

# **Molecular similarity and diversity analysis of bioactive small molecules using chemoinformatics approaches**

**Varun Khanna**

**A thesis submitted to Macquarie University  
in fulfilment of the degree of  
Doctor of Philosophy**

**Department of Chemistry and Biomolecular Sciences  
Macquarie University  
Sydney, Australia**

**March 2011**







**DEDICATED TO**  
**ALL THE PIONEERING SCIENTISTS OF ARYAVARTA (ANCIENT INDIA),**  
**WHO GAVE THE KNOWLEDGE OF VEDIC MATHEMATICS ,**  
**YOGA AND AYURVEDA TO THE WORLD**



## **DECLARATION**

This thesis contains original work, which was performed by me. Several aspects of this work have been carried out with the help and guidance of many researchers; these people have been acknowledged and their contributions recognised in the section in which their assistance was received. This thesis contains no material that has been accepted for the award of any higher degree or diploma at any University or Institution, and to the best of my knowledge, contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Varun Khanna

March 2011



## ACKNOWLEDGEMENTS

Working at a scientific problem can be fun, stimulating and great. However, at times it can also be tough, time consuming and persevering and in the tough times one often needs help. It is my pleasure to thank many people who extended their support to me in difficult times and made this thesis possible.

### Professional

- ❖ My supervisor, Prof. Shoba Ranganathan, for her constant support and invaluable suggestions during this work. I thank her for giving me the freedom to follow up on interesting topics in chemoinformatics but also keeping me on track. I am very grateful for her patience, motivation, never ending enthusiasm, intellectual support and especially for encouraging me to visit Strasbourg Summer School on Chemoinformatics, Obernai very early in my PhD to participate and interact with leading scientists in the area of computational chemistry. I would also like to thank her for educating me on time management in academic research.
- ❖ Prof. Joanne Jamie, my co-supervisor for the support throughout my PhD tenure, for the motivation and useful discussions during meetings. I also thank her for providing the excellent scientific and social environment at the department.
- ❖ Dr. Dominique Gorse, for giving an opportunity to visit QFAB and helping me out with the final manuscript. I also thank him for suggestions on post PhD plans.
- ❖ I am grateful to my past and current lab members Dr. Durgaprasad Bollina, Dr. Shivashankar Hiriyur Nagaraj, Javed Khan, Dr. Jitendra Gaikwad, Dr. Adrian Cootes, Gagan Garg, , Gaurav Kumar, Elsa Chacko, Ranjeeta Menon and FoS staff members Michael Baxter, Chris McRae (CBMS), Maria Hyland (CBMS) Catherine Wong (CBMS) for their support.

### Personal

- ❖ My deepest gratitude goes to my parents and all other family members for their love, care and encouragement when I was miles away from home. Without their support it would not been possible to reach this important milestone in my research career.



- ❖ I am deeply indebted to my dear younger brother Dheeraj Khanna for sharing wonderful times and emotional support.
- ❖ I would like to thank all my following friends for making these three years a memorable experience with their good company and for their moral support: Javed Iqbal Khan, Sosuke (gym buddy), Gulshan Kumar, Long Vu (table tennis pal), Annaf, John (for gym training), Sudhir, Vikas Dhiya (India), Bhupinder Jangara (India), Aneesh, Carlos, Rajesh Mudugal (Adelaide), Faraaz, Jens Moll, Lucky Tur, Nandan Deshpanday, Orisi, Pankaj, Ruksana Master.
- ❖ My hearty thanks to my best friend, Sumit Kalra, India. I got the opportunity to know him in B.Sc. and I thank him for the wonderful times we shared since then. I will especially never forget the train journeys which we took together.
- ❖ Thanks to another great friend Noge Pham, Australia who has always enjoyed all my publications and work even though he never understood anything about it.
- ❖ I would also like to thank all my science teachers especially, Ruby madam (4<sup>th</sup> and 5<sup>th</sup> standard), Mona madam (6<sup>th</sup> - 8<sup>th</sup> standard), Dr. Rajpal Yadav (9<sup>th</sup> and 10<sup>th</sup> standard), Dr. Monga (Graduation, Botany), Dr. Dharam Singh Chopra (Graduation, Biochemistry) and Dr. Sudhir Kumar (Masters, Bioinformatics) for helping me maintain interest in science and encouraging me to take science as a career.



## TABLE OF CONTENTS

<i>Declaration</i>		<i>ii</i>
<i>Acknowledgements</i>		<i>iii</i>
<i>Table of Contents</i>		<i>v</i>
<i>List of Abbreviations</i>		<i>viii</i>
<i>List of Figures</i>		<i>ix</i>
<i>List of Tables</i>		<i>xii</i>
<i>List of Publications included in this thesis</i>		<i>xiii</i>
<i>Abstract</i>		<i>xiv</i>
<b>CHAPTER 1</b>	<b>Introduction and literature survey</b>	<b>1</b>
1.1	Overview	1
1.2	Bioinformatics, chemoinformatics and drug discovery	2
1.2.1	Molecular similarity and diversity analyses	
1.2.2	Descriptors used in chemoinformatics	
1.2.2.1	1D descriptors	
1.2.2.2	2D descriptors	
1.2.2.3	3D descriptors	
1.2.3	Descriptor selection	
1.2.3.1	Wrapper methods	
1.2.3.2	Filter methods	
1.2.4	Molecular coefficients in chemoinformatics	
1.2.5	Comparison of drug-like and non drug-like compounds	
1.2.5.1	Physicochemical properties analyses	
1.2.5.2	Scaffold and molecular fragment data analysis	
1.2.5.3	Machine learning and datamining methods	
1.3	Databases and resources available	30
	<i>Publication 1: Molecular similarity and diversity approaches in drug discovery</i>	38
	<i>Publication 2: In silico methods for the analysis of metabolites and drug molecules.</i>	51
1.4	Objectives	73
<b>CHAPTER 2</b>	<b>Methods and Applications</b>	<b>74</b>



<b>CHAPTER 3</b>	<b>Development of CMKb chemoinformatics module to digitize and store chemical information.</b>	<b>76</b>
3.1	Summary	76
	<i>Publication 2: CMKb: a web-based prototype for integrating Australian Aboriginal customary medicinal plant knowledge.</i>	77
3.2	Conclusions	85
<b>CHAPTER 4</b>	<b>Comparison of physicochemical property space among human metabolites, drugs and toxins</b>	<b>86</b>
4.1	Summary	86
	<i>Publication 3: Physicochemical property space distribution among human metabolites, drugs and toxins</i>	87
4.2	Conclusions	107
<b>CHAPTER 5</b>	<b>Scaffold and fragment co-occurrence studies on datasets of biological interest.</b>	<b>108</b>
5.1	Summary	108
	<i>Publication 4: Scaffold and fragment co-occurrence studies on datasets of biological interest.</i>	109
5.2	Conclusions	149
<b>CHAPTER 6</b>	<b><i>Virtual screening of compounds active against parasitic nematodes of major socio-economic importance</i></b>	<b>150</b>
6.1	Summary	150
	<i>Publication 5: In silico approach to screen compounds active against parasitic nematodes of major socio-economic</i>	151
6.2	Conclusion	
<b>CHAPTER 7</b>	<b>General discussion and conclusion</b>	
7.1	Principal findings and conclusions	
7.2	Contributions to the field of chemoinformatics highlighting the significance of the work	
7.3	Future directions	
<b>References</b>		



## ABBREVIATIONS

<b>1D</b>	One-dimensional
<b>2D</b>	Two-dimensional
<b>3D</b>	Three-dimensional
<b>ACD</b>	Available Chemicals Directory
<b>ANN</b>	Artificial neural network
<b>AUC</b>	Area under the curve
<b>BCI</b>	Barnard Chemical Information
<b>BE</b>	Backward elimination
<b>BFGS</b>	Broyden–Fletcher–Goldfarb–Shanno
<b>CAS</b>	Chemical Abstract Service
<b>CASRN</b>	Chemical Abstract Service Reference Number
<b>CMC</b>	Comprehensive Medicinal Chemistry
<b>CMKb</b>	Customary Medicinal Knowledgebase
<b>CoMFA</b>	Comparative Molecular Field Analysis
<b>CoMSIA</b>	Comparative Molecular Similarity Indices Analysis
<b>CPDB</b>	The Carcinogenic Potency Database
<b>DT</b>	Decision trees
<b>DUD</b>	Directory of useful decoys
<b>ECFP</b>	Extended connectivity fingerprints
<b>FCFP</b>	Functional class fingerprints
<b>FS</b>	Forward selection
<b>GA</b>	Genetic algorithm
<b>GRIND</b>	Grid-Independent Descriptors
<b>HMDB</b>	Human Metabolome database
<b>HTS</b>	High-throughput screening
<b>InChI</b>	International Chemical Identifier
<b>IUPAC</b>	International Union of Pure and Applied Chemistry
<b>KEGG</b>	Kyoto Encyclopaedia of Genes and Genomes
<b>MACCS</b>	Molecular Access System
<b>MDDR</b>	MDL Drug Data Report
<b>MIP</b>	Molecular Interaction Potentials
<b>NCI</b>	National Cancer Institute



<b>PDB</b>	Protein Data Bank
<b>QSAR</b>	Quantitative structure activity relationships
<b>Ro5</b>	Rule of Five
<b>ROC</b>	Receiver Operation Characteristic
<b>SA</b>	Simulated annealing
<b>SCAM</b>	Statistical Classification of Activities of Molecules
<b>SD</b>	Structure data
<b>SMILES</b>	Simplified Molecular Input Line Entry Specification
<b>SOM</b>	Self-organising maps
<b>SVM</b>	Support vector machines
<b>TI</b>	Topological indices
<b>TTD</b>	Therapeutic Target Database
<b>VS</b>	Virtual screening



## LIST OF FIGURES

- Figure 1.1 Steps in a typical drug design program.** 3
- Identification and validation of biological target is followed by identification and optimization of leads. After testing the potential drug in the laboratory, it is approved by FDA to release into the market. The whole process on an average takes around 15 years and \$ 1billion of investment.
- Figure 1.2 Examples to show compounds with the same biological function but different scaffolds.** 4
- The Tanimoto similarity of methadone is quite low to typical opioid receptor agonist such as morphine and codeine; however, it performs the same function as others.
- Figure 1.3 Representation of a molecule  $M$  at position  $i$  in  $N$  dimensional descriptor space.** 5
- The molecules that lie close to each other in the descriptor space are likely to exhibit similar biological properties.
- Figure 1.4 Sample molecular descriptors for vasicine.** 6
- 1D descriptors are the simplest and mostly refer to whole molecule properties like molecular weight, while 2D descriptors refer to topological characteristics of the molecule. 3D descriptors on the other hand encode information regarding ligand-receptor binding.
- Figure 1.5. Representing a. a labelled graph as b. a bond (adjacency) matrix.** 9
- Connected nodes are represented as 1 while nodes with no connection are represented as 0.



- Figure 1.6. Connection table for a chemical compound, aspirin.** 9
- The connection table is divided into two blocks: atom block and bond block. The atom block comprises all the information regarding atoms, including 3D coordinates while the bond block contains information on all the bonds present in the molecule. The first row of the connection table contains the total number of atoms in a molecule followed by total number of bonds.
- Figure 1.7 Generic scheme for filter- and wrapper-based feature selection.** 15
- Filter-based methods are fast and easy to apply especially in case of large dataset whereas wrapper-based approaches require higher computational resources but are more accurate.
- Figure 1.8 Different levels of abstraction used in the literature for molecular scaffold analysis.** 22
- The scaffold is obtained by deleting all side chains. A molecular framework is obtained by replacing all heteroatoms by carbons from a scaffold while a carbon skeleton is generated by setting all bond orders to single bonds in molecular framework.
- Figure 1.9 A pictorial representation of a decision tree.** 25
- As shown in the decision tree, each internal node has a splitting predicate, with binary predicates being most common.
- Figure 1.10 An example of the basic ANN derived by emulating biological neurons.** 26
- Each input signal is associated with its own weight, which can be positive or negative. All the inputs are summed using a summing function and supplied to the activation unit. If the summation is greater than the threshold, then the output is 1, otherwise 0.



(A) There are  $n$  number of possible hyperplanes that can correctly classify the data. (B) The SVM algorithm seeks to maximise the margin around a hyperplane (optimum hyperplane) that separates a positive class (circles) from a negative class (in crossed circles).



## LIST OF TABLES

<b>Table 1.1</b>	<b>Common molecular coefficients used in chemoinformatics analysis.</b>	<b>21</b>
<b>Table 1.2:</b>	<b>List of public databases and resources used in chemoinformatics.</b>	<b>31</b>
<b>Table 2.1</b>	<b>Tools, resources and publications</b>	<b>74</b>



## LIST OF PUBLICATIONS INCLUDED IN THIS THESIS

The following publications are presented in their published form in this thesis and are referred to from this point onwards as listed in respective section of the thesis.

1. **Khanna V**, Ranganathan S: Molecular similarity and diversity approaches in drug discovery. *Drug Development Research*, 2010, 72(1), 74-84.  
Contributions to: (i) concept: VK 50%, SR 50%; (ii) data gathering: VK 100%, (iii) data analysis: VK 75%, SR 25%; and (iv) writing: VK 50%, SR 50%.
2. **Khanna V**, Ranganathan S: *In Silico* Methods for the Analysis of Metabolites and Drug Molecules, In *Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications*, eds. M. Elloumi and A.Y. Zomaya, Wiley, pp.363-383, in press.  
Contributions to: (i) concept: VK 50%, SR 50%; (ii) data gathering: VK 100%, (iii) data analysis: VK 75%, SR 25%; and (iv) writing: VK 50%, SR 50%.
3. Gaikwad J, **Khanna V**, Vemulpad S, Jamie J, Kohen J, Ranganathan S: CMKb: a web-based prototype for integrating Australian Aboriginal customary medicinal plant knowledge. *BMC Bioinformatics*, 2008, 9(Suppl 12):S25.  
Contributions to: (i) concept: GJ 50%, VK 20%, SR 30%; (ii) data gathering: GJ 75%, VK 25%; (iii) data analysis: 50%, VK 25%, SR 25%; and (iv) writing: GJ 50%, SR 50%.
4. **Khanna V**, Ranganathan S: Physicochemical property space distribution among human metabolites, drugs and toxins. *BMC Bioinformatics*, 2009, 10 (Suppl 15):S10  
Contributions to: (i) concept: VK 50%, SR 50%; (ii) data gathering: VK 100%, (iii) data analysis: VK 75%, SR 25%; and (iv) writing: VK 50%, SR 50%.
5. **Khanna V**, Ranganathan S: Scaffold and fragment co-occurrence studies on datasets of biological interest. (submitted)  
Contributions to: (i) concept: VK 50%, SR 50%; (ii) data gathering: VK 100%, (iii) data analysis: VK 75%, SR 25%; and (iv) writing: VK 50%, SR 50%.
6. **Khanna V**, Ranganathan S: *In silico* approach to screen compounds active against parasitic nematodes of major socio-economic importance. (under preparation)  
Contributions to: (i) concept: VK 50%, SR 50%; (ii) data gathering: VK 100%, (iii) data analysis: VK 75%, SR 25%; and (iv) writing: VK 50%, SR 50%.



## ABSTRACT

The search for pharmaceutically interesting compounds using computational methods is the core idea in chemoinformatics. With the advent of combinatorial synthesis and high-throughput screening (HTS), researchers and drug industries are currently able to screen millions of compounds each day. However, improvements in screening capabilities have failed to yield a proportionate increase in novel chemotypes. Given the magnitude of compounds in one of the most popular chemistry databases, PubChem, it is irrational to experimentally screen all compounds for a potential target.

This thesis aims to study the property space occupied by therapeutic compounds of economic importance obtained from public datasets, using chemoinformatics tools and computational technologies.

With this objective in mind, a comprehensive review of current chemoinformatics research, with a particular emphasis on drug discovery was carried out. In addition, the most commonly used, freely available small molecule databases and algorithms for small molecule analysis were also reviewed. Further, recent developments in computational library design techniques were summarized in a separate review article.

For web-based analysis and visualization of small molecules, I have developed the chemoinformatics analysis module for the Customary Medicinal Knowledgebase (CMKb; <http://www.biolinfo.org/cmkb>) which has served as a prototype to integrate the use of medicinal plant among Australian Aboriginals with bioactives, for identifying potential lead compounds.

In order to examine the similarity of current drug molecules with human metabolites and toxics, a preliminary comparative study based on several computed physicochemical properties and functional groups was carried out. We established that searching against complete datasets was comparable to results obtained from clustered data. We then used a multi-criteria approach to analyse physicochemical properties, scaffold architecture and fragment occurrence among large public datasets of biological interest *viz.* drugs, metabolites, toxics, natural products, lead compounds and the ChEMBL dataset. Fragments are often dependent on each other and therefore, fragment co-occurrences were further assessed by association analysis. Going beyond the general datasets, a nematode-specific



anthelmintic dataset was also analysed. Machine learning methods were used to screen potential anthelmintic compounds from public collections and novel anthelmintics have been identified.

From our preliminary analysis, it was established that although the physicochemical property space occupied by the drugs, human metabolites and toxics was distinct, present-day drugs are more akin to toxic compounds than to metabolites. This result was in accordance with high attrition rates in drug discovery projects. Furthermore, we concluded that empirical rules such as Lipinski's "rule of five" can be supplemented to include toxicity information. Following preliminary study on physicochemical properties, we corroborated our earlier finding that metabolites are least similar to current day drugs in our subsequent comprehensive analysis. However, in scaffold analysis we found that over 42.0% of the non-redundant metabolite scaffolds are represented among drugs which suggest that drugs and metabolites largely differ in side chains and linkers but vastly share the scaffold space. Additionally, a robust statistical technique known as association analysis was explored for the first time in chemoinformatics to carry out efficient mining and fragment co-occurrence analysis.







# Chapter 1: Introduction and literature survey

## 1.1 Overview

The drug discovery process is time consuming and an expensive endeavour. Moreover it is tedious and failure at any stage of development is a strong possibility. The total cost estimated to bring a drug to the market is estimated to be between \$500 million and \$1.2 billion dollars [1]. Traditionally, drugs were discovered by synthesizing compounds in a multi-step process, followed by screening against a number of biological targets *in vitro* and *in vivo*. Pharmacokinetic properties, toxicity, metabolism and efficacy were further accessed for promising candidates obtained from the screening experiments. Given the significant advances in genomics, proteomics, computational chemistry and virtual screening (VS), *in silico* drug design can help speed up some of the rate limiting steps in the traditional drug discovery pipeline. Briefly, *in silico* drug design refers to the use of chemical information to build computational models that can make predictions and suggest hypothesis which ultimately advance our knowledge of medicines and therapeutics. *In silico* methods can help in identifying drug targets, analysing their structure for possible binding sites, designing and screening virtual libraries, checking lead/drug-likeness, docking selected ligands to their respective targets, accessing binding affinities and further optimizing molecules for enhancing efficacy and reducing toxicity.

Following on from the significant advancements in combinatorial synthesis (parallel synthesis of molecules) and high-throughput screening (HTS) techniques, pharmaceutical industries are now able to synthesize and assay a vast number of compounds per target per year. Nevertheless, an exhaustive search of the biologically relevant chemical space, which is estimated to be the order of  $10^{60}$ , is not feasible by traditional methods [2]. In order to consider this vast chemical space for screening, it is essential to deploy computational resources and develop methods that can eliminate unwanted or redundant compounds early in the screening process. Consequently, drug-likeness and related concepts are being explored in detail since mid 1990's. However, these efforts and increasing screening capabilities have not yielded a proportionate increase in novel chemotypes [3]. In addition, it is becoming increasingly clear that too many promising compounds are eliminated during clinical testing due to various reasons, with the recent report that drugs often fail due to toxicity [4].



In this thesis, we study the property space occupied by freely available biologically relevant compound datasets, especially drugs, human metabolites, natural products and toxic compounds. In addition, some of the computational chemistry techniques applied in the early stages of drug discovery and their application to discover novel anthelmintic compounds is further investigated. Improved and faster methods to analyse large amounts of chemical data are suggested and more importantly, several novel potential anthelmintic compounds have been identified.

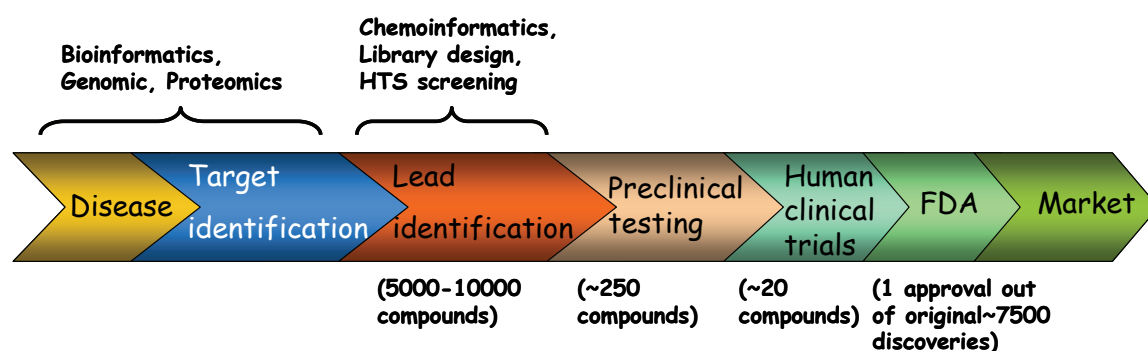
## 1.2 Bioinformatics, chemoinformatics and drug discovery

Modern drug design methodology has enormously benefited from the computational resources available today. Bioinformatics is defined as the application of computational techniques for the analysis of biological or sequence data obtained from experiments, modeling and database searching [5]. Until recently, drug development was restricted to a small fraction of possible targets since most of the human genes were unknown. Usually pharmaceutical companies follow those targets that are identified and well studied. Bioinformatics, in the drug design context, aims to facilitate the identification and validation of novel drug targets. Chemoinformatics, on the other hand, deals with structural information of small molecules and is defined as the use of computers and information techniques, applied in the field of chemistry with the intended purpose of guiding drug discovery and development [6]. It is an evolving area of research and has been frequently reviewed by number of authors in the past [6, 7] and in recent times [8, 9]. A good critique on the origin of chemoinformatics was given by Hann and Green [10]. Recently, a new term “Bio-Chemo-informatics” comprising both bioinformatics and chemoinformatics has emerged in the literature [11, 12]. It has been used to describe the research efforts on meeting the emerging need for the integration of bioinformatics and chemoinformatics [11, 12]. Due to the applied nature of the field, chemoinformatics has found many applications in drug discovery and development. Drug discovery generally follows a set of common stages as shown in Figure 1.1.

The first stage is target identification, usually by analysing the lifecycle of pathogen to identify the critical points of intervention followed by target validation. Druggability and similarity to human proteins is also analysed with the help of various bioinformatics techniques. A validated target is screened against millions of compounds in parallel in HTS assays and the successful results of these screens are known as *hits*. A number of *hits* are followed up as *leads* to determine whether any of these can be converted to *candidates*



with further optimization of absorption, distribution, metabolism, excretion, toxicity (ADMETox) properties and biological activity. Once appropriate *candidates* are indentified, they proceed further along the pipeline for preclinical development. Applications of chemoinformatics in drug design include molecular similarity and diversity analyses, data mining, library generation, lead screening and optimization, exploration of quantitative structure activity relationships (QSAR), and the analysis of drug-like features. All these studies require chemicals to be represented in a suitable format that can be recognized by computers. These representations range from one-dimensional (1D) line notations to three-dimensional (3D) molecular models which result in various descriptors. The following section describes molecular similarity and diversity analyses as a major application of chemoinformatics in drug discovery.

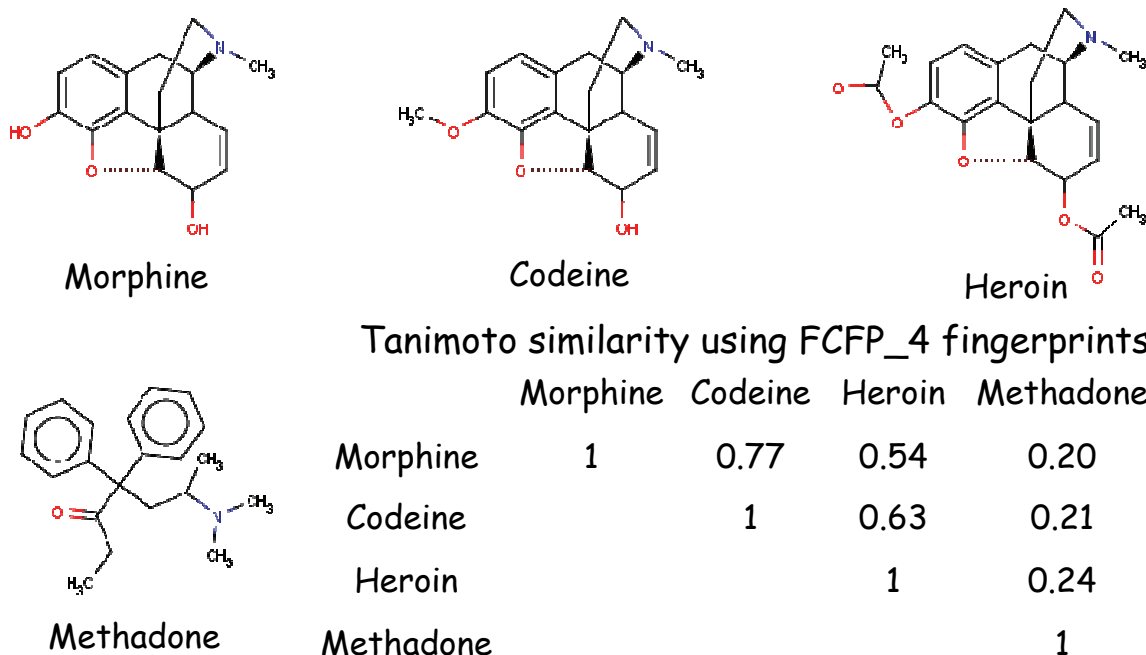


**Figure 1.1: Steps in a typical drug design program.** Identification and validation of biological target is followed by identification and optimization of leads. After testing the potential drug in the laboratory, it is approved by FDA to release into the market. The whole process on an average takes around 15 years and \$ 1billion of investment.

### 1.2.1 Molecular similarity and diversity analyses

Molecular similarity and diversity are the core concepts in chemoinformatics and have found particular favour in pharmaceutical industry. According to the “*similar-structure, similar-property principle*” which is often known as “*similarity property principle*” structurally similar compounds tend to have similar properties – both physicochemical and functional, more often than structurally dissimilar compounds [13, 14]. In large part, this is true however it can break down in certain cases [15]. The opioid ligands shown in Figure 1.2 are the good examples where “*similarity property principle*” breaks down:





**Figure 1.2: Examples to show compounds with the same biological function but different scaffolds.** The Tanimoto similarity of methadone is quite low to typical opioid receptor agonist such as morphine and codeine; however, it performs the same function as others.

Morphine, codeine and heroin all share the same basic scaffold and show similar bioactivity, i.e. opioid receptor agonist. Methadone also binds to the opioid receptor and acts as an opioid receptor agonist; however, it does not share any structural resemblance to other ligand members of the typical opioid receptor agonist family.

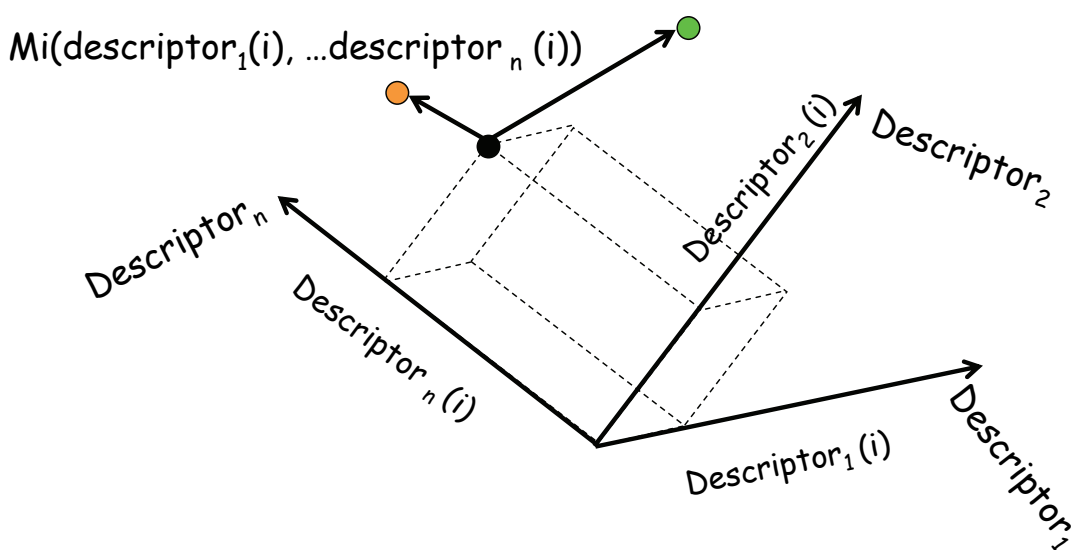
Molecular similarity provides a simple but elegant method for VS and lies at the core of all the clustering methods available [16]. On the other hand, molecular diversity explores the structural coverage of a set of molecules in the chemical space and underlies many approaches for novel compound selection and design of combinatorial libraries [16]. Molecular similarity and diversity are complementary concepts and the correct choice of similarity/diversity measures can help us place molecules at an optimal location in chemical space, in order to maximize diversity and minimize redundancy. Measuring molecular similarity and diversity involve, in general, three main components: *structural descriptors* that represent chemical compounds in a way that they can easily be compared, *molecular coefficients* that provide mathematical function to calculate similarity or diversity and *weighting schemes* that assign the relative importance to different structural



descriptors. While there are few reports on weighting schemes [17, 18] and their effects on the utility of molecular coefficients, much interest has been shown in the type of descriptors [16, 19, 20] and coefficients [21, 22] used for similarity or diversity analyses. Various descriptors and their uses in chemoinformatics are described below.

### 1.2.2 Descriptors used in chemoinformatics

Descriptors are used to encode a variety of structural features in a molecule. A descriptor places two molecules in a chemical space at a distance that is proportional to their distance in bioactivity or some other property under study as shown in Figure 1.3.

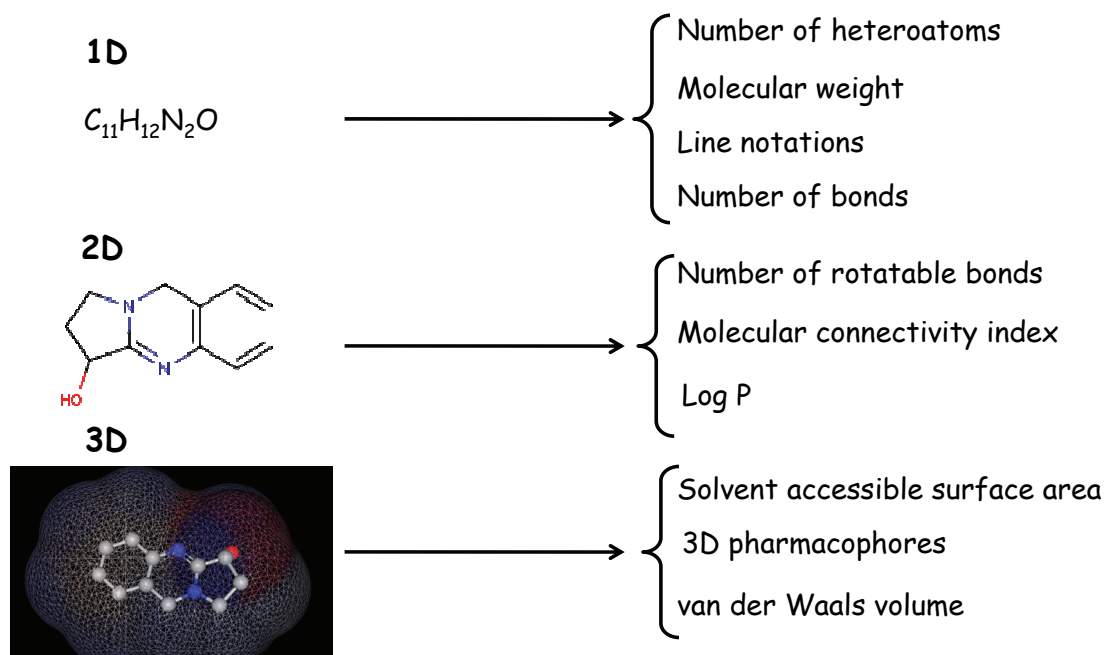


**Figure 1.3: Representation of a molecule  $M$  at position  $i$  in  $N$  dimensional descriptor space.** The molecules that lie close to each other in the descriptor space are likely to exhibit similar biological properties.

Descriptors can be determined using structure (constitution, configuration and conformations) or properties (physical, chemical or biological) of the molecule [16]. This section deals with the recent developments in the representation of a molecule in chemical space and the generation of molecular descriptors. A large number of descriptors have been developed that can be used in similarity calculations. Figure 1.4 shows three representations of the same molecule. The chemical formula (1D representation) conveys maximum information about the chemical constituents of the molecule while two-dimensional (2D) representations add knowledge about the molecular connectivity. 3D representations, on the other hand, deal with the conformation, without losing any information present in 1D descriptors. Nonetheless, it can be argued that the chemical composition is most easily seen in 1D representation than in 2D or 3D representations. The colour-coded isodensity surface (3D



representation) contains little, if any, information about the chemical composition. However, 3D representations contain by far the most information about the ability of the molecule to interact with biological target.



**Figure 1.4: Sample molecular descriptors for vasicine.** 1D descriptors are the simplest and mostly refer to whole molecule properties like molecular weight, while 2D descriptors refer to topological characteristics of the molecule. 3D descriptors on the other hand encode information regarding ligand-receptor binding.

Several software packages are available to calculate a wide variety of descriptors. Examples include the Chemistry Development Kit [23], JoELib [24], PowerMV [25], Pipeline Pilot [26], Molecular Operating Environment [27], Bioclipse [28] and CODESSA [29]. Owing to the large number of descriptors available, this discussion is confined only to descriptors that are relevant to the drug discovery paradigm. Classically, descriptors are categorized into three groups and vary in complexity and encoded information.

### 1.2.2.1 1D descriptors

1D property descriptors are the simplest descriptors and include physicochemical properties and the numerical count of features such as rings systems, hydrogen bond donors and hydrogen bond acceptors. Hydrophobicity is measured in terms of log P, defined as the logarithm of partition coefficient between n-octanol and water. Since no geometrical information is contained in 1D property descriptors, they are easily calculated from molecular structure alone and hence are often employed to predict physical properties. These descriptors are also called whole molecule descriptors because a single



value is derived from them (e.g. molecular weight, log P, number of heavy atoms), which describes the property of whole molecule.

#### 1.2.2.1.1 Representing molecules in 1D

1D line notations are quite popular representation where a molecule is represented as a linear string and nodes represent atoms. Molecular formula and linear notations such as Simplified Molecular Input Line Entry Specification (SMILES) are common examples of 1D line notations.

- ***Chemical Formula***

The most common way to represent the chemical structure is a chemical formula. It is not only compact and easy to interpret but also conveys chemical constituents and number of atoms in a compound. It, however, lacks information regarding connectivity, stereochemical configuration and 3D coordinates which are essential for advanced studies. If a molecule contains more than one atom of a particular element, the number is represented as a subscript following the chemical symbol in the formula. The chemical formula for vasicine (shown in Figure 1.4) is C<sub>11</sub>H<sub>10</sub>N<sub>2</sub>O. This gives a clear picture of atom constituents and molecular mass of the molecule.

- ***1D Line Notations***

Linear notation represents complete constitution and connectivity of the chemical compound as linear sequence of characters. A number of line notations have been introduced over the years and were popular during 1960 to early 70's. The most common are Wiswesser Line Notation [30], Representation of Organic Structures Description Arranged Linearly [31], SMILES [32] and SYBYL Line Notation [33]. Among these 1D notations, SMILES has found widespread use in representation and exchange of chemical information over the internet. Nevertheless, one major problem recognized early in SMILES representation was the lack of unique encoding because the same molecule can be represented by different SMILES codes. This issue has largely been resolved by using the Morgan algorithm, which was proposed in 1965 to provide canonical ordering of atoms in a molecule. Additionally, since its origin SMILES has been subjected to various modifications (SMARTS, SMIRKS) [34].



Beside these, International Union of Pure and Applied Chemistry (IUPAC) nomenclature and the recently introduced International Chemical Identifier (InChI) can also be considered as 1D Line notations. Although not widely accepted, InChI tries to overcome the limitations in IUPAC nomenclature but is more difficult to interpret.

### 1.2.2.2 2D descriptors

Descriptors derived from the knowledge of molecular topology are called 2D descriptors such as topological indexes (molecular connectivity index,  $\kappa$  shape index), 2D fingerprints (dictionary-based and hash-based) and 2D fragment descriptors (atom pairs, augmented atoms and atom sequence). Several of these descriptors are based on the representation of a molecular structure as a graph.

#### 1.2.2.2.1 Representing molecules in 2D

The most commonly used representations of compounds in 2D format are listed below:


- **Graphs**

The most acceptable way of representing chemical structure is by molecular graphs or structure diagrams. Similar to line notations, chemical graphs are also hydrogen-suppressed notations. The molecular graph is a collection of nodes representing atoms connected by edges, which are bonds, linking the atoms. Despite its popularity, there are problems related to this method and the most evident problem is graph isomorphism. Two graphs  $G_1$  and  $G_1'$  are said to be isomorphic if they contain same number of graph vertices (nodes) and are connected in the same manner, which could lead to two different chemical structures being isomorphic. Fortunately, there are well established algorithms that can be used to establish whether two molecular graphs are isomorphic or not. Further matrices, can be used to extend molecular graphs to include connectivity and chemical bond information.


- **Matrix**

There are two common ways to encode a molecule from a graph to a machine readable matrix format: graph adjacency or a bond matrix and a connection table. It is conventional to assume that the unlabeled nodes are carbon atoms, and hydrogens are added based on valence state of the atoms when computing an adjacency or bond matrix of the molecule. Such a matrix  $M$  is composed of one row and column for each



a. 

Labeled graph

b. 

Corresponding bond matrix

No. of atoms      No. of bonds

X, Y, Z coordinates

Atoms block

Atoms

Bonds block

M END

9



information on all the bonds present in the molecule. The first row of the connection table contains the total number of atoms in a molecule followed by total number of bonds.

The other way of representing a molecule in matrix form is a connection table, indexing all the atoms and the bonds which exist between them. Figure 1.6 illustrates the connection table for the aspirin molecule.

The first three columns in the connection table specify the X, Y, Z coordinates of the atoms. The next few column stores the atom symbols, formal charge, bond type, chirality and other atom or bond properties like isotope, charge, stereocode etc. Despite the effort to standardize the format of connection tables there are still various format available. Most commonly used connection table format is MDL a proprietary file format developed by MDL. There are many flavours available: Molfiles, Rgfiles, Rxnfiles, SDfiles , RDfiles and XDfiles.

- **Topological indices (TI) descriptors**

TIs are numerical quantities based on certain characteristics of the molecular graph and encode for molecular properties, such as ring structure, the number of atoms, branching, heteroatom content and bond order. They are easy to calculate and hence have found widespread use among researchers. Many indices have been reported in the literature. TIs are subdivided into three generations. The first generation TIs are the simplest and are based on integer graph properties like topological distances, degree of branching and overall shape. The most representative of this class is the Wiener index  $W$  and the centric indices of Balaban,  $B$  and  $C$ . Second generation indices, such as molecular connectivity indices are real numbers, derived from the integer graph properties. The most successful indices of this class are molecular connectivity indices, Kier and Hall indices [35] and  $\kappa$  shape indices [36]. Third generation indices are real value numbers derived from graph properties. They were recently introduced and offer a wide range of selection possibilities. Estrada and Uriarte have published an excellent review on the role of TI in drug discovery [37].

- **2D fingerprint descriptors**

Perhaps the most commonly used descriptors are 2D fingerprints. They encode the presence or absence of substructural fragments within a molecule as a binary vector. There are two types of fingerprints described in the literature: dictionary-based and



hash-based. The dictionary-based fingerprints consist of a binary vector with those bits set 1 (ON) that correspond to the substructural fragment found in the fragment dictionary. The fragments that are present in the dictionary should be chosen carefully as dictionary-based fingerprints are dataset-dependent. Hence, a limitation of the dictionary-based approach is that, every time a dataset changes, a new dictionary has to be created. Molecular complexity [38] is yet another complication that can introduce bias in fingerprint comparisons. More complex structures tend to have more bits which are set to 1 (ON) in the bit string than simple or topologically less complex structures. A recent paper by Wang and Bajorath [39], discusses a new approach, called bit silencing, for selecting important bits from a fingerprint based on the structural features. Examples of dictionary-based fingerprints are Barnard Chemical Information (BCI) fingerprints [40] and the Molecular Access System (MACCS) structural keys.

Hash-based fingerprints eliminate the reliance on a pre-defined list of substructural fragments and thus avoid the shortcomings of dictionary-based fingerprints. Patterns are generated from the molecule itself by enumerating all the paths in the molecule. The bits in different patterns may overlap, due to finite length of the bit string and the large number of possible patterns generated by the hashing function. The most commonly used hash-based fingerprints are Daylight [34] and UNITY fingerprints [41].

- **2D fragment descriptors**

The next group of descriptors are fragment descriptors. Examples of fragments include *atom pairs*, *augmented atoms*, *atom sequences*, *bond sequences* and various *ring fragments*. *Atom pairs* consist of all unique bonded atom pairs in a molecule. *Augmented atoms* consist of atoms and their neighbours, while *atoms sequences* consist of atom pairs along with their bond information. *Ring fragments*, on the other hand, include rings along with the atoms surrounding the rings. Recently, Pipeline Pilot [26] circular substructure fingerprints have proven useful, and provide fingerprints with extended connectivity (ECFP) and functional connectivity (FCFP). An extensive review on Pipeline Pilot fingerprints has been recently published [42].

### 1.2.2.3 3D descriptors

The physical, biochemical and the molecular recognition properties of a compound often depends on the conformations that it can adopt hence, it is quite informative to compare



molecules using their 3D characteristics. Since, 3D descriptors require conformational properties to be considered, they are more computationally demanding than 1D or 2D descriptors. The issues of conformation generation, sampling and refinement have hindered the use of 3D descriptors in VS. There are two generations of 3D descriptors: the traditional alignment-dependent, descriptors generated from Comparative Molecular Field Analysis (CoMFA) or Comparative Molecular Similarity Indices Analysis (CoMSIA) approach and the alignment-independent descriptors such as Grid-Independent Descriptors (GRIND) and VolSurf descriptors.

#### 1.2.2.3.1 Alignment-dependent descriptors

- **CoMFA**

The aim of the CoMFA [43] is to elucidate the correlation between biological activity and the Molecular Interaction Potentials (MIP) of a set of molecules with a common binding mode. The MIP contains the information regarding the interaction energies between the probes and the compound placed on a 3D grid. At each point on the grid, steric, electrostatic and hydrophobic field values are measured for each molecule by interaction with the probe atom. Collectively, all the above calculated fields are referred to as MIP, which contains a full set of information related to the interaction potential of a molecule. For comparing a series of molecules based on their MIP, it is important to align the molecules so that the same grid box can be used for all the compounds. Unfortunately, alignment is not a trivial task especially when molecules are structurally diverse. There are a number of techniques proposed for aligning diverse compounds with varied levels of similarity [44]. The quality of the alignment determines the quality of all further calculations. Therefore, the quality of alignment poses a major challenge in the use of alignment-dependent descriptors. Once the compounds are aligned, MIP can be calculated using a grid box which generally results in thousands of descriptors. Following the calculation of descriptors, multivariate techniques are used to analyse the data further.

- **CoMSIA**

The CoMSIA [45] is similar to CoMFA, where the molecular fields are expressed in terms of similarity indices between the compounds of interest calculated via a common probe atom. It is believed to be less affected by changes in molecular alignment. Compared to CoMFA, CoMSIA uses a different potential function called the Gaussian-type function instead of Lennard-Jones and Coulombic function which provide



accurate information in grid points and easily interpretable contour maps. Furthermore, the CoMSIA method takes into account various properties that potentially contribute significantly to ligand binding such as steric, electrostatic, hydrophobic, hydrogen bond acceptor and hydrogen bond donors. Like CoMFA, leave-one-out and other cross-validation procedures are used to validate the models developed by CoMSIA methodology.

#### **1.2.2.3.2 Alignment-independent descriptors**

Due to the limitation of the alignment step mentioned above, alignment-free descriptors were developed. The idea is to retain the MIP information as far as possible without the need for structural superimposition of the compounds under study. The simplest alignment-independent descriptor is the dipole moment. Another important descriptor of this class includes pharmacophoric fingerprints which makes use of the “lock and key” concept and relies on the internal distance of the molecule. A set of pharmacophoric features that have critical interaction with the receptor such as hydrophobic centres, hydrogen bond acceptors and hydrogen bond donors are calculated. Hundreds of combinations of two-point, three-point and four-point pharmacophores are computed. Two-point pharmacophores represent all the possible combinations of atom pairs in a molecule while three-point pharmacophores provide a clearer representation of the interatomic distances and their relative orientation, while four-point pharmacophores can distinguish between stereoisomers. Like 2D fingerprint descriptors, pharmacophore features present in the molecule are set to 1 otherwise 0 and can successfully be used for database searching.

- **VolSurf descriptors**

VolSurf [46] is a procedure to derive simple relevant physicochemical descriptors from 3D molecular field maps. The molecular descriptors obtained contain information regarding size and shape of hydrophilic and hydrophobic regions, as well as the balance between them. Critical packing, amphiphilic moment, hydrogen bonding polarisability and energy minima distances are other useful descriptors. VolSurf descriptors are hardly influenced by conformational sampling, fast to calculate and independent of alignment of molecules. In order to calculate VolSurf descriptors, the MIP is first calculated with a probe using the program GRID, following which it is analysed to extract the relevant information required to compute the VolSurf descriptors.



- **GRIND descriptor**

GRIND [47] descriptors are another type of 3D alignment-independent descriptors designed to characterize ligand-receptor interactions [47]. They are obtained from a set of molecular interaction fields which are first simplified and the results are encoded in alignment-independent descriptors, using a particular type of autocorrelation transformation. The descriptors calculated can be used in a variety of chemometric analyses such as principal component analysis or partial least square analysis, and even the analysis of correlograms, colour-coded according to chemical groups or activity. Like other alignment-free descriptors, GRIND descriptors are unaffected by the position and orientation of molecules in the space. The calculation of GRIND descriptors involves three steps: (i) computing a set of MIF, (ii) filtering the MIF to extract the most relevant regions that can describe the receptor site and, (iii) encoding the extracted regions into GRIND variables. After calculating the MIF, probes generally used for the calculating the GRIND descriptors are the hydrophobic probe (DRY), the hydrogen bond donor amide nitrogen probe and (N1) and the hydrogen bond acceptor carbonyl oxygen probe (O).

Many researchers have concluded that overall 2D descriptors perform better than 3D because they are conformation-independent and easy to calculate [48-50]. However, counterclaims have also been made by several others [51, 52]. A comparative study has shown that different type of descriptors have their own use and therefore, a choice must be made in individual cases [53].

