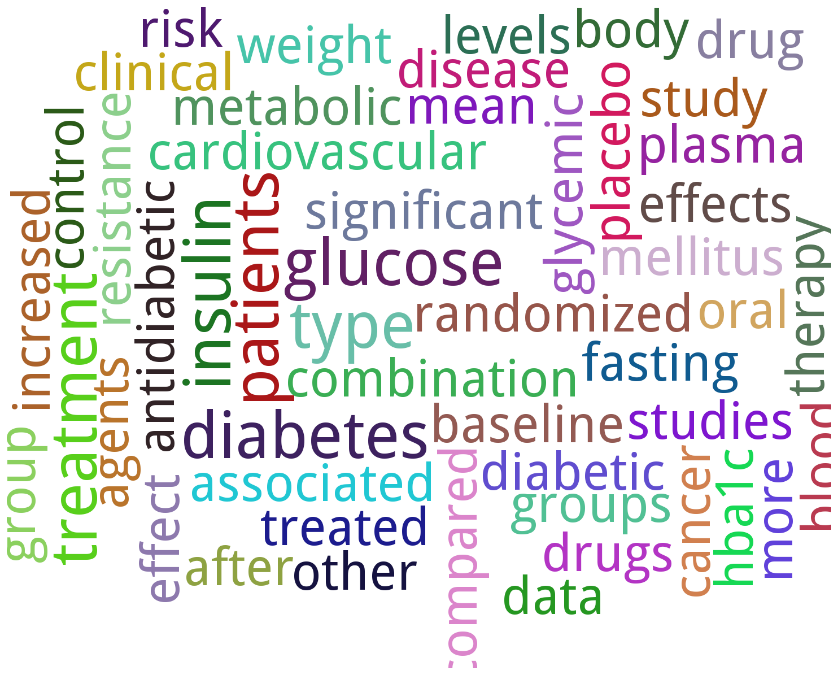


Fig. 3.6 Word cloud for 50 most frequently occurred co-occurances of Metformin

3.2 StreptomeDB 2.0

StreptomeDB initially published [141] in the year of 2012 with the then available more than 2400 unique and distinct compounds from more than 1900 *Streptomycetes* bacterial strains and substrains. *Streptomyces* genus is vital for the production of natural bioactive compounds such as an antitumor, antibiotic or immunosuppressant drugs. In StreptomeDB 2.0[4] [22] and update of the earlier version, we have presented several new compounds along with features such as Scaffold browser, Phylogenetic tree and experimentally predicted information such as NMR- and MS- spectra of thousands of compounds.

3.2.1 Largest database of *Streptomycetes*

Literature in PubMed (Figure 3.7) indicates a massive interest in the research community about the bacteria resulted from genus *Streptomyces* which are essential for the production of drugs, antitumor, immunosuppressant, antifungal or natural antibiotic compounds. On an average 354 publications (Figure 3.7), every year indicates the enormity of engagement around the *Streptomyces* bacteria; statistics suggest that over 66% of known natural antibiotics are from *Streptomyces* bacteria [142, 143]. *Streptomyces* is a genus of Gram-positive actinobacteria. The phylum *Actinobacteria* of the order *Actinomycetales* is one of the largest taxonomic units within recognized lineages domain of *Bacteria*. *Actinobacteria* upon *Gram stain* test produces positive test confirming their more receptive nature to antibiotics, and they are within high range (70%) of G+C content in their DNA [144]. The order *Actinomycetales* are nearly all from terrestrial soils, with distinct lineage living primarily as *saprophytes*, showing labeled chemical and morphological diversity [145].

[4] http://phabi.de/streptomedb2/

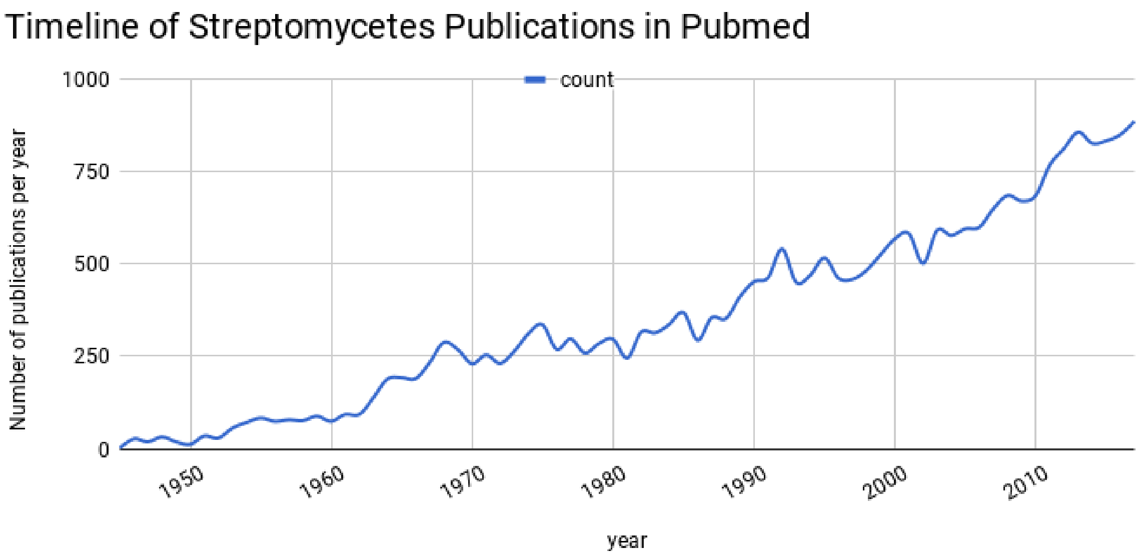


Fig. 3.7 Graph indicating the number of articles published in PubMed from 1945 to 2017

The abundant usage of the StreptomeDB which is published and made freely available to the scientific community in the year 2013 motivated us to facilitate more features and update the comprehensive database with the new discovered compounds and organisms. The compounds in the latest StreptomeDB 2.0 update [22] have risen to over 4000 compounds and host organisms to beyond 2500 along with content collected from thousands of abstracts and full papers by text mining methods and extensive manual curation by members of our group. More information about the latest StreptomeDB 2.0 content is in table 3.3. For curation purpose, our group members have programmed CoRSCurator software which is Java-based standalone software which annotates abstract titles and full texts along with background information such as compound name, source organism, biological activity, and synthesis routes. Dr. Kersten Döring [5] has improved CoRSCurator software further to obtain these results and export the collected data into PostgreSQL database (Section 2.5.2). Subsequently, PostgreSQL stores the data with determined schema and relationships as described in appendix B.1.

[5] http://phabi.de/main/members/

Descriptor	#	**Descriptor**	#
Compounds	4040	Scaffolds	4485
Organisms	2584	Organism Relationships	6717
Synthesis Pathways	12	Synthesis Pathway relationships	731
Activities	905	Activity relationships	3813
References	5486	Compounds with MS spectrum	1945
NMR spectrum	3989	Gene clusters	251
Number of genomes	693		

Table 3.3 Statistics of StreptomeDB 2.0

StreptomeDB 2.0 [22] being an update to the initially published database thus we decided to preserve the significant part of the data architecture of the database and more focus was given to the collection of new data of compounds and their completeness. The names of several compounds were not found in the original references and were also not found in PubChem service. Therefore they were manually painted and are stored in the database. Besides, web service also includes phylogenetic tree feature. Dr. Dennis Klementz [6] implemented the function, which bases on *16S rRNA* sequences, holds more than two-thirds of the incorporated host organisms.

[6] http://phabi.de/main/members/

3.2.2 Scaffold browser

As part of the new feature, we have introduced scaffold-based navigation system[7], which facilitates the investigation of diverse chemical compounds of StreptomeDB based on their structure to understand the character of the library. Additionally, implementation of Phylogenetic tree feature by Dr. Dennis Klementz [8], enhances the scaffold browser in their evolutionary context. Dr. Xavier Lucas [9] decomposed StreptomeDB compounds into their staged scaffolds with the subsidiary tool Canvas 2.3 provided by Schrödinger, LLC, NY, USA [146]. As discussed in section 2.1.8 scaffold representation of Canvas closely resembles the Murcko framework. Canvas first identifies unique scaffolds within the available set of structures. Similarly, it obtains the scaffold by stripping side-chain except for exocyclic and exolinker double bonds. Further, the Schrödinger tool [147] generates subscaffolds breaking scaffolds exhaustively. Subsequently for subscaffolds, for instance, R1, R2, and R3 are the ring systems breaking the scaffold R1-R2-R3 into R1-R2, R2-R3, R1, R2, and R3 subscaffolds. Depending upon the direction of hierarchical scaffold traversal it is either termed as subscaffolds or superscaffolds. Scaling of scaffolds is linear and quadratic depending upon the number of structures analyzed and unique scaffolds contained in those structures respectively. When many different backbones exist in the library, it would be a worst-case scenario consisting of a large number of independent structures. The rich diversity of chemical scaffolds produced by *Streptomycetes* allows their use as precursors for many semi-synthetic therapeutic approaches. An example of scaffold browser is in figure 3.22. StreptomeDB 2.0 has up to 22 scaffold levels at each scaffold level number of scaffolds, and related compounds are listed in table 3.4.

A SQL partition function (Listing 1) is applied to extract compounds of the selected scaffold from the scaffold grid or the scaffold browser. The SQL query (Listing 2) gets the number of super scaffolds of a selected scaffold or multiple scaffolds.

[7] http://132.230.56.4/streptomedb2/scaffold_grid/

[8] http://phabi.de/main/members/

[9] http://phabi.de/main/members/

```
;WITH queryscaffolds(qs) AS (
VALUES %s)
SELECT COUNT(DISTINCT(c.compound_id))
FROM (
        SELECT scaffold_id, compound_id, COUNT(scaffold_id) OVER
                (PARTITION BY compound_id) AS cnt
                FROM (
                        SELECT DISTINCT scaffold_id, compound_id
                        FROM web.compounds_to_scaffolds
                        WHERE scaffold_id IN (SELECT qs FROM queryscaffolds)) t ) res
JOIN web.compound c ON c.compound_id = res.compound_id
WHERE res.cnt = (SELECT COUNT(qs) FROM queryscaffolds);
```

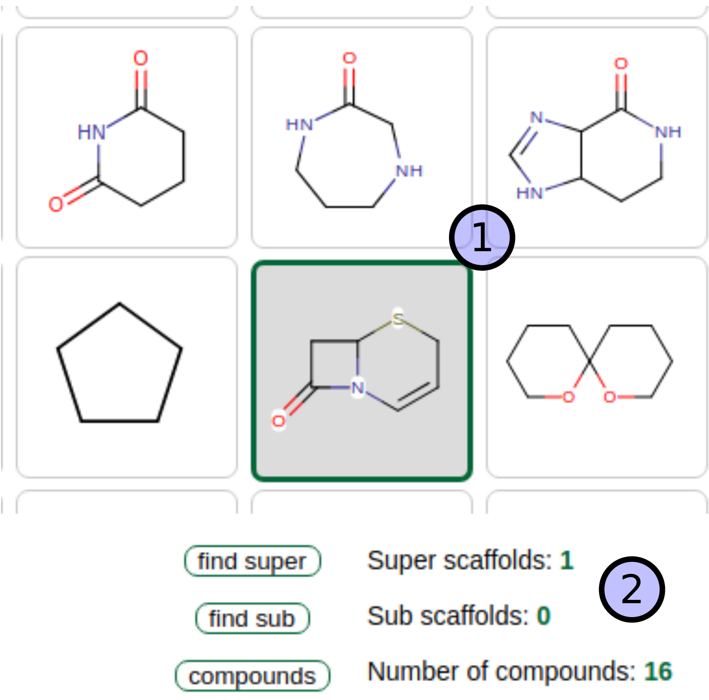
Listing 1 : SQL Query paritition function to get compounds of the selected scaffold in scaffold grid such as figure 3.8

3.2.2.1 Usage of scaffold browser

Scaffold browser plays a vital role in finding novel drugs from NPs. From the distinct set of molecules, scaffold acts as a uniting factor. For better understanding an example *β-lactam antibiotic* which is a class of broad-spectrum antibiotics, i.e., working on both gram-positive and gram-negative bacteria [148]. Broad-spectrum antibiotics are particularly useful when multiple groups of bacteria are suspected to be infected. It will be interesting to know how many compounds contain *β-lactam* antibiotic. Scaffold browser (Figure 3.8) can give us such a kind of information. In StreptomeDB 2.0 there are 16 compounds (Figure 3.9) which contain *β-lactam* scaffold.

```
;WITH queryscaffolds(qs) AS (
VALUES %s)
SELECT MIN(d.level)
FROM (
        SELECT scaffold_id, super, COUNT(scaffold_id) OVER (PARTITION BY super) AS cnt
        FROM (
                SELECT DISTINCT scaffold_id, super
                FROM web.super
                WHERE scaffold_id IN (SELECT qs FROM queryscaffolds)) t ) res
JOIN web.scaffolds s ON s.scaffold_id = res.super
JOIN web.detect d ON d.scaffold_id = res.super
WHERE res.cnt = (SELECT COUNT(qs) FROM queryscaffolds);
```

Listing 2 : SQL Query paritition function to get number of super scaffolds available for the selected single or multiple scaffold in scaffold browser such as figure 3.8



Super scaffolds: 1

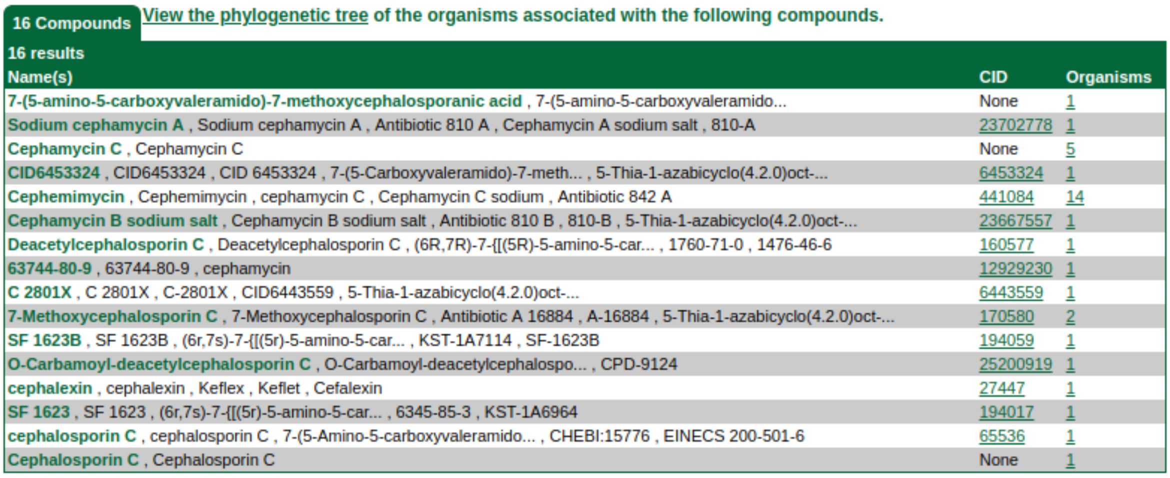
Sub scaffolds: 0

Number of compounds: 16

Fig. 3.8 Use case of scaffold browser: (1) selecting β-lactam scaffold. (2) Showing number of super, sub scaffolds available to β-lactam along with number of compounds containing the selected scaffold

Scaffold level	# of scaffolds	# of compounds	Scaffold level	# of scaffolds	# of compounds
0	1032	3707	12	65	6
1	732	1619	13	53	4
2	672	1048	14	35	5
3	559	709	15	22	4
4	469	478	16	14	2
5	314	270	17	9	2
6	213	177	18	6	2
7	137	115	19	2	1
8	96	51	20	1	1
9	91	30	21	1	1
10	84	21	22	1	1
11	72	7			

Table 3.4 Number of Scaffolds and compounds at each scaffold level in StreptomeDB 2.0



16 Compounds View the phylogenetic tree of the organisms associated with the following compounds.

16 results

Name(s)	CID	Organisms
7-(5-amino-5-carboxyvaleramido)-7-methoxycephalosporanic acid , 7-(5-amino-5-carboxyvaleramido...	None	1
Sodium cephamycin A , Sodium cephamycin A , Antibiotic 810 A , Cephamycin A sodium salt , 810-A	23702778	1
Cephamycin C , Cephamycin C	None	5
CID6453324 , CID6453324 , CID 6453324 , 7-(5-Carboxyvaleramido)-7-meth... , 5-Thia-1-azabicyclo(4.2.0)oct-...	6453324	1
Cephemimycin , Cephemimycin , cephamycin C , Cephamycin C sodium , Antibiotic 842 A	441084	14
Cephamycin B sodium salt , Cephamycin B sodium salt , Antibiotic 810 B , 810-B , 5-Thia-1-azabicyclo(4.2.0)oct-...	23667557	1
Deacetylcephalosporin C , Deacetylcephalosporin C , (6R,7R)-7-{[(5R)-5-amino-5-car... , 1760-71-0 , 1476-46-6	160577	1
63744-80-9 , 63744-80-9 , cephamycin	12929230	1
C 2801X , C 2801X , C-2801X , CID6443559 , 5-Thia-1-azabicyclo(4.2.0)oct-...	6443559	1
7-Methoxycephalosporin C , 7-Methoxycephalosporin C , Antibiotic A 16884 , A-16884 , 5-Thia-1-azabicyclo(4.2.0)oct-...	170580	2
SF 1623B , SF 1623B , (6r,7s)-7-{[(5r)-5-amino-5-car... , KST-1A7114 , SF-1623B	194059	1
O-Carbamoyl-deacetylcephalosporin C , O-Carbamoyl-deacetylcephalospo... , CPD-9124	25200919	1
cephalexin , cephalexin , Keflex , Keflet , Cefalexin	27447	1
SF 1623 , SF 1623 , (6r,7s)-7-{[(5r)-5-amino-5-car... , 6345-85-3 , KST-1A6964	194017	1
cephalosporin C , cephalosporin C , 7-(5-Amino-5-carboxyvaleramido... , CHEBI:15776 , EINECS 200-501-6	65536	1
Cephalosporin C , Cephalosporin C	None	1

Fig. 3.9 Compound results for β-lactam in StreptomeDB 2.0 arriving through scaffold browser (Figure 3.8). Figure shows 16 compounds resulting the search, along with their PubChem links and number of source organisms

3.2.3 MS spectra in StreptomeDB 2.0

The software CFM-ID generated MS values for the StreptomeDB 2.0 molecules. Typically identification of metabolite is by comparing the observed MS (Section 2.1.3) with the reference spectra available in the database and then ranking the candidates based on their closeness to the target under observation. The limitation of this method is lack of ample information in the databases. CFM-ID as proposed in Allen et al. [149] uses competitive fragmentation modeling (CFM) an alternative computation method. CFM computes of a probabilistic generative model for the electrospray tandem spectrometry fragmentation (ESI-MS/MS) is used to predict the fragmentation spectrum of molecules which are ≤30 heavy atoms in positive ionization mode at 10, 20, and 40V. Also in CFM for bombarding collision-induced dissociation (CID) is usually employed. It intentionally fragments molecules into smaller parts to analyze their structure. StreptomeDB 2.0 reports maximum five intense peaks, mapping the resulting peaks to their fragment structures at each given intensity level. An example of MS-spectra values shown in figure 3.10.

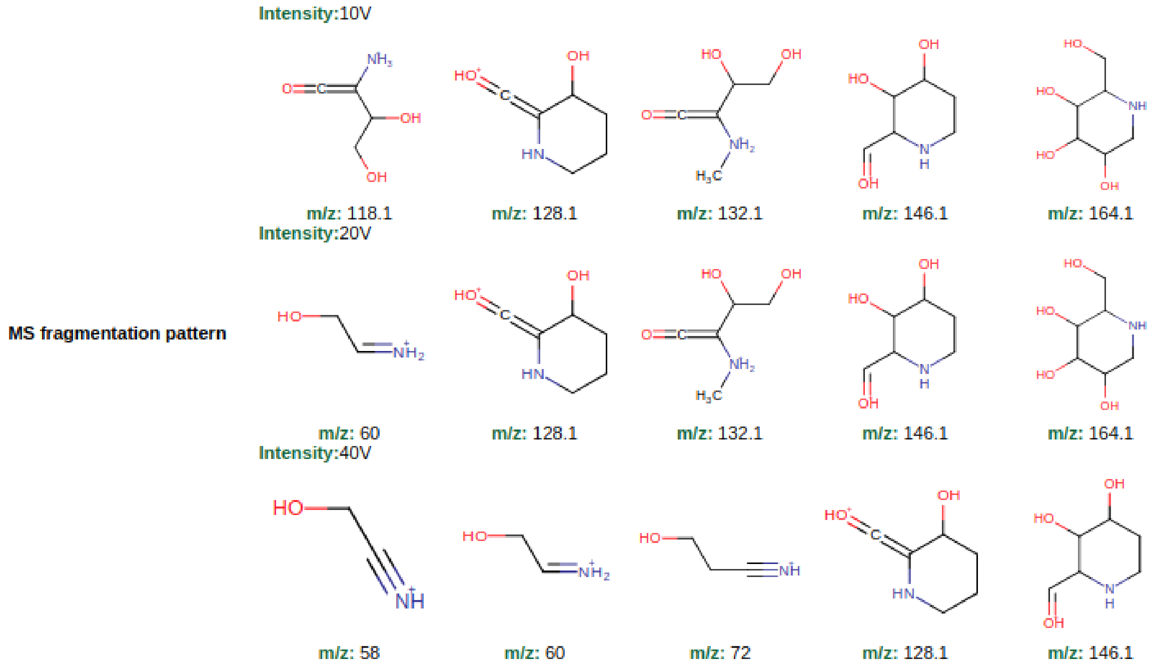


Fig. 3.10 StreptomeDB 2.0 MS-Spectra results for *1-DEOXYNOJIRIMYCIN*

Query	#	Notes
Number of DrugBank DM compounds	67	listed in table A.1
Number of StreptomeDB compounds	3991	available in StreptomeDB 2.0
Number of Fingerprints used	28	fingerprints table 2.1
Number of comparisons with all Fingerprints	5697720	Between DrugBank DM compounds and StreptomeDB 2.0 compounds
Number of comparison pairs with Tanimoto >0.85	4769	Fingerprint wise distribution listed in table 3.6
Number of unique comparison pairs with Tanimoto >0.85	4582	Top 10 pairs according to number of fingerprints predicted as similar listed in table 3.7

Table 3.5 Information generated with the molecular similarity between DrugBank DM and StreptomeDB 2.0 [22] compounds

3.2.4 DM compounds in StreptomeDB 2.0

In continuation of diabetes dataset from section 3.1.2 and the drive for the search of inhibitors associated with DM. The basic step of similarity between DrugBank DM compounds and StreptomeDB 2.0 presented information listed in the table 3.5.

The similarity distribution of the comparisons between above sets (DrugBank DM and StreptomeDB 2.0) with the several different fingerprints (Table 2.1) exhibited that there were very few comparisons which are similar to DrugBank DM compounds. It revealed 4769 comparisons (Table 3.5) were similar with Tanimoto coefficient >0.85, filtering almost 99.998% comparisons. The Kernel density estimation of Tanimoto coefficient for some of the chosen significant fingerprints (significant based on Figure 2.3) can be interpreted from the graph (Figure 3.11). Kernel density estimation is available for all the fingerprints in the appendix, figure B.4.

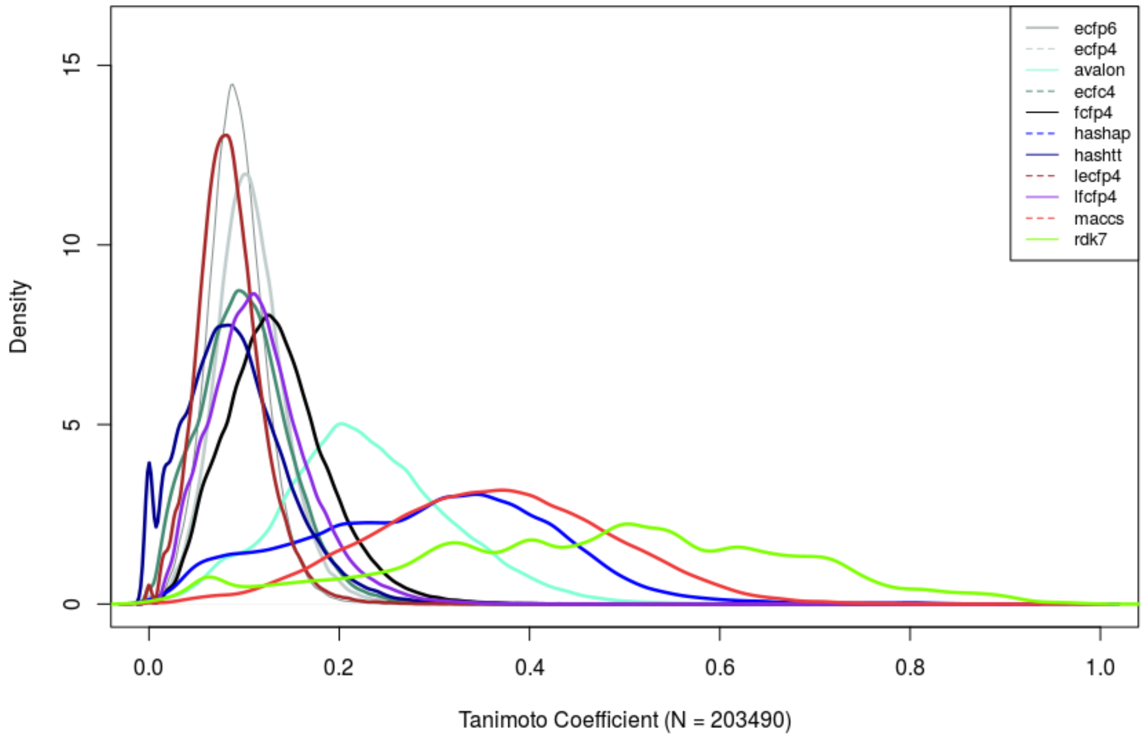


Fig. 3.11 Graph showing Kernel density estimation of tanimoto coefficient between DrugBank [132] DM compounds and StreptomeDB 2.0 [22] compounds. N indicates the number of comparisions with each fingerprint

The graph (Figure 3.11) showing very few comparisons having above 0.85 Tanimoto coefficient for different fingerprints. Table 3.6 presents the data of the graph (Figure 3.11). The graph has only those fingerprints which have shown better performance (Section 2.2.2). Observing the kernel density graph and the fingerprint wise distribution table fingerprints, i.e., TT, LECFP4, FCFP6, ECFP4, ECFP2, ECFC4, ECFC2, AP, LFCFP6, LFCFP4 which have shown better performance estimated only two comparisons which are above 0.85 Tanimoto coefficient. On the other side, it would be interesting to know which comparisons were estimated by multiple fingerprints. And the table 3.7 shows top 10 such pairs according to the number of fingerprints evaluated above 0.85 Tanimoto coefficient.

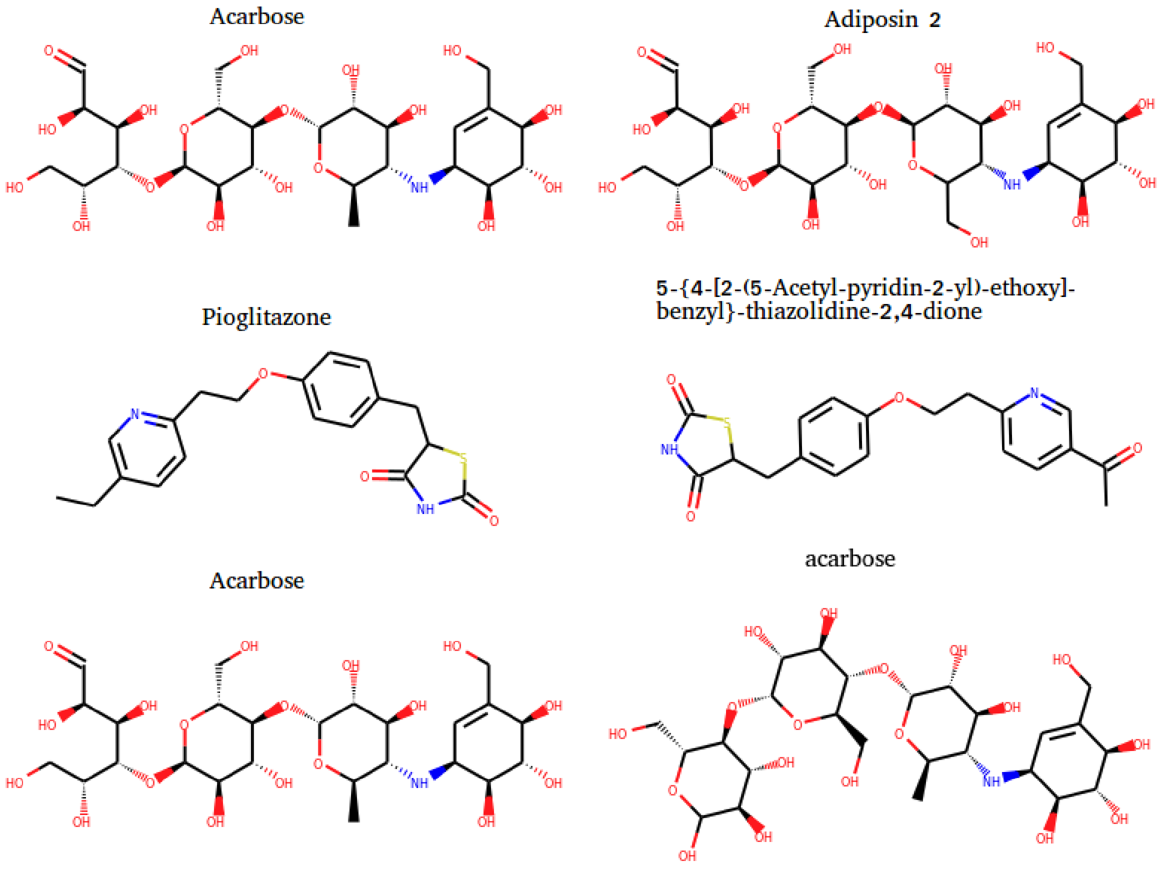
It is apparent that *Acarbose* is dominating the table (Table 3.7) and also *Acarbose* is available in the StreptomeDB 2.0 with ID number 778. *Acarbose* is an anti-diabetic drug for which

Fingerprint	# tanimoto >0.85	Fingerprint	# tanimoto >0.85
rdk7	4344	fcfc6	2
ecfp0	141	fcfp4	2
hashap	126	lfcfp4	2
fcfp2	46	lfcfp6	2
maccs	28	ap	1
rdk5	18	ecfc2	1
rdk6	18	ecfc4	1
avalon	9	ecfp2	1
fcfc2	7	ecfp4	1
laval	7	fcfp6	1
fcfc4	4	lecfp4	1
hashtt	3	tt	1
ecfc0	2		

Table 3.6 Fingerprint wise distribution of comparisions between DrugBank DM compounds and StreptomeDB 2.0 compounds which have tanimoto > 0.85

DM DrugBank ID	DrugBank drug name	StreptomeDB ID	StreptomeDB compound name	# of fingerprints predicted
DB00284	Acarbose	3313	Adiposin 2	25
DB01132	Pioglitazone	5208	CHEMBL1268	15
DB00284	Acarbose	778	acarbose	9
DB00284	Acarbose	3310	Trestatin B	7
DB00284	Acarbose	3311	Trestatin A	7
DB00284	Acarbose	3769	Trestatin C	6
DB00284	Acarbose	4506	acarviostatin I03	6
DB00284	Acarbose	4507	acarviostatin II03	5
DB00284	Acarbose	9258	isovalertatin M23	5
DB00035	Desmopressin	5079	Aspartocin	4

Table 3.7 Top 10 pairs according to number of fingerprints predicted as similar



Acarbose

Adiposin 2

Pioglitazone

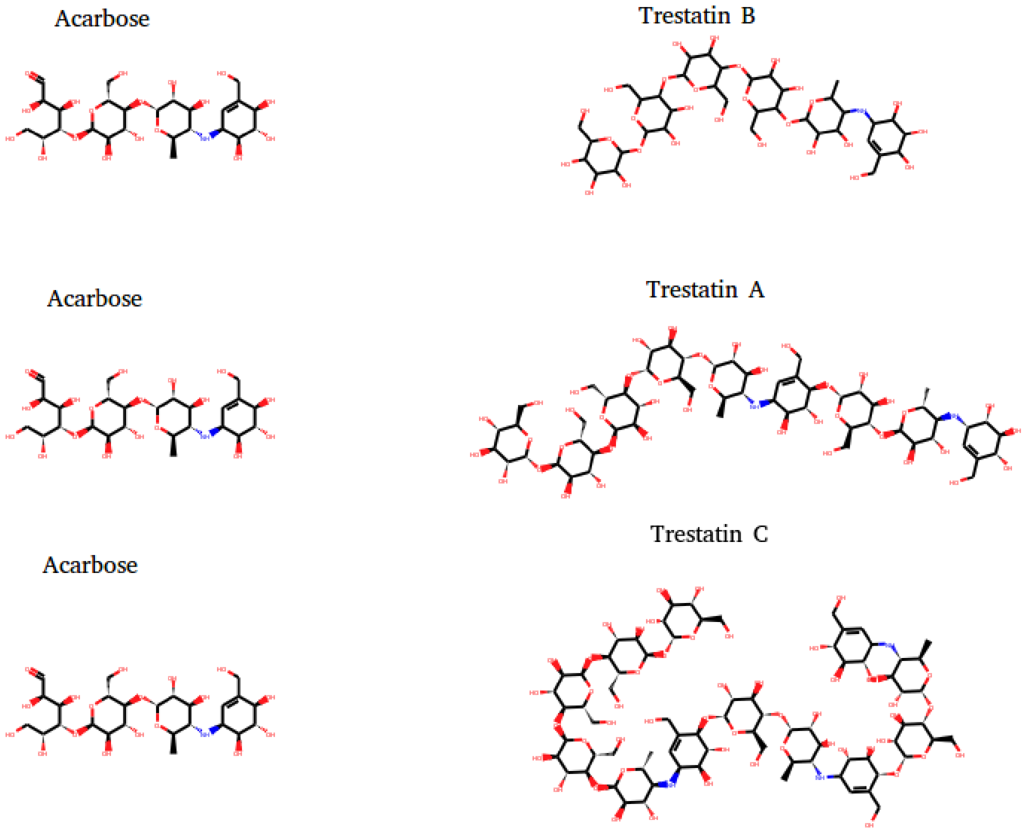
5-{4-[2-(5-Acetyl-pyridin-2-yl)-ethoxy]-benzyl}-thiazolidine-2,4-dione

Acarbose

acarbose

Fig. 3.12 Structures of top 10 pair compounds which are listed in table 3.7. DrugBank compounds (left), StreptomeDB 2.0 compounds (right). Figure: 1/4

16 similar StreptomeDB 2.0 compounds were detected, which have tanimoto coefficent greater than 0.85 compared with several DrugBank DM compounds. Indicating its most probable antidiabetic bioactivity. Combining PubMedPortable information (Table 3.2) and StreptomeDB 2.0 more information can be gathered that *Acarbose* is one of the top 10 compounds which has substantial literature with 1361 occurances of *Acarbose*. From the table 3.2 it can be observed that *Acarbose* has 1361 occurrences of publications. From the table 3.7, we can also see more StreptomeDB 2.0 compounds *Adiposin2*, *Trestatin B*, *Trestatin A*, *Trestatin C* and others. The structures of the top 10 comparisons (Table 3.7) can be seen in figures (3.12, 3.13, 3.14, 3.15). The FragPred (Section 3.6) tool also predicted a probable target for the compound *Trestatin C* it can be seen in table 3.25.



Acarbose

Trestatin B

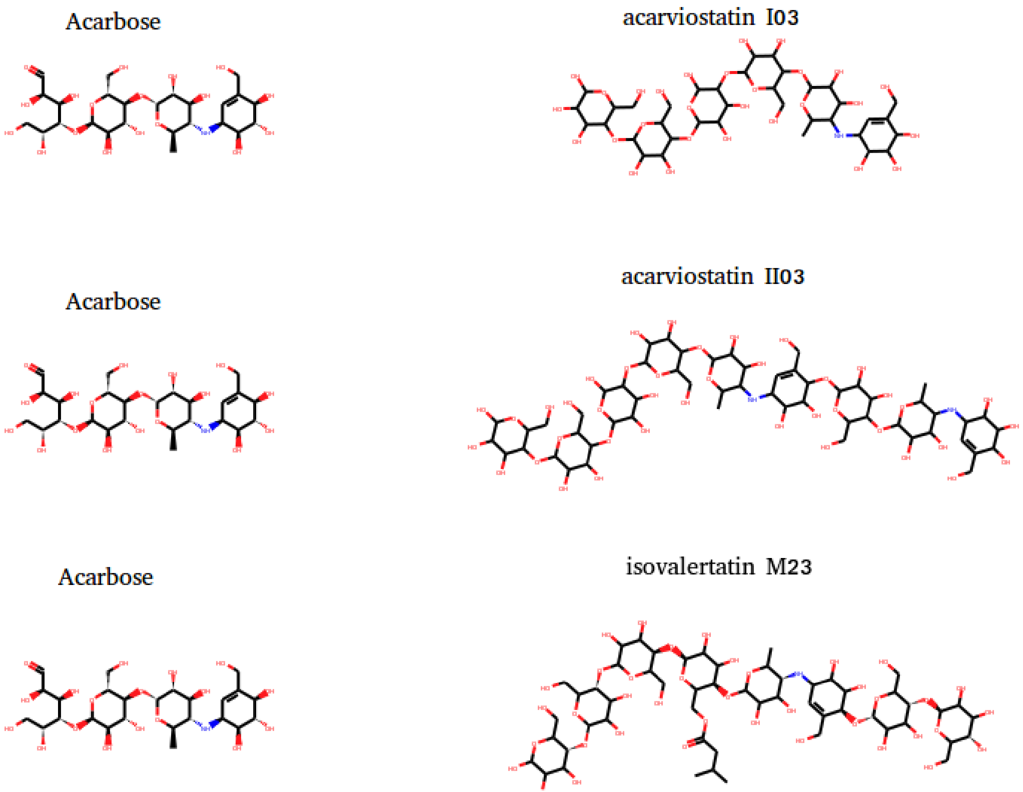
Acarbose

Trestatin A

Acarbose

Trestatin C

Fig. 3.13 Structures of top 10 pair compounds which are listed in table 3.7. DrugBank compounds (left), StreptomeDB 2.0 compounds (right). Figure: 2/4


Acarbose

acarviostatin I03

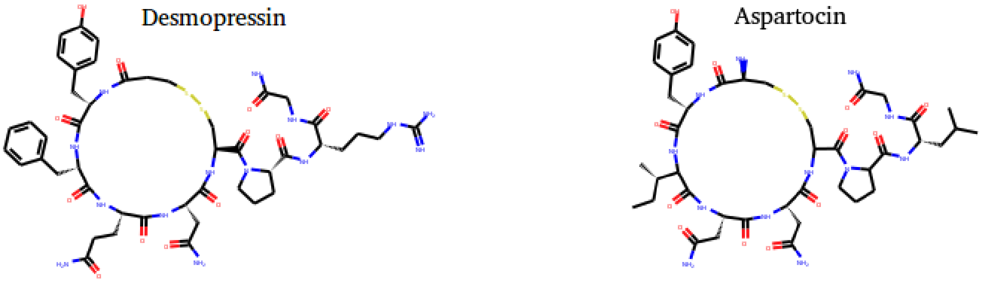
Acarbose

acarviostatin II03

Acarbose

isovalertatin M23

Fig. 3.14 Structures of top 10 pair compounds which are listed in table 3.7. DrugBank compounds (left), StreptomeDB 2.0 compounds (right). Figure: 3/4


Desmopressin

Aspartocin

Fig. 3.15 Structures of top 10 pair compounds which are listed in table 3.7. DrugBank compounds (left), StreptomeDB 2.0 compounds (right). Figure: 4/4

With the help of PubMedPortable a word cloud be generated for the 50 most frequently occuring co-occurances of *Acarbose* (Figure 3.16) compound. With the word cloud it can be understood that diabetes, cardiovascular, glucose are the frequently occuring terms assuring that *Acarbose* is a diabetic compound. Further, StreptomeDB 2.0 provides compound card (Figures 3.17, 3.18, 3.19, 3.20, 3.21, 3.22) of *Acarbose*[10] containing several useful information about the compound.

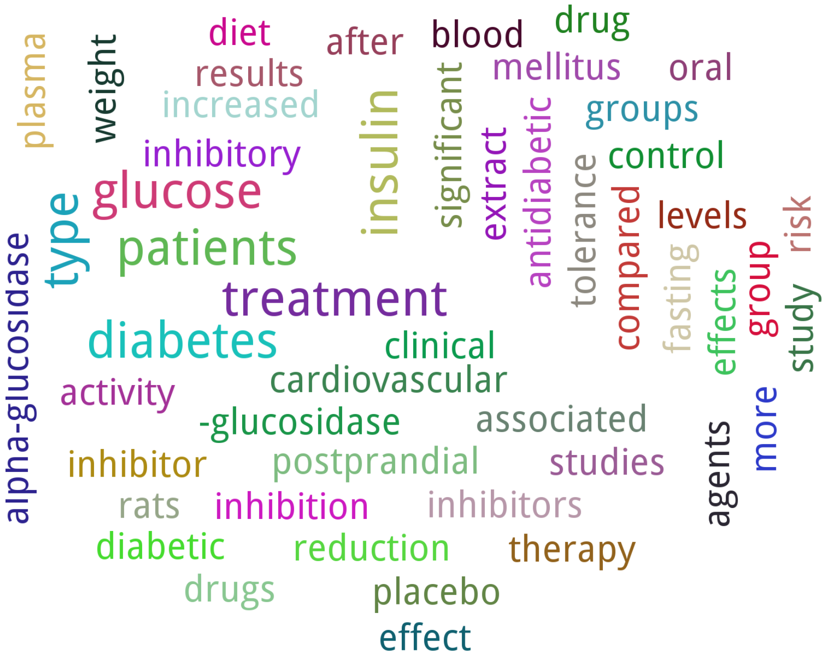
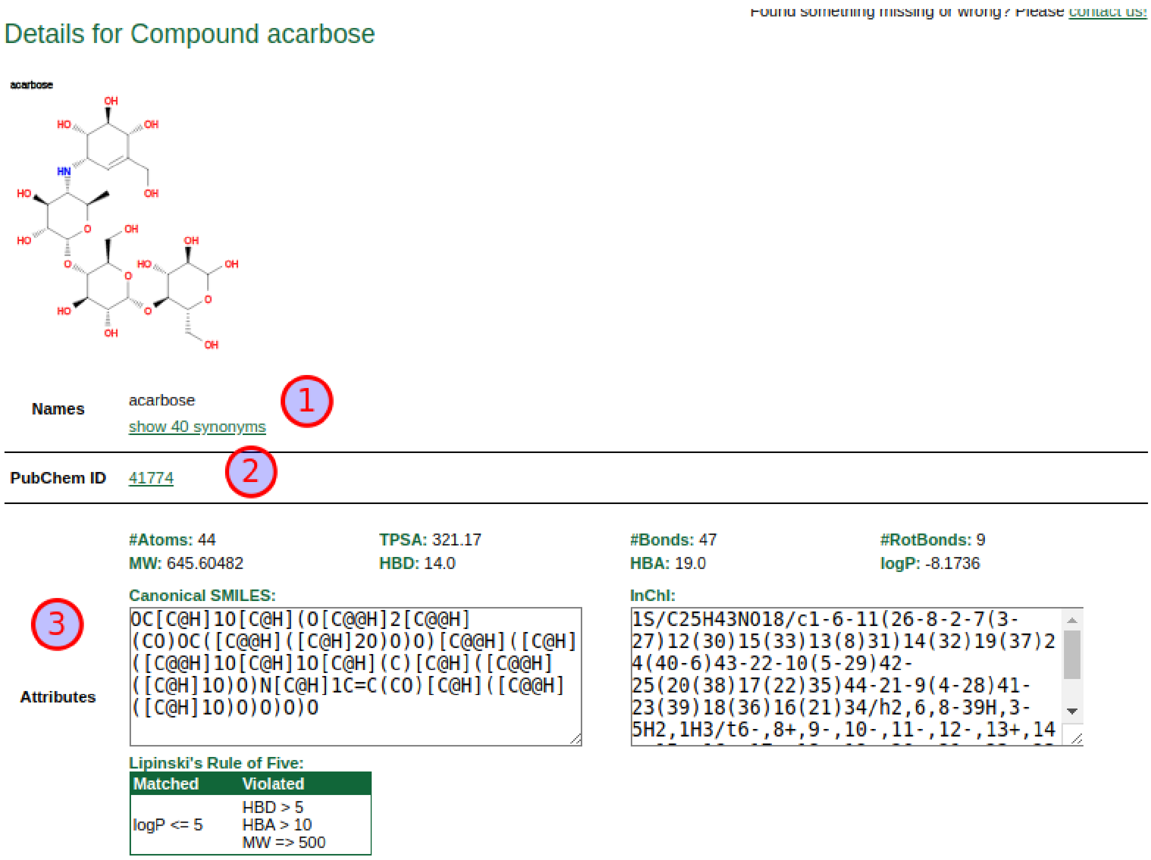


Fig. 3.16 Word cloud for 50 most frequently occurred cooccurances of Acarbose in the literature corpus generated (Section 3.1.2)

[10] http://132.230.56.4/streptomedb2/get_drugcard/778/



Found something missing or wrong? Please contact us!

Details for Compound acarbose

Names acarbose
show 40 synonyms

PubChem ID 41774

Attributes

#Atoms: 44 TPSA: 321.17 #Bonds: 47 #RotBonds: 9
MW: 645.60482 HBD: 14.0 HBA: 19.0 logP: -8.1736

Canonical SMILES:

```
OC[C@H]1O[C@H](O[C@@H]2[C@@H]
(CO)OC([C@@H]([C@H]2O)O)O)[C@@H]([C@H]
([C@@H]1O[C@H]1O[C@H](C)[C@H]([C@@H]
([C@H]1O)O)N[C@H]1C=C(CO)[C@H]([C@@H]
([C@H]1O)O)O)O)O
```

InChI:

```
1S/C25H43NO18/c1-6-11(26-8-2-7(3-
27)12(30)15(33)13(8)31)14(32)19(37)2
4(40-6)43-22-10(5-29)42-
25(20(38)17(22)35)44-21-9(4-28)41-
23(39)18(36)16(21)34/h2,6,8-39H,3-
5H2,1H3/t6-,8+,9-,10-,11-,12-,13+,14
```

Lipinski's Rule of Five:

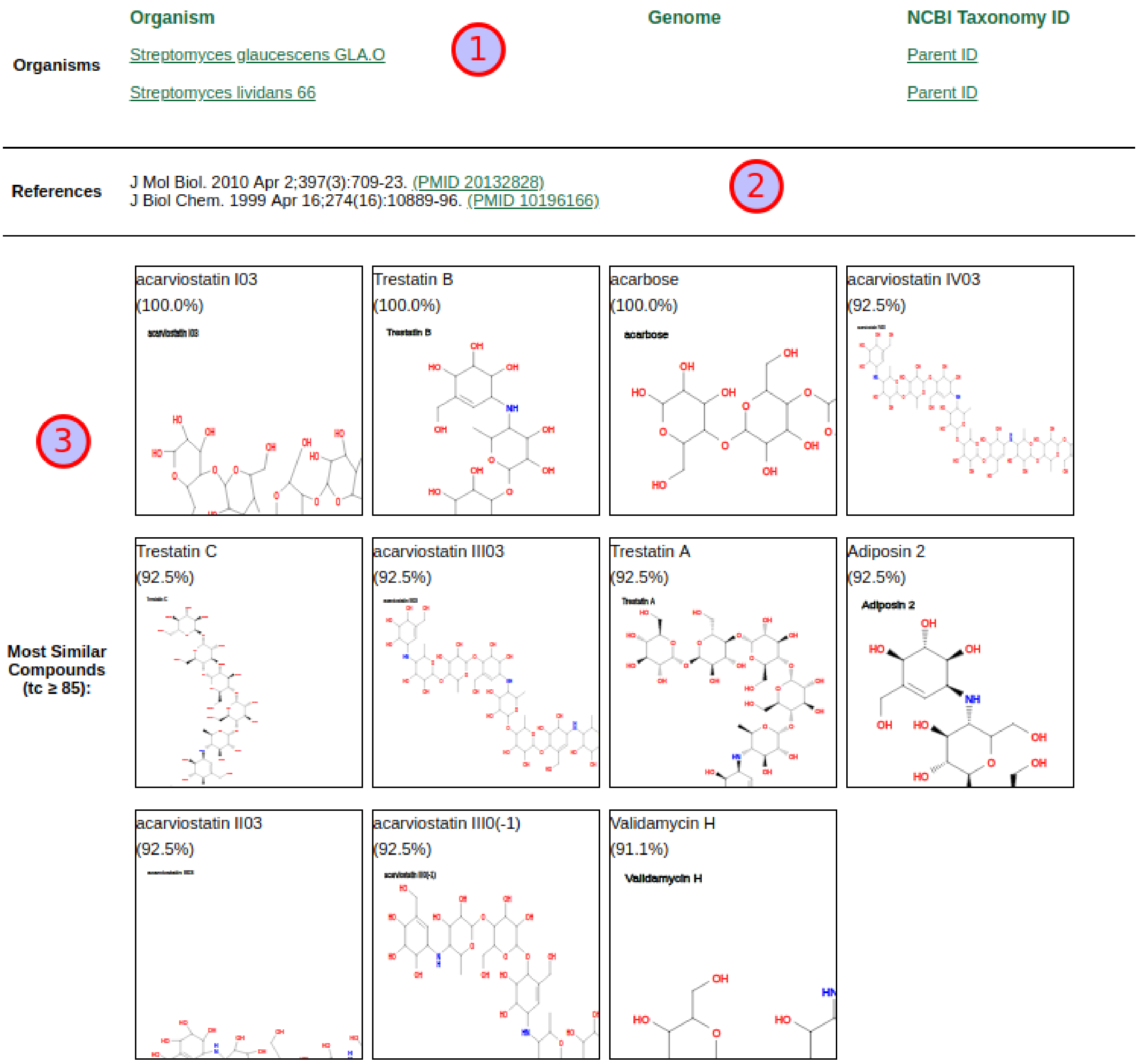
Matched	Violated
logP <= 5	HBD > 5 HBA > 10 MW => 500

Field 1: In this field, *Acarbose* image and link to its synonyms.

Field 2: PubChem ID [150] if available is provided.

Field 3: Some of the important properties and descriptors are listed.

Fig. 3.17 StreptomeDB 2.0 results for *Acarbose* providing its structural information, compound card: 1/5



Organisms	Organism	Genome	NCBI Taxonomy ID
	Streptomyces glaucescens GLA.O		Parent ID
	Streptomyces lividans 66		Parent ID

References

J Mol Biol. 2010 Apr 2;397(3):709-23. (PMID 20132828)
J Biol Chem. 1999 Apr 16;274(16):10889-96. (PMID 10196166)

Most Similar Compounds (tc ≥ 85):

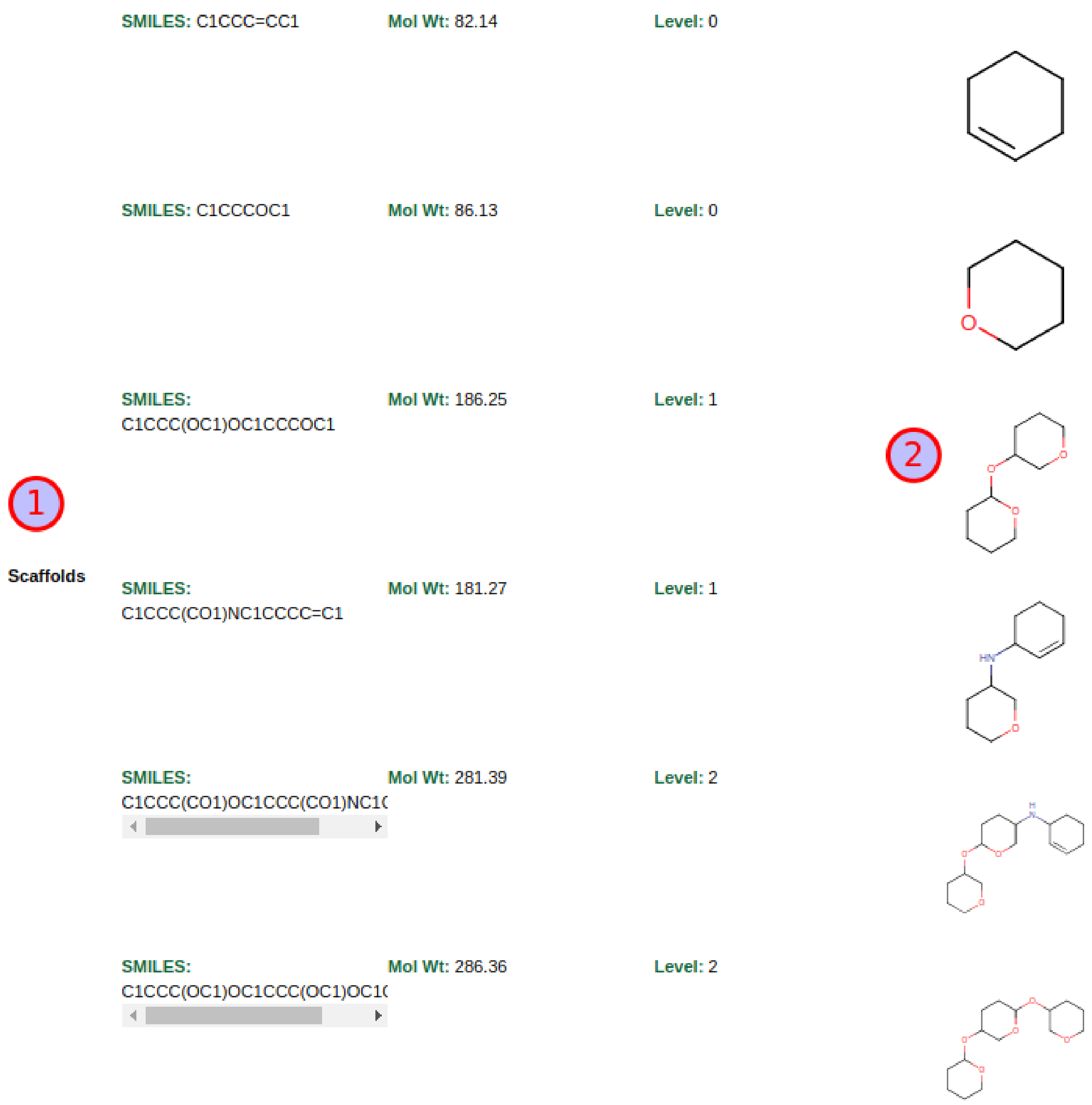
acarviostatin I03 (100.0%)	Trestatin B (100.0%)	acarbose (100.0%)	acarviostatin IV03 (92.5%)
Trestatin C (92.5%)	acarviostatin III03 (92.5%)	Trestatin A (92.5%)	Adiposin 2 (92.5%)
acarviostatin II03 (92.5%)	acarviostatin III0(-1) (92.5%)	Validamycin H (91.1%)	

Field 1: In this field, organisms, related to *Acarbose* are listed along with their taxonomical parent ID.