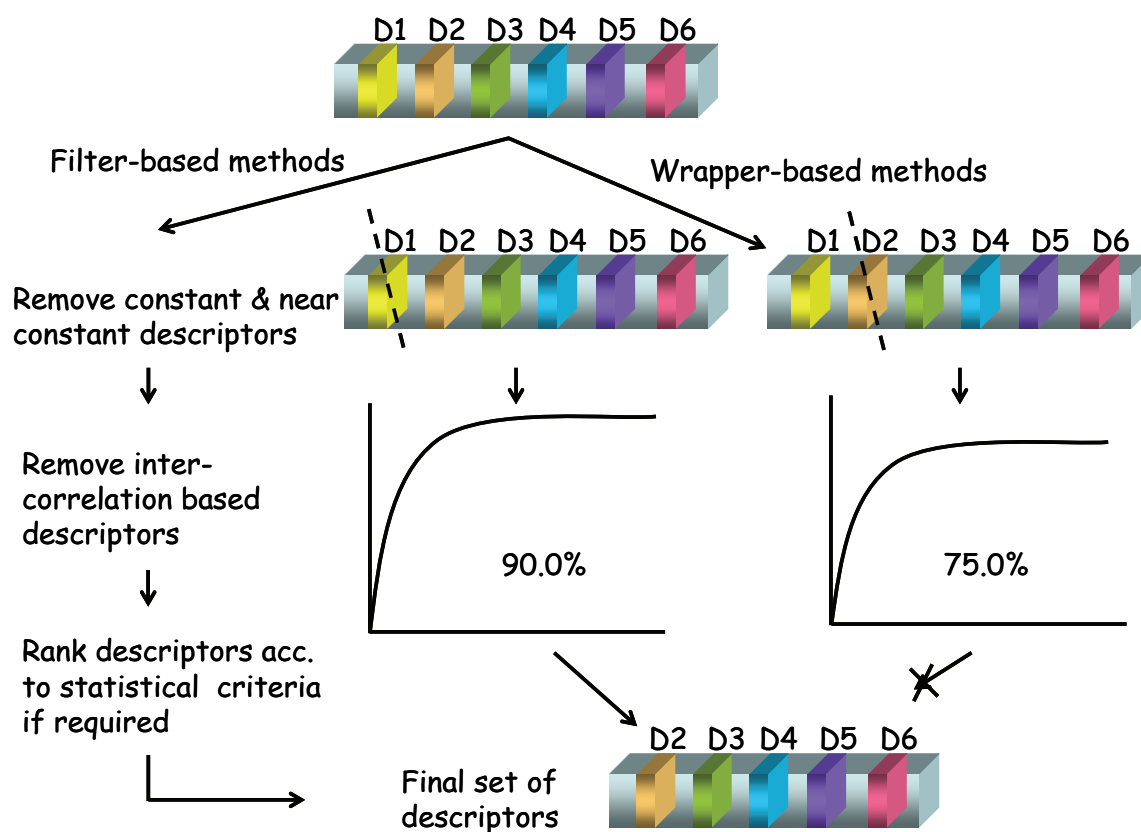
### **1.2.3 Descriptor selection (feature or variable selection)**

As the dimensionality of the data increases, the data becomes increasingly sparse in the space it occupies. This is referred to as the “*curse of dimensionality*” and can lead to problems for both supervised and unsupervised learning. Usually not all the descriptors contribute equally to explain the property of interest and some may even add noise to the model. Considering that so many descriptors are available, few important questions arise. Why and how should descriptors be selected for similarity or diversity analysis and QSAR studies? Are there any descriptors which perform best for a particular kind of study? Is the performance of simple descriptors comparable to complex 3D descriptors? It is too computationally expensive to examine all descriptors; moreover, some descriptors or combinations of descriptors are redundant and thus can be set aside, with no or little information loss. A group of descriptors might work brilliantly for one target, but may be



poor choice for another target. It is therefore, essential to select a good subset of descriptors that are suitable to study the bioactivity of compounds. This requires a systematic exploration of the descriptor space in order to examine all combinations of the best descriptor subsets. Several subset solutions are possible and thus, a tool is required that could inspect all the possible solutions that determine the best possible solution subset. Ideally descriptors calculated should be free of correlation ( $r < 0.6$ ) [54] proposed for further chemoinformatics analysis. For feature selection in unsupervised learning, learning algorithms aim to find a good subset of features that form high quality clusters.

There are two main approaches to select of descriptors in a supervised learning context (Figure 1.7). These are the wrapper-based methods and the filter-based methods.



**Figure 1.7 Generic scheme for filter- and wrapper-based feature selection.** Filter-based methods are fast and easy to apply especially in case of large dataset whereas wrapper-based approaches require higher computational resources but are more accurate.

### 1.2.3.1 Wrapper method

The first method of feature selection is the wrapper approach [55]. It consists of using a classifier as a black box for selecting the best subset of features and uses cross-validation



to compare the error rate of the candidate subsets. In the wrapper approach, the result relies heavily on the search algorithm and on the assessment of the performance. Most often, the performance criterion is the error rate. Nevertheless, other criteria can be used. This may be the cost if a misclassification cost matrix is used. The area under the curve (AUC) can also be used, when we assess the classifier using different Receiver Operation Characteristic (ROC) curves. There are two types of search strategies that can be employed by wrapper approaches for feature selection – statistical and optimization.

#### **1.2.3.1.1 Statistical approaches**

The well known statistical algorithms used for descriptor selection are forward selection (FS) and backward elimination (BE). In the FS, at each step, the variable that really contributes to the discrimination between the groups is determined. The feature is added to the selection group if its contribution is significant. The process stops when there is no feature to add in the model or the required number of features is reached. BE algorithm, successively eliminates descriptors starting from the complete set of descriptors. The algorithm searches for the less relevant features and removes it, if the removal does not significantly reduce the discrimination between groups. The process stops when there is no variable to remove. Both these algorithms have a drawback of “nesting”. In the FS context, nesting refers to the fact that once a particular feature is added it cannot be removed later on even if the removal of the feature may improve the objective function. More sophisticated algorithms with a backtracking phase after addition of a feature have been introduced [56] to overcome nesting. The optimization algorithm avoids the nesting problem by introducing some degree of randomness into the search strategy.

#### **1.2.3.1.2 Optimization approaches**

Optimization methods can be divided into two broad classes – deterministic and stochastic. The best known examples of deterministic methods are the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [57] and the Nelder-Mead simplex algorithm [58]. Examples of stochastic methods include the genetic algorithm (GA) [59], the simulated annealing (SA) algorithm [60] and particle swarm optimization [61]. GA mimics the biological process of evolution and natural selection therefore, much of the vocabulary from evolution has been adopted for the use in the field of GA. Thus a chromosome in the feature selection context refers to a subset of descriptors associated with a fitness value. A population is defined as the collection of such chromosomes. A number of models (which can be linear or non-linear models) are created randomly in the first generation, the best of which (as measured



by root mean square error of each model) are selected and interbred to create new generations. The genetic operators such as crossover and mutation are applied to generate new chromosomes with better features than the parent chromosomes. The objective function is optimized over the course of many generations. GA was first used for feature selection in QSAR by Rogers and Hopfinger [62]. Subsequently, GAs have been used in feature selection for QSAR with a number of mapping methods, such as Artificial Neural Network [63] and Random Forest [64].

Another stochastic method for function optimization applied in QSAR is SA. It was introduced independently by Kirkpatrick *et al.* [60] in 1983 and Gelatt and Vecchi [65] in 1985. The SA process is also inspired by nature. Annealing, in general, refers to the cooling of glass or metal; if the cooling is slow enough it results in ordered states and when the final temperature is reached the configuration of the atoms is at the most stable state, whereas rapid cooling produces states with low order. It should be noted that both SA and GA share the fundamental assumption that it is highly likely to find better solutions in the vicinity of optimal solutions, than by randomly searching the whole solution space. However, the key difference between SA and GA is that while SA creates new solutions by modifying only one solution, GA creates solutions by combining two different solutions. SA is a generalization of the Metropolis algorithm.

In the Metropolis algorithm, each point  $s$  of the search space represents a state of some physical system with internal energy,  $E(s)$ . The current state of the system is disturbed by iterative mutation  $s'$  and the change in energy, called  $\Delta E$  ( $\Delta E = E(s) - E(s')$ ) is evaluated. For a minimization task, the change is accepted if  $\Delta E$  is negative and in case it is positive, the new state is accepted with the probability equal to Boltzmann factor,  $\exp(-\Delta E/kT)$ . Occasional acceptance of positive  $\Delta E$  does not allow the algorithm to get stuck in a local minimum; therefore, the solutions obtained from SA are often of high quality. This step is repeated until the threshold criterion is reached, usually in terms of a predefined  $\Delta E$  cut-off. In SA, the temperature term is often added, so that SA consists of an outer temperature loop and an inner Metropolis algorithm.

### 1.2.3.2 Filter method

Although the wrapper approach may achieve better performance, it presents two main disadvantages: it requires more computational resources and is prone to overfitting. The second method of feature selection is the filter approach. It consists of selecting the best



subset of features, using an *ad hoc* criterion. Filter methods rank the subset of features independent of the classifier. As a result, feature selection needs to be carried out only once, and then different classifiers can be evaluated. For large datasets, filter approaches are more practical than wrapper approaches because they are much faster. However, the common disadvantage of filter methods is that most of the proposed techniques are univariate and thus cannot handle redundancies among features. In order to overcome this problem, a few multivariate techniques have also been proposed. There are many types of filter-based methods such as inter-descriptor correlation and ranking methods.

#### **1.2.3.2.1 Correlation-based filter methods**

With hundred of descriptors available, it is likely that many descriptors are inter-correlated. Therefore, Pearson's correlation coefficient may serve as a preliminary filter for discarding these descriptors. This can be done by measuring the association between pairs of descriptors and discarding the descriptors if their correlation coefficient exceeds a predefined threshold, rather than randomly discarding one of the descriptors [66]. However, if one of the descriptors from the pair is a topological descriptor, then the topological descriptor is preferentially retained while the other is discarded.

#### **1.2.3.2.2 Ranking-based filter methods**

Ranking methods are based on the association between descriptors and the target attribute. This association may be correlation-based ranking, based on a correlation coefficient or any other statistically relevant parameter. For correlation-based ranking, the correlation of each descriptor with the target attribute is calculated initially and then ranked according to the decreasing order of the correlation coefficient. Other methods attempt to rank features according to a different relevancy score such as the F-ratio [67].

### **1.2.4 Molecular coefficients in chemoinformatics**

The second main component required for similarity and diversity analysis or to compare two objects with a common set of attributes, are molecular coefficients. This section deals with those coefficients that are widely used in chemoinformatics. Some coefficients directly measure the similarity between the molecules and are termed as similarity coefficients; while others which measure the distance or dissimilarity are called distance coefficients. The similarity between two molecules can be judged from distance coefficients by subtracting the value from unity.



**Table 1.1: Common molecular coefficients used in chemoinformatics analysis.**

Molecular Coefficient	Formula	Range	Reference
<b>Associative coefficients</b>			
Tanimoto	$\frac{c}{a+b-c}$	0 to 1	[68]
Size modified Tanimoto	$\frac{2-p}{3}T_c + \frac{1+p}{3}T_{c,0}$	0 to 1	[69]
Russel-Rao	$\frac{c}{n}$	0 to 1	[70]
Cosine/Ochiai	$\frac{c}{\sqrt{ab}}$	0 to 1	[71]
Tversky	$\frac{c}{\alpha a + \beta b + c}$	0 to 1	[34]
Simpson	$\frac{a}{\min(a+b, a+c)}$	0 to 1	[72]
Forbes	$\frac{cn}{ab}$	0 to $\infty$	[73]
Fossum	$\frac{n(c-0.5)^2}{ab}$	0 to $\infty$	[34]
Dice/Sorenson/ Czekanowski	$\frac{2c}{a+b}$	0 to 1	[74]
<b>Correlation coefficients</b>			
Dennis	$\frac{(cd) - (ab)}{\sqrt{nab}}$	0 to $\infty$	[75]
Pearson	$\frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	-1 to 1	[76]
Yule	$\frac{ad - bc}{ad + bc}$	-1 to 1	[77]
<b>Distance coefficients</b>			
Euclidean distance	$\sqrt{a+b-2c}$	0 to 1	[78]
Soergel distance	$\frac{a+b-2c}{a+b-c}$	0 to 1	[22]
Hamming/Manhattan/City-block	$a+b-2c$	0 to 1	[76]



Comprehensive reviews on molecular coefficients are available [22, 76], with several studies comparing the performance of the different molecular coefficients [77, 79, 80], which can be characterized into three groups. *Association coefficients* are mainly used for binary representation and their values vary in direct proportion to the degree of similarity i.e. greater similarity is indicated by higher values. *Correlation coefficients* measure the degree of correlation and range from -1 to +1 where -1 indicates that any change in one property would be accompanied by equal and opposite change in the other and *vice versa*. *Distance coefficients* measure the degree of dissimilarity between the objects and their values are inversely proportional to the degree of similarity, so that the higher the value, the lower the similarity. A fundamental difference between *association* and *distance coefficients* is that the latter takes the common absence of features as an attribute of similarity. In chemoinformatics, it has been argued that, as many of the descriptor features are absent in the majority of molecules, the use of distance coefficients should be avoided during similarity and diversity analysis [34]. Table 1.1 lists the most common molecular coefficients used for various chemoinformatics analyses.

### **1.2.5 Comparison of drug-like and non drug-like compounds**

Many previous analyses have led to the widespread acceptance of concepts describing what makes a drug, to act as drug (drug-likeness) or a lead (lead-likeness) as well as what leads to a molecule not being a drug (non drug-likeness). More recently similarity to metabolites (metabolite-likeness) is increasingly being used as a drug design concept, to reduce attrition rates in drug discovery and development. Generally speaking, drug-like properties refer to the physicochemical, absorption, distribution, metabolism, excretion and toxicity properties of a molecule. Lacking drug-like properties often results in drug failures. A number of studies including simple counting schemes, shape analysis, statistical analysis and machine learning methods have been carried out, to characterize the properties of drug-like and non drug-like molecules

#### **1.2.5.1 Physicochemical properties analyses**

The pioneering work of Lipinski *et al.* [81], in recognizing and listing the important molecular descriptors that contribute to drug-likeness, is commendable. Lipinski's rule of five (Ro5) describes a set of simple criteria for bioavailability of drugs. The rule was derived from the analysis of 2,245 drugs obtained from the World Drug Index database, at that time. The assumption was that since these compounds have entered human clinical trials, they must therefore possess many of the desirable characteristics of drugs. The rule



states that poor absorption and permeation is more likely if compounds have more than 5 H-bond donors, 10 H-bond acceptors, a molecular weight of more than 500 and a logP value greater than 5. If a compound fails the Ro5 test, then there is a high probability that oral activity problems will occur. However, passing Ro5 is no guarantee that a compound is a drug. Since its publication, Ro5 has dominated drug design. The analysis carried out by Lesson and Davis [82] of the approved drugs released pre-Ro5 and post-Ro5 era gives an idea of the impact of Ro5 on drug discovery programmes. The publication of Ro5 spurred enormous interest in new approaches to classify drugs from non-drug molecules. However, there are exceptions to Ro5 and these mostly belong to a small number of therapeutic classes such as antibiotics, antifungal, vitamins, cardiac glycosides, macrolides and cyclic peptides.

Ghose *et al.* [83], extended Lipinski's original work by characterizing 6,304 compounds (taken from the Comprehensive Medicinal Chemistry database) and seven different subsets belonging to different classes of drug molecules based on computed physicochemical properties such as logP, molar refractivity, molecular weight, and number of atoms. The authors further characterized the occurrence of functional groups and important substructures in these compounds. They were able to establish qualifying ranges which cover more than 80% of the compounds in the set. Ranges were established for logP (−0.4 and 5.6), with an average value of 2.52, molecular weight (160 to 480), molecular refractivity (40 to 130) and for total number of atoms (20 to 70).

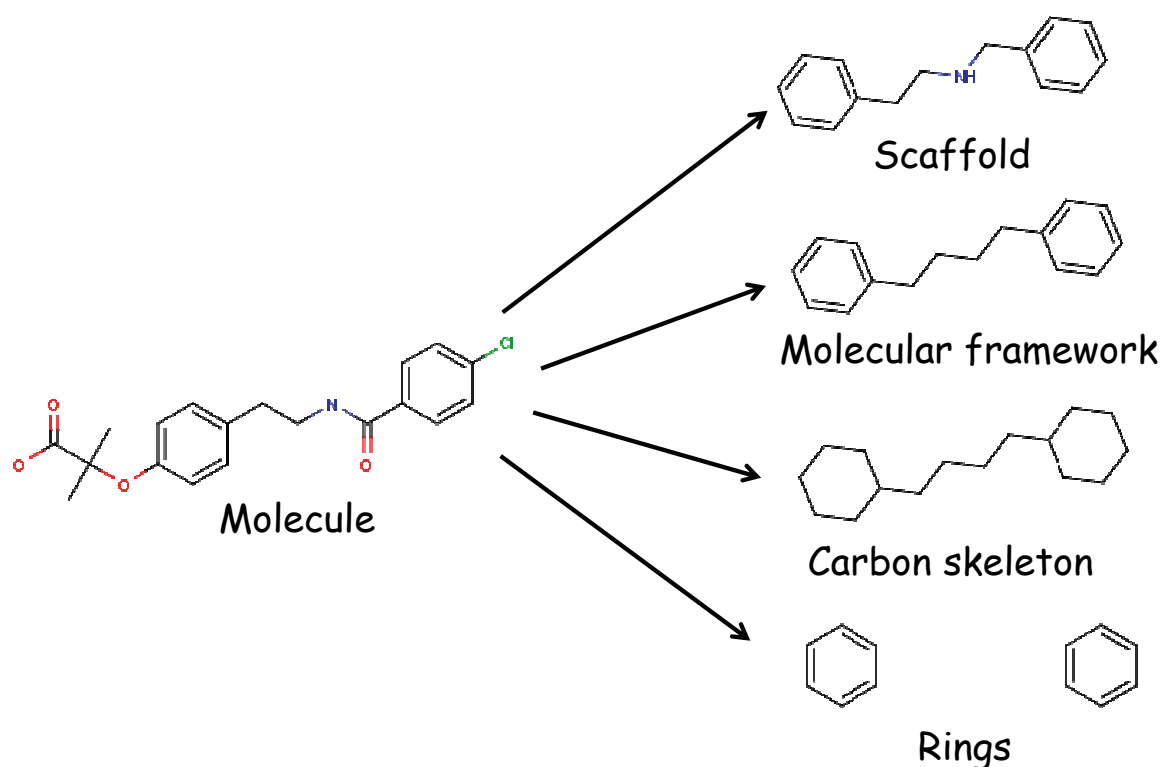
Other studies have led to the similar concept of lead-likeness [84], natural product-likeness [85], peptide-likeness [86] and more recently metabolite-likeness [87, 88]. A 'rule-of-three' (Ro3) [89] which states that molecular weight < 300, clogP < 3, hydrogen-bond donors < 3, hydrogen-bond acceptors < 3, rotatable bond count < 3 and polar surface area < 60, has been proposed to design lead fragments.

#### **1.2.5.2 Scaffold and molecular fragment data analysis**

Over the past decade, scaffold- and fragment-based analysis have been widely used in drug discovery [90-95]. Analysing the scaffold content of small molecule databases has led to the concepts of scaffold hopping [96] and privileged substructures [97]. As shown in Figure 1.8, it is possible to apply different levels of abstraction to the core molecule. Molecular graphs are decomposed by distinguishing ring assemblies, linkers connecting these ring assemblies, and side chains. The linker atoms form the direct path connecting



two rings while side chain atoms are any non-ring, non-linker atoms. Rings and linkers without side chains constitute molecular scaffolds. The molecular framework can be obtained from scaffolds by replacing all the heteroatoms with carbon atoms. Furthermore, bond types can be reduced to single bonds in order to obtain carbon skeletons, which represent the highest level of structural abstraction. Finally, ring assemblies can be obtained by breaking molecular scaffolds at linker atoms.



**Figure 1.8: Different levels of abstraction used in the literature for molecular scaffold analysis.** The scaffold is obtained by deleting all side chains. A molecular framework is obtained by replacing all heteroatoms by carbons from a scaffold while a carbon skeleton is generated by setting all bond orders to single bonds in molecular framework.

Similarly, there are different ways to get molecular fragments for example, by breaking the molecule at predefined bonds such as retro-synthetic criteria [98], random fragmentation approaches [99] and circular substructural fingerprints [42].

In one of the significant studies by Bemis and Murcko [90], 5,120 drugs (from the Comprehensive Medicinal Chemistry database) were analysed to identify common types of scaffolds. The compounds were fragmented into rings, linkers, frameworks and side chains. Using 2D topological graph-based molecular descriptors, the authors found 2,506 different frameworks for a set of 5,120 drug compounds, with the top 32 accounting for the



topologies of 50% of the database compounds. In conclusion, they suggested a skewed distribution of molecular frameworks in drugs. Recently, Franco *et al.* [91] examined the scaffold diversity of 16 datasets of active compounds, targeting five protein classes, using an entropy-based information metric. The authors concluded that the compounds targeted to the vascular endothelial growth factor receptor kinase, followed by compounds targeted to HIV reverse transcriptase and phosphodiesterase V, are maximally diverse. On the other hand, molecules in the glucocorticoid receptor, neuraminidase and glycogen phosphorylase datasets are least diverse. Wang *et al.* [100] structurally analysed molecular fragments in two drug datasets which they termed as “building block analysis”. The first dataset, ADDS, comprised 1,240 FDA-approved drugs, and the second drug dataset, EDDS, was a non-redundant collection of FDA-approved drugs and experimental drugs in different phases of clinical trials from several drug databases (6,932 entries). For each molecule in the two datasets, a brute force fragmentation method was applied to enumerate all possible fragments. Three kind of fragments were collected, namely, drug scaffolds, rings and small fragments. All the fragments were ranked according to the frequencies of occurrence in the dataset. The authors found that the top 50% of the fragments cover 52.6% and 48.6% of drugs in the ADDS and EDDS datasets, respectively. Hu and Bajorath [101] analyzed scaffolds and associated compound activity data in the public databases, namely ChEMBL and BindingDB, in order to compare their availability of target-selective scaffolds. The authors identified 143 scaffolds with varying complexity that are represented in multiple compounds and are promiscuous binders (i.e. compounds containing these scaffolds bind to multiple targets).

#### **1.2.5.3 Data mining and machine learning methods**

Data mining is the process of discovering the hidden predictive information and analysing it from different perspectives in order to summarize useful information. In other words, data mining is the process of finding correlations or patterns from large amount of data stored in data repositories. Data mining involves three major tasks. *Clustering* – is a task to group objects in such a manner that objects within the group possess high intra-group similarity while low inter-group similarity. *Classification and regression* – this involves learning the properties of known data and apply it to unknown data in order to classify objects in a classification task or to find a function that can model the data in a regression task. Common algorithms include decision trees, neural networks and support vector machines, and linear regression. *Association rule learning* – is a task to find correlations



and patterns among the attributes especially in a transaction database. In this section we describe common machine learning algorithms and association rule mining method.

Machine learning is a collection of methods that focuses on making machines learn from a given dataset and make predictions on unseen data. In recent years, machine learning methods have become increasingly popular in drug design, compared to conventional statistical and modelling methods. Machine learning methods can be divided into two categories, supervised and unsupervised learning. In supervised learning the data are labelled with predefined classes whereas in unsupervised learning, class labels of the data are unknown. If the target class is discrete, then the task is classification while if it is continuous, it poses a regression problem.

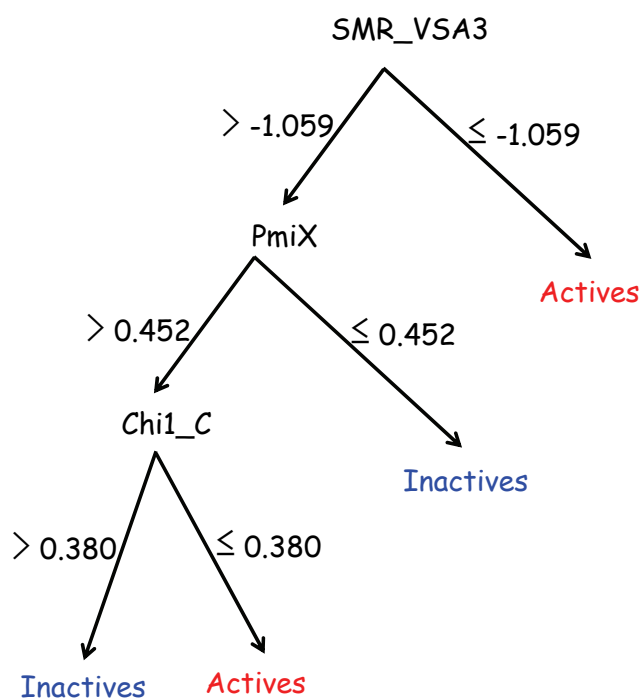
The goal of a supervised learning task is to optimize the mapping function that correlates the input descriptors or properties with the target variable. The function is optimized using the training data and validated using the validation set, which is withheld during training. In  $n$ -fold cross-validation, the training set is split into  $n$  subsets, and the model is built using  $n-1$  subsets while, the validation is done on the remaining subset. The process is repeated  $n$  times; therefore, each subset is used as the validation set at least once. In leave-one-out cross-validation,  $n$  equals the total number of objects (molecules) in the training set. Besides cross-validation, an external test dataset is also used which is independent of the training data and is not used for model building or optimization. The best known examples of supervised learning algorithms are decision trees, artificial neural networks, and support vector machines.

#### **1.2.5.3.1 Decision trees**

Decision trees (DT) are rule-based methods that classify patterns using a sequence of well defined rules [102]. The method uses a process called recursive partitioning [103], where each descriptor of the data is examined and ranked according to its ability to partition the remaining data. The best descriptors are selected to split the training samples into child nodes. The whole process is recursively repeated until some predefined completion criterion is met. The trained classifier has a tree-like structure and the nodes without children are called leaf nodes, while the others are internal nodes. An example of a decision tree network is shown below in Figure 1.9.



Wagener and Geerestein [104] successfully employed decision trees to discriminate between potential drugs and non-drugs, obtained from the Available Chemical Directory and the World Drug Index. They found that 75% of all drugs can be predicted based on the occurrence of six chemical groups (hydroxyl, tertiary or secondary amino, carboxyl, phenol and enol groups). Likewise, the majority of unsuitable compounds can be ruled out from further analysis based on the presence of specific chemical groups that result in a substance being reactive, toxic or difficult to synthesize. In addition, they found that non-drugs are mostly aromatic and characterized by a low content of functional groups, except halogens. Rusinko *et al.* [105] used recursive partitioning for the analysis of structure-activity data in diverse, large datasets. The authors created a program called SCAM (Statistical Classification of Activities of Molecules), which can be used to analyse large numbers of binary descriptors and partition the data into activity classes.



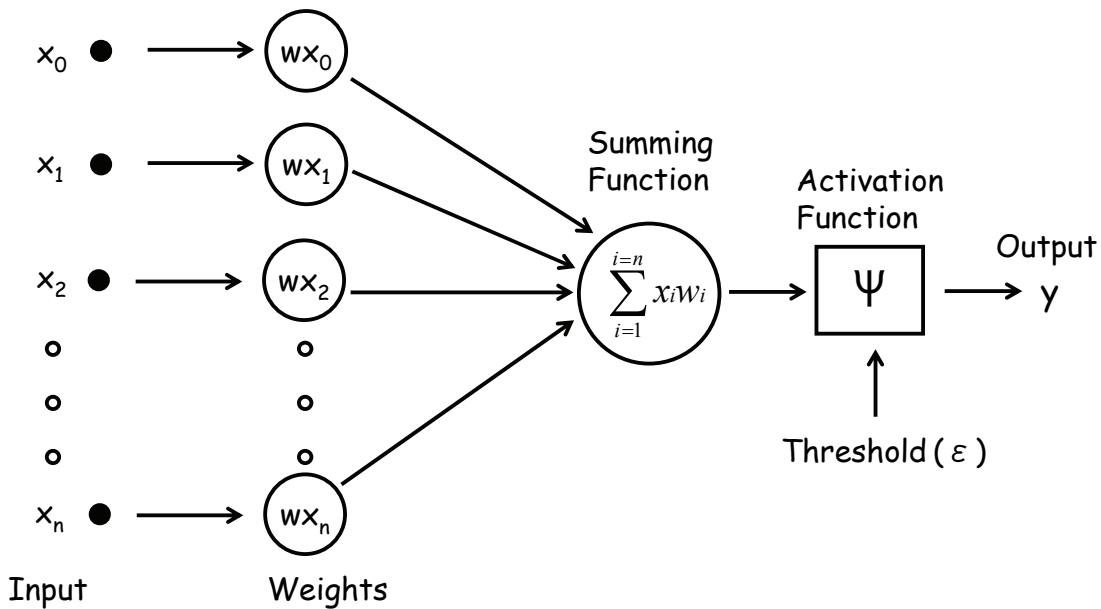
**Figure 1.9: A pictorial representation of a decision tree.** As shown in the decision tree, each internal node has a splitting predicate, with binary predicates being most common.

#### 1.2.5.3.2 Artificial neural networks

The human brain is composed of 100 billion basic units (cells) called *neurons*. An artificial neural network (ANN) is a system inspired from the operation of biological *neurons*. The neuron has three main components a *cell body*, branching extensions called *dendrites*, for receiving inputs and an *axon*, that carries an the neuron's output to other neurons. The *axon* of each neuron is connected to the *dendrites* of several others *neurons* via *synapses*



and communicates with them through electrochemical signals. The *neuron* continuously receives signals and sums up the all input signals to evaluate them against a threshold value. If the summation is greater than a threshold, a *neuron* is activated and *fires* an electrochemical signal, which generates a voltage and propagates the signal to other *neurons*. Emulating the design of biological neurons, ANN works in a similar manner. The *synapses* of the biological neuron are modeled as weights in ANN. A positive weight designates excitatory connection, while a negative weight corresponds to an inhibitory connection. All inputs are summed altogether and modified by the weights. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between -1 and 1, or it could be 0 and 1 as shown in Figure 1.10.



**Figure 1.10: An example of the basic ANN derived by emulating biological neurons.** Each input signal is associated with its own weight, which can be positive or negative. All the inputs are summed using a summing function and supplied to the activation unit. If the summation is greater than the threshold, then the output is 1, otherwise 0.

There are different types of neural networks available.

### 1. Feed forward network

The simplest type of neural networks is feed forward networks and involves a unidirectional flow of information. The data processing can extend over several units, but no feedback connections are present. There are no cycles or loops involved, so that information flow is from input to hidden layers to output.



## **2. Recurrent neural networks**

A feed forward ANN propagates data linearly whereas; recurrent ANN can propagate data from later processing stages to earlier stages.

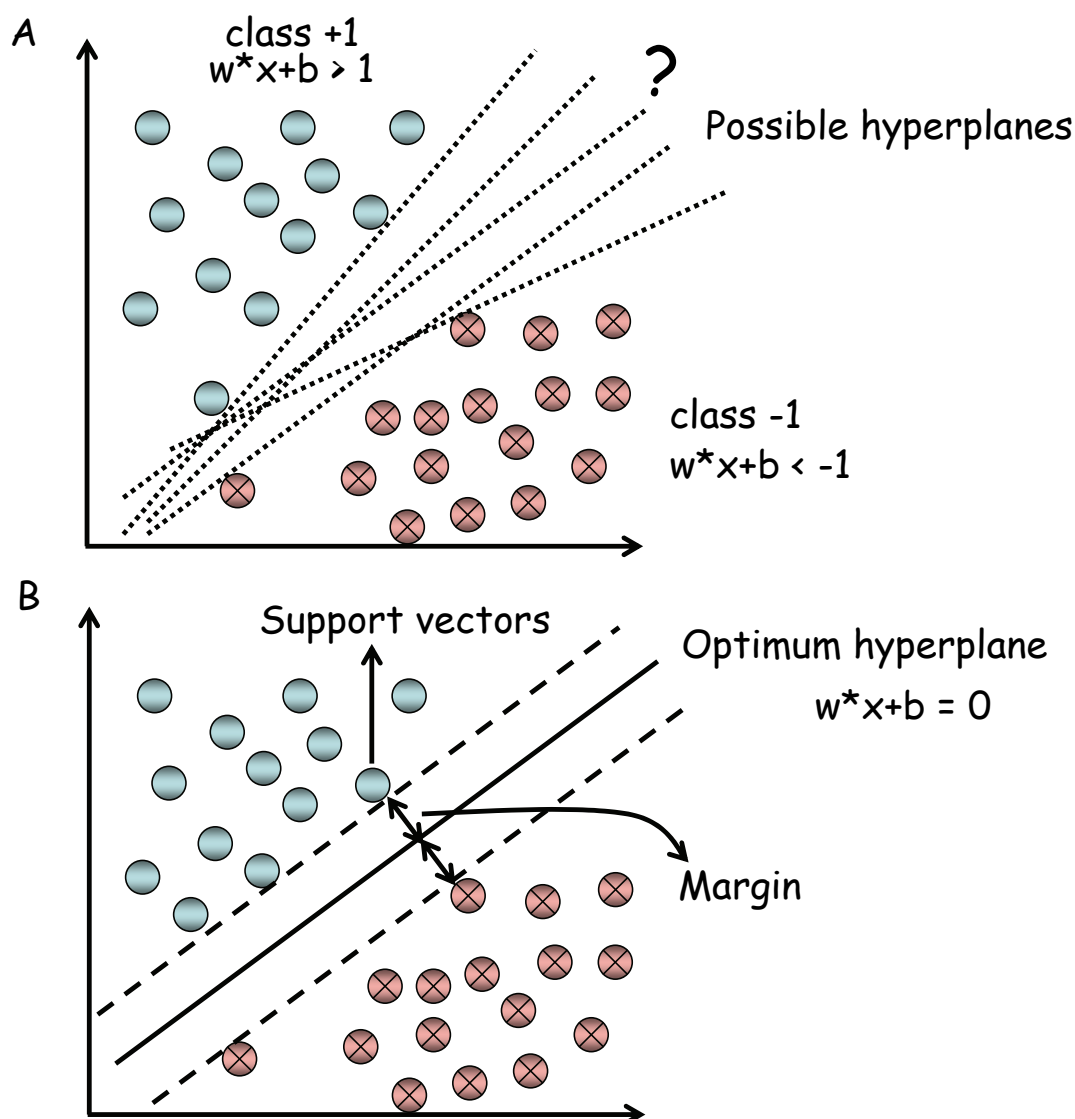
## **3. Kohonen self organizing maps**

Kohonen self-organising maps (SOMs) were invented by Teuvo Kohonen [106] and do not require a target output to be specified. SOMs provide an elegant way to represent multidimensional data in lower dimensions - usually one or two dimensions. In addition, SOM creates a network that stores information in such a way that, while reducing the data dimensionality, there is minimum loss of information within the training set.

More extensive introduction to the theory of neural networks and applications in chemoinformatics, QSAR and drug design are also available [107, 108]. Anzali et al. [109] have discussed two principal types of ANN that have application in combinatorial chemistry: SOMs and feed-forward neural networks. Kohonen maps, with their ability to represent high dimensional data in lower dimensions, can be used as a clustering tool, complementing the classical clustering techniques. Feed-forward networks can be used for classification of compounds or prediction of specific compound properties. A comparative study by Ajay et al. [110] that investigated the predictive performance of Bayesian neural networks and DT for classifying drugs and non-drug molecules revealed that Bayesian neural networks performed well in all instances, compared to decision trees. In their study, the training and test sets consisted of 3,500 and 2,000 compounds each, from the Comprehensive Medicinal Chemistry (CMC) and Available Chemicals Directory (ACD) databases, respectively; and each compound was described by an appropriate vector representation. The authors were able to classify 90% of the compounds in CMC and ACD and applied this method to classify 80% of the compounds from the MDL Drug Data Report (MDDR) database.

Often, ANN provides higher accuracy than DT; nevertheless, ANN approaches suffer from problems such as over-fitting the training data, lack of reproducibility of results and the lack of information regarding the classification produced. On the other hand, DT may also suffer from the over-fitting problem but provides an ample amount of information on the splitting criteria, in the form of predictive rules.





**Figure 1.11: Separation of data by Support Vector Machine hyperplanes.** (A) There are  $n$  number of possible hyperplanes that can correctly classify the data. (B) The SVM algorithm seeks to maximise the margin around a hyperplane (optimum hyperplane) that separates a positive class (circles) from a negative class (in crossed circles).

### 1.2.5.3.3 Support vector machines

Support vector machines (SVM) are a new type of machine learning algorithm, based on the structural risk minimization principle i.e. they search for the hypothesis with the lowest error on the training set [111, 112]. Recently, SVMs have been extensively applied in chemoinformatics, due to their robustness and ability to classify objects into two classes, as a function of their features [113-115]. Features encoding 1D, 2D and 3D properties of molecules, such as molecular weight, solvent accessibility, volume, charge, surface area, polarity and hydrophobicity are used to construct specific feature vectors that serve as



input to SVM. Binary classification of data (e.g. active, inactive or benign, malignant) within a linear SVM is done by separating the data optimally into categories, by constructing a hyperplane that divides the descriptor space into two parts. Actives lie in the positive half of the hyperplane, with inactive compounds in the negative half. In fact, a variety of different hyperplanes exist that may separate the data correctly as shown in Figure 1.11.

The margin between actives and inactives is maximized after mapping input vectors into a high dimensional feature space. Many studies in the past have shown SVM to be one of the best methods for correctly classifying molecules [116-119]. In one of the comparative studies where the SVM method was compared to other 16 classification methods and 9 regression methods, it was found that although SVM performed well in all the cases, the other methods also proved to be quite competitive [119]. Burbidge *et al.* [118] carried out a classification study that involved predication of the inhibition of dihydrofolate reductase by pyrimidines. The authors compared SVM with ANN and decision trees. They found that SVM outperformed most of the methods except a manually capacity-controlled neural network, although it took considerably longer to train. Nevertheless, SVM remains the most advanced machine learning method currently available.

#### 1.2.5.3.4 Association analysis

Association analysis [120] is an important data mining technique to discover hidden relationship among items and transactions. It is a supervised learning technique in the sense that we feed the association algorithm with a training data set (as called Experience  $e$  in machine learning context) to formulate hypothesis ( $h$ ). A typical and widely used example of association rule mining is “market basket analysis”. In retail stores data consists of large number of transaction records. Each record contains the information of all the goods purchased by a customer as a part of single bill. This kind of data is called “market basket data” and the analysis is termed as “market basket analysis”. In market basket analysis different buying habits of customers are identified and analysed to find association among items purchased by customers. For example customer who buy pencil are more likely to buy eraser. These kinds of patterns can be identified using association rules. Other applications include in the field of bioinformatics [121], graph mining [122] business intelligence [123], document analysis [124] and weblog mining [125].



An association rule is made of two parts, an antecedent (head) and a consequent (body):

**X** (buys, milk) antecedent  $\rightarrow$  **X** (buys, bread), [support =0.75%, confidence =50%]

where antecedent = Milk and consequent = Bread.

The above hypothetical rule can be interpreted as “Purchase of milk implies purchase of bread”. To measure the extent of implication, three measures are most commonly used. The first one is called *support* of the rule. The *support* is simply the number of transactions that include all items in the antecedent and consequent parts of the rule. The *support* is sometimes expressed as a percentage of the total number of records in the database. For example, if 75 transactions involve the above rule, then the support value is 0.75 which indicates the rule is significant. The second measure is the *confidence*. A *confidence* measure quantifies the confidence as a ratio of number of transaction holding this rule valid against the number of transactions involving this rule. Higher the value, more reliable is the rule. The third measure is the *lift* and it indicates the strength of an association rule over the random co-occurrences of the antecedent and the consequent, given their individual support. It can be calculated as  $lift = support\ (rule) / (support\ (antecedent) * support\ (consequent))$ . If an association rule has the *lift* less than 1, it suggests that the presence of the antecedent, causes a reduction in the probability of purchase of the consequent, compared to random chance and *vice versa*.

### 1.3 Databases and resources available

Availability of quality data is a vital first step towards any analysis. Unfortunately, public databases, large collaborative efforts to annotate the small molecules and analysis software are relatively scarce in chemical research [126], until 2005. In an analogy to the related field bioinformatics, the chemoinformatics equivalent of GenBank and BLAST had not been created in 2006 [127]. However, now there are increasing numbers of attempts to address the issue [128] and quite a few public databases have become available in chemoinformatics. In this section, the various public databases relevant to drug discovery are discussed. The most important and commonly used databases and resources for the chemoinformatics analysis are presented in Table 1.2 (current as of 15 March 2011). For a summary of first five databases, please refer to Publication 2, in this thesis. A brief description of the remaining databases can be found below.



### 1.3.1 ZINC

ZINC [129] is a public database for commercially available compounds for virtual screening. ZINC contains over 13 million compounds in 3D format, suitable for docking studies. A web-based query tool can be used for browsing using ZINC identifier or molecular properties such as xlogP, net charge, molecular weight and the number of rotatable bonds. 2D similarity and substructure searches are also possible with user specified structures. Various pre-compiled subsets such as “drug-like”, “lead-like” and “fragment-like” are particularly useful. Furthermore, users can customize their own subsets for download.

**Table 1.2: List of public databases and resources used in chemoinformatics.**

<i>Name</i>	<i>Homepage</i>	<i>Number of compounds</i>	<i>Data format<sup>#</sup></i>	<i>Data type</i>
PubChem [130]	pubchem.ncbi.nlm.nih.gov	>50,000,000	SDF	Small molecules
ChEBI [131]	ebi.ac.uk/chebi	19,000	mol	Biologically relevant molecules
ChemBank [132]	chembank.broad.harvard.edu	>800,000	Text	Small molecules
ChemIDplus [133]	chem.sis.nlm.nih.gov/chemidplus	>370,000	mol	Small molecules
ChemDB [127]	cdb.ics.uci.edu/index.htm	>5,000,000	SMILES, SDF	Small molecules
ZINC [129]	zinc.docking.org	>13,000,000	SMILES, SDF or mol	Small molecules
KEGG ligand [134]	genome.jp/kegg/ligand.html	16,948	SDF	Small molecules
DUD dataset [135]	dud.docking.org	98,266	SDF	Small molecules
Ligand.Info [136]	ligand.info	1,159,274	SDF	Small molecules



<i>Name</i>	<i>Homepage</i>	<i>Number of compounds</i>	<i>Data format<sup>#</sup></i>	<i>Data type</i>
NCI open [137]	cactus.nci.nih.gov/ncidb2	260,071	SDF	Toxicity
SuperToxic [138]	bioinf-services.charite.de/supertoxic	>60,000	mol	Toxicity
CPDB [139]	epa.gov/NCCT/dsstox/index.html	>1,500	SDF	Toxicity
DrugBank [140]	www.drugbank.ca	>6,827	SDF	Drugs
HMDB [141]	www.hmdb.ca	>7,900	SDF	Human metabolites
BindingDB [142]	bindingdb.org/	284,206	SDF	Small molecules
TTD [143]	bidd.nus.edu.sg/group/ttd/ttd.asp	5,124	SDF, mol	Drugs, Protein targets
UniProt [144]	uniprot.org/	—	—	Protein targets
SuperTarget and Matador [145]	bioinf-tomcat.charite.de/supertarget/	1500	mol	Drugs, Protein targets
ChEMBL [146]	ebi.ac.uk/chembl/db/	>600,000	SDF	Small molecules
PharmGKB [147]	pharmgkb.org	>3,100	Text	Drugs

### 1.3.2 Kyoto Encyclopaedia of Genes and Genomes (KEGG) Ligand Database

KEGG Ligand [134] is a composite, public database and contain chemical compounds that are relevant to life. It is made up six sub-datasets: COMPOUND comprising small



molecules and metabolites; REACTION, the collection of substrate-product and other metabolic reactions; ENZYME representing the set of enzyme molecules; DRUG dataset for drug collection including different salt forms and drug carriers; GLYCAN, a collection for experimentally determined glycan structures and RPAIR consisting of reactant pair alignments. Currently, the database contains 16,948 entries in the COMPOUND dataset, 8,451 reactions and 5,342 enzymes, 9,724 entries in drug dataset and 10,978 glycan structures.

### **1.3.3 Directory of useful decoys (DUD) dataset**

The DUD dataset [135] is derived from ZINC database to benchmark docking algorithms by providing challenging decoys. The database spans 40 different receptors and contains a total of 2,950 compounds active against those receptors. For each active compound, a set of 36 drug-like molecules were chosen from ZINC database to serve as decoys. The decoys have similar physicochemical properties (molecular weight, hydrophobicity, number of hydrogen donor/acceptors) but different topology leading to the database of 98,266 compounds. The actives included for each target comprise as few as 10–20 ligands in some cases, up to as many as a few hundred, in others.

### **1.3.4 Ligand.Info**

Ligand.Info [136] is a collection of various publicly available databases of small molecules such as ChemBank (2,344 entries), KEGG ligands (10,005 entries), AsinexLtd (348,276 entries), ChemPDB (4,009 entries), Anti-HIV NCI (42,689 entries) and TimTec (7,500 entries) subset. The database contains over a million entries. The compounds are present in 3D format and contain bioactivity information if possible. Some molecules have additional information about FDA drug approval status or about anti-HIV activity. Ligand.Info allows interactive clustering and searching of similar molecules using a Java-based tool. The whole database or individual datasets can be downloaded in structure data (SD) format .

### **1.3.5 National Cancer Institute (NCI) open**

The NCI open database [137] is a public part of the NCI database that was built and is maintained by the Developmental Therapeutics Program Division of Cancer Treatment, National Cancer Institute, USA. The current size of the NCI database is around 500,000 entries. The public part (NCI open database) contains almost half the compounds



(260,071). The data is available for download in SD format. A web-based graphical user interface, called Enhanced NCI Database Browser, can be used to browse the database. It is also possible to carry out similarity and substructure searches, beside searching the database using the unique database identifies (NSC ID), physicochemical properties or functional groups.

### **1.3.6 SuperToxic**

The SuperToxic database [138] provides information on toxins obtained from different sources (such as animals, plants, synthetics). Besides providing the chemical properties of the searched toxins, it also presents the user with information on commercial availability of toxins. The data in the SuperToxic database is cross-linked to various other external databases, such as the Protein Data Bank (PDB), UniProt and KEGG, to allow easy identification of targets and pathways linked with the toxin. Currently the database contains 60,000 compounds with structural information. The database can be browsed using the alphabetic listing or can be searched by various search techniques, including structure and substructure search, property search using molecular properties e.g. logP, hydrogen bond acceptors/donors and molecular weight. The database can also be queried using name of the toxin, the Chemical Abstract Service Reference Number (CASRN) and measured values of toxicity.

### **1.3.7 The Carcinogenic Potency Database (CPDB)**

The CPDB database is a unique and widely recognized resource of the results of 6,540 chronic, long-term animal cancer tests on 1,547 chemicals. The CPDB provides easy access to the bioassay literature, with qualitative and quantitative analyses of both positive and negative experiments. The results of each experiment include a range of information that is important in the interpretation of bioassays: species, strain, and sex of test animal, route of administration, duration of dosing, average daily dose-rate in mg/kg body weight/day, target organ, tumor type, carcinogenic potency (TD<sub>50</sub>) and its statistical significance and literature citation. The result of carcinogenicity bioassays are represented by TD<sub>50</sub> values (the dose at which tumorogenesis was found in 50% of the tested animals) of each species (rats and mice). The data is available for download in SD format.



### 1.3.8 DrugBank

The Drugbank [140] database is one of the largest, richly annotated and blended resource of chemoinformatics and bioinformatics datasets. It contains detailed chemical, medical and biological information on over 6,827 drugs including more than 1,431 FDA-approved small molecule drugs, 133 FDA approved biotech (protein/peptide drugs), 83 nutraceuticals and 5,212 experimental drugs. Furthermore, around 4,481 non-redundant protein (drug-targets, enzymes) sequences are available for download. Each drug entry in the database has more than 150 data fields, with first half of the information dedicated to drug-chemical data and the remaining half on drug-target or protein data.

The database supports several search methods like Boolean text search, structure search, local BLAST type sequence search for drug-targets and a relational data extractor tool. In the data extractor tool, various drug fields can be turned on or off during the search and the output can be obtained in different formats. The database can also be searched for similar structures which can be a useful tool for virtual screening programs. The SD format files of drugs and corresponding drug-target sequence files are available for download for seven different categories, namely approved drugs, small molecule drugs, nutraceutical drugs, experimental, biotech, withdrawn and illicit drugs.

### 1.3.9 Human Metabolome database (HMDB)

The Human Metabolome database is a comprehensive, organism specific, highly annotated, and freely available electronic dataset which contains spectroscopic, quantitative, analytic and molecular scale information on human metabolites. The database is intended to contain three types of data namely chemical data, clinical data, and molecular biology/biochemistry data.

Currently, the database holds (>7,900) metabolites and approximately 7,200 protein (and DNA) sequences are linked to these metabolite entries. The protein sequences can be downloaded from the website in FASTA format and the molecular structures are available in SD format. Clinical and biochemical information related to the metabolites are present in a MetaboCard Flat Files which nearly contains 110 data fields with 66% devoted to chemical/clinical data while the rest dedicated to enzymatic or biochemical data. The database can be browsed through variety of methods which include search by pathways, diseases, chemical substructure, text, sequences or chemical classes (inbuilt chemical



ontology). The database also supports spectral search like NMR, MS/MS and GC/MS search.

### **1.3.10 BindingDB**

BindingDB is a publicly accessible small molecule-protein interaction database of experimentally determined binding affinities. Currently the database comprises of 648,915 protein–ligand complexes, for 5,662 protein targets and approximately 284,206 small molecule ligands. The data in BindingDB is collected from various scientific literatures and the focus of data collection is proteins that are either drug-targets or potential drug-targets. The BindingDB website can be queried through various methods like search by chemical structure, substructure search and similarity search. Search by protein sequence, ligand and protein names, affinity ranges and molecular weight is also possible. The results can be downloadable in SD format. The structural data in BindingDB is cross linked to PDB and chemical and sequence data to the literature sources. Virtual screening by SVM, Binary kernel discrimination and maximum similarity are also implemented.

### **1.3.11 Therapeutic Target Database (TTD)**

TTD is a molecular target database that provides comprehensive information about the known and explored therapeutic protein and nucleic acid targets published in the literature. In addition, the database also provides information on targeted disease conditions, pathway information and the corresponding drugs/ligands for each of these targets. The database is searchable by target name, drug/ligand name disease name, drug/ligand function or drug therapeutic classification. TTD is cross-linked to various other databases that contain information about the function, sequence, 3D structure, ligand binding properties, enzyme nomenclature and related literature of each target. The database currently contains 1,906 targets (358 are successful, 251 in clinical trials, 1,254 are research targets) and 5,124 drugs/ligands (1,511 are approved, 1,118 in clinical trials and 2,331 are experimental drugs).