Field 2: References related to *Acarbose* are listed here.

Field 3: All the similar compounds 2D structures which have tanimoto coefficient more than 0.85 are listed in this field.

Fig. 3.18 StreptomeDB 2.0 results for *Acarbose* providing its similar compounds, compound card: 2/5

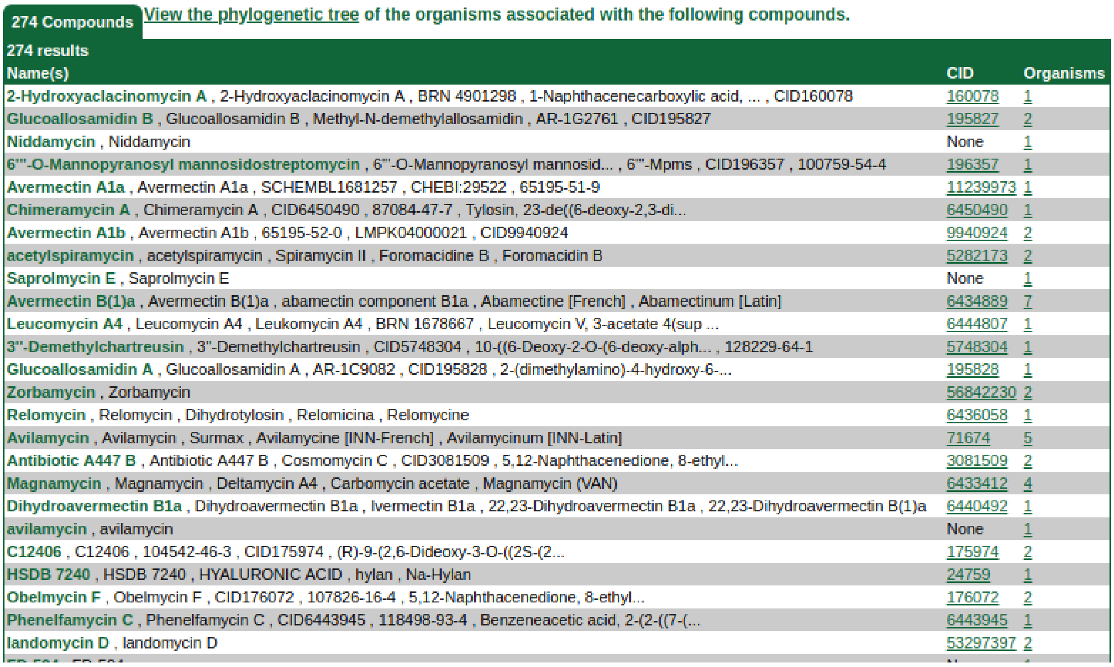


Scaffolds	SMILES: C1CCC=CC1	Mol Wt: 82.14	Level: 0
	SMILES: C1CCCOC1	Mol Wt: 86.13	Level: 0
	SMILES: C1CCC(OC1)OC1CCCOC1	Mol Wt: 186.25	Level: 1
	SMILES: C1CCC(CO1)NC1CCCC=C1	Mol Wt: 181.27	Level: 1
	SMILES: C1CCC(CO1)OC1CCC(CO1)NC1(	Mol Wt: 281.39	Level: 2
	SMILES: C1CCC(OC1)OC1CCC(OC1)OC1(	Mol Wt: 286.36	Level: 2

Field 1: In this field all the scaffolds of *Acarbose* are listed and linked to the compound listing having the selected scaffold.

Field 2: By selecting this scaffold will yield compounds (Figure 3.20) related to it. Conversely, the same scaffold can be selected from the scaffold browser (Figure 3.22).

Fig. 3.19 StreptomeDB 2.0 results for *Acarbose* providing its scaffolds, compound card: 3/5

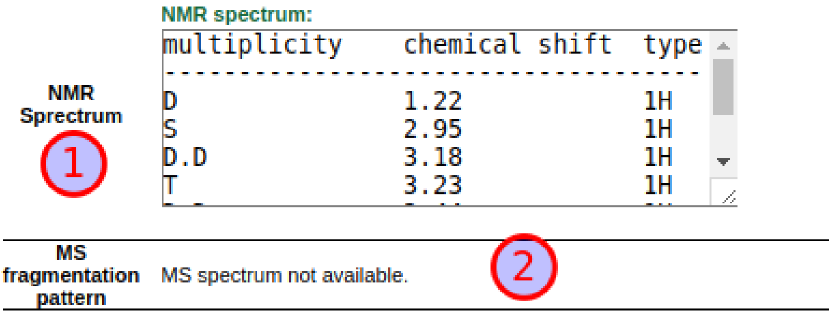


274 Compounds View the phylogenetic tree of the organisms associated with the following compounds.

274 results

Name(s)	CID	Organisms
2-Hydroxyaclacinomycin A , 2-Hydroxyaclacinomycin A , BRN 4901298 , 1-Naphthacenecarboxylic acid, ... , CID160078	160078	1
Glucoallosamidin B , Glucoallosamidin B , Methyl-N-demethylallosamidin , AR-1G2761 , CID195827	195827	2
Niddamycin , Niddamycin	None	1
6'''-O-Mannopyranosyl mannosidostreptomycin , 6'''-O-Mannopyranosyl mannosid... , 6'''-Mpms , CID196357 , 100759-54-4	196357	1
Avermectin A1a , Avermectin A1a , SCHEMBL1681257 , CHEBI:29522 , 65195-51-9	11239973	1
Chimeramycin A , Chimeramycin A , CID6450490 , 87084-47-7 , Tylosin, 23-de((6-deoxy-2,3-di...	6450490	1
Avermectin A1b , Avermectin A1b , 65195-52-0 , LMPK04000021 , CID9940924	9940924	2
acetylspiramycin , acetylspiramycin , Spiramycin II , Foromacidine B , Foromacidin B	5282173	2
Saprolmycin E , Saprolmycin E	None	1
Avermectin B(1)a , Avermectin B(1)a , abamectin component B1a , Abamectine [French] , Abamectinum [Latin]	6434889	7
Leucomycin A4 , Leucomycin A4 , Leukomycin A4 , BRN 1678667 , Leucomycin V, 3-acetate 4(sup ...	6444807	1
3''-Demethylchartreusin , 3''-Demethylchartreusin , CID5748304 , 10-((6-Deoxy-2-O-(6-deoxy-alph... , 128229-64-1	5748304	1
Glucoallosamidin A , Glucoallosamidin A , AR-1C9082 , CID195828 , 2-(dimethylamino)-4-hydroxy-6-...	195828	1
Zorbamycin , Zorbamycin	56842230	2
Relomycin , Relomycin , Dihydrotylosin , Relomicina , Relomycine	6436058	1
Avilamycin , Avilamycin , Surmax , Avilamycine [INN-French] , Avilamycinum [INN-Latin]	71674	5
Antibiotic A447 B , Antibiotic A447 B , Cosmomycin C , CID3081509 , 5,12-Naphthacenedione, 8-ethyl...	3081509	2
Magnamycin , Magnamycin , Deltamycin A4 , Carbomycin acetate , Magnamycin (VAN)	6433412	4
Dihydroavermectin B1a , Dihydroavermectin B1a , Ivermectin B1a , 22,23-Dihydroavermectin B1a , 22,23-Dihydroavermectin B(1)a	6440492	1
avilamycin , avilamycin	None	1
C12406 , C12406 , 104542-46-3 , CID175974 , (R)-9-(2,6-Dideoxy-3-O-((2S-(2...	175974	2
HSDB 7240 , HSDB 7240 , HYALURONIC ACID , hylan , Na-Hylan	24759	1
Obelmycin F , Obelmycin F , CID176072 , 107826-16-4 , 5,12-Naphthacenedione, 8-ethyl...	176072	2
Phenelfamycin C , Phenelfamycin C , CID6443945 , 118498-93-4 , Benzeneacetic acid, 2-(2-((7-(...	6443945	1
landomycin D , landomycin D	53297397	2

Fig. 3.20 Compounds which have the scaffold selected in figure 3.19 of *Acarbose*, compound card: 4/5



NMR spectrum:

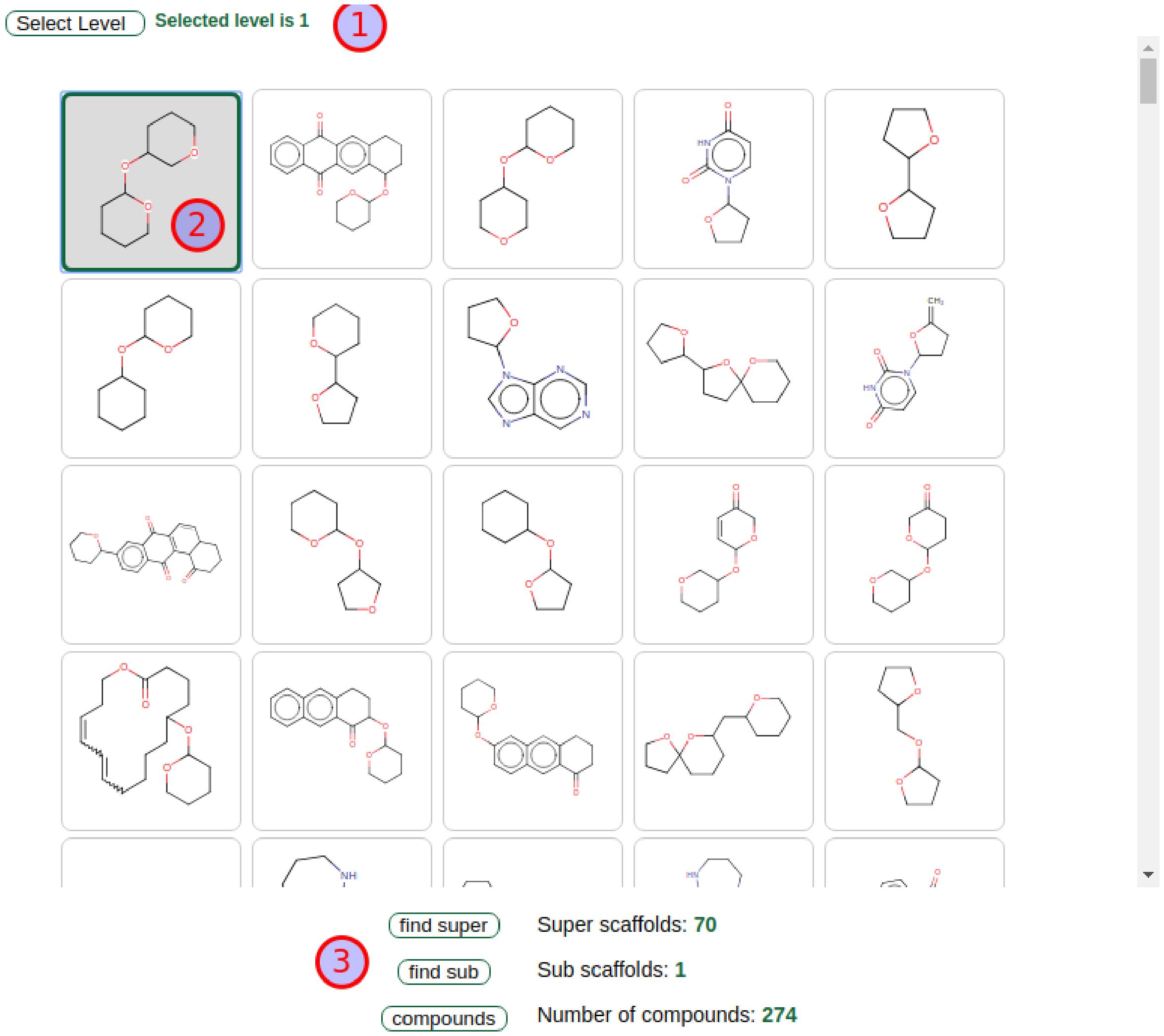
```
multiplicity     chemical shift   type
--------------------------------------
D                1.22             1H
S                2.95             1H
D.D              3.18             1H
T                3.23             1H
```

MS spectrum not available.

Field 1: NMR spectrum 2.1.2 values can be observed here.

Field 2: Mass spectrometry values can be observed here. For an example of MS-Spectra refer figure 3.10.

Fig. 3.21 StreptomeDB 2.0 results for *Acarbose* providing its NMR and MS information, compound card: 5/5



Selected level is 1

Super scaffolds: 70

Sub scaffolds: 1

Number of compounds: 274

Field 1: Here level of the scaffold is selected.

Field 2: Desired scaffold selected in scaffold browser (Section 3.2.2)

Field 3: Options to browse super, sub scaffolds, and compounds related to the selected scaffold. It also pre lists number of results found for each criterion.

Fig. 3.22 StreptomeDB scaffold browser for the *Acarbose* scaffold selected in figure 3.19

3.3 NANPDB

Northern African Natural Products Database (NANPDB)[11] is an accessible online database of Natural products from the region of Northern Africa. The data cover compounds isolated mostly from plants, with contributions from some endophyte, animal (e.g., coral), fungal, and bacterial sources. Computed physicochemical properties, often used to predict drug metabolism and pharmacokinetics, as well as predicted toxicity information, have been included for each compound in the data set. NANPDB is the most extensive collection of annotated natural compounds produced by native organisms from Northern Africa.

3.3.1 NANPDB introduction

As introduced the importance of NPs in the section 1.2, which are leading sources of drugs and drug leads and play an essential role in drug discovery by providing novel scaffolds. One of the critical strategy adopted in pharmacy research world to explore new NPs and increase the chemical diversity is to delve into untapped geographical sources [151, 152]. With this idea, we have invested our resources in exploring much less explored the geographic region of the world, i.e., North African region. Also, NPs procured from this region are either identified as drugs or clinically potential compounds [153].

Northern Africa or the Mediterranean countries which are in the northern-most geographical region of the African continent with an area of 9 million km^2 [154] of which Sahara desert covers more than 75 percent, and it is fastly getting desertified [155]. Being a desert area increases the expectancy of its unearthing unique natural products concerning structural diversity and bio-activity in the selected region compared to other parts of Africa. Plants, when exposed to extreme environments such as drought, develop an adaptive defense mechanism for their survival from abiotic stress which includes morphological as well as biogeochemical [156]. Similarly, microorganisms such as Fungi are majorly affected by the pH of their environment changes, due which organisms develop pH-responsive

[11] http://african-compounds.org/nanpdb/

signaling pathways [157]. Countries which combine Northern Africa are Algeria, Egypt, Libya, Morocco, Sudan, South Sudan, Tunisia, Western Sahara, and parts of Northern Mali.

Earlier to our work there were few attempts made to develop such databases and public resources for the NPs belonging to Africa. Some of these NPs are discussed here, from the Central Africa *CamMedNP* [158], chemical library contain generated 3D models of NPs. CamMedNP contains isolated medicinal plant species belonging to plant families from the *Cameroonian flora*. The NPs in CamMedNP are isolated from 224 plant species from 55 plant families [158]. CamMedNP majorly has compounds which are *terpenoids*, *flavonoids*. It also reported isolated secondary metabolites and their known biological activities. CamMedNP has a wider sampling of the chemical space compared to *ChemBridge*. Other 3D models generated natural product database from Africa is *ConMedNP* [159] it has 3200 compounds of natural origin. Compounds of ConMedNP reveal information from 376 distinct medicinal plants which are extracted from 79 different plant families from the central region of Africa to show the diversity ConMedNP showed the diversity analysis of physicochemical properties with DIVERSet ™ database which contained 48,651 compounds from ChemBridge Corporation [160]. It also showed the *pharmacokinetic profiling* (Section 2.1.5) of the compounds with the help of QikProp software (Section 2.1.4). *AfroDB* is another NP which has distinct NPs from multiple regions of Africa, among which dominating from Central Africa region. It is relatively a small molecule library of ≈1000 compounds [161]. Compounds of AfroDB exhibit biological activity in specific areas, e.g., *prolyl endopeptidase I* inhibition, *11β-hydroxysteroid dehydrogenase* inhibition, *α-glucosidase* inhibition. Generic classification of AfroDB includes anti-HIV, anti-salmonella, activity against *Onchocerca gutturosa* and several others. More than 50% compounds showed no Lipinski violations (Section 2.1.6). AfroDB compound library a subset of *ZINC* library [162] produced after virtual screening ZINC database. *p-ANAPL* [163] has the compound library which can be readily used in research experiments for virtual screening (Section 2.1) from the African medicinal plants [158, 159, 161]. It has the tiny library of just more than ≈500 compounds. It has also discussed the pharmacological profiles of few compounds. Compound type of *flavonoids* had more representation in pan-African Natural Products Library (p-ANAPL).

All of the discussed NP resources have several standard things such as all are from the African region, collected from the various areas. To establish their relevance they have examined *Lipinski's criteria* (Section 2.1.6), compared with DNP datasets [160], known biological activities, chemical composition by plant origin, diversity analysis, the similarity of compounds, prediction of dermal penetration, prediction of plasma-protein binding, metabolism prediction, availability of substructures. To indicate their utility they have provided 3D Models, Calculated molecular descriptors, pharmacokinetic profiling, classification of compounds, PCA, giving virtual libraries in several file formats, availability of compounds.

3.3.2 Information content of NANPDB

The Northern African Natural Products Database (NANPDB) can be browsed through lists and searches (Section 3.3.5.2). Natural products from Northern African species are available to download as SMILES. The data currently covers compounds derived mainly from plants, with contributions from some endophytes, animals (e.g., corals), fungi, and bacteria. The available PubChem [150] compounds links are in the database. For the description of the source species, NANPDB has been linked to available databases; for example, some plant sources have been linked to the Prota [164] and Tropicos [165] databases, while marine sources have been linked to the World Register of Marine Species (WoRMS) [166]. Bacterial sources are connected to GenBank [167] and fungal sources to the Mycobank database [168–170], while the taxonomic data for a majority of the compounds have been linked to the National Center for Biotechnology Information (NCBI) Taxonomy database [171]. Additionally, the availability of source organism storage data, e.g., in national and university herbaria and repositories, has been included, where possible. Literature information (e.g., authors, journal names, titles of publications, literature reference, doi, among others) has also been included, via links to PubMed [34] and journal reference links.

3.3.2.1 Properties of NANPDB compounds

NANPDB represents the largest collection of annotated NPs derived from natural species found in Northern Africa. We had a recent update of compounds after the NANPDB first published in 2017, in this thesis statistics include updated compounds. Although the database currently includes many known drugs and drug leads, the biological activity for the majority of compounds is not yet tested. Hence, assessing the medicinal potential of most of the molecules are needed. It opens a broad window of opportunity for further investigations in drug discovery studies, analysis of the bioactivity of selected compounds, and probing of the routes toward the biosynthesis of some of the identified metabolites. An additional asset of NANPDB is the inclusion of related information such as the terpenoids, flavonoids, and alkaloids, (Figure 3.23). A total of about 100 known biological activities have been recorded, along with 43 distinct modes of action. They have been grouped into 99 biological activity classes. The majority of the known bioactive compounds are anti-infective (e.g., exhibiting anti-HIV, other antiviral, antifungal, antitubercular, other antimycobacterial, and antibacterial activities), cytotoxic, and potential anticancer drugs. The latter class of bioactive molecules includes compounds exhibiting inhibitory activities against a broad range of cancer cell lines or having recorded antileukemic, kinase inhibitory, tumor anti-initiating, tumor-specific cytotoxic, antimetastatic, antiproliferative, and antitumor activities, while other compounds have shown clinical potential for cancer treatment. Other bioactive molecules include antifeedants or insect repellents, among others. Less than half of the dataset (2367 compounds) were present in PubChem at the time of data curation. Less than half of the 874 literature references are currently listed in PubMed, making the database of particular interest.

Besides, estimation of the toxicity predictions with the *pkCSM* model [172] for ten toxicity endpoints is summarized (Tables 3.8 and 3.9). The *pkCSM* is a *in silico* approach which uses graph-based signatures to study and predict for ADMET properties of the chemical compound. The Ames *Salmonella* mutagenicity assay (*Salmonella* test; Ames test) is a bacterial reverse mutation assay to detect molecules which unfold genetic damage. Boolean toxic properties can be understood from the table 3.8. A compound that tests positive for AMES mutagenicity is mutagenic and therefore may act as a carcinogen promoting

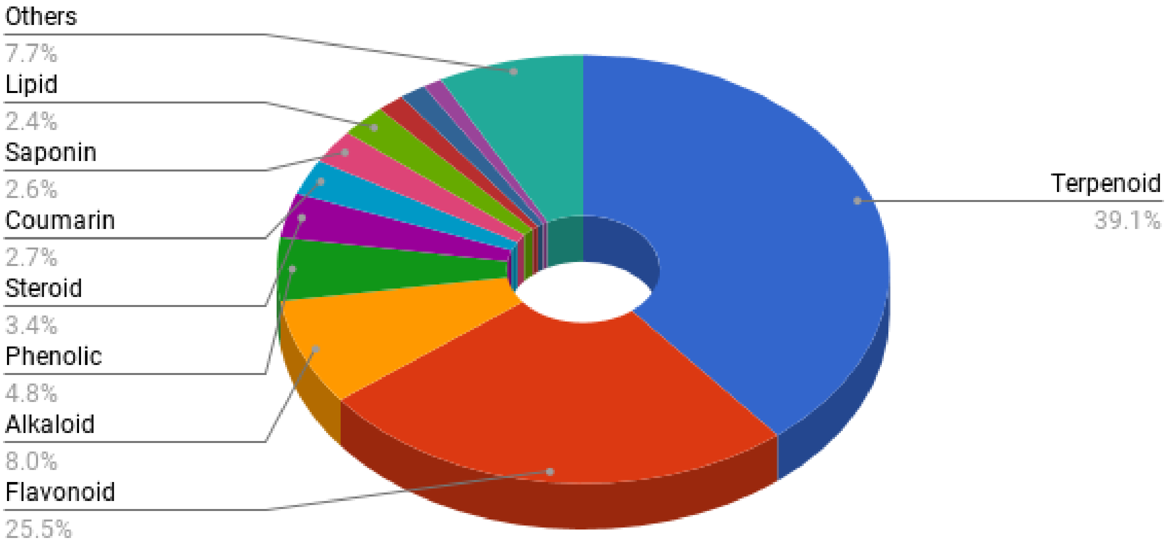


Fig. 3.23 Pie chart showing percentages of main compound classes of NANPDB

cancer [173]. NANPDB exhibits under 15% of compounds to be not passing through Ames test. An hERG I/II inhibitor could cause the development of acquired long QT syndrome, which leads to fatal ventricular arrhythmia [174] and the compounds did not exhibit any hERGI inhibition, 70% compounds did not exhibit hERGII inhibition. Hepatotoxicity leads to drug failures by disrupting the normal function of the liver [175]. A compound that tests positive has a high probability of in disrupting the normal function of the liver, almost 88% compounds are safe to the liver. Skin Sensitisation test determines if the compound is dermally safe. A compound that tests positive could have a high potential adverse effect for products applied to the skin, e.g., cosmetics and antifungals and 84% compounds are predicted to be safe to skin. The second table 3.9 summarizes further information on some more toxic points with their minimum, maximum and mean values. Human maximum tolerated recommended dose provides an estimate of the toxic threshold of the substance in humans. If the compound is $\leq 0.477\ \log(\mathrm{mg\,kg^{-1}\,d^{-1}})$ it is considered low anything greater than it is considered high. The maximum value of NANPDB compounds is 2.38 indicating safe values for all compounds. Rat oral acute toxicity indicates the toxic potency of the molecule. The median lethal dose (LD50) values are the standard measurement of acute toxicity. The model for acute toxicity is built on over 10000 compounds tested in rats [176]. Rat oral chronic toxicity or the lowest observed adverse effect log is prediction can influence the treatment strategy and this a relative term which depends on bioactive concentration and treatment lengths required [177]. T.*pyriformis* is a *protozoa* bacteria, its toxicity is used as the toxic point and only 1% NANPDB compounds satisfy this parameter. Compounds with value greater than -0.5 log µL are considered toxic.*Tetrahymena Pyriformis* is an attractive test organism for the assessment of environmental impact of toxicants [178]. Indicating not many NANPDB compounds passed the toxicological properties of industrial chemicals in the environmental hazard assessment. Minnow toxicity represents the concentration of a molecule necessary to cause death 50% of the Flathead Minnows toxic to nature [179]. Over 85% of compounds are not toxic satisfying the high acute toxicity threshold value of staying below -0.3mM.

Toxicity end point	**Predicted**	**% Compliance**
AMES test	Yes	87.05
	No	12.94
hERG I inhibition	Yes	0
	No	100
hERG II inhibition	Yes	30.22
	No	69.78
Hepatotoxicity	Yes	12.08
	No	87.92
Skin sensitization	Yes	15.89
	No	84.10

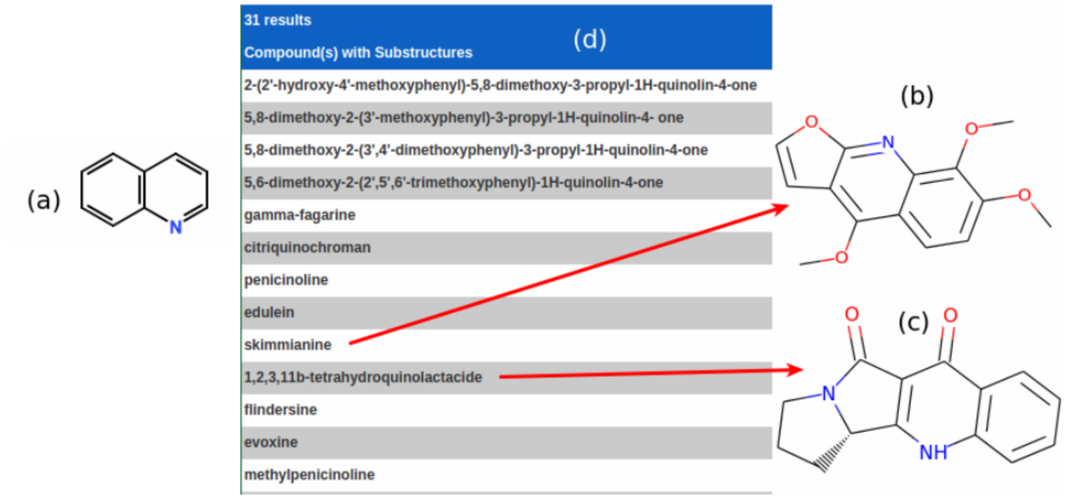
Table 3.8 Selected predicted toxicity parameters for NANPDB 1/2

Parameter	**Maximum**	**Minimum**	**Mean**
Human maximum tolerated recommended dose (log $\mathrm{mg\,kg^{-1}\,d^{-1}}$**)**	2.38	-3.09	0.13
Rat oral acute toxicity ($\mathrm{mg\,kg^{-1}}$)	4.99	0.69	2.39
Rat oral chronic toxicity (log mg/kg_bw/d)	44.95	-1.03	2.63
T. pyriformis toxicity (log µL)	2.76	-1.57	0.46
Minnow toxicity (log mM)	52.29	-9.64	2.55

Table 3.9 Selected predicted toxicity parameters for NANPDB 2/2

3.3.2.2 Use case of NANPDB

Natural products are a precious resource for the identification of new drugs and are important sources of chemical diversity [180–182]. In the diverse compounds, a more in-depth look at *privileged scaffolds* which are usually related to specific bioactivity can result in more bioactive compounds. An example, quinoline is a nitrogen-containing heterocyclic aromatic organic compound. It is a colorless hygroscopic liquid. And it appears in kinase inhibitors [183], show potent inhibitory activity at low micromolar concentrations [184], and quinoline compounds are anti-malarial and anti-cancer agents [185]. Quinolines have become more useful after having regioselective control over them [186]. A more in-depth study with PubMedPortable can reveal various information about the quinoline confirming its appearance in several drugs. A substructure search with quinoline (Figure 3.24-a) in NANPDB yields 31 different molecules (Figure 3.24-d). The bioactivity and mode of action of several resulted compounds can be at least partly understood, e.g., the anti-inflammatory properties of skimmianine (Figure 3.24-b) against such as carrageenan-induced acute inflammation [187]. For other molecules such as 1,2,3,11b-tetrahydroquinolactacide (Figure 3.24-c) a precursor of quinolactacide [188], the activity has to be tested. All the 31 molecules can be a good starting point for the search for novel drugs.



31 results

Compound(s) with Substructures

2-(2'-hydroxy-4'-methoxyphenyl)-5,8-dimethoxy-3-propyl-1H-quinolin-4-one
5,8-dimethoxy-2-(3'-methoxyphenyl)-3-propyl-1H-quinolin-4- one
5,8-dimethoxy-2-(3',4'-dimethoxyphenyl)-3-propyl-1H-quinolin-4-one
5,6-dimethoxy-2-(2',5',6'-trimethoxyphenyl)-1H-quinolin-4-one
gamma-fagarine
citriquinochroman
penicinoline
edulein
skimmianine
1,2,3,11b-tetrahydroquinolactacide
flindersine
evoxine
methylpenicinoline

Fig. 3.24 Identification of potential drugs in NANPDB by substructure search using quinoline as the *privileged scaffold*, leading to 31 hits. (a) Quinoline scaffold, (b) skimmianine, (c) 1,2,3,11b-tetrahydroquinolactacide, and (d) Names of the first 13 hits, indicating b and c compounds

3.3.2.3 NANPDB compared to other NP databases

The compounds in NANPDB were further compared with known drugs and other published natural products data sets, some of which contain compounds from other specific geographical regions (Figure 3.25), while others are collections of NPs with activities against specific diseases tuberculosis and cancer. The Venn diagrams in 3.26 show the overlap of NANPDB with the selected data sets (Table 3.10).

Dataset	**# SMILES**	**Nature of compounds**
NANPDB[12] [189]	4928	NPs from Northern African sources
IDAAPM[13] [190]	1617	FDA-approved drugs
DrugBank[14] [132]	7133	Known drugs
NuBBE[15] [191]	1749	NPs from Brazilian sources
StreptomeDB 2.0[16] [22]	4040	NPs from *Streptomyces* species
ConMedNP [192]	3177	NPs from Central African sources
BioPhytMol[17] [193]	633	Antimycobacterial NPs from around the world
NPACT[18] [194]	1574	Anticancer NPs from around the world

Table 3.10 Description of datasets used for comparative analysis of ADMET-related descriptors study

[12] http://african-compounds.org/nanpdb/
[13] http://idaapm.helsinki.fi/
[14] https://www.drugbank.ca/
[15] http://nubbe.iq.unesp.br/portal/nubbedb.html
[16] http://phabi.de/streptomedb2/
[17] http://ab-openlab.csir.res.in/biophytmol/
[18] http://crdd.osdd.net/raghava/npact/

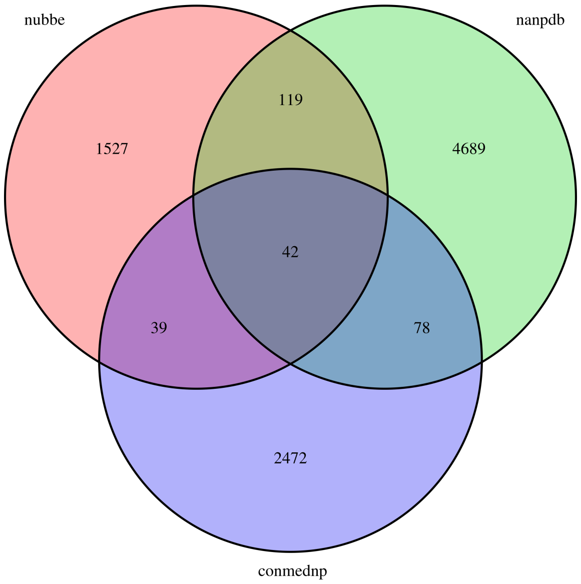


Fig. 3.25 NANPDB compared to NPs from other geographical regions

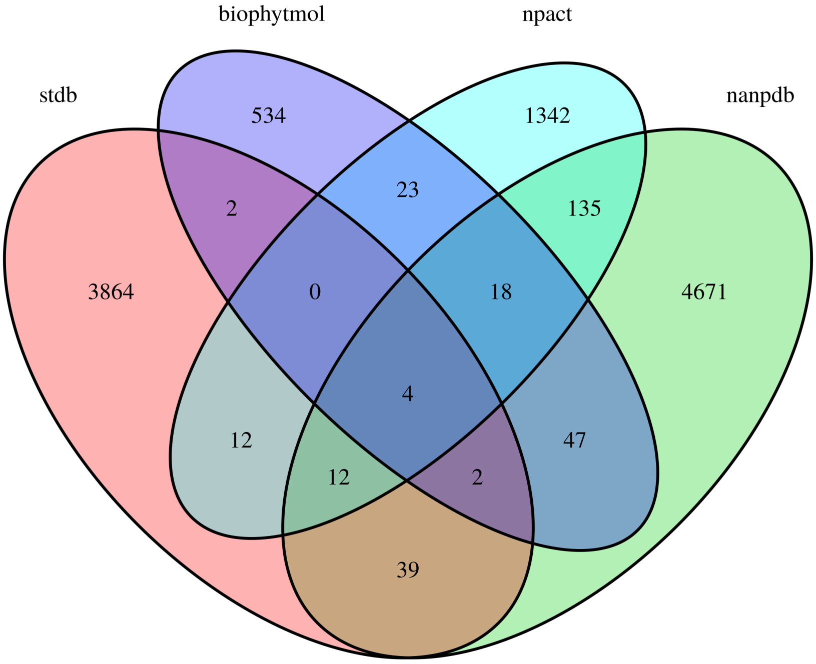


Fig. 3.26 NANPDB compared to NPs of bacterial origin, targeting specific disease tuberculosis and cancer

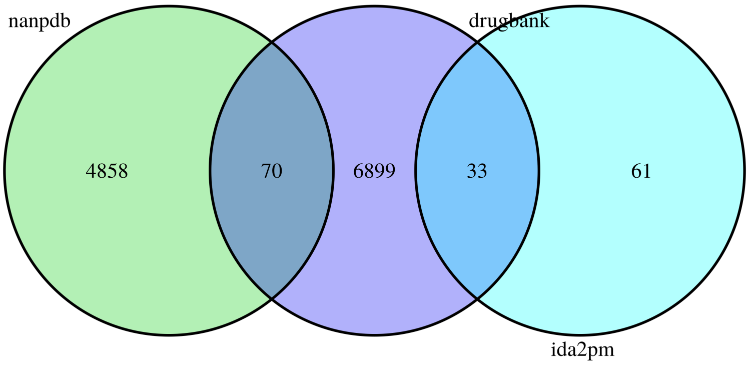


Fig. 3.27 NANPDB compared to drugbank

The present database is also compared with the FDA-approved drugs from the DrugBank and ida2pam (Figure 3.27). All the compared databases information are provided in the table 3.10. In all the comparisons a common phenomenon is that NANPDB has several new sets of compounds which are not available in other databases making NANPDB more interesting and unique. The evaluation of the physicochemical properties (Tables 3.11, 3.12) employed in Lipinsk's RO5 (Section 2.1.6) comparing with other data sets (Table 3.10) only known drugs (DrugBank and IDA2PM), along with NuBBE and BioPhytMol, showed a better drug-likeness. It was additionally observed that the mean clogP values of both NANPDB and FDA-approved drugs were equal to ≈2.15 units (Table 3.10)

Dataset	MW			clogP			HBA		
	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg
NANPDB	1625.67	55.07	419.15	20.05	-9.88	2.11	53	0	9.1
IDA2PM	1626.24	17.03	370.13	16.71	-12.49	2.16	64	0	6.98
DRUGBANK	1150.19	17.03	341.1	17.84	-16.04	1.56	69	0	7.23
NuBBE	1171.04	86.09	369.45	18.19	-5.31	3.22	38	0	6.29
StreptomeDB 2.0	1473.69	17.03	514.75	19.21	-18.67	1.03	71	0	12.51
ConMedNP	1439.6	84.16	426.7	22.29	-6.8	3.86	70	0	5.85
BioPhytMol	1084.73	74.08	347.43	13.45	-5.41	3.92	39	0	4.73
NPACT	1383.53	106.12	441.94	18.46	-6.32	3.19	51	0	8.2

Table 3.11 Comparison of RO5 parameters of NANPDB with approved drugs and five other NP datasets 1/2

Dataset	HBD			NRB			NLV		
	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg
NANPDB	24	0	3.06	58	0	7.86	3	0	0.85
IDA2PM	19	0	2.17	52	0	6.81	4	0	0.45
DRUGBANK	25	0	2.65	62	0	6.96	4	0	0.42
NuBBE	16	0	1.59	30	0	5.99	3	0	0.54
StreptomeDB 2.0	29	0	3.84	66	0	12.72	4	0	1.29
ConMedNP	37	0	2.39	56	0	5.51	4	0	0.71
BioPhytMol	17	0	1.12	51	0	5.78	4	0	0.46
NPACT	18	0	2.54	44	0	8.41	4	0	0.79

Table 3.12 Comparison of RO5 parameters of NANPDB with approved drugs and five other NP datasets 2/2

3.3.3 NANPDB webservice development

3.3.3.1 Data collection

Our collaborator implemented the data collection Dr. Ntie-Kang[19]. The curators and developers were in different locations, and so good planning was needed, and we followed a workflow (Figure 3.28) for the entire process. The data related to source organisms, geographical collection sites and chemical structures of derived compounds have been retrieved from literature sources from the major international journals on natural products and medicinal chemistry, alongside available Ph.D. theses, spanning the period 1962 to 2016. The journal article data sources are from the journals queried with country name as the search term. The resulting articles were checked to verify that the source species harvest from the Northern African region. The retained articles were downloaded and referred to henceforth as data sources. The data sources were arranged by taxonomic families of source organisms, to avoid duplicate curation. From each data source abstracts and full text are collected. The accuracy of the manually curated data is particularly checked for chemical information, biological activity data, and source species information.

[19] http://pc.pharmazie.uni-halle.de/medchem/mitarbeiter/fidele_ntie-kang/

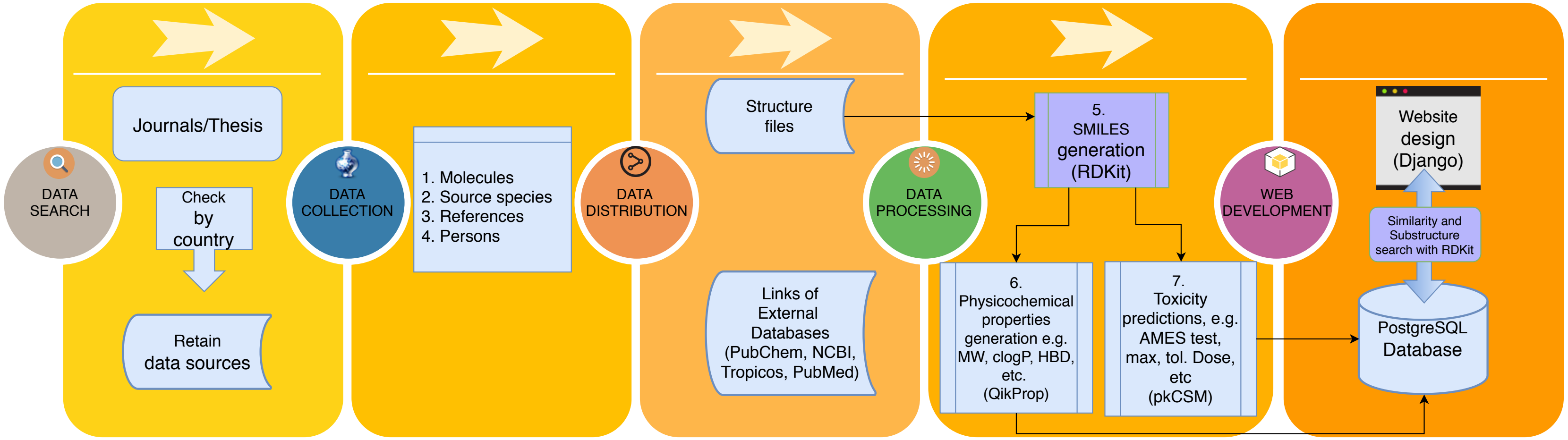

DATA SEARCH
Journals/Thesis
Check by country
Retain data sources
DATA COLLECTION
1. Molecules
2. Source species
3. References
4. Persons
DATA DISTRIBUTION
Structure files
Links of External Databases (PubChem, NCBI, Tropicos, PubMed)
DATA PROCESSING
5. SMILES generation (RDKit)
6. Physicochemical properties generation e.g. MW, clogP, HBD, etc. (QikProp)
7. Toxicity predictions, e.g. AMES test, max, tol. Dose, etc (pkCSM)
WEB DEVELOPMENT
Website design (Django)
Similarity and Substructure search with RDKit
PostgreSQL Database

Fig. 3.28 The complete workflow that captures the entire process of construction of NANPDB

3.3.3.2 Database implementation

The data in NANPDB is organized as a relational database. The main goal of the design was to access the information with minimum redundancy in the data readily. PostgreSQL 9.5.4 (Section 2.5.2) was applied as a database server [195] combined with the RDKit (Section 2.3.1) for processing small molecules [196] and Ubuntu 16.04 LTS as host operating system. Physicochemical properties of the molecules were generated using QikProp (version 2015) [56] and were used, together with the reported biological activity information, as molecular properties. Compound-related information stored in a single table also includes the chemical structure of the compounds in SMILES format as well as names and synonyms extracted from the PubChem database. The source species information such as kingdom, family, habitat, and availability is inserted into another table. Two other tables include literature reference information and information about the curators. Toxicity predictions were stored in a fifth table (Figure 3.29).

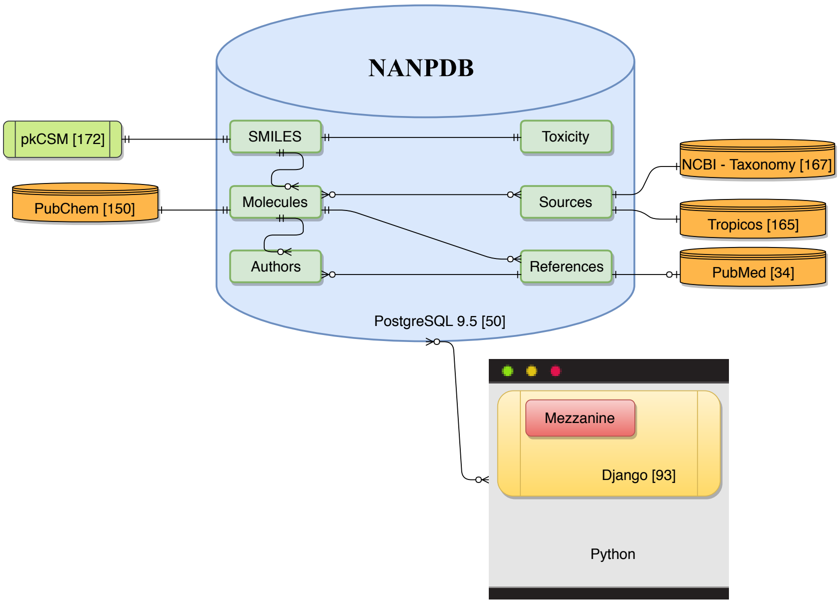


Fig. 3.29 Simplified database schema of NANPDB in PostgreSql

Appendix figure C.1 provides a summary of the main contents of each SQL table and how the tables are connected to each other. Compound names and synonyms were individually queried in PubChem [150], and related data were retrieved. A Python script was used to download individual 2D molecule (sdf) files from PubChem using the manually retrieved and verified PubChem ID (PCID) compound codes. Where compounds were not available in PubChem [150] ChemSpider [197] or MarvinSketch [198] the compounds were sketched (based on reported 2D structures) and compared with the structure (if available) in SciFinder [199]. The 2D sketches were carried out using the ChemDraw 8.0 Ultra tool [200] and saved as 2D mol files. Chemical structures were double checked for potential problems such as incorrect formulas, the presence of salts, and counterions. Unique (canonical) SMILES strings were generated using Open Babel [92] following the previously described methodology [141, 22]. Lipinski physicochemical property filters such as molecular weight (MW), computed logarithms of the n-octanol/water partition coefficients (clogP), number of hydrogen bond donors (HBDs) and acceptors (HBAs), and other properties related to drug metabolism and pharmacokinetics (DMPK) were computed using QikProp (Schrodinger, LLC) [55] after a preliminary ligand preparation treatment on the maestro interface [201] using LigPrep (Schrodinger, LLC) [202] Data were uploaded in a PostgreSQL 9.5.4 database.

3.3.4 DM compounds in NANPDB

Exercising further the diabetes dataset from section 3.1.2 calculating the molecular similarity of NANPDB compounds DrugBank DM compounds (Appendix table A.1) elucidated the numbers in the table 3.13.

Query	#	Notes
Number of DrugBank DM compounds	67	listed in table A.1
Number of NANPDB compounds	4928	available in NANPDB
Number of Fingerprints used	28	fingerprints table 2.1
Number of comparisons with all Fingerprints	7035756	Between DrugBank DM compounds and NANPDB compounds
Number of comparison pairs with Tanimoto >0.85	1275	Fingerprint wise distribution listed in table 3.14
Number of unique comparison pairs with Tanimoto >0.85	1266	Top 10 pairs according to number of fingerprints predicted as similar listed in table 3.15

Table 3.13 Informaiton generated with the molecular similarity between DrugBank [132] DM and NANPDB [189] compounds

The similarity distribution of the comparisons between chosen sets of compounds with the different diverse fingerprints (Table 2.1) showed that there were very few comparisons which are similar to DrugBank DM compounds. It revealed 1266 comparisons (Table 3.13) filtering almost 99.999% comparisons. The Kernel density estimation of Tanimoto coefficient for some of the significant fingerprints (Figure 2.3) can be explained by the graph (Figure 3.30). Kernel density estimation is available for all the fingerprints in the appendix, Figure C.2.

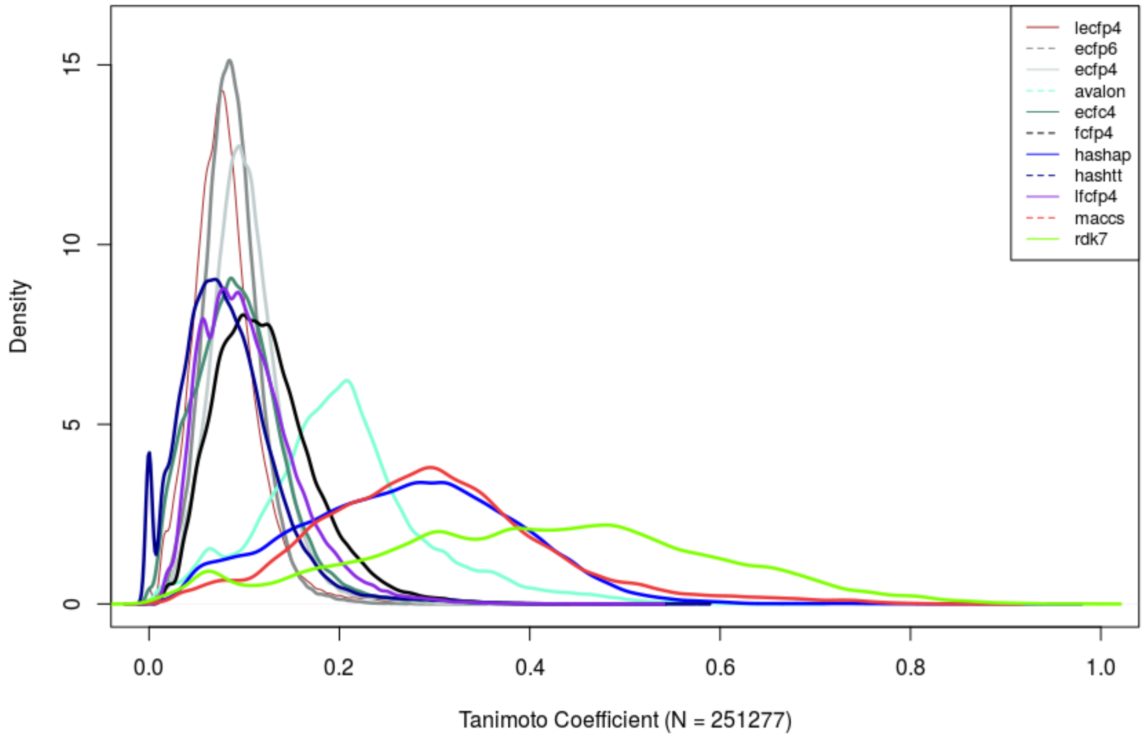


Fig. 3.30 Graph showing Kernel density estimation of tanimoto coefficient between DrugBank [132] DM compounds and NANPDB [189] compounds. N indicates the number of comparisions with each fingerprint. Major fingerprints are taken from 2.3.

The Kernel density graph (Figure 3.30) showing very few comparisons above 0.85 for different fingerprints are in table 3.14.

Fingerprint	# tanimoto >0.85
rdk7	902
fcfp2	168
ecfp0	152
hashap	42
maccs	10
fcfc2	1

Table 3.14 Fingerprint wise distribution of comparisions between DrugBank DM compounds and NANPDB compounds which have tanimoto > 0.85

In the table 3.14, there are several results for fingerprints rdk7, fcfp2, ecfp0, and hashap however on viewing the structures (Figure 3.31) they were not similar in several pairs (Appendix figures C.3, C.4, C.5, C.6). An hypotheses can be that if a pair indicated as similar by several fingerprints can be structurally also similar. However this was not the case for the pairs indicated by several fingerprints are similar, such as top 10 fingerprints estimated by multiple fingerprints are in table 3.15.

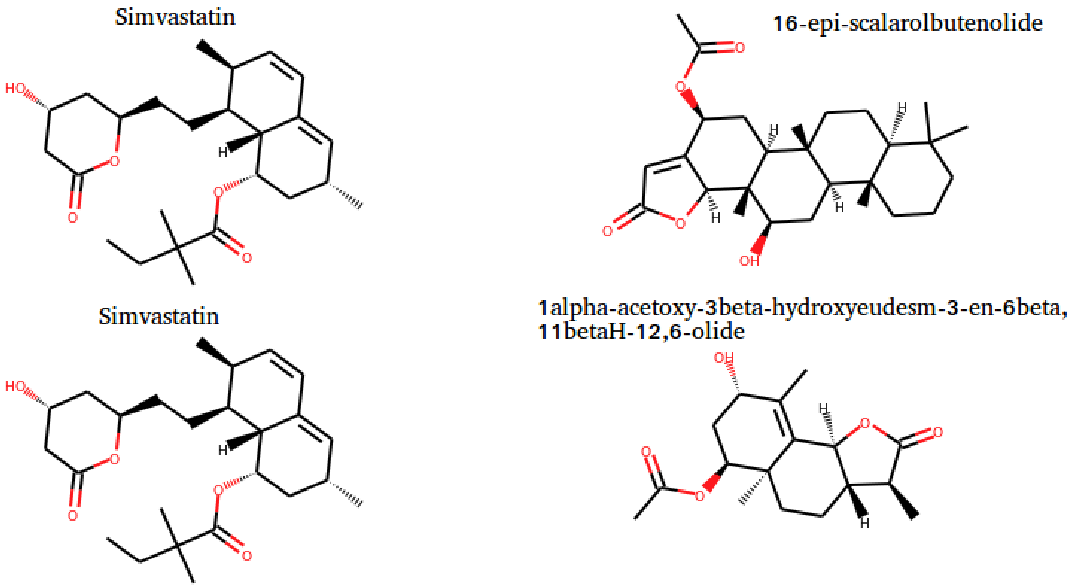

Simvastatin
16-epi-scalarolbutenolide
Simvastatin
1alpha-acetoxy-3beta-hydroxyeudesm-3-en-6beta, 11betaH-12,6-olide

Fig. 3.31 Structures of top 10 pair compounds which are listed in table 3.15, figure set: 1/5 and more compound structure images are in appendix C.3. DrugBank compounds (left), NANPDB compounds (right).