### **1.3.12 UniProt database**

UniProt [144] is a comprehensive, high-quality and freely accessible resource of protein sequence that provides a non-redundant high level of annotation. It is a unified knowledgebase that combines databases such as Swiss-Prot (a manually curated and



annotated protein sequence database), TrEMBL (Translated EMBL, a computer-annotated supplement to Swiss-Prot), UniRef (a database of protein sequence clusters, developed to speed up sequence similarity searches) and UniPrac (an archive for protein sequences, used to keep track of protein sequence identifiers and changes in protein sequences). The primary object of this database is proteins, for which sequence data, references and the taxonomic data is provided. Currently, the UniProt knowledgebase holds over 13 million sequences and is updated regularly.

### **1.3.13 SuperTarget and Matador**

SuperTarget [145] is drug-target interaction database and currently includes more than 2,500 target proteins which are annotated for about 7,300 drug-target interactions to 1,500 drugs. The database also includes information on adverse drug effects, drug metabolism, pathways, gene ontology terms and sequence comparison data of target proteins.

Matador (Manually Annotated Targets and Drugs Online Resource), which is a highly annotated subset of the SuperTarget database, lists 775 drugs and contains additional binding and indirect interaction information. The database can be searched using drug name, target name, The Chemical Abstract Service (CAS) Number, PDB Ligand identifier, KEGG Human Pathways and WHO developed Anatomical Therapeutic Chemical Classification System codes (ATC-codes). Complex queries with Boolean operators can also be easily performed. SuperTarget is also cross-referenced to other databases like KEGG for pathway information, SuperDrug for information on similar drugs and SuperLigand on drug-like ligand information. Tanimoto score is used for similarity calculation and fingerprints calculated from (Chemistry Development Kit) CDK allow for fast identification of drugs that may interact with same target protein.

### **1.3.14 ChEMBL**

The ChEMBL database [146] is a freely available chemogenomics data resource of bioactive drug-like compounds. The data is obtained from primary scientific literature and covers a significant portion of known targets, small molecules and their SAR information. The database can be searched for targets by using the text search for protein names and sequence search using BLAST which can also identify related proteins. The compounds can be searched using the compound name, SMILES, or chemical identifier. Substructure



or similarity search can also be employed for compound searching. Currently the database contains 8,091 targets, 658,075 distinct compounds and over 3 million activity records and is updated monthly.

### **1.3.15 PharmGKB**

The PharmGKB [147] is central repository on pharmacogenomics and pharmacogenetics data and promotes research into the relationship between human genotype, phenotype and drugs. Its main objective is to aid the researchers in understanding the genetic basis for variation in response to drugs. In addition to the data on the gene-drug relationship, the database also contains information on gene variations, genomics, drug-action and pathways. PharmGKB contains highly curated pathways documenting genes involved in pharmacodynamics and pharmacokinetics of selected drugs. The database currently holds information on (>3100) drugs, (>3200) diseases and (>27000) genes and 72 pathways. It also has the detailed information on gene variants (SNP data) affecting drug metabolism.

While a comprehensive overview of chemoinformatics and its applications in ligand based drug design has been presented above, an in-depth explanation pertaining to virtual library design is presented in publication 1 along with an overview of molecular similarity and diversity analyses. Following this, previous work on ligand based virtual screening methods including successful application of machine learning approaches [104, 110] has been reviewed in publication 2.



Due to copyright restrictions, from page 39 – 71, the following articles have been omitted from the thesis. Please refer to the following citations for details.

**Khanna V**, Ranganathan S: Molecular similarity and diversity approaches in drug discovery. *Drug Development Research*, 2010, 72(1), 74-84.

**Khanna V**, Ranganathan S: *In Silico* Methods for the Analysis of Metabolites and Drug Molecules, In *Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications*, eds. M. Elloumi and A.Y. Zomaya, Wiley, pp.363-383







## 1.4 Objectives

Currently, there is an immense need to develop quick and efficient methods for identification and characterization of biologically relevant molecules in order to cut down the time involved in a typical drug discovery pipeline. Although information on the traditional use of plants is available through databases and monographs, relatively little information is available for Australia Aboriginal medicinal plants [148]. Over 25% of the currently used modern medicines are a direct result of ethnopharmacological studies [149]. Natural products can thus play a major role in discovering biologically relevant molecules. Given the exponential expansion of PubChem [130], experimental screening of potential ligands using traditional methods would be an impossible and irrational exercise. However, recently developed complementary computer based drug design approaches such as similarity-based virtual screenings have looked promising. Similarity-based virtual screening has been particularly useful when information regarding the biological target is scarce. Where the information is available, structure-based screening methods such as docking have been successfully employed [150].

Since the publication by Lipinski et al. [81], a number of studies have attempted to identify the chemical space occupied by drug-like molecules. These include comparing physicochemical properties, scaffold and fragment data analysis, machine learning and building statistical models. Until now only a few studies have used human metabolites and toxics compounds in their analysis. This might be primarily due to the limited availability of high quality, public human metabolome and toxic data. Furthermore, only a couple of studies involving scaffold and fragment data analysis have taken into consideration the influence of co-occurring fragments in virtual screening. Also, little work has been done to compare public databases of bioactive compounds often used in various chemoinformatics analysis.

The above objectives were sub-divided into specific aims, described in the following sections and addressed in detail in five publications presented in this thesis:

- I. Review the current status of chemoinformatics and its potential application in drug discovery and development (Publications 1 and 2).



- II. To develop a web-based chemoinformatics module for Customary Medicinal Knowledgebase (CMKb database) in order to store, analyse and visualize natural products available from Australian Aboriginal medicinal plants (Publication 3).
- III. Preliminary comparison of the distribution of physicochemical properties in current drugs, human metabolites and toxic compounds so as to examine the similarity of current drugs with human metabolites and toxic molecules (Publication 4).
- IV. Detailed analysis of publicly available chemoinformatics databases for drugs, human metabolites, toxics, natural products and current lead compounds in order to confirm our preliminary findings and to identify frequently occurring scaffolds or fragments. In addition we also wanted to study the possible pairs of co-occurring fragments in these datasets (Publication 5).
- V. Virtually screening for compounds active against parasitic nematodes using machine learning approaches and fragment co-occurrence data (Publication 6).



## Chapter 2: Methodology and Implementation

A list of the methods and analyses carried out during this study are provided in Table 2.1. The ensuing publications have also been listed and included in the relevant chapter.

**Table 2.1: Tools, resources and publications**

Methods/Applications	Chapter	Refer to publication
Molecular similarity and diversity approaches in drug discovery.	1	1
<i>In silico</i> methods for the analysis of metabolites and drug molecules.	1	2
CMKb: a web-based prototype for integrating Australian Aboriginal customary medicinal plant knowledge.	3	3
Physicochemical property space distribution among human metabolites, drugs and toxins.	4	4
Scaffold and fragment co-occurrence studies on datasets of biological interest.	5	5
<i>In silico</i> approach to screen compounds active against parasitic nematodes of major socio-economic importance.	6	6



## **Chapter 3: Development of the CMKb chemoinformatics module to digitize and store chemical information.**

### **3.1. Summary**

Indigenous medicinal plants are a major resource for safe alternative medicines and new drugs. Approximately 80% of the drugs derived from plants were discovered as a direct result of the study on ethno-medicinal aspects of plants used by humans [151]. Australian customary medicinal knowledge aggregates traditional (native) and contemporary (exotic) use of medicinal plants, which is unfortunately diminishing as elders die and is poorly documented. Systematic documentation of this knowledge can lead to novel drug discovery. Customary Medicinal Knowledgebase (CMKb) is a web-based relational database system to document multidisciplinary data including ethnobotany, taxonomy, biogeography, medicinal and chemical information. In order to digitize and store chemical information, obtained from medicinal plants used by Australian Aborigines, we developed the chemoinformatics analysis and visualization module for the Customary Medicinal Knowledgebase (<http://www.biolinfo.org/cmkb>).

Therefore, this paper presents CMKb, a relational database, with user-defined search capabilities. The chemoinformatics module comprises the chemical table which includes information on various physicochemical properties (molecular weight, logP, melting point, boiling point etc) of the molecule. It also stores the IUPAC name, common names, CAS number, SMILES, biological activity, reported literature (DSN) and external database links such as PubChem and ChEBI chemical identifier. In addition, the chemoinformatics module incorporates Jmol, a java applet for visualization and Marvin sketch applet for creating and editing chemical structures. Chemical information in CMKb can be queried based on the IUPAC name, the CAS number and the common name using queries. The view link provides platform-independent structure visualization using Jmol.



Research

Open Access

## CMKb: a web-based prototype for integrating Australian Aboriginal customary medicinal plant knowledge

Jitendra Gaikwad<sup>1,2</sup>, Varun Khanna<sup>1,2</sup>, Subramanyam Vemulpad<sup>3</sup>,  
Joanne Jamie<sup>1</sup>, Jim Kohen<sup>4</sup> and Shoba Ranganathan\*<sup>1,2,5</sup>

Address: <sup>1</sup>Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW 2109, Australia, <sup>2</sup>Australian Research Council (ARC) Centre of Excellence in Bioinformatics, Macquarie University, Sydney, NSW 2109, Australia, <sup>3</sup>Department of Health and Chiropractic, Macquarie University, Sydney, NSW 2109, Australia, <sup>4</sup>Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia and <sup>5</sup>Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597

Email: Jitendra Gaikwad - jgaikwad@cbms.mq.edu.au; Varun Khanna - vkhanna@cbms.mq.edu.au;  
Subramanyam Vemulpad - subramanyam.vemulpad@mq.edu.au; Joanne Jamie - joanne.jamie@mq.edu.au;  
Jim Kohen - jim.kohen@mq.edu.au; Shoba Ranganathan\* - shoba.ranganathan@mq.edu.au

\* Corresponding author

from Asia Pacific Bioinformatics Network (APBioNet) Seventh International Conference on Bioinformatics (InCoB2008)  
Taipei, Taiwan. 20–23 October 2008

Published: 12 December 2008

BMC Bioinformatics 2008, 9(Suppl 12):S25 doi:10.1186/1471-2105-9-S12-S25

This article is available from: <http://www.biomedcentral.com/1471-2105/9/S12/S25>

© 2008 Gaikwad et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** The customary medicinal plant knowledge possessed by the Australian Aboriginal people is a significant resource. Published information on it is scattered throughout the literature, in heterogeneous data formats, and is scattered among various Aboriginal communities across Australia, due to a multiplicity of languages. This ancient knowledge is at risk due to loss of biodiversity, cultural impact and the demise of many of its custodians. We have developed the Customary Medicinal Knowledgebase (CMKb), an integrated multidisciplinary resource, to document, conserve and disseminate this knowledge.

**Description:** CMKb is an online relational database for collating, disseminating, visualising and analysing initially public domain data on customary medicinal plants. The database stores information related to taxonomy, phytochemistry, biogeography, biological activities of customary medicinal plant species as well as images of individual species. The database can be accessed at <http://biolinfo.org/cmkb>. Known bioactive molecules are characterized within the chemoinformatics module of CMKb, with functions available for molecular editing and visualization.

**Conclusion:** CMKb has been developed as a prototype data resource for documenting, integrating, disseminating, analysing multidisciplinary customary medicinal plant data from Australia and to facilitate user-defined complex querying. Each species in CMKb is linked to online resources such as the Integrated Taxonomic Information System (ITIS), NCBI Taxonomy, Australia's SpeciesLinks-Integrated Botanical Information System (IBIS) and Google images. The bioactive compounds are linked to the PubChem database. Overall, CMKb serves as a single knowledgebase



for holistic plant-derived therapeutics and can be used as an information resource for biodiversity conservation, to lead discovery and conservation of customary medicinal knowledge.

---

## Background

Australia is among the 34 biodiversity hotspot countries in the world [1] endowed with unique endemic plant diversity. It is estimated that 85 percent of over 21,000 vascular plant species are endemic to Australia [2]. More than 40,000 years of Aboriginal inhabitation [3] has led to the use of medicinal plants from this vast bioresource for maintaining and treating health-related problems [4]. Aboriginal remedies vary between clans and in different parts of the country, with no single set of aboriginal medicines and remedies [5]. The indigenous knowledge has been passed on from one generation to the next orally through traditional songs, stories, poetry and legends [6]. Unfortunately, Aboriginal customary medicinal knowledge is poorly documented and is on the verge of being lost due to dislocation and the westernisation of the communities [7,8].

Documented Australian medicinal plant knowledge is in the main, fragmented, restricted to specific locales and of limited applicability, usually to pharmacology or phytochemistry. Several studies have focussed on the Northern Territory, where the use of medicinal plants has been documented, with limited data on chemical components and pharmacological assay work [9,10]. A database of plants used as bush foods and medicines by New South Wales Aboriginal communities comprises information largely obtained from published sources or early manuscripts [11], but does not include chemical or pharmacological data. The CSIRO Australian phytochemical database comprises a compendium of published work, searchable by plant and chemical names alone [12]. Thus, there is no single comprehensive inventory of Aboriginal medicinal plants available similar to initiatives such as Native American Ethnobotany database [13] and Prelude Medicinal Plants Database from Africa [14]. The available information in published literature is species-specific, scattered and in different formats, making data integration challenging.

Customary knowledge of medicinal plants and practices is a significant contributor to scientific research and development in pharmaceuticals, cosmetics, foodstuffs, agricultural products and a wide range of other biologically based products and processes [15]. Access to public domain information on Australian customary medicinal plants will advance research in bioinformatics, ethnobotany, taxonomy, biogeography and phytochemistry. Here, we report the development of a comprehensive knowledgebase for Australian customary medicinal plants,

CMKb. To the best of our knowledge, this is the first such knowledgebase of its kind.

## Construction and content

### System architecture

The goal was to design a database which could be flexible and could accommodate heterogeneous data from published literature or bibliographic search. CMKb is developed using MySQL 5 relational database [16] for systematic and efficient content management. The user-friendly interface, consisting of dynamic web pages, is developed using PHP 5 [17] for data visualisation and data management. The chemoinformatics module incorporates Jmol, a Java based applet program [18] for visualization and Marvin Sketch [19] for drawing and editing of chemical structures. The data is served using Apache web-server [20] (Figure 1).

### Construction method

Before developing the database schema, end user and data resource availability assessment was carried out. The assessment results showed that the potential end users range from members of Aboriginal communities to scientists with interests in ethnobotany, phytochemistry, biology and microbiology. The major data resource is the information collated from an exhaustive literature survey.

We have created a novel schema for integrating multidisciplinary information on medicinal plant species, such as taxonomy, habit and habitat, phytochemistry, bioactivity, biogeography, data sources, medicinal preparation methods and usage, community information, and images into CMKb. Since the species name is the fundamental biological descriptor [21], all the information is linked to the scientific name. Thus, the species information table is central to our schema, and is connected to the other tables (Figure 2). CMKb is designed with the possibility of future expansion including scaling to accommodate very large datasets, and the addition of other multidisciplinary components, described later.

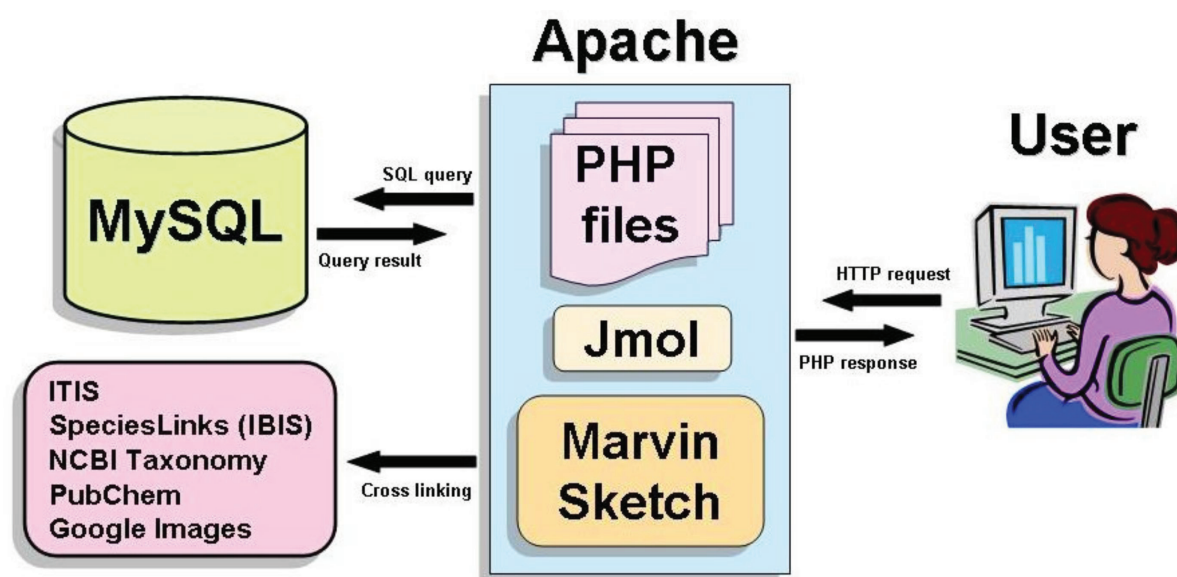
### Content of the database

Information related to medicinal plant species is stored in seven major tables (Figure 2) which are briefly described below. Mandatory information comprises the species name, the published reference and the medicinal use.

#### 1. Species information

Information related to customary medicinal plant species such as kingdom, family, scientific name, synonym, com-





**Figure 1**  
Schematic presentation of system architecture of CMKb.

mon name, native language name, habit and habitat as well as author citation, is stored in this table. This table is the hub to which all other tables are connected. The scientific name from this table is also used for cross-linking to external data portals, such as IBIS, ITIS and NCBI Taxonomy.

#### 2. Data Source Number (DSN)

Each published article in the literature used to collate and populate the database, is assigned a unique DSN identifier. The DSN table contains fields such as the title of the article, reference type (such as thesis, journal or book), names of authors and citation details.

#### 3. Medicinal information

Species-specific customary medicinal information such as the parts of the plant used, preparation method, taste, odour, colour, application and storage method, is collated in this table.

#### 4. Biological activity information

This table records the biological activity associated with the medicinal plant. The type of assay used to identify biological activity (such as antifungal, antiviral, antibacterial), the specific assay used, assay targets (such as cell line, enzyme and organism name) are recorded in this table.

#### 5. Chemical information

This table is used to store the chemical information and structure of bioactives derived from the medicinal plants such as IUPAC name, CAS number, PubChem [20] identifier, common chemical name, chemical structures in SMILES and MOL formats, biological activity related to that chemical compound, spectral data and other physical properties. The chemical structures are created locally using Marvin Sketch and are displayed using Jmol, a freely available Java applet. PubChem identifier stored in this table is used to link to PubChem database [22] from CMKb (Figure 3).

#### 6. Biogeography information

The biogeography table collates observational data of the species from the published literature such as locality name, latitude and longitude in decimal units, district/town, state and country.

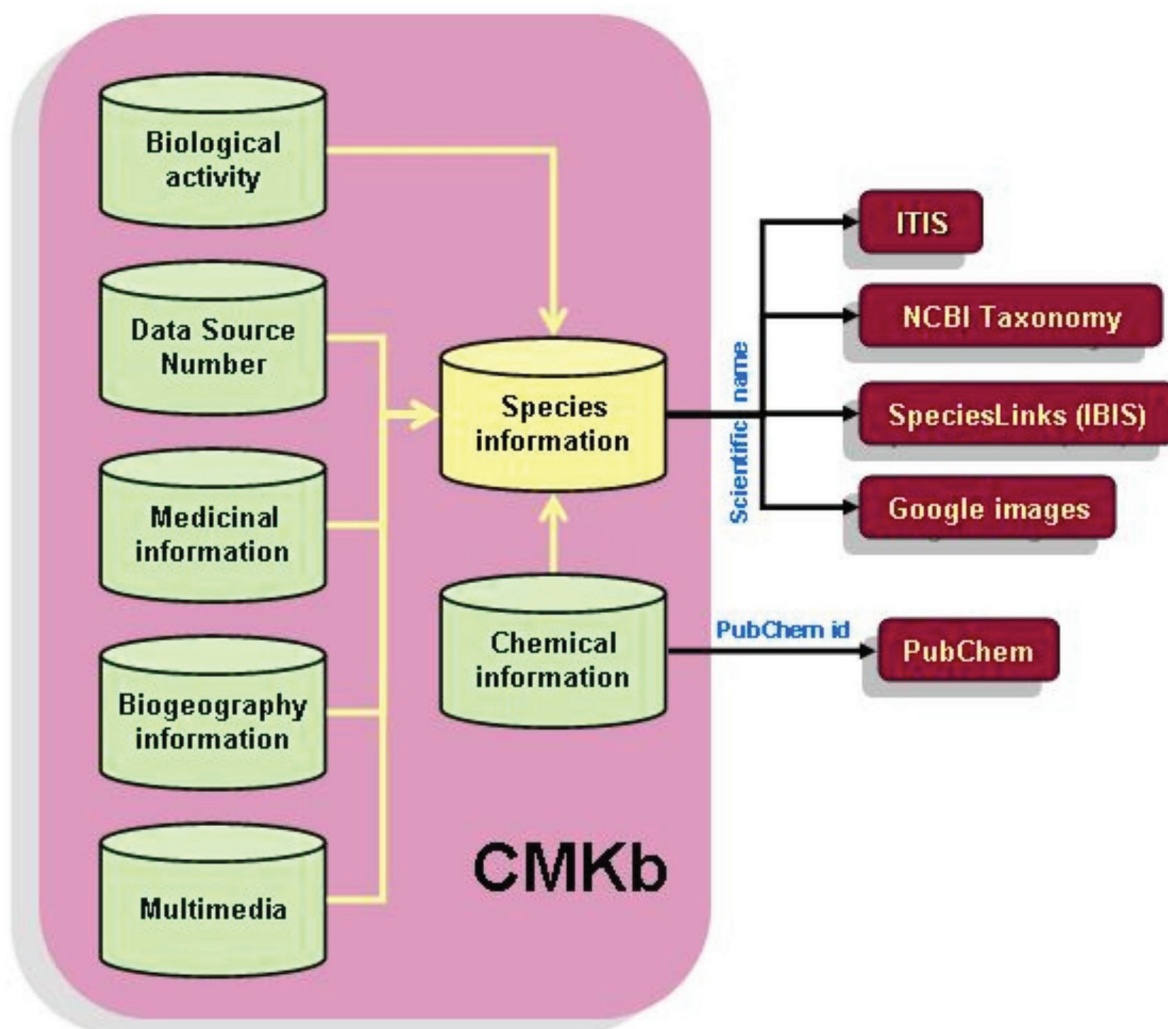
#### 7. Multimedia information

CMKb will also accept data in multimedia formats. This table is used to store multimedia information for each species, in the form of videos, drawings and photographs. Multimedia file formats such as jpeg, mpeg and avi can be uploaded to the database, with detailed text description.

#### Utility and discussion

CMKb provides a user friendly web interface for accessing and managing the customary medicinal plant data. The





**Figure 2**  
Dataflow in CMKb, showing external links.

database consists of three main modules: *Browse*, *Search* and *Data Management*. Links to these modules are provided as a menu on the LHS of the CMKb website, as "Browse," "Search" and "Login," respectively. A brief description of each module is given below.

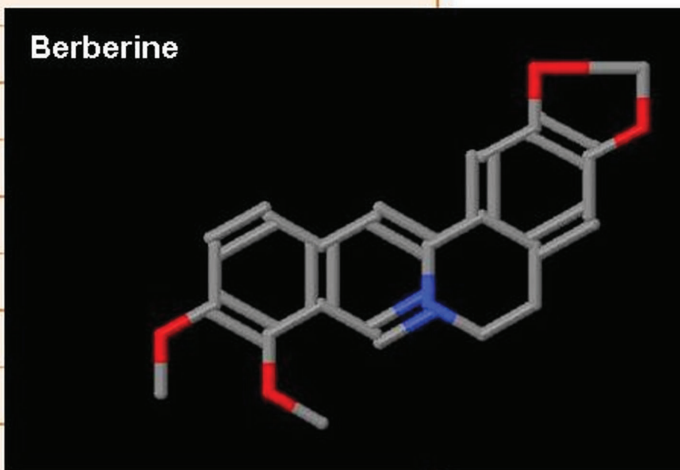
• **Browse module**

The database contents can be browsed (Figure 4a) using the alphabetical listing of scientific names and these are hyperlinked to a species list (Figure 4b), each of which is linked to a detailed information page.

Since CMKb is a web-based application, we have provided external links to other relevant global databases and data portals. Linking with other databases providing taxonomic, geospatial and molecular information, and search engines for images will help in data mining and facilitate the exploration of questions that, at present, cannot readily be answered [23] and would provide additional value to the information. Using the scientific name from CMKb we have provided external links to public domain data portals such as Integrated Botanical Information System (IBIS) [24] which provides links to a range of Australian data portals, Integrated Taxonomy Information System



<b>IUPAC/Chemical name</b>	Berberine
<b>Chemical common name</b>	Berberine <a href="#">View</a>
<b>CAS#</b>	633-66-9
<b>PubChemID</b>	<a href="#">2353</a>
<b>Formula</b>	C <sub>19</sub> H <sub>14</sub> NO <sub>4</sub> <sup>+</sup>
<b>SMILES</b>	
<b>InChI</b>	
<b>Melting point</b>	
<b>Boiling point</b>	
<b>Spectral data report</b>	
<b>Molecular weight</b>	336.36
<b>Biological activity</b>	Immunostimulatory, anti inflammatory, anti microbial, anti HIV
<b>Log P</b>	
<b>IC50 value</b>	
<b>Extraction method</b>	
<b>DSN</b>	<a href="#">44</a>



**Figure 3**  
Chemical information page with structure visualization.

(ITIS) [25], NCBI Taxonomy [26,27] and Google images [28] for species images.

#### • Search module

The database can be searched using its comprehensive search engine. The "Quick Search" option provides users with the facility to query the database by scientific name, species common name, native name, locality or chemical name using different logical parameters such as "contains", "begins with", "ends with" and "is" (Figure 5a).

For more complex queries, the "Advanced Search" option can be used, where the user can combine different search fields, using AND as the logical parameter (Figure 5b)

#### • Data management module

Efficient online content management is coordinated by CMKb's data management module, accessible to authorized users via the Login link. The data management module is provided with ADD, EDIT and DELETE functionality for managing data present in different tables.





**Figure 4**  
**Browsing the CMKb database.** a. Alphabetical listing of species in the Browse module, and b. a list of species starting with "V".

The overall contents of the database can be accessed from the "Content Summary" link.

## Conclusion

Customary Medicinal Knowledgebase (CMKb) is a prototype for collating, integrating, visualising, disseminating and analysing multidisciplinary public domain data on customary medicinal plants. It is a holistic knowledgebase with data on taxonomy, biogeography, ethnobotany, phytochemistry, and bioactivity of the customary medicinal plants used by the Australian Aboriginals. The goal of CMKb is to collate information from scientific publications which are peer reviewed along with documenting and conserving the dwindling customary medicinal plant knowledge. The data will be constantly scrutinised by the experts and will be updated accordingly. Overall, CMKb is developed as a single knowledgebase for holistic plant-derived therapeutic substances and can be used as an integrated resource by researchers, policy makers, students and Aboriginal communities. As the database grows CMKb can be used for research in areas such as Geographical Information System (GIS) studies, chemoinformatics and biodiversity informatics. Further, the goal is to help address global and national priorities of biodiversity conservation, better human health, and smart use of information using information technology.

## Availability and requirements

CMKb is freely available online at <http://biolinfo.org/cmkb/>

## List of abbreviations used

CAS: Chemical Abstracts Service; CMKb: Customary Medicinal Knowledgebase; CSIRO: Commonwealth Scientific and Industrial Research Organisation; DSN: Data Source Number; FDSN: Field Data Source Number; GIS: Geographical Information System; IBIS: Integrated Botanical Information System; ITIS: Integrated Taxonomic Information System; IUPAC: International Union of Pure and Applied Chemistry; JRE: Java Runtime Environment; LHS: Left Hand Side; NCBI: National Center for Biotechnology Information; SMILES: Simplified Molecular Input Line Entry Specification

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

SR, JK, JJ and SV conceived the database concept. JG developed and constructed the database. VK contributed to the web interface and developed the Chemical information module. JG and SR wrote the paper. All authors approved the manuscript and declare that there is no conflict of interest.



**a**

characters string for searching the knowledgebase.

**b**

Please select the community

Genus

Species

Locality

Chemical Name

Medicinal Uses

**Figure 5**  
User-defined querying facilities. a. Quick search and b. Advanced searching for expert users.

## Acknowledgements

JG and VK are grateful to Macquarie University for the award of MQRES PhD scholarships. We thank Doan Le, Moa Ek, David Harrington and Elsa Chacko (Macquarie University); Karen Wilson (Royal Botanic Gardens, Sydney, Australia); Vishwas Chavan (Global Biodiversity Information Facility, Denmark) and Greg Whitbread (Australian Centre for Plant Biodiversity Research) for their assistance and suggestions. This work was initially supported by a Macquarie University grant (MURDG A006661). Funding to pay the Open Access publication charges for this article was provided by Macquarie University.

This article has been published as part of *BMC Bioinformatics* Volume 9 Supplement 12, 2008: Asia Pacific Bioinformatics Network (APBioNet) Seventh International Conference on Bioinformatics (InCoB2008). The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2105/9?issue=S12>.

## References

1. **Biodiversity Hotspots** [<http://www.biodiversityhotspots.org>]
2. Mummary J, Hardy N: **Australia's Biodiversity: an overview of selected significant components.** *Biodiversity series – Information*

- about the conservation of Australia's biodiversity 2004): [<http://www.environment.gov.au/biodiversity/publications/series/paper2/index.html>].
3. Roberts-Thomson JM, Martinson JJ, Norwich JT, Harding RM, Clegg JB, Boettcher B: **An Ancient Common Origin of Aboriginal Australians and New Guinea Highlanders Is Supported by  $\alpha$ -Globin Haplotype Analysis.** *Am J Hum Genet* 1996, **58**:1017-1024.
4. Barr A, Chapman J, Smith N, Beveridge M, Knight T, Alexander V, Andrews M: **Traditional Bush Medicines: An Aboriginal Pharmacopoeia. Aboriginal Communities of the Northern Territory of Australia.** Richmond, Victoria: Greenhouse Publications; 1988.
5. **Traditional aboriginal Bush Medicine** [<http://www.aboriginalaronline.com/culture/medicine.php>]
6. Dayalan AM: **Traditional Aboriginal Medicine Practice in the Northern Territory.** Paper Presented at: International Symposium on Traditional Medicine: 11–13 September 2000; Awaji island Japan 2000:1-16 [[http://www.maningrida.com/mac/bwc/documents/traditional\\_aboriginal\\_medicine\\_practice.pdf](http://www.maningrida.com/mac/bwc/documents/traditional_aboriginal_medicine_practice.pdf)].
7. Smith NM: **Ethnobotanical field notes from the Northern Territory, Australia.** *J Adelaide Bot Gard* 1991, **14**:1-65.
8. Brouwer N, Liu Q, Harrington D, Kohen J, Vemulpad S, Jamie J, Randall M, Randall D: **An Ethnopharmacological Study of Medicinal Plants in New South Wales.** *Molecules* 2005, **10**:1252-1262.
9. Barr A, Chapman J, Smith N, Wightman G, Knight T, Mills L, Andrews M, Alexander V: **Traditional Aboriginal Medicines in the Northern Territory of Australia by Aboriginal Communities**



- of the Northern Territory. Darwin, Australia: Conservation Commission of the Northern Territory, 1993.
10. Levitt D: **Plants and people: Aboriginal uses of plants on Groote Eylandt.** Canberra: Australian Institute for Aboriginal Studies; 1981.
  11. Gott B: **NSWUSE – Database of NSW Plants Utilised by Aboriginals.** Lodged at: Australian Institute of Torres Strait Islander Studies (AIATSIS) Library; 1996.
  12. Collins DJ, Culvenor CCJ: **PhytoChem Australia: A database of Australian plant chemistry 1940–2000.** Collingwood, Australia: CSIRO Publishing; 2003.
  13. **Native American Ethnobotany** [<http://herb.umd.umich.edu/>]
  14. **Prelude Medicinal Plant Database** [<http://www.metafro.be/prelude/>]
  15. Davis M: **Biological Diversity and Indigenous Knowledge.** 1998 [<http://www.aph.gov.au/LIBRARY/pubs/RP/1997-98/98rp17.htm>]. Parliamentary Library, Parliament of Australia
  16. **MySQL database management system** [<http://www.mysql.com/>]
  17. **PHP** [<http://www.php.net/>]
  18. **Jmol: an open-source Java viewer for chemical structures in 3D** [<http://www.jmol.org/>]
  19. Marvin: **Marvin was used for drawing, and displaying chemical structures, Marvin 5.0.3.** ChemAxon 2008 [<http://www.chemaxon.com>].
  20. **Apache** [<http://www.apache.org/>]
  21. Sarkar IN: **Biodiversity informatics: organizing and linking information across the spectrum of life.** *Brief Bioinform* 2007, 8:347-357.
  22. **PubChem** [<http://pubchem.ncbi.nlm.nih.gov/>]
  23. Edwards JL, Lane MA, Nielsen ES: **Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop.** *Science* 2000, 289:2312-2314.
  24. **SpeciesLinks of Integrated Botanical Information System (IBIS)** [<http://www.anbg.gov.au/ibis/specieslinks.html>]
  25. **Integrated Taxonomy Information System (ITIS)** [<http://www.itis.gov/>]
  26. **NCBI Taxonomy** [<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi>]
  27. Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, Tatusova TA, BA R: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 28(1):10-14. 2000 January 1.
  28. **Google images** [<http://images.google.com/>]

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)





### **3.2. Conclusion**

The chemoinformatics module is a part of our effort to help preserve customary medicinal knowledge and serves to integrate chemoinformatics data with the customary knowledge. This will assist in discovery of novel uses of medicinal plants other than customarily known and thus could be helpful in new lead discovery. The module also provides a prototype for developing other small molecule databases that integrate traditional medicinal knowledge with the current biochemical data. Tools for the creation, editing, and interactive visualization are integrated in web pages. The chemical records in the database are also cross referenced to pertinent chemoinformatic databases.



## **Chapter 4: Comparison of physicochemical property space among human metabolites, drugs and toxins**

### **4.1 Summary**

With the advent of combinatorial chemistry and HTS techniques in 1980's, it is now possible to simultaneously design and screen thousands or even million of compounds in a day. Nevertheless, even with these techniques it would take thousands of years to thoroughly explore the chemical universe. Fortunately, biology survives with only a tiny percentage of small molecules and a surprisingly small number of proteins. It is therefore necessary to carry out the chemical expeditions to the pharmaceutically interesting island in chemical space. Many studies in the past have compared the chemical space occupied by various molecular datasets including drugs, natural products, synthetic compounds and lead libraries.

Subsequently, the knowledge of the physicochemical properties of chemical compounds has lead to the concept of drug-like molecules. In this paper, we examined the similarity of current drug molecules with human metabolites and toxics, using a range of computed molecular descriptors and functional groups. Moreover, the effect of using clustered data compared to complete datasets was also investigated



## Physicochemical property space distribution among human metabolites, drugs and toxins

Varun Khanna<sup>1</sup> and Shoba Ranganathan\*<sup>1,2</sup>

Addresses: <sup>1</sup>Dept. of Chemistry and Biomolecular Sciences & ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney, Australia and <sup>2</sup>Dept. of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

E-mail: Varun Khanna - [vkhanha@cbms.mq.edu.au](mailto:vkhanha@cbms.mq.edu.au); Shoba Ranganathan\* - [shoba.ranganathan@mq.edu.au](mailto:shoba.ranganathan@mq.edu.au)

\*Corresponding author

from Asia Pacific Bioinformatics Network (APBioNet) Eighth International Conference on Bioinformatics (InCoB2009) Singapore 7-11 September 2009

Published: 3 December 2009

BMC Bioinformatics 2009, **10**(Suppl 15):S10 doi: 10.1186/1471-2105-10-S15-S10

This article is available from: <http://www.biomedcentral.com/1471-2105/10/S15/S10>

© 2009 Khanna and Ranganathan; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** The current approach to screen for drug-like molecules is to sieve for molecules with biochemical properties suitable for desirable pharmacokinetics and reduced toxicity, using predominantly biophysical properties of chemical compounds, based on empirical rules such as Lipinski's "rule of five" (Ro5). For over a decade, Ro5 has been applied to combinatorial compounds, drugs and ligands, in the search for suitable lead compounds. Unfortunately, till date, a clear distinction between drugs and non-drugs has not been achieved. The current trend is to seek out drugs which show metabolite-likeness. In identifying similar physicochemical characteristics, compounds have usually been clustered based on some characteristic, to reduce the search space presented by large molecular datasets. This paper examines the similarity of current drug molecules with human metabolites and toxins, using a range of computed molecular descriptors as well as the effect of comparison to clustered data compared to searches against complete datasets.

**Results:** We have carried out statistical and substructure functional group analyses of three datasets, namely human metabolites, drugs and toxin molecules. The distributions of various molecular descriptors were investigated. Our analyses show that, although the three groups are distinct, present-day drugs are closer to toxin molecules than to metabolites. Furthermore, these distributions are quite similar for both clustered data as well as complete or unclustered datasets.

**Conclusion:** The property space occupied by metabolites is dissimilar to that of drugs or toxin molecules, with current drugs showing greater similarity to toxins than to metabolites. Additionally, empirical rules like Ro5 can be refined to identify drugs or drug-like molecules that are clearly distinct from toxic compounds and more metabolite-like. The inclusion of human metabolites in this study provides a deeper insight into metabolite/drug/toxin-like properties and will also prove to be valuable in the prediction or optimization of small molecules as ligands for therapeutic applications.



## Background

To search for biologically active compounds, with favorable ADMET [1] (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties from the immense "chemical space" is a non-trivial task [2]. Drug-likeness has been dominated, in the past decade, by Lipinski's "Rule of Five" (Ro5) [3], which states that a compound is likely to be "non-drug-like" if it has more than five hydrogen bond donors, more than 10 hydrogen bond acceptors, molecular mass is greater than 500 and lipophilicity is above 5. The analysis carried out by Leeson and Davis [4] of the approved drugs released before 1983 (i.e. pre-Ro5 era) and the drugs released in between 1983 and 2002 clearly indicates the impact of Ro5 on drug discovery projects.

However, Lipinski's rule has many exceptions and in one of the studies [5] it was shown that using the above criteria, only 66% of approved drugs in the MDL Drug Data Report (MDDR) database, were classified as drug-like; whereas 75% of the theoretically non-drug-like compounds from the Available Chemical Directory (ACD) were in fact regarded as drug-like by Ro5. Moreover, Ro5 does not select metabolites because metabolite-likeness is a recent measure, since Ro5 was formulated a decade ago, with little knowledge on metabolites and pathways. Similar studies have spurred the quest for new approaches to classify drugs from non-drug molecules [6,7], and to characterize the properties of drug-like or lead-like compounds [8,9]. Subsequently, the "rule-of-three" (Ro3) [10] was proposed for fragment-based lead discovery. Ro3 states that successful hits possess an average Molecular weight  $\leq 300$ , the number of hydrogen bond donors  $\leq 3$ , the number of hydrogen bond acceptors  $\leq 3$  and Clog P  $\leq 3$ . In addition, the number of rotatable bonds  $\leq 3$  and the polar surface area  $\leq 60$  are also useful in characterizing drug-like and non-drug-like molecules. In past few years, researchers have developed a range of indices, such as the natural product index [11], the metabolite index [12], peptide-likeness [13], lead-likeness [14-16], and drug-likeness [3], in an attempt to achieve a better classification between drugs and non-drugs. In conjunction with machine learning techniques, like Artificial Neural Networks (ANN) [6,7], Support Vector Machine (SVM) [17] and Hidden Markov Models (HMM), statistical [18] and substructure analyses have become widely accepted to characterize the properties of drug-like datasets and reduce the attrition rates in drug development.

### Drug-likeness in natural products and synthetic compounds

In this section, we present a summary of analysis reports primarily focused on identifying drug-likeness in natural

products and synthetic organic compounds, derived from combinatorial functional group replacement. Henkel et al. [18] carried out statistical analysis to determine the properties and structural differences between natural products (NPs) and combinatorial molecules. In their analysis, NPs were derived from Chapman and Hall Dictionary of Natural products and the bioactive natural product database (BNPD) obtained from Szentor Management Consulting Company. These were compared with synthetic compounds from the Available Chemical Directory (ACD) and Bayers database and representative bioactive molecules from drug databases. Stahura et al. [19] used Shannon entropy to analyze the differences between NPs obtained from the Dictionary of Natural Products and synthetic molecules obtained from ACD database. Feher and Schmidt [20] examined representative set of molecules from NPs obtained from four databases namely BioSPECS natural product database, ChemDiv natural product database, Interbio-screen IBS2001N and HTS-NC database, drugs obtained from (Chapman and Hall Dictionary of Drugs) and combinatorial molecules obtained from (MayBridge HTS database, ChemBridge EXPRESS-Pick database ComGenex Collection, ChemDiv Collection and SPECS screening compound database). The authors concluded that the number of chiral centers, the number of rotatable bonds and the ratio of aromatic atoms to ring atoms are the most distinguishing features among the three classes of compounds. In their study, drugs occupied the property space between NPs and combinatorial compounds, consistent with drugs being obtained from NPs as well as combinatorial libraries. The first three principal components accounted for about 66% of the variance. Feher and Schmidt were thus the first to introduce the idea of NP-like filters. Lee and Schneider [21] utilized Self Organizing Maps (SOM) for the classification of drugs, non-drugs and NPs. Their study revealed several pharmacophoric patterns in common between NPs and drugs, suggesting the use of such patterns for exploring drug relevant pharmacophoric space.

### Metabolite-likeness as the criterion for lead discovery

With the growing knowledge of biochemical pathways and their cognate metabolites, Hattori et al. [22] analyzed the molecular diversity of KEGG (Kyoto Encyclopedia of Genes and Genomes) Ligand database which includes 9,383 chemical structures. Nobeli et al. [23] have produced an interesting classification of *Escherichia coli* metabolome according to fragment-based fingerprints and maximum common subgraphs. Gupta and Aires-de-Sousa [12] compared the structural coverage of the metabolite molecules from the KEGG database and purchasable molecules from the ZINC



library, a free database of commercially available compounds. They reported the use of various machine learning techniques like Kohonen maps, random forest (RFs) and classification trees to distinguish between metabolites and non-metabolites. Cherkasov [24] and coworkers derived 20 binary classifiers and achieved 99% of the accurate separation between drugs, drug-like compounds ("druglikes"), bacterial and human metabolites and antimicrobial compounds, and proposed metabolite-likeness as a potential tool for discovering novel antimicrobials. Recently, Dobson et al. [25] compared different molecular properties among human metabolites, drugs and "predrugs" (precursor drug molecules). They concluded that although metabolites are a distinct class of compounds, metabolites and drugs occupy a significant amount of common property space. They further suggested that metabolite-likeness may be used as a filter for designing drugs which are functionally similar to metabolites and thus have better ADMET properties.

The several excellent studies described above have each compared different datasets, using a variety of chemoinformatics tools and molecular descriptors. Furthermore, some of the studies used datasets that were clustered [25], while others have searched or compared complete (unclustered) datasets [24]. Most importantly, the property space of toxic compounds has not been included in any of these studies, whereas one of the basic tenets of drug development to reduce or eliminate toxicity [26].

The analysis carried out of the drug failures during past few decades have shown that over 90% of the failures are due to high toxicity [27,28]. It is therefore essential that the property space of toxins is explored along with drugs and metabolites to develop filters for toxicity.

Our aim is to compare freely available datasets of metabolites, drugs and toxins, as benchmark datasets, using a range of available molecular descriptors, to identify the property space occupied by these three data types. We also present analysis results from complete datasets, as well as clustered datasets, to determine whether clustering molecules would affect the analysis results. Our results indicate that clustering does not

affect property distributions to a significant level and that unclustered datasets can be used in drug discovery pipelines. We also report, for the first time to the best of our knowledge, that current drug molecules are more akin to toxins than to metabolites, in physicochemical property space.

## Results

### Rule of five (Ro5) analysis

The number of molecules adhering to Ro5 was calculated and the results are reported in Table 1. It is surprising to note that although Ro5 was formulated to pick out drugs or drug-like molecules, it actually does well in identifying toxin molecules. Over 90% of the toxin molecules satisfy all Ro5 criteria. On the other hand, metabolites perform worst among the three datasets while drugs do fairly well, as expected due to the predominance of Ro5 over the past decade. It should also be noted that among the four properties compared, the numbers of hydrogen bond donor and acceptor seem to be more robust properties, as over 84% of the molecules in all the datasets satisfy Ro5 requirements.

### Examining the molecular properties of three datasets

The distribution of various descriptors (properties) among drugs, human metabolites and toxin molecules are available from Table 2 and Fig. 1, 2, 3, 4, based on the analyses of clustered datasets (details in the Methods section). There is very little overlap in the clustered datasets and so no further reduction in redundant data has been carried out (details in the Methods section and Fig. 5).

While there is a multitude of molecular descriptors available for carrying out comparison studies, given the large size of the datasets, we need a set of rapidly computable molecular descriptors, for efficient analysis. Furthermore, to account for 70% of the drugs, Oprea et al. [9] used simple descriptors such as the count of rings and rotatable bonds along with Lipinski descriptors. We have considered a range of 1D and 3D properties for the current analysis. The results are presented as Lipinski (Ro5) properties, 1D properties (non-Ro5 measures) and 3D properties.

**Table 1: Distribution of molecules following Lipinski's rule**

Datasets	Lipinski Properties			
	Molecular weight < 500 Da	H-bond Donor <=5	H-bond Acceptor <=10	Log P < 5
HMDB (Metabolites)	34%	84%	84%	35%
DDB (Drugs)	84%	86%	87%	92%
CPDB (Toxins)	94%	98%	97%	92%



**Table 2: Comparison of molecular properties among the three datasets**

Molecular Property	Mean (Median) $\pm$ std. dev.		
	Metabolites	Drugs	Toxins
<b>Lipinski properties</b>			
<b>Molecular weight</b>	<b>621 (701) <math>\pm</math> 322</b>	<b>355 (309) <math>\pm</math> 259</b>	<b>275 (239) <math>\pm</math> 167</b>
<b>Alog P</b>	<b>7 (10) <math>\pm</math> 7</b>	<b>.08 (1) <math>\pm</math> 3.5</b>	<b>2 (2) <math>\pm</math> 2</b>
<b>Lipinski HB acceptors</b>	<b>9 (9) <math>\pm</math> 6</b>	<b>7 (6) <math>\pm</math> 7</b>	<b>5 (4) <math>\pm</math> 4</b>
<b>Lipinski HB donor</b>	<b>3 (3) <math>\pm</math> 3</b>	<b>3 (3) <math>\pm</math> 4</b>	<b>2 (1) <math>\pm</math> 2</b>
<b>ID properties</b>			
Number of atoms	43 (51) $\pm$ 22	24 (21) $\pm$ 8	16 (14) $\pm$ 11
<b>Number of carbon atoms</b>	<b>34 (41) <math>\pm</math> 18</b>	<b>16 (14) <math>\pm</math> 12</b>	<b>12 (10) <math>\pm</math> 9</b>
Number of hydrogen atoms	60 (72) $\pm$ 33	23 (19) $\pm$ 18	16 (12) $\pm$ 12
<b>Number of nitrogen atoms</b>	<b>1 (1) <math>\pm</math> 2</b>	<b>3 (2) <math>\pm</math> 3</b>	<b>2 (1) <math>\pm</math> 2</b>
<b>Number of oxygen atoms</b>	<b>8 (8) <math>\pm</math> 5</b>	<b>5 (4) <math>\pm</math> 5</b>	<b>3 (2) <math>\pm</math> 3</b>
<b>Number of rings</b>	<b>1 (0) <math>\pm</math> 2</b>	<b>3 (2) <math>\pm</math> 2</b>	<b>2 (2) <math>\pm</math> 2</b>
Number of ring assemblies	1 (0) $\pm$ 1	2 (2) $\pm$ 1	1 (1) $\pm$ 1
<b>Number of rotatable bonds</b>	<b>27 (37) <math>\pm</math> 20</b>	<b>6 (4) <math>\pm</math> 7</b>	<b>3 (2) <math>\pm</math> 4</b>
<b>Number of aromatic bonds</b>	<b>1 (0) <math>\pm</math> 4</b>	<b>8 (6) <math>\pm</math> 7</b>	<b>6 (6) <math>\pm</math> 6</b>
Log D	6 (9) $\pm$ 7	0.4 (0.9) $\pm$ 4	2 (1.4) $\pm$ 2.6
Mol. solubility	-10 (-13) $\pm$ 8	-3 (-3) $\pm$ 3	-3 (-2) $\pm$ 3
<b>3D properties</b>			
Mol. SA	651 (788) $\pm$ 343	364 (316) $\pm$ 252	270 (233) $\pm$ 159
Mol. volume	450 (548) $\pm$ 244	245 (214) $\pm$ 170	179 (153) $\pm$ 110
<b>Mol. polar SA</b>	<b>143 (126) <math>\pm</math> 94</b>	<b>121 (95) <math>\pm</math> 117</b>	<b>84 (63) <math>\pm</math> 76</b>
Mol. SA volume	866 (1051) $\pm$ 420	510 (464) $\pm$ 272	401 (366) $\pm$ 164
Mol. polar sa SA	216 (195) $\pm$ 138	191 (156) $\pm$ 173	126 (105) $\pm$ 91
Mol. sa SA	1034 (1205) $\pm$ 472	578 (523) $\pm$ 313	451 (408) $\pm$ 187

For each dataset, the mean, median and standard deviation values are provided, with properties ordered as Lipinski properties, ID properties and 3D properties. HB: hydrogen bond; Mol.: Molecular; sa: solvent accessible; SA: Surface Area.