DM DrugBank ID	DrugBank drug name	NANPDB ID	NANPDB compound name	# of fingerprints predicted
DB00641	Simvastatin	3950	16-epi-scalarolbutenolide	3
DB00641	Simvastatin	1800	1alpha-acetoxy-3beta-hydroxyeudesm-3-en-6beta,11betaH-12,6-olide	2
DB00641	Simvastatin	3949	sesterstatin 7	2
DB00641	Simvastatin	5158	3,21-dipalmitoyloxy-16beta,21alpha-dihydroxy-beta-amyrine	2
DB09265	Lixisenatide	2259	euphornin D	2
DB09265	Lixisenatide	2262	euphornin G	2
DB09265	Lixisenatide	2316	euphorhelin	2
DB09265	Lixisenatide	2562	cocciferin T2	2
DB00178	Ramipril	3760	3-(3'-methoxytropoyloxy)tropane	1
DB00178	Ramipril	3765	littorine	1

Table 3.15 Top 10 pairs of DrugBank-NANPDB according to number of fingerprints predicted as similar with Tanimoto coefficient >0.85

Comparing the two results of similarity, one compounds of DrugBank DM with NANPDB (Table 3.13) and the other DrugBank DM with StreptomeDB 2.0 (Table 3.5), StreptomeDB 2.0 had better results although NANPDB had more compounds. Indicating StreptomeDB 2.0 has a closer structural relationship than NANPDB compound with DM compounds. However, in NANPDB there are several compounds (Table 3.16) with antidiabetic bioactivity. Conversely when their best Tanimoto estimated is queried among all calculated fingerprints only 25% compounds (Table 3.17) were having Tanimoto higher than 0.85, and their structures are in appendix C.4.

NANPDB ID	Molecule name	NANPDB ID	Molecule name
696	diosmetin 7-O-beta-L-arabinofuranosyl (12) beta-D-apiofuranoside	2773	caryatin
697	diosmetin 7-O-beta-D-apiofuranoside	2774	caryatin-3'-sulphate
749	berberine	2775	caryatin-3'-methyl ether-7-O-beta-D-glucoside
1126	quercetin-3-O-beta-D-glucopyranoside	3092	excelside B
1127	isoquercetrin	3094	(2S,4S,3E)-methyl 3-ethylidene-4-{2-[2-(4-hydroxyphenyl)ethyl]oxy-2-oxoethyl}-2-[(6-O-beta-D-glucopyranosyl-beta-d-glucopyranosyl)oxy]-3,4-dihydro-2H-pyran-5-carboxylate
1171	quercetin-3'-methoxy-3-O-(4"-acetylrhamnoside)-7-O-alpha-rhamnoside	3095	nuzhenide
1172	kaempferol-4'-methoxy-3,7-dirhamnoside	3096	GI-3
2769	protocatechuic acid	3097	GI-5
2770	quercetin	3100	oleoside dimethyl ester
2771	caryatin-3'-methyl ether	3693	scropolioside D
2772	azaleatin		

Table 3.16 NANPDB antidiabetic compounds

DrugBank ID	DrugBank name	NANPDB msrid	NANPDB name	Fingerprint type	Tanimoto coefficient
DB01200	Bromocriptine	1171	quercetin-3'-methoxy-3-O-(4"-acetylrhamnoside)-7-O-alpha-rhamnoside	rdk7	0.886608
DB09265	Lixisenatide	615	saponin 8	hashap	0.86608
DB01200	Bromocriptine	1172	kaempferol-4'-methoxy-3,7-dirhamnoside	rdk7	0.855044
DB01200	Bromocriptine	2775	caryatin-3'-methyl ether-7-O-beta-D-glucoside	rdk7	0.852395
DB00284	Acarbose	3100	oleoside dimethyl ester	ecfp0	0.833333
DB00284	Acarbose	3092	excelside B	ecfp0	0.833333
DB00284	Acarbose	3096	GI-3	ecfp0	0.833333
DB11827	Ertugliflozin	697	diosmetin 7-O-beta-D-apiofuranoside	ecfp0	0.833333
DB00641	Simvastatin	697	diosmetin 7-O-beta-D-apiofuranoside	ecfp0	0.833333
DB00284	Acarbose	3097	GI-5	ecfp0	0.833333
DB00284	Acarbose	3095	nuzhenide	ecfp0	0.833333
DB00641	Simvastatin	1126	quercetin-3-O-beta-D-glucopyranoside	ecfp0	0.818182
DB01200	Bromocriptine	749	berberine	rdk7	0.787046
DB01200	Bromocriptine	2774	caryatin-3'-sulphate	rdk7	0.777126
DB00284	Acarbose	3693	scropolioside D	ecfp0	0.769231
DB01200	Bromocriptine	2771	caryatin-3'-methyl ether	rdk7	0.733595
DB01200	Bromocriptine	2773	caryatin	rdk7	0.7238
DB01200	Bromocriptine	2772	azaleatin	rdk7	0.68952
DB01200	Bromocriptine	162	quercetin	rdk7	0.657843
DB00966	Telmisartan	307	protocatechuic acid	ecfp0	0.5
DB01124	Tolbutamide	307	protocatechuic acid	ecfp0	0.5

Table 3.17 NANPDB compounds which are marked as antidiabetic and their best tanimoto coefficient

3.3.5 NANPDB features

With the rich resource of African compounds compilation, *NANPDB* provides small molecule results in various user-friendly formats. Namely, they are as follows:

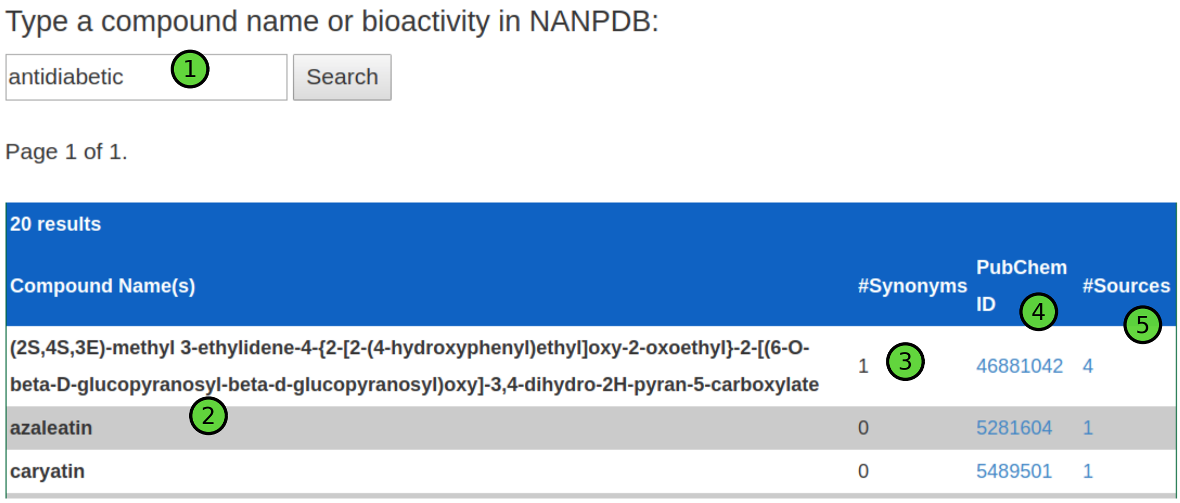
3.3.5.1 Compounds - Bioactivity search

The essential utility of databases such as NANPDB is to be able to search for compounds. This feature allows to search based on compound name or its synonyms or the bioactivity. When bioactivity is given as a query, all the compounds exhibiting the queried bioactivity are listed. Figure 3.32 shows how compound - bioactivity[20] search is performed. The code listing 3 produces compounds - bioactivity result.

```
moleculeQuerySetList = moleculeQuerySetList.filter(
            Q(mol_name__icontains=query) |
            Q(bio_activity__icontains=query)
            ).select_related('pubchem').distinct('mol_name')
```

Listing 3 QuerySet (Section 2.5.3) code of compound - bioactivity search

[20] http://african-compounds.org/nanpdb/compounds_search/



Type a compound name or bioactivity in NANPDB:

antidiabetic

Page 1 of 1.

20 results

Compound Name(s)	#Synonyms	PubChem ID	#Sources
(2S,4S,3E)-methyl 3-ethylidene-4-{2-[2-(4-hydroxyphenyl)ethyl]oxy-2-oxoethyl}-2-[(6-O-beta-D-glucopyranosyl-beta-d-glucopyranosyl)oxy]-3,4-dihydro-2H-pyran-5-carboxylate	1	46881042	4
azaleatin	0	5281604	1
caryatin	0	5489501	1

Field 1: In this field, a compound name, a synonym or bioactivity can be entered. Here *cancer* is given.

Field 2: These are compounds which are similar to the given compound name or compounds which contain the queried bioactivity.

Field 3: For the given query the search yields compounds or synonyms indicated by the displayed number. To view the output, one can click on the compound name on Field (2).

Field 4: In this field if a PubChem entry of the compopund is availble its link is along with PubChem ID is provided.

Field 5: The number on Field (5) indicates the number of source species containing the corresponding molecule or synonym.

Fig. 3.32 Compounds or the Bioactivity search

3.3.5.2 List searches

List search feature starts with alphabetically ascending sorted list, and one can easily switch to any required alphabet group of compounds. In these lists as shown in figure 3.33 corresponding to each compound its related information is provided. List searches are available for compounds (Figure 3.33), families (Figure 3.34), references (Figure 3.35), species (Figure 3.36), authors (Figure 3.37).

3.3.5.2.1 Compounds list:

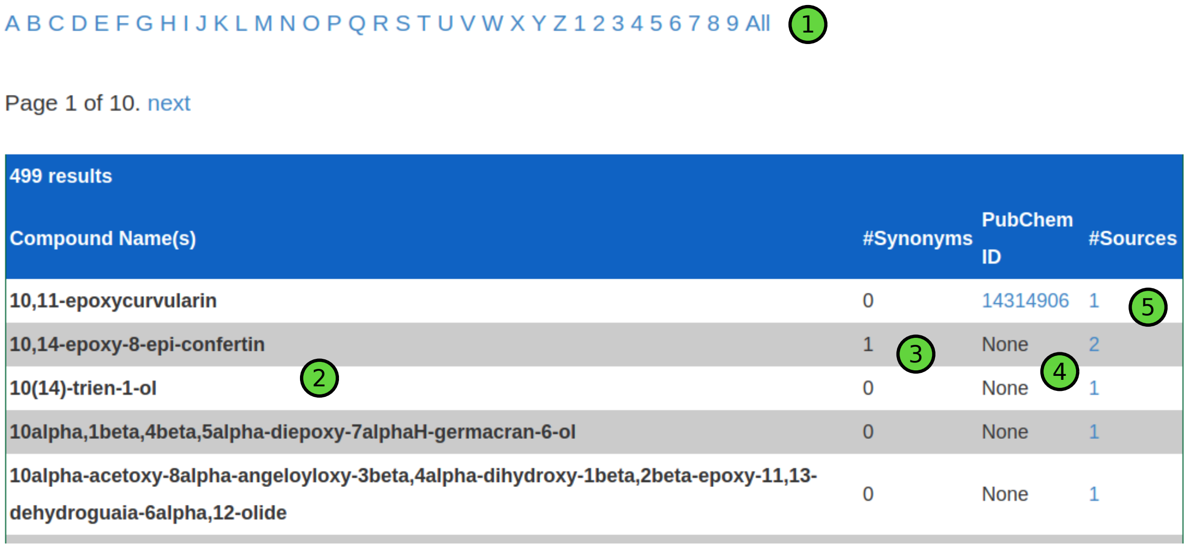
A typical compounds list[21] search result would look like Figure 3.33. Compounds list serves the purpose of browsing based on a name of the small molecule or the number of sources available to it. In this example alphabet *S* is clicked and subsequently 265 results i.e compounds found with letter *S*. The code listing 4 lists the compound list, the results can be seen in the figure 3.33.

```
if currentInitial == "All":
        moleculeQuerySetList = moleculeQuerySetList.distinct('mol_name')
else:
    moleculeQuerySetList = Molecule.objects.filter(
        Q(mol_name__istartswith=currentInitial) |
        Q(mol_name__istartswith=currentInitial.lower())
        ).select_related('pubchem').distinct('mol_name')

molecules = Molecule.objects.filter(Q(mol_name__istartswith=currentInitial) |
            Q(mol_name__istartswith=currentInitial.lower())
            ).select_related('pubchem').annotate(smiles_count=Count('smiles'))
```

Listing 4 QuerySet code of compound list search

[21] http://african-compounds.org/nanpdb/compounds_list/



A B C D E F G H I J K L M N O P Q R S T U V W X Y Z 1 2 3 4 5 6 7 8 9 All

Page 1 of 10. next

499 results

Compound Name(s)	#Synonyms	PubChem ID	#Sources
10,11-epoxycurvularin	0	14314906	1
10,14-epoxy-8-epi-confertin	1	None	2
10(14)-trien-1-ol	0	None	1
10alpha,1beta,4beta,5alpha-diepoxy-7alphaH-germacran-6-ol	0	None	1
10alpha-acetoxy-8alpha-angeloyloxy-3beta,4alpha-dihydroxy-1beta,2beta-epoxy-11,13-dehydroguaia-6alpha,12-olide	0	None	1

Field 1: The NANPDB compounds have been arranged alphabetically and numerically, in *Field (1)*. By clicking any of the letters (A-Z) or numbers (1-9), results in the list of compounds whose names begin with the matched letter or number.

Field 2: Here selection or search for the compound of interest on the list in Field (2). To see compound card.

Field 3: The number in Field (3) indicates how many synonyms there are for the corresponding molecule. It can be viewed by clicking on the number in Field (3).

Field 4: If a PubChem ID of the corresponding compound is available, it is shown in Field (4). Clicking the number in Field(4) redirects to PubChem reference.

Field 5: The number on Field (5) indicates the number of source species containing a particular molecule or its synonyms.

Fig. 3.33 Compounds list search

3.3.5.2.2 Families list

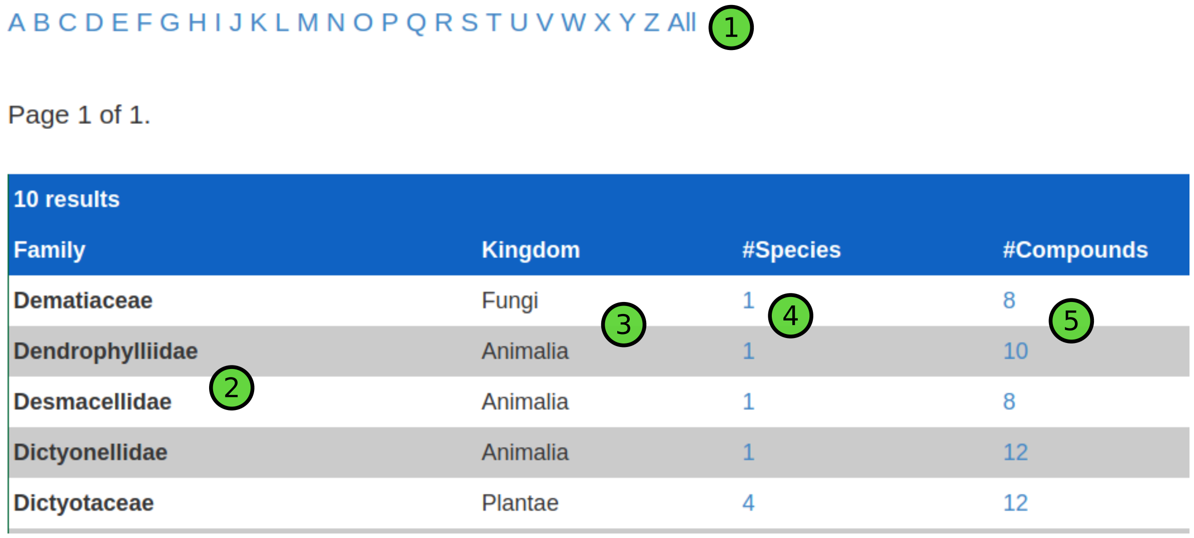
Families list[22] search can be utilized when the number of compounds allocated to the family is to be known. The code listing 5 lists the families list, the results can be seen in the figure 3.34.

```
if currentInitial == "All":
    familyQuerySetList = familyQuerySetList.distinct('family')
else:
    familyQuerySetList = Sources.objects.filter(
        Q(family__istartswith=currentInitial) |
        Q(family__istartswith=currentInitial.lower())
        ).distinct('family')

sourceQuerySetList = \
    sourceQuerySetList.annotate(species_count = Count('source_species'))
compoundCount = \
    Sources.objects.values('family').annotate(compound_count = Count('family'),
                    species_count=Count('source_species'))
```

Listing 5 QuerySet code of families list search

[22] http://african-compounds.org/nanpdb/families_list/



A B C D E F G H I J K L M N O P Q R S T U V W X Y Z All

Page 1 of 1.

10 results			
Family	**Kingdom**	**#Species**	**#Compounds**
Dematiaceae	Fungi	1	8
Dendrophylliidae	Animalia	1	10
Desmacellidae	Animalia	1	8
Dictyonellidae	Animalia	1	12
Dictyotaceae	Plantae	4	12

Field 1: The NANPDB species have been arranged alphabetically into their respective families, Field (1).

Field 2: By clicking the first letter of the respective family name, or All, an output is produced, Field (2). Families for letter A are shown in the picture. By clicking on a particular family, all its similar species are shown.

Field 3: corresponds to the kingdom of the corresponding family.

Field 4: The number on Field (4) indicates the number of species in a particular family.

Field 5: The number on Field (5) indicates the number of compounds found in all source species of the corresponding family.

Fig. 3.34 Families list search

3.3.5.2.3 References list

References list[23] search can be utilized when the number of compounds described or mentioned in a reference is to be known. The code listing 6 lists the references list, the results can be seen in the figure 3.35. Further for each reference, its PubMed ID is also mapped.

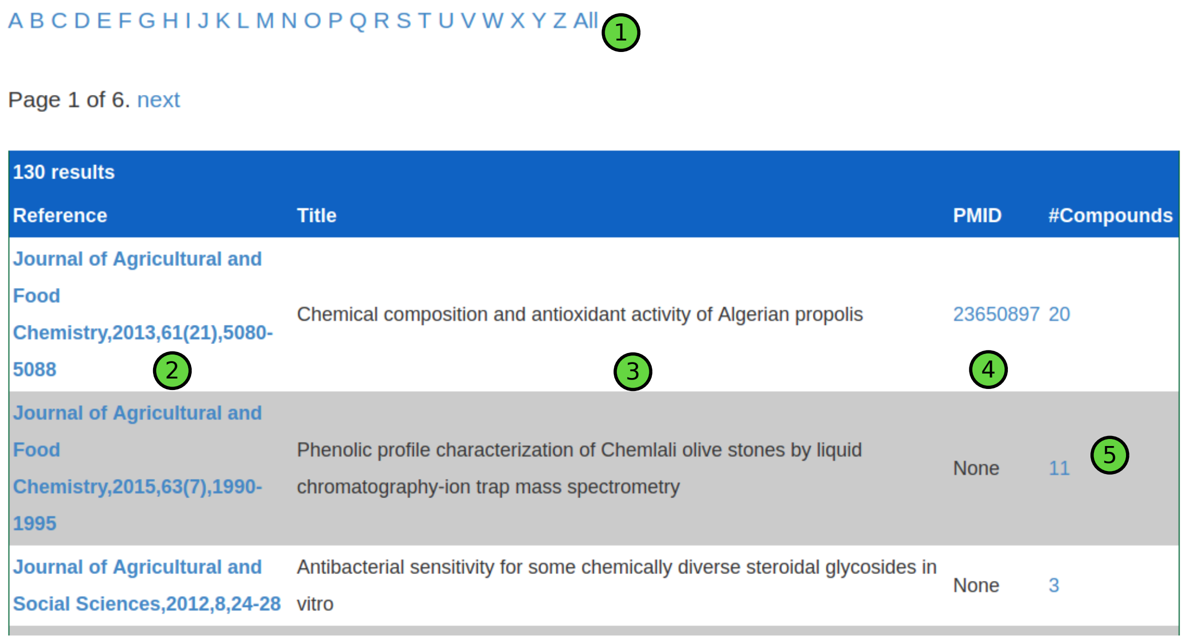
```
if currentInitial == "All":
        referenceQuerySetList = referenceQuerySetList.distinct('reference')

else:
    referenceQuerySetList = Reference.objects.filter(
        Q(reference__istartswith=currentInitial) |
        Q(reference__istartswith=currentInitial.lower())
        ).distinct('reference')

compoundCount = Reference.objects.values('reference').annotate(c=Count('reference'))
```

Listing 6 QuerySet code of reference list search

[23] http://african-compounds.org/nanpdb/references_list/



A B C D E F G H I J K L M N O P Q R S T U V W X Y Z All

Page 1 of 6. next

130 results			
Reference	Title	PMID	#Compounds
Journal of Agricultural and Food Chemistry,2013,61(21),5080-5088	Chemical composition and antioxidant activity of Algerian propolis	23650897	20
Journal of Agricultural and Food Chemistry,2015,63(7),1990-1995	Phenolic profile characterization of Chemlali olive stones by liquid chromatography-ion trap mass spectrometry	None	11
Journal of Agricultural and Social Sciences,2012,8,24-28	Antibacterial sensitivity for some chemically diverse steroidal glycosides in vitro	None	3

Field 1: The NANPDB references have been arranged alphabetically, Field (1). By clicking any of the letters (A-Z) or All, all references which begin with a particular letter will be shown.

Field 2: shows the results for references beginning with letter *M*. By clicking on a particular reference, it navigates to the reference, e.g., on the journal website.

Field 3: gives the title of the article, thesis or conference paper from the corresponding reference.

Field 4: The PubMed ID of the article is shown on Field (4), which upon clicking will navigate to the corresponding article on the PubMed website.

Field 5: The number on Field (5) indicates the number of compounds curated from the corresponding title.

Fig. 3.35 Reference list search

3.3.5.2.4 Species list

Species list[24] search can be utilized to know the number of compounds associated with a species is to be known. The code listing 7 lists the species list, the results can be seen in the figure 3.36. Further, the description of the source species is linked to available databases; for example, some plant sources have been linked to the Prota [164] and Tropicos [165] databases.

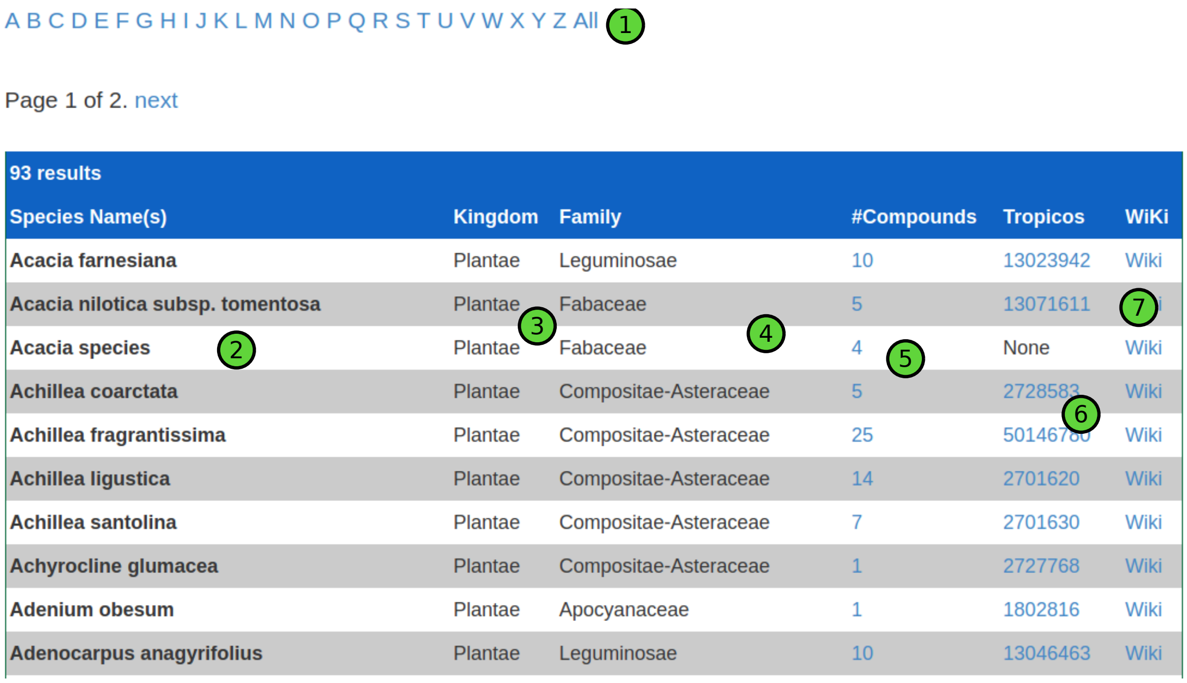
```python
if currentInitial == "All":
        sourceQuerySetList = sourceQuerySetList.distinct('source_species')

else:
    sourceQuerySetList = Sources.objects.filter(
        Q(source_species__istartswith=currentInitial) |
        Q(source_species__istartswith=currentInitial.lower())
        ).distinct('source_species')

compoundCount = \
    Sources.objects.values('source_species').annotate(c=Count('source_species'))
```

Listing 7 QuerySet code of species list search

[24] http://african-compounds.org/nanpdb/species_list/



A B C D E F G H I J K L M N O P Q R S T U V W X Y Z All

Page 1 of 2. next

93 results

Species Name(s)	Kingdom	Family	#Compounds	Tropicos	WiKi
Acacia farnesiana	Plantae	Leguminosae	10	13023942	Wiki
Acacia nilotica subsp. tomentosa	Plantae	Fabaceae	5	13071611	Wiki
Acacia species	Plantae	Fabaceae	4	None	Wiki
Achillea coarctata	Plantae	Compositae-Asteraceae	5	2728583	Wiki
Achillea fragrantissima	Plantae	Compositae-Asteraceae	25	50146780	Wiki
Achillea ligustica	Plantae	Compositae-Asteraceae	14	2701620	Wiki
Achillea santolina	Plantae	Compositae-Asteraceae	7	2701630	Wiki
Achyrocline glumacea	Plantae	Compositae-Asteraceae	1	2727768	Wiki
Adenium obesum	Plantae	Apocyanaceae	1	1802816	Wiki
Adenocarpus anagyrifolius	Plantae	Leguminosae	10	13046463	Wiki

Field 1: The NANPDB species have also been arranged alphabetically, Field (1). By clicking any of the letters (A-Z) or All, species whose names begin with that letter will be shown in the results.

Field 2: Species names are listed in Field (2). Upon clicking the species name, the detailed information on the species in the species card is provided.

Field 3: The respective kingdoms of the corresponding species are displayed on Field (3).

Field 4: The respective families of the corresponding species are displayed on Field (4).

Field 5: shows the number of compounds contained in the corresponding species.

Field 6: Provides a link to the Tropicos [165] sources database.

Field 7: provides the link to the Wikipedia page describing the species, if available.

Fig. 3.36 Species list search

3.3.5.2.5 Authors list

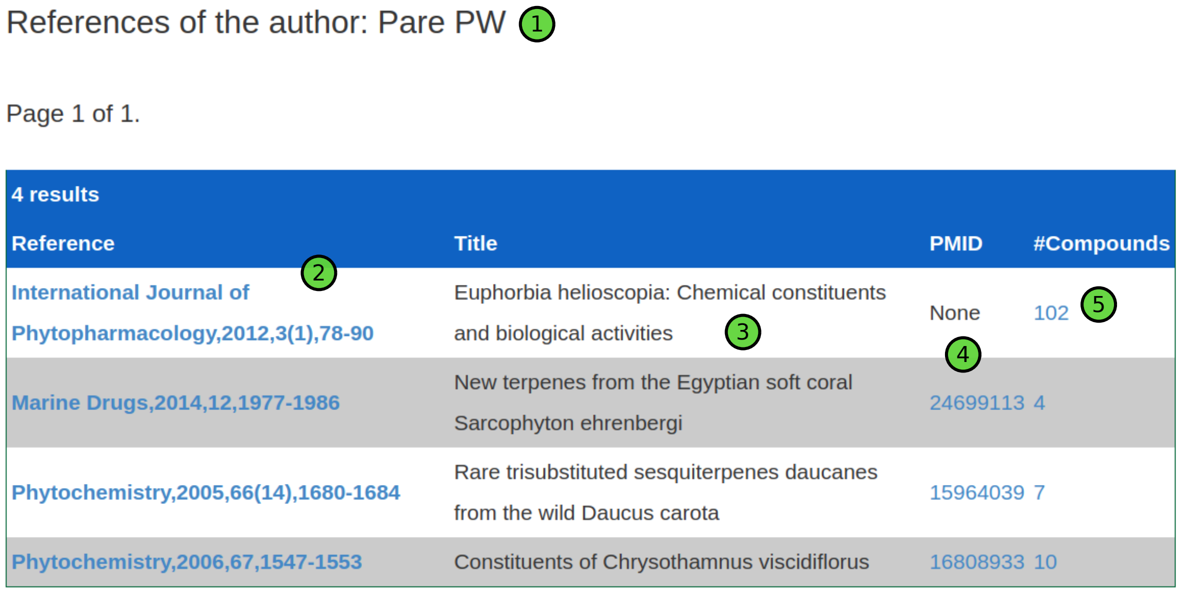
Authors list search can be utilized to know the number of references and the number of compounds associated with the queried author. This is available via compound card (Section 3.3.5.3.1) on clicking the author name. The code listing 8 lists the authors list, the results can be seen in the figure 3.37.

```
referenceQuerySetList = Reference.objects.filter(
        molecule__msridtopersons__person_id=personID).distinct('reference')
author = Persons.objects.get(person_id=personID)
author = author.person_name

compoundCount = Reference.objects.values('reference').annotate(c=Count('reference'))

table, compoundCountOfReferencesList = \
    getReferencesResultTable(referenceQuerySetList)
```

Listing 8 QuerySet code of author list search



References of the author: Pare PW

Page 1 of 1.

4 results

Reference	Title	PMID	#Compounds
International Journal of Phytopharmacology,2012,3(1),78-90	Euphorbia helioscopia: Chemical constituents and biological activities	None	102
Marine Drugs,2014,12,1977-1986	New terpenes from the Egyptian soft coral Sarcophyton ehrenbergi	24699113	4
Phytochemistry,2005,66(14),1680-1684	Rare trisubstituted sesquiterpenes daucanes from the wild Daucus carota	15964039	7
Phytochemistry,2006,67,1547-1553	Constituents of Chrysothamnus viscidiflorus	16808933	10

Field 1: On each compound card and species card, a list of authors and curators is provided. To see compounds which were identified by a particular author, click on the author's name on the compound card or species card. Author's can also be done directly with author's name on the keywords search. Only the author's surnames (last names) are provided. The other names are abbreviated. In this example, the author's name is Elgamal HMA, Field (1).

Field 2: provides all references corresponding to the selected author.

Field 3: provides the title of the reference.

Field 4: gives the PubMed ID of the reference. Upon clicking, it redirects to the article page in the PubMed website.

Field 5: indicates the number of compounds curated from the corresponding reference/title.

Fig. 3.37 Authors list search

3.3.5.3 Cards

Cards are the detailed information of compound or source species.

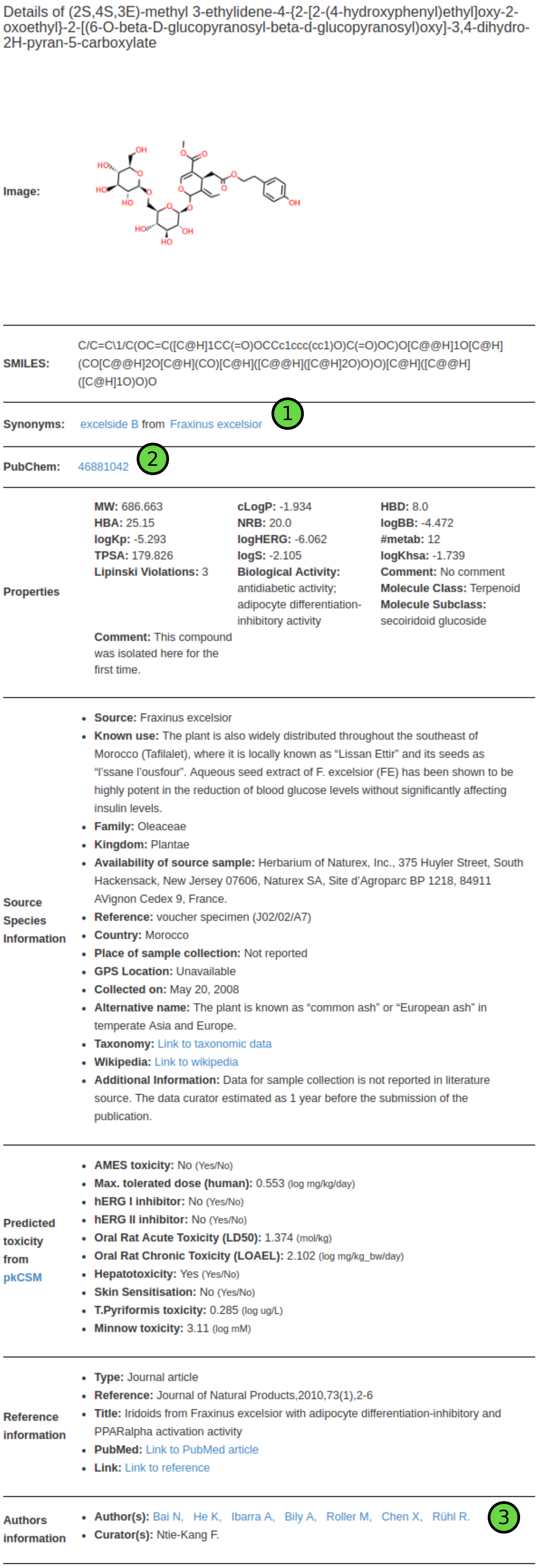
3.3.5.3.1 Compound card

Following are the fields marked in Compound card (Figure 3.38)

Field 1: To see that the number of synonyms of the given compound is displayed along with its source species information. One can click on either the highlighted synonym name or its species names (in blue) to access the additional information.

Field 2: shows the PubChem ID of the compound. Upon clicking this link, it will navigate to the PubChem page, where additional information about this compound can be accessed. It is followed by computed physicochemical information (from QikProp), source species information, the predicted toxicity data (from pkCSM) and the reference information.

Field 3: shows the information about the authors of the article from which the compound is obtained. Upon clicking on the author's name, it redirects to the author's references list.



Details of (2S,4S,3E)-methyl 3-ethylidene-4-{2-[2-(4-hydroxyphenyl)ethyl]oxy-2-oxoethyl}-2-[(6-O-beta-D-glucopyranosyl-beta-d-glucopyranosyl)oxy]-3,4-dihydro-2H-pyran-5-carboxylate

Image:

SMILES: C/C=C\1/C(OC=C([C@H]1CC(=O)OCCc1ccc(cc1)O)C(=O)OC)O[C@@H]1O[C@H](CO[C@@H]2O[C@H](CO)[C@H]([C@@H]([C@H]2O)O)O)[C@H]([C@@H]([C@H]1O)O)O

Synonyms: excelside B from Fraxinus excelsior

PubChem: 46881042

Properties

MW: 686.663
HBA: 25.15
logKp: -5.293
TPSA: 179.826
Lipinski Violations: 3

cLogP: -1.934
NRB: 20.0
logHERG: -6.062
logS: -2.105
Biological Activity: antidiabetic activity; adipocyte differentiation-inhibitory activity

HBD: 8.0
logBB: -4.472
#metab: 12
logKhsa: -1.739
Comment: No comment
Molecule Class: Terpenoid
Molecule Subclass: secoiridoid glucoside

Comment: This compound was isolated here for the first time.

Source Species Information

- **Source:** Fraxinus excelsior
- **Known use:** The plant is also widely distributed throughout the southeast of Morocco (Tafilalet), where it is locally known as "Lissan Ettir" and its seeds as "l'ssane l'ousfour". Aqueous seed extract of F. excelsior (FE) has been shown to be highly potent in the reduction of blood glucose levels without significantly affecting insulin levels.
- **Family:** Oleaceae
- **Kingdom:** Plantae
- **Availability of source sample:** Herbarium of Naturex, Inc., 375 Huyler Street, South Hackensack, New Jersey 07606, Naturex SA, Site d'Agroparc BP 1218, 84911 AVignon Cedex 9, France.
- **Reference:** voucher specimen (J02/02/A7)
- **Country:** Morocco
- **Place of sample collection:** Not reported
- **GPS Location:** Unavailable
- **Collected on:** May 20, 2008
- **Alternative name:** The plant is known as "common ash" or "European ash" in temperate Asia and Europe.
- **Taxonomy:** Link to taxonomic data
- **Wikipedia:** Link to wikipedia
- **Additional Information:** Data for sample collection is not reported in literature source. The data curator estimated as 1 year before the submission of the publication.

Predicted toxicity from pkCSM

- **AMES toxicity:** No (Yes/No)
- **Max. tolerated dose (human):** 0.553 (log mg/kg/day)
- **hERG I inhibitor:** No (Yes/No)
- **hERG II inhibitor:** No (Yes/No)
- **Oral Rat Acute Toxicity (LD50):** 1.374 (mol/kg)
- **Oral Rat Chronic Toxicity (LOAEL):** 2.102 (log mg/kg_bw/day)
- **Hepatotoxicity:** Yes (Yes/No)
- **Skin Sensitisation:** No (Yes/No)
- **T.Pyriformis toxicity:** 0.285 (log ug/L)
- **Minnow toxicity:** 3.11 (log mM)

Reference information

- **Type:** Journal article
- **Reference:** Journal of Natural Products,2010,73(1),2-6
- **Title:** Iridoids from Fraxinus excelsior with adipocyte differentiation-inhibitory and PPARalpha activation activity
- **PubMed:** Link to PubMed article
- **Link:** Link to reference

Authors information

- **Author(s):** Bai N, He K, Ibarra A, Bily A, Roller M, Chen X, Rühl R.
- **Curator(s):** Ntie-Kang F.

Fig. 3.38 Compound card

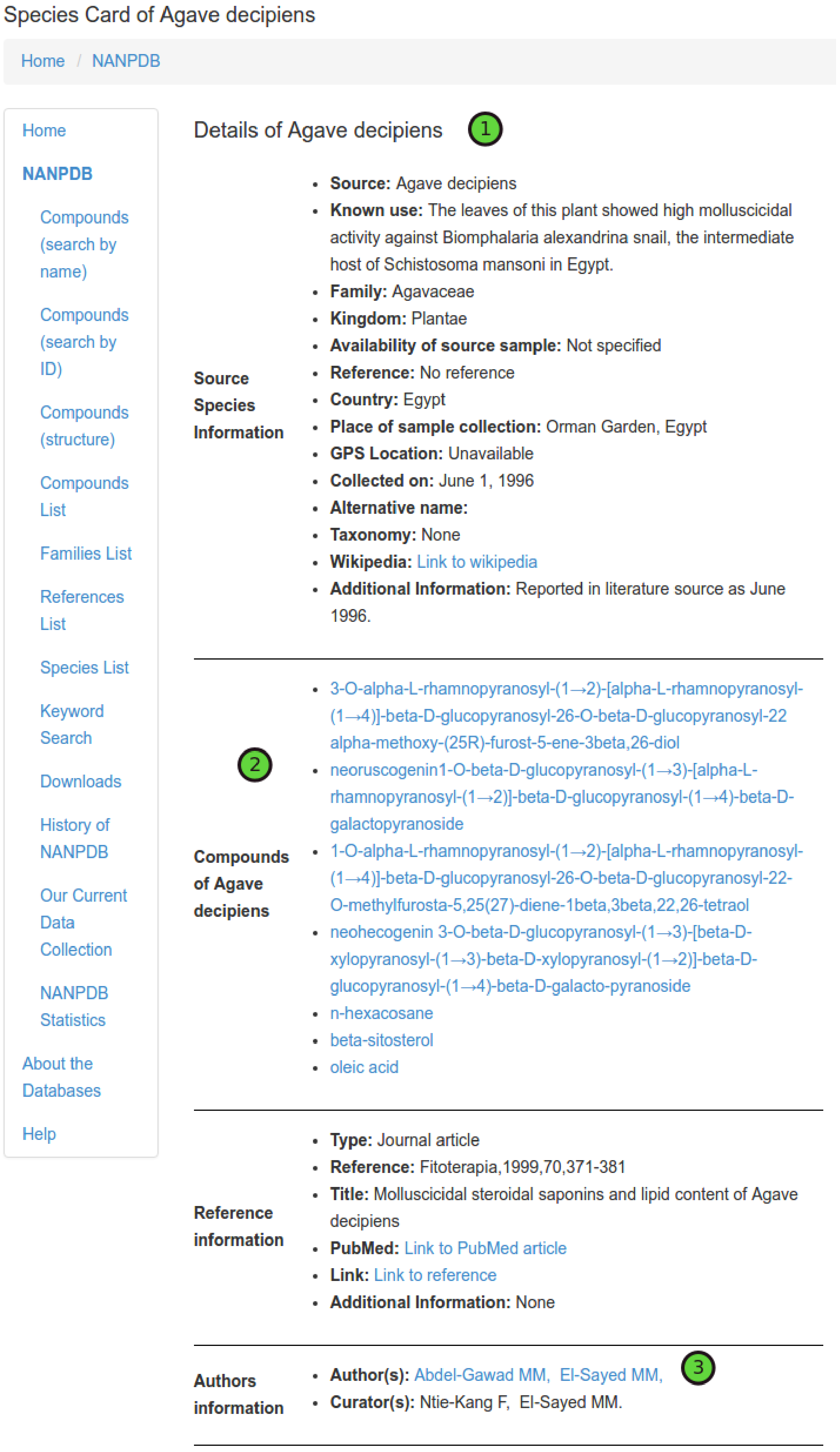
3.3.5.3.2 Species card

Following are the fields marked in Species card 3.39

Field 1: gives the name of the selected species.

Field 2: gives the names of all compounds identified from the selected species. Each compound card can be accessed by simply clicking on the highlighted compound links.

Field 3: shows the names of the authors of the article from which information about the selected species is taken. On clicking shows the author's references.



Species Card of Agave decipiens

Home / NANPDB

- Home
- NANPDB
  - Compounds (search by name)
  - Compounds (search by ID)
  - Compounds (structure)
  - Compounds List
  - Families List
  - References List
  - Species List
  - Keyword Search
  - Downloads
  - History of NANPDB
  - Our Current Data Collection
  - NANPDB Statistics
- About the Databases
- Help

Details of Agave decipiens

Source Species Information

- **Source:** Agave decipiens
- **Known use:** The leaves of this plant showed high molluscicidal activity against Biomphalaria alexandrina snail, the intermediate host of Schistosoma mansoni in Egypt.
- **Family:** Agavaceae
- **Kingdom:** Plantae
- **Availability of source sample:** Not specified
- **Reference:** No reference
- **Country:** Egypt
- **Place of sample collection:** Orman Garden, Egypt
- **GPS Location:** Unavailable
- **Collected on:** June 1, 1996
- **Alternative name:**
- **Taxonomy:** None
- **Wikipedia:** Link to wikipedia
- **Additional Information:** Reported in literature source as June 1996.

Compounds of Agave decipiens

- 3-O-alpha-L-rhamnopyranosyl-(1→2)-[alpha-L-rhamnopyranosyl-(1→4)]-beta-D-glucopyranosyl-26-O-beta-D-glucopyranosyl-22 alpha-methoxy-(25R)-furost-5-ene-3beta,26-diol
- neoruscogenin1-O-beta-D-glucopyranosyl-(1→3)-[alpha-L-rhamnopyranosyl-(1→2)]-beta-D-glucopyranosyl-(1→4)-beta-D-galactopyranoside
- 1-O-alpha-L-rhamnopyranosyl-(1→2)-[alpha-L-rhamnopyranosyl-(1→4)]-beta-D-glucopyranosyl-26-O-beta-D-glucopyranosyl-22-O-methylfurosta-5,25(27)-diene-1beta,3beta,22,26-tetraol
- neohecogenin 3-O-beta-D-glucopyranosyl-(1→3)-[beta-D-xylopyranosyl-(1→3)-beta-D-xylopyranosyl-(1→2)]-beta-D-glucopyranosyl-(1→4)-beta-D-galacto-pyranoside
- n-hexacosane
- beta-sitosterol
- oleic acid

Reference information

- **Type:** Journal article
- **Reference:** Fitoterapia,1999,70,371-381
- **Title:** Molluscicidal steroidal saponins and lipid content of Agave decipiens
- **PubMed:** Link to PubMed article
- **Link:** Link to reference
- **Additional Information:** None

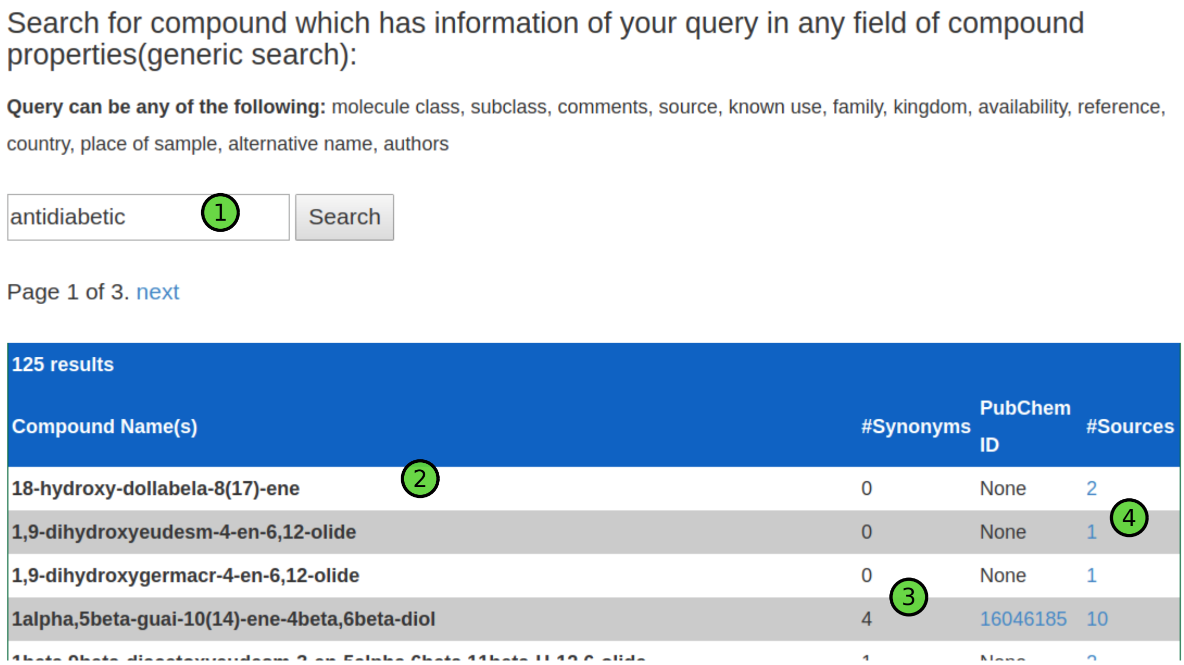
Authors information

- **Author(s):** Abdel-Gawad MM, El-Sayed MM,
- **Curator(s):** Ntie-Kang F, El-Sayed MM.

Fig. 3.39 Species card

3.3.5.4 Keyword search

Keyword search[25] is useful when the term has to be searched in entire database irrespective of which field because *keyword search* searches for the queried term in all fields of the database. The result of keyword search will give properties of the compound.



Search for compound which has information of your query in any field of compound properties(generic search):

Query can be any of the following: molecule class, subclass, comments, source, known use, family, kingdom, availability, reference, country, place of sample, alternative name, authors

antidiabetic

Page 1 of 3. next

125 results

Compound Name(s)	#Synonyms	PubChem ID	#Sources
18-hydroxy-dollabela-8(17)-ene	0	None	2
1,9-dihydroxyeudesm-4-en-6,12-olide	0	None	1
1,9-dihydroxygermacr-4-en-6,12-olide	0	None	1
1alpha,5beta-guai-10(14)-ene-4beta,6beta-diol	4	16046185	10
1beta,9beta-diacetoxyeudesm-3-en-5alpha,6beta,11betaH-12,6-olide	1	None	2

Field 1: In this field, any keyword can be entered in Field (1).

Field 2: By clicking a compound in the column of Field (2), the page will navigate to the compound card of the corresponding compound.

Field 3: The number on Field (3) indicates how many synonyms exist for the corresponding molecule. To view, those synonyms click the number.

Field 4: The number on Field (4) indicates from the number of sources from which the corresponding molecule or its synonyms are obtained. To view, those sources click the number.

Fig. 3.40 Keyword search

[25] http://african-compounds.org/nanpdb/keyword_search/

3.3.5.5 Similarity search

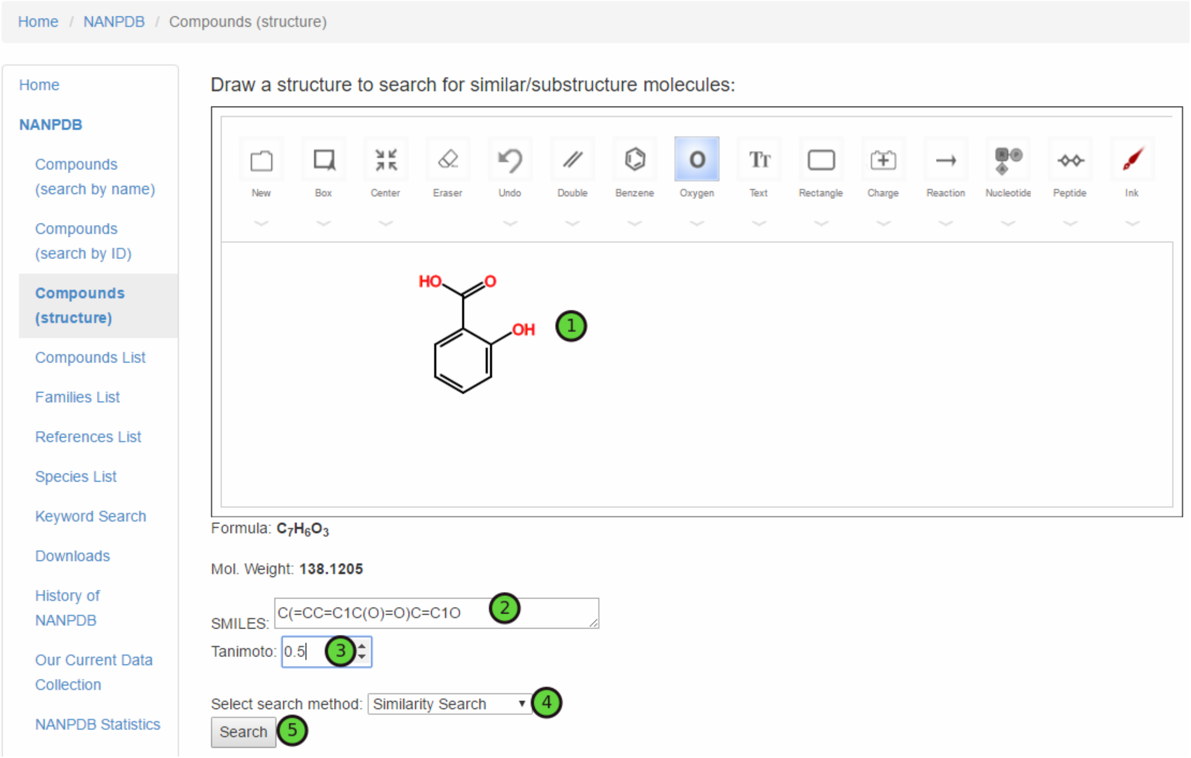
The user can seek to the similarity search[26] when similar compounds are required. For the queried compound NANPDB searches all the compounds using fingerprints technique (Section 2.2.1) here we have chosen Circular Morgan (Table 2.1) fingerprint technique. The steps are listed in figure 3.41.

3.3.5.6 Sub-structure search

In the Substructure search[27] module, one can search for molecules with a particular substructure or for a fragment of a compound. A substructure search is performed in a similar way to the Similarity search. The only difference is that at field 4 of the Similarity search (Figure 3.41), substructure search instead of Similarity search is chosen.

[26] http://african-compounds.org/nanpdb/compounds_structure/

[27] http://african-compounds.org/nanpdb/compounds_structure/



Draw a structure to search for similar/substructure molecules:

Formula: $C_7H_6O_3$

Mol. Weight: 138.1205

SMILES: C(=CC=C1C(O)=O)C=C1O

Tanimoto: 0.5

Select search method: Similarity Search

Field 1: To draw the structure for either the similarity or the substructure search. The structure depicted on the Field (1) window is 2-hydroxybenzoic acid. It can be illustrated by picking the benzene ring on the top panel, followed by the single bond, then the oxygen atom. The carboxylic group can be drawn by clicking on the single bond, then attaching the single bond, double bond, and oxygen atoms, placing each one at the appropriate position.

Field 2: This automatically generates the SMILES string for the drawn molecule on Field (2).

Field 3: a preferred Tanimoto filter value can be given, which is useful for similarity searches.

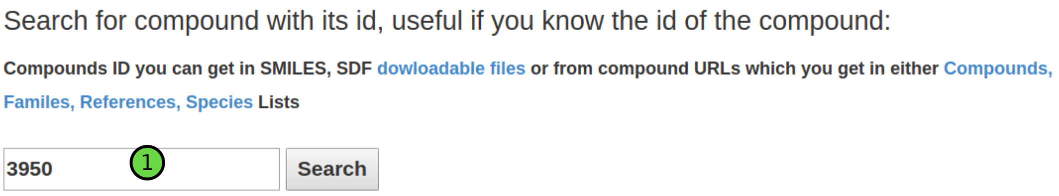
Field 4: Selecting either the similarity or substructure search after drawing the desired molecule.

Field 5: Clicking the Search button submits the query.

Fig. 3.41 Similarity and substructure search

3.3.5.7 ID search

In the ID search[28] known NANPDB id can be directly given to see its compound card. NANPDB ids are availble in the SMILES list which can downloaded[29].



Search for compound with its id, useful if you know the id of the compound:

Compounds ID you can get in SMILES, SDF dowloadable files or from compound URLs which you get in either Compounds, Familes, References, Species Lists

3950

Field 1: Each molecule has a compound ID which is given in the download file. The ID can be used in Field (1) to access a molecule directly.

Fig. 3.42 Compound ID search

[28] http://african-compounds.org/nanpdb/msrid_search/
[29] http://african-compounds.org/nanpdb/downloads/

3.4 SeMPI

SeMPI is a genome-based secondary metabolite prediction and identification web server[30]. SeMPI identifies encoding gene clusters from genomic data and predicts the basic structure of related natural products synthesized by polyketide synthases of type I modular.

3.4.1 Idea of SeMPI

Natural products (Section 1.2) which are primary sources of drugs and bioactive compounds and play an essential role in drug discovery. Conventionally use of chemical, and analytical methods contributed to cultivating antibiotic microorganisms. Since the arrival of HGP and the amount of whole-genome sequencing technologies, the field of natural compounds is attempting to explore the new direction of omics technologies (genomics, transcriptomics, proteomics, and metabolomics) based methods to locate, distinguish and describe lead molecules for drug discovery. This lead to the development of several genome mining and computational tools in the identification of biosynthetic gene clusters [203]. The prediction of new compounds based on genome mining become ever more significant seeing the trends of the number of bases and sequence records submitted each year to GenBank. From the graphs (Figures 3.43 and 3.44), it is evident that since ascertaining the sequence of nucleotide bases that make up human genetic blueprint (around the year 2003) a drastic increase in determining gene sequences of various species. The graph also indicates about the *Shotgun sequencing* is a technique used for sequencing longer DNA strands, in this technique DNA is split into several random smaller segments to produce a representative library of the more extended DNA strand, in the end, a final sequence is made up of overlapping sequence segments [204]. Adding to it the decreasing costs of DNA sequencing (Figure 3.45) contributing to shifting in the paradigm, the new paradigm utilizes biosynthetic methods. Natural products are the building blocks of biosynthetic methods. However, instability of intermediates in biosynthetic pathways are still larger hurdles [205]. Many natural products which are known to be secondary metabolites provided by polyketides having a significant

[30] http://phabi.de/sempi

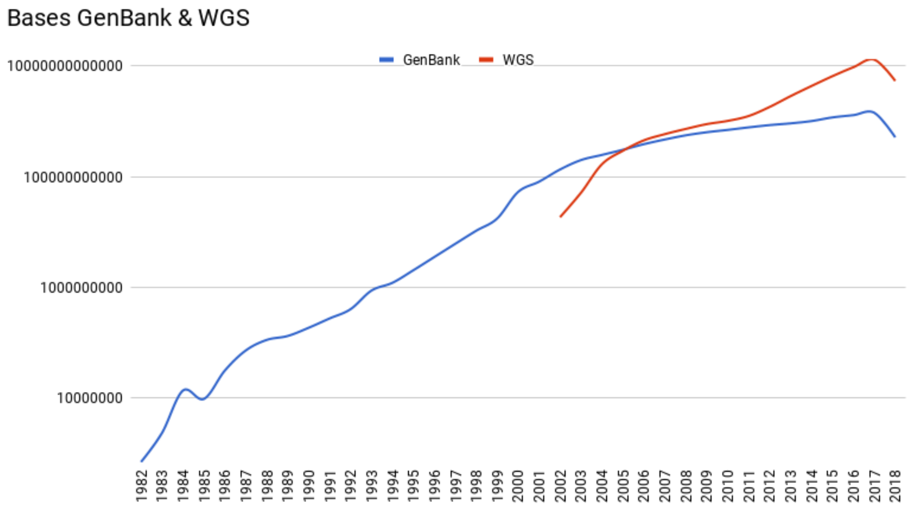


Fig. 3.43 GenBank and Whole Genome Shotgun sequences (WGS) Statistics are showing the trend of increasing number of Bases submitted each year to GenBank [171].

relationship with fatty acid biosynthesis and have the important therapeutic effects. The scientific enthusiasm in natural products accelerated resolve structures of compounds yielding from them, but identification of their engaged genome clusters and structure elucidation of the actual secondary metabolite is a challenging task [206, 207]. The built web server Secondary Metabolite Prediction and Identification Pipeline (SeMPI) identify genome clusters and predict structures also show corresponding matched similar natural compounds from StreptomeDB 2.0, allowing for a more thorough investigation.

3.4.2 SeMPI pipeline

Our group members Paul F Zierep[31] and Natàlia Padilla Sirera[32] majorly accomplish the implementation of SeMPI [206] pipeline. SeMPI adopts a streamlined pipeline to predict best matching compounds for the given gene cluster (Section 1.3). The flowchart (Figure 3.46) begins with accepting raw DNA code in the formats of FASTA, GenBank files. With the received data antiSMASH 3.0 [209] identifies the PKS Gene cluster. *Polyketide synthases* (PKS) are the multi-domain enzymes that produce *Polyketides* a class of secondary metabolites produced

[31] http://phabi.de/main/members/ziereppaul/
[32] http://phabi.de/main/members/

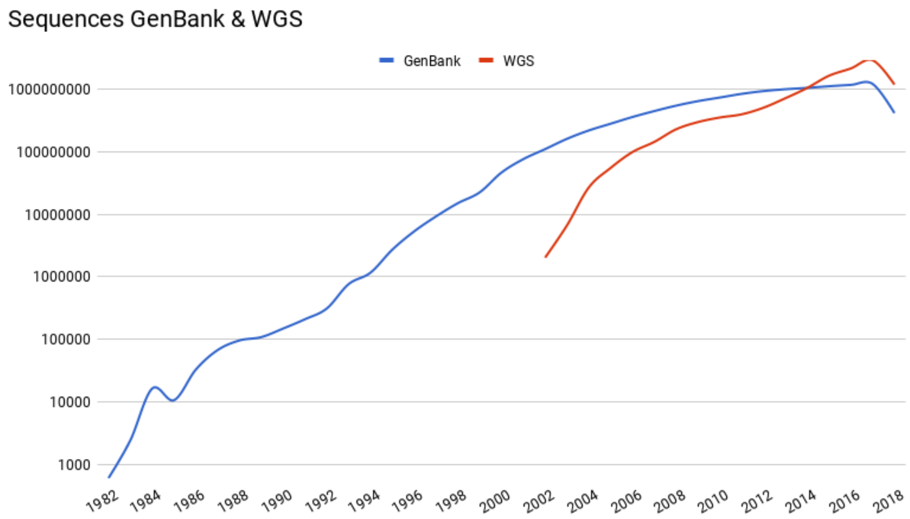


Fig. 3.44 GenBank and Whole Genome Shotgun sequences (WGS) Statistics are showing the trend of increasing number of Sequences submitted each year to GenBank [171].

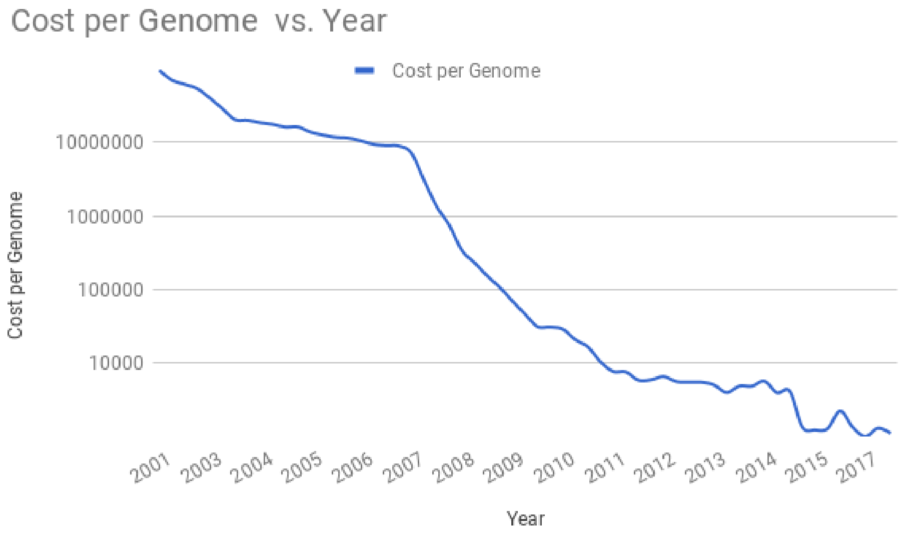


Fig. 3.45 DNA Sequencing costs changing trends [208]

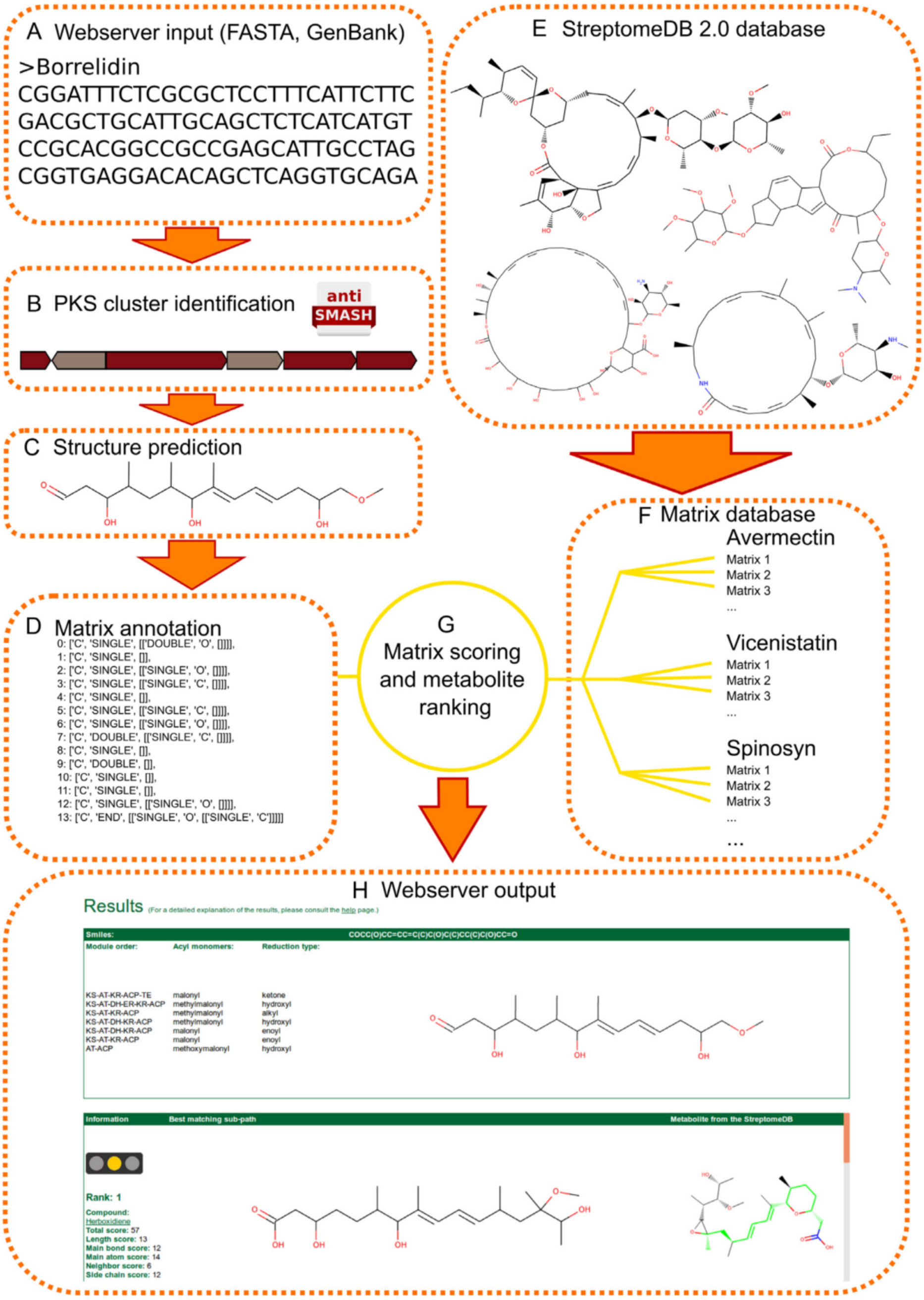
by living organisms for their survival. Polyketides are a diverse group of compounds that are produced by similar synthesis pathways and are of three types, SeMPI identifies type-I PKS which can be divided into structured modules and are also a well-structured building-block composition that eases during product formation [210, 211]. Polyketide chain prediction step follows with a dual step involving prediction of PK-chain and identification of putative molecule scaffolds that fit the predicted PK-chain (Figure 3.46 C). The predicted PK-structure represents the initial PKS biosynthesis without any cyclization or further post-modifications. The calculated paths of the molecule are stored in matrix format (Figure 3.46 D) which enables quick comparison. To limit the possible structures of PKS products, we developed an automated workflow which compares the predicted polyketide chain to pre-processed annotated natural products in a large database of natural products (StreptomeDB 2.0 [22], Figure 3.46 E). Creation of another matrix annotation of the predicted molecule from the NP database via the same algorithm (Figure 3.46 D) and the results are stored in the matrix again for the predicted compound (Figure 3.46 F). Comparisons of both matrices are scored (Figure 3.46 G) according to the maximum common sub-paths within each molecule utilizes the RDKit Chem module (Section 2.3.1) MCS package. Finally ten best matching compounds along with their domain information.

3.4.3 Implementation of website

The SeMPI[33] project is an in-house developed API based content management system (CMS). The CMS has been developed using Django web framework (Section 2.5.3). The CMS and its API is optimized, and its adaptation applied to multiple databases ([212, 22, 206]) that exist in our group, it is small and modular according to the requirements. RDKit (Section 2.3.1) an open-source toolkit for cheminformatics supplied definitions, subroutines, and protocols to access the structures of the natural compound from StreptomeDB.

Apache a free and open-source cross-platform web server assisted in serving the data via Hypertext Transfer Protocol (HTTP). Apache can work in wide variety of platforms and environments through its modular design and has several ways to implement its features.

[33] http://phabi.de/sempi



```
>Borrelidin
CGGATTTCTCGCGCTCCTTTCATTCTTC
GACGCTGCATTGCAGCTCTCATCATGT
CCGCACGGCCGCCGAGCATTGCCTAG
CGGTGAGGACACAGCTCAGGTGCAGA
```

```
0: ['C', 'SINGLE', [['DOUBLE', 'O', []]]],
1: ['C', 'SINGLE', []],
2: ['C', 'SINGLE', [['SINGLE', 'O', []]]],
3: ['C', 'SINGLE', [['SINGLE', 'C', []]]],
4: ['C', 'SINGLE', []],
5: ['C', 'SINGLE', [['SINGLE', 'C', []]]],
6: ['C', 'SINGLE', [['SINGLE', 'O', []]]],
7: ['C', 'DOUBLE', [['SINGLE', 'C', []]]],
8: ['C', 'SINGLE', []],
9: ['C', 'DOUBLE', []],
10: ['C', 'SINGLE', []],
11: ['C', 'SINGLE', []],
12: ['C', 'SINGLE', [['SINGLE', 'O', []]]],
13: ['C', 'END', [['SINGLE', 'O', [['SINGLE', 'C']]]]]
```

Results (For a detailed explanation of the results, please consult the help page.)

Smiles: COCC(O)CC=CC=C(C)C(O)C(C)CC(C)C(O)CC=O

Module order:	Acyl monomers:	Reduction type:
KS-AT-KR-ACP-TE	malonyl	ketone
KS-AT-DH-ER-KR-ACP	methylmalonyl	hydroxyl
KS-AT-KR-ACP	methylmalonyl	alkyl
KS-AT-DH-KR-ACP	methylmalonyl	hydroxyl
KS-AT-DH-KR-ACP	malonyl	enoyl
KS-AT-KR-ACP	malonyl	enoyl
AT-ACP	methoxymalonyl	hydroxyl

Information | Best matching sub-path | Metabolite from the StreptomeDB

Rank: 1

Compound:
Herboxidiene
Total score: 57
Length score: 13
Main bond score: 12
Main atom score: 14
Neighbor score: 6
Side chain score: 12

Fig. 3.46 Flowchart of the SeMPI pipeline taken from Zierep et al. [206]. (A) Data can be submitted as FASTA, GenBank files, or raw DNA code. (B) Gene cluster is identified with antiSMASH 3.0. (C) Structure of the carbon chain is predicted. (D) Carbon chain is translated into the matrix annotations. (E) StreptomeDB 2.0 provides the natural compounds for the path similarity search. (F) A set of matrices was computed for each compound. (G) The matrix from the prediction is compared with each matrix set of the database. (H) The web server output displays the prediction and domain information as well as the ten best matching compounds.

The modular design allows the web server to have a specific configuration by selecting which modules to load at compile or runtime. Apache has Multi-Processing Modules (MPMs) which are a group of parameters that are responsible for binding to network ports on the server, handling the number of requests that server can serve, and protocols for dispatching child processes to handle the web requests [213]. SeMPI runs on the web server configuration which uses event MPM which is threaded highly scalable category of MPM. The Apache server is resting on the Debian open-source operating system connected to a PostgreSQL database via the psycopg2 wrapper for database access.

Since multiple web servers run on PostgreSQL and it is an Open Source SQL database it is incredibly flexible. However, the flexibility feature becomes a drawback as there are no default settings which are already optimized. StreptomeDB (Section 3.2) requires more of accessing the static data compared to SeMPI and SeMPI requires more computing power compared to StreptomeDB. Considering the complexity of multiple databases in the design plays a vital role in fine-tuning. In fine-tuning maximum connections [214] that database can have at any given time is keenly observed for duration of time and has been adapted accordingly. Fine tuning reduces the influence of excessive lousy web bots which are small predefined web programmable scripts, and they run by tracking some keywords or location-based servers.

3.4.4 DM dataset with SeMPI

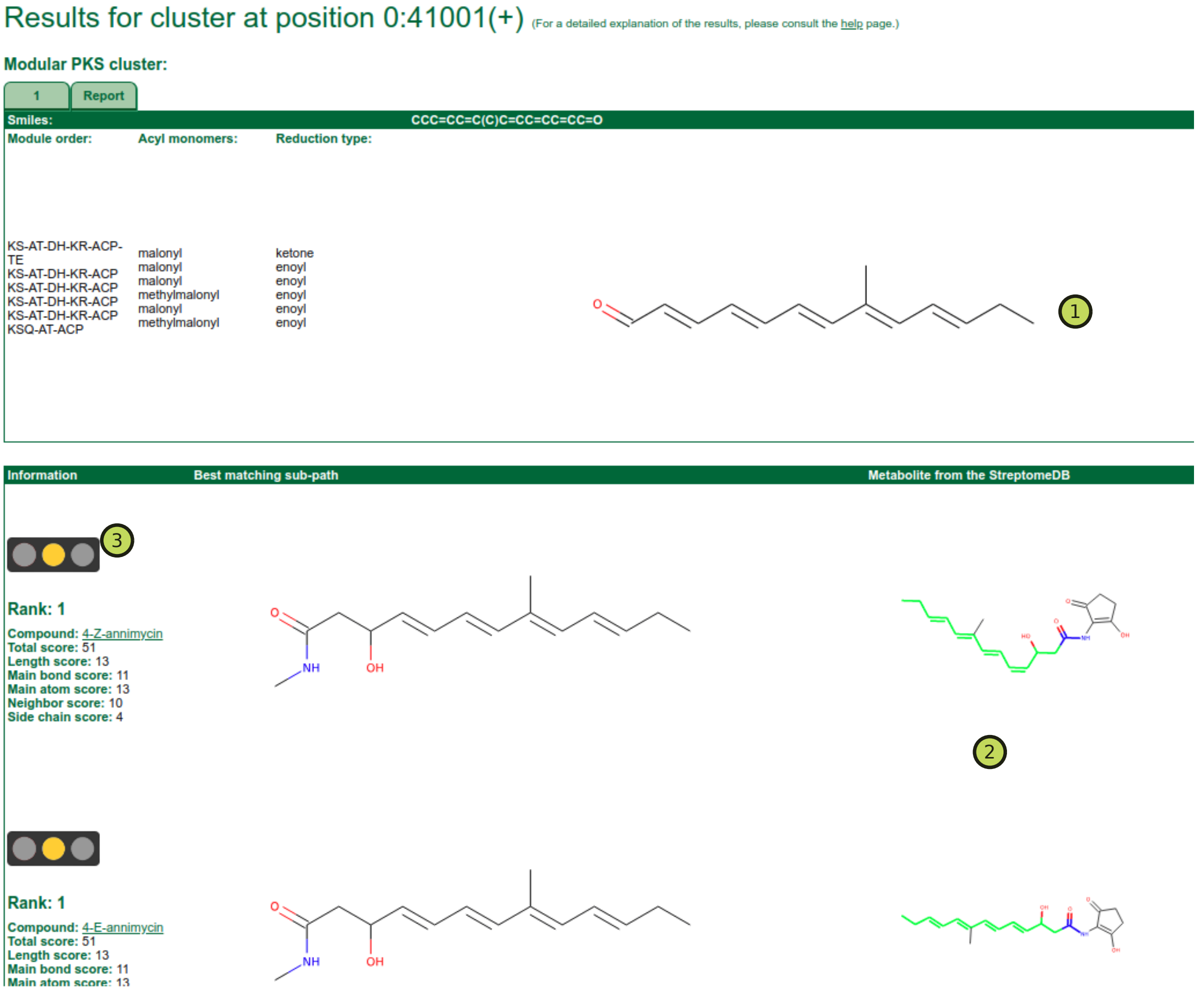
The basic functionality of SeMPI is to take genome sequences as input and identify the encoding gene clusters and predicts the basic structure of the related natural products. To understand the usage of SeMPI a converse SeMPI scheme is assumed. Using the DrugBank DM dataset from StreptomeDB (Section 3.2.4) and holding the DM compounds in StreptomeDB 2.0 (Table 3.7), examined if SeMPI predicted this DM related secondary metabolites from *Streptomycetes* genomic data acquired from MIBiG [215] database. Minimal Information about a Biosynthetic Gene cluster (MIBiG) is a data standard which facilitates consistent and systematic deposition and retrieval of metabolic data on biosynthetic gene clusters. Regarding the input sequences of the top ten DM similar StreptomeDB 2.0 molecules, none of the specific strains are available in MIBiG. Here parent strains of the specific strains were considered. Few organisms had structure predictions (Table 3.18), and several others had issues such as not finding genome sequences in MIBiG, No signature of modular KS, too many NRPS signatures, no gene cluster found, and flaws in input sequences. A complete list of results are in appendix D.1. SeMPI could predict five among 21 found DNA sequences. Further observing results eleven had issues during the polyketide chain prediction individually while screening for domain signatures of modular ketoacyl synthase (KS) after antiSMASH 3.0 screened for gene clusters of *PKS type I* and three of them had too many Nonribosomal peptide synthetases (NRPS) signatures. For SeMPI it is essential to have KS domain to select the start codon which is considered to be 30-40 residues from KS domain [216], and it is the length of N-terminal docking domain. In five DNA sequences, antiSMASH 3.0 could not find gene clusters of *PKS type I*.

StDB ID	**StDB compound name**	**StDB organism**	**MIBiG accession**	**MIBiG organism**	**SeMPI result**
3313	Adiposin 2	*Streptomyces calvus TM-521*	BGC0001297	*Streptomyces calvus*	No signature of modular KS, Too many NRPS signatures
			BGC0001298		Predicted
			BGC0001387		No gene cluster found
5208	CHEMBL1268	*Streptomyces hygroscopicus UC 11099*	BGC0000345	*Streptomyces hygroscopicus*	No signature of modular KS, Too many NRPS signatures
			BGC0000603		No signature of modular KS
			BGC0000066		Predicted
			BGC0000068		Predicted
			BGC0000074		Predicted
			BGC0000388		No signature of modular KS, Too many NRPS signatures
			BGC0000698		No gene cluster found

Table 3.18 SeMPI results of DM compounds of StreptomeDB 2.0 [22] (StDB) 1/4 other three parts of the table is in appendix D.1

For example from the table 3.18 compound *Adipoisin 2*[34] with id 3313 with source organism *Streptomyces calvus TM-521* was not available in MIBiG and *Streptomyces calvus* gene cluster is given as input to SeMPI. For *Streptomyces calvus* there was nonribosomal peptides (NRP) class cluster (BGC0001297), polyketide (PK) class cluster (BGC0001298) and lastly another one with general class (BGC0001387). For the NRP class, SeMPI reported an issue of too many NRPS. The PK class was able to predict the structure (Figure 3.47), and with general class, SeMPI reported that it could not find gene cluster. In the predicted result (Figure 3.47-1) on the top, it shows the structure predicted, and under it, two of top ten results are in the picture (Figure 3.47-2). The traffic lights (Figure 3.47-3) gives a quick visual evaluation of presented molecule quality. Here the color is yellow indicating quality between 50-100. Other colors red shows a score of <50, and green >100. However, prediction of *Adipoisin 2* compound did not happen. The explanation can be that input is parent organism and not the specific strain *Streptomyces calvus TM-521*.

[34] http://132.230.56.4/streptomedb2/get_drugcard/3313/



Results for cluster at position 0:41001(+) (For a detailed explanation of the results, please consult the help page.)

Modular PKS cluster:

1 Report

Smiles: CCC=CC=C(C)C=CC=CC=CC=O

Module order:	Acyl monomers:	Reduction type:
KS-AT-DH-KR-ACP-TE	malonyl	ketone
KS-AT-DH-KR-ACP	malonyl	enoyl
KS-AT-DH-KR-ACP	malonyl	enoyl
KS-AT-DH-KR-ACP	methylmalonyl	enoyl
KS-AT-DH-KR-ACP	malonyl	enoyl
KS-AT-DH-KR-ACP	methylmalonyl	enoyl
KSQ-AT-ACP		

Information	Best matching sub-path	Metabolite from the StreptomeDB
○●○ Rank: 1 Compound: 4-Z-annimycin Total score: 51 Length score: 13 Main bond score: 11 Main atom score: 13 Neighbor score: 10 Side chain score: 4		
○●○ Rank: 1 Compound: 4-E-annimycin Total score: 51 Length score: 13 Main bond score: 11 Main atom score: 13		

Fig. 3.47 SeMPI result of *Streptomyces calvus*, a parent organism of *Streptomyces calvus TM-521*, it being a source organism of *Adiposin 2*

3.5 Dynamic Virtual Screening

Dynamic virtual screening (DVS) offers an algorithm which understands the physicochemical properties of the chemical library and serves to narrow down the chemical space that has to be screened to identify putative drugs against a specific target.

3.5.1 Requirement of DVS

Virtual screening is a far-reaching technique used in drug discovery which can substitute time-consuming *in vitro* assays to identify inhibitors for a protein. Chemical space (Section 1.1) contains over 35 million drug-like molecules [162, 6] molecule library for virtual screening. Usually, brute-force-technique are employed, i.e., all compounds are docked one after the other. It is a time-consuming process considering the scale of data available [217]. To optimize the technique and accelerate the structural based virtual screening (Section 2.1) process we developed a rational approach called dynamic virtual screening (*DVS*).

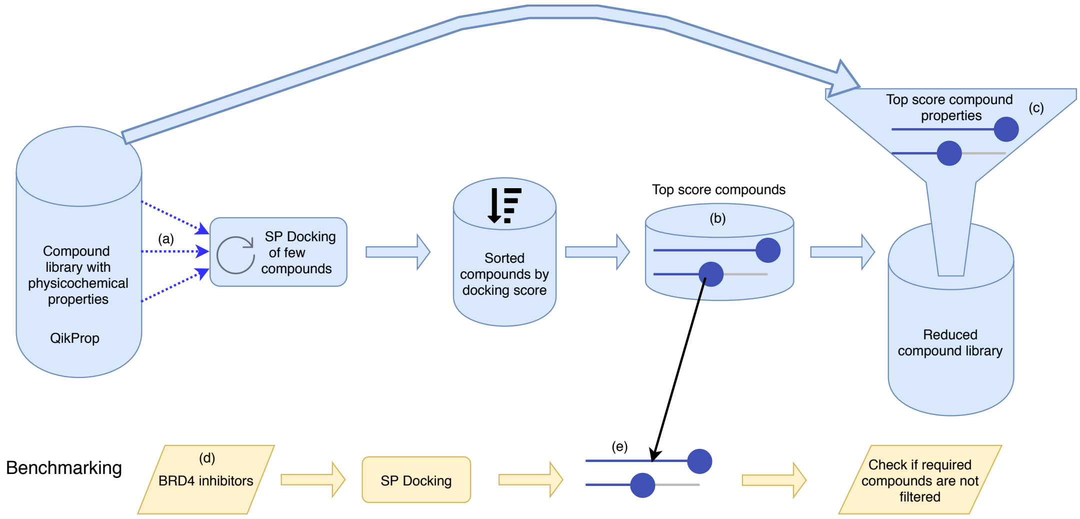

Compound library with physicochemical properties
QikProp
(a)
SP Docking of few compounds
Sorted compounds by docking score
Top score compounds
(b)
Top score compound properties
(c)
Reduced compound library
Benchmarking
(d)
BRD4 inhibitors
SP Docking
(e)
Check if required compounds are not filtered

Fig. 3.48 Dynamic Virtual Screening workflow. (a) Arbitrary compounds representing diverse library acquired from the compound library for SP docking. The step can repeatedly be performed to better-configured model. (b) Compounds arranged by their docking score and are ready to be configured by their properties to build a model. (c) The configured model applied on compounds to filter the compound library. (d) The input of BRD4 inhibitors for benchmarking DVS algorithm. (e) Model configured in the earlier flow is applied on BRD4 inhibitors to test DVS workflow

3.5.2 Acquiring data for DVS

As seen the importance of chemical space in section 1.1. In DVS first prominent step is to collect the compound library. Generation of ligand library is with *ChemicalToolBoX* [85] by merging several compound databases which are from Dictionary of Natural Products [35], the ChEMBL database [218], and the purchasable dataset of ZINC [162]. The compounds yielding from the databases and compounds sources are duly processed using LigPrep 2.4 (Schrödinger, LLC) in combined generated a dataset of 9.4 million molecules [217].

3.5.3 Preparatioin for filtering data

To filter the compounds a model is required and to generate the model representing the compounds space their physicochemical properties are needed, *QikProp* (Section 2.1.4) computes the necessary physicochemical properties of the compound library. Further, based on the docking score the screening of the library is performed. *SP Docking* (Section 2.1.7) in *in silico* evaluates the affinity of a protein-ligand complex and estimates the docking score. In this step, docking calculations were performed using the crystal structure of BRD4. The BRD4 protein belongs to BET (bromodomain and extra-terminal domain) family. The human genome encodes up to 61 different bromodomains which are part of chromatin modifying enzymes [217]. Bromodomain is a protein domain recognizing acetylated lysine residues, those on the N-terminal tails of histones [219]. Histones are globular proteins with a flexible N terminus that protrudes from the nucleosome and is subject to a wide range of potential covalent modifications that are believed to have an essential influence on chromatin structure and the accessibility of the proximal DNA to transcription [220]. This step ensures to select putative binders of BRD4. However, *SP Docking* is performed only to few compounds (Figure 3.48 a) which arbitrarily represent the complete chemical library, here we have chosen randomly distributed hundred thousand compounds. In the next step model generation with the help of physicochemical properties is implemented.

[35] http://dnp.chemnetbase.com

3.5.3.1 Physicochemical properties range configuration

In the next step the appropriate configuration of physicochemical properties (Figure 3.48 b) give the representation of high-affinity compounds. To achieve this, the SP docked compounds are ordered by their docking score to take top ten thousand high-affinity compounds along with their selected QikProp descriptors for filtering. Among 150 QikProp descriptors [56], prominent guiding descriptors (Table 3.19) are chosen to estimate the information gain of each descriptor. To each descriptor further interquartile range is calculated to consolidate information of each descriptor. Here we have taken 0.9 and 0.1 as upper and lower percentile respectively. In the end, Shanon entropy (Section 2.4.2) determines the information gain of each descriptor so that we can choose descriptors which have high information gain. It is an information theory metric, which indicates dispersion of the physicochemical property. In our method, we have taken properties which have Shannon entropy <0.7. With the help of algorithm applying the configured filter (Figure 3.48 c), the library is reduced by 18%. Steps selecting random SP docked compounds and defining quartile value of descriptors can be repeated several times until the best model (Figure 3.48 b) is available at final filtering level (Figure 3.48 c).

Property	Description
#rotor	Number of non-trivial (not CX3), non-hindered (not alkene, amide, small ring) rotatable Bonds
mol_MW	Molecular weight of the molecule
SASA	Total solvent accessible surface area (SASA) in square angstroms using a probe with a 1.40Å Radius
volume	Total solvent-accessible volume in cubic angstroms using a probe with a 1.40Å radius
donorHB	Estimated number of hydrogen bonds that would be donated by the solute to water molecules in an aqueous solution. Values are averages taken over a number of configurations, so they can be non-integer.
accptHB	Estimated number of hydrogen bonds that would be accepted by the solute from water molecules in an aqueous solution. Values are averages taken over a number of configurations, so they can be non-integer.
glob	Globularity descriptor, where r is the radius of a sphere with a volume equal to the molecular volume. Globularity is 1.0 for a spherical molecule.
QPlogPo/w	Predicted octanolwater partition coefficient.
IP(eV)	PM3 (Parameterized Model number 3) calculated ionization potential.
EA(eV)	PM3 (Parameterized Model number 3) calculated electron affinity
PSA	Van der Waals surface area of polar nitrogen and oxygen atoms.
#NandO	Number of nitrogen and oxygen atoms.
RuleOfFive	Number of violations of Lipinski's rule of five. The rules are: mol_MW <500, QPlogPow <5, $donorHB \leq 5$, $accptHB \leq 10$. Compunds that satisfy these rules are considered drug-like. (The *five* refers to the limits, which are multiples of 5.)
#nonHatm	Number of heavy atoms (nonhydrogen atoms).
RuleOfThree	Number of violations of Jorgensen's rule of three. The three rules are: QPlogS >-5.7, QPPCaco >22 nms, Primary Metabolites <7.Compounds with fewer (and preferably no) violations of these rules are more likely to be orally available.

Table 3.19 Selected QikProp properties & descriptors taken from [56], and there more descriptors in the specified source

3.5.4 Benchmarking

To benchmark method we have chosen Bromodomain BRD4 inhibitors [221]. BET and BRD4 inhibitors (Figure 3.48 d) are the class of drugs with immunosuppressive, antitumor effects in drug development [112, 222, 220]. Drugs inhibiting BRD4 are in the later stages of clinical trials [223]. Again these compounds are subjected to Glide to generate docking scores against BRD4. The filtering of compounds is with the physicochemical and glide score model created in earlier step (Figure 3.48 b). The compounds which fall in the range of 10% lower percentile and 90% percentile of model yield minimum and maximum values for all the guiding properties (Table 3.20). Evaluating Shannon entropy (Section 2.4.2) of all the physicochemical properties resulted in ranges of properties (Table 3.20 sorted by information gain) serve as the information of the dataset we have chosen. The table shows properties Rotatable bonds (#rotor) should be in between 2 and 7, Hydrogen bond donors (donorHB) between 0 and 3, Ionization potential (IP(eV)) ≈ between 8 and 9, Van der Waals surface area (PSA) ≈ between 50 and 116, Number of Nitrogen and oxygen atoms (#NandO) between 4 and 8 were the top five ranked descriptors which had significant information. Molecular weight (mol_MW) being the least significant *shannon* (information bit). The list of BRD4 binders also had known hits [217] (Lucas_XD14, JQ1, I-BET) and was expected to be not filtered. Taking the constraint of satisfying at least three properties among top 5 ranked information bits the model is applied in benchmarking step (Figure 3.48 e). Figure 3.49 represents information bits of the resultant compounds. The column information gain (Table 3.20) tells the percentage of compounds satisfying the recommended property range values. And it indicates the model is gaining the information of all those compounds.

Property or descriptor	Range or recommended values	Minimum	Maximum	Shannon entropy	Information gain (%)
donorHB	0.0 – 6.0	0	3	0.23	96.20%
RuleOfThree	maximum is 3			0.23	96.20%
#NandO	2 – 15	4	8	0.51	88.61%
#rotor	0 – 15	2	7	0.55	87.34%
RuleOfFive	maximum is 4			0.58	86.08%
IP(eV)	7.9 – 10.5	8.383	9.469	0.67	82.28%
PSA	7.0 – 200.0	50.381	116.8445	0.70	81.01%
accptHB	2.0 – 20.0	3.5	8.5	0.77	77.22%
glob	0.75 – 0.95	0.78	0.87	0.80	75.95%
QPlogPo/w	–2.0 – 6.5	1.2545	4.485	0.80	75.95%
SASA	300.0 – 1000.0	505.05	732.25	0.82	74.68%
EA(eV)	–0.9 – 1.7	0.227	1.433	0.85	72.15%
#nonHatm		19	30	0.85	72.15%
volume	500.0 – 2000.0	851.67	1309.91	0.87	70.89%
mol_MW	130.0 – 725.0	266.33	428.52	0.89	69.62%

Table 3.20 Guiding properties of the benchmarking compounds BRD4 inhibitors [221] sorted by information gain. Information gain is the percentage of compounds which satisfy the specified range of property.

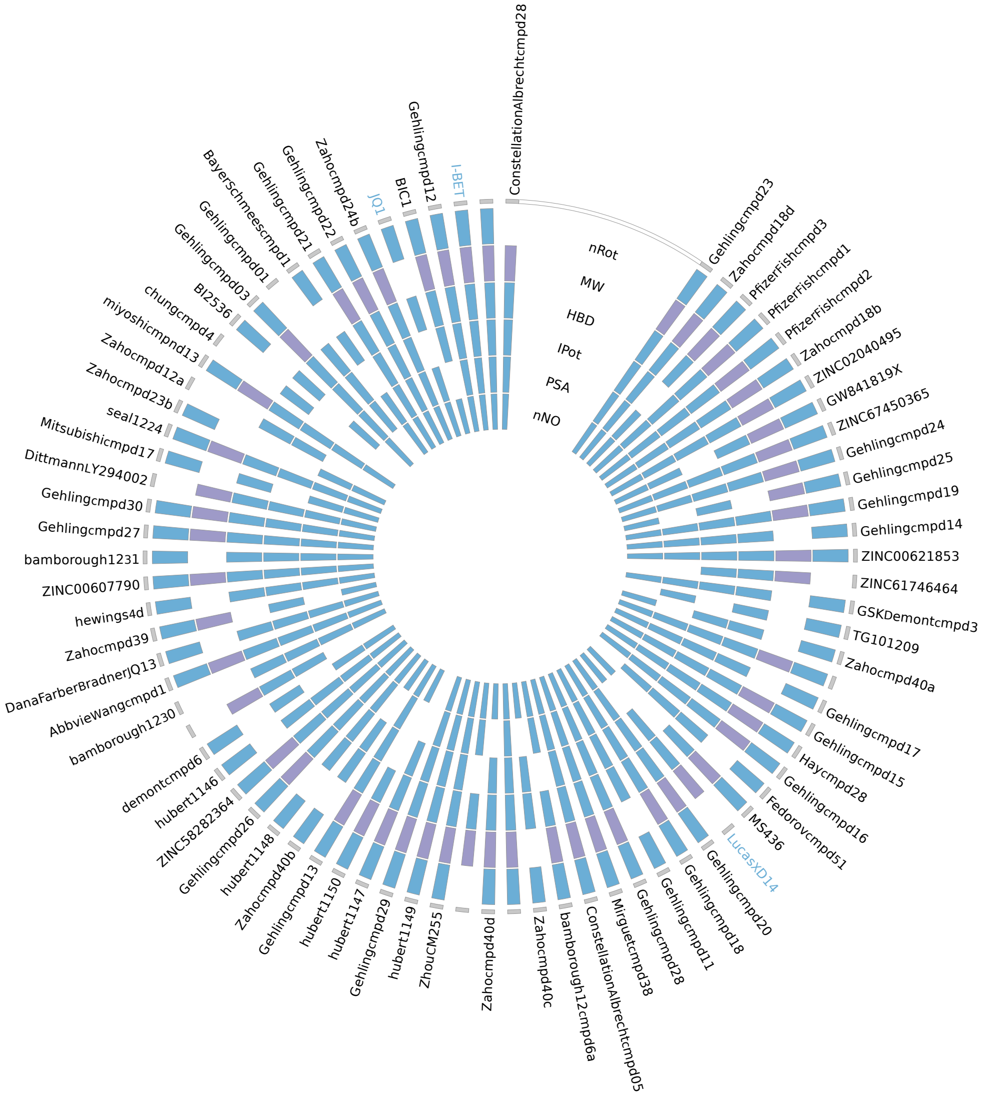


Fig. 3.49 The graph shows how the method performed against Bromodomain inhibitors. Molecular had a modest role compared to nRot, HBD, IP(eV), PSA, and nN+O. Compounds which are marked in blue color are the hits which should pass through the filter to indicate model is working. Purple color strips show which are in the range of recommended values of molecular weight and molecular weight being the least informative property for the model.

3.6 Fragment based target prediction

Fragment-based target prediction or the FragPred predicts the activity of compounds based on the knowledge of contained active substructures in the chosen chemical library to expedite the drug discovery.

3.6.1 Preamble to FragPred

The initial study in chemical space (Section 1.1) subsequently probe into the NP libraries sections 3.2 and 3.3 instigated data to foster a hypothesis that the biological pathway or activation of a protein or inhibition can result in a therapeutic effect in a disease state. The outcome of this process leads to the selection of target which further needs to be validated ere it progresses to the lead discovery phase (Section 1.5). Before the lead discovery phase compounds binding to targets with high selectivity and affinity are selected, which increases the probability of the druggable targets [224]. As discussed various methods of VS in section 2.1 and is known that *in silico* VS methods [225] dominate other ways. Ligand-based approaches mainly confide in the global structure of *druggable* molecules and their overall similarity to existing ligands. Although ligand-based methods yield reasonable results when it comes to finding structurally similar molecules, problems arise when no globally similar compounds with activity exist [47, 226]. Scaffold hopping is directed at recognizing isofunctional and structurally different compound substances, which find the maximum number of a diverse set of active compounds from the available chemical space. However, this remains a hurdle with the existing similarity methods [227, 226, 228]. Scaffold retrieval is an essential step one of such technique available in StreptomeDB 2.0 (Section 3.2.2) as a diverse set of chemotypes among the potential hits improve the possibilities for lead optimization and thus for successful completion of the drug discovery pathway [227, 226]. The hypothesis in this study is that common structures yielded from fragmentation could be a significant addition to global similarity searches in cases of globally different structures because fragments might sometimes be more relevant for target-binding than overall structure. The substructures approach can act as an incentive in drug discovery process [71, 229]. Breaking compounds into fragments is taken as an initial step in fragment informatics analysis. Fragmentation of molecules can enable several cheminformatics similarity approaches. Fragment-based drug discovery (FBDD), both in vitro and in silico, has become an import alternative and complement to HTS [230, 231]. Screening and identification of fragments that bind the target of interest from fragment libraries for higher binding affinity complete the classification of hits, which further take part in drug discovery process (Section 1.5). Moreover, the integration of chemical background knowledge information of substructures improves ligand-based similarity screenings [232]. The advent of several target prediction tools based on the compound-target interactions has increased as the chemical space in the drug research increased. In the Similarity Ensemble Approach (SEA) [233] although ligands do not share any ligands, they are appraised as similar based on the ligand annotations for their drug targets. This annotation to targets forms a collection of sets which are made independent of their size and chemical constitution using a technique similar to BLAST, which enables connection of composition sets to their corresponding protein targets in a minimal spanning tree forming a network of pharmacological space. Further, this approach propagates one that most ligand sets are related to only a few others also a majority of ligand sets are unrelated. Two, sufficient connections link almost all sets together forming a compelling chemical space. And three clustering of targets which are biologically connected blend together on the pharmacological map. There are also works which illustrated that a supervised learning prediction model could be devised between compounds and target proteins [234]. Considering the popularity of FBDD and target prediction models the question it was inquisitive to know if fragments or the substructures are also relevant for the prediction of interactions between drug-like compounds and biological targets. Integrating global similarity of compounds with fragment-based prediction improves the prediction of target at the binding pocket. The project is in its final stages of completion.

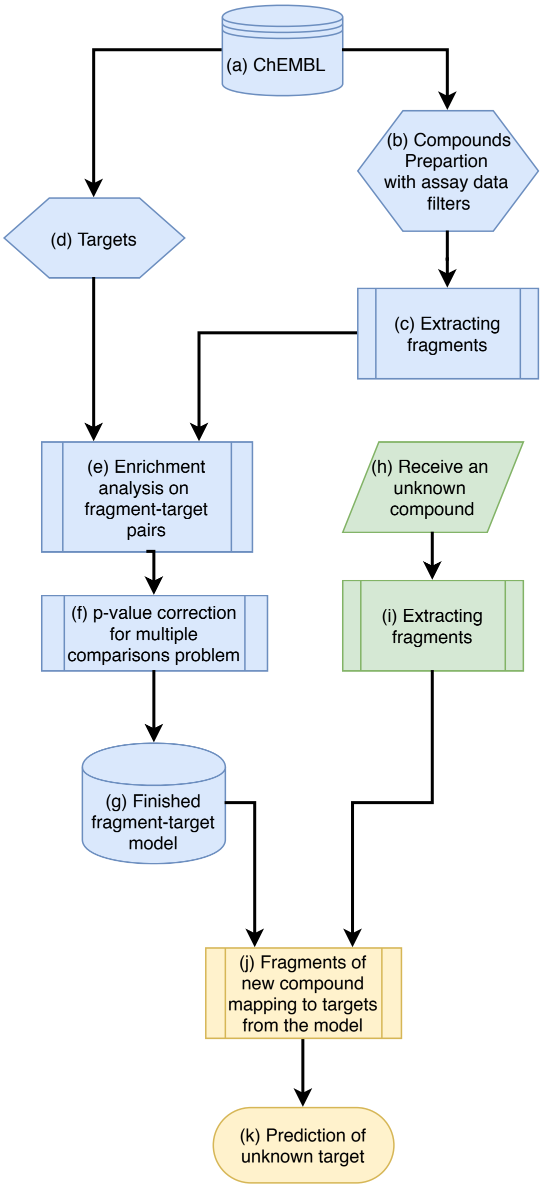

(a) ChEMBL
(b) Compounds Prepartion with assay data filters
(c) Extracting fragments
(d) Targets
(e) Enrichment analysis on fragment-target pairs
(f) p-value correction for multiple comparisons problem
(g) Finished fragment-target model
(h) Receive an unknown compound
(i) Extracting fragments
(j) Fragments of new compound mapping to targets from the model
(k) Prediction of unknown target

Fig. 3.50 Fragment based target prediction workflow.

3.6.2 Data collection

The initial task of data collection of majorly required chemical compounds, activity information, and biological targets information. We have chosen ChEMBL [235] as it is an open large-scale bioactivity database which is freely accessible and frequently maintained, the database with on an average more than one yearly update. ChEMBL data model includes data generated from pre-clinical and clinical phases of the drug discovery, specifically drug metabolism and disposition data. Learning from the failed drug candidates, ChEMBL has a mapping of drug-target annotations including clinical candidates to their therapeutic indications. The richness of assay information increases with functional or phenotypic assays which were challenging in earlier versions to indicate a molecular target are improved. The authenticity of ChEMBL increases with the inclusion of approved drugs (e.g., U.S. FDA, Orange Book (ora)). Sources such as *Journal of Medicinal Chemistry, Bioorganic Medicinal Chemistry Letters* and *Journal of Natural Products* were used for manually extracting full-text data [218, 236]. ChEMBL compounds have dose-response measurements (e.g., IC_{50}, K_i) of an affinity which particularly infuse biologically relevant entities, the response measurements have been extended to publicly available assay data of BioAssay agreeable database PubChem [150], which are exclusive to PubChem. Categorization of assays by BAO [237], improve the information strength of the data. Drug indicators or the efficacious drug-disease pairs can be a useful basis for therapeutic information [238], ChEMBL collected these indicators from various sources (DailyMed [36], Anatomical Therapeutic Chemical [37] and ClinicalTrials.gov [38]) for its database.

The complete process of FragPred is illustrated with the flow chart (Figure 3.50). ChEMBL (Table 3.21) provides data in several methods, for our purpose we used PostgreSQL (Section 2.5.2) method. The database is built up with the downloaded ChEMBL data, such that compounds and their target interaction can be accessible from the database. Subsequently a prefiltering (Figure 3.50 b) step exercised to improve the efficiency of the dataset by accepting only high-affinity targets [239, 224, 240]. The filtering process involves discarding all the

[36] https://dailymed.nlm.nih.gov/dailymed/

[37] https://www.whocc.no/atc_ddd_methodology/history/

[38] https://clinicaltrials.gov/

data with affinity values of IC_{50}, K_i, K_d, EC_{50} higher than 20µM, null values, and data which does not have affinity values. Further discarding molecules involving hexafluorophosphate ligands completes the filtering process.

Item	ChEMBL 22.1	ChEMBL 23
Compound records	2,036,512	2,101,843
Distinct compounds	1,686,695	1,735,442
Activities	14,371,197	14,675,320
Assays	1,246,683	1,302,147
Targets	11,224	11,538

Table 3.21 ChEMBL statistics in different used versions and detail information about the data available at ChEMBL 22.1[39] and 23[40] pages.

3.6.3 Fragmentation

The filtered library is now ready for fragmentation (Figure 3.50 c), in fragmentation, a fragment tool principally breaks small molecules into chemically meaningful fragments. RECAP (Section 2.1.1) based rules guide which bonds of small molecules are to be broken in both the tools (molBlocks [241], Fragmenterr [242]) used in this work.

3.6.3.1 molBlocks

The molBlocks [241] has a limitation which takes into account SMILES which are up to 200 characters. In molBlocks, four atoms being the minimum number of atoms in a fragment. Two types of fragmentation methods available in molBlocks, extensive and nonextensive and molBlocks recommends using extensive fragmentation [243]. In the project this flag is enabled, which ensures following steps during fragmentation of each molecule:

[39] ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_22_1/chembl_22_1_release_notes.txt

[40] ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_23/chembl_23_release_notes.txt

- identify all cleavable bonds in the molecule
- build a graph representation of the cleavable bonds, where there is an edge between cleavable bonds if they can be cleaved simultaneously
- identify all the maximal cliques in the graph; these cliques can be overlapping
- fragment the original molecule by breaking all the bonds in each maximal clique, one clique at a time

Further, if bonds are cuttable at the same time, then the cleavable bonds are represented as nodes in an undirected graph, with an edge between two nodes. Subsequently, the Bron-Kerbosch algorithm [244] is used to identify all maximal cliques. Finally, all the possible fragments are generated by cutting the bonds within each maximal group, one clique at a time.

Without the *-e* flag, the bonds are applied sequentially until no more fragments can be produced. In general, molBlocks recommends using the *-e* flag.

3.6.3.2 Fragmenter

Fragmenter [245] is a tool part of ChemicalToolBoX [85]. Fragmenter does not have the limitation of SMILES to be 200 characters. It also takes molecules which have a minimum of four atoms. It parses the molecules as SMILES and first looks for rotatable bonds, ring patterns, and the RECAP patterns by marking their starting and ending points. Subsequently, it makes fragments while recursively searching for its children.

Combining c and d steps in the flowchart (Figure 3.50) targets of molecules and fragments resulting fragmentation process, and their parent molecules kept in the database for enrichment analysis.

3.6.4 Enrichment analysis

Enrichment analysis (Figure 3.50 e) takes into account the dataset (Table 3.22) to perform Fisher's exact test (2.4.1.1) provided by SciPy stats python API [246]. Computation of *p-value* of each fragment per target is according to the figure 2.5 where:

- N is the total number of fragments produced by the ChEMBL compounds
- K is number of times specific fragments appearing in N
- n is the number of fragments per target
- k is the number of times specific fragment associated with a target.

Number of	molBlocks	Fragmenter
Distinct compounds (a)	1600407	1600407
Compounds after filtering (b)	430648	433009
Fragments (c)	2719016	2086288
Distinct fragments (c)	220141	215586
Distinct targets	10980	10980
Targets after filtering (d)	4751	4817
Enriched fragments after Benjamin Hochberg (f)	64540	63992
Enriched fragments after Bonferroni (f)	33336	31405
Enriched targets (g)	4474	3643

Table 3.22 Number of compounds, fragments at various steps in the workflow. Alphabets in first column parenthesis refer to the workflow (Figure 3.50)

To control the number of false positives, Bonferroni (Section 2.4.1.2) and Benjamini-Hochberg (Section 2.4.1.2) procedures were applied using python StatsModels API [247]. It has options of selecting Bonferroni correction or Benjamini-Hochberg test model. The function of StatsModels also has options to set the level of probability of committing *Type I error* (α) to determine if *p-value* is high or low. Here the traditional value of α is chosen is set to 0.05 or the 5% probability.

As shown in the table 3.22, a relaxed model Benjamini-Hochberg (Section 2.4.1.2) yields more number enriched fragments compared to stricter Bonferroni correction (Section 2.4.1.2) almost 50% fragments were filtered.