#### Lipinski properties

##### Molecular weight

Metabolites follow a bimodal distribution in molecular weight, with the first peak at 100-400 (almost 31% of the dataset) and the second and larger peak at 700-1000, containing 48% of the dataset. On the other hand, the molecular weight of drugs follows a Gaussian distribution with the majority of drugs (82%) under the range of 500. This is in accordance with the Lipinski restriction of weight less than or equal to 500. Despite the Ro5 restriction, 18% of the drug molecules possess a molecular weight in excess of 500. Toxin molecules more or less follow the same pattern as drugs, with the gradual decrease in number of compounds as molecular weight increases from 100 to 500 (Fig. 1a).

From the calculated mean and median values for the molecular weight, it appears that the metabolite data is skewed towards high molecular weight compounds whereas drugs and toxin molecules prefer a lower molecular weight distribution. The statistics of the molecular weight property for the three datasets are available in Table 2.

##### Lipophilicity (Alog P)

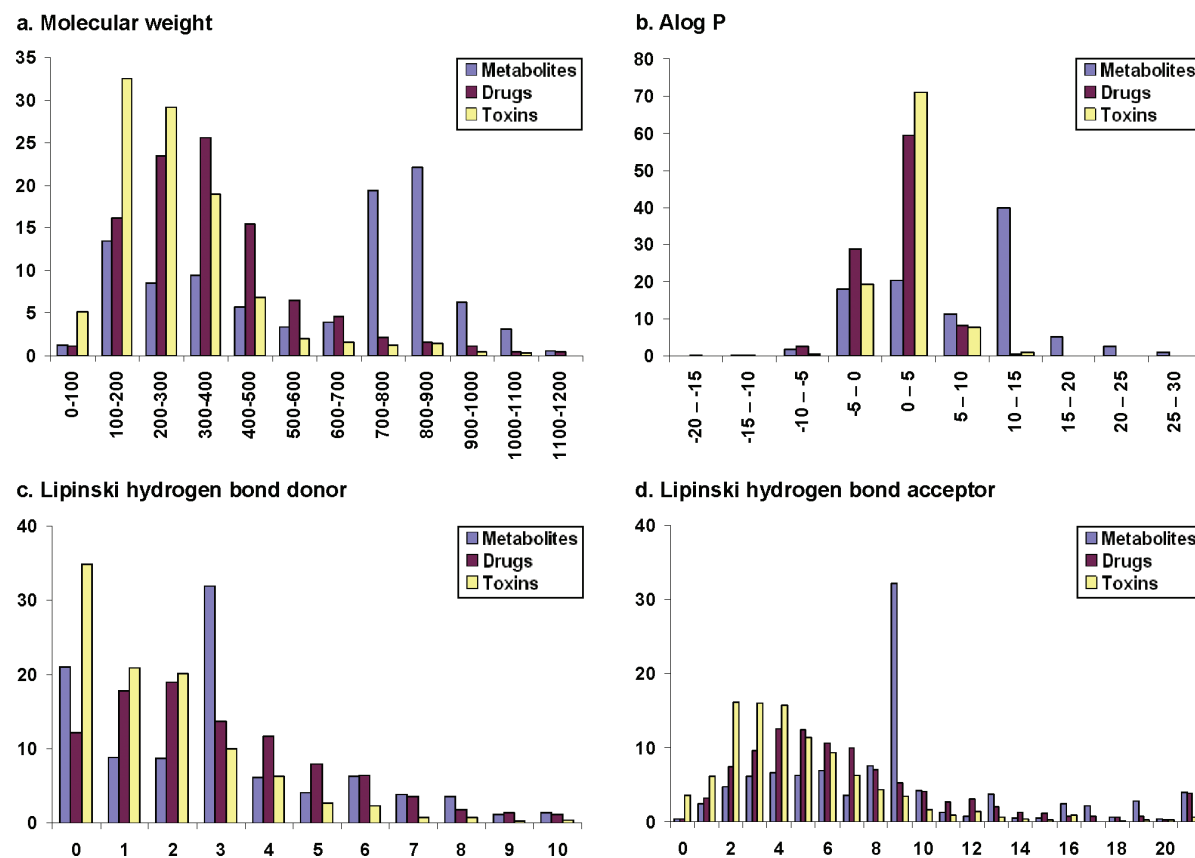
Lipid solubility is a direct measure of transport abilities of the compound across biological membranes [29].

Drug molecules should have enough solubility to traverse the membrane but should not be too soluble so as to get trapped in them. Thus, lipophilicity of a compound is of special significance in drug discovery programs. The most commonly used parameter to evaluate lipid solubility is the n-octanol/water partition coefficient (Alog P). Positive values of this partition coefficient correspond to a preference for lipophilic or hydrophobic environment while negative values indicate a preference for lipophobic or hydrophilic environment. It is clear from Table 2 and Fig. 1b, that metabolites in general are more lipophilic than drugs or toxic compounds. Only 17% of the metabolites have negative Alog P values confirming that the majority of the metabolites are lipophilic. On the other hand, 39% of the drugs have Alog P values in negative territory, indicating that two-fifths of the drugs are lipophobic. Like metabolites, only 19% of the toxin molecules have negative Alog P values while the majority of the molecules are in the range 0 to +5 which is much smaller range as compared to metabolites (Fig. 1b).

##### Lipinski hydrogen bond donors

Lipinski hydrogen bond donors (LHBDs) are determined by counting the numbers of OH and NH bonds in each molecule [3]. Approximately 21% of the metabolites, 12% of the drugs and 34% of the toxin molecules do not



**Figure 1**

**Comparison of Lipinski properties among human metabolites, drugs and toxins.** Compared properties include a. Molecular weight, b. AlogP, c. Number of Lipinski hydrogen bond donors and d. Number of Lipinski hydrogen bond acceptors.

possess any LHBDs. Almost the same percentage of molecules in the drug (~41%) and toxin (~36%) dataset have one or two LHBDs, respectively, while only 17% of the metabolite dataset has the same number of LHBDs. Only 5% of the toxins, 14% of the drugs and 16% of the metabolites have LHBD greater than five (Fig. 1c).

#### Lipinski hydrogen bond acceptor

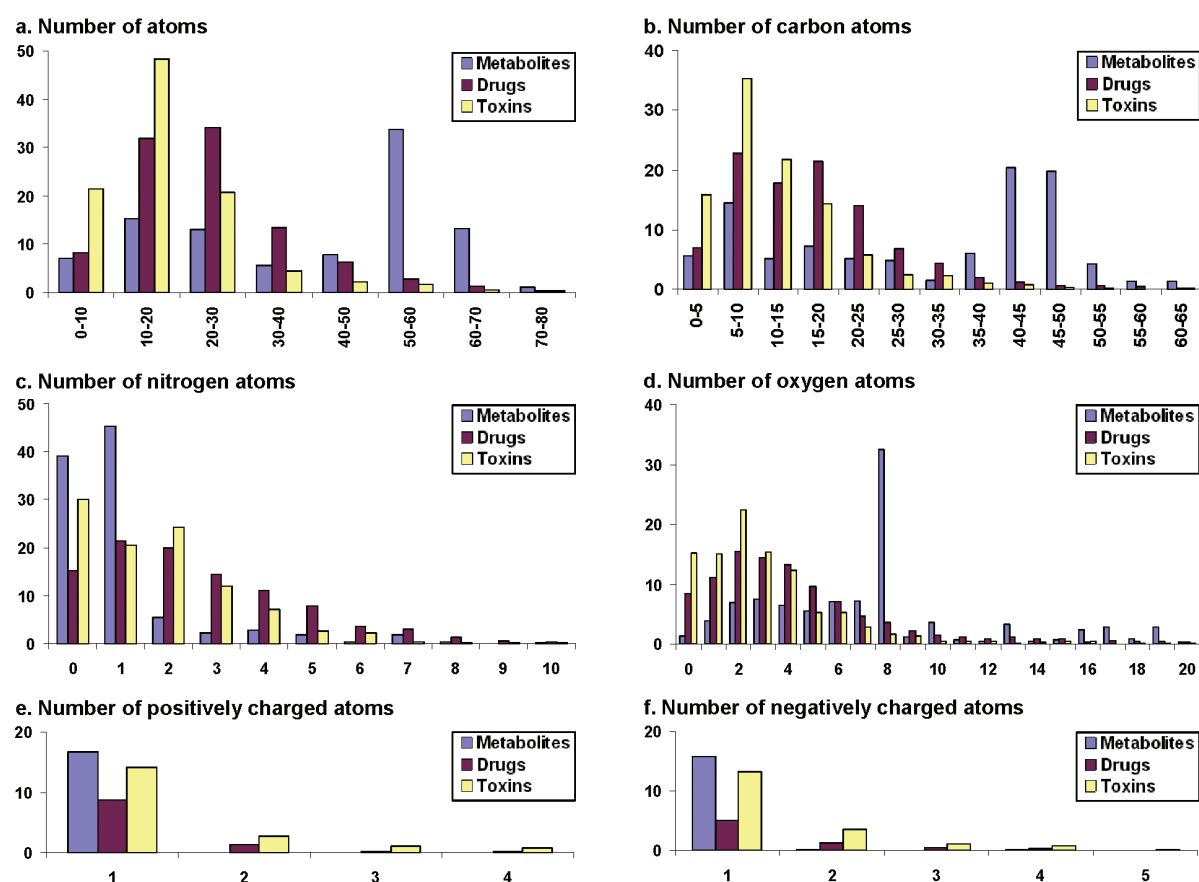
Only a fraction of molecules in all the datasets (0.35% of metabolites, 0.40% of drugs and 3.6% of toxins) do not possess Lipinski hydrogen bond acceptors (LHBAs), computed by summing the numbers of nitrogen and oxygen atoms in each molecule [3]. Drugs and toxins follow almost the same distribution with the highest percentage of molecules in the range 2-7 LHBA atoms per molecule. On the other hand, metabolites have a wide spread distribution with an unusually high peak at 9 LHBA (Fig. 1d).

#### ID properties

##### Total number of Atoms

The distribution of the total number of atoms in metabolites follows a bimodal pattern (Fig. 2a), with the larger peak at 50-70 atoms, containing 47% percent molecules and the smaller peak at 10-30 atoms, containing 28% of molecules. The maximum number of atoms in a metabolite molecule is 124, while the mean value is 43 atoms per molecule. In contrast to human metabolites, the drug dataset follows a bell-shaped curve, skewed towards low numbers of atoms per molecule. Approximately 79% of drugs contain 10-40 atoms per molecule. The average number of atoms per molecule in the drug dataset is 24, while in metabolites, the average is 43. Like drugs, toxin molecules also favor smaller numbers of atoms per molecule, with a mean of 16 and a gradual decrease in the number of compounds as the number of atoms increases per molecule. The majority of



**Figure 2**

**ID Atomic property differences between human metabolites, drugs and toxins.** Compared properties include a. Number of atoms, b. Number of carbon atoms c. Number of nitrogen atoms d. Number of oxygen atoms e. Number of positively charged atoms f. Number of negatively charged atoms

the toxin dataset (91%) contains 10-30 atoms per molecule while only 9% of toxin molecules contain 30 or more atoms per molecule. The overall statistics of the three datasets is given in Table 2 and show that metabolites tend to have more atoms than drugs and toxin molecules.

#### Carbon content

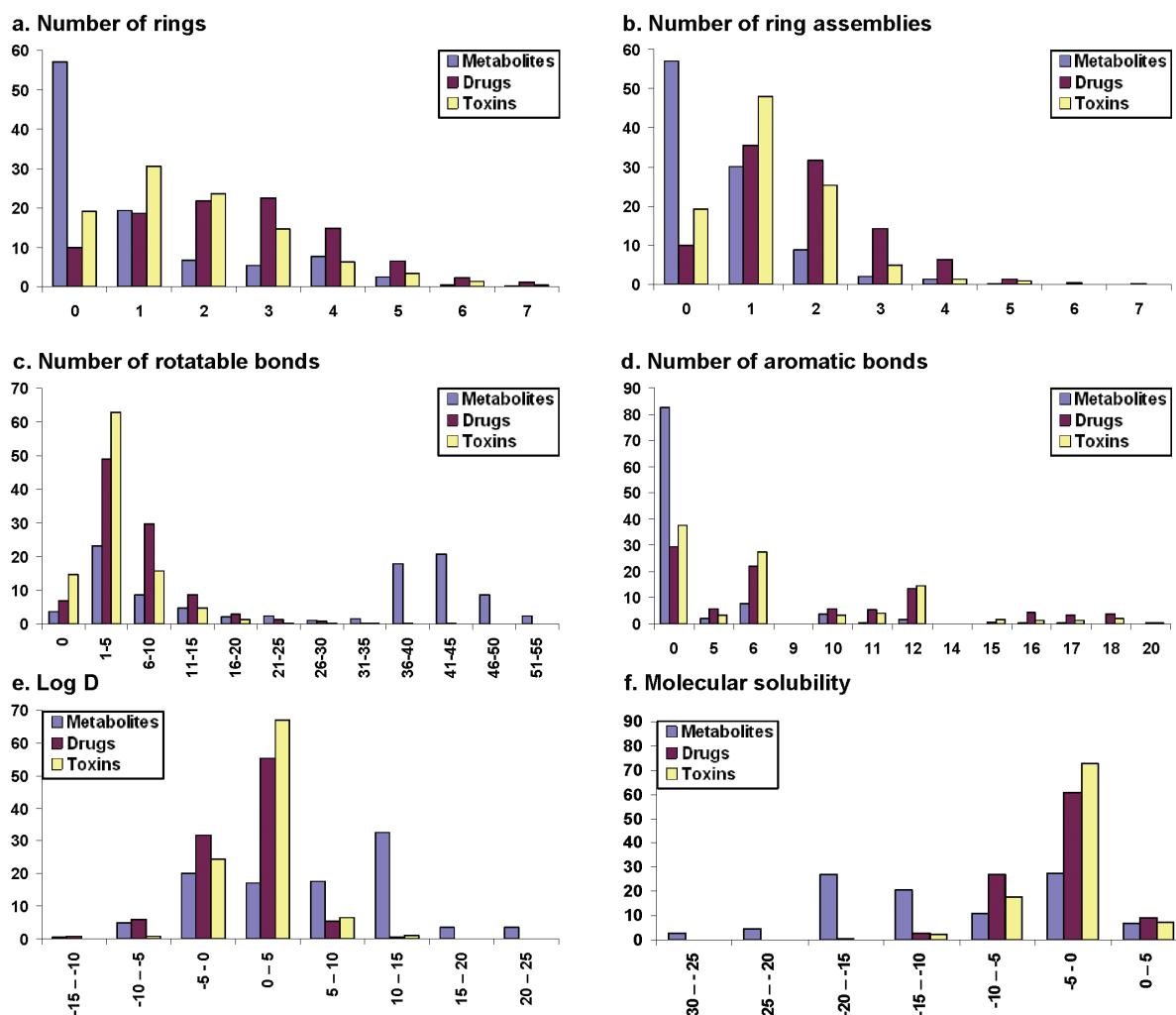
Almost half of the molecules in the metabolite dataset have carbon atoms in the range 35-55 while 32% have 5-25 carbon atoms per molecule (Fig. 2b). The carbon atom distribution in metabolites has a mean of 33 atoms and a maximum of 100. On the other hand, drugs have a mean of 18 carbon atoms per molecule, with a maximum of 256 and 76% of drugs have carbon atoms in the range 5-25. Similar to drugs, toxin molecules also seem to prefer fewer carbons. In the toxin dataset, 77% of the molecules have 5-25 carbon

atoms, while 16% have five or fewer carbon atoms. Only 7% of the molecules have more than 25 carbon atoms in toxin dataset. The distribution of carbon atoms in the toxin dataset has a mean of 12 and a maximum of 62. From Table 2, we note that metabolites contain more carbon atoms than drugs, which in turn have greater carbon content than toxin molecules.

#### Nitrogen content

Approximately 40% of metabolites do not have any nitrogen atom (Fig. 2c), while 45% have only one nitrogen atom and 16% have two or more nitrogen atoms per molecule. In sharp contrast to metabolites, only 15% of drug molecules do not possess nitrogen atoms while 74% of the molecules have nitrogen atoms in the range 1-5. On the other hand, 30% of the toxin molecules are devoid of any nitrogen atom while 66% of



**Figure 3**

**Other ID properties compared among human metabolites, drugs and toxins.** Compared properties include a. Number of rings, b. Number of ring assemblies c. Number of rotatable bonds, d. Number of aromatic bonds, e. Log D, f. Molecular solubility.

toxin molecules contain nitrogen atoms in the range 1-5 and only 3% have six or more nitrogen atoms. From Table 2 and the values presented above, drugs molecules clearly possess the most number of nitrogen atoms, followed by toxin molecules and lastly, metabolites.

#### Oxygen content

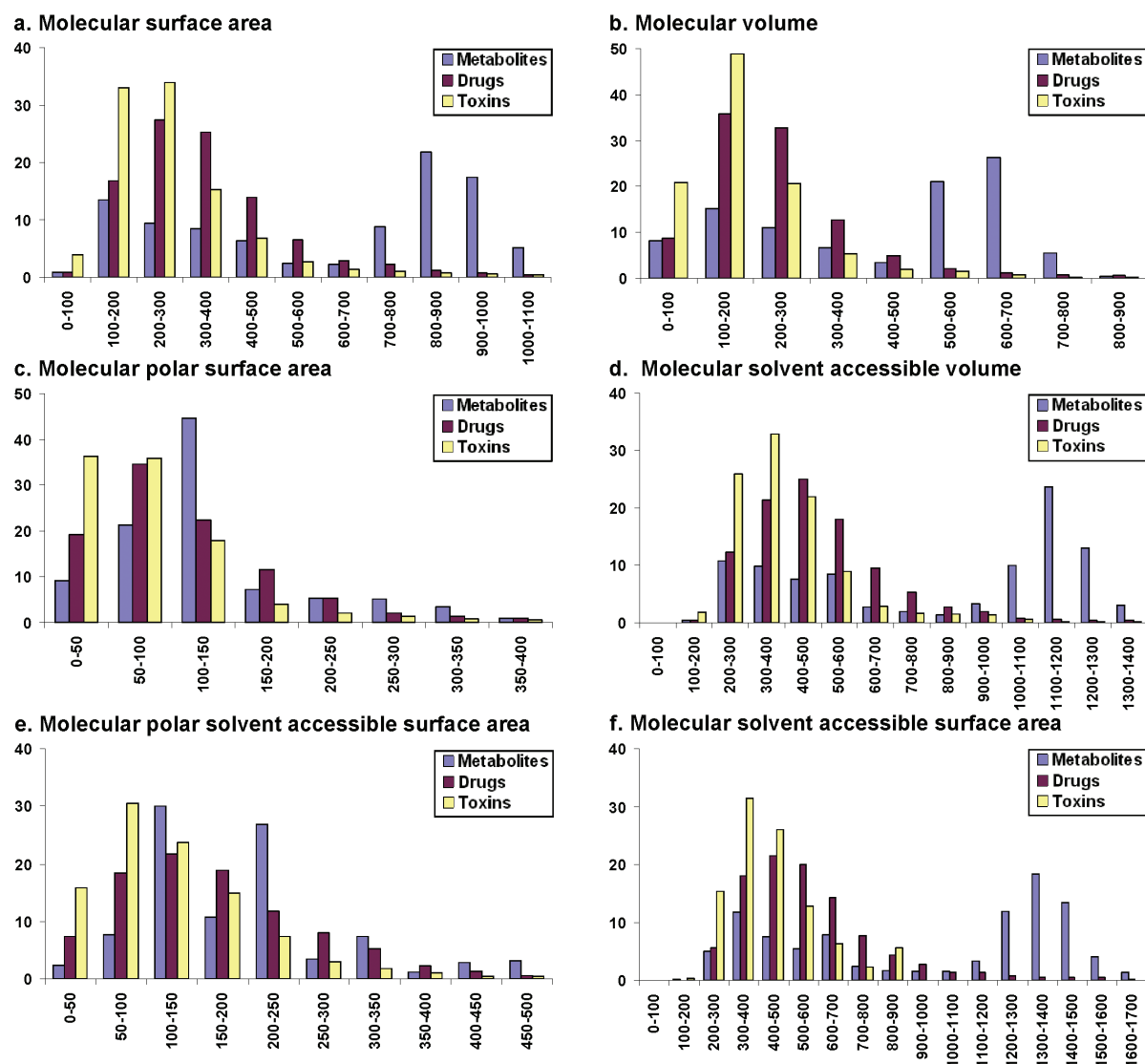
For the three datasets, there is a clear reversal of the trend for the oxygen atom distribution compared to the nitrogen atom distribution presented in the previous section. Only 1% of the metabolite molecules do not

have an oxygen atom as compared to 8% of drugs and 15% of toxin molecules (Fig. 2d). Furthermore, in metabolite dataset, 73% of the molecules possess oxygen atoms in the range 2-8, compared to 68% of drugs and 65% of the toxins. Metabolites comprise more oxygen atoms than drugs, followed by toxic compounds, with mean values of eight, five and three, respectively (Table 2).

#### Number of negatively and positively charged atoms

The fraction of molecules with a single negatively charged atom in the metabolite dataset (16%) is almost



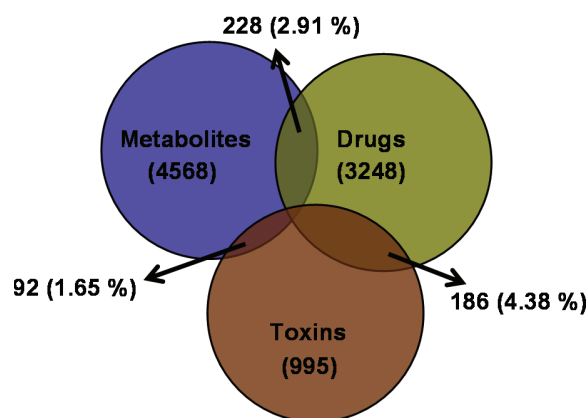
**Figure 4**

**Comparison of 3D properties among human metabolites, drugs and toxins.** Compared properties include a. Molecular surface area, b. Molecular volume c. Molecular polar surface area, d. Molecular solvent accessible volume, e. Molecular polar solvent accessible surface area, f. Molecular solvent accessible surface area.

the same as that containing one positively charged atom (17%). While the metabolite dataset contains molecules with more than one negatively charged atom, there are no molecules with more than one positively charged atom. The percentage of negatively charged atoms is smaller in the drug dataset as compared to metabolite dataset (Fig. 2e and Fig. 2f). Only 5% of drug molecules contain a negatively charged atom, with only 2% containing two or more negatively charged atoms,

whereas 8% contain one positively charged atom. On the other hand, in the toxin dataset, 13% of the molecules contain one negatively charged atom per molecule and 5% of molecules contain two or more negatively charged atoms, whereas 14% of the molecules in the same dataset contain one positively charged atom per molecule and 4% contain two or higher positively charged atoms. The trend of charged atoms among the three datasets is Metabolites > Toxin molecules > Drugs,





**Figure 5**  
Venn diagram showing the overlap between the three clustered datasets.

with the drug dataset favouring negatively charged atoms over positively charged ones.

#### Number of rings

The distribution of number of rings is shown in Fig. 3a. Although more than 55% of the molecules in the metabolite dataset are acyclic, 19% contain a single ring and 21% contain 2-5 rings. In sharp contrast to metabolites, only 9% of the drugs are acyclic, while almost 60% contains rings, with a three-way distribution (~20% each) between one, two and three rings per molecule. The remaining 23% of drug molecules contains 4-6 rings, the maximum number of rings being 38. In the toxin dataset, 19% of the molecules are acyclic, whereas 68% contain 1-3 rings per molecule. The remaining 10% of toxins contain four or more rings per molecule. Thus, the pattern of ring distribution among the three datasets is Drugs > Toxin molecules > Metabolites.

#### Number of ring assemblies

After removing the non-ring bonds from a molecule the remaining backbone is termed as the ring assembly. As shown in the Fig. 3b, more than half of the molecules (57%) in the metabolite dataset have no ring assembly, while 30% of the molecules have one ring assembly and 13% have two or more ring assemblies per molecule. On the other hand, in the drug dataset, only 10% of the molecules are free of ring assembly, whereas 36% and 32% have one and two ring assemblies, respectively. Furthermore, in the same dataset, 23% molecules contain more than three ring assemblies per molecule. Similar to drugs, most of the toxins possess ring assemblies with only 19% are devoid of any ring

assembly, while 45% molecule have a single ring assembly. The percentage of molecules with two ring assemblies in the same dataset is 21% whereas 6% of the molecules have three or more ring assemblies. The pattern of ring assemblies is similar to that obtained for ring distribution in three datasets, being Drugs > Toxin molecules > Metabolites.

#### Number of rotatable bonds

The number of rotatable bonds is a measure of molecular flexibility and is important in determining oral bioavailability of the drugs [30]. Only 4% of the molecules in the human metabolite dataset have no rotatable bonds, whereas 32% have 1-10 rotatable bonds and 47% of the molecules have rotatable bonds in the range 36-50 (Fig. 3c). The mean value for rotatable bond distribution in metabolite dataset is 27, with the maximum number of rotatable bonds in a metabolite molecule being 83 (Table 2). Among the drug molecules, 7% are devoid of rotatable bonds, while 79% of molecules have 1-10 rotatable bonds. Another 12% of the molecules in this dataset have rotatable bonds in the range of 10-20. The mean value for rotatable bonds per molecule in drugs is 6, with a maximum of 170. In contrast to metabolites and drugs, 15% of toxin molecules do not possess any rotatable bonds, while 79% of the molecules contain rotatable bonds in the range of 1-10. The mean value for rotatable bonds in toxin molecules is 3 and the maximum number of rotatable bonds in a toxin dataset is 31. Thus, metabolites are more flexible than drugs and toxins.

#### Number of aromatic bonds

More than 80% of metabolites do not possess any aromatic bond. The remaining metabolites have several aromatic bonds, usually as multiples of five or six. As shown in Fig. 3d, 6% of the molecules have either five or ten aromatic bonds, while 8% of molecules have either six or twelve aromatic bonds. The maximum number of aromatic bonds in metabolites is 36.

In contrast to metabolites, only 29% molecules of drugs have no any aromatic bonds. Of the remaining, 12% and 36%, respectively, of drug molecules have aromatic bonds as multiple of five and six. The maximum number of aromatic bonds in the drug dataset is 62. On the other hand, toxin molecules are predominantly aromatic (61%) with 7% and 42% having aromatic bonds as multiples of five and six, respectively. The maximum number of aromatic bonds in the toxin dataset is 46. The order of aromatic bond distribution is Drugs ≈ Toxin molecules > Human metabolites, with almost half the aromatic bonds in all the three datasets being multiples of five or six.



### Log D

For solutes that can ionize, the distribution coefficient (D) is the ratio of the sum of the concentrations of all forms of the compound (ionized plus un-ionized) in each of the two solution phases. logD is thus considered a better measure of lipophilicity than Alog P. However, for all three datasets, logD follows the same distribution as Alog P (Fig 3e and Table 2).

### Molecular solubility

Human metabolites have large range of molecular solubility values for example more than 85% of the molecules in metabolite dataset have molecular solubility in the range of -20 to 0. Drug molecules have a smaller range of molecular solubility as compared to metabolites, with 87% of the molecules in drug dataset having molecular solubility values spanning -10 to 0. Similarly, more than 90% of toxin molecules have a solubility value in the range -10 to 0. Thus, the most preferred and common range of molecular solubility among the three datasets is -10 to 0, which comprises ~85% of the drug and toxin datasets but only 38% of the metabolite dataset (Fig. 3f and Table 2). The molecular solubility of metabolites is more than that of drugs, followed by toxin molecules, which suggests that metabolites tend to dissolve more easily than drugs and toxins *in vivo* (aqueous media).

### Chirality

Chirality seems to be a distinguishing feature among the three datasets. The majority of the molecules in the metabolite dataset (74%) are chiral. Chirality falls sharply in drugs and toxic compounds to 31% and 14%, respectively.

### Number of halogen atoms per molecule

As expected, toxin molecules have the highest number of halogen atoms per molecule compared to metabolites and drugs. 31% of molecules in toxin dataset possess a single halogen atom (F, Cl, Br, I) per molecule while in case of drugs close to 18% contain halogen atoms. In sharp contrast, to these two datasets, metabolites have far fewer halogen atoms per molecule. Only 15 out of 4568 molecules, i.e. only 0.3% of the molecules, studied are reported to have any halogen atom. The trend for halogens is Toxin molecules > Drugs >> Metabolites. The

statistics provided in Table 3 provides information on the number of halogen containing molecules in each dataset.

### Number of Sulphur and Phosphorus atoms per molecule

Only 5% of the molecules in metabolite dataset, 20% of the drugs and almost the same percentage (16%) of toxin molecules contain one or more sulphur atoms in their molecule. For sulphur atoms, the trend is Drugs ≈ Toxin molecules > Metabolites. The trend gets reversed in the case of phosphorus atoms, with 46% of molecules in metabolites, 13% of drug molecules and only 3% of the toxic dataset having one or more phosphorus atoms. So the trend in phosphorus atom distribution is Metabolites >> Drugs > Toxin molecules.

### Average bond length per molecule

The metabolite molecules form two groups, with 48% having a mean value of 0.82 Å bond length, while another 51% have 0.83 Å. The majority of drug molecules (65%) also have either 0.82 or 0.83 Å bond length. In sharp contrast to the other two datasets, the average bond length for 92% of molecules in toxic dataset is 1.33 Å, while another 7% of the molecules have 1.40 Å as the average bond length. As far as average bond length is concerned, metabolites and drugs have a much shorter average bond length compared to toxin molecules, the trend being Toxin molecules > Metabolites ≈ Drugs.

### 3D Descriptors: molecular volume and surface area

#### Molecular surface area

Molecular surface area distribution in metabolites is bimodal (Fig. 4a) with the first smaller peak at 100-400 Å<sup>2</sup>, containing 37% of the molecules and the second larger peak at 700-1100 Å<sup>2</sup>, with 53% of the molecules. On the other hand, 83% of drugs molecules have molecular surface area between 100-500 Å<sup>2</sup>. A similar distribution is obtained for toxin molecules with 89% of the compound in the toxin dataset having a molecular surface area in the range 100-500 Å<sup>2</sup> and only 4% are in the range 0-100 Å<sup>2</sup>. From these values and the statistics in Table 2, metabolites have greater molecular surface area than drugs and toxin molecules.

**Table 3: Halogen atom frequency distribution. The number of times different halogens are reported in each of the dataset is listed below**

Database	Fluorine	Chlorine	Bromine	Iodine
Metabolites	15	32	0	27
Drugs	496	477	110	63
Toxins	62	473	38	5



#### Molecular volume

The results of molecular volume distribution in three datasets are reflected in the related property of molecular weight distribution. As depicted in Fig. 4b the molecular volume range in metabolites is much wider and in accordance with molecular weight data when compared to the other two datasets. Almost 47% of the molecules have molecular volume in the range of 500-700 Å<sup>3</sup>. The majority of molecular volume distribution of the drug dataset is narrow compared to that of human metabolites with 81% of the molecules are in the range from 100-400 Å<sup>3</sup>, although the tails extend further, with some molecules found to have volumes above 1700 Å<sup>3</sup>. The molecular volume range is even more restricted in toxic compounds with 90% of the molecules in the range 0-300 Å<sup>3</sup> with ~49% of these having a molecular volume of 100-200 Å<sup>3</sup>. So, the trend for molecular volume distribution is the same as that observed for molecular weight distribution among the three datasets: Metabolites > Drugs > Toxin molecules.

#### Molecular polar surface area

The polar surface area is defined as the surface area summed over all polar atoms, (usually oxygen and nitrogen), including the attached hydrogen atoms. It is often correlated with drug transport capabilities and is important for penetrating the blood-brain barrier (BBB). As most of the metabolites do not need to be shuttled through barriers like BBB, they can afford to have more polar surface area than drugs and toxins. More than 95% of the metabolites have polar surface area in the range 0-350 Å<sup>2</sup> (Fig. 4c) while 92% of polar surface area of drugs is contained within 0-250 Å<sup>2</sup>. The distribution is even narrower for the toxin dataset with 90% of the molecules in the range 0-150 Å<sup>2</sup>.

#### Molecular solvent accessible volume

Molecular solvent accessible volume distribution is similar to the distribution of the molecular volume. In the case of metabolites (Fig. 4d), it also follows a bimodal distribution with a smaller peak of 36% molecules around 200-600 Å<sup>3</sup> and a larger peak containing 46% of the molecules around 1000-1300 Å<sup>3</sup>. However, there is no molecule with accessible volume less than or equal to 100 Å<sup>3</sup>. Unlike metabolites, drugs molecules have only one peak covering almost the entire dataset. About 91% of the drug molecules have solvent accessible volume from 200 to 800 Å<sup>3</sup>. Like metabolites there is no molecule with solvent accessible volume less than or equal to 100 in drug dataset. The distribution of solvent accessible volume in toxin molecules is even thinner with 89% of the molecules in the range 200-600 Å<sup>3</sup>. Other 7% are present in the range 600-1000 Å<sup>3</sup>. According to the statistics shown in Table 2 and Fig. 4d,

the order of molecular solvent accessible volume is Metabolites > Drugs > Toxin molecules.

#### Molecular polar solvent accessible surface area

Drugs and toxin molecules follow a perfect Gaussian distribution for polar solvent accessible surface area while metabolites follow a bimodal pattern (Fig. 4e). The maximum number of molecules in toxic dataset has molecular polar solvent accessible surface area is in the range 0-200 Å<sup>2</sup> while for drugs the range is 0-350 Å<sup>2</sup>. On the other hand maximum numbers of metabolites are covered in between 100-250 Å<sup>2</sup>. The statistics in Table 2 suggests that metabolites tend to have larger molecular polar solvent accessible surface area compared to drugs which in turn are larger than toxins.

#### Molecular solvent accessible surface area

Differences among metabolites, drugs and toxin molecules are readily observable for molecular solvent accessible surface area. Metabolites follow a bimodal distribution whereas drugs and toxins follow a Gaussian distribution (Fig. 4f). Toxin molecules peak at 300-500 Å<sup>2</sup> while drugs peak at 400-600 Å<sup>2</sup>. Metabolites, on the other hand, form a lower peak at 300-400 Å<sup>2</sup> with a second larger peak at 1200-1500 Å<sup>2</sup>. From Table 2, metabolites have clearly larger values for molecular solvent accessible surface area than drugs and toxins.

#### Functional group analysis

The frequency of functional group occurrence among the three datasets was carried out in this study with the Scitegic Pipeline pilot software (details in the Methods section). The occurrence of specific functional groups of interest to drug design is given in Table 4 and Additional file 1. Aromatic atoms are a prominent feature among drugs and toxins while only a sixth of metabolites have aromatic atoms. The same trend is observed in benzene ring distribution among the datasets. Further, primary and quaternary amines occur more frequently in metabolites than secondary and tertiary amines when compared to drugs and toxin molecules, respectively. Additionally, drugs are found to possess a greater number of amides than metabolites or toxins. Finally, toxic functional groups (like nitro, azo and cyanide) are only found in toxins while they are either absent or very limited in drugs and metabolites.

#### Clustered vs. unclustered datasets

We have compared all the above property distributions for clustered and unclustered (raw) datasets (data not shown). Correlation coefficients were calculated for all the properties and eight properties which are not significantly correlated are presented here, viz. Alog P, molecular weight, the number of oxygen atoms, the



**Table 4: Occurrence of functional groups in the three datasets**

Functional Group	Metabolite dataset	Drugs dataset	Toxin dataset
Alkyl halide	<0.5%	<0.5%	3.2%
Aromatic atom	17.4%	70.6%	62.3%
Benzene	10.3%	56.0%	53%
Steroid backbone	2.9%	0.6%	<0.5%
HBA Ester	56.3%	13.8%	15.4%
Pyridine	1.2%	6.4%	5.3%
Pyrimidine	3.2%	7.5%	1.9%
Enamine	3.2%	10.31%	3.41%
Primary amine	28%	14.4%	12.0%
Secondary amine	11.4%	64.0%	41.2%
Tertiary amine	44.6%	80.0%	60.0%
Quaternary Amine	15.3%	2.1%	0.5%
Primary amide	1.5%	4.5%	3.9%
Secondary amide	11.4%	31.0%	14.5%
Tertiary amide	2.8%	16.8%	9.2%
Imines	4.1%	14.0%	6.4%
Azo	0%	<0.5%	3.4%
Carbamic acid	<0.5%	3.1%	1.9%
Urea	2.5%	8.0%	6.5%

Those functional groups which can discriminate between the three datasets are presented here. The complete list is in Additional File 1.

number of nitrogen atoms, molecular polar surface area, molecular solubility, the number of rings and the number of aromatic bonds (Figs. 6, 7, 8, 9). Alog P and molecular weight values (Fig. 6) do not deviate significantly with clustering. Nitrogen atom distribution (Fig. 7) for clustered and unclustered molecules also remains fairly similar for all the datasets. The analysis also shows that the number of aromatic bonds (Fig. 8) and the molecular solubility distribution (Fig. 9) are also fairly conserved between clustered and unclustered datasets. We note that, by and large, the two distributions are very similar except in following cases:

#### Number of oxygen atoms

There is an exception at five oxygen atoms per molecule in the unclustered metabolite dataset (Fig. 7b).

#### Number of rings

The number of molecules with zero rings drops for drugs (~8% decrease) and toxins (~9% decrease) whereas metabolites follow a similar distribution in clustered and unclustered dataset comparison (Fig. 8a).

#### Molecular polar surface area

Clustered metabolites show a 10% decrease in molecules with polar surface area in the range 50-100 Å<sup>2</sup> while clustered toxins show a 15% increase in the number of molecules with polar surface area between 0 to 50 Å<sup>2</sup>. Drugs, on the other hand, follow a similar distribution for clustered and unclustered datasets (Fig. 9a).

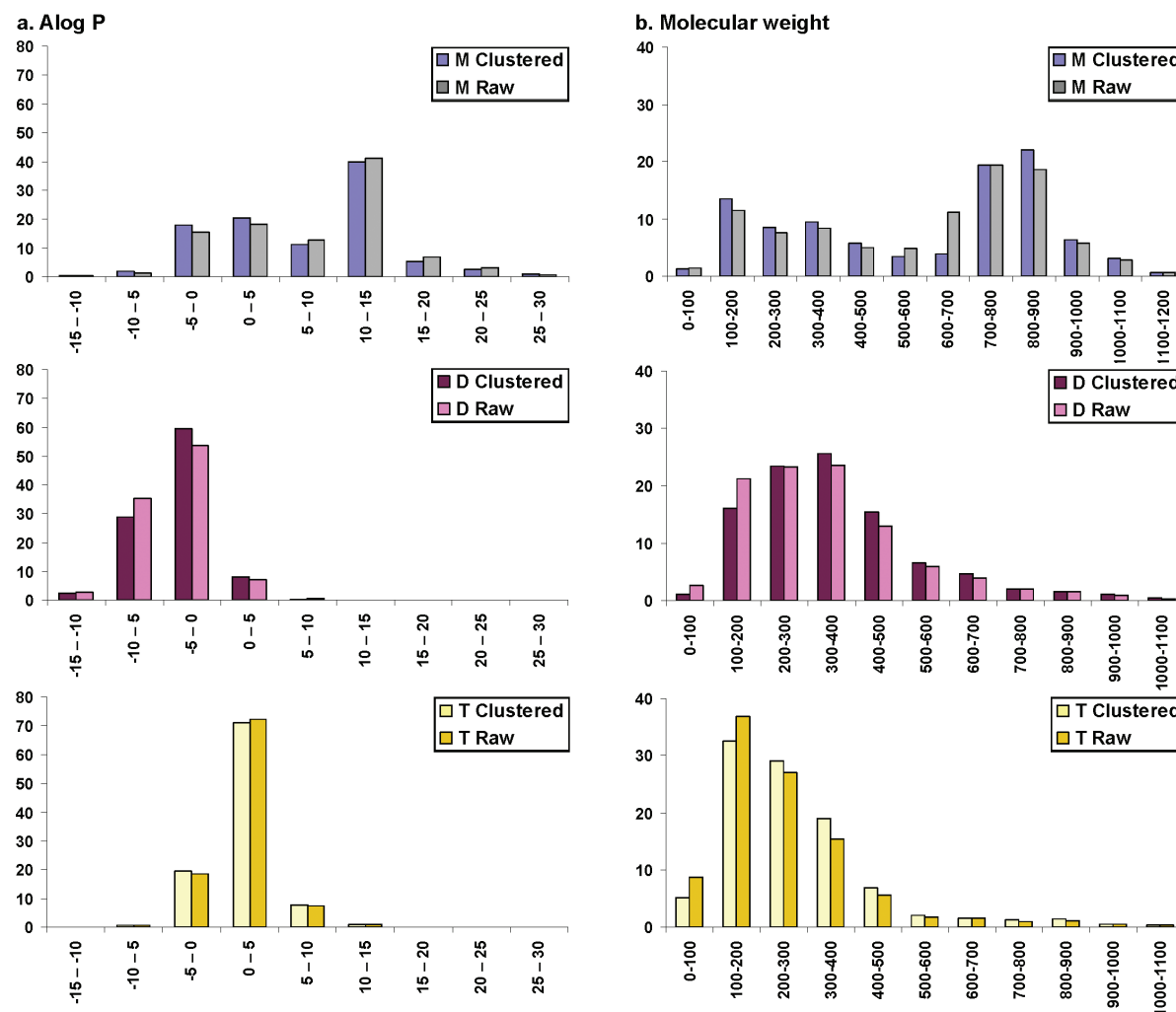
## Conclusion

We have carried out a comprehensive analysis of three publicly available datasets, comprising drug, metabolite and toxin molecules. We have also, for the first time, to the best of our knowledge, compared the distributions of various properties for complete datasets (unclustered data) as well as reduced or clustered datasets. We note that, in the main, the distributions for the two data groups, clustered and unclustered, are very similar, supporting the use of clustered datasets, except in the case of the number of oxygen atoms, the molecular polar surface area and the number of rings. Based on this result, these properties should be treated with caution for lead discovery in drug discovery pipelines with unclustered datasets.

From the analyses of clustered datasets, we find that two-thirds of the human metabolites lie outside the Lipinski universe. On the other hand, over 90% of the toxin molecules abide by Lipinski's rule, implying that since Ro5 does not explicitly take toxicity into account, present-day drugs are consequently similar to toxins than to metabolites.

Results from the analysis of 1D and 3D molecular properties consolidate our finding of drugs and toxins sharing a larger property space, than drugs and metabolites. 1D properties such as the total number of atoms advocate that metabolites are bulky, with more carbon and hydrogen atoms than drug and toxins. This is consistent with the idea that metabolites are produced at the required subcellular location and thus do not need to be transferred from one location to another. In order to design metabolite-like drugs, it would be beneficial to attempt alternative ways for drug delivery, since traditionally, drugs are required to pass through the blood-brain-barrier, which limits the size of drug molecules. Considering the numbers of nitrogen and oxygen atoms, metabolites prefer oxygen over nitrogen containing groups. Above 50% of the metabolites are acyclic while only 9% of the drugs and 19% of the toxin molecules are acyclic. The number of rotatable bonds measuring molecular flexibility and consequently, oral bioavailability, suggests that metabolites are far more flexible than drugs and toxin molecules. Over 70% of the drugs and 62% of toxin molecules are aromatic while only 20% of the metabolites are aromatic. This result is in accordance with the fact that drugs are derived from various sources including NPs which are mostly aromatic in nature. In all the datasets examined, the majority of molecules have negative solubility values, suggesting that a large proportion of these compounds are soluble in aqueous solutions. Chirality falls sharply from metabolites to drugs and toxin molecules while as expected, the number of halogen atoms are found to



**Figure 6**

**Comparison of example Lipinski properties for clustered and unclustered (raw) data.** Properties compared are a. Alog P, b. Molecular weight, for human metabolites (M), drugs (D) and toxin molecules (T).

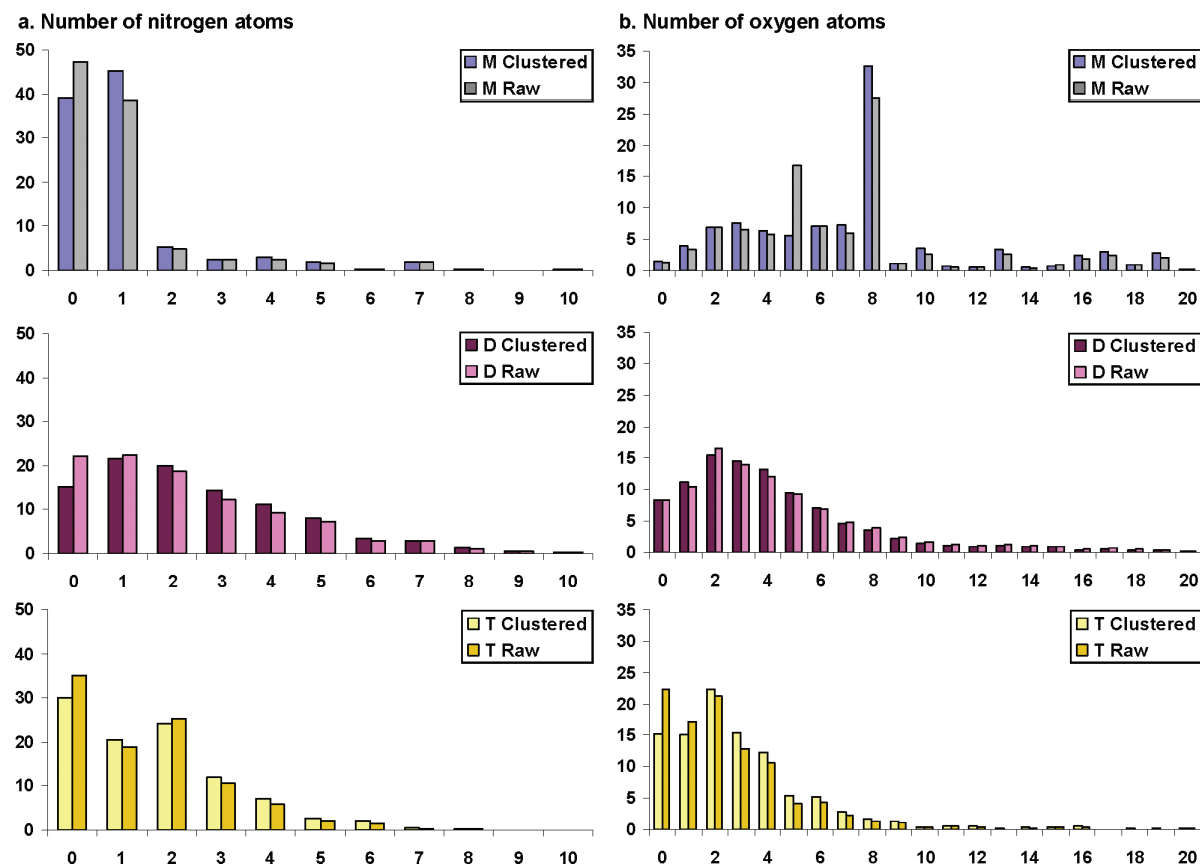
be higher in toxins than in drugs and metabolites. The average bond length of 90% metabolites and more than 65% of drugs is much smaller than majority of toxins, suggesting multiple bonds in the former datasets. The analysis results from 3D descriptors such as molecular volume and molecular surface area are reflected in the related property of molecular weight and confirm that present day drugs are more like toxins than metabolites.

The analysis also shows that although drugs share a relatively larger property space with toxins than with metabolites, drugs and toxins are two different classes of

compounds as reflected in specific physicochemical characteristics. Drugs tend to have higher values for properties such as molecular weight, the number of oxygen atoms, the number of rotatable bonds and molecular polar surface area whereas toxin molecules have considerably higher Alog P and Log D values.