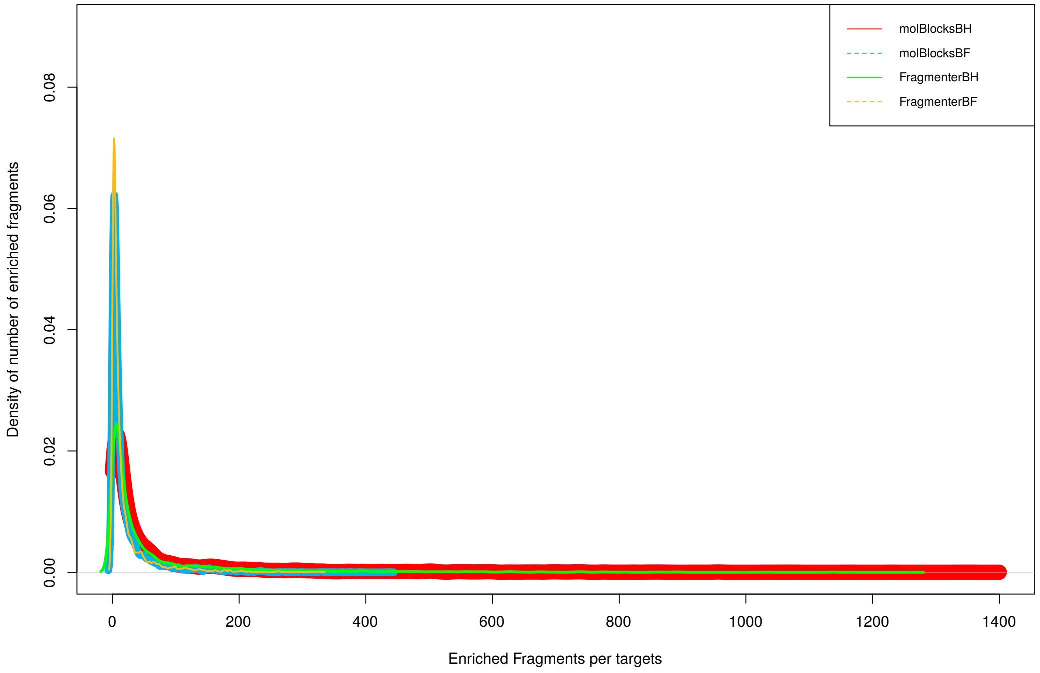


Fig. 3.51 Enrichment analysis of fragments per target with molBlocks, Fragmenter tools and applying Benjamini-Hochberg (BH), Bonferroni (BF) model

The Kernel density estimation of enrichment analysis indicates the number of enriched fragments per target in the figure 3.51. The graph suggests that in all the four methods, i.e., molBlocks with Benjamini-Hochberg and Bonferroni, Fragmenter with Benjamini-Hochberg and Bonferroni have fewer targets with several enriched fragments and many targets with few enriched sub-molecules. The list of top 25 proteins of the peak spike in the graph is in table 3.23. *hERG* (human *ether-a-go-go*-related gene) with 1261 fragments has the highest number of fragments is aα-subunit ion channel of a potassium ion channel. It is an off-target protein, and inherited mutations in *hERG* are the cause of long QT syndrome, a cardiac repolarization disorder called *arrhythmia*. In drug discovery inhibition of *hERG* is avoided [248]. *Vascular endothelial growth factor receptor 2* [249] (VEGFR-2) is a prominent kinase insert domain receptor (KDR) has 1169 fragments associated with it, and it belongs to type III receptor tyrosine kinase. KDR is a protein-encoding gene and is associated with diseases Capillary Infantile and Hemangioma. VEGFR-2 is a potential cell mitogen which regulates blood and lymphatic vessel development and homeostasis [250, 251]. Inhibition of VEFGR-2 by SU5416 significantly reduces parasites in the blood [252]. A high number of fragments associated with the target mean that they can be potential targets to several molecules. Another receptor Dopamine receptor D_2 (D2R) encoded by DRD2 gene and a GPCR (G protein-coupled receptor). GPCRs form a large protein family of receptors and are targets for more than 50% of current drugs in the market [253, 254]. Antipsychotic drugs are antagonists for D2R [255]. 1028 fragments associated with the D2R GPCR. Cytochrome $P_4$50 3A4 (CYP3A4 [256]) or CYP is an important heme-containing enzyme in the body, primarily found in the liver and the intestine. It oxidizes small foreign organic molecules (xenobiotics), such as toxins or drugs, so that they can be removed from the body. CYP3A4 gene encodes protein CYP34A, and the gene is part of Cytochrome $P_4$50 cluster [257]. CYPs are a major enzyme family metabolizing almost 60% of prescribed drugs [258] and are a major source of variability in drug response. CYP3A4 have almost thousand fragments association. Structures of the fragments listed in table 3.23 are in figure 3.52.

Taking the converse view of the enriched analysis data obtained from the figure 3.51 gives the number of targets associated with the fragments (Table 3.24). Their structures can be viewed in figure 3.53.

Target name	# of fragments	Min p-value	Min p-value fragment
HERG	1261	2.23E-133	NCc1nc2c(OCC(=O)N2)cc1
Vascular endothelial growth factor receptor 2	1169	2.39E-249	c1cc(NC=O)ccc1
Thrombin	1137	4.29E-282	NCc1ccc(C(=N)N)cc1
Coagulation factor X	1119	1.01E-302	C(=O)c1sc(Cl)cc1
Beta-secretase 1	1039	4E-261	NC(C(C)O)Cc1ccccc1
Dopamine D2 receptor	1028	1.06E-201	CCCCN
Cannabinoid CB2 receptor	998	1.55E-185	NC12CC3CC(CC(C3)C2)C1
PI3-kinase p110-alpha subunit	897	5.09E-294	c1nc(N)nc(C)n1
Cytochrome P450 3A4	894	4.51E-171	OCc1scnc1
Serotonin transporter	892	4.48E-118	N1CCNCC1
Cannabinoid CB1 receptor	885	6.41E-228	c1c(Cl)cc(Cl)cc1
Epidermal growth factor receptor erbB1	859	2.23e-315	Nc1cc(Cl)c(F)cc1
Serotonin 1a (5-HT1a) receptor	844	2.23E-114	COc1ccccc1
Carbonic anhydrase II	788	1.78E-214	[O-][Cl](=O)(=O)=O
MAP kinase p38 alpha	787	7.93E-185	CSc1[nH]ccn1
Kappa opioid receptor	775	9.49E-185	CC1CCC1
Melanin-concentrating hormone receptor 1	772	6.79E-164	N1CCCC1
Mu opioid receptor	772	2.11E-178	CC1CCC1
Adenosine A2a receptor	750	1.9E-134	OCC=O
Acetylcholinesterase	742	1.79E-130	Nc1c2c(cccc2)nc2CCCCc12
Dopamine D3 receptor	740	9.73E-270	CCCCN
Adenosine A1 receptor	738	1.72E-159	[nH]1c2c(cncn2)nc1
Serotonin 2a (5-HT2a) receptor	738	1.70E-119	N1CCC(c2c3c(cccc3)oc2)CC1
Peroxisome proliferator-activated receptor gamma	734	1.48E-293	C(=O)c1cc2c(c([nH]c2cc1)C)C
Carbonic anhydrase I	732	4.78E-168	[O-][Cl](=O)(=O)=O

Table 3.23 List of top 25 targets with highest number of fragments associated with it, corresponding enriched fragment and its *p-value*. The peak spike of the number of fragments can be observed in figure 3.51 obtained by fragmenter tool and Benjamini-Hochberg model

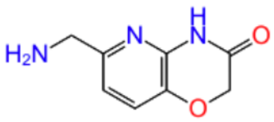
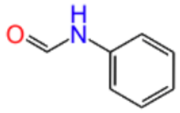
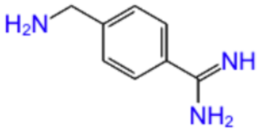
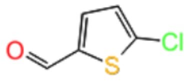
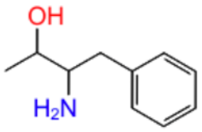
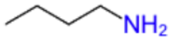
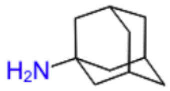
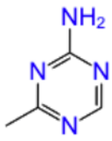
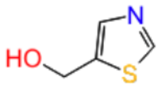
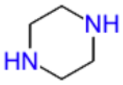
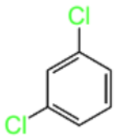
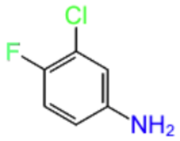
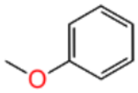
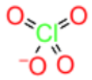
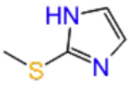
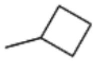
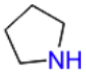
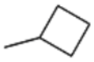
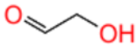
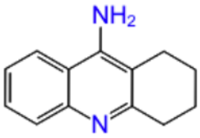
 (a) NCc1nc2c(OCC(=O)N2)cc1 (b) HERG (c) 1261	 (a) c1cc(NC=O)ccc1 (b) Vascular endothelial growth factor receptor 2 (c) 1169	 (a) NCc1ccc(C(=N)N)cc1 (b) Thrombin (c) 1137	 (a) C(=O)c1sc(Cl)cc1 (b) Coagulation factor X (c) 1119
 (a) NC(C(C)O)Cc1ccccc1 (b) Beta-secretase 1 (c) 1039	 (a) CCCCN (b) Dopamine D2 receptor (c) 1028	 (a) NC12CC3CC(CC(C3)C2)C1 (b) Cannabinoid CB2 receptor (c) 998	 (a) c1nc(N)nc(C)n1 (b) PI3-kinase p110-alpha subunit (c) 897
 (a) OCc1scnc1 (b) Cytochrome P450 3A4 (c) 894	 (a) N1CCNCC1 (b) Serotonin transporter (c) 892	 (a) c1c(Cl)cc(Cl)cc1 (b) Cannabinoid CB1 receptor (c) 885	 (a) Nc1cc(Cl)c(F)cc1 (b) Epidermal growth factor receptor erbB1 (c) 859
 (a) COc1ccccc1 (b) Serotonin 1a (5-HT1a) receptor (c) 844	 (a) [O-][Cl](=O)(=O)=O (b) Carbonic anhydrase II (c) 788	 (a) CSc1[nH]ccn1 (b) MAP kinase p38 alpha (c) 787	 (a) CC1CCC1 (b) Kappa opioid receptor (c) 775
 (a) N1CCCC1 (b) Melanin-concentrating hormone receptor 1 (c) 772	 (a) CC1CCC1 (b) Mu opioid receptor (c) 772	 (a) OCC=O (b) Adenosine A2a receptor (c) 750	 (a) Nc1c2c(cccc2)nc2CCCCc12 (b) Acetylcholinesterase (c) 742

Fig. 3.52 List of top 20 targets with highest number of fragments associated with it. (a) SMILES of the enriched fragment. (b) Target name. (c) Number of fragments associated with the target. The peak spike of the number of fragments can be observed in figure 3.51 obtained by fragmenter tool and Benjamini-Hochberg model

Fragment	# of targets	Min p-value	Min p-value target name
c1ccccc1	787	1E-298	P2X purinoceptor 3
Nc1ccccc1	473	1.1E-53	Purinergic receptor P2Y1
N1CCOCC1	450	7.42E-299	PI3-kinase p110-delta subunit
C(=O)c1ccccc1	360	5.5E-89	Cyclin-dependent kinase 5/CDK5 activator 1
N1CCCCC1	355	4.39E-191	Calcitonin gene-related peptide type 1 receptor
N1CCCC1	304	6.79E-164	Melanin-concentrating hormone receptor 1
c1ccncc1	293	1.59E-176	MAP kinase p38
N1CCN(C)CC1	269	9.78E-47	Tyrosine-protein kinase SRC
C(=O)CN	227	4.25E-267	C-C chemokine receptor type 2
OCCC	226	2.24E-58	Hepatocyte growth factor receptor
CCNCC	209	1.05E-93	Delta opioid receptor
NS(=O)(=O)c1ccccc1	200	2.81E-142	Matrix metalloproteinase-2
c1ncccn1	196	1.76E-86	Isocitrate dehydrogenase [NADP] cytoplasmic
NCc1ccccc1	195	4.74E-116	Transitional endoplasmic reticulum ATPase
c1ncccc1	194	1.31E-251	Cyclin T1
OC(F)(F)F	191	4.03E-130	Glucagon receptor
c1cnccc1	189	1.82E-238	Cytochrome P450 11B2
c1ncncc1	187	5.22E-176	Isocitrate dehydrogenase [NADP] cytoplasmic
CNC1C(OC)C2(n3c4c(c5c(c6c7c(n(C(O2)C1)c46)cccc7)C(=O)NC5)c1ccccc31)C	178	0.000907	Mitogen-activated protein kinase kinase kinase 7-interacting protein 1
C(=O)C(N)C	177	5.48E-100	Glucagon-like peptide 1 receptor
c1ccc(N)cc1	174	5.67E-36	Cyclin-dependent kinase 4
C(=O)C1NCCC1	170	1.54E-157	Thrombin
CC(=O)O	168	1.34E-123	Retinoid X receptor alpha
Cc1c(C=O)c(C)c(C)[nH]1	167	0.000114	Serine/threonine-protein kinase SRPK2
N1C(C=O)CCC1	165	9.49E-252	Thrombin

Table 3.24 List of top 25 fragments with highest number of targets associated with it, corresponding enriched target name and its *p-value*, obtained by fragmenter tool and Benjamini-Hochberg model

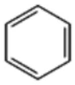
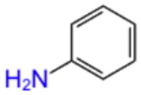
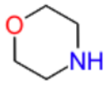
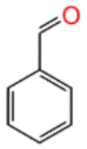
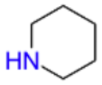
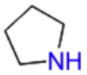
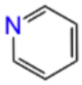
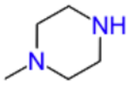
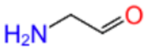
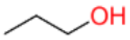
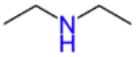
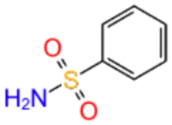
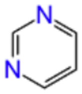
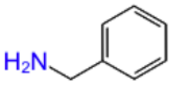
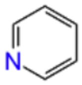
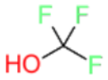
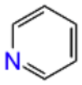
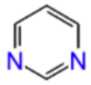
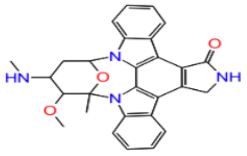
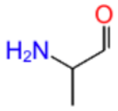
 (a) c1ccccc1 (b) P2X purinoceptor 3 (c) 787	 (a) Nc1ccccc1 (b) Purinergic receptor P2Y1 (c) 473	 (a) N1CCOCC1 (b) PI3-kinase p110-delta subunit (c) 450	 (a) C(=O)c1ccccc1 (b) Cyclin-dependent kinase 5/CDK5 activator 1 (c) 360
 (a) N1CCCCC1 (b) Calcitonin gene-related peptide type 1 receptor (c) 355	 (a) N1CCCC1 (b) Melanin-concentrating hormone receptor 1 (c) 304	 (a) c1ccncc1 (b) MAP kinase p38 (c) 293	 (a) N1CCN(C)CC1 (b) Tyrosine-protein kinase SRC (c) 269
 (a) C(=O)CN (b) C-C chemokine receptor type 2 (c) 227	 (a) OCCC (b) Hepatocyte growth factor receptor (c) 226	 (a) CCNCC (b) Delta opioid receptor (c) 209	 (a) NS(=O)(=O)c1ccccc1 (b) Matrix metalloproteinase-2 (c) 200
 (a) c1ncccn1 (b) Isocitrate dehydrogenase [NADP] cytoplasmic (c) 196	 (a) NCc1ccccc1 (b) Transitional endoplasmic reticulum ATPase (c) 195	 (a) c1ncccc1 (b) Cyclin T1 (c) 194	 (a) OC(F)(F)F (b) Glucagon receptor (c) 191
 (a) c1cnccc1 (b) Cytochrome P450 11B2 (c) 189	 (a) c1ncncc1 (b) Isocitrate dehydrogenase [NADP] cytoplasmic (c) 187	 (a) CNC1C(OC)C2(n3c4c(c5c(c6c7c(n(C(O2)C1)c46)cccc7)C(=O)NC5)c1ccccc31)C (b) Mitogen-activated protein kinase kinase kinase 7-interacting protein 1 (c) 178	 (a) C(=O)C(N)C (b) Glucagon-like peptide 1 receptor (c) 177

Fig. 3.53 List of top 20 fragments with highest number of targets associated with it. (a) SMILES of the enriched target. (b) Target name. (c) Number of targets associated with the fragment. The structures obtained by fragmenter tool and Benjamini-Hochberg model

3.6.5 Target prediction

The final phase in the Fragment-based target prediction is to predict a target for a given compound. For this purpose I have taken compounds diabetes workflow from StreptomeDB 2.0 diabetes-related compounds (Table 3.7) and NANPDB diabetes-related compounds (Table 3.15). These compounds serve as an input (Figure 3.50 h). Fragments of these compounds are extracted with Fragmenter tool and are mapped (Figure 3.50 j) to the targets with the help of earlier established model (Figure 3.50 g). The expected results are to be the target proteins which exhibit antidiabetic activity. The inputs are unknown to the model as the compounds from StreptomeDB 2.0 and NANPDB are not part of the initially acquired data. Prediction of targets can be observed in tables (Table 3.7 for StreptomeDB 2.0 and Tables E.1, E.2 for NANPDB). Multiple targets are predicted as there were multiple fragments for each compound and the targets *pancreatic alpha-amylase*, *salivary alpha-amylase* [259], *Sodium/glucose cotransporter 2* [260], *HMG-CoA reductase* [261], are antidiabetic activity targets. In the table 3.25 for the compound *Trestatin C* (with StDB ID 3769) has nine fragments, and all have the Tanimoto 1 indicating the very similar match of the fragment found in the model. There are also duplicate rows in the nine fragments, indicating that *Trestatin C* has same fragment multiple number of times. The compound *Trestatin C* is also similar to DrugBank compound *Acarbose*, one of the most commonly found compound in literature (Tables 3.2, 3.7) and its comparitive structure is in figure 3.13. For the nine fragments, eight fragments were the representatives in the model as two of nine fragments has the same representative. Comparing initial mapping (Figure 3.50 e) before the enrichment model these eight fragments were mapped to 543 targets. With the enriched model obtained after FDR (Benjamini-Hochberg test) these fragments were mapping to 103 targets. From which after checking with *p-value* highly ranked fragments mapping targets were chosen resulting the four targets (*pancreatic alpha-amylase*, *salivary alpha-amylase*, and *sodium/glucose cotransporter 2* in Homo sapiens, *Adhesin protein fimH* in *E.coli K-12*). However, one has to distinguish that model creation is with the principle of the number of fragments enriched per target. Checking the database conversely revealed that these four targets were attached to 1849 fragments before enrichment analysis and after enrichment analysis, these were 212 fragments ranked by their *p-value*.

StDB ID	Compound name	Tanimoto between 2 fragments	p-value	Organism	Predicted target where fragment is enriched	Diabetic target (Y/N)
1098	pravastatin	1	7.78E-57	*Rattus norvegicus*	HMG-CoA reductase	Y [261]
	pravastatin	0.84	5.35E-05	*Mus musculus*	Voltage-gated L-type calcium channel alpha-1C subunit	—
3313	Adiposin 2	1	2.02E-20	*Homo sapiens*	Pancreatic alpha-amylase	Y [262]
	Adiposin 2	0.52	7.37E-20	*Human immunodeficiency virus 1*	Human immunodeficiency virus type 1 protease	Y [263]
	Adiposin 2	1	2.09E-17	*Homo sapiens*	Pancreatic alpha-amylase	Y [262]
	Adiposin 2	1	1.32E-07	*Homo sapiens*	Salivary alpha-amylase	Y [264]
3769	Trestatin C	1	2.39E-123	*Homo sapiens*	Sodium/glucose cotransporter 2	Y [260]
	Trestatin C	1	2.78E-64	*Escherichia coli K-12*	Adhesin protein fimH	—
	Trestatin C	1	2.02E-20	*Homo sapiens*	Pancreatic alpha-amylase	Y [262]
	Trestatin C	1	2.02E-20	*Homo sapiens*	Pancreatic alpha-amylase	Y [262]
	Trestatin C	1	2.09E-17	*Homo sapiens*	Pancreatic alpha-amylase	Y [262]
	Trestatin C	1	2.51E-17	*Homo sapiens*	Pancreatic alpha-amylase	Y [262]
	Trestatin C	1	6.92E-15	*Homo sapiens*	Pancreatic alpha-amylase	Y [262]
	Trestatin C	1	1.10E-08	*Homo sapiens*	Pancreatic alpha-amylase	Y [262]
	Trestatin C	1	1.45E-04	*Homo sapiens*	Salivary alpha-amylase	Y [264]

Table 3.25 Target prediction of StreptomeDB 2.0 DM compounds (Table 3.7). In the coloumn Tanimoto between 2 fragments, it is the similarity between the fragment of the unknown input molecule and the fragmend found in the database.

Chapter 4

Discussion and outlook for the future

Drug discovery includes drug designing and development, is a diverse and expensive effort, where least number of drugs that pass the clinical trials makes it to market. The chemical space being almost infinite multiplies the challenge to the pharmaceutical industry. The primary hurdle being is it expedient to use such a chemical universe. Combined with compound space research happening around the world in the pharmaceutical industry and producing voluminous data. PubMed alone produces half million records every year. Subsequently, how to organize such an extensive data. The success rate is one in ten thousand compounds. And only one out of twenty approved drugs brings revenue to the academia and industry. The age of a drug discovery being ten years. Although challenges being several, compared to the number of drugs discovered in the last couple of decades as part of solutions are very few. Software-based drug discovery and development methods evolving from it have vital importance in the development of bioactive compounds. The computational techniques provide speed and accuracy to the experimental findings and mechanisms of action. The HGP added another dimension of a solution to the problem. The promise of BigData after Human Genome project adds to the advantage. The pharmaceutical industry adopts the modular approach in resolving challenges of drug discovery. Phases of drug discovery (Section 1.3) signify the modular approach.

The presented thesis endeavored to built cheminformatic tools which could form as modules in the drug discovery. A localized text mining tool (Section 3.1), ligand and structure-based virtual screening (Section 2.1) tools, databases (Sections 3.2, 3.3) incorporating essential natural products and convenient way of accessing them with several related features. The present work further explains how the sub-modular tools can be kept into practice by explaining diabetes drug research with a hypothesis as a case study. And to contemplate all the tools the doctoral research presents all the applications in a pipeline. The thesis focus is to present ligand based tools for drug discovery from natural products.

4.1 PubMedPortable: Incorporating into a pipeline

PubMedPortable is a modular based program and has modules for downloading the disease-specific dataset in multiple methods, creating the required schema and importing the downloaded dataset into the PostgreSQL database. Building up a full-text index with Xapian and Lucene [265] as an alternative option. Because of its portable and multiplatform capability, we have tested the complete workflow in Fedora Operating system, and it's complete instructions are available in the project page [129].

The use cases of PubMedPortable in publication and this thesis with pancreatic cancer [114] and diabetes mellitus (Section 3.1.2) respectively indicate some of the features of the PubMedPortable. The project is wholly open-source and free to be efficiently adapted to the requirement of the user. In the research field, it's always useful to get the latest information about the topic we are studying. In locally built PubMedPortable as of now, the recommendation is to make a new instance of the local database instead of updating the existing database; this is a drawback as of now. It is a future endeavor to provide options to build a robust PubMedPortable which can be quickly updated with every day growing PubMed literature. The modularity of software makes tiresome to execute several scripts, if a GUI is provided to the complete workflow or integrating into Galaxy framework it will be an easy tool to use. In the use case of DM (Section 3.1.2), PubMedPortable uses named entity recognition of chemical entity on DrugBank and for protein entity on UniProt. As having the synonyms is an essential part of the PubMedPortable it will be preferable to connect standard NER sources of bio-concepts such as from PubTator.

From this PubMedPortable following conclusions can be drawn:

- PubMedPortable is a handy tool which can be used to build hypotheses from the PubMed literature.
- The tool being highly modular is an advantage at every step which can be customized quickly for the hypothesis. However, same modularity can be disadvantageous to realize several steps.

- There is a great need of applying a framework for PubMedPortable so that the user is comfortable.

4.2 Natural compound databases

In the present work, I have presented two natural compound databases namely StreptomeDB 2.0 and NANPDB. Both the natural databases are significant concerning actinobacteria and North African region respectively. The former has plenty of literature already available in the research field due to interest in its genus actinobacteria. Whereas later one does not have much research available to it; it's relatively new to the online world although research is happening (oldest research reference available is from the year 1908, Herbarium, Cairo University (Egypt)) their knowledge was limited to printed literature in libraries of African universities. Curation is one of the time-consuming step in natural compounds collection. The curation becomes more difficult for a database like NANPDB compared to StreptomeDB as there is a limited volume of literature. We made a double check of every detail and also provided an extensive option of giving feedback[1] where one can also submit new data.

Tanimoto similarity estimated in the two NP databases explored uses two different tools, StreptomeDB 2.0 uses Openbabel and NANPDB uses RDKit. Both the databases have same time complexity. However, ease of installing and setting up on a server and hosting natural products databases RDKit performs better by calculating the Tanimoto distance on the fly. One of the reasons being RDKit works in Python environment whereas Openbabel is a C++ based tool. RDKit has several fingerprints techniques enabled, it is simpler to select different kinds of Fingerprints simultaneously, and RDKit has a very active community and working on parallel execution solution since the arrival of PostgreSQL 9.6 which has Parallel Querying feature. Scaffolds are of particular interest in drug discovery [63, 266], as on average each commercialized drug contains a novel scaffold, thus stressing the crucial relevance of examining both natural sources and literature for new drugs.

[1] http://african-compounds.org/nanpdb/downloads/

4.2.1 StreptomeDB update

StreptomeDB 2.0 update in 2015 is significant update after its first inception in 2012. Considering the amount of research (Figure 3.7) happening around *Streptomycetes* the update is deemed to be overdue update. StreptomeDB till now had two updates, and the latest update is work in this thesis which provided features like scaffolds browsing, phylogenetic tree, and vital information NMR, MS-spectra data for VS improving the chemical and evolutionary context of the library.

We could identify hundreds of naturally occurring level 1 scaffolds not found in purchasable compounds [6]. Scaffolds aid in compound design in medicinal chemistry. Dimova et al. [267] uses series-based scaffolds substituting multiple sites into account. Initially, it builds analog series of same core structures having different substituents at one or more sites. These analog series are used to build an analog series-based scaffold. The analog series based scaffold capture target information of analog series. For such scaffold applications, scaffolds browser is beneficial. In future works of scaffold providing such a series based scaffold would be high interest for the research community. The present update of the database offers MS and NMR data improving the quality of filtering compounds based on characterization measures. These measures are useful in primary structure determination by delivering strategies in differentiating isomers [268] of secondary metabolites. Now users of StreptomeDB 2.0 gets the opportunity to select compounds based on the MS and NMR spectra. However, as of now, information is provided under each compound a range querying browser would be more beneficial.

In the diabetes mellitus case study of StreptomeDB 2.0, it is observed that there were several fingerprints (Table 3.6) which estimated similar compounds (Tanimoto >0.85). However, compounds listed with antidiabetic activity were not the results Agn-PC-0BKWXC, Nfat-133, Platensimycin with best Tanimoto values and exhibiting 0.50, 0.20, and 0.25 as their best Tanimoto coefficient respectively irrespective of any fingerprint.

4.2.2 NANPDB: A North African resource

NANPDB is the first African database among the series[2] of databases from the region of Africa. Initiated by our collaborator Dr. Ntie-Kang[3]. Further, there are other areas (West, East) of compounds will be implemented as part of this series. For all comprehensive databases, one website[4] is chosen.

After the recent update of NANPDB number compounds found in PubChem [150] increased to 52%, indicating the uniqueness and importance of the database. About 20% compounds have shown at least one biological activity, and 2% of compounds are with the mode of action. Almost 18% of compounds required research to know their mode actions as their bioactivity is already known. An evaluation of the physicochemical properties (Tables 3.11, 3.12) operated in Lipinski's RO5 (Section 2.1.6) suggests that 53% of NANPDB compounds showed no Lipinski violations, while 71% of the NPs showed less than one violation. Indicating high probability of being drug compounds. Almost 25% of molecules are heavy molecules which are not suitable for DMPK/ADMET predictive models. The predictions are, therefore, guiding information could be of interest for future drug discovery. As with the intention of updating the compound information with continuous data curation, there has been already an update of compounds in NANPDB. However in future updates, we intend to expand are more features by including more computed molecular descriptors, experimental data leading to the characterization of the NPs, e.g., NMR (Section 2.1.2) and MS data (Section 2.1.3), melting and boiling points, and possible biosynthesis pathways with the included metabolites. PubChem vendor information can be retrieved. In future updates, we would include data for compound sample availability and vendor information for samples scattered in academic laboratories. NANPDB is planned to be upgraded annually.

For the exclusive database resources new and missing compounds are always interesting. For this purpose, we have provided new data submission pipeline[5]. Here we present a

[2] http://african-compounds.org/about/
[3] http://pc.pharmazie.uni-halle.de/medchem/mitarbeiter/fidele_ntie-kang/
[4] http://african-compounds.org
[5] http://african-compounds.org/nanpdb/downloads/

spreadsheet consisting of all the fields required to upload new or correct the existing data. The submissions provided by pipeline improves the quality of the database.

In the DM case study (Section 3.3.4) of NANPDB compounds it is found that maximum of three fingerprints predicted DrugBank DM similar compounds (Table 3.15). However, the structures (Figures 3.31, C.3, C.4, C.5, C.6) were not promising. Although there are antidiabetic compounds (Table 3.16) in NANPDB they were only predicted by one fingerprint with Tanimoto coefficient >0.85 (Table 3.17). This indicates that available fingerprints were not able to estimate antidiabetic compounds or NANPDB has contrasting compounds compared to DrugBank compounds which have to be investigated and validated.

4.2.3 Django QuerySets

The QuerySets (Section 2.5.3) are very useful in database resources. NANPDB takes advantage of these QuerySets. The compound list search (Figure 3.33) lists the compounds for the selected alphabet. The code snippet (9) is QuerySet access the database only when it sees the Q object (line 2 of the listing) and then filters it according to the expression in the Q object. QuerySet reduces database accessing time and accesses it only when it required to retrieve the specific expression data.

```
1 moleculeQuerySetList = Molecule.objects.filter(
2         Q(mol_name__istartswith=currentInitial) |
3         Q(mol_name__istartswith=currentInitial.lower())
4         ).select_related('pubchem').distinct('mol_name')
```

Listing 9 QuerySet example

From natural product databases following conclusions can be drawn:

- Scaffold-based molecular representations is an important tool in medicinal chemistry [63] and allow classifying and characterize the StreptomeDB 2.0 library.
- With phylogenetic tree conceiving the distribution of a particular scaffold, chemotype, bioactivity or synthetic route becomes easier in the evolutionary context.

- StreptomeDB 2.0 facilitates building relationship between different data, such as scaffolds, activities or phylogenetic distribution.

- NANPDB compounds do not have compounds which are much similar to DrugBank diabetic compounds, but it has several antidiabetic compounds.

- Natural compound databases are starting point with several additional tools for the discovery of new compounds, virtual screening campaigns, biochemical engineering and cheminformatics.

- In both presented NP databases antidiabetic compounds were not the best similar (Tanimoto coefficient >0.85) compounds affirming the lesser role of Tanimoto measures. However, antidiabetic active compounds do exist and they have to be validated with experimental measures.

4.3 SeMPI prospects

SeMPI a structure prediction tool predominantly depends on the algorithm used to measure structural similarity and the minimum threshold which is required to consider a molecule from the test dataset as a putative match to a subjected gene cluster. SeMPI could rank 67.5% actual gene cluster products within first ten ranks among 2839 candidates compared to antiSMASH could rank only 12.5% gene cluster products [206]. Demonstrating the efficiency of the algorithm to detect the correct paths in a large number of molecules. There were two limitations for the poor ranking one being path algorithm reaches its limitations in gene clusters which have multiple tailoring reactions. Another difficulty being the wrong identification of starting unit for the path algorithm due to an unexpected tailoring reaction of the unit. In subsequent updates, SeMPI will add more databases such as NANPDB and other NP databases to manifest a maximum number of putative annotated molecules that come into inquiry for subjected gene clusters. SeMPI depends on NP databases, and so maintenance of such prediction tools is complex on the servers as yet NP databases do not have any standards in providing as a service. ChEMBL can lead for such a kind of solution as it already provides ChEMBL RDF model [269]. Which will make easier for tools like SeMPI to integrate themselves into natural products chemical space.

In the case study of DM compounds of StreptomeDB 2.0 with SeMPI several DNA sequences were not available. Indicating the need of annotation of DNA sequences of natural databases. For the possible DNA sequences, SeMPI could not predict structures due to too many NRPS (Nonribosomal peptide synthetases) signatures, No signature of modular KS (ketoacyl synthase). The problem with NRPS genes is that they are built template independent building blocks called nonproteinogenic amino acids which are about 500 naturally occurring identified amino acids [270]. And every module in the NRPS has specificity for each amino acid with various functions such as antibiotics, pigments, etc. A possible solution could be categorizing building blocks into few classes and working based on the specific type of blocks. And for the problem of no signature of modular KS or the start codon by having unique open reading frames (ORFs or start codons) to each domain [271].

The SeMPI and StreptomeDB services use Apache web server, although Apache and another popularly available web server Nginx are well in use. One has to decide the web server based on the application requirements. NANPDB, on the other hand, is using Nginx. Seeing the experience with both the web servers for Python based application Nginx performance is far better. Nginx combined with gunicorn a Python-based dynamic content WSGI HTTP server works fast consuming fewer resources compared to Apache. Apache requires higher memory usage with comparatively lower performance. Nginx, even more, better when it is serving static files from a Python-based framework. Apache configuration settings differ from with operating system and is another disadvantage. In configuring the number of processors and amount of serving connections to the server, Apache has several options whereas Nginx has mere few possibilities. Apache web server would help for more customized environment according to the need compared to Nginx.

From SeMPI following conclusions can be drawn:

- SeMPI through its benchmark test proved better than antiSMASH 3.0 and could improve the prediction ability for the products of modular *PKS type I* gene clusters [206].
- SeMPI combines multiple tools to that are available for polyketide prediction and joins the prediction with NP databases assuring compound availability in nature.
- SeMPI is in its early stages in the process of solving problems such as too many NRPS signatures.
- Tools such as SeMPI emphasizes the importance of standardizing the natural product databases.
- Choosing a good web server environment is one of the key factors for the server to work efficiently.

4.4 DVS discussion

Dynamic virtual screening is a method of structure-based drug discovery. It takes into account the structural information generated by Glide [61] tool with SP Docking. And is mainly dependent on the score it makes (Figure 3.48 a). Glide is optimized to work on generic cases over a wide range of systems. As per the Glide documentation [272], Glide has three modes of docking HTVS, SP, and XP. All the three employ a hierarchical filter to search for prospective locations of the ligand in the binding site of a receptor. The exhaustive search of ligand torsions lists a collection of ligand conformations. From these confirmations, the method of detecting poses differs in each mode. Each of them takes 2, 10, and 120 seconds respectively for each compound. These options balances between speed and accuracy. Implementation of SP docking is against the target BRD4 limits the docking score applicability to bromodomain inhibitors. However, the specific nature of algorithm enables greater customization of parameters while building the model particular to the inhibitors in focus. Benchmarking step that we have used is a particular dataset of Bromodomains and their inhibitors, and so the application of the model of physicochemical properties with top scored compounds has to be tested further with other datasets.

From DVS following conclusions can be drawn:

- DVS requires high intensive computing environment and good docking tools for generating efficient docking scores.
- It is a good strategy to filter vast chemical spaces if the large chemical spaces are available such as PurchasableBoX [6].
- Once the docking scores are available then the algorithm can be dynamically reconfigured according to the requirement and repeat the arbitrary compound collection step until an optimum level of physicochemical properties model arrives.
- It can be a lengthy process to attain an optimum model.

4.5 FragPred antecedents and future

The fragment-based virtual screening (Section 2.1.1) technique applied to FragPred (Section 3.6) initially started with the idea of clustering to reduce the number of different substructures. Implementation of clustering step was before enrichment analysis (Figure 3.50 e). In clustering, fragments were clustered and matched to their biological target that their parent compounds interact. Among several clustering methods, we have considered Taylor Butina clustering and k-means clustering. Taylor Butina clustering Butina [273] which is a Javis-Patrick's algorithm takes fingerprints of the fragments and identifies latent cluster centroids and gathers substructures by measure their Tanimoto distance near to the potential centroid. In this process, it is very likely some of the fragments are lost or instead left alone as singleton sets. The advantage of being less number of heterogeneous clusters compared to J-P Taylor Butina. Creation of false singletons sets and there were 19% singleton sets compared to the method adopted without clustering had 4% significance difference. The initial step of comparing all molecules against all is also a CPU intensive step becomes a drawback. Also, Butina clustering wrapper provided by RDKit (Section 2.3.1) was not able to cluster all compounds. We have also tried k-means clustering, for this, we have employed Canvas [146] module of Schrödinger software. It uses the algorithm of Lloyd [274], in the k-means clustering algorithm several numbers of iterations are performed on the dataset, and each iteration starts by selecting k number of data points at arbitrary centroids of the cluster. Subsequently, assigning of data points to the closest centroid at every iteration. Also dynamically the algorithm updates the position of centroid by calculating the average position of all its existing data points. The process continues until a predefined number of iterations are reached, or until a convergence condition. In the end, cluster points belonging to the least cost run are the final clustered set. The selection of points being random it could be possible that selected points lie in between clusters and those points could be wrongly mapped. This reason pushes back the k-means algorithm to be efficient only locally. With k-means clustering, singleton sets were 13%. FragPred is the running project, although results (Section 3.6.5) test case shows good prediction with the help of StreptomeDB 2.0 diabetes compound. FragPred has to undergo some more tests as the results of NANPDB compounds (Appendix tables E.1) were not as good as StreptomeDB 2.0. In principle, clustering idea is a useful technique however to achieve an optimized algorithm number of clustering techniques have to be tested. Locality-sensitive hashing (LSH) can be one such clustering technique it generates random hyperplanes. To every data point, generation of hash-code happens in every iteration, as per the position of the data point in the hyperplane. Clustering of similar data points will be in the same hyperplane, and the significant advantage of LSH is it eliminates false positives during the assigning process. Fingerprint-based Tanimoto calculation is restrictive and may not represent the whole chemical similarity of a structure as the occurrence of same elements is rather important than functional groups.

In the target prediction phase (Section 3.6.5) to understand the working of FragPred compounds of StreptomeDB 2.0 and NANPDB which were similar to the DrugBank diabetes compounds given as input. StreptomeDB 2.0 compounds showed right predictions, NANPDB compound did not show diabetes targets although compounds were already antidiabetic nature. Many predictions have to be studied with more number test cases.

From FragPred following conclusions can be drawn:

- FragPred is an initial step towards filtering chemical space with the aid of existing bioactive compounds and predicting a putative target.
- Enrichment analysis can filter the considerable amount of compounds. It improved further with minimizing the Type 1 error.
- Applying cluster techniques did not show better results, however with better algorithms; there is a high probability of improvement in results.
- Although initial results of diabetic compounds predicted diabetic targets they are yet to be validated to confirm the working of FragPred.

All the cheminformatic tools presented in this doctoral research are useful however there are several things such as understanding how the tool work, what are the inputs to the applications, sometimes visiting several websites to obtain results. If the tools presented in the thesis are joined in one framework using Galaxy framework (Section 2.3.3) with series of actions would be extremely beneficial for pharmacy research. The galaxy workflow can take input about the disease and download all the PubMed data related to disease entity to the local machine. In the next step, it generates series of basic statistics related to the disease and provides options to moderate statistics according to the user. If alternative less toxic natural compounds are possible, it can check available natural databases (Section 3.2, 3.3) and offer several available open natural compound databases. Produce ready to use links to the features provided by the web resources. If the Genbank [167] data of the organism is available, then a secondary metabolite prediction can be instantly made and check if similar compounds are available in natural compound databases. After predicting a synthetic compound, FragPred (Section 3.6) can take it as input and predict if any suitable targets are available exhibiting bioactivity. There is enormous scope for farther development and working on the projects provided in the thesis.

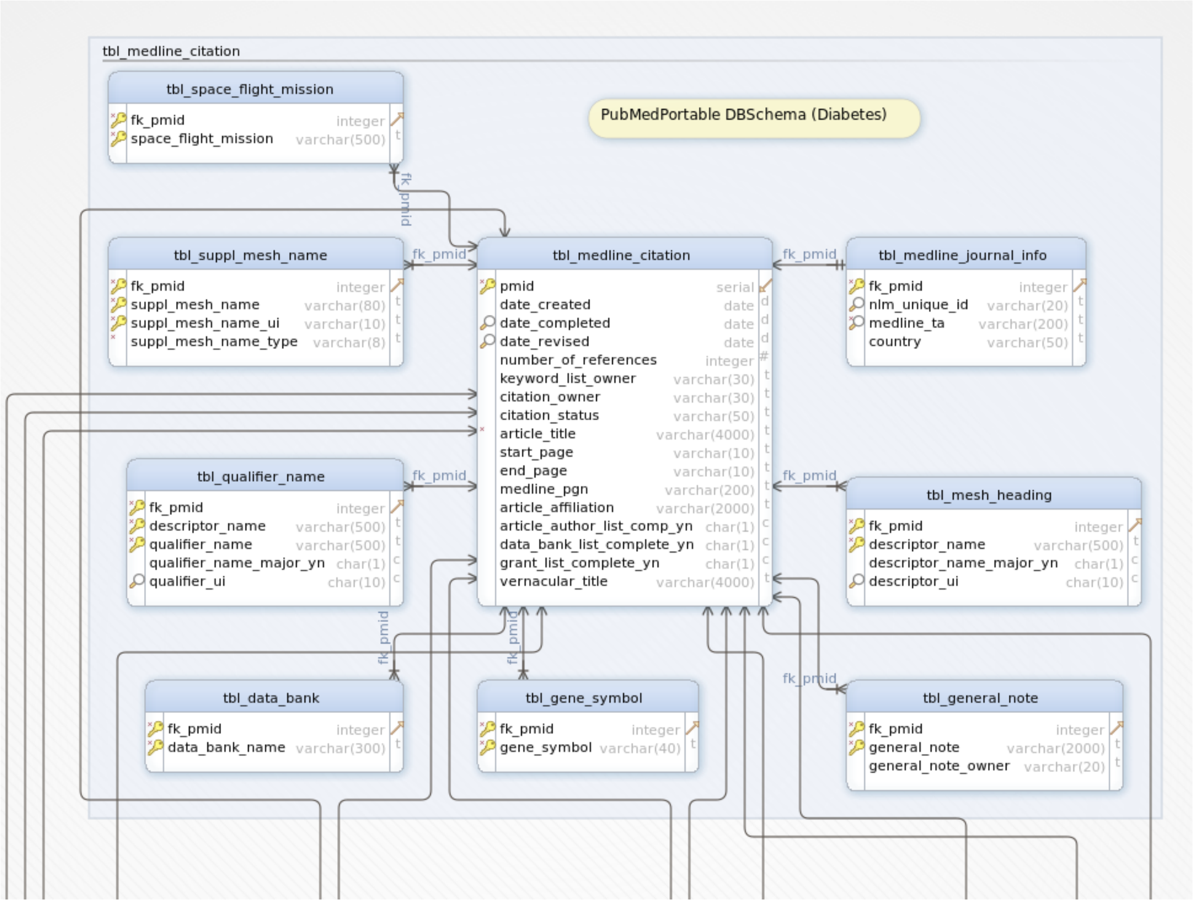
If computer applications developed are user-friendly to biologists, chemists, and pharmacists the drug research can progress much faster and reduce the age of new drug development and secure the health of humankind. The accomplishment of the complete work of this doctoral thesis is in the premises of university indicating the strength of the universities and their contribution to research in life sciences.

I conclude my thesis by thanking everybody named in acknowledgments. I respectfully appreciate the academia and all the funding agencies for the doctoral dissertation.

Appendix A

PubMedPortable

A.1 PubMedPortable detailed database schema



tbl_space_flight_mission	
fk_pmid	integer
space_flight_mission	varchar(500)

tbl_suppl_mesh_name	
fk_pmid	integer
suppl_mesh_name	varchar(80)
suppl_mesh_name_ui	varchar(10)
suppl_mesh_name_type	varchar(8)

tbl_medline_citation	
pmid	serial
date_created	date
date_completed	date
date_revised	date
number_of_references	integer
keyword_list_owner	varchar(30)
citation_owner	varchar(30)
citation_status	varchar(50)
article_title	varchar(4000)
start_page	varchar(10)
end_page	varchar(10)
medline_pgn	varchar(200)
article_affiliation	varchar(2000)
article_author_list_comp_yn	char(1)
data_bank_list_complete_yn	char(1)
grant_list_complete_yn	char(1)
vernacular_title	varchar(4000)

tbl_medline_journal_info	
fk_pmid	integer
nlm_unique_id	varchar(20)
medline_ta	varchar(200)
country	varchar(50)

tbl_qualifier_name	
fk_pmid	integer
descriptor_name	varchar(500)
qualifier_name	varchar(500)
qualifier_name_major_yn	char(1)
qualifier_ui	char(10)

tbl_mesh_heading	
fk_pmid	integer
descriptor_name	varchar(500)
descriptor_name_major_yn	char(1)
descriptor_ui	char(10)

tbl_data_bank	
fk_pmid	integer
data_bank_name	varchar(300)

tbl_gene_symbol	
fk_pmid	integer
gene_symbol	varchar(40)

tbl_general_note	
fk_pmid	integer
general_note	varchar(2000)
general_note_owner	varchar(20)

Fig. A.1 Part 1/3 of complete PubMedPortable database schema

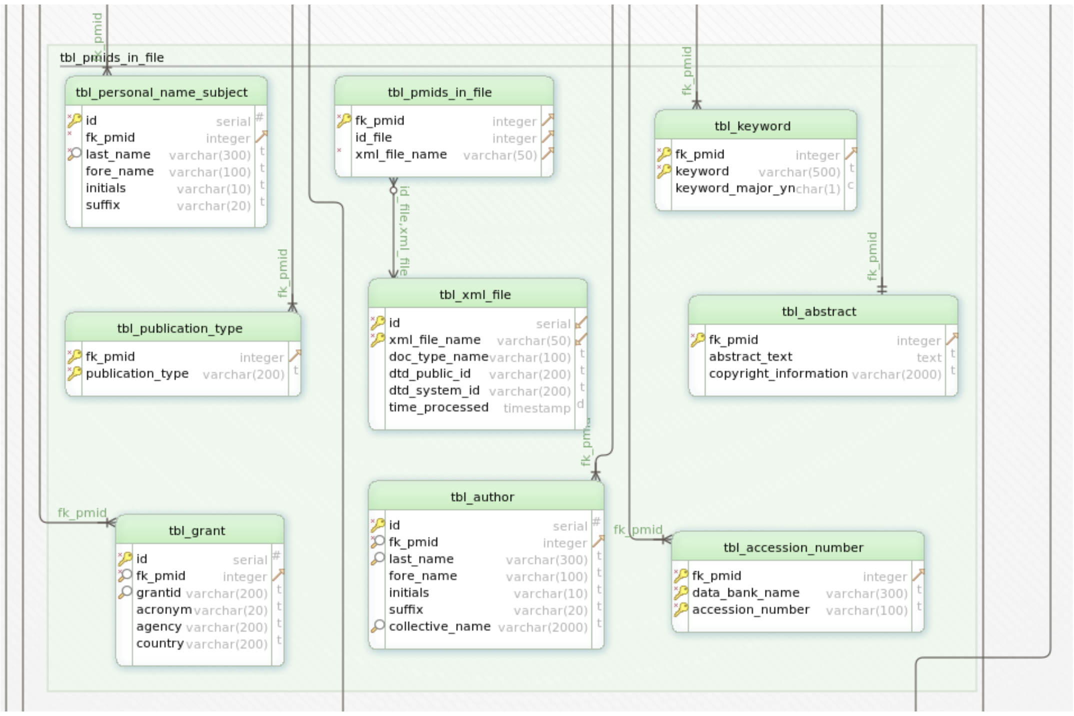


Fig. A.2 Part 2/3 of complete PubMedPortable database schema

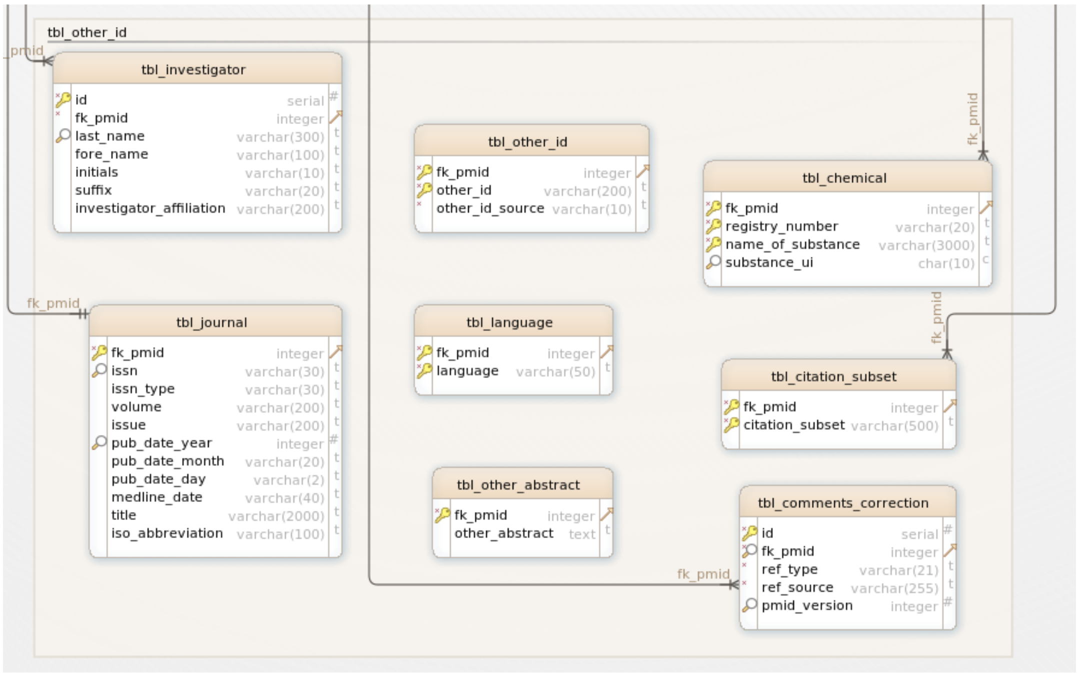


Fig. A.3 Part 3/3 of complete PubMedPortable database schema

A.2 Diabetes mellitus data used in PubMedPortable

DB ID	Drug name	Occurances	DB ID	Drug name	Occurances
DB00331	Metformin	11491	DB00071	Insulin Pork	427
DB01132	Pioglitazone	2723	DB00912	Repaglinide	427
DB00412	Rosiglitazone	2512	DB00790	Perindopril	358
DB01124	Tolbutamide	2496	DB01029	Irbesartan	345
DB01016	Glyburide	2443	DB09564	Insulin Degludec	345
DB00047	Insulin Glargine	2226	DB00722	Lisinopril	316
DB06655	Liraglutide	1625	DB00731	Nateglinide	315
DB01261	Sitagliptin	1604	DB06203	Alogliptin	314
DB00284	Acarbose	1361	DB09265	Lixisenatide	263
DB01276	Exenatide	1337	DB01200	Bromocriptine	261
DB00672	Chlorpropamide	1275	DB01278	Pramlintide	245
DB00046	Insulin Lispro	1102	DB00275	Olmesartan	240
DB01076	Atorvastatin	1085	DB01309	Insulin Glulisine	219
DB01306	Insulin Aspart	1040	DB01233	Metoclopramide	204
DB00641	Simvastatin	1031	DB00491	Miglitol	177
DB00678	Losartan	976	DB09045	Dulaglutide	161
DB01120	Gliclazide	893	DB00519	Trandolapril	137
DB00222	Glimepiride	866	DB00930	Colesevelam	131
DB01307	Insulin Detemir	825	DB09043	Albiglutide	129
DB00999	Hydrochlorothiazide	777	DB08885	Aflibercept	128
DB01197	Captopril	727	DB00839	Tolazamide	119
DB00584	Enalapril	722	DB00834	Mifepristone	97
DB04876	Vildagliptin	684	DB00542	Benazepril	85
DB01067	Glipizide	657	DB04861	Nebivolol	85
DB08907	Canagliflozin	571	DB00796	Candesartan cilexetil	80
DB09038	Empagliflozin	556	DB01184	Domperidone	79
DB00178	Ramipril	539	DB00492	Fosinopril	72
DB06292	Dapagliflozin	527	DB00881	Quinapril	68
DB00035	Desmopressin	513	DB00102	Becaplermin	42
DB00177	Valsartan	492	DB00876	Eprosartan	35
DB01270	Ranibizumab	472	DB09477	Enalaprilat	22
DB06335	Saxagliptin	457	DB11827	Ertugliflozin	18
DB08882	Linagliptin	450	DB09054	Idelalisib	1
DB00966	Telmisartan	434			

Table A.1 DM drugs from DrugBank [132] and their number of occurances in PubMed using PubMedPortable [114]