Additionally, empirical rules like the "rule of five" can be refined to increase the coverage of drugs or drug-like molecules that are clearly not close to toxic compounds, because toxicity reduction is one of the key aspects of drug discovery programs. Our results have implications



**Figure 7**

**Comparison of example ID atomic properties for clustered and unclustered (raw) data.** Properties compared a. Number of nitrogen atoms, b. Number of oxygen atoms, for human metabolites (M), drugs (D) and toxin molecules (T).

for the analysis of novel compounds in lead discovery pipelines, to uncover novel target molecules.

## Methods

### Preparation of the dataset

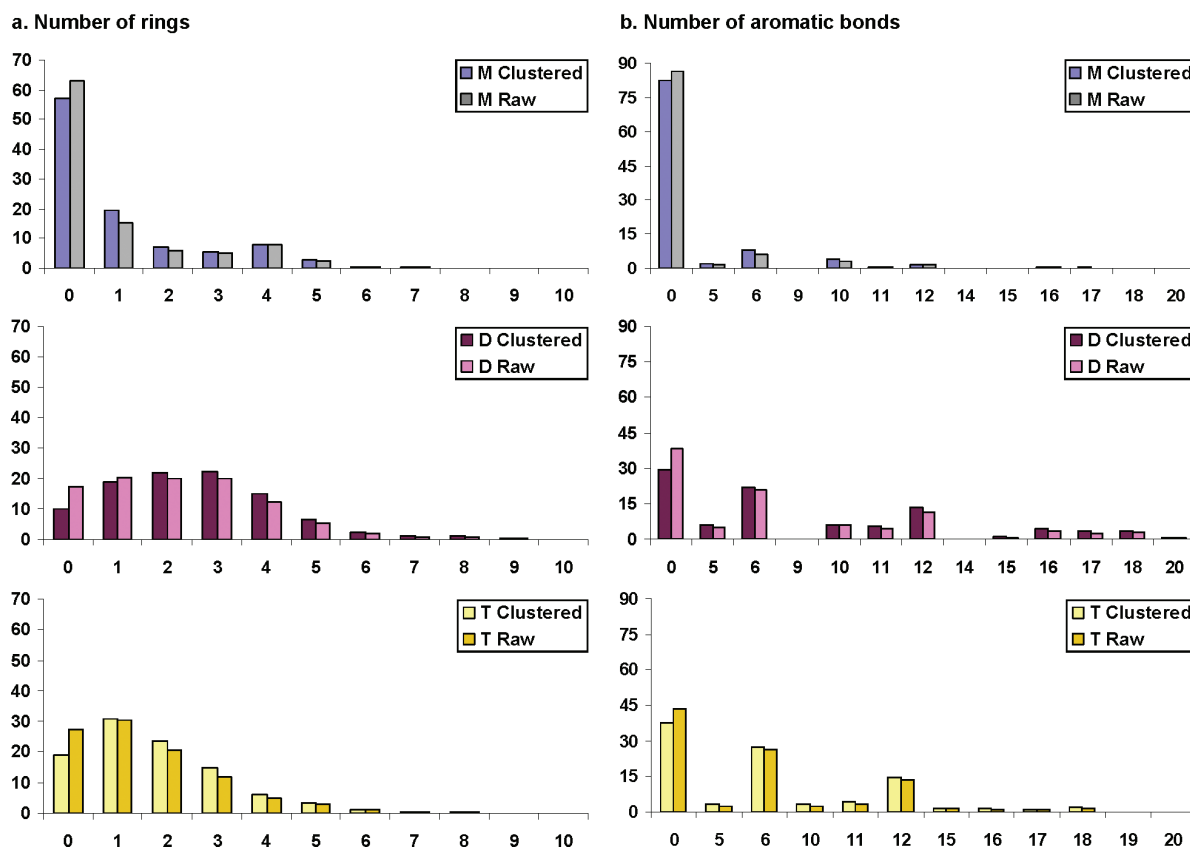
Three publicly available databases, relevant to human diseases and their treatment have been used in this study. The human metabolome database [31] contains information on nearly 7000 small molecule metabolites found in human body. Similarly, DrugBank [32] is a comprehensive resource on drugs and drug targets, with detailed chemical, pharmaceutical and medical information on nearly 3000 drug targets and 4800 drugs including >1,350 FDA-approved small drugs and experimental drugs derived from the PDB-Ligand database [33], containing compounds bound to biomolecules. Distributed Structure-Searchable Toxicity (DSSTox)

Carcinogenic Potency Database [34] is hosted by the US Environmental Protection Agency's National Center for Computational Toxicology aiming to provide a public data repository on toxicity data. DSSTox contains experimental results and carcinogenicity information for 1547 substances tested against different species.

Preliminary datasets containing 6668 human metabolites from the human metabolome database (as on 23-Dec-2008), 4883 drugs from DrugBank (as on 6-Jan-2009) and 1547 toxin molecules from DSSTox (as on 16-Jan-2009) were extracted.

From these preliminary datasets duplicates and inorganic molecules (individual atoms, metal salts, inorganic oxides, hydroxides, cations and anions) were removed. Any "missing" compounds (either with no or incomplete structure) were also removed. The "cleaned" collections of



**Figure 8**

**Comparison of example 1D aromatic properties for clustered and unclustered (raw) data.** Properties compared are a. Number of rings, b. Number of aromatic bonds, for human metabolites (M), drugs (D) and toxin molecules (T).

unique compounds were compiled into analysis datasets containing 6582 metabolites, 4829 drug molecules and 1448 toxin molecules. Finally, clusters were generated from each dataset, using the Cluster "Clara" algorithm embedded in the Scitegic Pipeline Pilot software [35], which is an approximate version of "partitioning around medoids" (pam) method comprising 70% of the entire raw data, similar to that reported in Dobson et al. [25]. Clustering was performed to address the issue of possible overrepresentation of the chemical space, which might bias the analysis results towards these redundant molecules. Representative sets of molecules were produced by employing the extended connectivity fingerprint (ECFP) [36,37] as a molecular descriptor and Euclidean distance was the distance metric selected. ECFP generates an array of structural features by encoding each atom and its molecular environment within a sphere of specified diameter. Cluster centres were selected as the representatives, for clusters containing more than one molecule while singletons were directly used as cluster

centres in non-cluster situations. The contents of unclustered and clustered datasets, prepared for analyses are presented in Table 5.

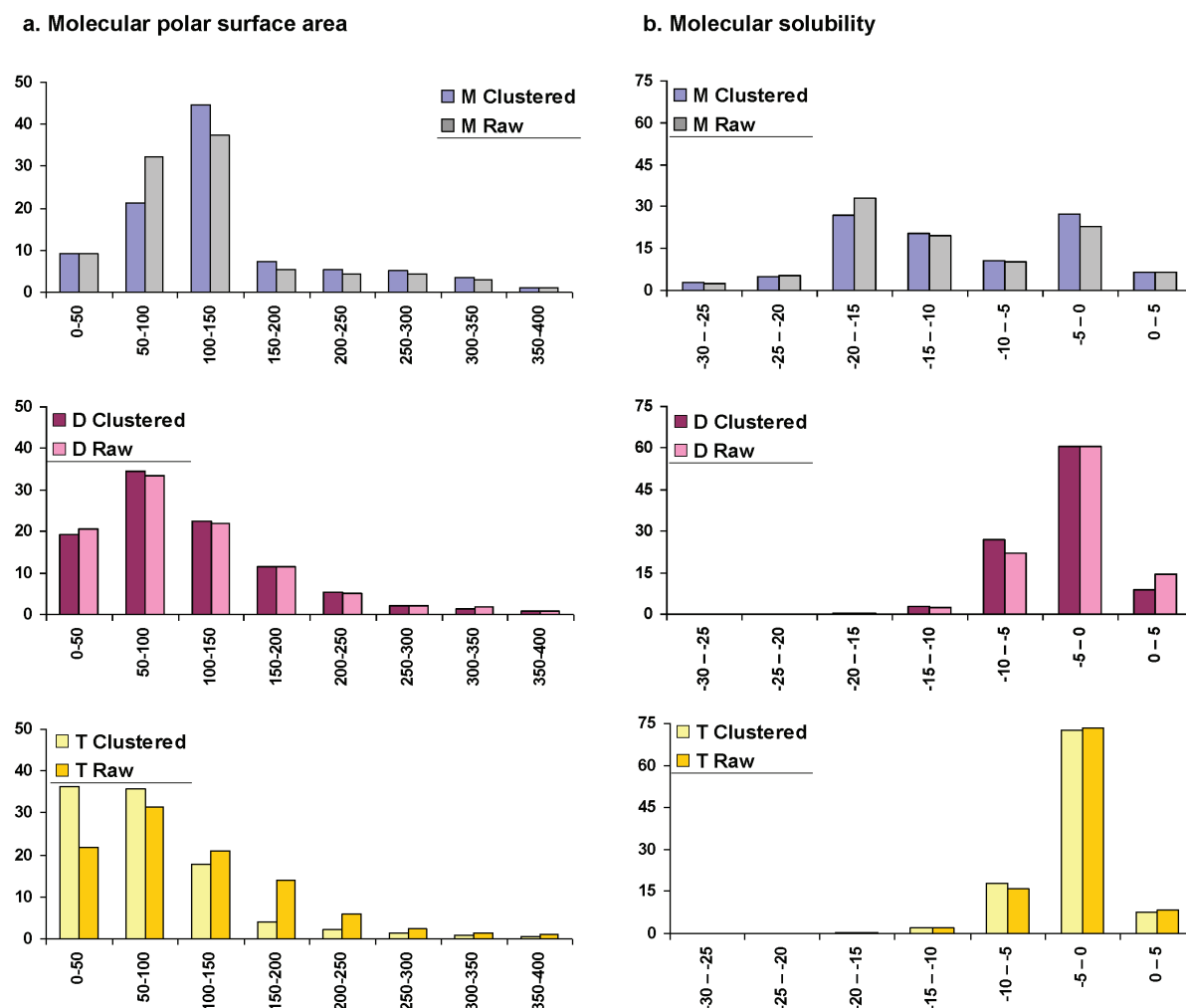
The overlap among the three clustered datasets (CM, CD and CT) was calculated and it was found that more compounds are common in between drugs and toxin molecules than any other combination. The results are displayed in Figure 5. As the binary overlap is very small (<5%) and the ternary overlap is negligible, the datasets were retained as such, without further size reduction.

#### Calculations of the physicochemical properties

The calculation of all the molecular properties was carried out through the Scitegic pipeline pilot [35] and in-house Perl scripts.

Two types of hydrogen bond acceptors and donors were taken into account. Firstly, the Lipinski type donors (sum





**Figure 9**  
**Comparison of example molecular properties important in drug design, for clustered and unclustered (raw) data.** Properties compared are a. Molecular polar surface area, b. Molecular solubility, for human metabolites (M), drugs (D) and toxin molecules (T).

**Table 5: Clustered and Unclustered datasets**

Dataset	Metabolites	Drugs	Toxin molecules
Unclustered	M: 6582	D: 4829	T: 1448
Clustered	CM: 4568	CD: 3248	CT: 995

of OH and NH) and acceptors (sum of N and O atoms) were calculated as defined by Lipinski et al. [3] and then, all available hydrogen bond donors and acceptors were summed up.

The octanol-water partition coefficient was either retained if provided with the data, or was calculated from Scitegic software. The hydrophobicity measure, Alog P, was calculated using the Ghose-Crippen method [38] which takes into account the group contribution to Log P. Another partition coefficient, Log D (the distribution coefficient), which take into account unionized and ionized species, was also calculated. Log D is equal to Log P for unionizable compounds but with ionized species, Log D is considered better than Log P, as it takes ionized species into account, along with unionized forms. A positive value of Log P or Log D suggests a



preference to lipophilic surroundings, whereas a negative value indicates preference to lipophobic (or hydrophilic) environment.

$$\text{Log D} = \sum [C_i]_{\text{oct}} / \sum [C_i]_{\text{aq}} \quad (1)$$

Other simple count-based molecular descriptors enumerating aromatic bonds, atoms, carbon atoms, nitrogen atoms, oxygen atoms, hydrogen atoms and rings were also calculated. Beside these, one-dimensional (1D) descriptors calculated include molecular weight and molecular solubility. Three-dimensional (3D) descriptors like molecular volume, molecular surface area, molecular polar surface area and molecular solvent accessible surface area were also computed. The molecular polar surface area is defined as the sum of all the polar atoms (usually oxygen and nitrogen atoms, and the attached hydrogen atoms). This descriptor is often correlated with drug transport capabilities and is important in penetrating the blood-brain barrier.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

VK curated the dataset and conducted the analysis work; SR directed the study and both authors prepared and approved the manuscript.

### Note

Other papers from the meeting have been published as part of BMC Genomics Volume 10 Supplement 3, 2009: Eighth International Conference on Bioinformatics (InCoB2009): Computational Biology, available online at <http://www.biomedcentral.com/1471-2164/10?issue=S3>.

### Additional material

#### Additional file 1

Table S1. Occurrence of discriminatory functional groups in the three datasets.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-10-S15-S10-S1.pdf>]

### Acknowledgements

We thank Assoc. Prof. J. Jamie for useful discussions during this study. VK is grateful to Macquarie University for the award of MQRES research scholarship. Open access publication charges for this article were covered by Macquarie University.

This article has been published as part of BMC Bioinformatics Volume 10 Supplement 15, 2009: Eighth International Conference on Bioinformatics (InCoB2009): Bioinformatics. The full contents of the supplement are available online at <http://www.biomedcentral.com/1471-2105/10?issue=S15>.

### References

- Hodgson J: **ADMET - turning chemicals into drugs.** *Nat Biotechnol* 2001, **19**(8):722-726.
- Lipinski C and Hopkins A: **Navigating chemical space for biology and medicine.** *Nature* 2004, **432**(7019):855-861.
- Lipinski CA, Lombardo F, Dominy BW and Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Deliv Rev* 2001, **46**(1-3):3-26.
- Leeson PD and Davis AM: **Time-related differences in the physical property profiles of oral drugs.** *J Med Chem* 2004, **47**(25):6338-6348.
- Frimurer TM, Bywater R, Naerum L, Lauritsen LN and Brunak S: **Improving the odds in discriminating "drug-like" from "non drug-like" compounds.** *J Chem Inf Comput Sci* 2000, **40**(6):1315-1324.
- Ajay A, Walters VJP and Murcko MA: **Can we learn to distinguish between "drug-like" and "nondrug-like" molecules?.** *J Med Chem* 1998, **41**(18):3314-3324.
- Sadowski J and Kubinyi H: **A scoring scheme for discriminating between drugs and nondrugs.** *J Med Chem* 1998, **41**(18):3325-3329.
- Brustle M, Beck B, Schindler T, King W, Mitchell T and Clark T: **Descriptors, physical properties, and drug-likeness.** *J Med Chem* 2002, **45**(16):3345-3355.
- Oprea TI: **Property distribution of drug-related chemical databases.** *J Comput Aided Mol Des* 1999, **14**:251-264.
- Congreve M, Carr R, Murray C and Jhoti H: **A 'rule of three' for fragment-based lead discovery?.** *Drug Discov Today* 2003, **8**(19):876-877.
- Erd P, Roggo S and Schuffenhauer A: **Natural product-likeness score and its application for prioritization of compound libraries.** *J Chem Inf Model* 2008, **48**(1):68-74.
- Gupta S and Aires-de-Sousa J: **Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness.** *Mol Divers* 2007, **11**(1):23-36.
- Eckert H and Bajorath J: **Exploring peptide-likeness of active molecules using 2D fingerprint methods.** *J Chem Inf Model* 2007, **47**(4):1366-1378.
- Oprea TI: **Cheminformatics and the quest for leads in drug discovery.** *Handbook of Chemoinformatics* Wiley-VCH: J Gasteiger, Weinheim 2003, 1508-1531.
- Oprea TI: **Current trends in lead discovery: are we looking for the appropriate properties?.** *J Comput Aided Mol Des* 2002, **16**(5-6):325-334.
- Oprea TI, Davis AM, Teague SJ and Leeson PD: **Is there a difference between leads and drugs? A historical perspective.** *J Chem Inf Comput Sci* 2001, **41**(5):1308-1315.
- Jorissen RN and Gilson MK: **Virtual screening of molecular databases using a support vector machine.** *J Chem Inf Model* 2005, **45**(3):549-561.
- Henkel T, Brunne RM, Muller H and Reichel F: **Statistical investigation into the structural complementarity of natural products and synthetic compounds.** *Angewandte Chemie-International Edition* 1999, **38**(5):643-647.
- Stahura FL, Godden JW, Xue L and Bajorath J: **Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations.** *J Chem Inf Comput Sci* 2000, **40**(5):1245-1252.
- Fehér M and Schmidt JM: **Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry.** *J Chem Inf Comput Sci* 2003, **43**(1):218-227.
- Lee ML and Schneider G: **Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries.** *J Comb Chem* 2001, **3**(3):284-289.
- Hattori M, Okuno Y, Goto S and Kanehisa M: **Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways.** *J Am Chem Soc* 2003, **125**(39):11853-11865.



23. Nobeli I, Ponstingl H, Krissinel EB and Thornton JM: **A structure-based anatomy of the E. coli metabolome.** *J Mol Biol* 2003, **334**(4):697-719.
24. Karakoc E, Sahinalp SC and Cherkasov A: **Comparative QSAR- and fragments distribution analysis of drugs, druglikes, metabolic substances, and antimicrobial compounds.** *J Chem Inf Model* 2006, **46**(5):2167-2182.
25. Dobson PD, Patel Y and Kell DB: **'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries.** *Drug Discov Today* 2009, **14**(1-2):31-40.
26. Gleeson MP: **Generation of a set of simple, interpretable ADMET rules of thumb.** *J Med Chem* 2008, **51**(4):817-834.
27. Schuster D, Laggner C and Langer T: **Why drugs fail—a study on side effects in new chemical entities.** *Curr Pharm Des* 2005, **11**(27):3545-3559.
28. Gut J and Bagatto D: **Theragenomic knowledge management for individualised safety of drugs, chemicals, pollutants and dietary ingredients.** *Expert Opin Drug Metab Toxicol* 2005, **1**(3):537-554.
29. Hansch C, Bjorkroth JP and Leo A: **Hydrophobicity and central nervous system agents: on the principle of minimal hydrophobicity in drug design.** *J Pharm Sci* 1987, **76**(9):663-687.
30. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW and Kopple KD: **Molecular properties that influence the oral bioavailability of drug candidates.** *J Med Chem* 2002, **45**(12):2615-2623.
31. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D and Sawhney S, et al: **HMDB: the Human Metabolome Database.** *Nucleic Acids Res* 2007, **35** Database: D521-526.
32. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z and Woolsey J: **DrugBank: a comprehensive resource for in silico drug discovery and exploration.** *Nucleic Acids Res* 2006, **34** Database: D668-672.
33. Shin JM and Cho DH: **PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures.** *Nucleic Acids Res* 2005, **33** Database: D238-241.
34. Gold LS, Sawyer CB, Magaw R, Backman GM, de Veciana M, Levinson R, Hooper NK, Havender WR, Bernstein L and Peto R, et al: **A carcinogenic potency database of the standardized results of animal bioassays.** *Environ Health Perspect* 1984, **58**:9-319.
35. **SciTegic Pipeline Pilot.** Accelrys, Inc., San Diego, CA, USA; <http://accelrys.com/products/scitegic/>.
36. Zhou D, Alelyunas Y and Liu R: **Scores of extended connectivity fingerprint as descriptors in QSPR study of melting point and aqueous solubility.** *J Chem Inf Model* 2008, **48**(5):981-987.
37. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E and Schuffenhauer A: **Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures.** *Org Biomol Chem* 2004, **2**(22):3256-3266.
38. Ghose AK and Crippen GM: **Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions.** *J Chem Inf Comput Sci* 1987, **27**(1): 21-35.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:

[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)





## Additional File 1

### Physiochemical property space distribution among human metabolites, drugs and toxins

Varun Khanna, Shoba Ranganathan

**Table S1. Occurrence of discriminatory functional groups in the three datasets.**

The functional groups which are most distinguishable among the three datasets are shown in bold.

Functional Group	Metabolite dataset	Drugs dataset	Toxin dataset
Aldehyde	2.1%	1.5%	1.0%
<b>Alkyl halide</b>	<b>&lt;0.5%</b>	<b>&lt;0.5%</b>	<b>3.2%</b>
<b>Aromatic atom</b>	<b>17.4%</b>	<b>70.6%</b>	<b>62.3%</b>
<b>Benzene</b>	<b>10.3%</b>	<b>56.0%</b>	<b>53%</b>
Benzoic Acid	1.0%	3.4%	3.5%
Benzoic acid amide	0.5%	6.3%	1.5%
Indole	2.7%	3.8%	1.5%
Flavone core	0.5%	0.7%	0.7%
Lactam	0.6%	3.6%	4.4%
Prostaglandins	0.7%	<0.5 %	<0.5%
<b>Steroid backbone</b>	<b>2.9%</b>	<b>0.6%</b>	<b>&lt;0.5%</b>
<b>HBA Ester</b>	<b>56.3%</b>	<b>13.8%</b>	<b>15.4%</b>
SP hybrid atom	<0.5%	2.5%	2.9%
<b>Pyridine</b>	<b>1.2%</b>	<b>6.4%</b>	<b>5.3%</b>
<b>Pyrimidine</b>	<b>3.2%</b>	<b>7.5%</b>	<b>1.9%</b>
Carboxylic acid	21.0%	24.1%	10.3%
<b>Enamine</b>	<b>3.2%</b>	<b>10.31%</b>	<b>3.41%</b>
Enol	<0.5%	1.5%	1.3%
Enol-Ether	5.5%	3.0%	3.3%
Guanadine	2.3%	4.8%	1.8%



Functional Group	Metabolite dataset	Drugs dataset	Toxin dataset
Primary amine	28%	14.4%	12.0%
Secondary amine	11.4%	64.0%	41.2%
Tertiary amine	44.6%	80.0%	60.0%
Quaternary Amine	15.3%	2.1%	0.5%
Primary amide	1.5%	4.5%	3.9%
Secondary amide	11.4%	31.0%	14.5%
Tertiary amide	2.8%	16.8%	9.2%
Imines	4.1%	14.0%	6.4%
Isourea	0%	<0.5%	<0.5%
Nitrate	0%	0%	0%
Nitrite	0%	<0.5%	<0.5%
Nitro	0%	0%	0%
Nitroso	<0.5 %	0.6%	8.4%
Oxime-Ether	<0.5%	1.1%	<0.5%
Semicarbazide	0%	0.7%	2.9%
Isocynate	0%	0%	<0.5%
Hydrazine	0%	<0.5%	2.1%
Diazo	0%	<0.5%	<0.5%
Azide	0%	0%	0%
Azo	0%	<0.5%	3.4%
Carbamic acid	<0.5%	3.1%	1.9%
Carbamic acid ester	<0.5%	2.3%	1.0%
Urea	2.5%	8.0%	6.5%



## 4.2 Conclusion

A number of physicochemical descriptors have been utilized to analyze three publicly available datasets, comprising drug, metabolite and toxics. We have also, for the first time, to the best of our knowledge, compared the distributions of various properties for unclustered data as well as clustered datasets. We note that distribution of the two groups clustered and unclustered is quite similar, arguing the case for the use of clustered datasets. From our analysis, we note that over 70% of the metabolites do not follow Lipinski's rule while over 90% of the toxics lie within Lipinski's universe implying that Ro5 does explicitly takes toxicity into consideration, which could explain the high attrition rates due to drug failures from toxicity.

We also found that physicochemical property space occupied by current drugs is relatively similar to toxics as compared to metabolites. In our study over 50% of the metabolites are acyclic while only 9% and 19% of the drugs and toxics are acyclic respectively. The number of rotatable bonds measuring molecular flexibility and consequently, oral bioavailability, suggests that metabolites are far more flexible than drugs and toxics molecules. Further we note that only 20% of the metabolites are aromatic while 70% of the drugs and 62% of the toxic molecules are aromatic. Finally we note that although drugs share a relatively larger physicochemical property space with toxics than with metabolites, drugs and toxics are two different classes of compounds as reflected in specific physicochemical characteristics. In our analysis we find that drugs have higher values for properties such as molecular weight, the number of oxygen atoms, the number of rotatable bonds and molecular polar surface area whereas toxics have considerably higher Alog P and Log D. Furthermore, human metabolites are a distinct class of compounds as compared to drugs. Therefore, metabolite-like space could be explored while designing lead libraries.

Empirical rules like the Ro5 can be refined and complemented with other measures to include the toxicity information so as to increase the coverage of drugs or drug-like molecules that are clearly not close to toxic compounds. The distribution of many physicochemical properties is similar in clustered and unclustered datasets, except in the case of the number of oxygen atoms, the molecular polar surface area and the number of rings.



## **Chapter 5: Scaffold and fragment co-occurrence studies on datasets of biological interest.**

### **5.1 Summary**

The drug discovery process screens potential lead compounds using dictionary-based and hash-based binary fingerprints, for drug-likeness or lead-likeness. With the availability of the human metabolome, metabolite-likeness is now increasingly used as a drug design concept [88] for targeting specific pathways.

In chapter 4, we reported a comprehensive analysis of the physicochemical property space occupied by drugs, metabolites and toxics, using public datasets suggesting that present day drugs are more akin to toxics than to metabolites [152]. In this manuscript, we used a multi-criteria approach to study the differences among various datasets of biological interest viz. drugs, human metabolites, toxics, natural products and lead compounds. Further, we included molecules from two well known public databases, NCI and ChEMBL. We extended our earlier physicochemical analysis to include fingerprints, in order to identify commonly found fragments in bioactive compounds found in these datasets. Thus, we analysed physicochemical properties, scaffold architecture and fragment co-occurrence data in these datasets. In order to prioritize the enormous fingerprint data obtained, we have applied association or market basket analysis, to determine statistically significant co-occurring fingerprints. Association rules were generated, with “support” and “confidence” values for frequently occurring fragments.



# Scaffold analysis and fragment co-occurrence studies on public datasets of biological interest

*Varun Khanna<sup>1</sup>, Shoba Ranganathan<sup>1,2\*</sup>*

<sup>[1]</sup> Dept. of Chemistry and Biomolecular Sciences and ARC Centre of Excellence in Bioinformatics,  
Macquarie University, Sydney, Australia.

<sup>[2]</sup> Dept. of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore,  
Singapore.

Email: [varun.khanna@mq.edu.au](mailto:varun.khanna@mq.edu.au); [shoba.ranganathan@mq.edu.au](mailto:shoba.ranganathan@mq.edu.au)



## **Abstract**

### **Background**

The recent availability of the human metabolome has revitalized “metabolite-likeness” as a drug design concept to design lead libraries, targeting specific pathways. Many reports have analyzed the property space of biologically important datasets, with only a few studying the co-occurrences of fragments. With large collections of high quality public data currently available, we carried out a comparative analysis of current day leads with other biologically relevant datasets.

### **Results**

In this study, we note a two-fold enrichment of metabolite scaffolds in drug dataset (42%) as compared to presently used lead libraries (23%). Further, we note that only a small percentage (5%) of natural product scaffolds space is shared by the lead dataset. We have identified specific scaffolds that are present in metabolites and natural products, with close counterparts in the drug dataset, but missing in the lead dataset. To determine statistically significant co-occurring fragments, we applied association rules. Here, we note that metabolites produce a large number of association rules (96, *support* 0.1) compared to drugs and toxics (4 and 2 respectively, *support* 0.1 each), signifying that metabolites extensively reuse fragments to produce novel molecules.

### **Conclusions**

Currently used lead libraries make little use of the metabolites and natural products scaffold space. We believe that since metabolites, natural products are recognized by at least one protein in the biosphere, sampling the fragment, and scaffold space of these compounds, along with the knowledge of co-occurring fragments, can result in better lead libraries. Nevertheless, metabolites have a limited distribution in chemical space that limits the usage of metabolites in library design.



## Background

An established idea of similarity-based virtual screening is that similar structures tend to have similar properties [1]. Diversifying the compound library collection for *in silico* and *in vitro* high-throughput screening without compromising biological activity remains an active research area. Chemical space is enormous but mostly biologically insignificant [2] and therefore, uninteresting from a drug design perspective. Given the large number of currently available chemical compounds in PubChem [3], it is impossible and irrational to screen all known compounds for potential ligands. One key methodology, fragment-based virtual screening (FBVS) or fragment-based drug discovery (FBDD), is an emerging area to identify novel, small molecules for preclinical studies. In FBDD, the starting points are small low molecular weight, drug-like fragments. Examples of such fragments are ring systems, functional groups, side chains, linkers and fingerprints.

Over the past decade, substructures contributing to drug-like or lead-like properties have governed library design [4]. In one of the pioneering works to understand the distribution of common fragments in drugs, Bemis and Murcko [5] fragmented a drug dataset (taken from the Comprehensive Medicinal Chemistry database) into rings, linkers, frameworks and side chains. Using two-dimensional topological graph-based molecular descriptors, they found 2506 different frameworks for a set of 5120 drug compounds, with the top 32 accounting for the topologies of 50% of the database compounds. They concluded a skewed distribution of molecular frameworks in drugs. Metabolite-likeness is increasingly being used as filter to design lead libraries similar to metabolites with better absorption, distribution, metabolism, elimination and toxicology (ADMET) properties [6]. Many recent studies have compared chemical space occupied by compounds of pharmaceutical interest [7-12]. Grabowski and Schneider [7] studied the molecular properties and chemotype diversity of drugs, pure natural products (NPs), and natural product derived compounds. Following the approach described by Bemis and Murcko [5], they virtually dissected the molecules into frameworks, corresponding to scaffolds and side-chains. The drug dataset was ranked most structurally diverse, followed by marine and plant derived NPs, respectively. However, in contrast to the observation of Bemis and Murcko that only 32 frameworks form the basis of nearly 50% of the



compounds in CMC drug database, Grabowski and Schneider found that 160 graph-based frameworks are needed to explain the chemotype of 50% of the compounds in the COBRA drug dataset. In the same year, Siegel and Vieth [8] examined a set of 1386 marketed drugs and found that 15% of the drugs are embedded within other larger drugs, differing by one or more chemical fragments while 30% of drugs contain other drugs as building blocks. Recently, Franco *et al.* [9] analyzed scaffold diversity of 16 datasets of active compounds, targeting five protein classes, using an entropy-based information metric. They found that compounds targeted to the vascular endothelial growth factor receptor kinase, followed by compounds targeted to HIV reverse transcriptase and phosphodiesterase V, are maximally diverse. On the other hand, molecules in the glucocorticoid receptor, neuraminidase and glycogen phosphorylase  $\beta$  datasets are least diverse. Singh *et al.* [10] employed multiple criteria to compare libraries of drugs, small molecules and NPs, in terms of physicochemical properties, molecular scaffolds and fingerprints. The degree of overlap between libraries was assessed using the R-NN curve technique and the biologically relevant chemical space occupied by various compound datasets delineated. Hert *et al.* [11] compared a comprehensive dataset of 26 million compounds (i.e. the full chemical space) with 25810 purchasable screening compounds, metabolites, and natural product dataset. They found that almost 1300 ring systems present in NPs are missing in current day screening or lead libraries and suggest introducing bias in screening libraries towards molecules that are likely to bind protein targets. Khanna and Ranganathan [12] compared current day drugs with toxics and metabolites and found that drugs are more similar to toxics than to metabolites in physicochemical property space distribution.

However, there are only a few reports discussing the co-occurrence of fragments in pharmaceutically interesting datasets [13, 14], to identify an extensive list of fragments that may be used in drug discovery and to also provide a deeper insight into biologically relevant chemical space. In addition, questions such as how the occurrence of one fragment in a molecule is related to the occurrence of another and what are the most common fragments in a class of compounds need to be addressed. We believe that fragment co-occurrence data will help researchers to determine which



pairs of fragments are known to co-occur more often and therefore are more suitable for synthetic feasibility, and *vice-versa*.

In this study, we aim to answer questions such as 1) What is the physicochemical property space distribution of compounds for the datasets under comparison? 2) Are there any scaffolds or fragments missing in current lead libraries? 3) Are there any preferred or frequently occurring fragments and scaffolds in these datasets? 4) What fragments co-occur in drugs, metabolites and NPs? 5) What is the percentage similarity of drugs to other datasets?

We also report, for the first time in chemoinformatics, to the best of our knowledge, results on fragment co-occurrences, based on a data mining technique called association analysis (AA) [15]. We found patterns of commonly occurring fragments using extended connectivity functional class fingerprint (FCFP\_4; details in Methods section). FCFP is a variant of extended connectivity atom type (ECFP) fingerprint and the only difference between ECFP and FCFP fingerprint is the assignment of initial code [16]. The highly specific initial atoms types in ECFP fingerprints are replaced with more general atom types, with functional meaning in FCFP fingerprints. For example, a single initial code is assigned for all halogens in FCFP fingerprints as they can often substitute each other functionally. In accord with their definition, ECFP fingerprints are a better choice to measure diversity. Therefore, we used ECFP fingerprints for diversity analysis while the more generic FCFP fingerprints have been selected for Tanimoto and co-occurrence analyses. The fragment list and co-occurrence information obtained by association analysis can be used to design effective combinatorial lead libraries.

## Results and Discussion

Five different types of pharmaceutically relevant public molecular datasets were selected for this study: drugs, human metabolites, toxics, natural products and a sample of currently used lead compounds. Furthermore, we have also considered two popular small molecule databases *viz.* National Cancer Institute (NCI) database and ChEMBL database (details in the Methods section). Our results are presented in three sections, *viz.* preliminary analysis (calculating physicochemical



properties, measuring diversity and Tanimoto similarity), scaffold analysis and fragment co-occurrence analysis based on association rules.

After carefully pruning and filtering the datasets, all the datasets were clustered (see Methods section) to avoid biased results due to overrepresentation of similar molecules.

## **1. Preliminary analysis**

### **1.1 Physicochemical property analysis**

#### **1.1.1 Lipinski's properties for "rule of five" (Ro5) compliance:**

Ro5 has dominated drug design since 1997 and therefore, we believe it would be useful to analyze these datasets for compliance with the Ro5 test. Ro5 predicts passive and oral absorption based on log P, molecular weight, hydrogen bond donors and hydrogen bond acceptors. We have performed this test on both the clustered datasets as well as a subset of 2000 molecules, randomly selected from the clustered datasets, in order to check whether the random subsets are sufficient for other physicochemical property analyses. We report in Table 1, the percentage of molecules "failing" the Ro5 test, i.e. at least not meeting one condition of the Ro5 test. The results are comparable for both kinds of datasets, showing that randomly selected subsets are representative of the clustered datasets. Also, for the clustered datasets, initially, over 25% of drugs do not adhere to Ro5 while 20% of the metabolites are outside Lipinski's universe. Further, we found that similar to drugs, only 26.5% of the toxics fail the Ro5 test. This result highlights one of the shortcomings of Lipinski's rule, which overlooks toxicity resulting, in high attrition rates during drug discovery programs as has been reported in the literature [17, 18]. This is due to the fact that Lipinski's rule was originally designed to estimate bioavailability of compounds rather than toxicity. Further, we found that only 16% of NPs failed Lipinski's test. Many other related studies on NPs have reported similar results [7, 19]. Grabowski and Schneider [7] analyzed pure natural products from MEGAAbolite and Interbioscreen, natural products and derivatives from BioSpecs and marine natural products from the literature. They found that 18% of natural products, 30% of the marine natural products and only 8% of the natural product derived compounds violate Lipinski's rule, averaging 18.7%. While Grabowski and



Schneider have reported results very similar to ours, Ganesan [19] analyzed a focused set of 24 natural products that were the starting point for marketed drugs in the 25-year period from 1981-2006 and found that 50% of these failed Lipinski's rule. Overall, the results of Ganesan are in accord with our study that NPs adhere to Lipinski's rule, although in general, NPs do not necessarily abide by Lipinski's rule because they are thought to enter the human body not by passive diffusion but by more complex mechanisms such as active transportation, and so are not expected to comply with the rules for bioavailability. The probable explanation of our results could be the manner in which the NP dataset is pooled at the ZINC database. ZINC is a public database for commercially available compounds and NPs present in ZINC are pre-filtered to cover more drug-like space, contributing towards Ro5-like characteristics. Lead molecules on the other hand did reasonably well in the Ro5 test. This is in accordance with the lead-likeness concept proposed earlier [20] which states that leads should be simple, low molecular weight molecules and thus, should fall well within Lipinski's universe. Further, our results indicate that, NCI compounds follow Lipinski's rule more strictly than compounds present in ChEMBL dataset.

#### 1.1.2 Lipinski's properties as boxplots:

Box plots for Lipinski properties for random subsets are available from Additional File 1. We find that the mean value for the molecular weight in the metabolite dataset is relatively low when compared to the other datasets such as drugs, leads and natural products. We also observe that the lead dataset is well within Lipinski's universe and covers a fair amount of drug space. Further, we find a noticeable difference in lipophilicity values of metabolites as compared to drugs and leads. The mean value of lipophilicity (measured as AlogP) suggests that metabolites prefer a hydrophilic environment. Our results are comparable to the recent study using similar datasets [6]. In this study, lipophilicity (measured by a similar parameter, clogD) for drugs, metabolites and library compounds showed that the distribution of library compounds is similar to that of drugs, but differ markedly from metabolites and that metabolites are more hydrophilic than both drugs and library compounds.



### 1.1.3 Other physicochemical properties:

To comprehensively study the physicochemical property space distribution, we computed four more common whole molecule descriptors: the molecular polar surface area, the number of rotatable bonds, the molecular solubility and the number of rings (details in the Methods section). Distributions of these physicochemical properties as box plots are available from Additional File 2. We note that metabolites show relatively higher solubility, higher molecular polar surface area but lower complexity (less rings, less rotatable bonds and lower molecular weight) compared to drugs. Further, our results indicate that, in general, NCI molecules are also low molecular weight compounds with less complexity and slightly higher solubility than drug molecules. In addition, we note that a large part of the ChEMBL database contain drug-like compounds with higher molecular weight and more complex molecules than drugs.

## 1.2 Diversity analysis

In order to compare the diversity of features (fragments) present in each dataset, we have plotted the total number of non-redundant (nr) fingerprint features calculated, using ECFP fingerprints, up to order 8 (Figure 1). Our results indicate that overall, the ChEMBL dataset generates the maximum number of fragments and is highly diverse, while the metabolite dataset is the least diverse. From Figure 1a, we note that initially toxics outnumber other molecular datasets in generating features. This could be due to the high heteroatom content in toxics, resulting in large numbers of ECFP features generated during the first iteration step of fingerprinting. Similarly, the NCI dataset contains a large number of features during the initial iteration step of fingerprint feature generation. Metabolites, on the other hand, produce the least number of features, which suggests a limited occupancy of chemical space. Drugs were moderately diverse throughout and we find an increase in fragment diversity with increasing order of fingerprints.

## 1.3 Tanimoto analysis

The Tanimoto similarity coefficient (equation 1) compares two molecules, A and B, having  $N_A$  as the number of features in A,  $N_B$  as the number of features in B, and  $N_{AB}$  as the number of features common to both A and B:



$$T = \frac{N_{AB}}{(N_A + N_B - N_{AB})} \dots (1)$$

We extend this concept to compare different datasets used in this study. To calculate how similar two datasets are, we first calculated the Scitegic Pipeline Pilot connectivity fingerprints, FCFP\_4 (details in the Methods section) for all the datasets. Subsequently, a set of  $n_r$  fingerprint features ( $N_A$  and  $N_B$ ) was extracted for each dataset. Finally, the common features present in both datasets ( $N_{AB}$ ) were counted, by comparing the  $n_r$  fingerprint sets generated above, to determine  $T$ .

For the five different datasets described in the Methods section, as well as the two reference datasets, NCI and ChEMBL, the Tanimoto coefficient values are shown in Table 2. We note that the FCFP fingerprint patterns (of order 4; FCFP\_4) found in drugs are most similar to toxics (FCFP\_4: 0.3) than to any other dataset, except for the fingerprint patterns found in ChEMBL dataset. On the other hand, drugs are least similar to metabolites (FCFP\_4: 0.23). These observations are consistent with our earlier study on smaller datasets [12]. Further, we note that ChEMBL contains more drug-like fragments than any other biologically relevant fragment type present in this study (FCFP\_4: 0.36). Additionally, with the increasing order of fingerprints (FCFP\_6 and so on), although the number of fragments generated increases, the Tanimoto similarity coefficient values fall for all the datasets compared (data not shown). This suggests an inverse relationship between the size of the fragment and the probability of its occurrence in two separate datasets, i.e. the larger the fragment, the less likely that it will be found in the two datasets being compared.

## 2. Scaffold or cyclic system analysis

It is quite informative to study the molecular frameworks while comparing different datasets of chemical compounds. Since the publication of Bemis and Murcko [5], many attempts have been made to explore the chemical space occupied by bioactive scaffolds [21] as scaffold hopping remains an active area under research [22]. In this study, we define scaffolds as the core structure of the molecule after removing side chains but not the linkers, similar to the definition of *atomic frameworks* used by Bemis and Murcko. A detailed analysis of the total number of  $n_r$  scaffolds



present in the different datasets is available in Table 3. The percentage of singletons (scaffolds occurring only once) relative to the total number of scaffolds in a dataset has also been reported. In addition, we have tabulated the proportion of nr scaffolds containing aromatic and non-aromatic rings.

The drug dataset generates the largest proportion of nr scaffolds (50.0%) relative to the dataset size, followed by the toxics (42%), ChEMBL (33.4%), leads (32%) and NCI dataset (28%). Exceptionally low number of scaffolds in metabolites (14.3%) and natural products (21.2%) suggest lower scaffold diversity in these datasets. The higher scaffold diversity in drugs could be attributed to the fact that drugs are derived from various biologically relevant compounds thereby contributing to the scaffold diversity. Similarly, large number of scaffolds in toxic compound set is indicative of the high diversity of compounds with toxicity potential. Further, we note that distribution of scaffolds in all the datasets is highly skewed with large number of them occurring only once (singletons). In fact, almost 70% of the scaffolds in drugs, toxics, NCI and ChEMBL dataset occur only once. We also found that natural products comprise maximum number of recurring scaffolds ( $100 - \% \text{ of singletons} = 64\%$ ) followed by metabolites (38.9%) and leads (35.7%) suggesting that the compounds in these datasets revolve around certain preferred types of scaffolds. Our results agree with the recent study using similar natural product and drug dataset [10]. In their study, authors found high scaffold diversity in drugs (39.7%) while low diversity in natural products (17.9%) which is in accordance with our results. By counting the number of aromatic rings in nr scaffolds, we note that metabolites contain least number of aromatic rings (only 47.3% contain one or more aromatic rings in a scaffold) as compared to other datasets. 85% of the drugs on the other hand have scaffolds with aromatic rings. Furthermore, we note that 97.2% of the scaffolds found in lead dataset contain aromatic rings. There seems to be a bias towards aromatic ring containing scaffolds in presently used lead libraries.

The top five scaffolds and their relative percentages based on the total number of scaffolds found in each dataset are shown in Figure 2. Benzene is the most abundant scaffold system in all the datasets, particularly in metabolites (over 36%). Apart from metabolites, toxics (15%) and NCI



compounds (13%) also contain benzene in high percentages. Drugs and leads, on the other hand contain benzene in moderate amounts (10% and 7% respectively). While benzene is the most common scaffold type in NP (2.2%) and ChEMBL datasets (3.4%), the relative abundance of benzene in these datasets is far lower than that in the other datasets. Following benzene, pyridine is the second most commonly occurring scaffold type in four out of seven datasets: metabolites (5.2%), drugs (1%), leads (1%), and NCI (1.2%). We also note that steroid derivatives are largely present in drugs and NPs. Similarly, most of the fused large scaffolds are found in NPs (four of the top five scaffolds) followed by drugs and the ChEMBL dataset. Metabolites, on the other hand, seem to prefer smaller, less complex systems. Likewise, toxics and lead compounds also have few complex fused systems. Other commonly occurring scaffold systems are purine and purine derivatives (found mainly in metabolites and ChEMBL dataset), imidazole and biphenyls.

In Table 4, we tabulate the percentages of nr shared scaffolds between pairs of different datasets. From Table 4 we note that drugs and metabolites share 6% of the total nr scaffolds whereas NPs, leads and toxics share overall 2.4%, 1.4% and 7.5% of scaffolds with drugs, respectively. It is interesting to note that metabolites and leads do not share as many scaffolds (0.3%) as drugs and metabolites (6%). Due to the uneven size of the datasets, we have also reported the contribution of each dataset to the set of shared scaffolds. We find that of the total 296 nr scaffolds found in metabolites (Table 3), 123 (42%) are shared by drugs whereas only 68 (23%) are shared by the lead dataset. This suggests that lead compounds need further optimization to become more metabolite-like. Similarly, there seems to be little overlap between the scaffolds of presently used lead libraries and NPs (2.1%). Since metabolites and NPs are recognized by at least one protein in the biosphere, they seem to be appropriate candidates in lead library design. Our results however, indicate that neither metabolites nor NP scaffolds are being sampled enough while designing lead libraries. In addition, we note that over 7% of scaffolds are shared between drugs and toxics while metabolites and toxics share over 6% of the scaffolds, suggesting the recurrence of common scaffolds between these datasets. Compounds in the NCI and ChEMBL datasets are quite diversified; however, the NCI



dataset clearly contains more toxic scaffolds than the ChEMBL dataset. Furthermore, we note that large part of the drug scaffold space is present in NCI (45%) and ChEMBL (72%) implying that these datasets cover good amount of drug-like compounds. We also note that a large part of metabolite scaffold space is present in natural product (47%), NCI (78%) and ChEMBL (73%) datasets.

We expect that lead libraries biased towards molecules that biological targets have evolved to recognize, would yield better hits rates, than unbiased or universal libraries. Metabolites and NPs could potentially provide suitable lead molecules. Consequently, we further analyzed these datasets for the type of scaffolds that are currently missing in lead libraries. In fact, we note a very slight overlap in the scaffold space of lead libraries and these datasets as discussed above. We therefore, suggest that with the optimum coverage of biologically relevant scaffold space, hit rates in high throughput screening experiments can be improved. We report a set of scaffolds that occur in NPs (Figure 3) and metabolites (Figure 4), with a minimum Tanimoto similarity of 0.9 to the scaffolds found in drugs, which are actually missing in currently used lead datasets.

### 3. Co-occurrence (or Association) analysis

It would be interesting to determine which fragments occur together and in what order. Co-occurrence can be evaluated using association rules (AR) [23]. While Piatetsky-Shapiro [24] define AR analysis as the general problem of finding recurrent patterns in data, much of the work in the past has been concerned with finding relations between different items in a database of a sales transaction, typically of supermarket data. An AR is composed of two parts, an antecedent (head) and a consequent (body), and is usually denoted as antecedent  $\rightarrow$  consequent, where the presence of an antecedent in a database implies to some extent, the presence of the consequent. To determine the extent of this implication, two measures called *support* and *confidence* are most commonly reported. The value of *support* for a rule tells us in how many instances (rows or records) the rule (both antecedent and consequent) can be observed, usually as a fraction of the total number of instances. The value of *confidence* of the rule tells us what percentage of records containing the antecedent also contains the consequent of the rule. *Confidence* gives us an idea of the strength of the influence that



an antecedent has on the presence of a consequent of the rule. Clearly, the confidence measure is merely an estimate of the conditional probability of the consequent given the antecedent. The higher the confidence value, the more often the items co-occur. We report *support* and *confidence* as appropriate measures for fingerprint analysis.

The AR mining problem consists of finding all association rules existing in a database, having *support* and *confidence* values greater than a pre-defined threshold. In part A of Table 5, each row represents an item set and each column denotes an item. A Boolean variable is associated with each item. The presence of a specific item in a transaction is indicated by a value of “1” while “0” denotes its absence. Using the three items X, Y, Z in part A, we can develop example association rules (part B of Table 5), for which we can compute *support* and *confidence* values. *Support* for the first rule, where  $X,Y \rightarrow Z$ , is 0.5 (50%) with a *confidence* of 0.75 (75%) which implies that rule has 50% probability of occurrence and we can be 75% confident that Z will be present if both X and Y are present.

ARs were generated for all the datasets using FCFP\_4 fingerprints. Although there are many other fingerprints available we believe more the abstract nature of FCFP\_4 fingerprints serve our purpose well. It is often desirable to consider all halogens at par during lead design process and FCFP fingerprints are capable of doing the same. We tabulate the total numbers of association rules generated per dataset in the Table 6.

We observe that with an increase in the *support* value, the number of co-occurring fragments (or association rules) decreases. Metabolites produce the largest number of association rules (using FCFP\_4 fingerprints) even with the high *support* value of 0.2, while toxics and drugs produced an insignificant number of rules (1 and 0 respectively; *support* = 0.2). This signifies that metabolites occupy limited chemical space and therefore, tend to use selected fragments combinatorially to produce novel metabolites. Association analysis thus corroborates our earlier observation from our Tanimoto study that metabolites have limited diversity. On the other hand, the minimal number of ARs in drugs and toxics implies excellent diversity in these datasets. Table 7 lists the top two co-occurring fragments generated per dataset using FCFP\_4 fingerprints while for top five co-occurring



(see additional file 3). From Table 7 and additional file 3, we note that fragments appearing in different datasets are largely restricted to the dataset, which could prove useful in library design.