Gene/Protein	Occurances	Gene/Protein	Occurances	Gene/Protein	Occurances
PPAR	36467	ABCC8	24653	RET	21205
ACE	36198	HNF1A	24623	AVPR2	21139
DB	35818	OX	24563	CAPN10	21105
VEGF	34787	PON1	24517	SOD2	21072
AMPK	33960	CTLA4	24409	ICA512	20899
INS	33267	PTPN22	24149	CAV	20791
OB	32973	WFS1	24149	GIPR	20791
GLUT4	32817	ADIPOQ	24132	CCR5	20791
RAGE	31423	IRS2	24132	GLP1R	20644
CCL2	30216	PERK	24048	ENPP1	20644
CP	30153	NGN3	23979	PDK4	20606
PERI	29907	HMGB1	23944	ATGL	20569
GAD65	29666	RAD	23909	LMNA	20492
AVP	28573	SLC30A8	23909	VEGFA	20334
TCF7L2	28350	GAL	23673	CEL	20211
LPL	28331	INSR	23579	XBP1	20211
IAPP	28155	ZNT8	23560	NEUROD	20170
C3	28155	CDKAL1	23521	PC1	20128
FOXP3	28135	LEP	23502	APOA5	20043
STAT3	28109	HFE	23404	GCG	19912
PDX1	28061	HNF4A	23364	NAMPT	19822
GLUT2	27656	CART	23242	IFIH1	19777
KCNJ11	27442	KCNQ1	23138	V2R	19542
HLA-DRB1	27411	HLA-DQA1	22900	GCGR	19493
GLUT1	27363	UCP3	22878	FLT	19444
GCK	27299	FABP4	22576	IGRP	19344
PKB	26972	PAX4	22479	GSP	19344
IRS1	26444	PTC	22405	GPX1	19294
VDR	26095	AKT2	22329	IPEX	19030
CO2	26085	AIRE	22253	PPARA	18920
UCP1	26042	AKT1	22253	MCP1	18864
DRG	25865	VP	22068	CD38	18864
NDI	25809	SUR	21903	PCK1	18808
LEPR	25705	gamma2	21875	GNAS	18808
PRL	25682	PON	21846	IL2RA	18750
PPARG	25611	NEUROD1	21818	HT2A	18750
SUR1	25587	OGT	21731	DCP	18692
UCP2	25390	IGF2BP2	21702	GLIS3	18692
HLA-A	25340	AS160	21673	IPF1	18633
UCP	25276	FOXA2	21522	TCF2	18573
POMC	25276	HNF1B	21492	NKX2.2	18573
EPO	25185	CPM	21430	EIF2AK3	18573
HLA-DQB1	25105	GCKR	21398	SGK1	18512
NLRP3	24828	RAC	21303	SLC2A2	18512
AQP2	24800	ASC	21238	COX2	18450

Table A.2 Extensive list of genes taken from UniProt and their number of occurances in DM PubMed literature using PubMedPortable

Appendix B

StreptomeDB 2.0

B.1 StreptomeDB 2.0 detailed database schema

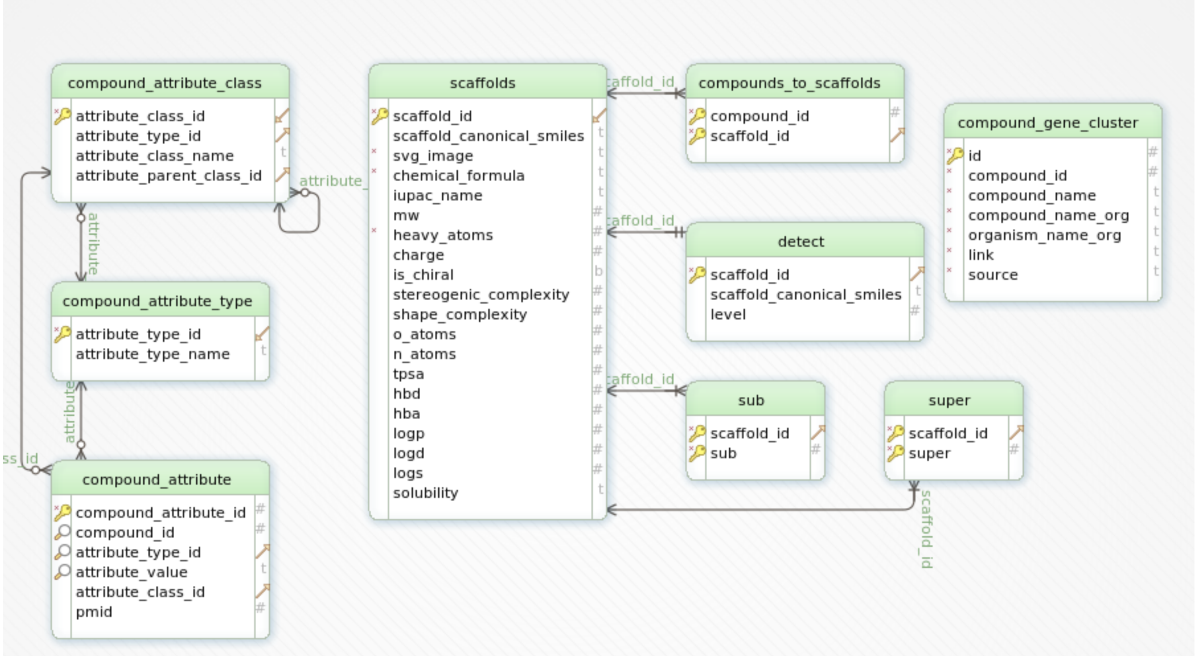


Fig. B.1 Part 1/3 of complete StreptomeDB2 database schema

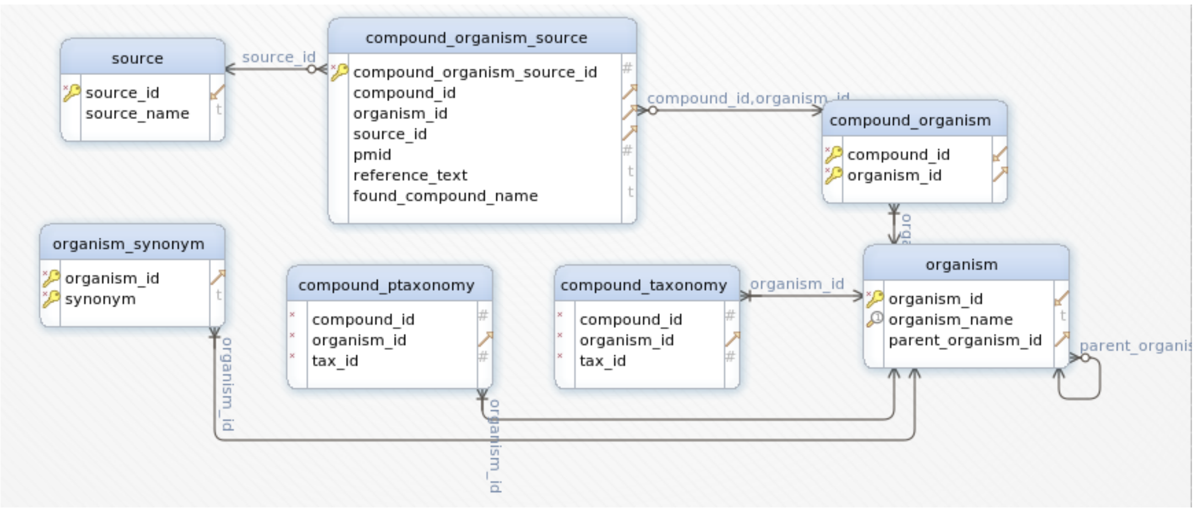


Fig. B.2 Part 2/3 of complete StreptomeDB2 database schema

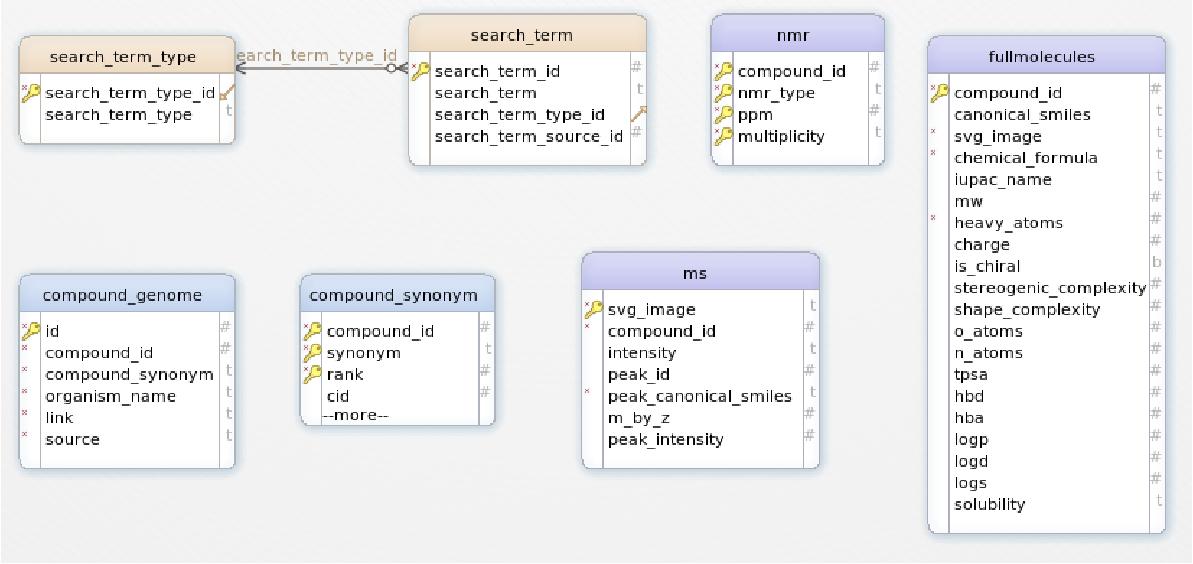


Fig. B.3 Part 3/3 of complete StreptomeDB2 database schema

B.2 Kernel density function of similarities

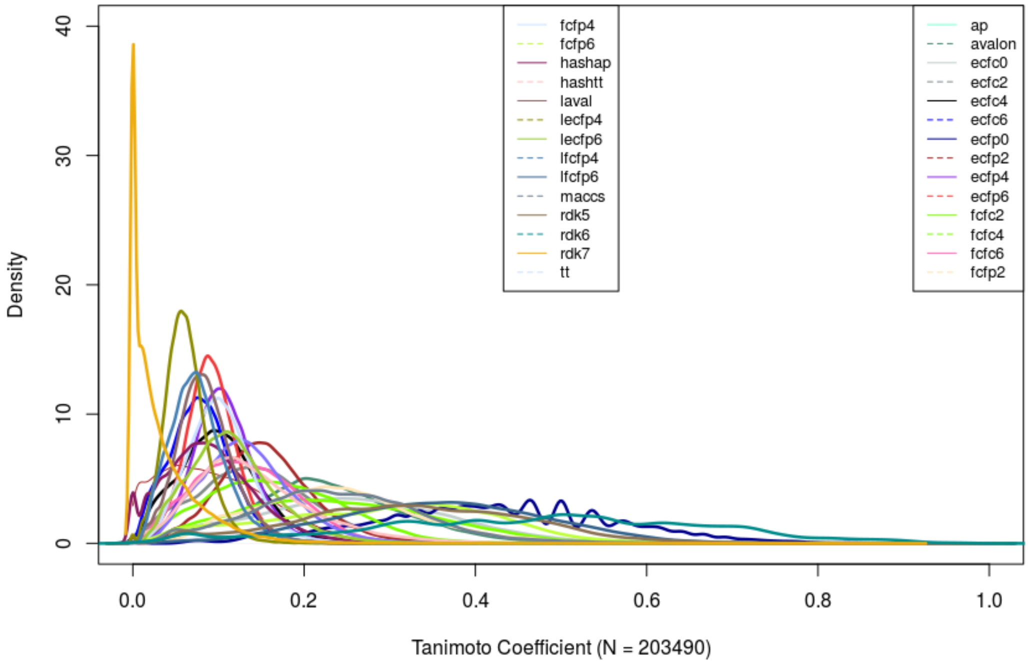


Fig. B.4 Graph showing Kernel density estimation of tanimoto coeficcient between DrugBank [132] DM compounds and StreptomeDB 2.0 [22] compounds. N indicates the number of comparisions with each fingerprint.

Appendix C

NANPDB

C.1 NANPDB detailed database schema

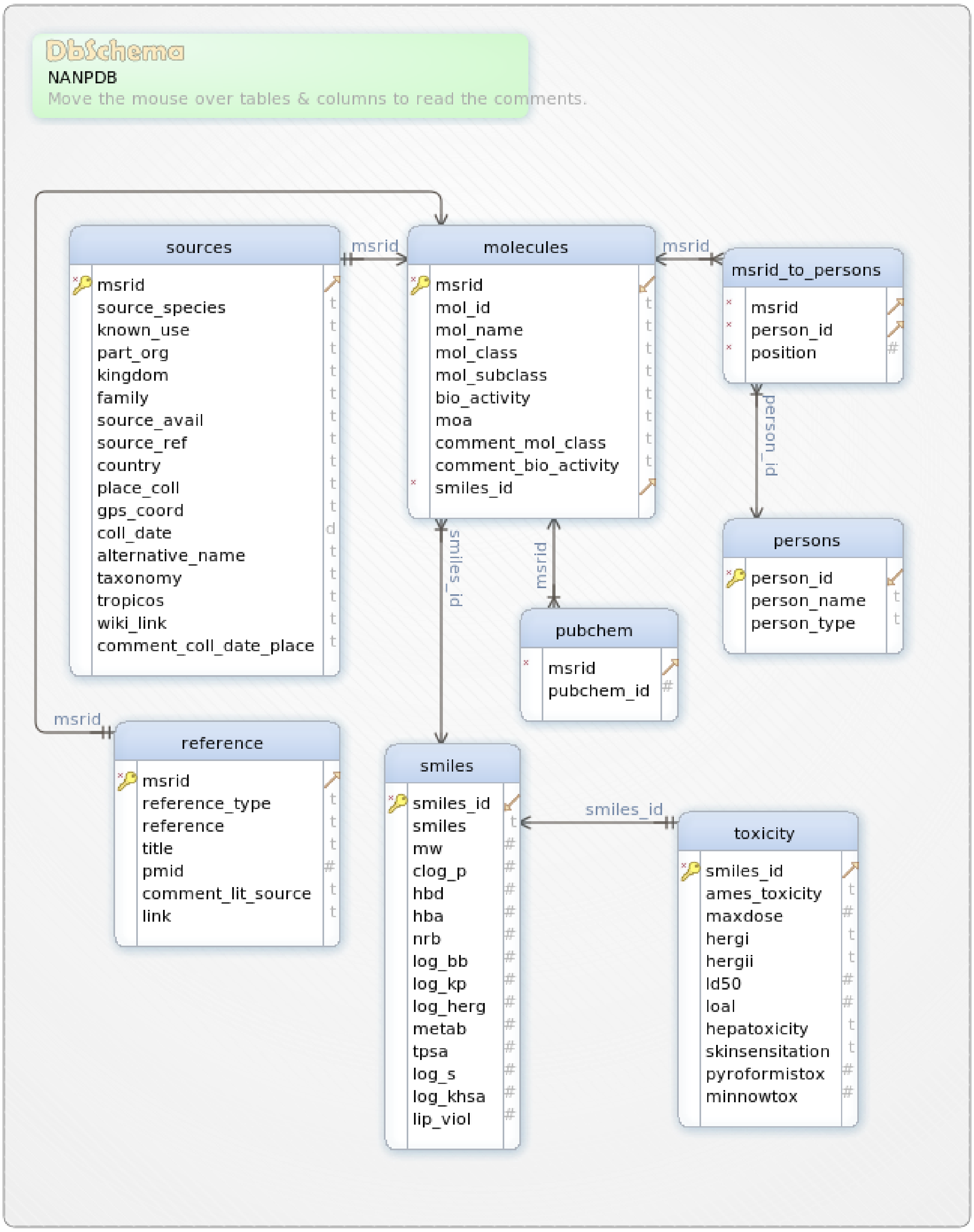


Fig. C.1 Complete overview of NANPDB database schema

C.2 Kernel density function of similarities

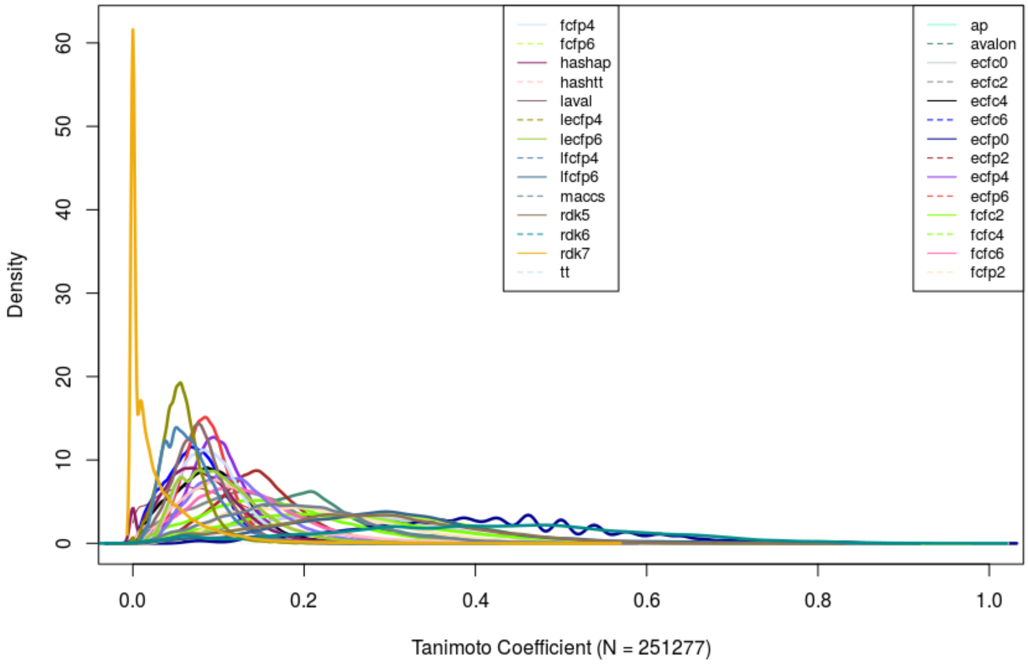


Fig. C.2 Graph showing Kernel density estimation of tanimoto coefficient between Drug-Bank [132] DM compounds and NANPDB [189] compounds. N indicates the number of comparisions with each fingerprint. All fingerprint listed in table 2.1.

C.3 Top 10 similar pair structures

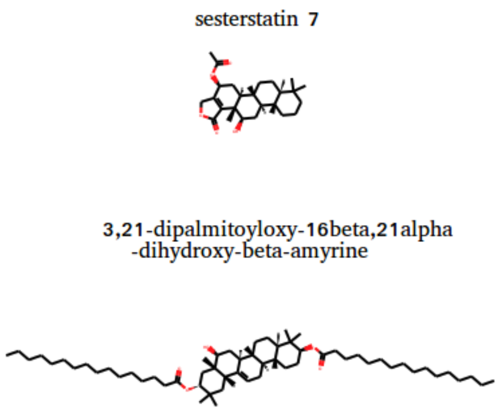


Fig. C.3 Structures of top 10 pair compounds which are listed in table 3.15, figure set: 2/5. DrugBank compounds (left), NANPDB compounds (right).

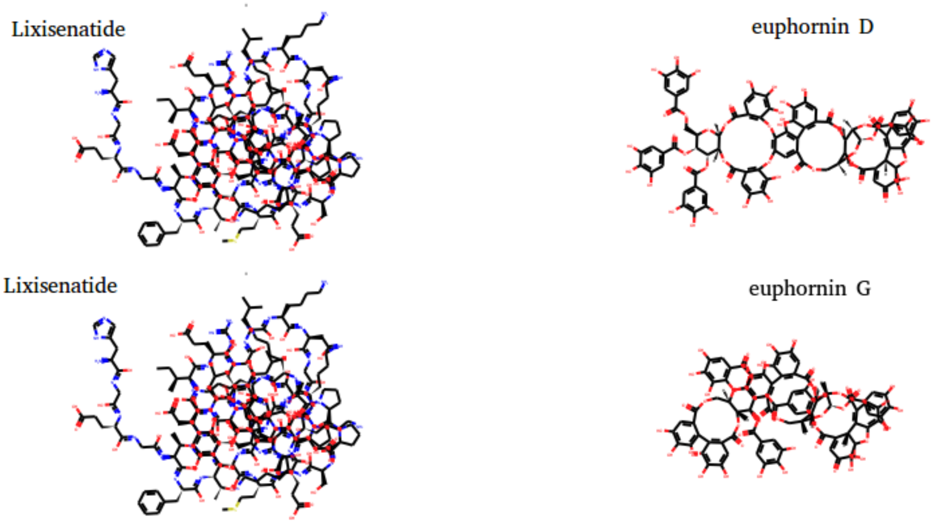


Fig. C.4 Structures of top 10 pair compounds which are listed in table 3.15, figure set: 3/5. DrugBank compounds (left), NANPDB compounds (right).

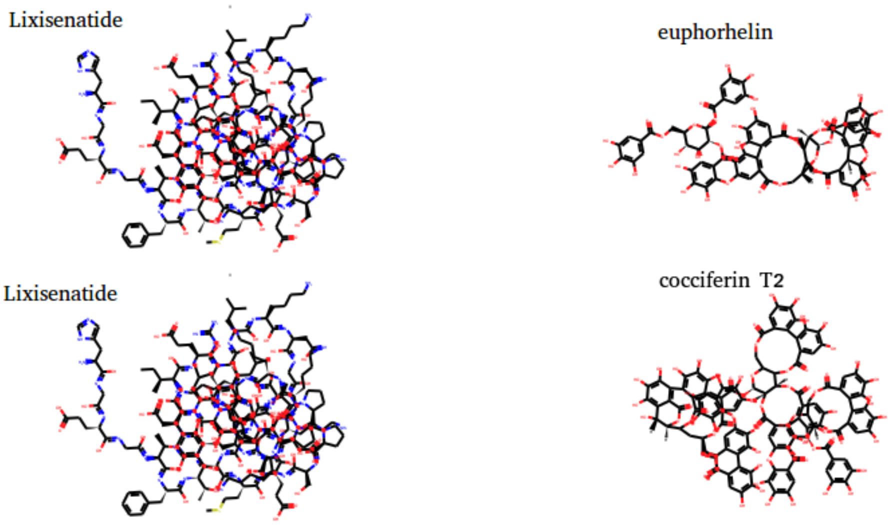


Fig. C.5 Structures of top 10 pair compounds which are listed in table 3.15, figure set: 4/5. DrugBank compounds (left), NANPDB compounds (right).

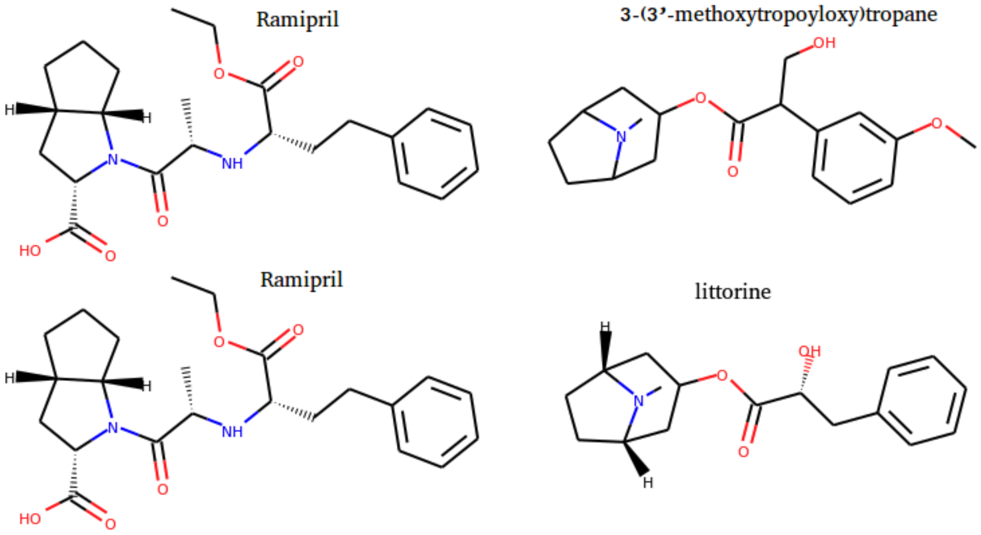


Fig. C.6 Structures of top 10 pair compounds which are listed in table 3.15, figure set: 5/5. DrugBank compounds (left), NANPDB compounds (right).

C.4 NANPDB antidiabetic indicated compounds

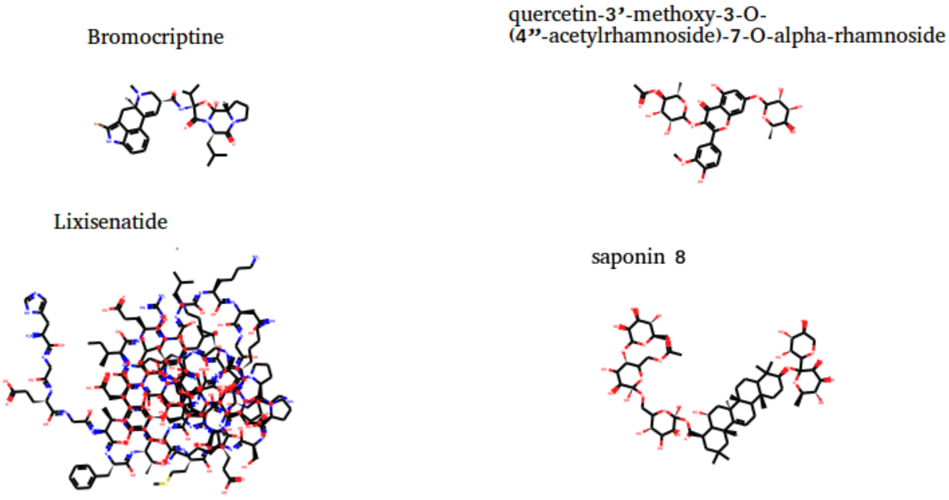

Bromocriptine
quercetin-3’-methoxy-3-O-
(4”-acetylrhamnoside)-7-O-alpha-rhamnoside
Lixisenatide
saponin 8

Fig. C.7 Structures of DrugBank DM compounds (left) compared with NANPDB antidiabetic indicated compounds (right) which are listed in table 3.17 Figure: 1/11

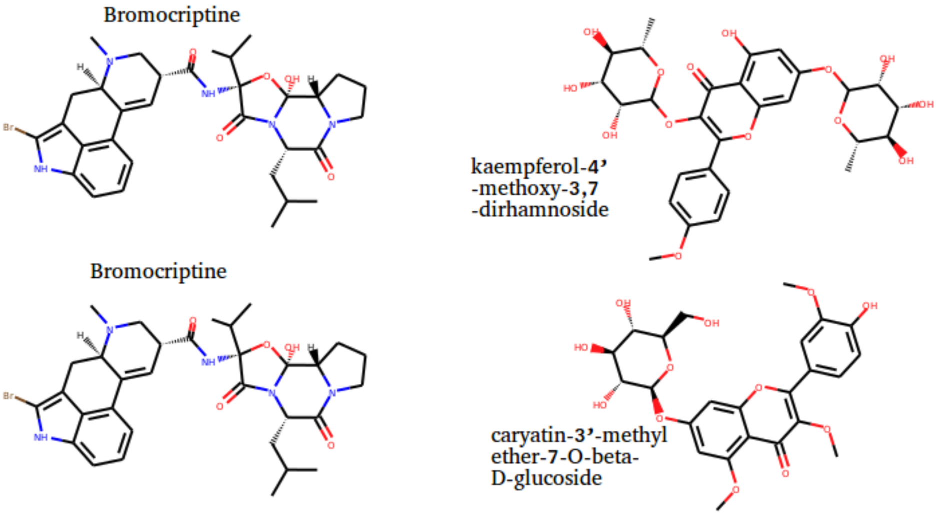
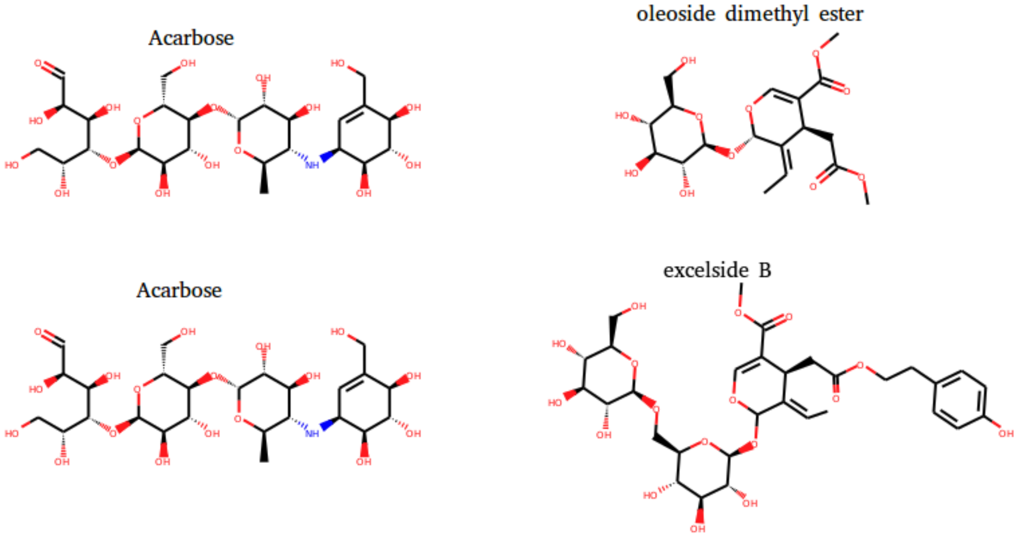


Fig. C.8 Structures of DrugBank DM compounds (left) compared with NANPDB antidiabetic indicated compounds (right) which are listed in table 3.17 Figure: 2/11



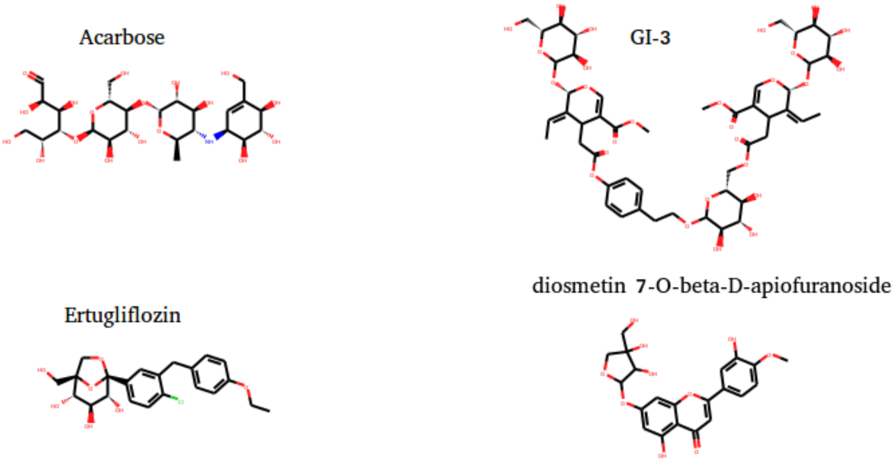
Acarbose

oleoside dimethyl ester

Acarbose

excelside B

Fig. C.9 Structures of DrugBank DM compounds (left) compared with NANPDB antidiabetic indicated compounds (right) which are listed in table 3.17 Figure: 3/11



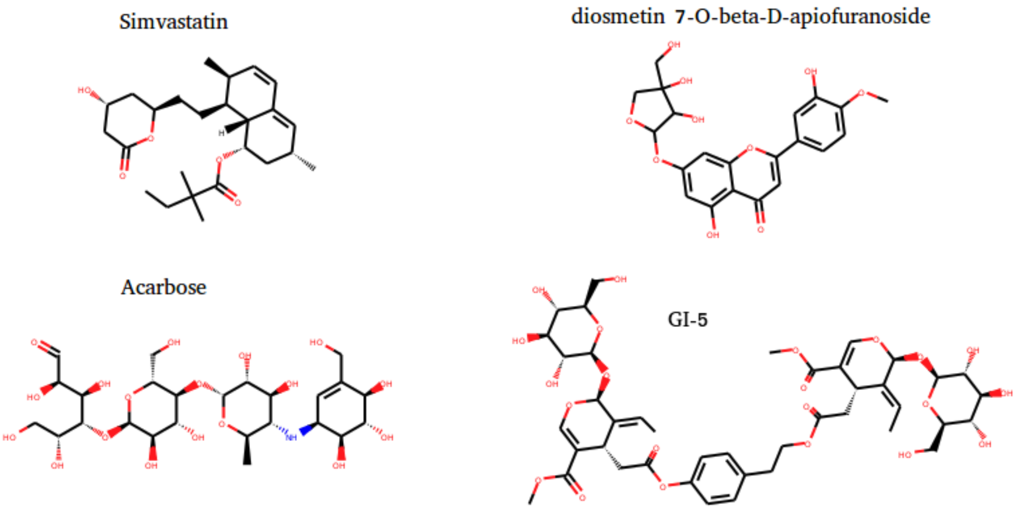
Acarbose

GI-3

Ertugliflozin

diosmetin 7-O-beta-D-apiofuranoside

Fig. C.10 Structures of DrugBank DM compounds (left) compared with NANPDB antidiabetic indicated compounds (right) which are listed in table 3.17 Figure: 4/11



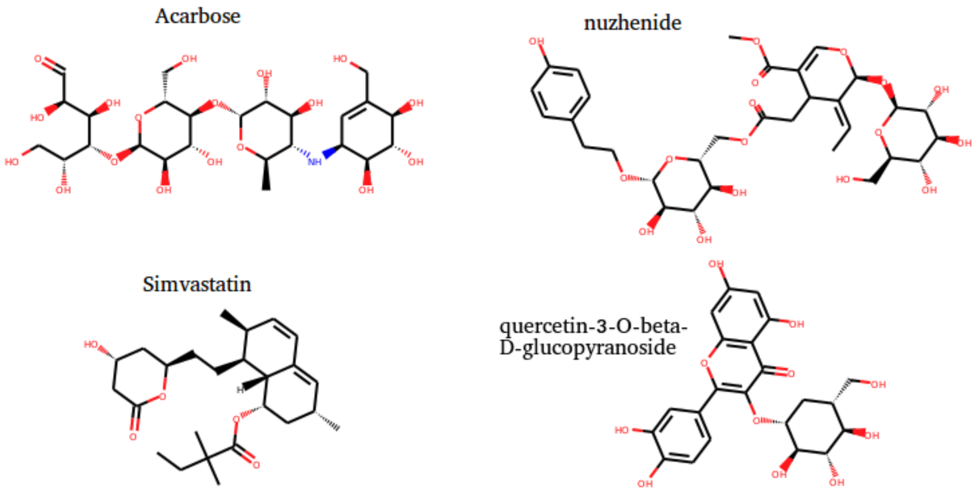
Simvastatin

diosmetin 7-O-beta-D-apiofuranoside

Acarbose

GI-5

Fig. C.11 Structures of DrugBank DM compounds (left) compared with NANPDB antidiabetic indicated compounds (right) which are listed in table 3.17 Figure: 5/11



Acarbose

nuzhenide

Simvastatin

quercetin-3-O-beta-D-glucopyranoside

Fig. C.12 Structures of DrugBank DM compounds (left) compared with NANPDB antidiabetic indicated compounds (right) which are listed in table 3.17 Figure: 6/11

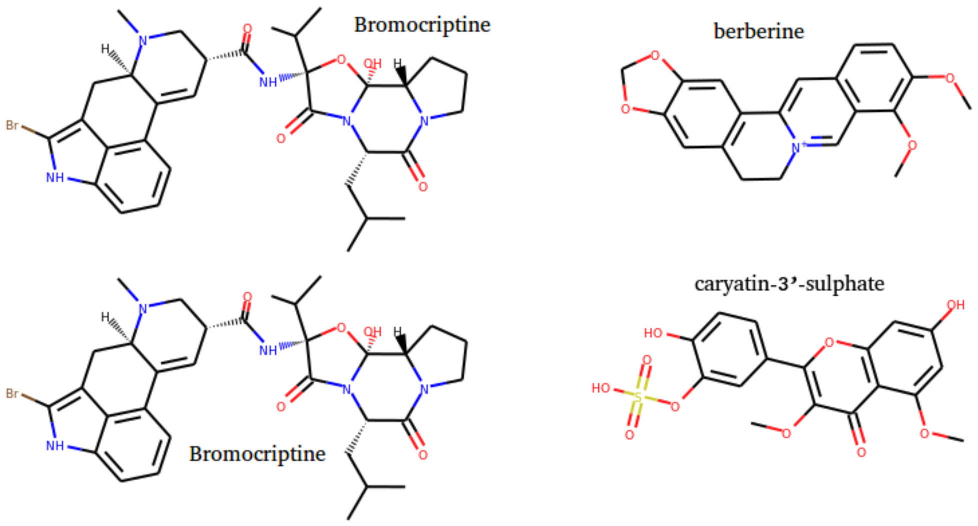


Fig. C.13 Structures of DrugBank DM compounds (left) compared with NANPDB antidiabetic indicated compounds (right) which are listed in table 3.17 Figure: 7/11

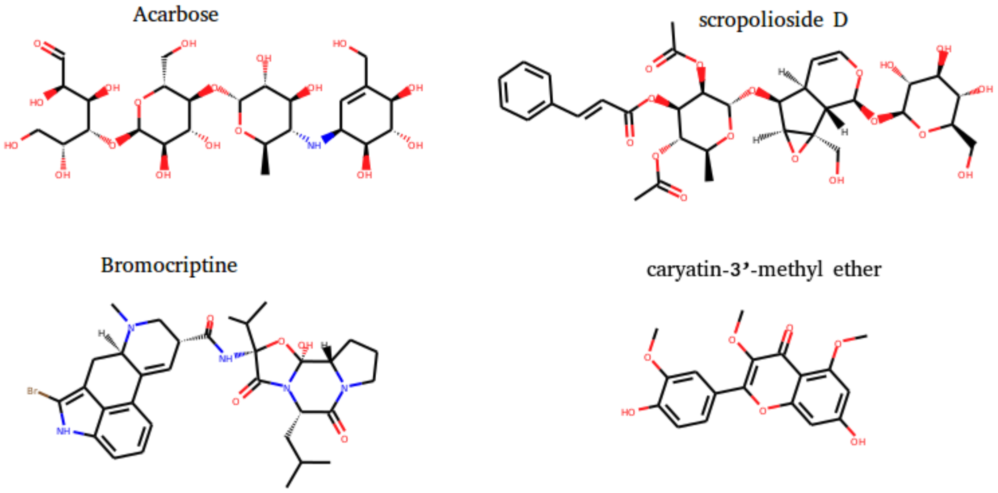


Fig. C.14 Structures of DrugBank DM compounds (left) compared with NANPDB antidiabetic indicated compounds (right) which are listed in table 3.17 Figure: 8/11

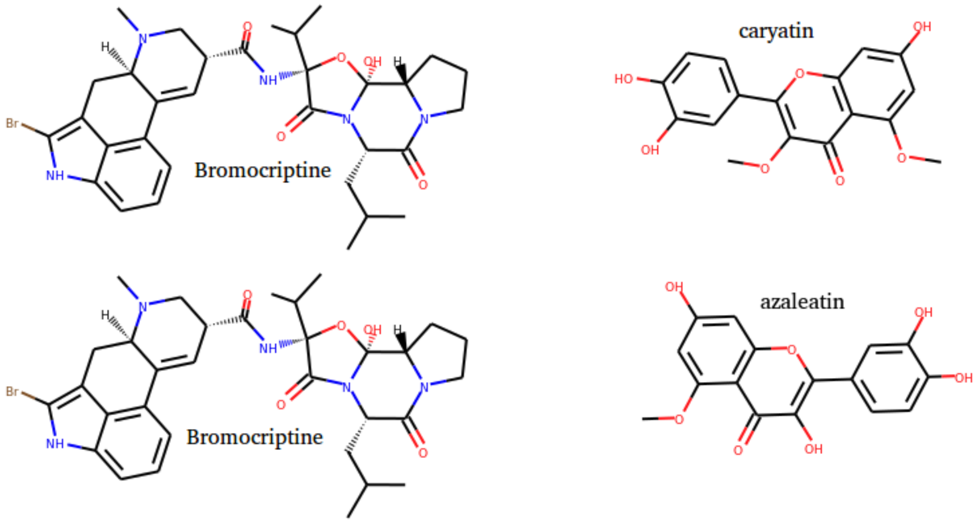


Fig. C.15 Structures of DrugBank DM compounds (left) compared with NANPDB antidiabetic indicated compounds (right) which are listed in table 3.17 Figure: 9/11

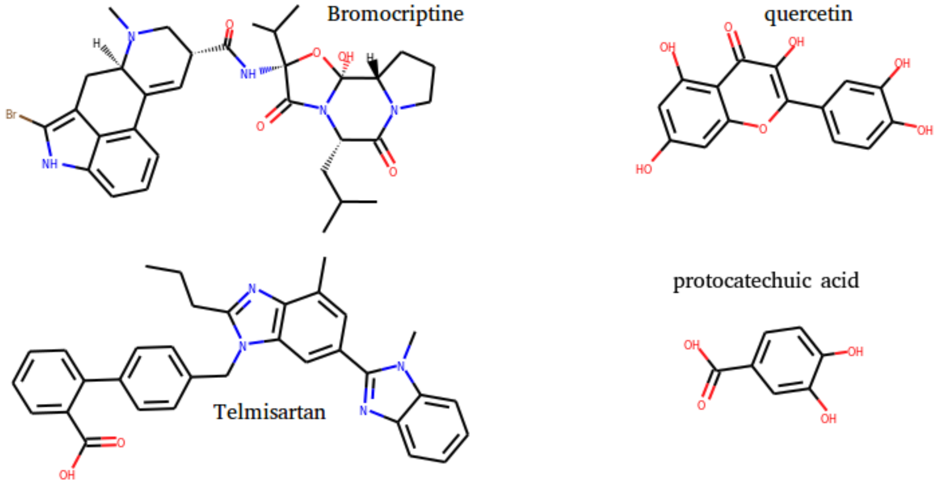


Fig. C.16 Structures of DrugBank DM compounds (left) compared with NANPDB antidiabetic indicated compounds (right) which are listed in table 3.17 Figure: 10/11

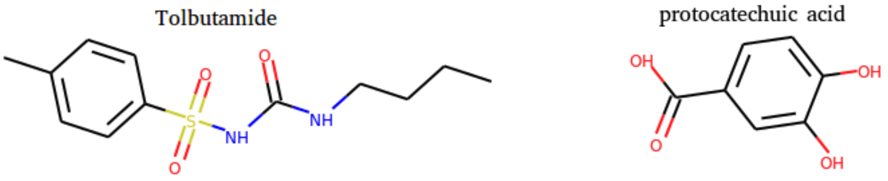


Fig. C.17 Structures of DrugBank DM compounds (left) compared with NANPDB antidiabetic indicated compounds (right) which are listed in table 3.17 Figure: 11/11

Appendix D

SeMPI

D.1 SeMPI results of DM compounds

StDB ID	**StDB compound name**	**StDB organism**	**MIBiG accession**	**MIBiG organism**	**SeMPI result**
5208	CHEMBL1268	*Streptomyces hygroscopicus UC 11099*	BGC0000994	*Streptomyces hygroscopicus subsp. Ascomyceticus*	Too many NRPS signatures
			BGC0000067	*Streptomyces hygroscopicus subsp. Duamyceticus*	Predicted
			BGC0000699	*Streptomyces hygroscopicus subsp. Hygroscopicus*	No signature of modular KS
			BGC0000722	*Streptomyces hygroscopicus subsp. Jinggangensis*	No signature of modular KS
			BGC0000701	*Streptomyces hygroscopicus subsp. Yingchengensis*	No signature of modular KS
778	acarbose	*Streptomyces glaucescens GLA.O*	BGC0000275	*Streptomyces glaucescens*	No signature of modular KS
			BGC0000690		No signature of modular KS

Table D.1 SeMPI results of DM compounds of StreptomeDB 2.0 [22] 2/4

StDB ID	StDB compound name	StDB organism	MIBiG accession	MIBiG organism	SeMPI result
778	acarbose	*Streptomyces lividans 66*	BGC0001168	*Streptomyces lividans 1326*	Unfortunatly no gene cluster could be found in the file you submitted. Pleace control your input file for possible flaws and check if there are clusters present.
			BGC0001283		No signature of modular KS
			BGC0000596	*Streptomyces lividans TK24*	Unfortunatly no gene cluster could be found in the file you submitted. Pleace control your input file for possible flaws and check if there are clusters present.
3310	Trestatin B	*Streptomyces dimorphogenes NR 320-OM7HB*	Does not exist		
3311	Trestatin A	*Streptomyces dimorphogenes NR 320-OM7HB*	Does not exist		
3769	Trestatin C	*Streptomyces dimorphogenes NR 320-OM7HB*	Does not exist		

Table D.2 SeMPI results of DM compounds of StreptomeDB 2.0 [22] 3/4

StDB ID	**StDB compound name**	**StDB organism**	**MIBiG accession**	**MIBiG organism**	**SeMPI result**
4506	acarviostatin I03	*Streptomyces coelicoflavus var. nankaiensis*	BGC0000804	*Streptomyces coelicoflavus ZG0656*	Unfortunatly no gene cluster could be found in the file you submitted. Pleace control your input file for possible flaws and check if there are clusters present.
4507	acarviostatin II03	*Streptomyces coelicoflavus var. nankaiensis*			Unfortunatly no gene cluster could be found in the file you submitted. Pleace control your input file for possible flaws and check if there are clusters present.
4507	acarviostatin II03	*Streptomyces coelicoflavus ZG0656*			Unfortunatly no gene cluster could be found in the file you submitted. Pleace control your input file for possible flaws and check if there are clusters present.
9258	isovalertatin M23	*Streptomyces luteogriseus*	Does not exist		
5079	Aspartocin	*Streptomyces*	Does not exist		

Table D.3 SeMPI results of DM compounds of StreptomeDB 2.0 [22] 4/4

Appendix E

FragPred

E.1 NANPDB DM compound results from FragPred

NANPDB ID	Compound name	Tanimoto between 2 fragments	p-value	Organism	Predicted target where fragment is enriched	Diabetic target (Y/N)
1800	1alpha-acetoxy-3beta-hydroxyeudesm-3-en-6beta,11betaH-12,6-olide	0.36	2.17E-04	*Homo sapiens*	Mu opioid receptor	Y [275]
2259	euphornin D	1.00	1.29E-87	*Rattus norvegicus*	Squalene monooxygenase	N
	euphornin D	0.55	4.66E-04	*Rattus norvegicus*	Squalene monooxygenase	N
2262	euphornin G	1.00	1.29E-87	*Rattus norvegicus*	Squalene monooxygenase	N
	euphornin G	0.58	1.49E-15	*Homo sapiens*	Protein kinase C (PKC)	Y [276]
	euphornin G	0.64	8.42E-05	*Homo sapiens*	Induced myeloid leukemia cell differentiation protein Mcl-1	N
	euphornin G	1.00	8.68E-04	*Rattus norvegicus*	Squalene monooxygenase	N
2316	euphorhelin	1.00	2.39E-123	*Homo sapiens*	Sodium/glucose cotransporter 2	Y [260]
	euphorhelin	1.00	1.29E-87	*Rattus norvegicus*	Squalene monooxygenase	N
	euphorhelin	0.59	1.49E-15	*Homo sapiens*	Protein kinase C (PKC)	Y [276]
	euphorhelin	0.64	8.42E-05	*Homo sapiens*	Induced myeloid leukemia cell differentiation protein Mcl-1	N

Table E.1 Target prediction of NANPDB DM compounds(Table 3.15) 1/2

NANPDB ID	Compound name	Tanimoto between 2 fragments	p-value	Organism	Predicted target where fragment is enriched	Diabetic target (Y/N)
2562	cocciferin T2	1.00	1.29E-87	*Rattus norvegicus*	Squalene monooxygenase	N
	cocciferin T2	1.00	1.29E-87	*Rattus norvegicus*	Squalene monooxygenase	N
	cocciferin T2	0.47	1.49E-15	*Homo sapiens*	Protein kinase C (PKC)	Y [276]
	cocciferin T2	0.87	1.49E-15	*Homo sapiens*	Protein kinase C (PKC)	Y [276]
	cocciferin T2	1.00	5.35E-05	*Homo sapiens*	Alpha-(1,3)-fucosyltransferase 7	Y [277]
	cocciferin T2	0.64	8.42E-05	*Homo sapiens*	Induced myeloid leukemia cell differentiation protein Mcl-1	N
	cocciferin T2	1.00	8.68E-04	*Rattus norvegicus*	Squalene monooxygenase	N
3760	3-(3'-methoxy tropoyloxy) tropane	1.00	1.11E-56	*Rattus norvegicus*	Serotonin 3 (5-HT3) receptor	N
		0.64	4.50E-06	*Homo sapiens*	Glutaminase kidney isofor mitochondrial	N
3765	littorine	1.00	6.18E-75	*Homo sapiens*	Renin	Y [278]
	littorine	1.00	1.11E-56	*Rattus norvegicus*	Serotonin 3 (5-HT3) receptor	N
3949	Sesterstatin 7	0.34	2.80E-05	*Homo sapiens*	Protein-tyrosine phosphatase 1B	Y [279]
3950	16-epi-scalarol butenolide	0.39	2.80E-05	*Homo sapiens*	Protein-tyrosine phosphatase 1B	Y [279]
5158	3,21-dipalmitoyloxy-16beta,21alpha-dihydroxy-beta-amyrine	1.00	7.49E-70	*Homo sapiens*	Lysophosphatidic acid receptor Edg-7	N
		1.00	7.49E-70	*Homo sapiens*	Lysophosphatidic acid receptor Edg-7	N
		0.65	1.04E-05	*Homo sapiens*	CD81 antigen	—

Table E.2 Target prediction of NANPDB DM compounds(Table 3.15) 2/2

Appendix F

Research contributions

F.1 Publications List

F.1.1 Publications

- Ntie-Kang, F., Telukunta, K. K., Döring, K., Simoben, C. V, A Moumbock, A. F., Malange, Y. I., Günther, S. (2017). NANPDB: A Resource for Natural Products from Northern African Sources. Journal of Natural Products, 80(7), 2067–2076. http://doi.org/10.1021/acs.jnatprod.7b00283

- Döring, K., Grüning, B. A., Telukunta, K. K., Thomas, P., & Günther, S. (2016). PubMed-Portable: A framework for supporting the development of text mining applications. PLoS ONE, 11(10). http://doi.org/10.1371/journal.pone.0163794

- Klementz, D., Döring, K., Lucas, X., Telukunta, K. K., Erxleben, A., Deubel, D., Günther, S. (2016). StreptomeDB 2.0 - An extended resource of natural products produced by streptomycetes. Nucleic Acids Research, 44(D1). http://doi.org/10.1093/nar/gkv1319

- Zierep, P. F., Padilla, N., Yonchev, D. G., Telukunta, K. K., Klementz, D., & Günther, S. (2017). SeMPI: a genome-based secondary metabolite prediction and identification web server. Nucleic Acids Research, 45(W1), W64–W71. http://doi.org/10.1093/nar/gkx289

F.1.2 Posters

- Telukunta, KK; Lucas, X; Döring, K; Flemming, S; Günther S. CoRS - A Comprehensive Research Information System For Small Molecules. Oct. 9, 2013, Meeting of German Pharmaceutical Societies 2013, Freiburg, Germany

- Telukunta KK, Lucas X, Döring K, Flemming S, Günther, S. CoRS: Dynamic Information System for Small Molecules. Nov. 10, 2013, German Conference on Chemoinformatics, Fulda, Germany

- Telukunta KK, Lucas X, Flemming S, Günther S. Dynamic virtual screening: reducing the search space within a ligand library. Sep. 24, 2014, DPhG, Frankfurt, Germany

- Telukunta, KK; Ntie-Kang, F; Döring, K; Simoben, C.V; Moumbock, A.F.A; Malange, Y.I; Njume, L.E; Yong, J.N; Sippl, W; Günther S. NANPDB: A database of Natural Products from Northern Africa. Jul. 14, 2017, Tag der Forschung 2017(Day of Science), Freiburg, Germany

- Telukunta KK; Zierep P, Ntie-Kang F; Purschke L; Konrad M; Günther S. RDKit habitation in Pharmaceutical Bioinformatics. Sep. 20, 2017, RDKit User group meeting 2017 (RDKit UGM 2017)

- Fidele Ntie-Kang, Kiran Telukunta, Kersten Döring, Philip N. Judson, Stefan Günther, Joseph N. Yong, Luc M. Mbazee and Wolfgang Sippl. Development of a database of chemical components of African Traditional Medicine: Focus on Northern Africa. Sep. 7, 2015, 2nd European Conference on Natural Product, Frankfurt am Main

- Fidele Ntie-Kang, Kiran K. Telukunta, Kersten Döring, Aurélien F. Adié à Moumbock, Yvette I. Malange, Leonel E. Njume, Wolfgang Sippl, Stefan Günther. Construction Of A Natural Products Database From North African Sources. Sep 4-8, 2016, 21st European Symposium on Quantitative Structure-Activity Relationship Where Molecular Simulations Meet Drug Discovery, Verona, Italy.