### **Conclusions:**

In this study, we have carried out a detailed analysis of commonly occurring fragments in various datasets of biological interest. For the first time, to the best of our knowledge, the data mining technique known as association analysis, was employed in chemoinformatics to search for patterns in fragments space. The results obtained give the frequencies and co-occurrences of these fragments in each dataset.

Dataset comparison using the Tanimoto coefficient shows that drugs and toxics share a large number of topological fragments whereas drugs are least similar to metabolites than to any other dataset studied. However, in scaffold analysis we found that current drugs and metabolites share 7.0% of the total nr scaffolds, i.e. over 42% of the metabolite scaffolds are present in drugs, whereas only 23% of the metabolite scaffolds are shared between leads and metabolites. This shows that although drugs and metabolites share many scaffolds, they largely differ in topological fragment space. Further, we conclude that current lead libraries do not cover much of metabolite scaffold space.

Library design is a multi-class optimization problem. It often presents a trade-off between several factors, including diversity and ADMET properties. Since metabolites and NPs are already optimized by millions of years of evolution to bind to at least one biological macromolecule therefore, it is highly likely that libraries designed based on the scaffolds and fragments occurring in metabolite and NP space will result in molecules with better ADMET properties. However, it should be kept in mind that metabolites occupy a limited space in chemical universe that limits the usage of metabolites in library design.

From physicochemical properties analysis, we note that there is a need to diversify present day lead libraries in order to optimize the coverage of chemical space. Our studies on scaffolds systems suggest that drugs are most diverse (50% scaffolds relative to the dataset size) and prefer aromatic to



non-aromatic ring-containing scaffolds. Metabolites, on the other hand, have a very narrow distribution of scaffolds (only 14.3% scaffolds relative to the dataset size) of which 38.9% recur. The exceptionally low number of cyclic systems in metabolites implies lower scaffold diversity in metabolites. This reaffirms our conclusions from Tanimoto and association analyses. Furthermore, we confirm earlier reports of skewed distribution of scaffolds, with many more singletons than recurring scaffolds.

Following on from these results, co-occurrence studies also revealed that metabolites utilize only a very limited set of fragments, while drug and toxic datasets contain a wide variety of fragments, indicating high molecular diversity. This result is in accordance with the fact that drugs are produced from different sources, like synthetic molecules, natural products, leads libraries, and hence, are diverse in distribution. The fragment list and co-occurrence information obtained by association analysis can be used to enrich combinatorial lead libraries for enhancing metabolite-likeness.

## **Methods**

### **Preparation of datasets**

Five different types of biologically relevant molecular datasets have been considered in this study. Beside these, the contents of public databases like NCI and ChEMBL were also analyzed. Table 8 presents a summary of all the databases used in this study. The drug dataset was assembled by merging molecules obtained from the DrugBank [25] and a subset of Kyoto Encyclopedia of Genes and Genomes database (KEGG DRUG) [26]. DrugBank is a comprehensive resource on drugs and includes over 1350 FDA-approved small drugs. KEGG is a bioinformatics resource and currently provides 19 databases; we used the KEGG DRUG subset as it contains all the drugs approved in the USA and Japan. It not only contains prescription drugs but also “over-the-counter” (OTC) drugs. Similarly, for metabolite dataset we used the Human Metabolome Database (HMDB) [27], HumanCYC [28] database and BiGG [29] database. HMDB contains information on nearly 8,000 metabolites found in the human body. HumanCYC is a bioinformatics database that combines human metabolic pathway and genome information, providing KEGG, PubChem and ChEBI



identifiers for the metabolites present in this database. BiGG stores manually annotated human metabolic network information, with links to KEGG metabolites.

Likewise, for the toxics dataset, compounds from various public sources were integrated to make a single dataset focusing largely on carcinogenic molecules. The Distributed Structure-Searchable Toxicity (DSSTox) Carcinogenic Potency Database [30] contains experimental results and carcinogenicity information for 1547 substances tested against different species. Contrera *et al.* [31] published a dataset of 282 human pharmaceuticals obtained from FDA database for carcinogenicity studies on mouse and rat. They reported 125 (44% of the above 282) of the positive chemicals that were used in this study. Toxicology Excellence for Risk Assessment (TERA) is an independent non-profit organization dedicated to the public health. Since 1996, TERA has maintained an International Toxicity Estimate for Risk database [32] which provides chronic human risk assessment data from organization around the world for over 650 chemicals [33]. Finally, ~1000 molecules with medium and high toxicity were downloaded from the SuperToxic database [34]. The dataset for NPs was obtained from the ZINC database[35]. These molecules can be searched under the subset tab, as “Meta subsets”. For lead dataset, we merged two independent screening sets obtained from BioNET [36] and Maybridge database [37]. The molecules in these two databases are well diversified and we integrated them to form a dataset of lead compounds as found in pharmaceutical collections. Further, we included molecules from NCI open database [38]. The latest September 2003 release of the database stores 260071 organic compounds tested by NCI for anticancer activity. Since many of the compounds are experimental, have not been tested for human consumption and covers high diversity therefore, we believe it would be good choice to include this dataset in our study. One other public dataset, ChEMBL [39] was used as the reference dataset for biologically interesting molecules. ChEMBL is a chemogenomics data resource with over 8000 targets and about 622,884 bioactive compounds.

All datasets are current as of 10-November-2010.



### **Cleaning and processing of the datasets**

We followed a standard cleaning procedure (see additional file 4) to obtain a nr dataset in each category. Finally, clustering was performed to address the issue of possible overrepresentation of the chemical space, which might bias the analysis results towards similar molecules [6]. Clusters were generated, using the Cluster “Clara” algorithm embedded in the Pipeline Pilot (PP) software [40] by employing an atom type fingerprint as a chemical descriptor and Euclidean distance was the distance metric selected. Cluster centers served as the representatives for clusters containing more than one molecule while singletons were directly used as cluster centers. This resulted in 30% decreases of each dataset. Upon further analysis, we found that clustered metabolite set contains lipids in large numbers. In order to remove the bias towards lipids and large molecules, we filtered out lipids resulting in 2072 molecules in the “lipid-free” metabolite dataset, used for analysis in this study.

To simplify the analysis, we randomly selected 2000 compounds from each of the clustered datasets and lipid-free metabolite dataset in case of metabolites. The majority of the analysis was carried out using the clustered datasets and lipid-free metabolite dataset, except for preliminary analysis, where these randomly selected molecules were used and in the case of Ro5 test, where both datasets were compared.

### **Molecular descriptors**

All the descriptors were calculated using PP. Beside the four Lipinski properties: molecular weight, the number of hydrogen bond acceptors, AlogP (a hydrophobicity measure) and the number of hydrogen bond donors [4], other descriptors such as molecular polar surface area (MPSA), molecular solubility (MS), the number of rings (NR) and the number of rotatable bonds (NRB) were also computed. AlogP was calculated using the Ghose-Crippen method [41] which takes into account the group’s contribution to Log P. MPSA is defined as the sum over all the polar atoms. This descriptor is correlated with drug transport capabilities and is important in penetrating the blood-brain barrier. The NRB is a direct measure of the flexibility of molecules thus related to MPSA. Binary descriptors (ECFP\_4 and FCFP\_4) were calculated using a structural property calculator embedded in PP. Initially, each atom is assigned a code based on its properties and



connectivity. With increasing iteration, each atom code is combined with the code of its immediate neighbours to produce the next order code. This process is repeated until the desired number of iterations has been achieved, typically to four iterations, generating ECFP<sub>4</sub>, or FCFP<sub>4</sub> fingerprints.

### **Cyclic systems**

In addition to examining the physicochemical properties, each dataset was also explored for the frequent scaffold systems. We used an inbuilt PP protocol to identify the most common fragments, by setting “FragmentType” to MurckoAssemblies and adjusting “MaxFragSize” parameter at the required level.

### **Association rules**

We employed the cover rules algorithm described by Cristofor and Simovici [42] for calculating ARs that are implemented in the java-based open source software ARtool [43]. ARtool has a special file format requirement for the data analysis. Therefore, all the files containing fingerprints were converted into the correct input format. Before that, a nr fingerprint set was extracted from the fingerprint files containing information regarding fingerprints and corresponding SMILES patterns. Since higher order extended connectivity fingerprints include all the information of lower order fingerprints therefore, it is important to filter out lower order fingerprints from the nr set of higher order fingerprints, in our case we filtered out FCFP<sub>2</sub> fingerprints from FCFP<sub>4</sub> set. Finally, an inbuilt utility in ARtool called asc2db was used to convert these asc files into db files that can be read by ARtool for further processing.

### **Author contributions**

VK curated the dataset and conducted the analysis work; SR directed the study and both authors prepared and approved the manuscript.

**Conflict of interest:** none declared.

### **Acknowledgements**



VK is grateful to Macquarie University for the award of MQRES research scholarship. Open access publication charges for this article were covered by Macquarie University.



## References:

1. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE: **Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors.** *J Med Chem* 1996, **39**(16):3049-3059.
2. Dobson CM: **Chemical space and biology.** *Nature* 2004, **432**(7019):824-828.
3. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH: **PubChem: a public information system for analyzing bioactivities of small molecules.** *Nucleic Acids Res* 2009, **37**(Web Server issue):W623-633.
4. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Deliv Rev* 2001, **46**(1-3):3-26.
5. Bemis GW, Murcko MA: **The properties of known drugs. 1. Molecular frameworks.** *J Med Chem* 1996, **39**(15):2887-2893.
6. Dobson PD, Patel Y, Kell DB: **'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries.** *Drug Discov Today* 2009, **14**(1-2):31-40.
7. Grabowski K, Schneider G: **Properties and Architecture of Drugs and Natural Products Revisited.** *Curr Chem Biol* 2007, **1**:115-127.
8. Siegel MG, Vieth M: **Drugs in other drugs: a new look at drugs as fragments.** *Drug Discov Today* 2007, **12**(1-2):71-79.
9. Medina-Franco JL, Martinez-Mayorga K, Bender A, Sciore T: **Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure.** *QSAR Comb Sci* 2009, **28**(11-12):1551-1560.
10. Singh N, Guha R, Giulianotti MA, Pinilla C, Houghten RA, Medina-Franco JL: **Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository.** *J Chem Inf Model* 2009, **49**(4):1010-1024.
11. Hert J, Irwin JJ, Laggner C, Keiser MJ, Shoichet BK: **Quantifying biogenic bias in screening libraries.** *Nat Chem Biol* 2009, **5**(7):479-483.
12. Khanna V, Ranganathan S: **Physiochemical property space distribution among human metabolites, drugs and toxins.** *BMC Bioinformatics* 2009, **10** Suppl 15:S10.
13. Lameijer EW, Kok JN, Back T, Ijzerman AP: **Mining a chemical database for fragment co-occurrence: discovery of "chemical cliches".** *J Chem Inf Model* 2006, **46**(2):553-562.
14. Batista J, Bajorath J: **Mining of randomly generated molecular fragment populations uncovers activity-specific fragment hierarchies.** *J Chem Inf Model* 2007, **47**(4):1405-1413.
15. Brin S, Motwani R, Ullman J, Tsur S: **Dynamic itemset counting and implication rules for market basket data.** In: *Int Conf on Management of Data: 1997; New York, USA: ACM SIGMOD 1997* 1997: 255-264.
16. Rogers D, Hahn M: **Extended-connectivity fingerprints.** *J Chem Inf Model* 2010, **50**(5):742-754.
17. Schuster D, Laggner C, Langer T: **Why drugs fail--a study on side effects in new chemical entities.** *Curr Pharm Des* 2005, **11**(27):3545-3559.
18. Gut J, Bagatto D: **Theragenomic knowledge management for individualised safety of drugs, chemicals, pollutants and dietary ingredients.** *Expert Opin Drug Metab Toxicol* 2005, **1**(3):537-554.
19. Ganesan A: **The impact of natural products upon modern drug discovery.** *Curr Opin Chem Biol* 2008, **12**(3):306-317.
20. Oprea TI, Davis AM, Teague SJ, Leeson PD: **Is there a difference between leads and drugs? A historical perspective.** *J Chem Inf Comput Sci* 2001, **41**(5):1308-1315.
21. Wetzel S, Klein K, Renner S, Rauh D, Oprea TI, Mutzel P, Waldmann H: **Interactive exploration of chemical space with Scaffold Hunter.** *Nat Chem Biol* 2009, **5**(8):581-583.
22. Krueger BA, Dietrich A, Baringhaus KH, Schneider G: **Scaffold-hopping potential of fragment-based de novo design: the chances and limits of variation.** *Comb Chem High Throughput Screen* 2009, **12**(4):383-396.



23. Aggarwal C, Yu P: **A new framework for itemset generation**. In: *In Proc of the 17th Symposium on Principles of Database Systems: 1998; Seattle, WA; 1998*: 18-24.
24. Piatetsky-Shapiro G: **Discovery, analysis, and presentation of strong rules**. In: *Knowledge Discovery in Databases*. Edited by Frawley W. Cambridge, MA: AAAI/MIT Press; 1991.
25. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic Acids Res* 2008, **36**(Database issue):D901-906.
26. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T *et al*: **KEGG for linking genomes to life and the environment**. *Nucleic Acids Res* 2008, **36**(Database issue):D480-484.
27. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S *et al*: **HMDB: a knowledgebase for the human metabolome**. *Nucleic Acids Res* 2009, **37**(Database issue):D603-610.
28. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD: **Computational prediction of human metabolic pathways from the complete human genome**. *Genome Biol* 2005, **6**(1):R2.
29. Schellenberger J, Park JO, Conrad TM, Palsson BO: **BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions**. *BMC Bioinformatics*, **11**:213.
30. Richard AM, Williams CR: **Distributed structure-searchable toxicity (DSSTox) public database network: a proposal**. *Mutat Res* 2002, **499**(1):27-52.
31. Contrera JF, Jacobs AC, DeGeorge JJ: **Carcinogenicity testing and the evaluation of regulatory requirements for pharmaceuticals**. *Regul Toxicol Pharmacol* 1997, **25**(2):130-145.
32. **International Toxicity Estimate for Risk database (TERA)**. [<http://www.tera.org/iter>]
33. Wullenweber A, Kroner O, Kohrman M, Maier A, Dourson M, Rak A, Wexler P, Tomljanovic C: **Resources for global risk assessment: the International Toxicity Estimates for Risk (ITER) and Risk Information Exchange (RiskIE) databases**. *Toxicol Appl Pharmacol* 2008, **233**(1):45-53.
34. Schmidt U, Struck S, Gruening B, Hossbach J, Jaeger IS, Parol R, Lindequist U, Teuscher E, Preissner R: **SuperToxic: a comprehensive database of toxic compounds**. *Nucleic Acids Res* 2009, **37**(Database issue):D295-299.
35. Irwin JJ, Shoichet BK: **ZINC--a free database of commercially available compounds for virtual screening**. *J Chem Inf Model* 2005, **45**(1):177-182.
36. **BioNET**. [<http://www.keyorganics.ltd.uk/scdownloads.htm>]
37. **Maybridge**. [<http://www.maybridge.com/default.aspx>]
38. **National Cancer Institute (NCI)**. [<http://cactus.nci.nih.gov/download/nci/>]
39. Overington J: **ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr**. *J Comput Aided Mol Des* 2009, **23**(4):195-198.
40. **SciTegic Pipeline Pilot**. [<http://accelrys.com/products/scitegic/>]
41. Ghose AK, Crippen GM: **Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions**. *J Chem Inf Comput Sci* 1987, **27**(1):21-35.
42. Cristofor L, Simovici D: **Generating an informative cover for association rules**. In: *In Proc of the IEEE International Conference on Data Mining: 2002; 2002*.
43. **ARtool**. [<http://www.cs.umb.edu/~laur/ARtool/>]
44. Milne GW, Nicklaus MC, Driscoll JS, Wang S, Zaharevitz D: **National Cancer Institute Drug Information System 3D database**. *J Chem Inf Comput Sci* 1994, **34**(5):1219-1224.



## FIGURE LEGENDS

**Figure 1.** The number of non-redundant fingerprint features as a function of ECFP fingerprint order.

Fingerprints of orders 2, 4, 6 and 8 for datasets comprising drugs, metabolites, toxics, natural products, leads, NCI and ChEMBL are presented.

**Figure 2.** Top 5 scaffolds derived from A. drugs, B. metabolites, C. toxins, D. natural products, E. leads, F. NCI and G. ChEMBL. The extent of occurrence of the scaffold relative to the total number of scaffolds in the dataset (as %) are listed.

**Figure 3.** A set of scaffolds present in metabolites but are missing in lead dataset. The Tanimoto distance of these scaffolds with the closest counterparts in drugs is also reported.

**Figure 4.** A set of scaffolds present in NPs but are missing in lead dataset. The Tanimoto distance of these scaffolds with the closest counterparts in drugs is also reported.

### Additional File 1: (\*.pdf)

**Figure S1:** Box plots for the Lipinski physicochemical properties: (a) molecular weight, (b) the number of hydrogen bond acceptors, (c) AlogP and (d) the number of hydrogen bond donors.

### Additional File 2: (\*.pdf)

**Figure S2:** Box plots for other significant physicochemical properties: (a) molecular polar surface area, (b) the number of rotatable bonds, (c) molecular solubility and (d) the number of rings.

### Additional File 3: (\*.pdf)

**Table S1.** Top five association rules for various datasets using FCFP\_4 fingerprints at minimum support 0.1 and confidence 0.5.

### Additional File 4: (\*.pdf)

**Figure S3:** Flowchart adapted for the overall methodology.



**Table 1. Number of molecules failing Lipinski's "rule of five" (Ro5).**

<b>Dataset</b>	<b>Total no. of molecules (in clustered dataset)</b>	<b>% of molecules failing Ro5 in clustered datasets</b>	<b>% of molecules failing Ro5 in randomly selected sets</b>
<b>Drugs</b>	3788	25.7	23.0
<b>Metabolites</b>	6124	68.0	20.0*
<b>Toxics</b>	2166	26.5	21.5
<b>NPs</b>	61972	16.2	15.0
<b>Leads</b>	67983	19.8	19.5
<b>NCI</b>	161336	19.5	15.5
<b>ChEMBL</b>	379827	36.4	36.0

\*Metabolite dataset excluding lipids and large molecules (details in the Methods section)



**Table 2: Tanimoto similarity values using circular connectivity fingerprint descriptors for different datasets under study.**

The upper half of the diagonal contains similarity values calculated using FCFP\_4 fingerprint.

Datasets	Drugs	Metabolites	Toxics	NPs	Leads	NCI	ChEMBL
Drugs	1	0.23	0.30	0.26	0.26	0.30	0.36
Metabolites		1	0.19	0.20	0.16	0.18	0.19
Toxics			1	0.22	0.21	0.27	0.27
NPs				1	0.24	0.24	0.25
Leads					1	0.27	0.28
NCI						1	0.29
ChEMBL							1



**Table 3. Scaffold analysis of various datasets under study.**

Frequency of occurrence for non-redundant scaffolds (relative to the dataset size) and number of aromatic ring containing scaffolds (relative to the total number of nr scaffolds) have been reported in table.

Dataset	Occurrence of scaffolds (% relative to dataset size)		No. of singletons (% relative to number of scaffolds)		Aromatic scaffolds (% relative to number of scaffolds)	
	No.	%	No.	%	No.	%
Drugs	1874	50.0	1411	75.3	1588	85.0
Metabolites	296	14.3	181	61.1	140	47.3
Toxics	905	42.0	689	76.1	656	72.3
NPs	13151	21.2	6053	46.0	11776	90.0
Leads	21621	32.0	13819	64.0	21057	97.4
NCI	44324	28.0	31880	72.0	36778	83.0
ChEMBL	126843	33.4	87750	69.2	119419	94.1



**Table 4: Scaffolds shared between pairs of datasets.** The overall percentage of shared scaffolds is given in the brackets, along with percentages of shared scaffolds from each contributing dataset. D: Drugs, M: Metabolites, T: Toxics, P: Natural Products, L: Leads, N: NCI, C: ChEMBL.

Datasets	D	M	T	P	L	N	C
D	100%	123 (6%; D: 7%, M: 42%)	192 (7.5%; D: 10%, T: 21%)	347 (2.4%; D: 19%, P: 3%)	310 (1.4%; D: 17%, L: 1%)	840 (2%; D: 45%, N: 2%)	1347 (1.0%; D: 72%, C: 1%)
M		100%	71 (6.3%; M: 24%, T: 8%)	140 (1.1%; M: 47%, P: 1%)	68 (0.3%; M: 23%, L: 0.3%)	230 (0.5%; M: 78%, N: 0.5%)	215 (0.2%; M: 73%, C: 0.2%)
T			100%	174 (1.3%; T: 19%, P: 1%)	144 (0.7%; T: 16%, L: 1%)	534 (1.2%; T: 59%, N: 1%)	532 (0.4%; T: 59%, C: 0.4%)
P				100%	706 (2.1%; P: 5%, L: 3%)	1734 (3.1%; P: 13%, L: 8%)	1947 (1.4%; P: 15%, C: 1.5 %)
L					100%	2753 (4.4%; L: 13%, N: 6%)	3470 (2.4%; L: 16%, C: 3%)
N						100%	7600 (5.0%; N: 17%, C: 6%)
C							100%



**Table 5. An example Boolean matrix and association rules derived from it.**

X, Y and Z are items and A and C refer to the antecedent and consequent in association rules involving these items.

Category	X	Y	Z
<b>A. Example of transactions</b>	1	1	1
	0	1	1
	1	0	1
	1	1	0
	1	1	1
	1	1	1
<b>B. Examples of association rules using the above transactions</b>	<b>A <math>\rightarrow</math> C</b>	<b>Support</b>	<b>Confidence</b>
	X, Y $\rightarrow$ Z	3/6 = 0.5	3/4 = 0.75
	X $\rightarrow$ Y, Z	3/6 = 0.5	3/5 = 0.6

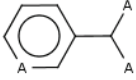
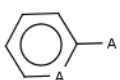
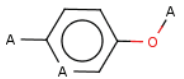
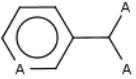
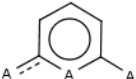
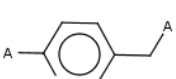
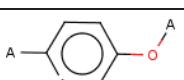
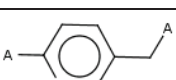
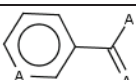
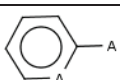
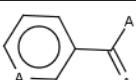
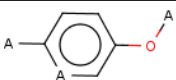
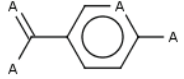
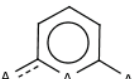
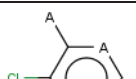
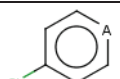
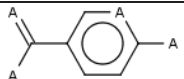
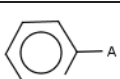
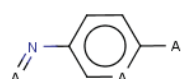
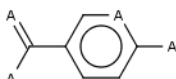
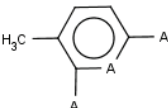
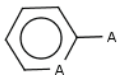
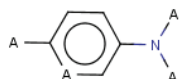
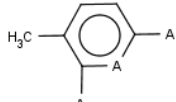


**Table 6.** Association rules generated using binary fingerprints at different support levels.

Datasets	Number of association rules generated			
	Using FCFP_4 at following min. support values and confidence 0.5			
	0.02	0.05	0.1	0.2
Drugs	147	24	4	1
Metabolites	>40000	146	96	27
Toxics	66	8	2	0
NP	532	94	21	5
Leads	220	44	13	4
NCI	77	13	3	1
ChEMBL	298	52	11	1



**Table 7.** Top two association rules for various datasets using FCFP<sub>4</sub> fingerprints at minimum support 0.1 and confidence 0.5.

Datasets	S.No.	Antecedent	Consequent	Support	Confidence
<b>Drugs</b>	1.			0.31	0.64
	2.			0.14	0.67
<b>Toxics</b>	1.			0.19	0.60
	2.			0.11	0.59
<b>NPs</b>	1.			0.39	0.65
	2.			0.34	0.56
<b>Leads</b>	1.			0.35	0.59
	2.			0.26	0.73
<b>NCI</b>	1.			0.32	0.69
	2.			0.15	0.64
<b>ChEMBL</b>	1.			0.42	0.67
	2.			0.22	0.73

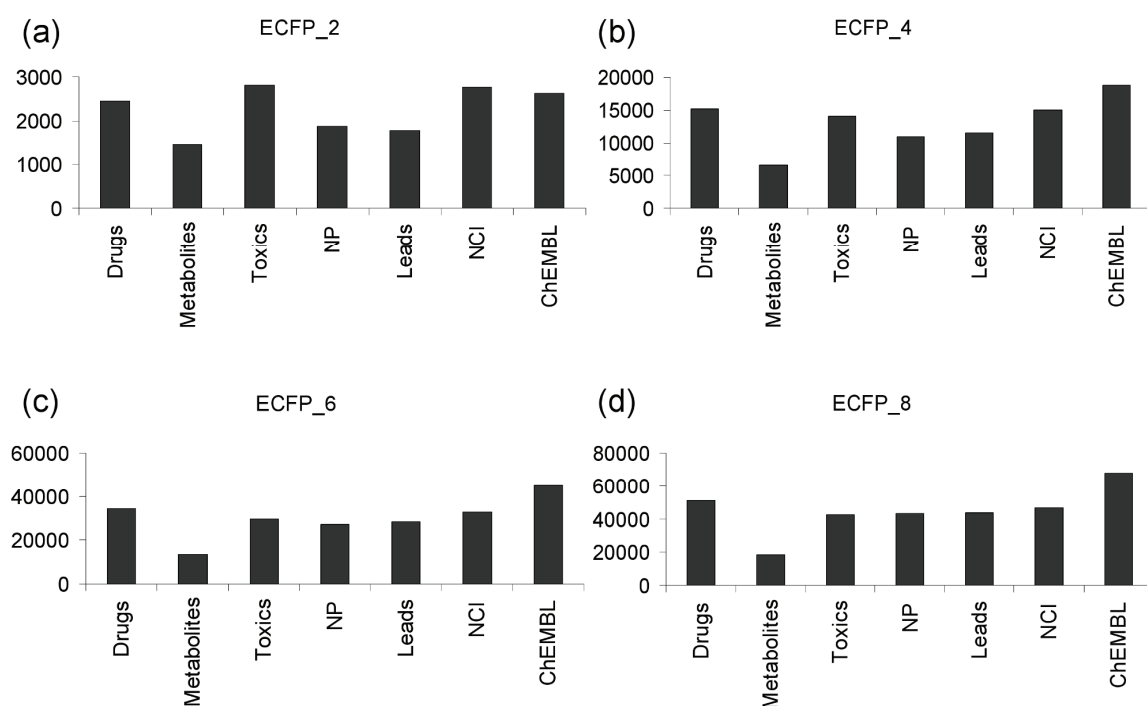


**Table 8:** Databases used in this study

Datasets		Number of molecules	Clustered dataset	Reference
Drugs	DrugBank	1372	3788	[25]
	KEGG drugs	7057		[26]
Metabolites	HMDB	7888	6124, 2072*	[27]
	HumanCYC	984		[28]
	BiGG	730		[29]
Toxics	DSSTox	582	2166	[30]
	FDA Carcinogenicity	125		[31]
	ITER	514		[33]
	SuperToxic	1097		[34]
NPs	ZINC NP database	89425	61972	[35]
Leads	BioNET	42699	67983	[36]
	Maybridge	60550		[37]
NCI	NCI database	260071	161336	[44]
ChEMBL	ChEMBL dataset	600625	379827	[39]

\*Metabolite dataset excluding lipids and large molecules (details in the Methods section)

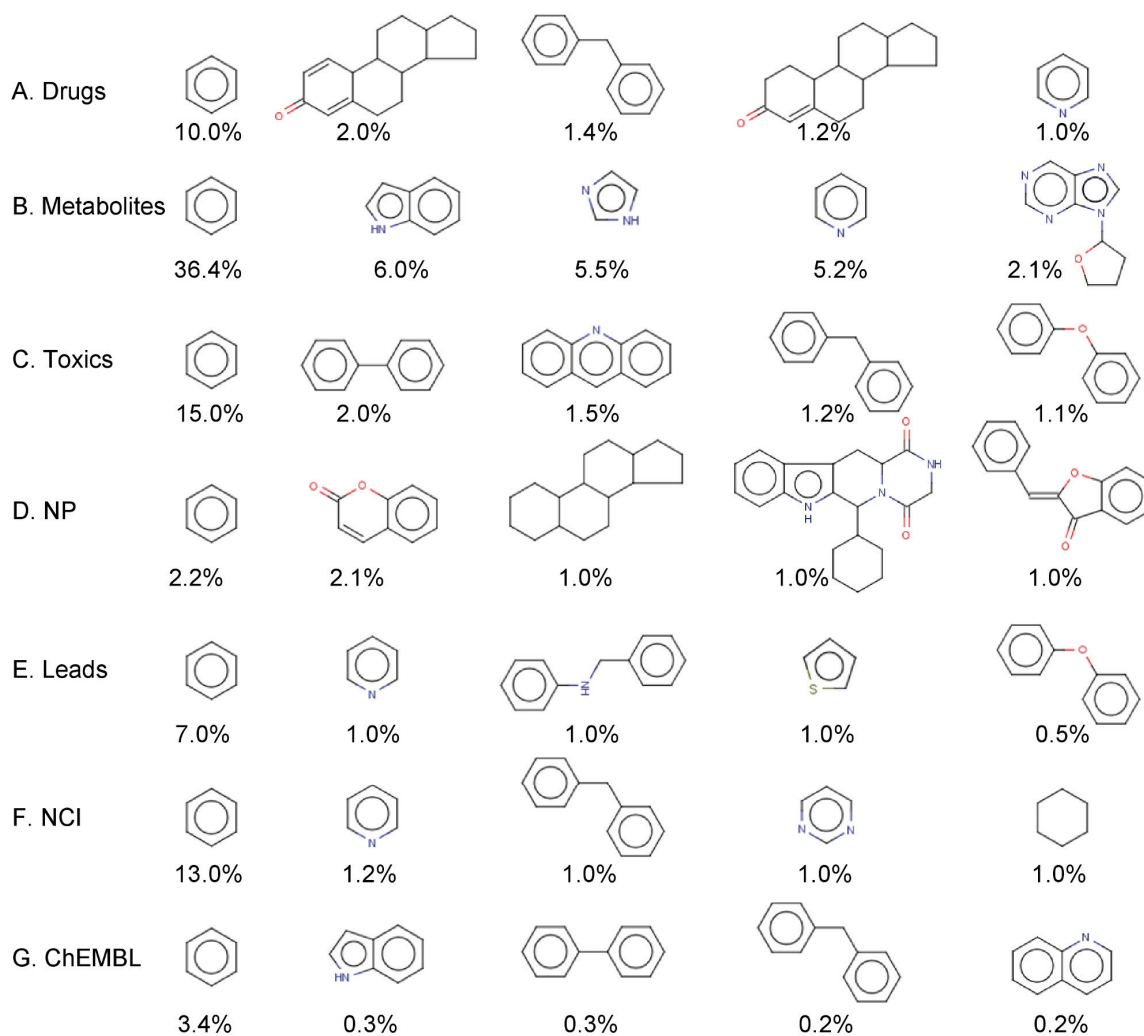




**Figure 1. The number of non-redundant fingerprint features as a function of ECFP fingerprint order.**

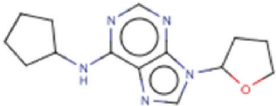
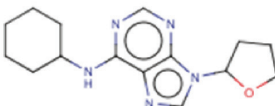
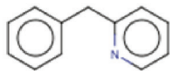
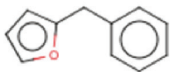
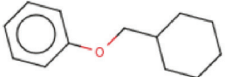
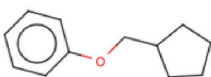


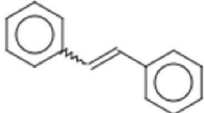
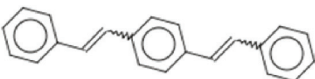
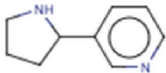
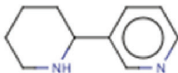
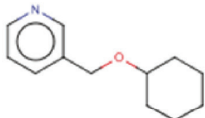
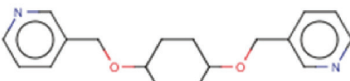
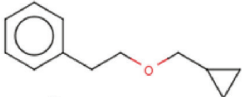
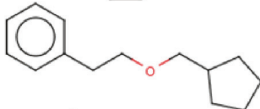
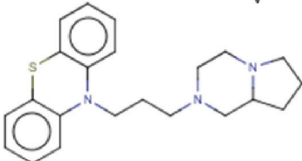
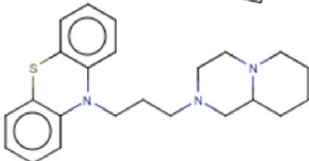
Fingerprints of orders 2, 4, 6 and 8 for datasets comprising drugs, metabolites, toxics, natural products, leads, NCI and ChEMBL are presented.





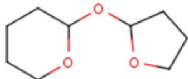
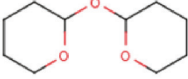
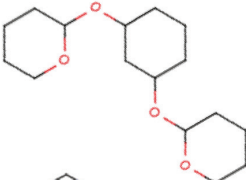
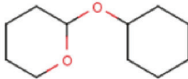
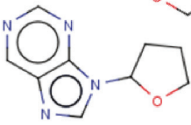
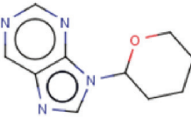
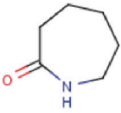
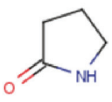
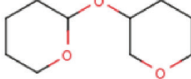
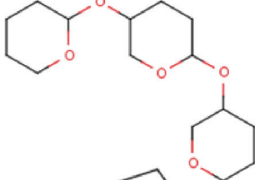
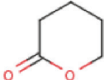
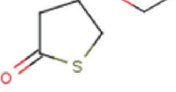
**Figure 2.** Top 5 scaffolds derived from A. drugs, B. metabolites, C. toxins, D. natural products, E. leads, F. NCI and G. ChEMBL. The extent of occurrence of the scaffold relative to the total number of scaffolds in the dataset (as %) are listed.



Scaffolds found in drugs	Scaffolds found in NPs but not in lead dataset	Tanimoto distance using FCFP <sub>4</sub>
		0.98
		0.96
		0.96
		0.95
		0.93
		0.91
		0.91
		0.90
		0.90

**Figure 3.** A set of scaffolds present in metabolites but are missing in lead dataset. The Tanimoto distance of these scaffolds with the closest counterparts in drugs is also reported.



Scaffolds found in drugs	Scaffolds found in metabolites but not in lead dataset	Tanimoto distance using FCFP_4
		0.95
		0.94
		0.93
		0.93
		0.93
		0.92

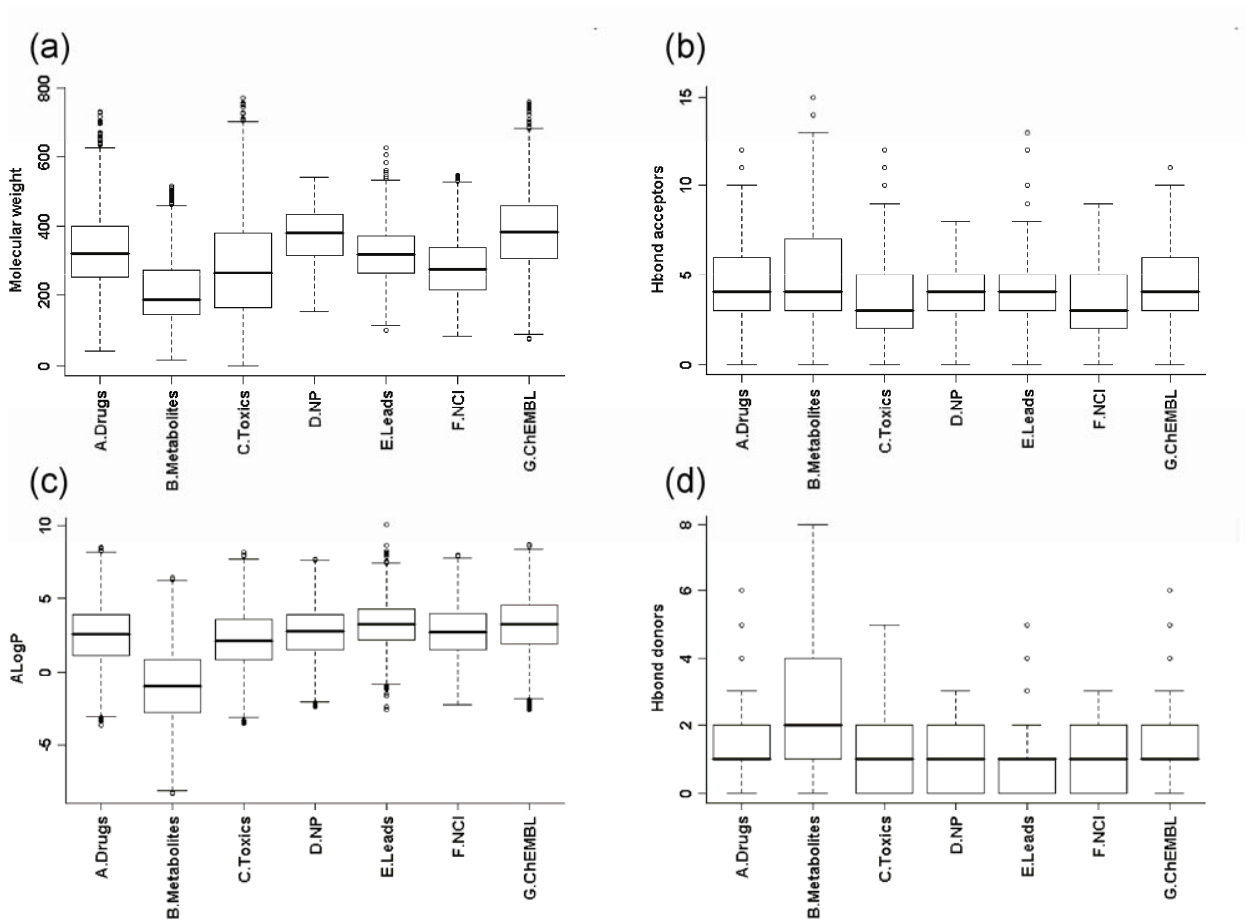
**Figure 4.** A set of scaffolds present in NPs but are missing in lead dataset. The Tanimoto distance of these scaffolds with the closest counterparts in drugs is also reported.



## Additional File 1

### Scaffold and fragment co-occurrence studies on datasets of biological interest

Varun Khanna, Shoba Ranganathan



**Figure S1: Box plots for the Lipinski physicochemical properties.**

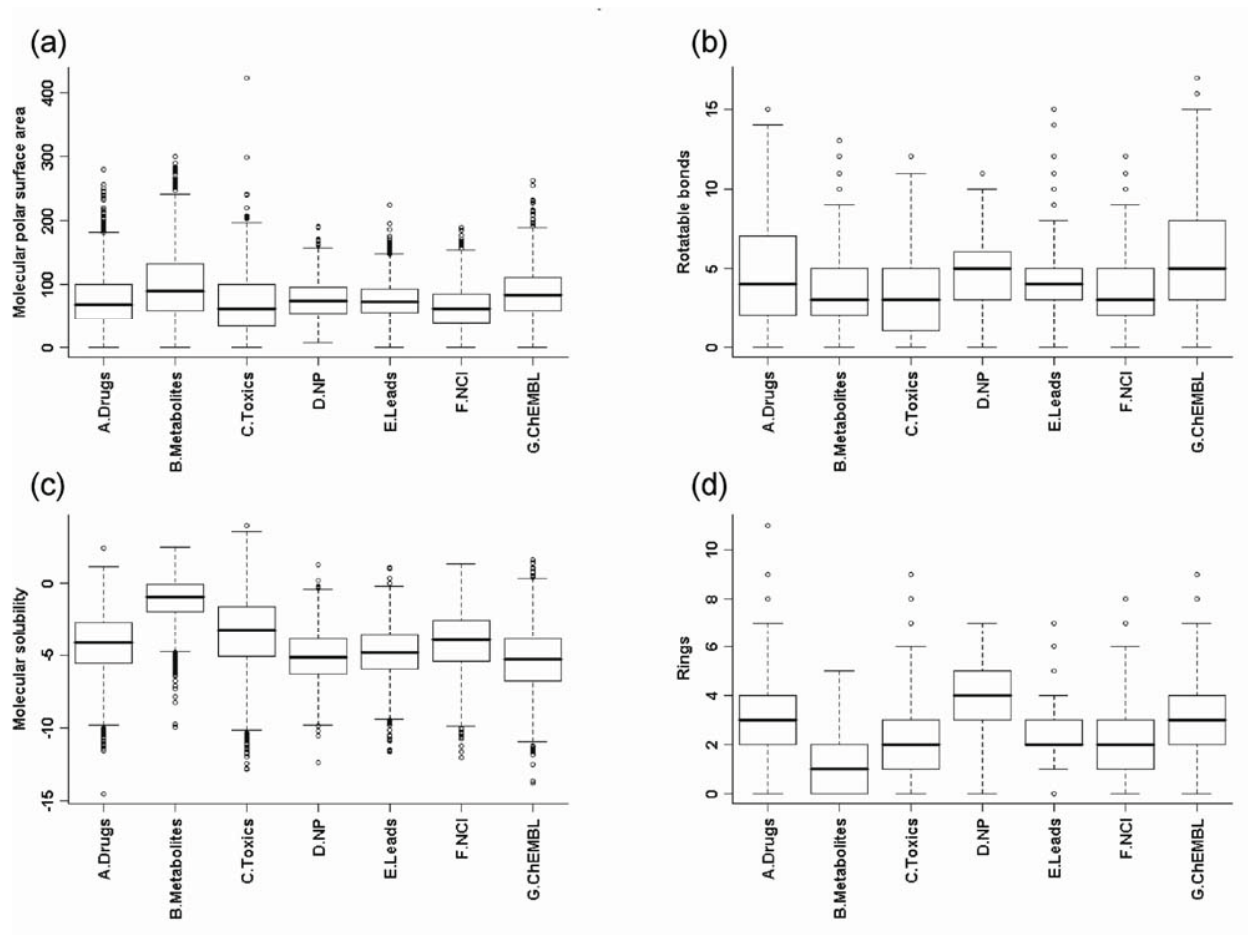
The randomly selected subsets of drugs, metabolites, toxics, natural products, NCI and ChEMBL are obtained from their respective clustered datasets. The physicochemical properties plotted are (a) molecular weight, (b) the number of hydrogen bond acceptors, (c) AlogP and (d) the number of hydrogen bond donors.



## Additional File 2

### Scaffold and fragment co-occurrence studies on datasets of biological interest

Varun Khanna, Shoba Ranganathan



**Figure S2: Box plots for other significant physicochemical properties.**

The randomly selected subsets of drugs, metabolites, toxics, natural products, NCI and ChEMBL are obtained from their respective clustered datasets. The physicochemical properties plotted are (a) molecular polar surface area, (b) the number of rotatable bonds, (c) molecular solubility and (d) the number of rings.

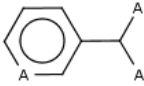
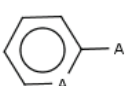
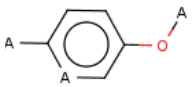
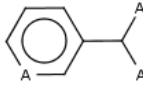
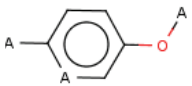
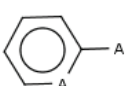
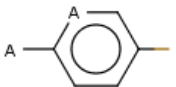
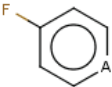
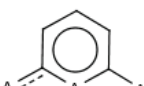
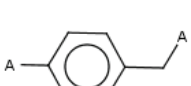
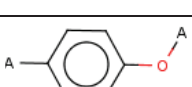
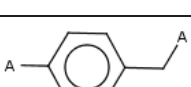
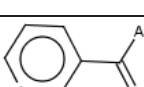
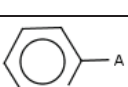
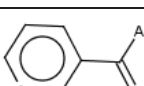
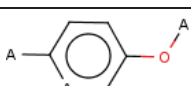
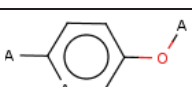
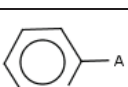


## Additional File 3

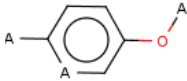
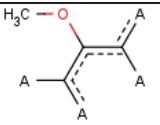
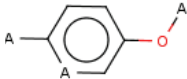
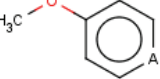
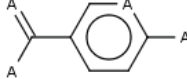
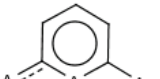
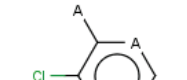
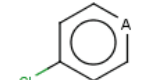
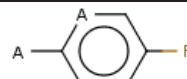
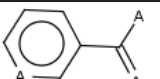
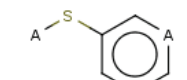
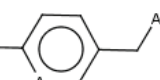
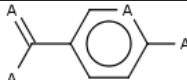
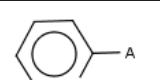
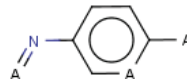
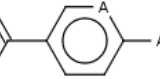
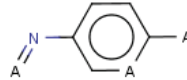
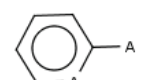
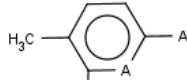




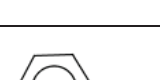
### Scaffold and fragment co-occurrence studies on datasets of biological interest

Varun Khanna, Shoba Ranganathan

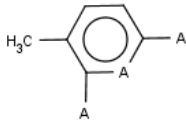
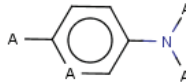
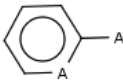
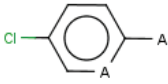
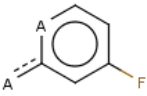
**Table S1. Top five association rules for various clustered datasets used in this study.** FCFP<sub>4</sub> is the fingerprint employed to calculate association rules at minimum support level 0.1 and confidence 0.5.

Datasets	S.No.	Antecedent	Consequent	Support	Confidence
<b>Drugs</b>	1.			0.31	0.64
	2.			0.14	0.67
	3.			0.12	0.56
	4.			0.11	0.78
<b>Toxics</b>	1.			0.19	0.60
	2.			0.11	0.59
<b>NPs</b>	1.			0.39	0.65
	2.			0.34	0.56
	3.			0.23	0.55



	4.			0.22	0.53
	5.			0.21	0.50
<b>Leads</b>	1.			0.35	0.59
	2.			0.26	0.73
	3.			0.21	0.61
	4.			0.21	0.60
<b>NCI</b>	1.			0.32	0.69
	2.			0.15	0.64
	3.			0.14	0.60
<b>ChEMBL</b>	1.			0.42	0.67
	2.			0.22	0.73
	3.			0.18	0.60



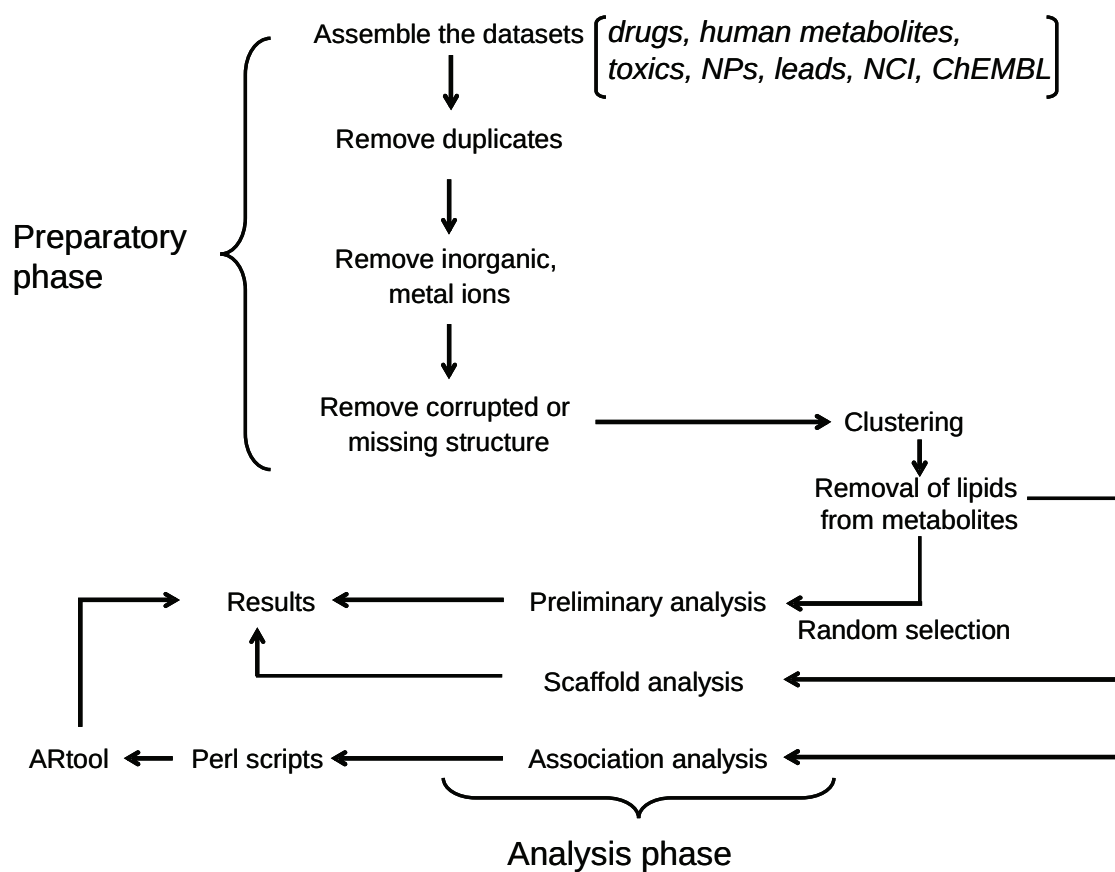
	4.	  	0.14	0.64
	5.	 	0.14	0.72



## Additional File 4

### Scaffold and fragment co-occurrence studies on datasets of biological interest

Varun Khanna, Shoba Ranganathan



**Figure S3: Flowchart adapted for the overall methodology.**



## 5.2 Conclusion

From physicochemical analysis, we corroborated our earlier finding (Chapter 4 and Publication 4) that metabolites occupy a distinct physicochemical property space. In this study, we have carried out a detailed analysis of commonly occurring fragments in various datasets of biological interest.