- Klementz D.; Döring K.; Lucas X.; Telukunta K.; Thomas O.; Günther S. The StreptomeDB 2.0 Knowledge database of secondary metabolites produced by streptomycetes. July 3, 2015, Freiburg, Day of Science 2015

- Paul Zierep, Natàlia Padilla, Dimitar Yonchev, Kiran Telukunta, Dennis Klementz and Stefan Günther. SeMPI– a genome-based Secondary Metabolite Prediction and Identification web server. Jul. 14, 2017, Tag der Forschung 2017(Day of Science), Freiburg, Germany

F.2 Accademic assistance

F.2.1 Supervised theses

- Moritz Konrad: Fragmentbasierte Analyse von Struktur-Wirkungsbeziehungen kleiner Moleküle (B.Sc), 2016

- Laura van Hazendonk: Protein-interacting drugs: An exploration of the potential of fragment-based target prediction (B.Sc), 2016

- Lea Purschke: Development of a target prediction pipeline based on substructures of pharmaceutical compounds (M.Sc), 2017

F.2.2 Teaching assistance

- Bioinformatik I - In the semesters WS2012/13, WS2013/14, WS2014/15, WS2015/16, WS2016/17

- Bioinformatik II - In the semester SS2013

- Bioinformatik III - In the semesters WS2013/14, WS2014/15, WS2015/16, WS2016/17, WS2017/18

- Molecular Modelling Praktikum - In the semesters WS2013/14, WS2015/16, WS2017/18

F.3 International Graduate Academy (IGA) courses

- Project Management for Research. Jun. 12 & 26, 2013, Freiburg.
- Deutsch als Fremdsprache IV. Apr. 2013, Freiburg.
- Powerful Scientific Presentations. Jun-Jul. 2013, Freiburg.
- Self-Organisation for PhD Candidates. Aug. 1, 2014, Freiburg.
- Berufsoption Selbstständigkeit - Von der Idee zum eigenen Unternehmen. Jan. 25, 2016, Freiburg.

F.4 Presentations

- Chemiebaukasten Spiel, 3D Visualization. Freiburger Wissenschaftsmarkt, Jul. 12-13 2013, Freiburg
- CORS: Dynamic information system for small molecules (Doktorandenseminar(PhD Seminar), Mar. 24 2014, Freiburg)

List of figures

List of tables

List of Listings

Nomenclature

Acronyms / Abbreviations

ADME	Absorbsion Distribution Metabolism And Excretion
API	Application programming interface
BAO	BioAssay Ontology
BET	Bromodomain and extra terminal domain
BGC	Biosynthetic gene clusters
BioC	BioCreative
CFM	Competitive fragmentation modeling
CID	Collision-induced dissociation
CKD	Chronic kidney disease
CMS	Content management system
CPI	Compound Protein interactions
CTB	ChemicalToolBoX
CVD	Cardiovascular disease
DDL	Data definition language
DM	Diabetes mellitus

DML	Data manipulation language
DMPK	Drug metabolism and pharmacokinetics
DTD	Document type definition
DUP	Drug use process
DVS	Dynamic virtual screening
FBDD	Fragment-based drug discovery
FDA	Food and Drug Administration
FPG	Fasting plasma glucose
FragPred	Fragment based target prediction
FDR	False discovery rate
FWER	Familywise error rate
GATE	General Architechture for Text Engineering
H2L	Hit to lead
HGP	Human Genome Project
HTS	High throughput screening
HTTP	Hypertext Transfer Protocol
HTVS	High throughput virtual screening
IDDM	Insulin-dependent diabetes mellitus
IFG	Impaired fasting glucose
IGT	Impaired glucose tolerence
IHD	Ischemic heart disease
KS	Ketoacyl Synthase

LBVS	Ligand-based virtual screening
LOAEL	Lowest observed adverse effect log
LSH	Locality-sensitive hashing
MD	Molecular dynamics
MGC	Metabolic gene clusters
MIBiG	Minimum Information about a Biosynthetic Gene cluster
MPM	Multi-Processing Modules
MS	Mass spectrometry
MVT	Model view template
NANPDB	Northern African Natural Products Database
NER	Named entity recognigion
NIDDM	Non insulin-dependent diabetes mellitus
NLM	National librariy of medicine
NLP	Natural language processing
NMR	Nuclear mangetic resonance
NPs	Natural products
NRPS	Nonribosomal peptide synthetases
OGTT	Oral glucose tolrence test
ORDBMS	Object-relational database management system
ORM	Object relational mapping
PCA	Principle Component Analysis
PG	Plasma glucose

PKS	Polyketide synthases
PPIs	Protein-Protein interactions
RDBMS	Relational Database Management System
RDB	Relational Database
RECAP	Retrosynthetic Combinatorial Analysis Procedure
SBDD	Structure-based drug discovery
SeMPI	Secondary metabolite prediction and identification
SQL	Structured query language
TAEs	Text Analysis Engines
UIMA	Unstructured Information Management Architecture
VS	Virtual Screening
XML	Extensive Markup Language
XRD	X-ray diffraction

References

[1] Lisa B. English. *Combinatorial Library Methods and Protocols*, volume 201. Humana Press, New Jersey, aug 2002. ISBN 1-59259-285-6. doi:10.1385/1592592856. URL http://link.springer.com/10.1385/1592592856.

[2] Asher Mullard. The drug-maker's guide to the galaxy. *Nature*, 549(7673):445–447, sep 2017. ISSN 0028-0836. doi:10.1038/549445a. URL http://www.nature.com/doifinder/10.1038/549445a.

[3] Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, nov 2012. ISSN 1549-9596. doi:10.1021/ci300415d. URL http://pubs.acs.org/doi/10.1021/ci300415d.

[4] Ricardo Macarron, Martyn N. Banks, Dejan Bojanic, David J. Burns, Dragan A. Cirovic, Tina Garyantes, Darren V. S. Green, Robert P. Hertzberg, William P. Janzen, Jeff W. Paslay, Ulrich Schopfer, and G. Sitta Sittampalam. Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, 10(3):188–195, mar 2011. ISSN 1474-1776. doi:10.1038/nrd3368. URL http://www.nature.com/articles/nrd3368.

[5] NHGRI. 2004 release: Ihgsc describes finished human sequence - national human genome research institute (nhgri). https://www.genome.gov/12513430/, October 2004. (Accessed on 05/12/2018).

[6] Xavier Lucas, Björn A. Grüning, Stefan Bleher, and Stefan Günther. The Purchasable Chemical Space: A Detailed Picture. *Journal of Chemical Information and Modeling*, 55(5):915–924, may 2015. ISSN 1549-9596. doi:10.1021/acs.jcim.5b00116. URL http://www.ncbi.nlm.nih.gov/pubmed/25894297http://pubs.acs.org/doi/10.1021/acs.jcim.5b00116.

[7] Eric Patridge, Peter Gareiss, Michael S. Kinch, and Denton Hoyer. An analysis of FDA-approved drugs: natural products and their derivatives. *Drug Discovery Today*, 21(2):204–207, feb 2016. doi:10.1016/J.DRUDIS.2015.01.009. URL https://www.sciencedirect.com/science/article/pii/S1359644615000318?via{%}3Dihub.

[8] Young-Won Chin, Marcy J. Balunas, Hee Byung Chai, and A. Douglas Kinghorn. Drug discovery from natural sources. *The AAPS Journal*, 8(2):E239–E253, jun 2006. ISSN 1550-7416. doi:10.1007/BF02854894. URL http://www.springerlink.com/index/10.1007/BF02854894.

[9] Gordon M. Cragg, Paul G. Grothaus, and David J. Newman. Impact of Natural Products on Developing New Anti-Cancer Agents †. *Chemical Reviews*, 109(7):3012–3043, jul 2009. doi:10.1021/cr900019j. URL http://pubs.acs.org/doi/abs/10.1021/cr900019j.

[10] Gordon M. Cragg, , * David J. Newman, and Kenneth M. Snader. Natural Products in Drug Discovery and Development. 1997. doi:10.1021/NP9604893. URL https://pubs.acs.org/doi/abs/10.1021/np9604893.

[11] A Ganesan. The impact of natural products upon modern drug discovery. *Current Opinion in Chemical Biology*, 12(3):306–317, jun 2008. doi:10.1016/J.CBPA.2008.03.016. URL https://www.sciencedirect.com/science/article/pii/S1367593108000574?via{%}3Dihub.

[12] David J. Newman and Gordon M. Cragg. Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products*, 79(3):629–661, mar 2016. doi:10.1021/acs.jnatprod.5b01055. URL http://pubs.acs.org/doi/10.1021/acs.jnatprod.5b01055.

[13] David J. Newman* And and Gordon M. Cragg. Natural Products as Sources of New Drugs over the Last 25 Years. 2007. doi:10.1021/NP068054V. URL https://pubs.acs.org/doi/abs/10.1021/np068054v.

[14] Seth I Berger and Ravi Iyengar. Role of systems pharmacology in understanding drug adverse events. *Wiley interdisciplinary reviews. Systems biology and medicine*, 3(2):129–35, 2011. ISSN 1939-005X. doi:10.1002/wsbm.114. URL http://www.ncbi.nlm.nih.gov/pubmed/20803507http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3057924.

[15] Jiangyong Gu, Yuanshen Gui, Lirong Chen, Gu Yuan, Hui-Zhe Lu, and Xiaojie Xu. Use of natural products as chemical library for drug discovery and network pharmacology. *PloS one*, 8(4):e62839, apr 2013. ISSN 1932-6203. doi:10.1371/journal.pone.0062839. URL http://dx.plos.org/10.1371/journal.pone.0062839http://www.ncbi.nlm.nih.gov/pubmed/23638153http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3636197.

[16] Paul R Jensen. Natural Products and the Gene Cluster Revolution. *Trends in microbiology*, 24(12):968–977, 2016. ISSN 1878-4380. doi:10.1016/j.tim.2016.07.006. URL http://www.ncbi.nlm.nih.gov/pubmed/27491886http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5123934.

[17] Ben Shen. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current opinion in chemical biology*, 7(2):285–95, apr 2003. ISSN 1367-5931. URL http://www.ncbi.nlm.nih.gov/pubmed/12714063.

[18] Michael A. Fischbach and Christopher T. Walsh. Assembly-Line Enzymology for Polyketide and Nonribosomal Peptide Antibiotics: Logic, Machinery, and Mechanisms. *Chemical Reviews*, 106(8):3468–3496, aug 2006. ISSN 0009-2665. doi:10.1021/cr0503097. URL http://www.ncbi.nlm.nih.gov/pubmed/16895337http://pubs.acs.org/doi/abs/10.1021/cr0503097.

[19] Christian Hertweck. The Biosynthetic Logic of Polyketide Diversity. *Angewandte Chemie International Edition*, 48(26):4688–4716, jun 2009. ISSN 14337851. doi:10.1002/anie.200806121. URL http://www.ncbi.nlm.nih.gov/pubmed/19514004http://doi.wiley.com/10.1002/anie.200806121.

[20] Stephan A. Sieber and Mohamed A. Marahiel. Molecular Mechanisms Underlying Nonribosomal Peptide Synthesis: Approaches to New Antibiotics. *Chemical Reviews*, 105(2):715–738, feb 2005. ISSN 0009-2665. doi:10.1021/cr0301191. URL http://www.ncbi.nlm.nih.gov/pubmed/15700962http://pubs.acs.org/doi/abs/10.1021/cr0301191.

[21] Eric J N Helfrich and Jörn Piel. Biosynthesis of polyketides by trans-AT polyketide synthases. *Natural product reports*, 33(2):231–316, feb 2016. ISSN 1460-4752. doi:10.1039/c5np00125k. URL http://www.ncbi.nlm.nih.gov/pubmed/26689670.

[22] Dennis Klementz, Kersten Döring, Xavier Lucas, Kiran K Telukunta, Anika Erxleben, Denise Deubel, Astrid Erber, Irene Santillana, Oliver S Thomas, Andreas Bechthold, and Stefan Günther. StreptomeDB 2.0–an extended resource of natural products produced by streptomycetes. *Nucleic acids research*, 44(D1):D509–14, jan 2016. ISSN 1362-4962. doi:10.1093/nar/gkv1319. URL http://www.ncbi.nlm.nih.gov/pubmed/26615197http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4702922.

[23] Heval Atas, Nurcan Tuncbag, and Tunca Doğan. Phylogenetic and Other Conservation-Based Approaches to Predict Protein Functional Sites. In *Methods in molecular biology (Clifton, N.J.)*, volume 1762, pages 51–69. 2018. doi:10.1007/978-1-4939-7756-7_4. URL http://www.ncbi.nlm.nih.gov/pubmed/29594767http://link.springer.com/10.1007/978-1-4939-7756-7{_}4.

[24] Mohd Shukri Baba, Noraziah Mohamad Zin, Zainal Abidin Abu Hassan, Jalifah Latip, Florence Pethick, Iain S. Hunter, RuAngelie Edrada-Ebel, and Paul R. Herron. In vivo antimalarial activity of the endophytic actinobacteria, Streptomyces SUK 10. *Journal of Microbiology*, 53(12):847–855, dec 2015. ISSN 1225-8873. doi:10.1007/s12275-015-5076-6. URL http://www.ncbi.nlm.nih.gov/pubmed/26626355http://link.springer.com/10.1007/s12275-015-5076-6.

[25] Alexander Heifetz, Gebhard F X Schertler, Roland Seifert, Christopher G Tate, Patrick M Sexton, Vsevolod V Gurevich, Daniel Fourmy, Vadim Cherezov, Fiona H Marshall, R Ian Storer, Isabel Moraes, Irina G Tikhonova, Christofer S Tautermann, Peter Hunt, Tom Ceska, Simon Hodgson, Mike J Bodkin, Shweta Singh, Richard J Law, and Philip C Biggin. GPCR structure, function, drug discovery and crystallography: report from Academia-Industry International Conference (UK Royal Society) Chicheley Hall, 1-2 September 2014. *Naunyn-Schmiedeberg's archives of pharmacology*, 388(8):883–903, aug 2015. ISSN 1432-1912. doi:10.1007/s00210-015-1111-8. URL http://www.ncbi.nlm.nih.gov/pubmed/25772061http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4495723.

[26] Leah I Elizondo, Paymaan Jafar-Nejad, J Marietta Clewing, and Cornelius F Boerkoel. Gene clusters, molecular evolution and disease: a speculation. *Current genomics*, 10(1):64–75, mar 2009. ISSN 1389-2029. doi:10.2174/138920209787581271. URL http://www.ncbi.nlm.nih.gov/pubmed/19721813http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2699835.

[27] Narayanan Raghupathy and Dannie Durand. Gene cluster statistics with gene families. *Molecular biology and evolution*, 26(5):957–68, may 2009. ISSN 1537-1719. doi:10.1093/molbev/msp002. URL http://www.ncbi.nlm.nih.gov/pubmed/19150803http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2668827.

[28] Anne Osbourn. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. *Trends in Genetics*, 26(10):449–457, oct 2010. doi:10.1016/J.TIG.2010.07.001. URL https://www.sciencedirect.com/science/article/pii/S0168952510001447.

[29] Kozo Ochi and Takeshi Hosaka. New strategies for drug discovery: activation of silent or weakly expressed microbial gene clusters. *Applied microbiology and biotechnology*, 97(1):87–98, jan 2013. ISSN 1432-0614. doi:10.1007/s00253-012-4551-9. URL http://www.ncbi.nlm.nih.gov/pubmed/23143535http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3536979.

[30] Guido F Pauli, Shao-Nong Chen, J Brent Friesen, James B McAlpine, and Birgit U Jaki. Analysis and purification of bioactive natural products: the AnaPurNa study. *Journal of natural products*, 75(6):1243–55, jun 2012. ISSN 1520-6025. doi:10.1021/np300066q. URL http://www.ncbi.nlm.nih.gov/pubmed/22620854http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3381453.

[31] Hsiao-Hang Chung, Yi-Chang Sung, and Lie-Fen Shyur. Deciphering the Biosynthetic Pathways of Bioactive Compounds In Planta Using Omics Approaches. In *Medicinal Plants - Recent Advances in Research and Development*, pages 129–165. Springer Singapore, Singapore, 2016. doi:10.1007/978-981-10-1085-9_5. URL http://link.springer.com/10.1007/978-981-10-1085-9{_}5.

[32] K. A. Houck, D. J. Dix, R. S. Judson, R. J. Kavlock, J. Yang, and E. L. Berg. Profiling Bioactivity of the ToxCast Chemical Library Using BioMAP Primary Human Cell Systems. *Journal of Biomolecular Screening*, 14(9):1054–1066, oct 2009. doi:10.1177/1087057109345525. URL http://jbx.sagepub.com/cgi/doi/10.1177/1087057109345525.

[33] Alastair J Fischer and Gemma Ghelardi. The Precautionary Principle, Evidence-Based Medicine, and Decision Theory in Public Health Evaluation. *Frontiers in public health*, 4:107, 2016. ISSN 2296-2565. doi:10.3389/fpubh.2016.00107. URL http://www.ncbi.nlm.nih.gov/pubmed/27458575http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4935673.

[34] NCBI. Home - pubmed - ncbi. https://www.ncbi.nlm.nih.gov/pubmed.

[35] Joanna Owens. Target validation: Determining druggability. *Nature Reviews Drug Discovery 2007 6:3*, mar 2007.

[36] J P Hughes, S Rees, S B Kalindjian, and K L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–49, mar 2011. ISSN 1476-5381. doi:10.1111/j.1476-5381.2010.01127.x. URL http://www.ncbi.nlm.nih.gov/pubmed/21091654http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3058157.

[37] Sandra Fox, Shauna Farr-Jones, Lynne Sopchak, Amy Boggs, Helen Wang Nicely, Richard Khoury, and Michael Biros. High-Throughput Screening: Update on Practices and Success. *Journal of Biomolecular Screening*, 11(7):864–869, oct 2006. ISSN 1087-0571. doi:10.1177/1087057106292473. URL http://journals.sagepub.com/doi/10.1177/1087057106292473.

[38] A. Lavecchia and C Di Giovanni. Virtual screening strategies in drug discovery: a critical review. *Current medicinal chemistry*, 20(23):2839–60, jun 2013. ISSN 1875-533X. doi:10.2174/09298673113209990001. URL http://www.eurekaselect.com/openurl/content.php?genre=article{&}issn=0929-8673{&}volume=20{&}issue=23{&}spage=2839http://www.ncbi.nlm.nih.gov/pubmed/23651302.

[39] Hu Ge, Yu Wang, Chanjuan Li, Nanhao Chen, Yufang Xie, Mengyan Xu, Yingyan He, Xinchun Gu, Ruibo Wu, Qiong Gu, Liang Zeng, and Jun Xu. Molecular Dynamics-Based Virtual Screening: Accelerating the Drug Discovery Process by High-Performance Computing. *Journal of Chemical Information and Modeling*, 53(10):2757–2764, oct 2013. doi:10.1021/ci400391s. URL http://pubs.acs.org/doi/10.1021/ci400391s.

[40] Evanthia Lionta, George Spyrou, Demetrios K Vassilatis, and Zoe Cournia. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current topics in medicinal chemistry*, 14(16):1923–38, 2014. ISSN 1873-4294. doi:10.2174/1568026614666140929124445. URL http://www.ncbi.nlm.nih.gov/pubmed/25262799http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4443793.

[41] Andrew L Hopkins, György M Keserü, Paul D Leeson, David C Rees, and Charles H Reynolds. The role of ligand efficiency metrics in drug discovery. *Nature reviews. Drug discovery*, 13(2):105–121, jan 2014. ISSN 1474-1784. doi:10.1038/nrd4163. URL http://www.nature.com/doifinder/10.1038/nrd4163http://www.ncbi.nlm.nih.gov/pubmed/24481311.

[42] Markus Hartenfeller and Gisbert Schneider. De Novo Drug Design. pages 299–323. Humana Press, Totowa, NJ, 2010. doi:10.1007/978-1-60761-839-3_12. URL http://link.springer.com/10.1007/978-1-60761-839-3{_}12.

[43] Peter Gedeck, Bernhard Rohde, , and Christian Bartels. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. 2006. doi:10.1021/CI050413P. URL https://pubs.acs.org/doi/abs/10.1021/ci050413p.

[44] Fidele Ntie-Kang, Conrad Veranso Simoben, Berin Karaman, Valery Fuh Ngwa, Philip Neville Judson, Wolfgang Sippl, and Luc Meva'a Mbaze. Pharmacophore modeling and in silico toxicity assessment of potential anticancer agents from African medicinal plants. *Drug design, development and therapy*, 10:2137–54, jul 2016. ISSN 1177-8881. doi:10.2147/DDDT.S108118. URL http://www.ncbi.nlm.nih.gov/pubmed/27445461http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4938243https://goo.gl/gqa5p6.

[45] Bruno O. Villoutreix, Nicolas Renault, David Lagorce, Matthieu Montes, and Maria A. Miteva. Free Resources to Assist Structure-Based Virtual Ligand Screening Experiments. *Current Protein & Peptide Science*, 8(4):381–411, aug 2007. ISSN 13892037. doi:10.2174/138920307781369391. URL http://www.eurekaselect.com/openurl/content.php?genre=article{&}issn=1389-2037{&}volume=8{&}issue=4{&}spage=381.

[46] Andreas Bender and Robert C Glen. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2(22):3204–18, nov 2004. ISSN 1477-0520. doi:10.1039/B409813G. URL http://www.ncbi.nlm.nih.gov/pubmed/15534697.

[47] Sereina Riniker and Gregory A Landrum. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of cheminformatics*, 5(1):26, may 2013. ISSN 1758-2946. doi:10.1186/1758-2946-5-26.

URL http://jcheminf.springeropen.com/articles/10.1186/1758-2946-5-26http://www.ncbi.nlm.nih.gov/pubmed/23721588http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3686626.

[48] Mark D. Mackey and James L. Melville. Better than random? The chemotype enrichment problem. *Journal of chemical information and modeling*, 49(5):1154–62, may 2009. ISSN 1549-9596. doi:10.1021/ci8003978. URL http://pubs.acs.org/doi/abs/10.1021/ci8003978http://www.ncbi.nlm.nih.gov/pubmed/19397275.

[49] Juan C Del Álamo, Derek Lemons, Ricardo Serrano, Alex Savchenko, Fabio Cerignoli, Rolf Bodmer, and Mark Mercola. High throughput physiological screening of iPSC-derived cardiomyocytes for drug development. *Biochimica et biophysica acta*, 1863(7 Pt B):1717–27, jul 2016. ISSN 0006-3002. doi:10.1016/j.bbamcr.2016.03.003. URL http://www.ncbi.nlm.nih.gov/pubmed/26952934http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4885786.

[50] Carlo Baggio, Parima Udompholkul, Elisa Barile, and Maurizio Pellecchia. Enthalpy-Based Screening of Focused Combinatorial Libraries for the Identification of Potent and Selective Ligands. *ACS chemical biology*, 12(12):2981–2989, dec 2017. ISSN 1554-8937. doi:10.1021/acschembio.7b00717. URL http://pubs.acs.org/doi/10.1021/acschembio.7b00717http://www.ncbi.nlm.nih.gov/pubmed/29094589.

[51] Xiao Qing Lewell, Duncan B Judd, Stephen P Watson, and Michael M Hann. RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *Journal of Chemical Information and Computer Sciences*, 38(3):511–522, may 1998. ISSN 0095-2338. doi:10.1021/ci970429i. URL http://www.ncbi.nlm.nih.gov/pubmed/9611787http://pubs.acs.org/doi/abs/10.1021/ci970429i.

[52] Peter J Hore. *Nuclear magnetic resonance*. Oxford University Press, USA, 2015.

[53] Maurizio Pellecchia, Ivano Bertini, David Cowburn, Claudio Dalvit, Ernest Giralt, Wolfgang Jahnke, Thomas L James, Steve W Homans, Horst Kessler, Claudio Luchinat, Bernd Meyer, Hartmut Oschkinat, Jeff Peng, Harald Schwalbe, and Gregg Siegal. Perspectives on NMR in drug discovery: a technique comes of age. *Nature reviews. Drug discovery*, 7(9):738–45, sep 2008. ISSN 1474-1776. doi:10.1038/nrd2606. URL http://www.ncbi.nlm.nih.gov/pubmed/19172689http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2891904.

[54] Bill Boggess. Mass Spectrometry Desk Reference (Sparkman, O. David). *Journal of Chemical Education*, 78(2):168, feb 2001. ISSN 0021-9584. doi:10.1021/ed078p168.2. URL http://pubs.acs.org/doi/abs/10.1021/ed078p168.2.

[55] Schrödinger Release 2017-4. Qikprop schrdinger, llc, new york. https://www.schrodinger.com/qikprop, 07 2014. (Accessed On 4/1/2018 21:15).

[56] LLC-New York QikProp, Schrödinger. Qikprop descriptors and properties. http://gohom.win/ManualHom/Schrodinger/Schrodinger_2012_docs/general/qikprop_props.pdf.

[57] Soraya Dhillon and Kiren Gill. *Basic Pharmacokinetics*. 2009. URL https://www.dandybooksellers.com/acatalog/9780853695714.pdf.

[58] Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 46(1-3):3–26, mar 2001. ISSN 0169-409X. doi:10.1016/S0169-409X(96)00423-1. URL http://www.sciencedirect.com/science/article/pii/S0169409X96004231?via{%}3Dihubhttp://www.ncbi.nlm.nih.gov/pubmed/11259830.

[59] Christopher A. Lipinski. Lead- and drug-like compounds: the rule-of-five revolution. *Drug discovery today. Technologies*, 1(4):337–41, dec 2004. ISSN 1740-6749. doi:10.1016/j.ddtec.2004.11.007. URL https://www.sciencedirect.com/science/article/pii/S1740674904000551?via{%}3Dihubhttp://www.ncbi.nlm.nih.gov/pubmed/24981612.

[60] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, 2004. ISSN 00222623. doi:10.1021/jm0306430. URL https://pubs.acs.org/doi/abs/10.1021/jm0306430.

[61] New York USA Schrödinger, LLC. Glide 5.6. schrödinger, llc., new york, usa. https://www.schrodinger.com/glide, 07 2014.

[62] Richard A. Friesner, Robert B. Murphy, Matthew P. Repasky, Leah L. Frye, Jeremy R. Greenwood, Thomas A. Halgren, Paul C. Sanschagrin, and Daniel T. Mainz. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry*, 49(21):6177–6196, 2006. ISSN 00222623. doi:10.1021/jm051256o. URL https://pubs.acs.org/doi/abs/10.1021/jm051256o.

[63] Sarah R Langdon, Nathan Brown, and Julian Blagg. Scaffold diversity of exemplified medicinal chemistry space. *Journal of chemical information and modeling*, 51(9):2174–85, sep 2011. ISSN 1549-960X. doi:10.1021/ci2001428. URL http://www.ncbi.nlm.nih.gov/pubmed/21877753http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3180201.

[64] Guy W. Bemis And and Mark A. Murcko. The Properties of Known Drugs. 1. Molecular Frameworks. 1996. doi:10.1021/JM9602928. URL https://pubs.acs.org/doi/abs/10.1021/jm9602928.

[65] Ansgar Schuffenhauer, Peter Ertl, Silvio Roggo, Stefan Wetzel, Marcus A. Koch, and Herbert Waldmann. The scaffold tree–visualization of the scaffold universe by hierarchical scaffold classification. *Journal of chemical information and modeling*, 47(1):47–58, jan 2007. ISSN 1549-9596. doi:10.1021/ci600338x. URL http://pubs.acs.org/doi/abs/10.1021/ci600338xhttp://www.ncbi.nlm.nih.gov/pubmed/17238248.

[66] Robert P Sheridan. The most common chemical replacements in drug-like compounds. *Journal of chemical information and computer sciences*, 42(1):103–8, 2002. ISSN 0095-2338. doi:10.1021/CI0100806. URL https://pubs.acs.org/doi/abs/10.1021/ci0100806http://www.ncbi.nlm.nih.gov/pubmed/11855973.

[67] Robert P. Sheridan*. Finding Multiactivity Substructures by Mining Databases of Drug-Like Compounds. 2003. doi:10.1021/CI030004Y. URL https://pubs.acs.org/doi/abs/10.1021/ci030004y.

[68] Anne Mai Wassermann, Mathias Wawer, and Jürgen Bajorath. Activity Landscape Representations for Structure-Activity Relationship Analysis. *Journal of Medicinal Chemistry*, 53(23):8209–8223, dec 2010. ISSN 0022-2623. doi:10.1021/jm100933w. URL http://pubs.acs.org/doi/abs/10.1021/jm100933w.

[69] C A. James and D. Weininger. Daylight theory manual. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html, 1995.

[70] Noel M. O'Boyle and Roger A. Sayle. Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of Cheminformatics*, 8(1):36, dec 2016. doi:10.1186/s13321-016-0148-0. URL http://jcheminf.springeropen.com/articles/10.1186/s13321-016-0148-0.

[71] Raymond E. Carhart, Dennis H. Smith, and R. Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Modeling*, 25(2):64–73, may 1985. ISSN 1549-9596. doi:10.1021/ci00046a002. URL http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00046a002.

[72] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, may 2010. ISSN 1549-9596. doi:10.1021/ci100050t. URL http://pubs.acs.org/doi/abs/10.1021/ci100050t.

[73] Shereena M. Arif, John D. Holliday, and Peter Willett. Analysis and use of fragment-occurrence data in similarity-based virtual screening. *Journal of Computer-Aided Molecular Design*, 23(9):655–668, sep 2009. doi:10.1007/s10822-009-9285-0. URL http://link.springer.com/10.1007/s10822-009-9285-0.

[74] Dr. Alex M. Clark. Open source ecfp/fcfp circular fingerprints in cdk – cheminformatics 2.0. https://cheminf20.org/2014/02/21/open-source-ecfpfcfp-circular-fingerprints-in-cdk/. (Accessed on 05/07/2018).

[75] OpenEye. Glossary — toolkits – python. https://docs.eyesopen.com/toolkits/python/graphsimtk/glossary.html#term-maccs. (Accessed on 05/07/2018).

[76] Ramaswamy Nilakantan, Norman Bauman, J. Scott Dixon, and R. Venkataraghavan. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *Journal of Chemical Information and Modeling*, 27(2):82–85, may 1987. ISSN 1549-9596. doi:10.1021/ci00054a008. URL http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00054a008.

[77] RDKit. rdkit.chem.rdmolops. http://rdkit.org/Python_Docs/rdkit.Chem.rdmolops-module.html#RDKFingerprint. (Accessed on 05/07/2018).

[78] John MacCuish, Christos Nicolaou, , and Norah E. MacCuish. Ties in Proximity and Clustering Compounds. 2000. doi:10.1021/CI000069Q. URL https://pubs.acs.org/doi/abs/10.1021/ci000069q.

[79] Greg Landrum. An overview of the rdkit — the rdkit 2018.03.1 documentation. http://www.rdkit.org/docs/Overview.html, . (Accessed on 05/09/2018).

[80] R Core Team. R: A language and environment for statistical computing, 2018. URL https://www.R-project.org.

[81] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10): R80, 2004. ISSN 1474-760X. doi:10.1186/gb-2004-5-10-r80. URL http://www.ncbi.nlm.nih.gov/pubmed/15461798http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC545600.

[82] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11): 1422–3, jun 2009. ISSN 1367-4811. doi:10.1093/bioinformatics/btp163. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp163http://www.ncbi.nlm.nih.gov/pubmed/19304878http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2682512.

[83] Jason E Stajich, David Block, Kris Boulez, Steven E Brenner, Stephen A Chervitz, Chris Dagdigian, Georg Fuellen, James G R Gilbert, Ian Korf, Hilmar Lapp, Heikki Lehväslaiho, Chad Matsalla, Chris J Mungall, Brian I Osborne, Matthew R Pocock, Peter Schattner, Martin Senger, Lincoln D Stein, Elia Stupka, Mark D Wilkinson, and Ewan Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome research*, 12(10): 1611–8, oct 2002. ISSN 1088-9051. doi:10.1101/gr.361602. URL http://www.ncbi.nlm.nih.gov/pubmed/12368254http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC187536.

[84] Galaxy | tool shed. https://toolshed.g2.bx.psu.edu/.

[85] Bjoern Gruening. galaxytools/chemicaltoolbox at master · bgruening/galaxytools. https://github.com/bgruening/galaxytools/tree/master/chemicaltoolbox, Apr 2013.

[86] R. A. Fisher. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87, jan 1922. ISSN 09528385. doi:10.2307/2340521. URL https://www.jstor.org/stable/2340521?origin=crossref.

[87] Alan Agresti and Alan Agresti. A survey of exact inference for contingency tables. 1992. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.296.874.

[88] Jelle J. Goeman and Aldo Solari. Multiple hypothesis testing in genomics. *Statistics in Medicine*, 33(11):1946–1978, may 2014. ISSN 02776715. doi:10.1002/sim.6082. URL http://doi.wiley.com/10.1002/sim.6082.

[89] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, 1995. URL https://www.jstor.org/stable/2346101.

[90] Brilliant.org. Entropy (information theory) | brilliant math & science wiki. https://brilliant.org/wiki/entropy-information-theory/, 06 2018.

[91] E. F. Codd and E. F. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6):377–387, jun 1970. ISSN 00010782. doi:10.1145/362384.362685. URL http://portal.acm.org/citation.cfm?doid=362384.362685.

[92] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):33, oct 2011. ISSN 1758-2946. doi:10.1186/1758-2946-3-33. URL http://www.jcheminf.com/content/3/1/33http://www.ncbi.nlm.nih.gov/pubmed/21982300http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3198950.

[93] Lawrence Journal-World. Django (version 1.10) [computer software]. url-https://www.djangoproject.com/, 2017.

[94] Mezzanine. Mezzanine - the best django cms. http://mezzanine.jupo.org/.

[95] World Life Expectancy. World rankings-total deaths. http://www.worldlifeexpectancy.com/world-rankings-total-deaths. (Accessed on 06/11/2018).

[96] Abbas E Kitabchi, Guillermo E Umpierrez, John M Miles, and Joseph N Fisher. Hyperglycemic crises in adult patients with diabetes. *Diabetes care*, 32(7):1335–43, jul 2009. ISSN 1935-5548. doi:10.2337/dc09-9032. URL http://www.ncbi.nlm.nih.gov/pubmed/19564476http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2699725.

[97] T. Nakagami and the DECODA Study Group. Hyperglycaemia and mortality from all causes and from cardiovascular disease in five populations of Asian origin. *Diabetologia*, 47(3):385–394, mar 2004. ISSN 0012-186X. doi:10.1007/s00125-004-1334-6. URL http://link.springer.com/10.1007/s00125-004-1334-6.

[98] E Bonora, S Kiechl, J Willeit, F Oberhollenzer, G Egger, R Bonadonna, and M Muggeo. Plasma glucose within the normal range is not associated with carotid atherosclerosis: prospective results in subjects with normal glucose tolerance from the Bruneck Study. *Diabetes care*, 22(8):1339–46, aug 1999. ISSN 0149-5992. doi:10.2337/diacare.27.12.2836. URL http://www.ncbi.nlm.nih.gov/pubmed/10480780.

[99] Agnese Fiori, Vincenzo Terlizzi, Heiner Kremer, Julian Gebauer, Hans-Peter Hammes, Martin C. Harmsen, and Karen Bieback. Mesenchymal stromal/stem cells as potential therapy in diabetic retinopathy. *Immunobiology*, feb 2018. ISSN 0171-2985. doi:10.1016/J.IMBIO.2018.01.001. URL https://www.sciencedirect.com/science/article/pii/S0171298518300019?via{%}3Dihub.

[100] Elizabeth J. Mayer-Davis, Dana Dabelea, and Jean M. Lawrence. Incidence Trends of Type 1 and Type 2 Diabetes among Youths, 2002-2012. *The New England journal of medicine*, 377(3):301, apr 2017. ISSN 1533-4406. doi:10.1056/NEJMc1706291. URL http://www.nejm.org/doi/10.1056/NEJMoa1610187http://www.ncbi.nlm.nih.gov/pubmed/28723318http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5639715.

[101] B E Metzger and D R Coustan. Summary and recommendations of the Fourth International Workshop-Conference on Gestational Diabetes Mellitus. The Organizing Committee. *Diabetes care*, 21 Suppl 2:B161–7, aug 1998. ISSN 0149-5992. URL http://www.ncbi.nlm.nih.gov/pubmed/9704245.

[102] DECODE Study Group, the European Diabetes Epidemiology Group. Glucose tolerance and cardiovascular mortality: comparison of fasting and 2-hour diagnostic criteria. *Archives of internal medicine*, 161(3):397–405, feb 2001. ISSN 0003-9926. URL http://www.ncbi.nlm.nih.gov/pubmed/11176766.

[103] Kay-Tee Khaw, Nicholas Wareham, Sheila Bingham, Robert Luben, Ailsa Welch, and Nicholas Day. Association of Hemoglobin A <sub>1c</sub> with Cardiovascular Disease and Mortality in Adults: The European Prospective Investigation into Cancer in Norfolk. *Annals of Internal Medicine*, 141(6):413, sep 2004. ISSN 0003-4819. doi:10.7326/0003-4819-141-6-200409210-00006. URL http://annals.org/article.aspx?doi=10.7326/0003-4819-141-6-200409210-00006.

[104] Maliheh Safavi, Alireza Foroumadi, and Mohammad Abdollahi. The importance of synthetic drugs for type 2 diabetes drug discovery. *Expert Opinion on Drug Discovery*, 8(11):1339–1363, nov 2013. ISSN 1746-0441. doi:10.1517/17460441.2013.837883. URL http://www.ncbi.nlm.nih.gov/pubmed/24050217http://www.tandfonline.com/doi/full/10.1517/17460441.2013.837883.

[105] Anand-Krishna Singh, Rameshwar Jatwa, Ashok Purohit, and Heera Ram. Synthetic and phytocompounds based dipeptidyl peptidase-IV (DPP-IV) inhibitors for therapeutics of diabetes. *Journal of Asian Natural Products Research*, 19(10):1036–1045, oct 2017. ISSN 1028-6020. doi:10.1080/10286020.2017.1307183. URL http://www.ncbi.nlm.nih.gov/pubmed/28351157https://www.tandfonline.com/doi/full/10.1080/10286020.2017.1307183.

[106] Shuai Xue, Jianli Yin, Jiawei Shao, Yuanhuan Yu, Linfeng Yang, Yidan Wang, Mingqi Xie, Martin Fussenegger, and Haifeng Ye. A Synthetic-Biology-Inspired Therapeutic Strategy for Targeting and Treating Hepatogenous Diabetes. *Molecular Therapy*, 25(2):443–455, feb 2017. ISSN 15250016. doi:10.1016/j.ymthe.2016.11.008. URL http://www.ncbi.nlm.nih.gov/pubmed/28153094http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5368401http://linkinghub.elsevier.com/retrieve/pii/S1525001616453918.

[107] Paulina Strzyz. Designer cells tackle diabetes. *Nature Reviews Molecular Cell Biology*, 18(2):69–69, feb 2017. ISSN 1471-0072. doi:10.1038/nrm.2016.175. URL http://www.ncbi.nlm.nih.gov/pubmed/28053346http://www.nature.com/articles/nrm.2016.175.

[108] Lingman Ma, Na Lu, and Guanzhong Wu. Antiplatelet aggregation and endothelial protection of I <sub>4</sub> , a new synthetic anti-diabetes sulfonylurea compound. *Platelets*, 26(4):342–348, may 2015. ISSN 0953-7104. doi:10.3109/09537104.2014.912749. URL http://www.ncbi.nlm.nih.gov/pubmed/24832568http://www.tandfonline.com/doi/full/10.3109/09537104.2014.912749.

[109] Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS computational biology*, 9(2):e1002854, feb 2013. ISSN 1553-7358. doi:10.1371/journal.pcbi.1002854. URL http://dx.plos.org/10.1371/journal.pcbi.1002854http:

//www.ncbi.nlm.nih.gov/pubmed/23408875http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3567135.

[110] Domonkos Tikk, Illés Solt, Philippe Thomas, and Ulf Leser. A detailed error analysis of 13 kernel methods for protein-protein interaction extraction. *BMC bioinformatics*, 14(1):12, jan 2013. ISSN 1471-2105. doi:10.1186/1471-2105-14-12. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-14-12http://www.ncbi.nlm.nih.gov/pubmed/23323857http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3680070.

[111] Luis Tari, Saadat Anwar, Shanshan Liang, James Cai, and Chitta Baral. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. *Bioinformatics (Oxford, England)*, 26(18): i547–53, sep 2010. ISSN 1367-4811. doi:10.1093/bioinformatics/btq382. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btq382http://www.ncbi.nlm.nih.gov/pubmed/20823320http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2935409.

[112] Christian Senger, Björn A. Grüning, Anika Erxleben, Kersten Döring, Hitesh Patel, Stephan Flemming, Irmgard Merfort, and Stefan Günther. Mining and evaluation of molecular relationships in literature. *Bioinformatics (Oxford, England)*, 28(5): 709–14, mar 2012. ISSN 1367-4811. doi:10.1093/bioinformatics/bts026. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts026http://www.ncbi.nlm.nih.gov/pubmed/22247277.

[113] Fabio Rinaldi, Simon Clematide, Hernani Marques, Tilia Ellendorff, Martin Romacker, and Raul Rodriguez-Esteban. OntoGene web services for biomedical text mining. *BMC Bioinformatics*, 15(Suppl 14):S6, nov 2014. ISSN 1471-2105. doi:10.1186/1471-2105-15-S14-S6. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S14-S6.

[114] Kersten Döring, Björn A. Grüning, Kiran K. Telukunta, Philippe Thomas, and Stefan Günther. PubMedPortable: A Framework for Supporting the Development of Text Mining Applications. *PLOS ONE*, 11(10):e0163794, oct 2016. ISSN 1932-6203. doi:10.1371/journal.pone.0163794. URL http://dx.plos.org/10.1371/journal.pone.0163794.

[115] Donald C. Comeau, Riza Theresa Batista-Navarro, Hong-Jie Dai, Rezarta Islamaj Doğan, Antonio Jimeno Yepes, Ritu Khare, Zhiyong Lu, Hernani Marques, Carolyn J. Mattingly, Mariana Neves, Yifan Peng, Rafal Rak, Fabio Rinaldi, Richard Tzong-Han Tsai, Karin Verspoor, Thomas C. Wiegers, Cathy H. Wu, and W. John Wilbur. BioC interoperability track overview. *Database : the journal of biological databases and curation*, 2014(0):bau053–bau053, jun 2014. ISSN 1758-0463. doi:10.1093/database/bau053. URL https://academic.oup.com/database/article-lookup/doi/10.1093/database/bau053http://www.ncbi.nlm.nih.gov/pubmed/24980129http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4074764.

[116] Donald C. Comeau, Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, Alfonso Valencia, Karin Verspoor, Thomas C. Wiegers, Cathy H. Wu, and W. John Wilbur. BioC: a minimalist approach to interoperability for biomedical text processing. *Database : the journal of biological databases and curation*, 2013(0):bat064, sep 2013.

ISSN 1758-0463. doi:10.1093/database/bat064. URL https://academic.oup.com/database/article-lookup/doi/10.1093/database/bat064http://www.ncbi.nlm.nih.gov/pubmed/24048470http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3889917.

[117] Eric Sayers. The e-utilities in-depth: Parameters, syntax and more - entrez programming utilities help - ncbi bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK25499/, 2 2018. (Accessed on 02/17/2018).

[118] Alias i. 2008. Lingpipe home. http://alias-i.com/lingpipe/, 2008. (Accessed on 26/02/2018).

[119] Rafal Rak, Riza Theresa Batista-Navarro, Jacob Carter, Andrew Rowley, and Sophia Ananiadou. Processing biological literature with customizable Web services supporting interoperable formats. *Database : the journal of biological databases and curation*, 2014(0):bau064–bau064, jul 2014. ISSN 1758-0463. doi:10.1093/database/bau064. URL https://academic.oup.com/database/article-lookup/doi/10.1093/database/bau064http://www.ncbi.nlm.nih.gov/pubmed/25006225http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4086403.

[120] DAVID FERRUCCI and ADAM LALLY. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348, sep 2004. ISSN 1351-3249. doi:10.1017/S1351324904003523. URL http://www.journals.cambridge.org/abstract{_}S1351324904003523.

[121] Yoshinobu Kano, William A. Baumgartner, Luke McCrohon, Sophia Ananiadou, K. Bretonnel Cohen, Lawrence Hunter, and Jun'ichi Tsujii. U-Compare: share and compare text mining tools with UIMA. *Bioinformatics (Oxford, England)*, 25 (15):1997–8, aug 2009. ISSN 1367-4811. doi:10.1093/bioinformatics/btp289. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp289http://www.ncbi.nlm.nih.gov/pubmed/19414535http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2712335.

[122] Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. Beyond accuracy: creating interoperable and scalable text-mining web services. *Bioinformatics (Oxford, England)*, 32 (12):1907–10, jun 2016. ISSN 1367-4811. doi:10.1093/bioinformatics/btv760. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv760http://www.ncbi.nlm.nih.gov/pubmed/26883486http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4908316.

[123] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. Gate: an architecture for development of robust hlt applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 168, Morristown, NJ, USA, 2001. Association for Computational Linguistics. doi:10.3115/1073083.1073112. URL http://portal.acm.org/citation.cfm?doid=1073083.1073112.

[124] Ritu Khare, Chih-Hsuan Wei, Yuqing Mao, Robert Leaman, and Zhiyong Lu. tmBioC: improving interoperability of text-mining tools with BioC. *Database : the journal of biological databases and curation*, 2014(0):bau073–bau073, jul 2014. ISSN 1758-0463. doi:10.1093/database/bau073. URL https://academic.oup.com/database/article-lookup/doi/10.1093/database/bau073http://www.ncbi.nlm.nih.gov/pubmed/25062914http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4110697.

[125] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(Web Server issue): W518–22, jul 2013. ISSN 1362-4962. doi:10.1093/nar/gkt441. URL http://academic.oup.com/nar/article/41/W1/W518/1105731/PubTator-a-webbased-text-mining-tool-for-assistinghttp://www.ncbi.nlm.nih.gov/pubmed/23703206http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3692066.

[126] Wei Yu, Marta Gwinn, Melinda Clyne, Ajay Yesupriya, and Muin J Khoury. A navigator for human genome epidemiology. *Nature Genetics*, 40(2):124–125, feb 2008. doi:10.1038/ng0208-124. URL http://www.nature.com/articles/ng0208-124.

[127] Diane E. Oliver, Gaurav Bhalotia, Ariel S. Schwartz, Russ B. Altman, and Marti A. Hearst. Tools for loading MEDLINE into a local relational database. *BMC Bioinformatics*, 5(1):146, oct 2004. ISSN 14712105. doi:10.1186/1471-2105-5-146. URL http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-146.

[128] SimTK. Simtk: Medline parser - load xml medline data into rdbms: Project home. https://simtk.org/projects/medlineparser. (Accessed on 20/07/2016).

[129] Björn Grünning Kersten Döring, Kiran Telukunta. telukir/pubmedportable: Pubmed2go automatically builds up a postgresql relational database schema and a xapian full text index on medline/pubmed xml files. https://github.com/PhaBiFreiburg/PubMedPortable. (Accessed on 06/03/2018).

[130] Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics*, 7(Suppl 1 Text mining for chemistry and the CHEMDNER track):S3, jan 2015. ISSN 1758-2946. doi:10.1186/1758-2946-7-S1-S3. URL http://jcheminf.springeropen.com/articles/10.1186/1758-2946-7-S1-S3http://www.ncbi.nlm.nih.gov/pubmed/25810774http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4331693.

[131] Minlie Huang, Jingchen Liu, and Xiaoyan Zhu. GeneTUKit: a software for document-level gene normalization. *Bioinformatics (Oxford, England)*, 27(7): 1032–3, apr 2011. ISSN 1367-4811. doi:10.1093/bioinformatics/btr042. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr042http://www.ncbi.nlm.nih.gov/pubmed/21303863http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3065680.

[132] Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, Alexandra Tang, Geraldine Gabriel, Carol Ly, Sakina Adamjee, Zerihun T. Dame, Beomsoo Han, You Zhou, and David S. Wishart. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(Database issue):D1091–7, jan 2014. ISSN 1362-4962. doi:10.1093/nar/gkt1068. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1068http://www.ncbi.nlm.nih.gov/pubmed/24203711http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3965102.

[133] Alex Bateman, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, Benoit Bely, Mark Bingley, Carlos Bonilla, Ramona Britto, Borisas Bursteinas, Hema Bye-A-Jee, Andrew Cowley, Alan Da Silva, Maurizio De

Giorgi, Tunca Dogan, Francesco Fazzini, Leyla Garcia Castro, Luis Figueira, Penelope Garmiri, George Georghiou, Daniel Gonzalez, Emma Hatton-Ellis, Weizhong Li, Wudong Liu, Rodrigo Lopez, Jie Luo, Yvonne Lussi, Alistair MacDougall, Andrew Nightingale, Barbara Palka, Klemens Pichler, Diego Poggioli, Sangya Pundir, Luis Pureza, Guoying Qi, Alexandre Renaux, Steven Rosanoff, Rabie Saidi, Tony Sawford, Aleksandra Shypitsyna, Elena Speretta, Edward Turner, Nidhi Tyagi, Vladimir Volynkin, Tony Wardell, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Ioannis Xenarios, Lydie Bougueleret, Alan Bridge, Sylvain Poux, Nicole Redaschi, Lucila Aimo, Ghislaine Argoud-Puy, Andrea Auchincloss, Kristian Axelsen, Parit Bansal, Delphine Baratin, Marie-Claude Blatter, Brigitte Boeckmann, Jerven Bolleman, Emmanuel Boutet, Lionel Breuza, Cristina Casal-Casas, Edouard de Castro, Elisabeth Coudert, Beatrice Cuche, Mikael Doche, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Florence Jungo, Guillaume Keller, Vicente Lara, Philippe Lemercier, Damien Lieberherr, Thierry Lombardot, Xavier Martin, Patrick Masson, Anne Morgat, Teresa Neto, Nevila Nouspikel, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Monica Pozzato, Manuela Pruess, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne-Lise Veuthey, Cathy H Wu, Cecilia N Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, John S Garavelli, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A Natale, Karen Ross, C R Vinayaka, Qinghua Wang, Yuqi Wang, Lai-Su Yeh, and Jian Zhang. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, jan 2017. doi:10.1093/nar/gkw1099. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1099.

[134] Nisa M. Maruthur, Eva Tseng, Susan Hutfless, Lisa M. Wilson, Catalina Suarez-Cuervo, Zackary Berger, Yue Chu, Emmanuel Iyoha, Jodi B. Segal, and Shari Bolen. Diabetes Medications as Monotherapy or Metformin-Based Combination Therapy for Type 2 Diabetes. *Annals of Internal Medicine*, 164(11):740, jun 2016. doi:10.7326/M15-2650. URL http://annals.org/article.aspx?doi=10.7326/M15-2650.

[135] Jill P Crandall, William C Knowler, Steven E Kahn, David Marrero, Jose C Florez, George A Bray, Steven M Haffner, Mary Hoskin, David M Nathan, and Diabetes Prevention Program Research Group. The prevention of type 2 diabetes. *Nature clinical practice. Endocrinology & metabolism*, 4(7):382–93, jul 2008. ISSN 1745-8374. doi:10.1038/ncpendmet0843. URL http://www.ncbi.nlm.nih.gov/pubmed/18493227http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2573045.

[136] R S Hundal, M Krssak, S Dufour, D Laurent, V Lebon, V Chandramouli, S E Inzucchi, W C Schumann, K F Petersen, B R Landau, and G I Shulman. Mechanism by which metformin reduces glucose production in type 2 diabetes. *Diabetes*, 49(12):2063–9, dec 2000. ISSN 0012-1797. URL http://www.ncbi.nlm.nih.gov/pubmed/11118008http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2995498.

[137] Atanas G. Atanasov, Martina Blunder, Nanang Fakhrudin, Xin Liu, Stefan M. Noha, Clemens Malainer, Matthias P. Kramer, Amina Cocic, Olaf Kunert, Andreas Schinkovitz, Elke H. Heiss, Daniela Schuster, Verena M. Dirsch, and Rudolf Bauer. Polyacetylenes from Notopterygium incisum–New Selective Partial Agonists of Perox-

isome Proliferator-Activated Receptor-Gamma. *PLoS ONE*, 8(4):e61755, apr 2013. doi:10.1371/journal.pone.0061755. URL http://dx.plos.org/10.1371/journal.pone.0061755.

[138] Heba H. Mansour, Shereen M. El kiki, and Shereen M. Galal. Metformin and low dose radiation modulates cisplatin-induced oxidative injury in rat via PPAR-γ and MAPK pathways. *Archives of Biochemistry and Biophysics*, 616:13–19, feb 2017. doi:10.1016/J.ABB.2017.01.005. URL https://www.sciencedirect.com/science/article/pii/S0003986116305343?via{%}3Dihub.

[139] Vladimir Ljubicic and Bernard J. Jasmin. Metformin increases peroxisome proliferator-activated receptor γ Co-activator-1α and utrophin a expression in dystrophic skeletal muscle. *Muscle & Nerve*, 52(1):139–142, jul 2015. ISSN 0148639X. doi:10.1002/mus.24692. URL http://doi.wiley.com/10.1002/mus.24692.

[140] Stewart Bates. The role of gene expression profiling in drug discovery. *Current Opinion in Pharmacology*, 11(5):549–556, oct 2011. doi:10.1016/J.COPH.2011.06.009. URL https://www.sciencedirect.com/science/article/pii/S1471489211000828?via{%}3Dihub.

[141] Xavier Lucas, Christian Senger, Anika Erxleben, Björn A Grüning, Kersten Döring, Johannes Mosch, Stephan Flemming, and Stefan Günther. StreptomeDB: a resource for natural compounds isolated from Streptomyces species. *Nucleic acids research*, 41 (Database issue):D1130–6, jan 2013. ISSN 1362-4962. doi:10.1093/nar/gks1253. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks1253http://www.ncbi.nlm.nih.gov/pubmed/23193280http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3531085.

[142] Catherine Esnault, Thierry Dulermo, Aleksey Smirnov, Ahmed Askora, Michelle David, Ariane Deniset-Besseau, Ian-Barry Holland, and Marie-Joelle Virolle. Strong antibiotic production is correlated with highly active oxidative metabolism in Streptomyces coelicolor M145. *Scientific Reports*, 7(1):200, dec 2017. doi:10.1038/s41598-017-00259-9. URL http://www.nature.com/articles/s41598-017-00259-9.