We found that over 42% of the metabolite scaffolds are present in drugs, whereas only half of that amount is shared between leads and metabolites. Similarly, we found that current lead libraries also lack much of the natural product scaffold space. Metabolites and natural products bind to at least one protein in biosphere. Therefore, they can be a good source of pharmaceutically relevant molecular fragments and scaffolds. Further we found that drugs and toxics are most diverse while metabolites and natural products have a narrow distribution of scaffolds which limits their use in library design. We also found that ChEMBL database is quite diverse and contains many drug-like compounds hereby recommending its use in drug discovery programs.

For efficient mining of co-occurring fragments, *association rules* were generated using a statistical technique called *market basket analysis*. The fragment list and co-occurrence information obtained reveals fragments unique to its parent dataset which can be used to enrich combinatorial lead libraries for enhancing drug-likeness, metabolite-likeness or natural-product likeness.



## **Chapter 6: *In silico* approach to screen compounds active against parasitic nematodes of major socio-economic importance.**

### **6.1 Summary**

There are only three major classes of anthelmintic drugs available in the market. Benzimidazoles are a broad spectrum anthelmintics and inhibit the  $\beta$ -tubulin resulting in impaired microtubule formation during cell division [153]. The benzimidazoles have much more affinity for tubulin in helminth cells than the tubulin found in the cells of mammals. Macrocyclic lactones interact with a range of ion channels including glutamate-gated chloride channels [154],  $\gamma$ -aminobutyric acid-gated chloride channels [155] and acetylcholine-gated chloride channels [156]. Levamisole, pyrantel and morantel belong to the third class and bind to the nicotinic acetylcholine receptors and cause muscle paralysis due to prolonged muscle contraction and spastic paralysis of the parasite [157].

Unfortunately, resistance has been developed against two of these classes [158, 159]. In this manuscript, a systematic approach was used to screen for potentially antihelmintic compounds, using a robust machine learning method called support vector machine. Historically, anti-parasitic drugs were discovered by empirical screening against intact parasites, but due to the enormity of the task and availability of better computational facilities, there has been a shift towards computational screening. The compounds active against parasitic nematodes were collected from various literature sources, while inactive compounds were obtained from DrugBank database with no reported anthelmintic activity, as no true inactive compounds have been reported, to best of my knowledge. Also, we believe DrugBank is good choice for our inactive set because it will allow us to explore novel compounds by avoiding the already known drug space. Following on from our previous analysis results in Chapter 5, we used the ChEMBL database as a source of novel compounds, to obtain our prediction set. Currently, ChEMBL contains a diverse set of over 600,000 unique drug-like small molecules. We randomly selected a small portion of the ChEMBL dataset to obtain 10,000 molecules for our prediction set.



# ***In silico* approach to screen compounds active against parasitic nematodes of major socio-economic importance**

Varun Khanna<sup>1</sup>, Shoba Ranganathan<sup>1,2</sup>

<sup>[1]</sup> Dept. of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia.

<sup>[2]</sup> Dept. of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore

Email: [varun.khanna@mq.edu.au](mailto:varun.khanna@mq.edu.au); [shoba.ranganathan@mq.edu.au](mailto:shoba.ranganathan@mq.edu.au)

## **Abstract**

Infections due to parasitic nematodes are common causes of morbidity and fatality around the world especially in developing nations. At present however, there are only three major classes of drugs for treating human nematodes infections. Additionally the scientific knowledge on the mechanism of action and the reason for the resistance to these drugs is poorly understood. Commercial incentives to design drugs that are endemic to developing countries are limited therefore, virtual screening in academic settings can play a vital role in discovering novel drugs useful against neglected diseases. In this study we propose to build robust machine learning model to classify and screen compounds active against parasitic nematodes. Different learning algorithms were used for model development, and stratified 10-fold cross validation was used to evaluate the performance of each classifier. The best results were obtained using support vector machine (RBF kernel). Using the model developed above we were able to identify novel compounds with potential anthelmintic activity.



**Keywords:** Anthelmintics, virtual screening, machine learning, support vector machine, nematodes, scaffolds analysis

## Introduction

Besides malaria, infections due to nematodes are the leading cause of ailment to human beings. In particular, parasitic flatworms (trematodes and cestodes) and roundworms (nematodes) are a major cause of substantial suffering, particularly in children. The World Health Organization (WHO) estimates that 2.9 billion people are infected with nematodes.<sup>1</sup> Therefore, to search for nematode specific targets is an active area under research. With the availability of the completely sequenced nematode genomes, currently there is much interest to investigate drugs targeting their gene products.

At present however, only a couple of drugs are being used to control most worm infections in humans and animals. The excessive use of anthelmintics has resulted in serious problems with drug resistance in farm animals.<sup>2; 3</sup> Furthermore, with a limited number of drugs being used, it is a favourable environment for a resistant worm strains to evolve. In fact, there have been reports of resistance for the present day anthelmintic drugs in humans.<sup>4</sup> Nematodes infect most of the farm animals, as a consequence present a huge risk to livestock industry and exacerbate the global food shortages. It is therefore, not surprising that most of the anthelmintic drugs were originally developed to treat animal infections but were subsequently approved for human use with little or no modification.

Due to poor economic gains, it takes extraordinary incentives for a pharmaceutical industry to invest in tropical diseases like nematode infections. The most recent class of anthelmintic drug, the macrocyclic lactones, was



introduced in early 1980's.<sup>5</sup> Although some potent drugs are found in this class, their biological mode of action and resistance to this class by worms remains unclear more than 25 years after introduction. With the rapid progression in genomics and bioinformatics a plethora of tools are now available for target identification and validation. However, the challenge still remains, to identify novel chemical entities with desirable pharmacokinetics.

Historically, antiparasitic drugs were discovered by empirical screening against intact parasites, but due to the enormity of the task and availability of better computational facilities there has been a shift towards computational screening. Computational screening (also known as virtual screening) has inherent advantage over traditional and even experimental high throughput screening (HTS) due to its massive parallel processing ability; millions of compounds per week can be tested. Virtual screening (VS) has been widely used to discover new lead compounds by computationally identifying compounds with higher probability of strong binding affinity to the target protein. Successful studies have led to the identification of molecules either resembling the native ligands of a particular target or novel compounds.<sup>6; 7</sup> VS methods can be classified based on the amount of structural and bioactivity data available – structure-based and ligand-based. If the 3D structure of the receptor is known, one of the structure-based VS methods that can be used is high-throughput docking<sup>8</sup> but where the information of receptor is scant ligand-based methods<sup>9</sup> like similarity searching are commonly used. Moreover, ligand-based methods are popular because they are computationally inexpensive and easy to use. Furthermore, the assumption that structurally similar molecules exhibit similar biological activity than dissimilar or less similar molecules is generally valid. Thus with the little 3D information ligand-based methods can be



used at the beginning of the drug discovery projects. Nonetheless, VS still remains an unproven approach in the discovery of antiparasitic medicines.<sup>10</sup>

In this investigation, we have developed an *in silico* classification model using current machine learning approaches to predict potential anthelmintic leads targeted towards parasitic nematodes. Our model has an estimated accuracy of 82% for the test dataset. We have applied this model to a large public database to predict novel anthelmintic compounds.

## **Material and Methods**

### **Datasets for the classification study**

The quality of any machine learning model depends highly on the quality of the data available.<sup>11</sup> Our primary dataset contains 295 unique compounds (148 actives and 147 inactives). The library of active molecules (compounds active against parasitic nematodes) was carefully collated from Pubchem<sup>12</sup> and other literature sources.<sup>13-16</sup> For inactive compounds, we searched the DrugBank<sup>17</sup> database for similar molecules to the ones present in active set with a Tanimoto cut-off range of 0.25 to 0.75. As a result, compounds from various pharmacological uses (anticancer, antibacterial, sedatives, antifungal) were collected into inactive dataset. Since no true negatives (compounds without any anthelmintic activity) are reported in the literature, inactive compounds used in this study may possess residual anthelmintic activity. In Figure 1, we present representative active and inactive compounds used in this study for developing models. Further, the dataset was divided into training (80%) and testing sets (20%). The sampling was carried out at random and compounds in the test set were excluded from the model development. In Table 1, we present the composition of the dataset used in this study. For the prediction set we used the



ChEMBL<sup>18</sup> database, based on our previous study, where we reported that the ChEMBL database is quite diverse, contains many drug-like and interesting compounds. Currently, the database holds over 650,000 compounds with calculated physicochemical properties (log P, molecular weight, Lipinski properties) and abstracted bioactivities (binding constant, pharmacology and ADMET data). We downloaded the ChEMBL dataset in SD format. After cleaning the dataset of any inconsistencies and inorganic structures, we clustered the dataset to remove similar structures. Cluster centres were selected from each cluster while singletons were retained as such. For clustering, we employed functional class substructural fingerprint as implemented in Pipeline Pilot software<sup>19</sup> with the Tanimoto cut-off value 0.7. Further, compounds with 0.8 or greater Tanimoto similarity to the compounds in primary dataset were also removed. This reduced our dataset to around 300,000 compounds. Finally, we randomly selected 10,000 compounds from ChEMBL dataset for descriptor calculation and further analysis.

## **1. Training datasets**

Training datasets were used for optimizing the support vector machine<sup>20</sup> (SVM) parameters and for training the SVM classifier to predict unseen test examples. The training dataset contain 240 compounds (126 active and 114 inactive examples). All the compounds present in training set are available in Additional file 1.

## **2. Testing datasets**



Test datasets were used for evaluating the performance of the SVM method. The test dataset contains 55 compounds (22 active and 33 inactive examples).

### **Scaffold analysis**

In order to study the patterns in chemical compounds, it is important to decompose the molecules into fragments. There are a number of other ways to fragment molecules as described elsewhere.<sup>21</sup> We describe below the specific methods used in this study to obtain molecular scaffolds, where the term scaffold describes the core structure of the molecule (carbon skeleton). To obtain the carbon skeleton of the molecule, all the heavy atoms are represented as carbon and all bonds are converted to single bonds as shown in Figure 2.

### **Descriptor calculation and selection**

The determination of relevant features is an important step in any machine learning process.<sup>22</sup> Moreover, with hundreds of descriptors available it is essential to choose the best subset of descriptors because many of the descriptors are noisy and irrelevant to the target activity. Feature selection is the effective way to remove irrelevant descriptors and reduce the dimensionality of the feature space in order to avoid overfitting. This improves the prediction accuracy and leads to simple and robust computational models.

There are two main approaches for feature selection in a supervised learning context. The first one is the filter method.<sup>23</sup> Filter method is fast and easy to implement, selecting the best subset of features in an independent way, with *ad hoc* criteria. The drawback of the filter method is that there is no guarantee that the best subset of descriptors have been selected. The second method is the wrapper



approach<sup>24</sup>. This requires a predetermined learning algorithm and uses its performance as an evaluation criterion to select subset of features.

The Molecular Operating Environment<sup>25</sup> (MOE) software was used for descriptor calculation. It calculates 333 descriptors, which are classified as one-dimensional (physicochemical properties), two-dimensional (topological) and three-dimensional (volume and surface area) descriptors. Due to the large number of descriptors available, we first filtered out constant and near constant descriptors (descriptors with  $<0.3$  standard deviation). This resulted in the removal of 81 descriptors from the dataset. Following this, we removed descriptors with a correlation coefficient greater than or equal to 0.8. The removal of correlated descriptors resulted in a set of 113 descriptors. Before performing univariate analysis, we normalized the dataset using the z-transformation. We then performed the normality test to check for the distribution of the remaining descriptors in all the datasets. Those descriptors that passed the normality test were retained while the others were rejected. This reduced our set of previous 113 descriptors to 34 descriptors. For further selection of descriptors, we used the Stepwise Discriminant Analysis<sup>26</sup> (SDA) using a free data mining software Tanagra<sup>27</sup>. SDA is often associated with discriminant analysis but in fact, it could be useful for various linear models such as linear support vector machines and logistic regression. However, it is not adapted to non-linear models such as multi-layer neural networks and nearest neighbours. We implemented *forward* and *backward elimination* strategies. In the *forward* approach, at each step, all the variables are evaluated to determine which variables contribute maximum to the discrimination between the groups. Variables with significant contributions are included and the process starts again till there is no attribute to add to the model. In the *backward* approach, all the descriptors are included in the model and then,



at each step, the descriptor that contributes least to the discrimination is eliminated, terminating when there is no descriptor to remove. For our problem, we found that the *forward* approach performs better than the *backward elimination* strategy. As a termination criterion, we used *F* statistics with a predefined threshold value of 3.44, where the *F* value for a descriptor indicates its statistical significance to discriminate between the positive and negative data groups. This resulted in the selection of final 14 descriptors out of 34.

### SVM algorithm

The SVM algorithm was developed by Vapnik.<sup>28</sup> Recently, SVM has been applied to chemoinformatics, due to its robustness and ability to classify objects into two classes as a function of their features.<sup>29; 30</sup> Many studies in the past have shown SVM to be one of the best methods for correctly classifying molecules.<sup>31; 32</sup> A standard application of SVM involves defining two classes of objects, determining the set of features that distinguish these objects and use the trained SVM model to predict the classes of unknown data. Details of the SVM methodology can be obtained in literature.<sup>20; 33</sup> Briefly, SVM is based on structural risk minimization principle from statistical learning theory. Each molecule to be classified by SVM is represented by a feature vector  $x_i$  ( $i=1,2,...N$ ) of  $M$  real numbers (descriptors) with the corresponding label  $y_i \in \{+1,-1\}$ , where  $y_i = -1$  means inactive and  $y_i = +1$  means active. To classify the data, the SVM attempts to find the optimal hyperplane  $\{x \in R^m: w \cdot x + b = 0\}$  that best separates the input data into two classes in  $M$  dimensional space. The optimal hyperplane is defined in such a way that margin of separation between positive  $\{x \in R^m: w \cdot x + b \geq 0\}$  and negative  $\{x \in R^m: w \cdot x + b \leq 0\}$  examples is maximized with minimal error; where  $w$  is the normal vector of the hyperplane and  $b$  is the scalar. In other words, the optimal



hyperplane passes through the “midpoint” between these sets. The decision function for new predictions on unseen examples is given in equation 1.

$$f(x) = \text{sign} \left( \sum_{i=1}^N \alpha_i y_i K(x_i, x_j) + b \right) \quad (1)$$

where  $K(x_i, x_j)$  is the kernel function, and the parameters are determined by maximizing the following equation 2

$$\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (2)$$

under the conditions (equation 3),

$$\sum_{i=1}^N \alpha_i y_i = 0 \text{ and } 0 \leq \alpha_i \leq C \quad (3)$$

The penalty constant  $C$  serves as a regularization parameter and represents the trade-off between minimizing the training set error and maximizing the margin. The large number of support vectors is due to a small  $C$  and *vice versa*. If we use an extremely small  $C$  value, then all the samples have almost the same influence to build a decision boundary regardless of how close they are to a decision boundary. As a result, almost all the samples become support vectors. On the other hand, if we use a large  $C$  it may cause overfitting.

Since there are different types of kernels present (linear, polynomial, radial basis function, sigmoid) we explored various kernels for the efficacy of SVM prediction. We have chosen the radial basis function (RBF) kernel (equation 4) as it was found to be most effective (data not shown).



$$K(x_i, x_j) = \exp \left( \frac{-\|x_i - x_j\|^2}{2\gamma^2} \right) \quad (4)$$

Two parameters are required for optimization of SVM classifiers:  $\gamma$ , which determines the capacity of the RBF kernel and the regularization parameter, C. To optimize the SVM parameters, C and  $\gamma$ , we carried out an extensive grid search to build accurate models. The resulting optimized parameters were as follows C = 1.4 and  $\gamma = 0.43$ .

### Model validation

The prediction accuracy of the models developed was tested using ten-fold cross-validation technique. In a ten-fold cross-validation, the dataset was split into ten subsets of equal proportions. One of the subset was used as the test data while the rest were used for training the classifier. The trained classifier was tested using the test set. This was repeated ten times using a different subset for testing and thus ensuring that every compound was used in prediction once. The 295 compounds in the dataset were randomly divided into training set of 240 compounds and a test set of 55 compounds. The total number of actives and inactive compounds in the dataset are given in the Table 1.

### Performance measure

The prediction results from SVM were evaluated for test dataset using the following statistical measures.

- (i) TP, true positive – the number of correctly classified active compounds.



- (ii) TN, true negative – the number of correctly classified non-active compounds.
- (iii) FP, false positives – the number of incorrectly classified non-active compounds.
- (iv) FN, false negative – the number of incorrectly classified active compounds.

Using the variables above, a series of metrics were computed *sensitivity* (SN), *specificity* (SP), *balanced accuracy* (BA), *F-measure* and *Matthews correlation coefficient* (MCC).

The recall rate for the members of positive class (actives) is given by sensitivity, equation 5

$$sensitivity = \frac{TP}{TP + FN} * 100 \quad (5)$$

Similarly, the recall rate for the members of the negative class (inactives) is given by the specificity, equation 6

$$specificity = \frac{TN}{TN + FP} * 100 \quad (6)$$

Accuracy measures the ratio of correct predictions to the total number of classes evaluated. We calculated *balanced accuracy* which is given by the equation 7.

$$balanced\ accuracy = \frac{specificity + sensitivity}{2} \quad (7)$$

Further we calculated *F-measure* which is given by equation 8 or equation 9 if precision and recall are known

$$F - measure = \frac{2 TP}{2TP + FN + FP} \quad (8)$$

$$F = 2 * \frac{precision * recall}{precision + recall} \quad (9)$$



Finally we calculated MCC from equation 10, the coefficient returns a value between +1 and -1. The higher the value of MCC, the better the classification result.

$$\text{matthews correlation coefficient} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (10)$$

## Results

The main aim of this study was to classify and predict novel compounds active against parasitic nematodes. The various molecular descriptors (333 in total) were calculated initially, using MOE<sup>25</sup>. After removing insignificant attributes (standard deviation  $\leq 0.3$ ) and applying correlation test with a cutoff value of 0.8 we were able to reduce the total number of attributes to 113. Further, SDA algorithm was applied and finally a set of 14 descriptors was selected for the development of classification model (details in Methods section).

The obtained model correctly classified 87.56% of the active compounds and 85.30% of the inactive compounds with the overall accuracy of 86.43% in the training set while 81.82% in the test set. The *F-measure* of the training and test are 86.52% and 79.17% respectively. Table 2 depicts the result of the classification for the training set and testing set. In Figure 4, we present some of the compounds predicted active. All the predicted compounds can be found in Additional File 2.

## Discussion

The machine learning systems such as this could clearly reduce the cost involved in experimental methods involved in drug discovery pipeline. As the SVM algorithm has been successfully applied in various classification problems, we investigated the utility of SVM approach for the prediction of potential anthelmintic



lead compounds. Based on the extensive literature survey we compiled the database of active compounds with 148 unique structures. As there were no experimentally reported inactives against parasitic nematodes, we searched DrugBank database for compounds within a Tanimoto range of 0.25 (minimum) to 0.75 (maximum) of active compounds and with no reported anthelmintic activity. The idea was to build a robust model that can classify compounds into separate groups even with the similar chemistry. One inactive compound was extracted for each active compound. Since DrugBank cover most of the FDA approved drugs, we surmise that including DrugBank compounds in our inactive dataset would allow us to navigate to the unexplored regions of drug-like space. Together, a primary dataset of 295 compounds (consisting of 148 actives and 147 inactives) was constructed. Next, we divided the primary dataset into training and test set. The training set was used to optimize the parameters of the SVM kernel using a 10-fold cross-validation. We then developed a SVM-based prediction model for anthelmintic compounds. The accuracy of the training dataset may indicate the effectiveness of a prediction model however; it may not be able to accurately show how the model will perform on novel compounds. Therefore, it is critical to test the model on an independent dataset, not used in training. In our case we applied the SVM classifier, trained and optimized separately using the entire training set on the test set and evaluated the results. As shown in Table 3, for the test set the SVM model obtained an accuracy of 81.79%. To best of our knowledge there are not many reported studies on the prediction of anthelmintics compounds there we were able to compare our results with only one study. We find that our results are comparable to that study. Marrero-Ponce *et al.*<sup>34</sup> used linear discriminant analysis to classify anthelmintic and non-anthelmintic drug-like compounds. The authors reported the accuracy of around 90.4 % in the training set while 88.2% which is



slightly higher than ours. However, we believe our model is more robust because our selection criterion to pick inactive compounds was quite stringent. We selected molecules within the Tanimoto range of 0.25 to 0.75 of the compounds present in the active set which would make it relatively difficult to classify than if chosen randomly.

The results obtained are particularly interesting from a clinical perspective. From our scaffold analysis, we note that even though the size of both datasets (active and inactive) is same, yet the number of unique scaffolds found in the inactive set is almost twice the number of unique scaffolds found in the active set. This clearly indicates that the inactive set is more diverse than the active set. The number of unique scaffolds, along with the relative percentage according to the total number of molecules present in the dataset is reported in Table 3 and in Figure 3, we report top ten molecular scaffolds in both the datasets. We note that over 70.0% of the active compounds are represented by top 10 scaffolds whereas, 51.1% of the inactive compounds are represented by the same number of scaffolds, again suggesting high scaffold diversity in the inactive dataset. It should also be noted that five of the top ten scaffolds shown in Figure 3 are shared by both datasets.

In the 45 predicted compounds, we note that piperazine-like substructures appear frequently, suggesting that nitrogen in the piperazine ring might be involved in important bonding with the receptors of the drugs. Also, we note that many predicted compounds either contain benzimidazole scaffold or are derived from it for example as shown in Figure 4, six compounds out of twelve are a derivative product of the benzimidazole scaffold. This shows the validity of the above method since the benzimidazole class of compounds are well recognized for anthelmintic activity.<sup>35</sup> Further, we searched the ChEMBL database for the binding affinity, assay type and target information of the predicted compounds. We note that many



predicted compounds bind to the targets of interest in model organisms but experimental validation in the case of nematodes needs to be further carried out. Out of the total 45 predicted compounds, 6 compounds look particularly interesting. Compound 3 with antiviral activity, compound 10 with inhibitory activity against *Ancylostoma ceylanicum* (a nematode), compound 12, compound 37 with antimicrobial activity against *Staphylococcus aureus*, compound 26 with activity to inhibit SARS-CoV 3CL protease enzyme and compound 40 with activity against *Rhinovirus*. In addition, there are compounds that bind to nicotinic acetylcholine receptor and tubulin  $\beta$ -1 chain in rats or humans. Since these two receptors are successful targets in nematodes, the predicted compounds that bind to these targets can be used as leads to design novel compounds with high binding affinity to nematode nicotinic acetylcholine receptor and tubulin  $\beta$ -1 chain receptor.

### **Conclusion:**

In conclusion, we have compiled an extensive dataset of anthelmintic compounds as reported in literature for the development and validation of support vector machine model. We have rigorously tested the SVM approach for recognizing the potential compounds with anthelmintic activity. Our results show that the use of the SVM method is well suited for the prediction of anthelmintic (or antiparasitic) compounds. We were able to identify a number of interesting compounds with potential activity against parasitic nematodes however; experimental validation of the predicted compounds is needed.



## Author's contribution

VK curated the datasets and conducted the analysis work, SR directed the study and both the authors prepared the manuscript.

**Conflict of interest:** none declared

## Acknowledgements

We thank Dr. Dominique Gorse for useful discussions during this study. VK is grateful to Macquarie University for the award of MQRES research scholarship.

## References:

1. Ranganathan S, Menon R, Gasser RB: Advanced in silico analysis of expressed sequence tag (EST) data for parasitic nematodes of major socio-economic importance--fundamental insights toward biotechnological outcomes. *Biotechnol Adv* 2009; 27:439-448.
2. Sutherland IA, Leathwick DM: Anthelmintic resistance in nematode parasites of cattle: a global issue? *Trends in Parasitology* 2010; In Press, Corrected Proof.
3. James CE, Hudson AL, Davey MW: Drug resistance mechanisms in helminths: is it survival of the fittest? *Trends in Parasitology* 2009; 25:328-335.
4. Geerts S, Gryseels B: Drug resistance in human helminths: current situation and lessons from livestock. *Clin Microbiol Rev* 2000; 13:207-222.
5. Grant WN, Behm CA: Target identification and validation for anthelmintic discovery. *Expert Opinion on Drug Discovery* 2007; 2:S91-S98.
6. Reddy S, Pati P, Kumar P, Pradeep HN, Sastry N: Virtual screening in drug discovery -- a computational perspective. *Current protein & peptide science* 2007; 8:329-351.
7. Freitas RF, Oprea TI, Montanari CA: 2D QSAR and similarity studies on cruzain inhibitors aimed at improving selectivity over cathepsin L. *Bioorganic & Medicinal Chemistry* 2008; 16:838-853.
8. Sousa Sr, Fernandes P, Ramos M: Protein-ligand docking: Current status and future challenges. *Proteins* 2006; 65:15-26.
9. Geppert H, Vogt M, Bajorath J: Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *Journal of Chemical Information and Modeling* 2010; 50:205-216.
10. Woods D, Williams T: The challenges of developing novel antiparasitic drugs. *Invertebrate Neuroscience* 2007; 7:245-250-250.
11. Tropsha A: Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* 2010; 29:476-488.
12. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH: PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research* 2009; 37:W623-W633.



13. Holden-Dye L, Walker RJ: Anthelmintic drugs. *WormBook* 2007:1-13.
14. Mayer AM, Hamann MT: Marine pharmacology in 2001--2002: marine compounds with anthelmintic, antibacterial, anticoagulant, antidiabetic, antifungal, anti-inflammatory, antimalarial, antiplatelet, antiprotozoal, antituberculosis, and antiviral activities; affecting the cardiovascular, immune and nervous systems and other miscellaneous mechanisms of action. *Comp Biochem Physiol C Toxicol Pharmacol* 2005; 140:265-286.
15. Mayer AM, Rodriguez AD, Berlinck RG, Hamann MT: Marine pharmacology in 2003-4: marine compounds with anthelmintic antibacterial, anticoagulant, antifungal, anti-inflammatory, antimalarial, antiplatelet, antiprotozoal, antituberculosis, and antiviral activities; affecting the cardiovascular, immune and nervous systems, and other miscellaneous mechanisms of action. *Comp Biochem Physiol C Toxicol Pharmacol* 2007; 145:553-581.
16. Mayer AM, Rodriguez AD, Berlinck RG, Hamann MT: Marine pharmacology in 2005-6: Marine compounds with anthelmintic, antibacterial, anticoagulant, antifungal, anti-inflammatory, antimalarial, antiprotozoal, antituberculosis, and antiviral activities; affecting the cardiovascular, immune and nervous systems, and other miscellaneous mechanisms of action. *Biochim Biophys Acta* 2009; 1790:283-308.
17. Wishart D, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* 2008; 36:D901-906.
18. Overington J: ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J Comput Aided Mol Des* 2009; 23:195-198.
19. Pipeline Pilot Retrieved from <http://accelrys.com/>.
20. Trotter MWB, Holden SB: Support vector machines for ADME property classification. *Qsar & Combinatorial Science* 2003; 22:533-548.
21. Bemis GW, Murcko MA: The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* 1996; 39:2887-2893.
22. Dutta D, Guha R, Wild D, Chen T: Ensemble Feature Selection: Consistent Descriptor Subsets for Multiple QSAR Models. *Journal of Chemical Information and Modeling* 2007; 47:989-997.
23. Duch W: Filter Methods. In Guyon I, Gunn S, Nikravesh M, Zadeh L (eds): *Feature Extraction: Foundations and Applications*. Berlin, Germany: Springer, 2006.
24. Marchiori E, Moore J, Soto A, Cecchini R, Vazquez G, Ponzoni I: A Wrapper-Based Feature Selection Method for ADMET Prediction Using Evolutionary Computing. *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*: Springer Berlin / Heidelberg, 2008:188-199-199.
25. MOE(2009.10) Retrieved from <http://www.chemcomp.com/>.
26. Jennrich RI: Stepwise discriminant analysis. In Enslein K, Ralston A, Wilf HS (eds): *Statistical methods for digital computers* New York: Wiley, 1977:76-96.
27. Tanagra: free data mining software Retrieved from <http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>.
28. Cortes C, Vapnik V: Support-Vector Networks. *Machine Learning* 1995; 20:273-297.
29. Jorissen RN, Gilson MK: Virtual Screening of Molecular Databases Using a Support Vector Machine. *Journal of Chemical Information and Modeling* 2005; 45:549-561.



30. Liew CY, Ma XH, Liu X, Yap CW: SVM Model for Virtual Screening of Lck Inhibitors. *Journal of Chemical Information and Modeling* 2009; 49:877-885.
31. Warmuth MK, Liao J, Rätsch G, Mathieson M, Putta S, Lemmen C: Active Learning with Support Vector Machines in the Drug Discovery Process. *Journal of Chemical Information and Computer Sciences* 2003; 43:667-673.
32. Byvatov E, Fechner U, Sadowski J, Schneider G: Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J Chem Inf Comput Sci* 2003; 43:1882-1889.
33. Ivanciuc O: Applications of support vector machines in chemistry. *Reviews in Computational Chemistry* 2007; 23:291-400.
34. Marrero-Ponce Y, Castillo-Garit JA, Olazabal E, Serrano HS, Morales A, Castanedo N, Ibarra-Velarde F, Huesca-Guillen A, Jorge E, del Valle A, Torrens F, Castro EA: TOMOCOMD-CARDD, a novel approach for computer-aided 'rational' drug design: I. Theoretical and experimental assessment of a promising method for computational screening and in silico design of new anthelmintic compounds. *J Comput Aided Mol Des* 2004; 18:615-634.
35. Lacey E: Mode of action of benzimidazoles. *Parasitol Today* 1990; 6:112-115.



**Table 1: Composition of the datasets used in this study.**

Dataset	Training set	Testing set	<b>Total</b>
Active	126	22	<b>148</b>
Inactive	114	33	<b>147</b>
<b>Total</b>	<b>240</b>	<b>55</b>	<b>295</b>
Prediction set (from ChEMBL)	–	–	10,000



**Table 2: Performance measure of SVM classifier in training and test dataset.**

SN: sensitivity, SP: specificity, BA: balanced accuracy, MCC: Matthews correlation coefficient

<b>Dataset</b>	<b>SN (%)</b>	<b>SP (%)</b>	<b>BA (%)</b>	<b>F-measure (%)</b>	<b>MCC</b>
Training set	87.56	85.38	86.43	86.52	0.75
Test set	83.82	79.76	81.79	79.17	0.63



**Table 3: The number of unique scaffolds found in active and inactive sets along with the percentage relative to the dataset size.**

<b>Datasets</b>	<b>Size of the dataset</b>	<b>Non-redundant scaffolds</b>	<b>Percentage (relative to dataset size)</b>
Actives	148	48	32.43%
Inactives	147	80	54.42%

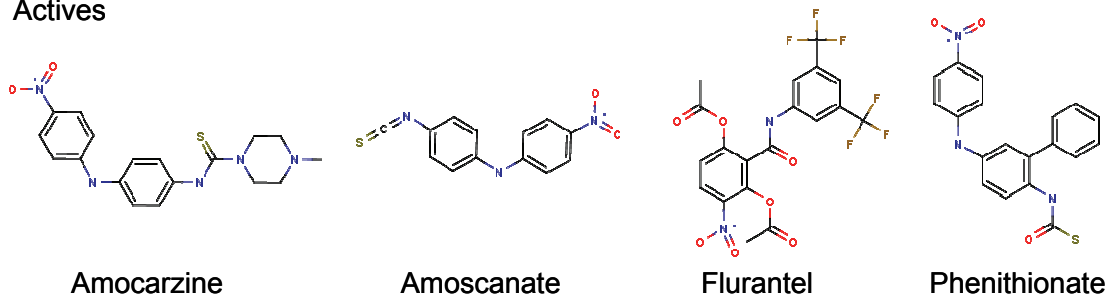


## **Additional file**

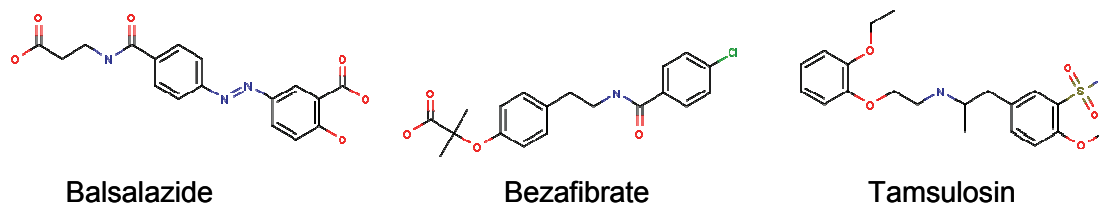
1. **Table S1: Dataset used for training, testing and validation of the model. (\*.pdf)**
2. **Table S2: Predicted compounds with AlogP, molecular weight and SMILES information. (\*.pdf)**



### Actives



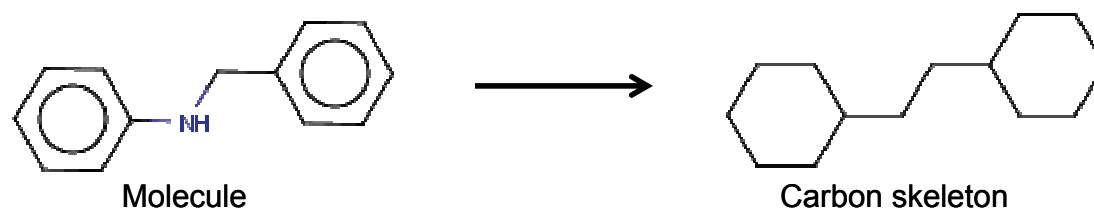
### Inactives



**Figure 1: Examples of active and inactive compounds used in this analysis.**

The active compounds are collected from various literature sources and PubChem database while inactive compounds are adapted from DrugBank.

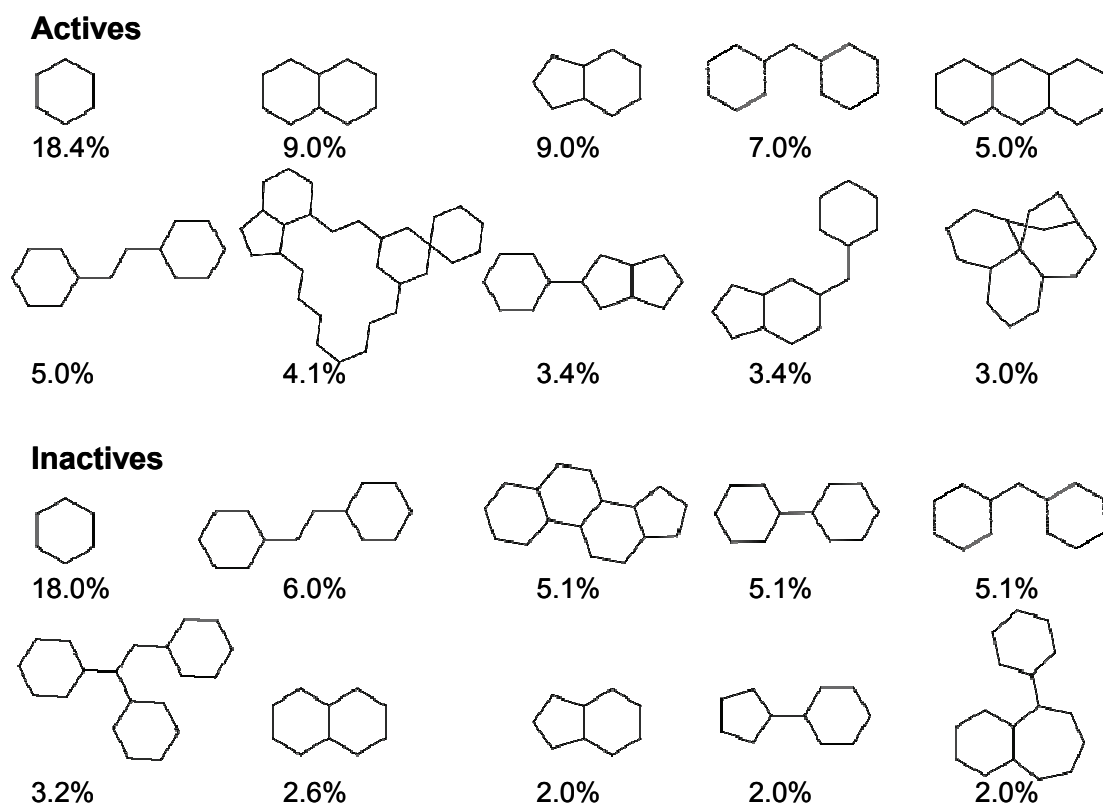




**Figure 2: Definition of the scaffold used in this study.**

The scaffold is obtained by iteratively removing side chains and converting all the bonds to single bonds

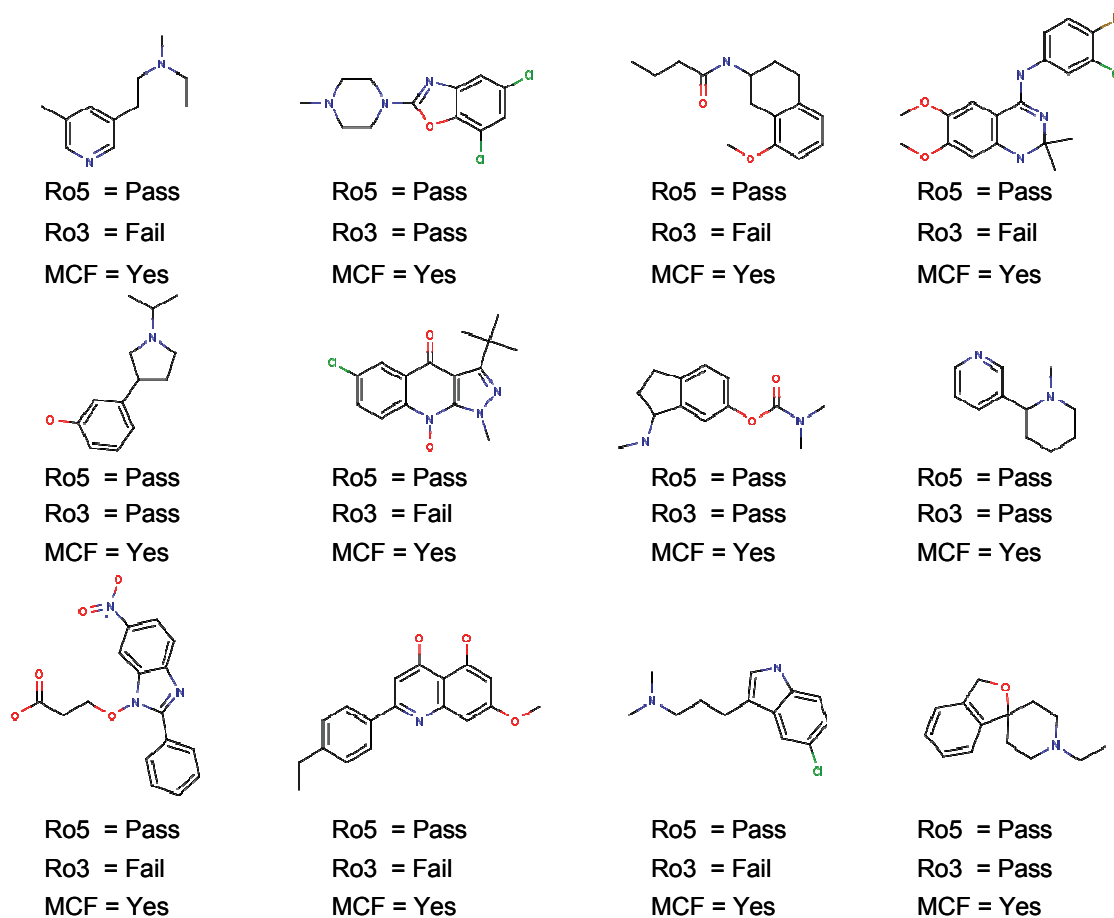




**Figure 3: Top ten scaffolds present in active and inactive dataset.**

Inactive dataset is more diverse than active dataset. Five out of top ten scaffolds are shared in both the datasets.





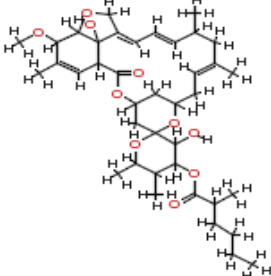
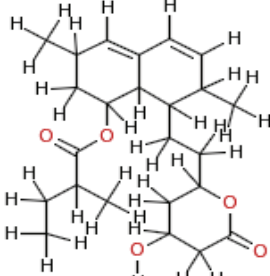
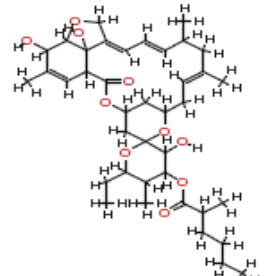
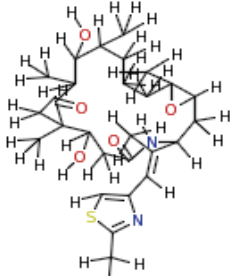
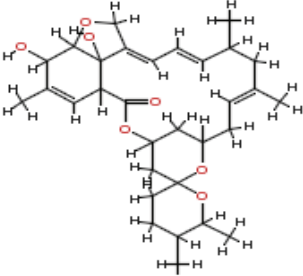
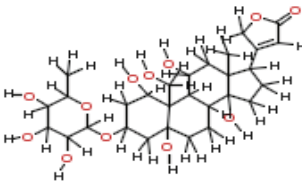
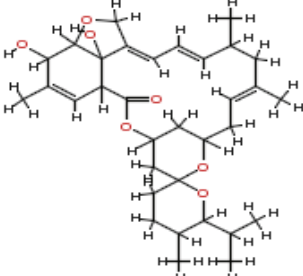
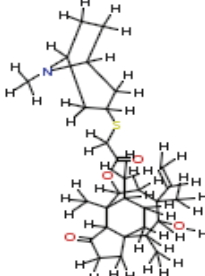
**Figure 4: Examples of the actives predicted in the prediction set derived from ChEMBL database.** All the molecules shown in the figure pass “rule of five” (ro5) test and are medicinal chemist friendly (MCF). Further a few of them also pass lead-likeness “rule of three” (Ro3) test.