[143] David A. Hopwood. Soil To Genomics: The Streptomyces Chromosome. *Annual Review of Genetics*, 40(1):1–23, dec 2006. ISSN 0066-4197. doi:10.1146/annurev.genet.40.110405.090639. URL http://www.ncbi.nlm.nih.gov/pubmed/16761950http://www.annualreviews.org/doi/10.1146/annurev.genet.40.110405.090639.

[144] Marco Ventura, Carlos Canchaya, Andreas Tauch, Govind Chandra, Gerald F Fitzgerald, Keith F Chater, and Douwe van Sinderen. Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiology and molecular biology reviews : MMBR*, 71(3):495–548, sep 2007. ISSN 1092-2172. doi:10.1128/MMBR.00005-07. URL http://www.ncbi.nlm.nih.gov/pubmed/17804669http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2168647.

[145] Hotam S Chaudhary, Jayprakash Yadav, Anju R Shrivastava, Smriti Singh, Anil K Singh, and Natrajan Gopalan. Antibacterial activity of actinomycetes isolated from different soil samples of Sheopur (A city of central India). *Journal of advanced pharmaceutical technology & research*, 4(2):118–23, apr 2013. ISSN 2231-4040. doi:10.4103/2231-4040.111528. URL http://www.ncbi.nlm.nih.gov/pubmed/23833752http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3696223.

[146] New York USA Schrödinger, LLC. Canvas 2.3. schrödinger, llc., new york, usa. https://www.schrodinger.com/canvas, 07 2017.

[147] Press Schrödinger. Canvas user manual. http://gohom.win/ManualHom/Schrodinger/Schrodinger_2015-2_docs/canvas/canvas_user_manual.pdf, .

[148] Edwin M. Ory and Ellard M. Yow. The Use and Abuse of the Broad Spectrum Antibiotics. *JAMA: The Journal of the American Medical Association*, 185(4):273, jul 1963. ISSN 0098-7484. doi:10.1001/jama.1963.03060040057022. URL http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.1963.03060040057022.

[149] Felicity Allen, Russ Greiner, and David Wishart. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, 11(1): 98–110, feb 2015. ISSN 1573-3882. doi:10.1007/s11306-014-0676-4. URL http://link.springer.com/10.1007/s11306-014-0676-4.

[150] Yanli Wang, Stephen H. Bryant, Tiejun Cheng, Jiyao Wang, Asta Gindulyte, Benjamin A. Shoemaker, Paul A. Thiessen, Siqian He, and Jian Zhang. PubChem BioAssay: 2017 update. *Nucleic Acids Research*, 45(Database issue):D955, 2017. doi:10.1093/NAR/GKW1118. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210581/.

[151] Miguel C. Leal, Ana Hilário, Murray H. G. Munro, John W. Blunt, and Ricardo Calado. Natural products discovery needs improved taxonomic and geographic information. *Natural product reports*, 33(6):747–50, jun 2016. ISSN 1460-4752. doi:10.1039/c5np00130g. URL http://xlink.rsc.org/?DOI=C5NP00130Ghttp://www.ncbi.nlm.nih.gov/pubmed/26892141.

[152] Joseph N Yong, Joseph N Yong, and Fidele Ntie-kang. ChemInform Abstract : The Chemistry and Biological Activities of Natural Products from Northern African Plant Families : From Ebenaceae to Solanaceae RSC Advances The chemistry and biological activities of natural products from Northern African plant fami. *RSC Advances*, 5 (June):26580–26595, 2015. ISSN 2046-2069. doi:10.1039/C4RA15377D. URL http://dx.doi.org/10.1039/C4RA15377D.

[153] Fidele Ntie-Kang, Leonel E Njume, Yvette I Malange, Stefan Günther, Wolfgang Sippl, and Joseph N Yong. The Chemistry and Biological Activities of Natural Products from Northern African Plant Families: From Taccaceae to Zygophyllaceae. *Natural products and bioprospecting*, 6(2):63–96, apr 2016. ISSN 2192-2195. doi:10.1007/s13659-016-0091-9. URL http://xlink.rsc.org/?DOI=C4RA11467Ahttp://www.ncbi.nlm.nih.gov/pubmed/26931529http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4805656.

[154] UN Stats. Unsd — methodology. https://unstats.un.org/unsd/methodology/m49/.

[155] American Geophysical Union. Sahara's abrupt desertification started by changes in earth's orbit, accelerated by atmospheric and vegetation feedbacks – sciencedaily. https://www.sciencedaily.com/releases/1999/07/990712080500.htm.

[156] Rafia Azmat, Saba Haider, Hajra Nasreen, Farha Aziz, and Marina Riaz. A VIABLE ALTERNATIVE MECHANISM IN ADAPTING THE PLANTS TO HEAVY METAL ENVIRONMENT. *Pak. J. Bot*, 41(6):2729–2738, 2009.

[157] Kyla Selvig and J. Andrew Alspaugh. pH Response Pathways in Fungi: Adapting to Host-derived and Environmental Signals. *Mycobiology*, 39(4):249–56, dec 2011. ISSN 2092-9323. doi:10.5941/MYCO.2011.39.4.249. URL http://synapse.koreamed.org/DOIx.php?id=10.5941/MYCO.2011.39.4.249http://www.ncbi.nlm.nih.gov/pubmed/22783112http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3385132.

[158] Fidele Ntie-Kang, James A Mbah, Luc Meva'a Mbaze, Lydia L Lifongo, Michael Scharfe, Joelle Ngo Hanna, Fidelis Cho-Ngwa, Pascal Amoa Onguéné, Luc C Owono Owono, Eugene Megnassan, Wolfgang Sippl, and Simon M N Efange. CamMedNP: building the Cameroonian 3D structural natural products database for virtual screening. *BMC complementary and alternative medicine*, 13(1):88, apr 2013. ISSN 1472-6882. doi:10.1186/1472-6882-13-88. URL http://bmccomplementalternmed.biomedcentral.com/articles/10.1186/1472-6882-13-88http://www.ncbi.nlm.nih.gov/pubmed/23590173http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3637470.

[159] Fidele Ntie-Kang, Lydia L Lifongo, James A Mbah, Luc C. Owono Owono, Eugene Megnassan, Luc Meva'a Mbaze, Philip N Judson, Wolfgang Sippl, and Simon M. N. Efange. In silico drug metabolism and pharmacokinetic profiles of natural products from medicinal plants in the Congo basin. *In silico pharmacology*, 1(1):12, nov 2013. ISSN 2193-9616. doi:10.1186/2193-9616-1-12. URL http://xlink.rsc.org/?DOI=C3RA43754Jhttp://www.ncbi.nlm.nih.gov/pubmed/25505657http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4230438.

[160] ChemBridge. Chembridge home. http://www.chembridge.com/, 05 1998. (Accessed On 4/1/2018 19:51).

[161] Fidele Ntie-Kang, Denis Zofou, Smith B. Babiaka, Rolande Meudom, Michael Scharfe, Lydia L. Lifongo, James A. Mbah, Luc Meva'a Mbaze, Wolfgang Sippl, and Simon M. N. Efange. AfroDb: a select highly potent and diverse natural product library from African medicinal plants. *PloS one*, 8(10):e78085, oct 2013. ISSN 1932-6203. doi:10.1371/journal.pone.0078085. URL http://dx.plos.org/10.1371/journal.pone.0078085http://www.ncbi.nlm.nih.gov/pubmed/24205103http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3813505.

[162] John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling*, 52(7):1757–68, jul 2012. ISSN 1549-960X. doi:10.1021/ci3001277. URL http://pubs.acs.org/doi/10.1021/ci3001277http://www.ncbi.nlm.nih.gov/pubmed/22587354http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3402020.

[163] Fidele Ntie-Kang, Pascal Amoa Onguéné, Ghislain W. Fotso, Kerstin Andrae-Marobela, Merhatibeb Bezabih, Jean Claude Ndom, Bonaventure T. Ngadjui, Abiodun O. Ogundaini, Berhanu M. Abegaz, and Luc Mbaze Meva'a. Virtualizing the p-ANAPL library: a step towards drug discovery from African medicinal plants. *PloS one*, 9(3):e90655, mar 2014. ISSN 1932-6203. doi:10.1371/journal.pone.0090655. URL http://dx.plos.org/10.1371/journal.pone.0090655http://www.ncbi.nlm.nih.gov/pubmed/24599120http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3944075.

[164] Prota4u. https://www.prota4u.org/. (Accessed on 05/14/2018).

[165] Tropicos MBG. Tropicos - home. http://www.tropicos.org/, May 2018. (Accessed on 05/14/2018).

[166] WoRMS Editorial Board. Worms - world register of marine species. http://www.marinespecies.org/. (Accessed on 05/14/2018).

[167] Eric W. Sayers, Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Ron Edgar, Scott Federhen, Michael Feolo, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J. Lipman, Thomas L. Madden, Donna R. Maglott, Vadim Miller, Ilene Mizrachi, James Ostell, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Eugene Yaschenko, and Jian Ye. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 37 (Database issue):D5–15, jan 2009. ISSN 1362-4962. doi:10.1093/nar/gkn741. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkn741http://www.ncbi.nlm.nih.gov/pubmed/18940862http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2686545.

[168] Vincent Robert, Duong Vu, Ammar Ben Hadj Amor, Nathalie van de Wiele, Carlo Brouwer, Bernard Jabas, Szaniszlo Szoke, Ahmed Dridi, Maher Triki, Samy Ben Daoud, Oussema Chouchen, Lea Vaas, Arthur de Cock, Joost A Stalpers, Dora Stalpers, Gerard J M Verkley, Marizeth Groenewald, Felipe Borges Dos Santos, Gerrit Stegehuis, Wei Li, Linhuan Wu, Run Zhang, Juncai Ma, Miaomiao Zhou, Sergio Pérez Gorjón, Lily Eurwilaichitr, Supawadee Ingsriswang, Karen Hansen, Conrad Schoch, Barbara Robbertse, Laszlo Irinyi, Wieland Meyer, Gianluigi Cardinali, David L. Hawksworth, John W. Taylor, and Pedro W. Crous. MycoBank gearing up for new horizons. *IMA fungus*, 4(2):371–9, dec 2013. ISSN 2210-6340. doi:10.5598/imafungus.2013.04.02.16. URL http://openurl.ingenta.com/content/xref?genre=article{&}issn=2210-6340{&}volume=4{&}issue=2{&}spage=371http://www.ncbi.nlm.nih.gov/pubmed/24563843http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3905949.

[169] G.; Stalpers J Robert, V.; Stegehuis. Mycobank database. http://www.mycobank.org/.

[170] G. Crous, P.W., Gams, W., Stalpers, J.A., Robert, V., Stegehuis. MYCOBANK: AN ONLINE INITIATIVE TO LAUNCH MYCOLOGY INTO THE 21ST CENTURY, 2004. URL https://www.narcis.nl/publication/RecordID/oai:pure.knaw.nl:publications{%}2F131e209d-bf03-4f24-acfa-07b360d167d0.

[171] NCBI. Genbank and wgs statistics. https://www.ncbi.nlm.nih.gov/genbank/statistics/, Apr 2018. (Accessed on 05/13/2018).

[172] Douglas E V Pires, Tom L Blundell, and David B Ascher. pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *Journal of medicinal chemistry*, 58(9):4066–72, may 2015. ISSN 1520-4804. doi:10.1021/acs.jmedchem.5b00104. URL http://pubs.acs.org/doi/abs/10.1021/acs.jmedchem.5b00104http://www.ncbi.nlm.nih.gov/pubmed/25860834http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4434528.

[173] Kristien Mortelmans and Errol Zeiger. The Ames Salmonella/microsome mutagenicity assay. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*,

455(1-2):29–60, nov 2000. ISSN 0027-5107. doi:10.1016/S0027-5107(00)00064-6. URL https://www.sciencedirect.com/science/article/pii/S0027510700000646?via{%}3Dihub.

[174] Yankang Jing, Alison Easter, David Peters, Norman Kim, and Istvan J Enyedy. In silico prediction of hERG inhibition. *Future Medicinal Chemistry*, 7(5):571–586, apr 2015. ISSN 1756-8919. doi:10.4155/fmc.15.18. URL http://www.future-science.com/doi/10.4155/fmc.15.18.

[175] Sebastian Greenhough and David C. Hay. Stem Cell-Based Toxicity Screening. *Pharmaceutical Medicine*, 26(2):85–89, apr 2012. ISSN 1178-2595. doi:10.1007/BF03256896. URL http://link.springer.com/10.1007/BF03256896.

[176] R. Gonella Diaza, S. Manganelli, A. Esposito, A. Roncaglioni, A. Manganaro, and E. Benfenati. Comparison of <i>in silico</i> tools for evaluating rat oral acute toxicity. *SAR and QSAR in Environmental Research*, 26(1):1–27, jan 2015. ISSN 1062-936X. doi:10.1080/1062936X.2014.977819. URL http://www.tandfonline.com/doi/abs/10.1080/1062936X.2014.977819.

[177] Jingzhuo Tian, Yan Yi, Yong Zhao, Chunying Li, Yushi Zhang, Lianmei Wang, Chen Pan, Jiayin Han, Guiqin Li, Xiaolong Li, Jing Liu, Nuo Deng, Yue Gao, and Aihua Liang. Oral chronic toxicity study of geniposide in rats. *Journal of Ethnopharmacology*, 213: 166–175, mar 2018. ISSN 03788741. doi:10.1016/j.jep.2017.11.008. URL http://www.ncbi.nlm.nih.gov/pubmed/29128573http://linkinghub.elsevier.com/retrieve/pii/S0378874117329549.

[178] Feixiong Cheng, Jie Shen, Yue Yu, Weihua Li, Guixia Liu, Philip W. Lee, and Yun Tang. In silico prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. *Chemosphere*, 82(11):1636–1643, mar 2011. doi:10.1016/J.CHEMOSPHERE.2010.11.043. URL https://www.sciencedirect.com/science/article/pii/S0045653510013500?via{%}3Dihub.

[179] Steven Broderius and Michael Kahl. Acute toxicity of organic chemical mixtures to the fathead minnow. *Aquatic Toxicology*, 6(4):307–322, jul 1985. ISSN 0166-445X. doi:10.1016/0166-445X(85)90026-8. URL https://www.sciencedirect.com/science/article/pii/0166445X85900268.

[180] Bhuwan B Mishra and Vinod K Tiwari. Natural products: an evolving role in future drug discovery. *European journal of medicinal chemistry*, 46(10):4769–807, oct 2011. ISSN 1768-3254. doi:10.1016/j.ejmech.2011.07.057. URL https://www.sciencedirect.com/science/article/pii/S0223523411005708?via{%}3Dihubhttp://www.ncbi.nlm.nih.gov/pubmed/21889825.

[181] Alan L. Harvey, RuAngelie Edrada-Ebel, and Ronald J. Quinn. The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery*, 14(2):111–129, feb 2015. ISSN 1474-1776. doi:10.1038/nrd4510. URL http://www.nature.com/articles/nrd4510.

[182] Tiago Rodrigues, Daniel Reker, Petra Schneider, and Gisbert Schneider. Counting on natural products for drug design. *Nature chemistry*, 8(6):531–41, jun 2016. ISSN 1755-4349. doi:10.1038/nchem.2479. URL http://dx.doi.org/10.1038/nrd4510http://www.ncbi.nlm.nih.gov/pubmed/27219696.

[183] Mahmoud S. Abdelbaset, Gamal El-Din A. Abuo-Rahma, Mostafa H. Abdelrahman, Mohamed Ramadan, Bahaa G.M. Youssif, Syed Nasir Abbas Bukhari, Mamdouh F.A. Mohamed, and Mohamed Abdel-Aziz. Novel pyrrol-2(3H)-ones and pyridazin-3(2H)-ones carrying quinoline scaffold as anti-proliferative tubulin polymerization inhibitors. *Bioorganic Chemistry*, 80:151–163, oct 2018. doi:10.1016/J.BIOORG.2018.06.003. URL https://www.sciencedirect.com/science/article/pii/S0045206818303651?via{%}3Dihub.

[184] Subhadeep Palit, Sanghamitra Mukherjee, Sougata Niyogi, Anindyajit Banerjee, Dipendu Patra, Amit Chakraborty, Saikat Chakrabarti, Partha Chakrabarti, and Sanjay Dutta. Quinoline-glycomimetic conjugates reducing lipogenesis and lipid accumulation in hepatocytes. *ChemBioChem*, jun 2018. doi:10.1002/cbic.201800271. URL http://doi.wiley.com/10.1002/cbic.201800271.

[185] Peng Teng, Chunhui Li, Zhong Peng, Vanderschouw Anne Marie, Alekhya Nimmagadda, Ma Su, Yaqiong Li, Xingmin Sun, and Jianfeng Cai. Facilely accessible quinoline derivatives as potent antibacterial agents. *Bioorganic & Medicinal Chemistry*, 26(12):3573–3579, jul 2018. doi:10.1016/J.BMC.2018.05.031. URL https://www.sciencedirect.com/science/article/pii/S0968089618307259?via{%}3Dihub.

[186] Chang You, Tingting Yuan, Yanzhen Huang, Chao Pi, Yangjie Wu, and Xiuling Cui. Rhodium-catalyzed regioselective C8-H amination of quinoline <i>N</i> -oxides with trifluoroacetamide at room temperature. *Organic & Biomolecular Chemistry*, 2018. doi:10.1039/C8OB01108G. URL http://xlink.rsc.org/?DOI=C8OB01108G.

[187] M. Ratheesh, G. Sindhu, and Antony Helen. Anti-inflammatory effect of quinoline alkaloid skimmianine isolated from Ruta graveolens L. *Inflammation research : official journal of the European Histamine Research Society ... [et al.]*, 62(4):367–76, apr 2013. ISSN 1420-908X. doi:10.1007/s00011-013-0588-1. URL http://link.springer.com/10.1007/s00011-013-0588-1http://www.ncbi.nlm.nih.gov/pubmed/23344232.

[188] Mona El-Neketi, Weaam Ebrahim, Wenhan Lin, Sahar Gedara, Farid Badria, Hassan-Elrady A. Saad, Daowan Lai, and Peter Proksch. Alkaloids and polyketides from Penicillium citrinum, an endophyte isolated from the Moroccan plant Ceratonia siliqua. *Journal of natural products*, 76(6):1099–104, jun 2013. ISSN 1520-6025. doi:10.1021/np4001366. URL http://pubs.acs.org/doi/10.1021/np4001366http://www.ncbi.nlm.nih.gov/pubmed/23713692.

[189] Fidele Ntie-Kang, Kiran K Telukunta, Kersten Döring, Conrad V Simoben, Aurélien F A Moumbock, Yvette I Malange, Leonel E Njume, Joseph N Yong, Wolfgang Sippl, and Stefan Günther. NANPDB: A Resource for Natural Products from Northern African Sources. *Journal of natural products*, 80(7):2067–2076, jul 2017. ISSN 1520-6025. doi:10.1021/acs.jnatprod.7b00283. URL http://www.ncbi.nlm.nih.gov/pubmed/28641017.

[190] Ashenafi Legehar, Henri Xhaard, and Leo Ghemtio. IDAAPM: integrated database of ADMET and adverse effects of predictive modeling based on FDA approved drug data. *Journal of Cheminformatics*, 8(1):33, 2016. ISSN 1758-2946. doi:10.1186/s13321-016-0141-7. URL http://jcheminf.springeropen.com/articles/10.1186/s13321-016-0141-7.

[191] Marilia Valli, Ricardo N. dos Santos, Leandro D. Figueira, Cíntia H. Nakajima, Ian Castro-Gamboa, Adriano D. Andricopulo, and Vanderlan S. Bolzani. Development

of a Natural Products Database from the Biodiversity of Brazil. *Journal of Natural Products*, 76(3):439–444, mar 2013. doi:10.1021/np3006875. URL http://pubs.acs.org/doi/10.1021/np3006875.

[192] Fidele Ntie-Kang, Pascal Amoa Onguéné, Michael Scharfe, Luc C. Owono Owono, Eugene Megnassan, Luc Meva'a Mbaze, Wolfgang Sippl, and Simon M. N. Efange. ConMedNP: a natural product library from Central African medicinal plants for drug discovery. *RSC Adv.*, 4(1):409–419, nov 2014. ISSN 2046-2069. doi:10.1039/C3RA43754J. URL http://xlink.rsc.org/?DOI=C3RA43754J.

[193] Arun Sharma, Prasun Dutta, Maneesh Sharma, Neeraj Kumar Rajput, Bhavna Dodiya, John J Georrge, Trupti Kholia, OSDD Consortium, and Anshu Bhardwaj. BioPhytMol: a drug discovery community resource on anti-mycobacterial phytomolecules and plant extracts. *Journal of cheminformatics*, 6(1):46, dec 2014. ISSN 1758-2946. doi:10.1186/s13321-014-0046-2. URL http://jcheminf.springeropen.com/articles/10.1186/s13321-014-0046-2http://www.ncbi.nlm.nih.gov/pubmed/25360160http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4206768.

[194] Manu Mangal, Parul Sagar, Harinder Singh, Gajendra P. S. Raghava, and Subhash M. Agarwal. NPACT: Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database. *Nucleic acids research*, 41(Database issue):D1124–9, jan 2013. ISSN 1362-4962. doi:10.1093/nar/gks1047. URL http://academic.oup.com/nar/article/41/D1/D1124/1052661/NPACT-Naturally-Occurring-Plantbased-Anticancerhttp://www.ncbi.nlm.nih.gov/pubmed/23203877http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3531140.

[195] PostgreSQL. Postgresql: The world's most advanced open source database. https://www.postgresql.org/.

[196] Greg Landrum. Rdkit: Open-source cheminformatics, . URL http://www.rdkit.org.

[197] Harry E. Pence and Antony Williams. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education*, 87(11):1123–1124, nov 2010. ISSN 0021-9584. doi:10.1021/ed100697w. URL http://pubs.acs.org/doi/abs/10.1021/ed100697w.

[198] ChemAxon. Marvinsketch 15.4.6.0; chemaxon: Cambridge, ma. http://www.chemaxon.com, 2015.

[199] Scifinder. Scifinder 2015; chemical abstracts service: Columbus, oh, 2015.

[200] Loren D. Mendelsohn*. ChemDraw 8 Ultra, Windows and Macintosh Versions. 2004. doi:10.1021/CI040123T. URL http://pubs.acs.org/doi/10.1021/ci040123t.

[201] Maestro, version 9.2; llc: New york, 2011.

[202] Schrödinger. Ligprep software, version 2.5; llc: New york, 2011.

[203] Tilmann Weber and Hyun Uk Kim. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synthetic and Systems Biotechnology*, 1(2):69–79, jun 2016. doi:10.1016/J.SYNBIO.2015.12.002. URL https://www.sciencedirect.com/science/article/pii/S2405805X15300156.

[204] S Anderson. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic acids research*, 9(13):3015–27, jul 1981. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pubmed/6269069http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC327328.

[205] David A. Hopwood. Preface. *Methods in Enzymology*, 459:xvii–xix, jan 2009. doi:10.1016/S0076-6879(09)04625-4. URL https://www.sciencedirect.com/science/article/pii/S0076687909046254?via{%}3Dihub.

[206] Paul F Zierep, Natàlia Padilla, Dimitar G Yonchev, Kiran K Telukunta, Dennis Klementz, and Stefan Günther. SeMPI: a genome-based secondary metabolite prediction and identification web server. *Nucleic acids research*, 45(W1):W64–W71, jul 2017. ISSN 1362-4962. doi:10.1093/nar/gkx289. URL http://www.ncbi.nlm.nih.gov/pubmed/28453782http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5570227.

[207] Gitanjali Yadav, Rajesh S. Gokhale, and Debasisa Mohanty. Towards Prediction of Metabolic Products of Polyketide Synthases: An In Silico Analysis. *PLoS Computational Biology*, 5(4):e1000351, apr 2009. doi:10.1371/journal.pcbi.1000351. URL http://dx.plos.org/10.1371/journal.pcbi.1000351.

[208] NHGRI. Dna sequencing costs: Data - national human genome research institute (nhgri). https://www.genome.gov/27541954/dna-sequencing-costs-data/, Apr 2018. (Accessed on 05/14/2018).

[209] Tilmann Weber, Kai Blin, Srikanth Duddela, Daniel Krug, Hyun Uk Kim, Robert Bruccoleri, Sang Yup Lee, Michael A. Fischbach, Rolf Müller, Wolfgang Wohlleben, Rainer Breitling, Eriko Takano, and Marnix H. Medema. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research*, 43(W1):W237–W243, jul 2015. doi:10.1093/nar/gkv437. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv437.

[210] Steven G Van Lanen and Ben Shen. Advances in polyketide synthase structure and function. *Current opinion in drug discovery & development*, 11(2):186–95, mar 2008. ISSN 1367-6733. URL http://www.ncbi.nlm.nih.gov/pubmed/18283606.

[211] Somnath Dutta, Jonathan R Whicher, Douglas A Hansen, Wendi A Hale, Joseph A Chemler, Grady R Congdon, Alison R H Narayan, Kristina Håkansson, David H Sherman, Janet L Smith, and Georgios Skiniotis. Structure of a modular polyketide synthase. *Nature*, 510(7506):512–7, jun 2014. ISSN 1476-4687. doi:10.1038/nature13423. URL http://www.ncbi.nlm.nih.gov/pubmed/24965652http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4278352.

[212] Björn A Grüning, Christian Senger, Anika Erxleben, Stephan Flemming, and Stefan Günther. Compounds In Literature (CIL): screening for compounds and relatives in PubMed. *Bioinformatics (Oxford, England)*, 27(9):1341–2, may 2011. ISSN 1367-4811. doi:10.1093/bioinformatics/btr130. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr130http://www.ncbi.nlm.nih.gov/pubmed/21414988.

[213] Apache. Multi-processing modules (mpms) - apache http server version 2.4. https://httpd.apache.org/docs/2.4/mpm.html.

[214] Frank Wiles. Performance tuning postgresql. https://www.revsys.com/writings/postgresql-performance.html.

[215] Marnix H Medema, Renzo Kottmann, Pelin Yilmaz, Matthew Cummings, John B Biggins, Kai Blin, Irene de Bruijn, Yit Heng Chooi, Jan Claesen, R Cameron Coates, Pablo Cruz-Morales, Srikanth Duddela, Stephanie Düsterhus, Daniel J Edwards, David P Fewer, Neha Garg, Christoph Geiger, Juan Pablo Gomez-Escribano, Anja Greule, Michalis Hadjithomas, Anthony S Haines, Eric J N Helfrich, Matthew L Hillwig, Keishi Ishida, Adam C Jones, Carla S Jones, Katrin Jungmann, Carsten Kegler, Hyun Uk Kim, Peter Kötter, Daniel Krug, Joleen Masschelein, Alexey V Melnik, Simone M Mantovani, Emily A Monroe, Marcus Moore, Nathan Moss, Hans-Wilhelm Nützmann, Guohui Pan, Amrita Pati, Daniel Petras, F Jerry Reen, Federico Rosconi, Zhe Rui, Zhenhua Tian, Nicholas J Tobias, Yuta Tsunematsu, Philipp Wiemann, Elizabeth Wyckoff, Xiaohui Yan, Grace Yim, Fengan Yu, Yunchang Xie, Bertrand Aigle, Alexander K Apel, Carl J Balibar, Emily P Balskus, Francisco Barona-Gómez, Andreas Bechthold, Helge B Bode, Rainer Borriss, Sean F Brady, Axel A Brakhage, Patrick Caffrey, Yi-Qiang Cheng, Jon Clardy, Russell J Cox, René De Mot, Stefano Donadio, Mohamed S Donia, Wilfred A van der Donk, Pieter C Dorrestein, Sean Doyle, Arnold J M Driessen, Monika Ehling-Schulz, Karl-Dieter Entian, Michael A Fischbach, Lena Gerwick, William H Gerwick, Harald Gross, Bertolt Gust, Christian Hertweck, Monica Höfte, Susan E Jensen, Jianhua Ju, Leonard Katz, Leonard Kaysser, Jonathan L Klassen, Nancy P Keller, Jan Kormanec, Oscar P Kuipers, Tomohisa Kuzuyama, Nikos C Kyrpides, Hyung-Jin Kwon, Sylvie Lautru, Rob Lavigne, Chia Y Lee, Bai Linquan, Xinyu Liu, Wen Liu, Andriy Luzhetskyy, Taifo Mahmud, Yvonne Mast, Carmen Méndez, Mikko Metsä-Ketelä, Jason Micklefield, Douglas A Mitchell, Bradley S Moore, Leonilde M Moreira, Rolf Müller, Brett A Neilan, Markus Nett, Jens Nielsen, Fergal O'Gara, Hideaki Oikawa, Anne Osbourn, Marcia S Osburne, Bohdan Ostash, Shelley M Payne, Jean-Luc Pernodet, Miroslav Petricek, Jörn Piel, Olivier Ploux, Jos M Raaijmakers, José A Salas, Esther K Schmitt, Barry Scott, Ryan F Seipke, Ben Shen, David H Sherman, Kaarina Sivonen, Michael J Smanski, Margherita Sosio, Evi Stegmann, Roderich D Süssmuth, Kapil Tahlan, Christopher M Thomas, Yi Tang, Andrew W Truman, Muriel Viaud, Jonathan D Walton, Christopher T Walsh, Tilmann Weber, Gilles P van Wezel, Barrie Wilkinson, Joanne M Willey, Wolfgang Wohlleben, Gerard D Wright, Nadine Ziemert, Changsheng Zhang, Sergey B Zotchev, Rainer Breitling, Eriko Takano, and Frank Oliver Glöckner. Minimum Information about a Biosynthetic Gene cluster. *Nature chemical biology*, 11(9):625–31, sep 2015. ISSN 1552-4469. doi:10.1038/nchembio.1890. URL http://www.ncbi.nlm.nih.gov/pubmed/26284661http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5714517.

[216] R.William Broadhurst, Daniel Nietlispach, Michael P Wheatcroft, Peter F Leadlay, and Kira J Weissman. The Structure of Docking Domains in Modular Polyketide Synthases. *Chemistry & Biology*, 10(8):723–731, aug 2003. doi:10.1016/S1074-5521(03)00156-X. URL https://www.sciencedirect.com/science/article/pii/S107455210300156X?via{%}3Dihub.

[217] Xavier Lucas, Daniel Wohlwend, Martin Hügle, Karin Schmidtkunz, Stefan Gerhardt, Rol Schüle, Manfred Jung, Oliver Einsle, and Stefan Günther. 4-Acyl pyrroles: Mimicking acetylated lysines in histone code reading. *Angewandte Chemie - International Edition*, 52(52):14055–14059, dec 2013. ISSN 14337851. doi:10.1002/anie.201307652. URL http://www.ncbi.nlm.nih.gov/pubmed/24272870.

[218] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P Overington. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(Database issue):D1100–7, jan 2012. ISSN 1362-4962. doi:10.1093/nar/gkr777. URL http://www.ncbi.nlm.nih.gov/pubmed/21948594http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3245175.

[219] Achilles Ntranos and Patrizia Casaccia. Bromodomains: Translating the words of lysine acetylation into myelin injury and repair. *Neuroscience Letters*, 625:4–10, jun 2016. ISSN 0304-3940. doi:10.1016/J.NEULET.2015.10.015. URL https://www.sciencedirect.com/science/article/pii/S0304394015301786?via{%}3Dihub.

[220] R.K. Prinjha, J. Witherington, and K. Lee. Place your BETs: the therapeutic potential of bromodomains. *Trends in Pharmacological Sciences*, 33(3):146–153, mar 2012. ISSN 0165-6147. doi:10.1016/J.TIPS.2011.12.002. URL https://www.sciencedirect.com/science/article/pii/S0165614711002227?via{%}3Dihub.

[221] Structural Genomics Consortium. Sgc | annotated phylogenetic trees - ligands. http://apps.thesgc.org/resources/phylogenetic_trees/ligands.php?target=BRD4&domain=BROMO.

[222] Xiangyang Li, Jian Zhang, Leilei Zhao, Yifei Yang, Huibin Zhang, and Jinpei Zhou. Design, Synthesis, and in vitro Biological Evaluation of 3,5-Dimethylisoxazole Derivatives as BRD4 Inhibitors. *ChemMedChem*, may 2018. doi:10.1002/cmdc.201800074. URL http://doi.wiley.com/10.1002/cmdc.201800074.

[223] Leilei Zhao, Yifei Yang, Yahui Guo, Lingyun Yang, Jian Zhang, Jinpei Zhou, and Huibin Zhang. Design, synthesis and biological evaluation of 7-methylimidazo[1,5-a]pyrazin-8(7H)-one derivatives as BRD4 inhibitors. *Bioorganic & Medicinal Chemistry*, 25 (8):2482–2490, apr 2017. ISSN 0968-0896. doi:10.1016/J.BMC.2017.03.008. URL https://www.sciencedirect.com/science/article/pii/S0968089617300111?via{%}3Dihub.

[224] S B Shuker, P J Hajduk, R P Meadows, and S W Fesik. Discovering high-affinity ligands for proteins: SAR by NMR. *Science (New York, N.Y.)*, 274(5292):1531–4, nov 1996. ISSN 0036-8075. doi:10.1126/SCIENCE.274.5292.1531. URL http://www.ncbi.nlm.nih.gov/pubmed/8929414.

[225] G C Terstappen and A Reggiani. In silico research in drug discovery. *Trends in pharmacological sciences*, 22(1):23–6, jan 2001. ISSN 0165-6147. doi:10.1016/S0165-6147(00)01584-4. URL http://www.ncbi.nlm.nih.gov/pubmed/11165668.

[226] Martin Vogt, Dagmar Stumpfe, Hanna Geppert, and Jürgen Bajorath. Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening. *Journal of medicinal chemistry*, 53(15):5707–15, aug 2010. ISSN 1520-4804. doi:10.1021/jm100492z. URL http://pubs.acs.org/doi/abs/10.1021/jm100492zhttp://www.ncbi.nlm.nih.gov/pubmed/20684607.

[227] Steffen Renner and Gisbert Schneider. Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem*, 1(2):181–5, feb 2006. ISSN 1860-7179. doi:10.1002/cmdc.200500005. URL http://doi.wiley.com/10.1002/cmdc.200500005http://www.ncbi.nlm.nih.gov/pubmed/16892349.

[228] Nathan Brown and Edgar Jacoby. On scaffolds and hopping in medicinal chemistry. *Mini reviews in medicinal chemistry*, 6(11):1217–29, nov 2006. ISSN 1389-5575. doi:10.2174/138955706778742768. URL http://www.eurekaselect.com/openurl/content.php?genre=article{&}issn=1389-5575{&}volume=6{&}issue=11{&}spage=1217http://www.ncbi.nlm.nih.gov/pubmed/17100633.

[229] Peter Willett, Vivienne Winterman, and David Bawden. Implementation of nearest-neighbor searching in an online chemical structure search system. *Journal of Chemical Information and Modeling*, 26(1):36–41, feb 1986. ISSN 1549-9596. doi:10.1021/ci00049a008. URL http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00049a008.

[230] Chunquan Sheng and Wannian Zhang. Fragment Informatics and Computational Fragment-Based Drug Design: An Overview and Update. *Medicinal Research Reviews*, 33(3):554–598, may 2013. doi:10.1002/med.21255. URL http://doi.wiley.com/10.1002/med.21255.

[231] Philip J. Hajduk and Jonathan Greer. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature reviews. Drug discovery*, 6(3):211–9, mar 2007. ISSN 1474-1776. doi:10.1038/nrd2220. URL http://www.nature.com/doifinder/10.1038/nrd2220http://www.ncbi.nlm.nih.gov/pubmed/17290284.

[232] Tobias Girschick, Lucia Puchbauer, and Stefan Kramer. Improving structural similarity based virtual screening using background knowledge. *Journal of cheminformatics*, 5(1):50, dec 2013. ISSN 1758-2946. doi:10.1186/1758-2946-5-50. URL http://jcheminf.springeropen.com/articles/10.1186/1758-2946-5-50http://www.ncbi.nlm.nih.gov/pubmed/24341870http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3928642.

[233] Michael J Keiser, Bryan L Roth, Blaine N Armbruster, Paul Ernsberger, John J Irwin, and Brian K Shoichet. Relating protein pharmacology by ligand chemistry. *Nature Biotechnology*, 25(2):197–206, feb 2007. doi:10.1038/nbt1284. URL http://www.nature.com/articles/nbt1284.

[234] Jongsoo Keum, Sunyong Yoo, Doheon Lee, and Hojung Nam. Prediction of compound-target interactions of natural products using large-scale drug and protein information. *BMC bioinformatics*, 17 Suppl 6(Suppl 6):219, jul 2016. ISSN 1471-2105. doi:10.1186/s12859-016-1081-y. URL http://www.ncbi.nlm.nih.gov/pubmed/27490208http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4965709.

[235] Anna Gaulton, Anne Hersey, Michał Nowotka, A Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P Overington, George Papadatos, Ines Smit, and Andrew R Leach. The ChEMBL database in 2017. *Nucleic acids research*, 45(D1):D945–D954, jan 2017. ISSN 1362-4962. doi:10.1093/nar/gkw1074. URL http://www.ncbi.nlm.nih.gov/pubmed/27899562http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5210557.

[236] A Patrícia Bento, Anna Gaulton, Anne Hersey, Louisa J Bellis, Jon Chambers, Mark Davies, Felix A Krüger, Yvonne Light, Lora Mak, Shaun McGlinchey, Michal Nowotka, George Papadatos, Rita Santos, and John P Overington. The ChEMBL bioactivity database: an update. *Nucleic acids research*, 42(Database issue):D1083–90, jan 2014.

ISSN 1362-4962. doi:10.1093/nar/gkt1031. URL http://www.ncbi.nlm.nih.gov/pubmed/24214965http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3965067.

[237] Saminda Abeyruwan, Uma D Vempati, Hande Küçük-McGinty, Ubbo Visser, Amar Koleti, Ahsan Mir, Kunie Sakurai, Caty Chung, Joshua A Bittker, Paul A Clemons, Steve Brudz, Anosha Siripala, Arturo J Morales, Martin Romacker, David Twomey, Svetlana Bureeva, Vance Lemmon, and Stephan C Schürer. Evolving BioAssay Ontology (BAO): modularization, integration and applications. *Journal of biomedical semantics*, 5 (Suppl 1 Proceedings of the Bio-Ontologies Spec Interest G):S5, 2014. ISSN 2041-1480. doi:10.1186/2041-1480-5-S1-S5. URL http://www.ncbi.nlm.nih.gov/pubmed/25093074http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4108877.

[238] Stuart J Nelson, Tudor I Oprea, Oleg Ursu, Cristian G Bologa, Amrapali Zaveri, Jayme Holmes, Jeremy J Yang, Stephen L Mathias, Subramani Mani, Mark S Tuttle, and Michel Dumontier. Formalizing drug indications on the road to therapeutic intent. *Journal of the American Medical Informatics Association*, 24(6):1169–1172, nov 2017. doi:10.1093/jamia/ocx064. URL http://academic.oup.com/jamia/article/24/6/1169/3977928/Formalizing-drug-indications-on-the-road-to.

[239] Melanie Brazil. High affinity good, lower affinity better. *Nature Reviews Drug Discovery*, 1(10):746–746, oct 2002. ISSN 1474-1776. doi:10.1038/nrd923. URL http://www.nature.com/articles/nrd923.

[240] Erik Kangas and Bruce Tidor. Optimizing electrostatic affinity in ligand–receptor binding: Theory, computation, and ligand properties. *The Journal of Chemical Physics*, 109(17):7522, oct 1998. ISSN 0021-9606. doi:10.1063/1.477375. URL https://aip.scitation.org/doi/abs/10.1063/1.477375.

[241] Dario Ghersi and Mona Singh. molBLOCKS: decomposing small molecule sets and uncovering enriched fragments. *Bioinformatics (Oxford, England)*, 30(14):2081–3, jul 2014. ISSN 1367-4811. doi:10.1093/bioinformatics/btu173. URL http://www.ncbi.nlm.nih.gov/pubmed/24681908http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4080744.

[242] B.Gruening TJ O'Donnell. galaxytools/fragmenter.py at master branch bgruening/galaxytools. https://github.com/bgruening/galaxytools/blob/master/chemicaltoolbox/fragmenter/fragmenter.py.

[243] Dario Ghersi. molblocksguide.pdf. http://compbio.cs.princeton.edu/molblocks/molblocksguide.pdf, Jan 2014.

[244] Coen Bron and Joep Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577, 1973. ISSN 0001-0782. doi:10.1145/362342.362367. URL https://dl.acm.org/citation.cfm?id=362367.

[245] Björn A Grüning. *Integrierte bioinformatische Methoden zur reproduzierbaren und transparenten Hochdurchsatz-Analyse von Life Science Big Data*. PhD thesis, Albert-Ludwigs-Universit{\"a}t Freiburg im Breisgau, 2015.

[246] SciPy ecosystem. scipy.stats.fisher_exact — scipy v0.14.0 reference guide. https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.fisher_exact.html.

[247] StatsModels. statsmodels.stats.multitest — statsmodels 0.9.0 documentation. http://www.statsmodels.org/dev/_modules/statsmodels/stats/multitest.html.

[248] Michael C. Sanguinetti and Martin Tristani-Firouzi. hERG potassium channels and cardiac arrhythmia. *Nature*, 440(7083):463–469, mar 2006. ISSN 0028-0836. doi:10.1038/nature04710. URL http://www.nature.com/articles/nature04710.

[249] A J Warner, J Lopez-Dee, E L Knight, J R Feramisco, and S A Prigent. The Shc-related adaptor protein, Sck, forms a complex with the vascular-endothelial-growth-factor receptor KDR in transfected cells. *The Biochemical journal*, 347(Pt 2):501–9, apr 2000. ISSN 0264-6021. URL http://www.ncbi.nlm.nih.gov/pubmed/10749680http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1220983.

[250] D T Connolly, D M Heuvelman, R Nelson, J V Olander, B L Eppley, J J Delfino, N R Siegel, R M Leimgruber, and J Feder. Tumor vascular permeability factor stimulates endothelial cell growth and angiogenesis. *The Journal of clinical investigation*, 84(5):1470–8, nov 1989. ISSN 0021-9738. doi:10.1172/JCI114322. URL http://www.jci.org/articles/view/114322http://www.ncbi.nlm.nih.gov/pubmed/2478587http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC304011.

[251] Lloyd Paul Aiello, Robert L. Avery, Paul G. Arrigg, Bruce A. Keyt, Henry D. Jampel, Sabera T. Shah, Louis R. Pasquale, Hagen Thieme, Mami A. Iwamoto, John E. Park, Hung V. Nguyen, Lloyd M. Aiello, Napoleone Ferrara, and George L. King. Vascular Endothelial Growth Factor in Ocular Fluid of Patients with Diabetic Retinopathy and Other Retinal Disorders. *New England Journal of Medicine*, 331(22):1480–1487, dec 1994. doi:10.1056/NEJM199412013312203. URL http://www.nejm.org/doi/abs/10.1056/NEJM199412013312203.

[252] Casper Hempel, Nils Hoyer, Trine Staalsø, and Jørgen A Kurtzhals. Effects of the vascular endothelial growth factor receptor-2 (VEGFR-2) inhibitor SU5416 on in vitro cultures of Plasmodium falciparum. *Malaria journal*, 13:201, may 2014. ISSN 1475-2875. doi:10.1186/1475-2875-13-201. URL http://www.ncbi.nlm.nih.gov/pubmed/24885283http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4046387.

[253] Zhijie Cheng, Denise Garvin, Aileen Paguio, Pete Stecha, Keith Wood, and Frank Fan. Luciferase Reporter Assay System for Deciphering GPCR Pathways. *Current chemical genomics*, 4:84–91, dec 2010. ISSN 1875-3973. doi:10.2174/1875397301004010084. URL http://www.ncbi.nlm.nih.gov/pubmed/21331312http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3040460.

[254] B Trzaskowski, D Latek, S Yuan, U Ghoshdastider, A Debinski, and S Filipek. Action of molecular switches in GPCRs–theoretical and experimental studies. *Current medicinal chemistry*, 19(8):1090–109, 2012. ISSN 1875-533X. doi:10.2174/092986712799320556. URL http://www.ncbi.nlm.nih.gov/pubmed/22300046http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3343417.

[255] Bertha K. Madras. History of the Discovery of the Antipsychotic Dopamine D2 Receptor: A Basis for the Dopamine Hypothesis of Schizophrenia. *Journal of the History of the Neurosciences*, 22(1):62–78, jan 2013. doi:10.1080/0964704X.2012.678199. URL http://www.tandfonline.com/doi/abs/10.1080/0964704X.2012.678199.

[256] Hisashi HASHIMOTO, Kenji TOIDE, Ryuji KITAMURA, Masako FUJITA, Sanae TAGAWA, Susumu ITOH, and Tetsuya KAMATAKI. Gene structure of CYP3A4, an adult-specific form of cytochrome P450 in human livers, and its transcriptional control. *European Journal of Biochemistry*, 218(2):585–595, dec 1993. doi:10.1111/j.1432-1033.1993.tb18412.x. URL http://doi.wiley.com/10.1111/j.1432-1033.1993.tb18412.x.

[257] Kiyoshi Inoue, Johji Inazawa, Hitoshi Nakagawa, Tsutomu Shimada, Hiroshi Yamazaki, F. Peter Guengerich, and Tatsuo Abe. Assignment of the human cytochrome P-450 nifedipine oxidase gene (CYP3A4) to chromosome 7 at band q22.1 by fluorescencein situ hybridization. *The Japanese Journal of Human Genetics*, 37(2):133–138, jun 1992. ISSN 0916-8478. doi:10.1007/BF01899734. URL http://www.ncbi.nlm.nih.gov/pubmed/1391968http://www.nature.com/doifinder/10.1007/BF01899734.

[258] Ulrich M. Zanger and Matthias Schwab. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacology & Therapeutics*, 138(1):103–141, apr 2013. ISSN 0163-7258. doi:10.1016/J.PHARMTHERA.2012.12.007. URL https://www.sciencedirect.com/science/article/pii/S0163725813000065?via{%}3Dihub.

[259] Sudha Ponnusamy, Saikat Haldar, Fayaj Mulani, Smita Zinjarde, Hirekodathakallu Thulasiram, and Ameeta RaviKumar. Gedunin and Azadiradione: Human Pancreatic Alpha-Amylase Inhibiting Limonoids from Neem (Azadirachta indica) as Anti-Diabetic Agents. *PLOS ONE*, 10(10):e0140113, oct 2015. ISSN 1932-6203. doi:10.1371/journal.pone.0140113. URL http://dx.plos.org/10.1371/journal.pone.0140113.

[260] Tushar Madaan, Mohd. Akhtar, and Abul Kalam Najmi. Sodium glucose CoTransporter 2 (SGLT2) inhibitors: Current status and future perspective. *European Journal of Pharmaceutical Sciences*, 93:244–252, oct 2016. doi:10.1016/J.EJPS.2016.08.025. URL https://www.sciencedirect.com/science/article/pii/S0928098716303141.

[261] Stephanie Ross, Hertzel Gerstein, and Guillaume Paré. The Genetic Link Between Diabetes and Atherosclerosis. *Canadian Journal of Cardiology*, 34(5):565–574, may 2018. doi:10.1016/J.CJCA.2018.01.016. URL https://www.sciencedirect.com/science/article/pii/S0828282X18300199?via{%}3Dihub.

[262] Tina Buchholz and Matthias F. Melzig. Medicinal Plants Traditionally Used for Treatment of Obesity and Diabetes Mellitus - Screening for Pancreatic Lipase and α-Amylase Inhibition. *Phytotherapy Research*, 30(2):260–266, feb 2016. ISSN 0951418X. doi:10.1002/ptr.5525. URL http://doi.wiley.com/10.1002/ptr.5525.

[263] Haralambos Tzoupis, Georgios Leonis, Grigorios Megariotis, Claudiu T. Supuran, Thomas Mavromoustakos, and Manthos G. Papadopoulos. Dual Inhibitors for Aspartic Proteases HIV-1 PR and Renin: Advancements in AIDS–Hypertension–Diabetes Linkage via Molecular Dynamics, Inhibition Assays, and Binding Free Energy Calculations. *Journal of Medicinal Chemistry*, 55(12):5784–5796, jun 2012. ISSN 0022-2623. doi:10.1021/jm300180r. URL http://pubs.acs.org/doi/10.1021/jm300180r.

[264] István Takács, Ákos Takács, Anikó Pósa, and Gyöngyi Gyémánt. HPLC method for measurement of human salivary α-amylase inhibition by aqueous plant extracts. *Acta Biologica Hungarica*, 68(2):127–136, jun 2017. ISSN 0236-5383.

doi:10.1556/018.68.2017.2.1. URL http://www.akademiai.com/doi/abs/10.1556/018.68.2017.2.1.

[265] Apache lucene - welcome to apache lucene. https://lucene.apache.org/.

[266] Barbara Zdrazil and Rajarshi Guha. The Rise and Fall of a Scaffold: A Trend Analysis of Scaffolds in the Medicinal Chemistry Literature. *Journal of Medicinal Chemistry*, page acs.jmedchem.7b00954, dec 2017. doi:10.1021/acs.jmedchem.7b00954. URL http://pubs.acs.org/doi/10.1021/acs.jmedchem.7b00954.

[267] Dilyana Dimova, Dagmar Stumpfe, and Jürgen Bajorath. Computational design of new molecular scaffolds for medicinal chemistry, part II: generalization of analog series-based scaffolds. *Future science OA*, 4(2):FSO267, feb 2018. ISSN 2056-5623. doi:10.4155/fsoa-2017-0102. URL http://www.ncbi.nlm.nih.gov/pubmed/29379641http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5778380.

[268] Paul Oguadinma, Francois Bilodeau, and Steven R. LaPlante. NMR strategies to support medicinal chemistry workflows for primary structure determination. *Bioorganic & Medicinal Chemistry Letters*, 27(2):242–247, jan 2017. doi:10.1016/J.BMCL.2016.11.066. URL https://www.sciencedirect.com/science/article/pii/S0960894X16312197?via{%}3Dihub.

[269] ChEMBL RDF model. Chembl < rdf platform < embl-ebi. URL https://www.ebi.ac.uk/rdf/documentation/chembl/.

[270] Christopher T Walsh, Robert V O'Brien, and Chaitan Khosla. Nonproteinogenic amino acid building blocks for nonribosomal peptide and hybrid polyketide scaffolds. *Angewandte Chemie (International ed. in English)*, 52(28):7098–124, jul 2013. ISSN 1521-3773. doi:10.1002/anie.201208344. URL http://www.ncbi.nlm.nih.gov/pubmed/23729217http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4634941.

[271] Sunil S. Chandran, Hugo G. Menzella, John R. Carney, and Daniel V. Santi. Activating Hybrid Modular Interfaces in Synthetic Polyketide Synthases by Cassette Replacement of Ketosynthase Domains. *Chemistry & Biology*, 13(5):469–474, may 2006. ISSN 1074-5521. doi:10.1016/J.CHEMBIOL.2006.02.011. URL https://www.sciencedirect.com/science/article/pii/S1074552106000810.

[272] Schrödinger. Docking and scoring | schrödinger, . URL https://www.schrodinger.com/science-articles/docking-and-scoring.

[273] Darko Butina. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. 1999. doi:10.1021/CI9803381. URL https://pubs.acs.org/doi/abs/10.1021/ci9803381.

[274] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, mar 1982. ISSN 0018-9448. doi:10.1109/TIT.1982.1056489. URL http://ieeexplore.ieee.org/document/1056489/.

[275] Shaaban A Mousa, Mohammed Shaqura, Baled I Khalefa, Christian Zöllner, Laura Schaad, Jonas Schneider, Toni S Shippenberg, Jan F Richter, Rainer Hellweg, Mehdi Shakibaei, and Michael Schäfer. Rab7 silencing prevents μ-opioid receptor lysosomal

targeting and rescues opioid responsiveness to strengthen diabetic neuropathic pain therapy. *Diabetes*, 62(4):1308–19, apr 2013. ISSN 1939-327X. doi:10.2337/db12-0590. URL http://www.ncbi.nlm.nih.gov/pubmed/23230081http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3609597.

[276] Pamela W Anderson, Janet B Mcgill, and Katherine R Tuttle. Protein kinase C b inhibition: the promise for treatment of diabetic nephropathy. *Curr Opin Nephrol Hypertens Current Opinion in Nephrology and Hypertension*, 16(16), 2007. URL https://insights.ovid.com/crossref?an=00041552-200709000-00002.

[277] Qiu-yan Wang, Ying Zhang, Hai-jiao Chen, Zong-hou Shen, and Hui-li Chen. Alpha 1,3-fucosyltransferase-VII regulates the signaling molecules of the insulin receptor pathway. *FEBS Journal*, 274(2):526–538, jan 2007. ISSN 1742464X. doi:10.1111/j.1742-4658.2006.05599.x. URL http://www.ncbi.nlm.nih.gov/pubmed/17229154http://doi.wiley.com/10.1111/j.1742-4658.2006.05599.x.

[278] Joshua J Joseph, Justin B Echouffo-Tcheugui, Rita R Kalyani, Hsin-Chieh Yeh, Alain G Bertoni, Valery S Effoe, Ramon Casanova, Mario Sims, Adolfo Correa, Wen-Chih Wu, Gary S Wand, and Sherita H Golden. Aldosterone, Renin, and Diabetes Mellitus in African Americans: The Jackson Heart Study. *The Journal of clinical endocrinology and metabolism*, 101(4):1770–8, 2016. ISSN 1945-7197. doi:10.1210/jc.2016-1002. URL http://www.ncbi.nlm.nih.gov/pubmed/26908112http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4880170.

[279] Hyeongjin Cho. Protein Tyrosine Phosphatase 1B (PTP1B) and Obesity. *Vitamins & Hormones*, 91:405–424, jan 2013. doi:10.1016/B978-0-12-407766-9.00017-1. URL https://www.sciencedirect.com/science/article/pii/B9780124077669000171?via{%}3Dihub.