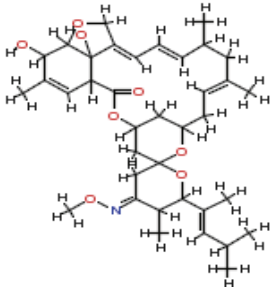
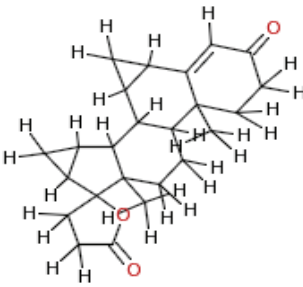
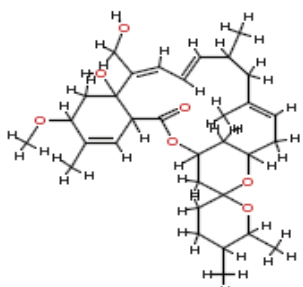
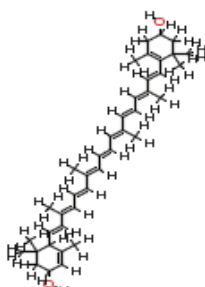
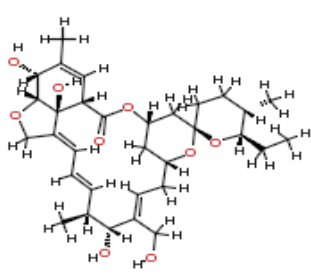
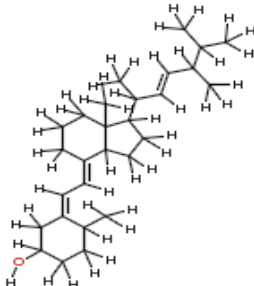
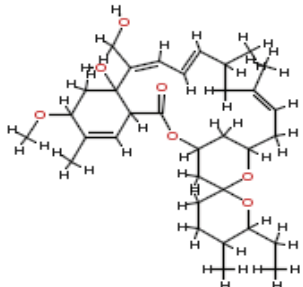
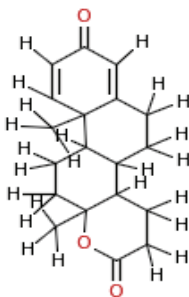
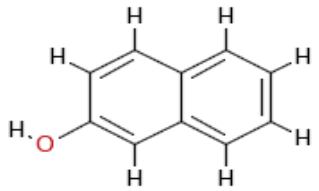
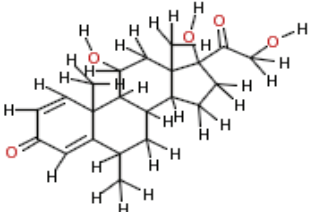
## Additional File 1

*In silico* approach to screen compounds active against parasitic nematodes of major socio-economic importance by Varun Khanna and Shoba Ranganathan

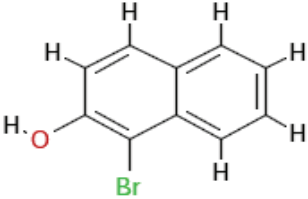
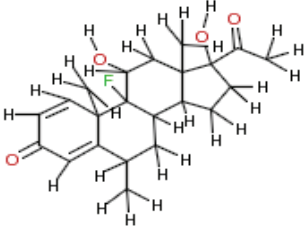
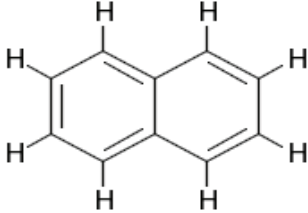
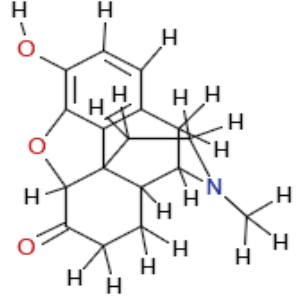
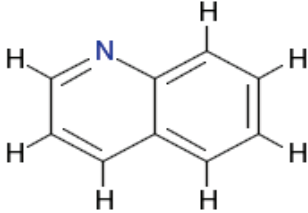
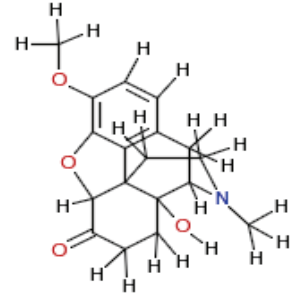
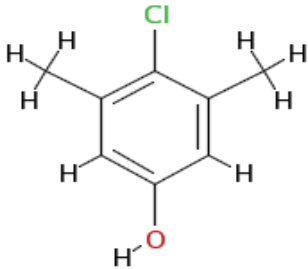
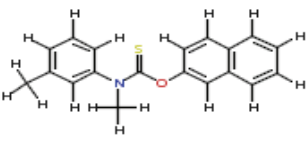
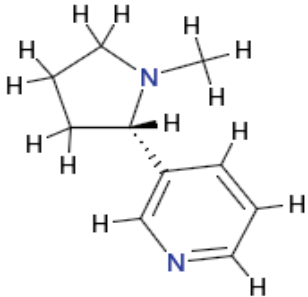
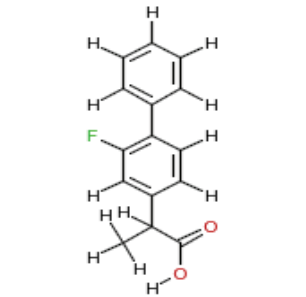
**Table S1: Dataset used for training, testing and validation of the model.**

ACTIVE COMPOUNDS		INACTIVE COMPOUNDS	
STRUCTURES	PUBCHEM ID	STRUCTURES	DRUGBANK ID
	6442491		DB00227
	6443028		DB04845
	6436638		DB01092
	6440579		DB01256

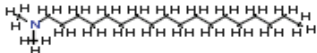
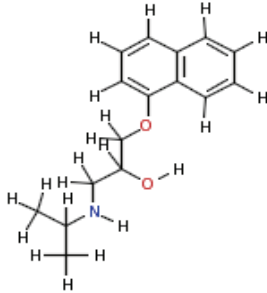
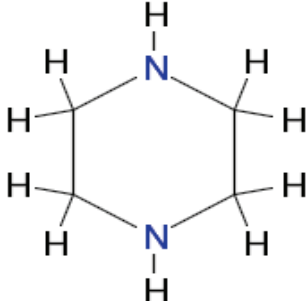
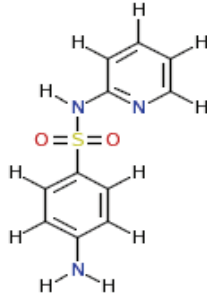
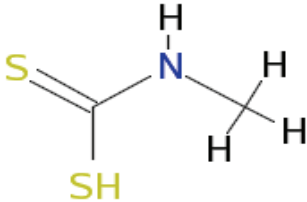
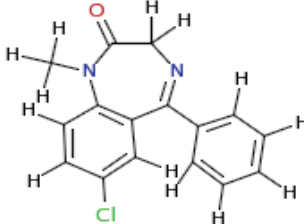
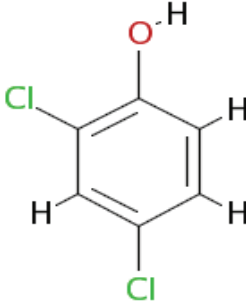
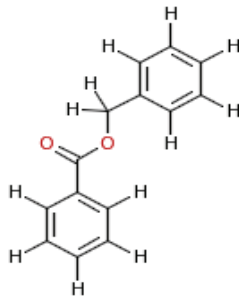
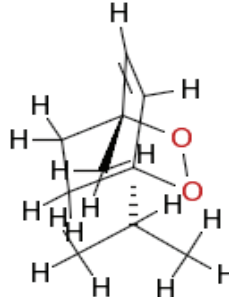
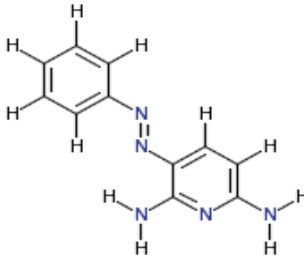


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	9571036		DB01395
	6450300		DB00137
	6441241		DB01070
	6443029		DB00894
	8663		DB00959



ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	11316		DB00324
	931		DB00327
	7047		DB00497
	2723		DB00525
	89594		DB00712

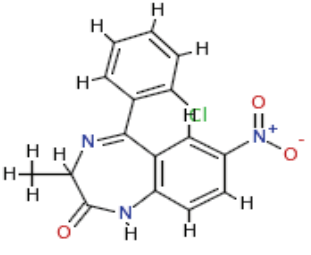
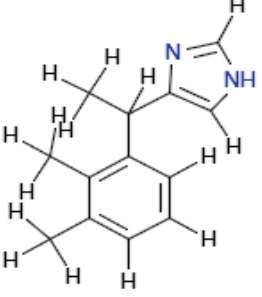
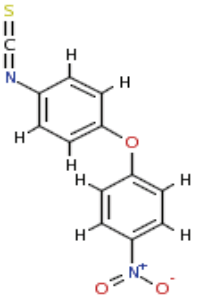
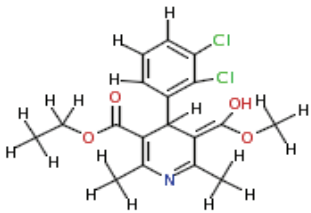
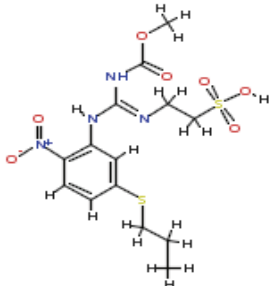
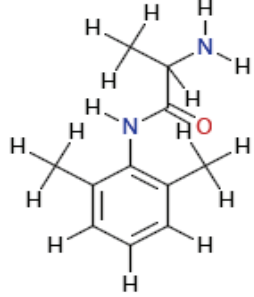
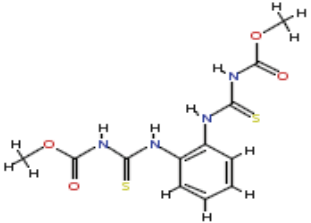
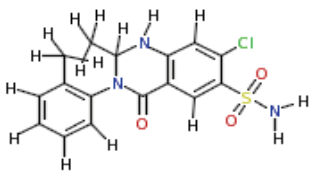
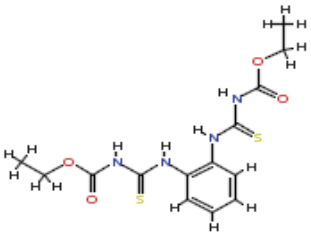
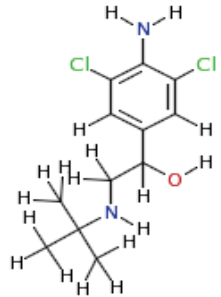


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	15365		DB00571
	3014216		DB00891
	3001858		DB00829
	8449		DB00676
	10545		DB01438

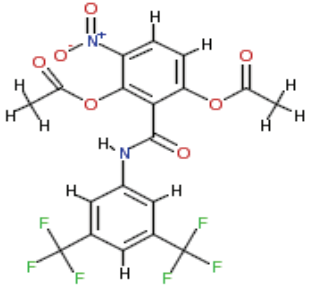
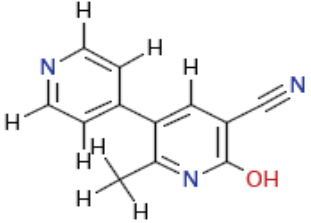
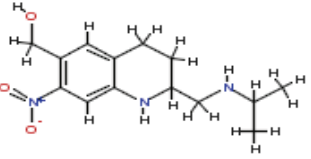
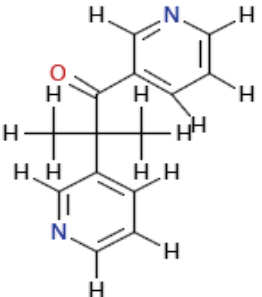
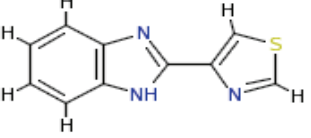
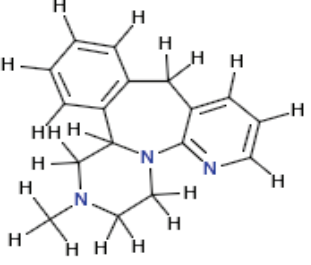
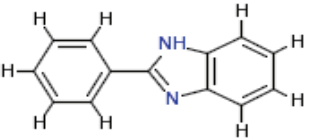
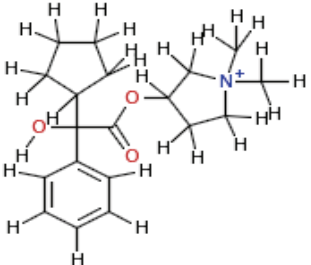
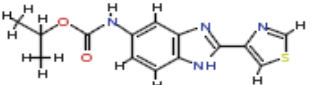
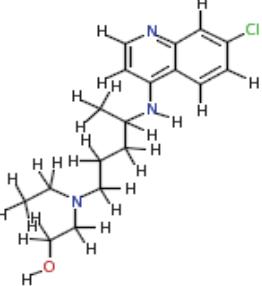




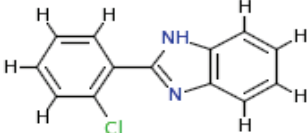
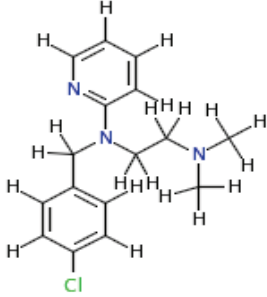
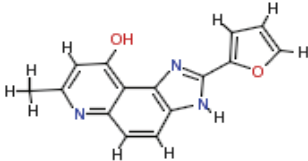
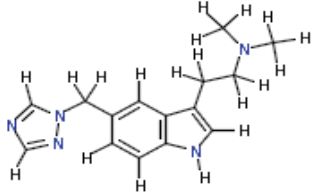
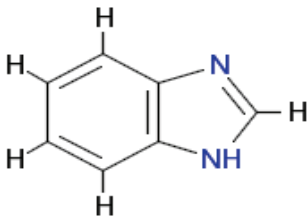
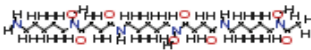
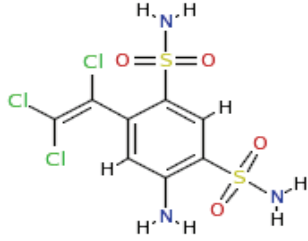
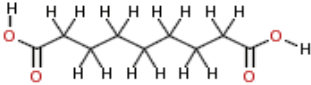
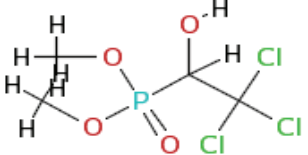
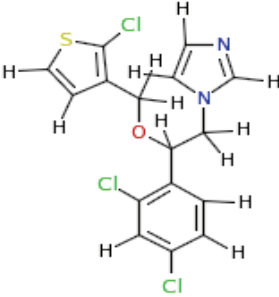


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	3033985		DB00633
	68547		DB01023
	71449		DB01056
	3032791		DB00524
	3032792		DB01407

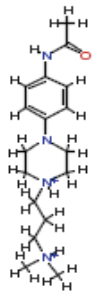
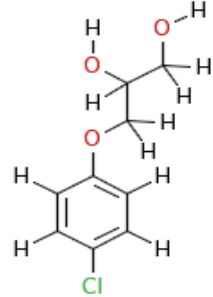
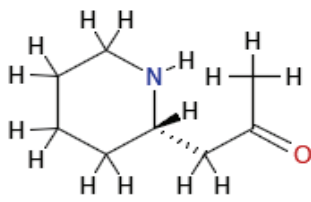
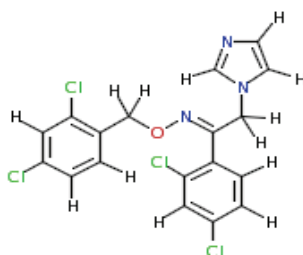
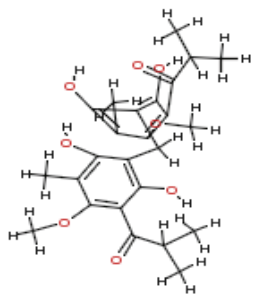
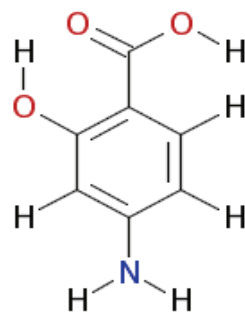
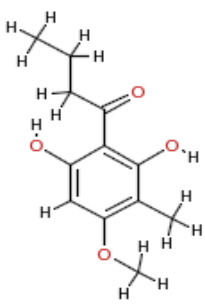
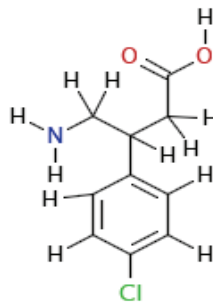
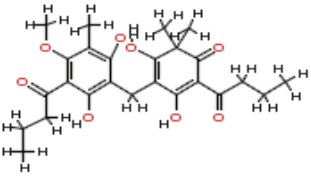
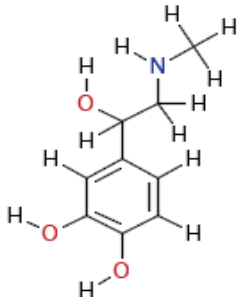


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	219070		DB00235
	4612		DB01011
	5430		DB00370
	12855		DB00986
	33309		DB01611

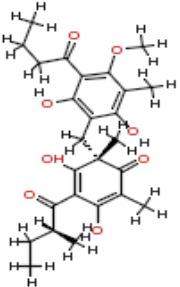
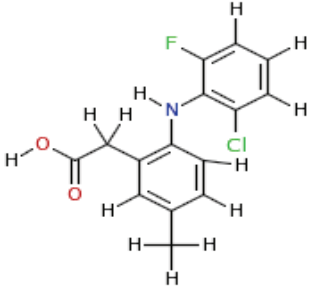
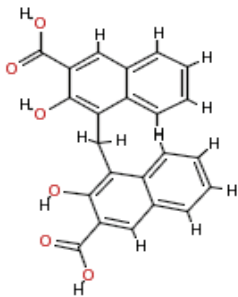
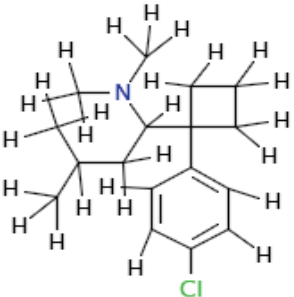
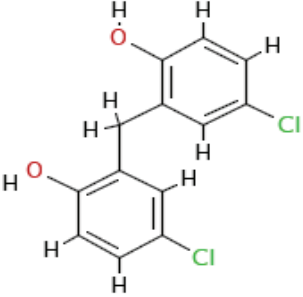
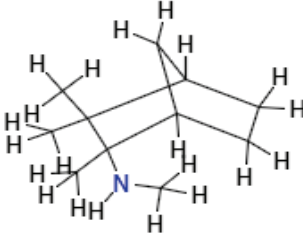
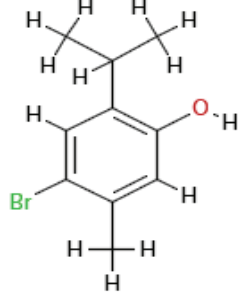
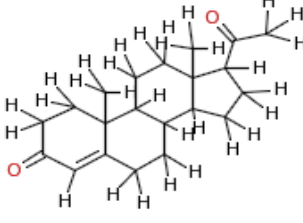
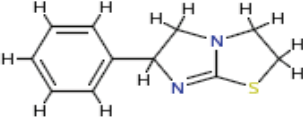
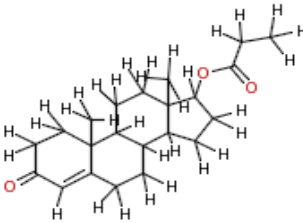


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	77123		DB08800
	312914		DB00953
	5798		DB00746
	43231		DB00548
	5853		DB01007

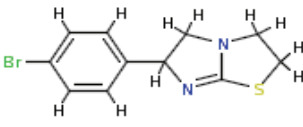
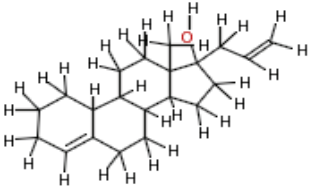
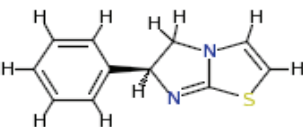
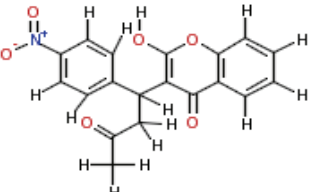
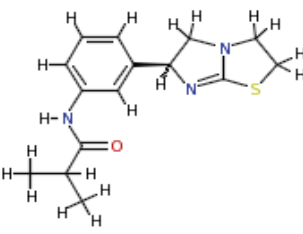
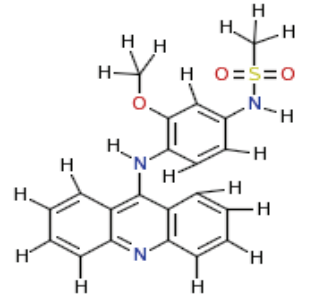
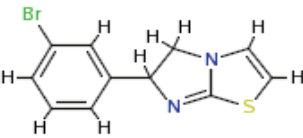
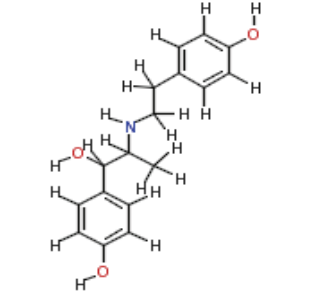
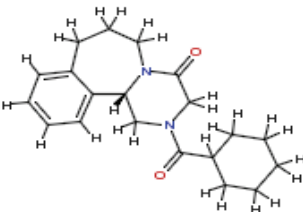
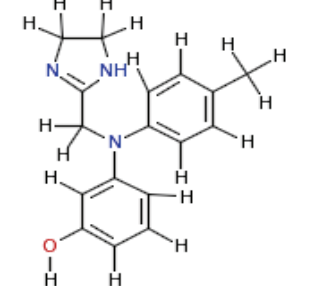


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	6433263		DB00856
	92987		DB00239
	160529		DB00233
	122841		DB00181
	120290		DB00668

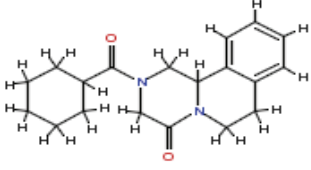
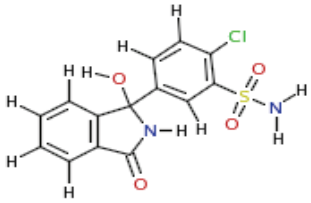
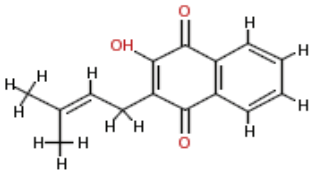
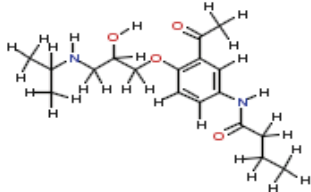
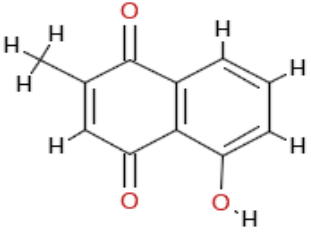
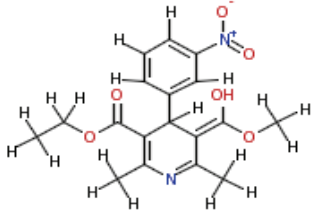
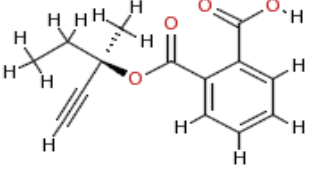
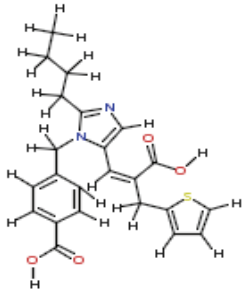
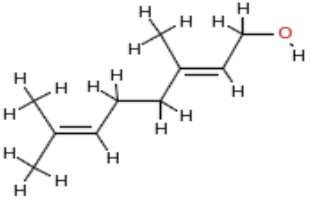
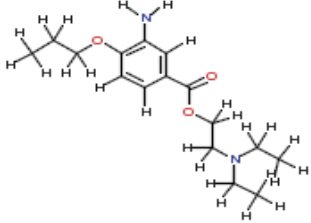


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	128292		DB01283
	5281033		DB01105
	3037		DB00657
	203726		DB00396
	27944		DB01420

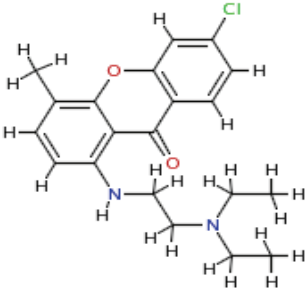
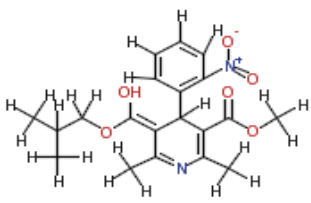
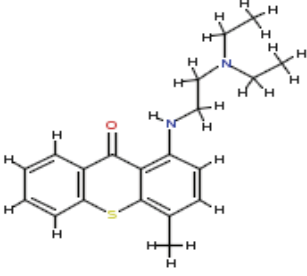
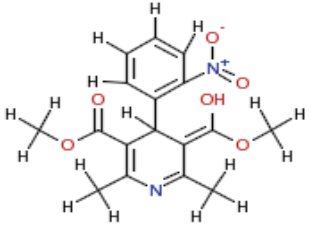
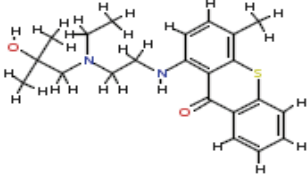
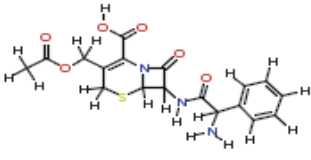
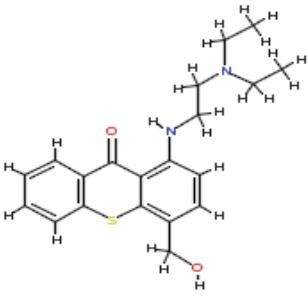
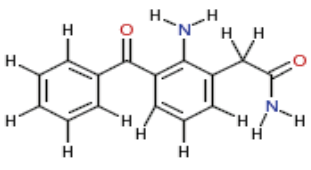
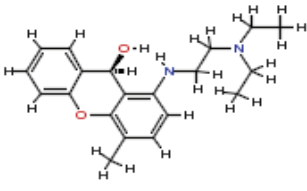
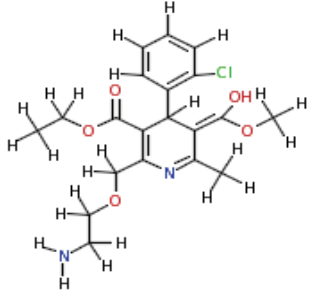


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	5121		DB01431
	170340		DB01418
	166572		DB00276
	3084926		DB00867
	72026		DB00692

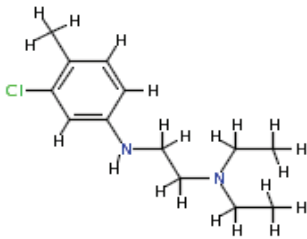
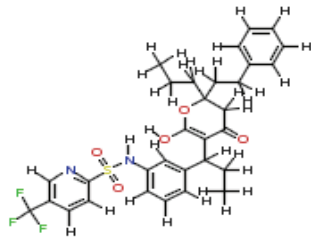
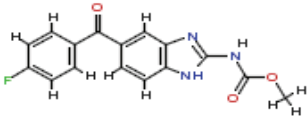
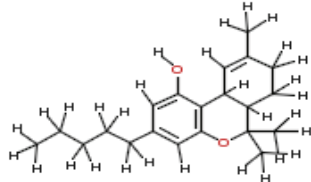
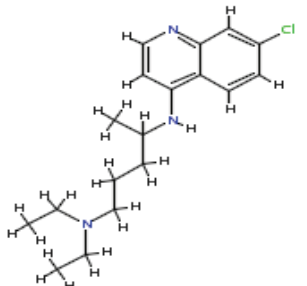
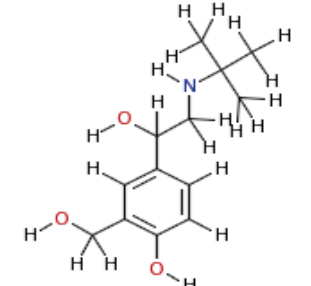
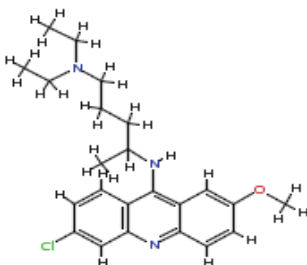
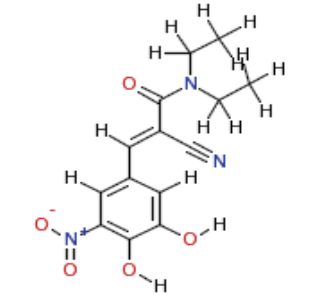
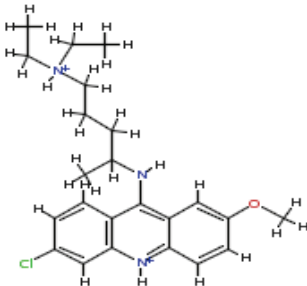
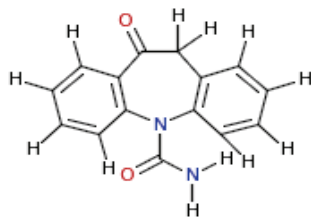


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	4891		DB00310
	24191630		DB01193
	10205		DB01054
	8573		DB00876
	637566		DB00807

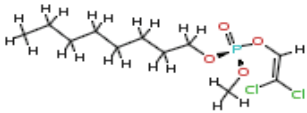
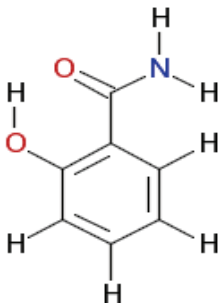
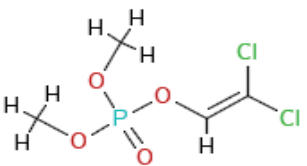
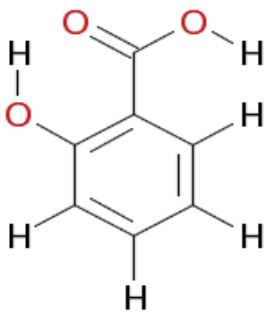
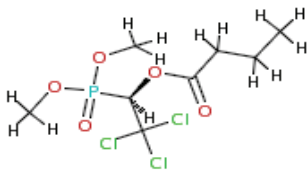
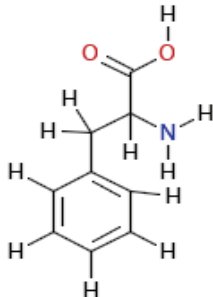
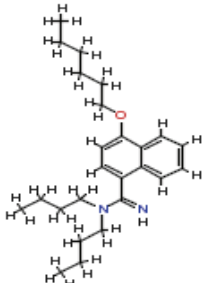
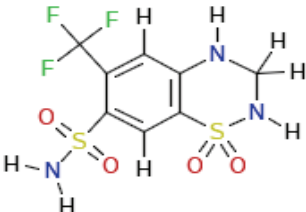
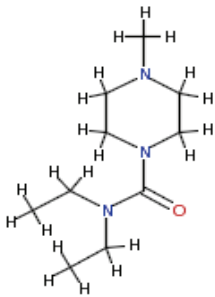
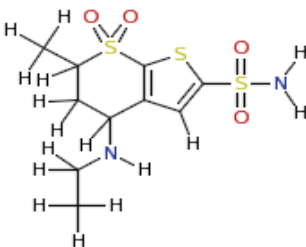


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	192786		DB00401
	5351142		DB01115
	21793		DB00689
	3634		DB06802
	192783		DB00381

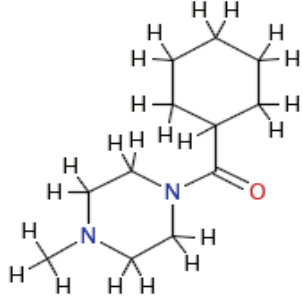
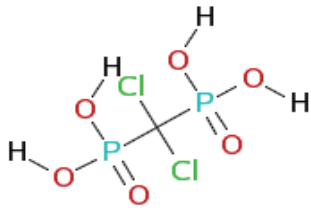
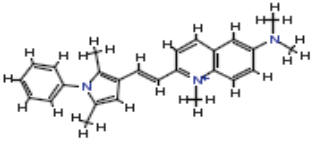
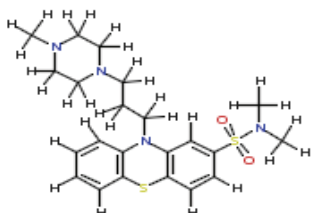
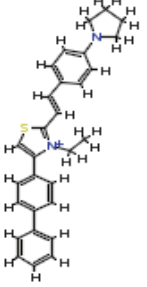
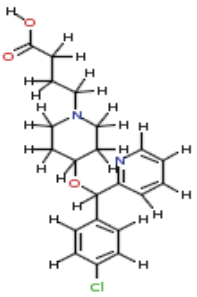
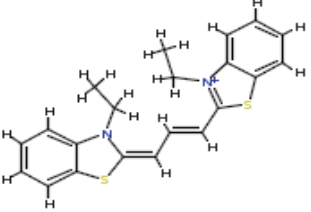
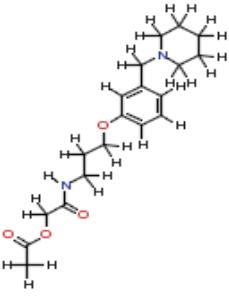
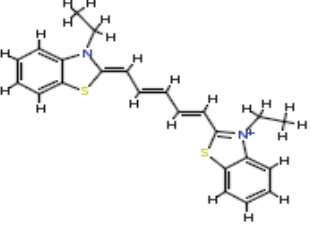
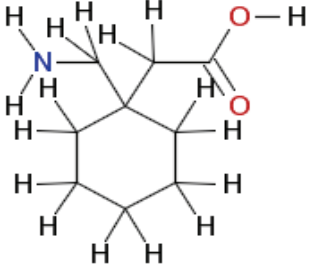


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	75648		DB00932
	35802		DB00470
	64927		DB01001
	23581813		DB00494
	15329120		DB00776



ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	66388		DB08797
	3039		DB00936
	31343		DB00120
	13986		DB00774
	15432		DB00869

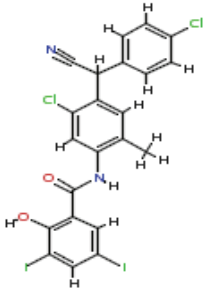
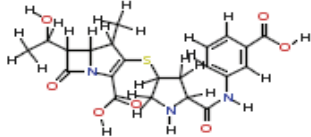
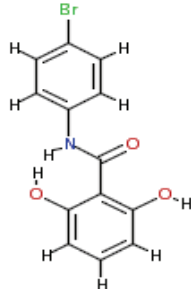
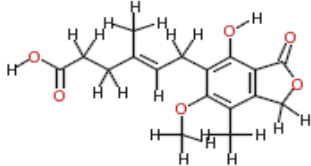
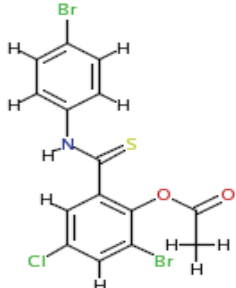
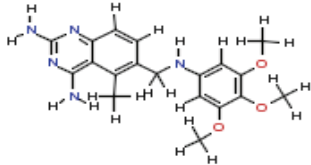
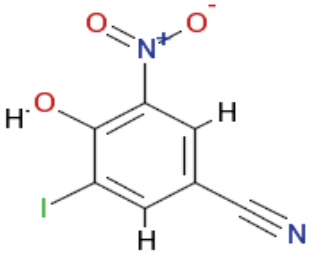
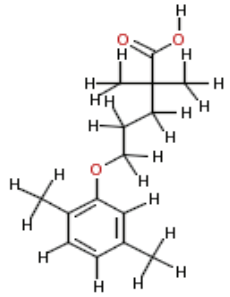
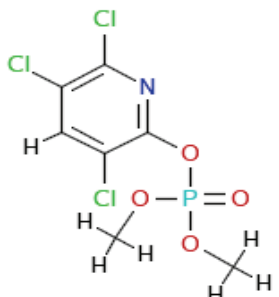
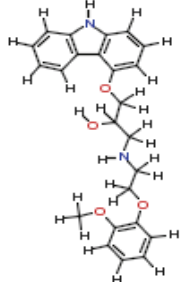


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	667497		DB00720
	11979707		DB01622
	6446785		DB04890
	6433449		DB08806
	24196442		DB00996

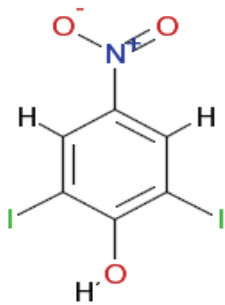
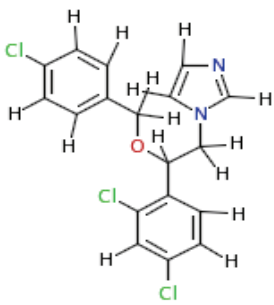
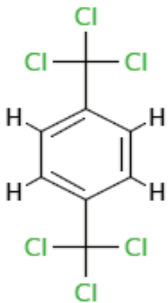
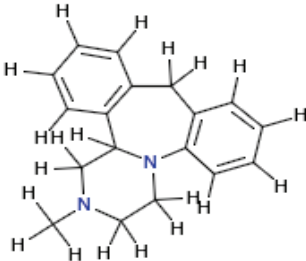
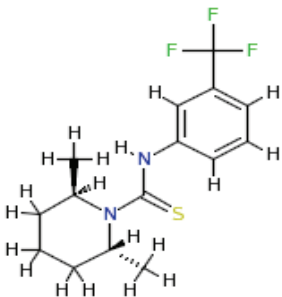
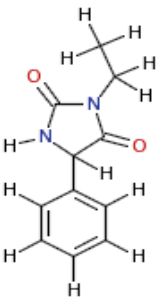
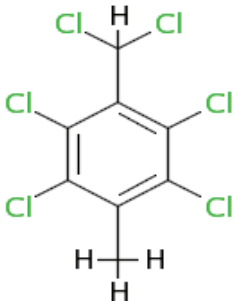
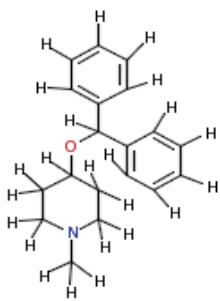
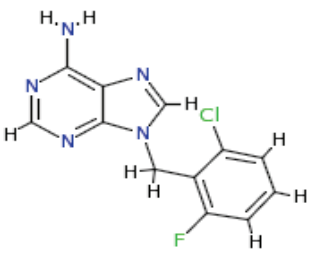
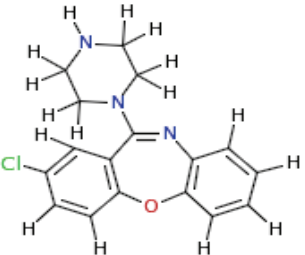




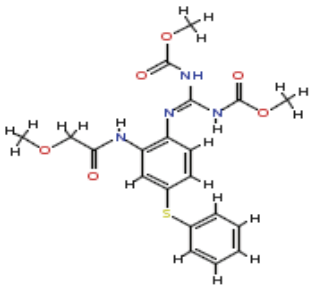
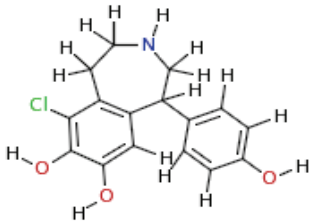
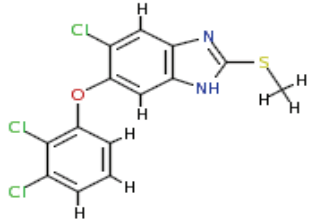
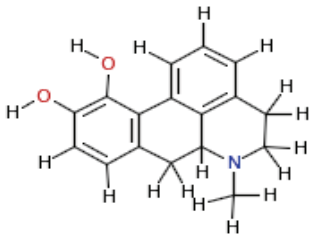
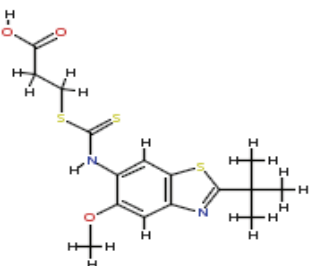
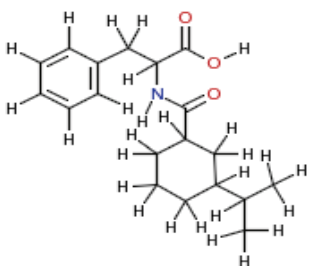
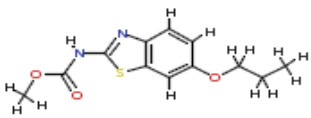
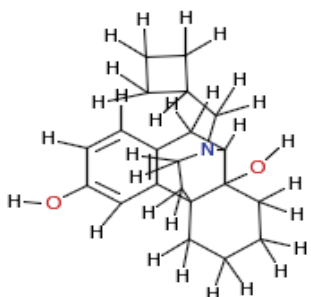
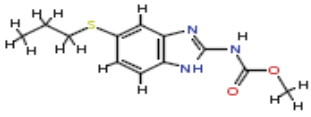
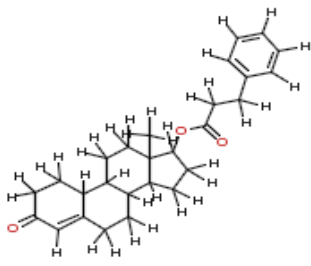


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	42574		DB00303
	65696		DB01024
	3034337		DB01157
	15532		DB01241
	21805		DB01136

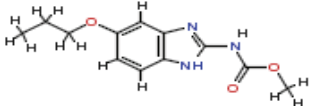
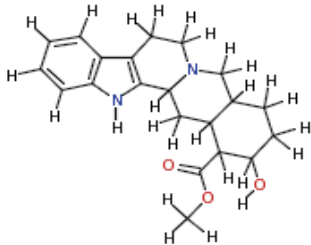
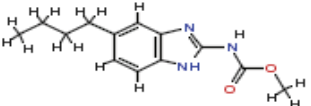
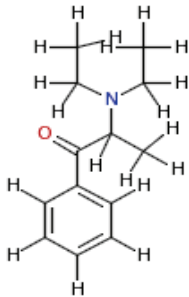
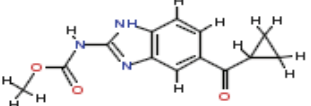
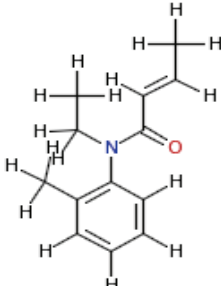
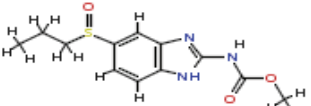
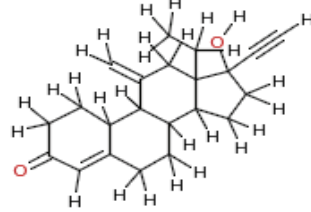
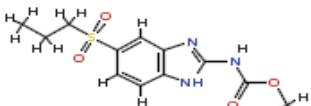
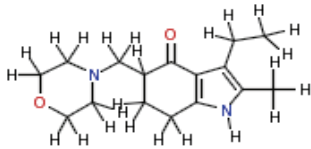


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	9370		DB01127
	6233		DB06148
	3034015		DB00754
	214321		DB01146
	41574		DB00543

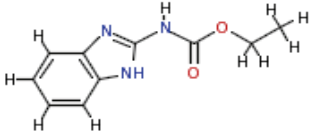
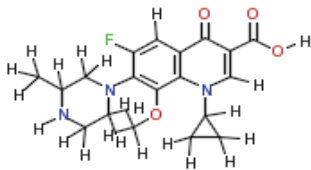
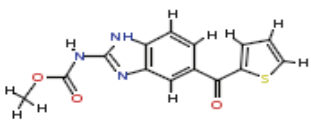
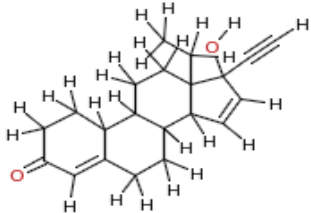
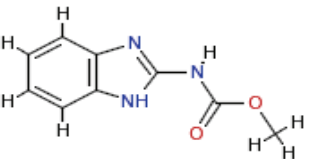
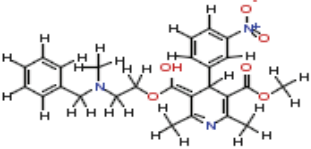
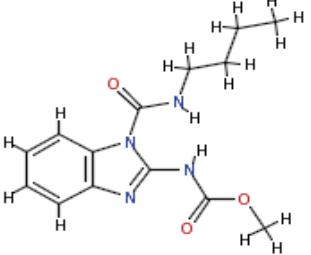
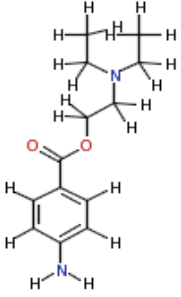
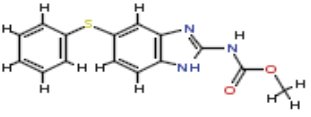
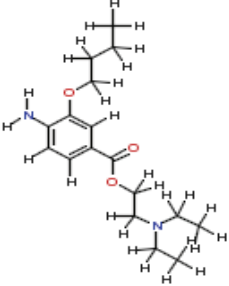


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	9570638		DB00800
	50248		DB00714
	3035447		DB00731
	72157		DB00611
	2082		DB00984

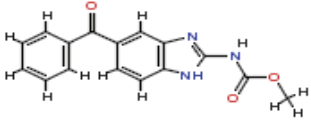
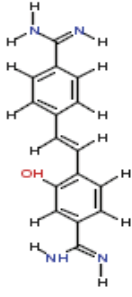
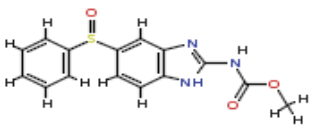
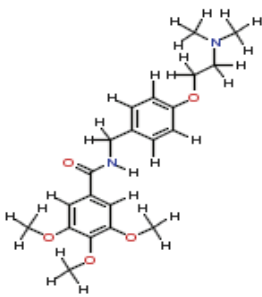
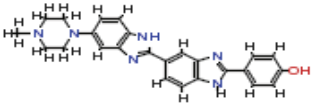
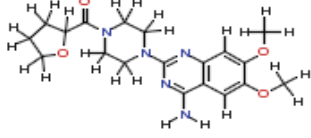
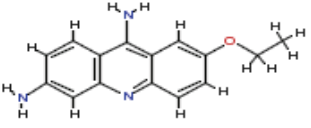
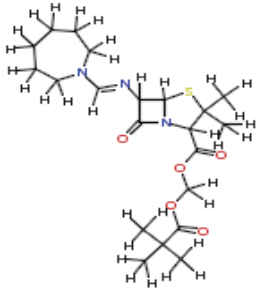
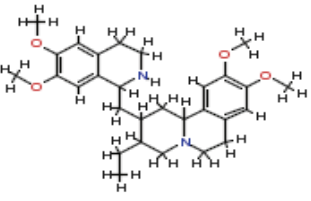
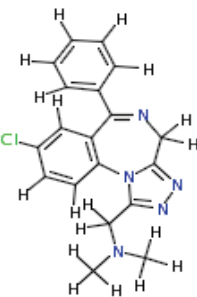


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	4622		DB01392
	26596		DB00937
	35803		DB00265
	83969		DB00294
	53174		DB01618

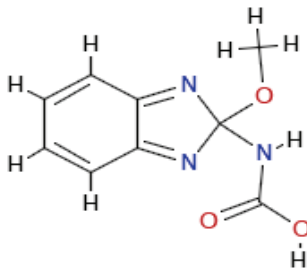
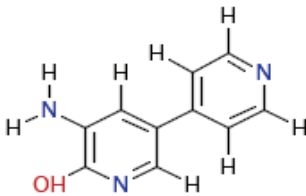
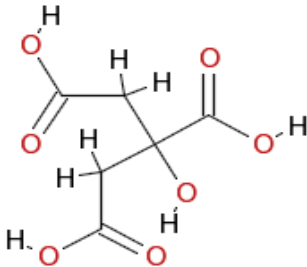
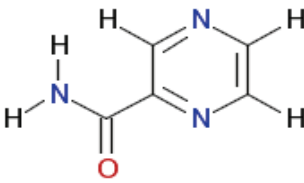
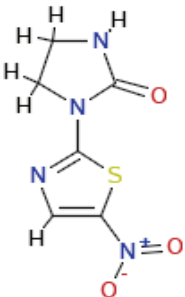
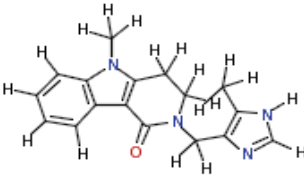
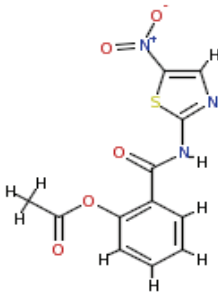
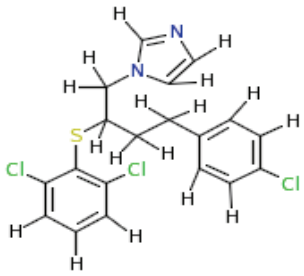
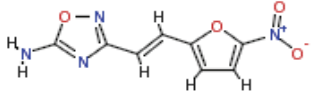
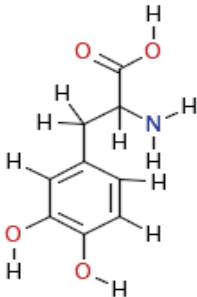


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	22752		DB01044
	4122		DB06730
	25429		DB00622
	28780		DB00721
	3334		DB00892

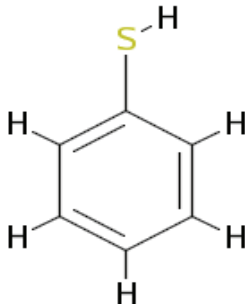
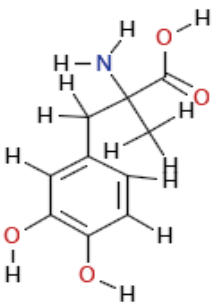
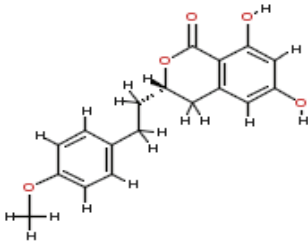
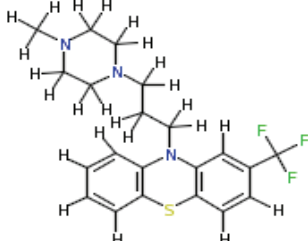
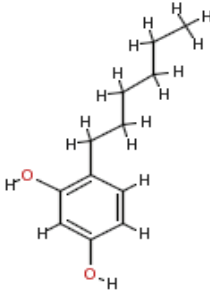
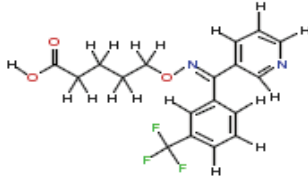
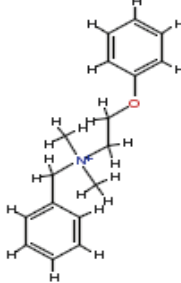
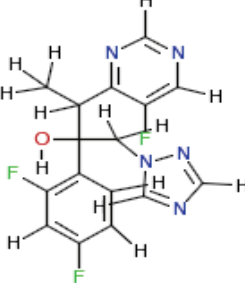
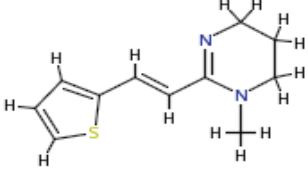
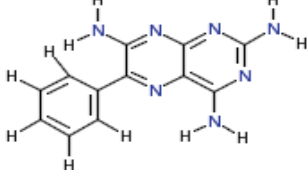


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	4030		DB01040
	40854		DB00662
	16218619		DB01162
	15789		DB01605
	3068143		DB00546

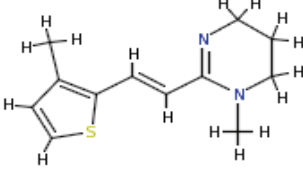
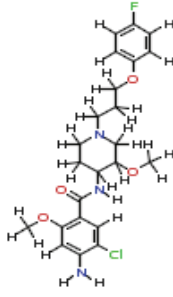
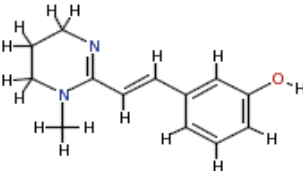
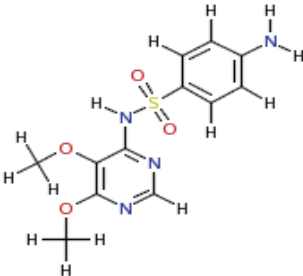
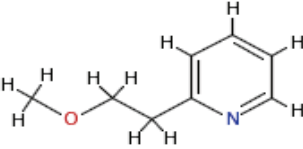

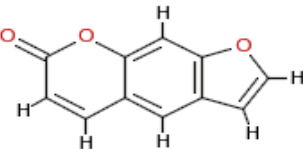
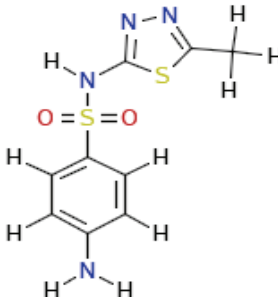
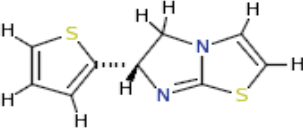
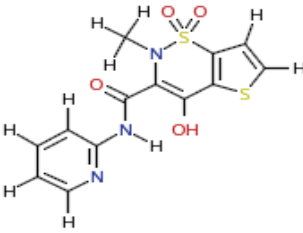


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	24901740		DB01427
	114763		DB00339
	6093		DB00969
	41684		DB00639
	5284340		DB01235

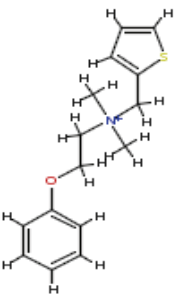
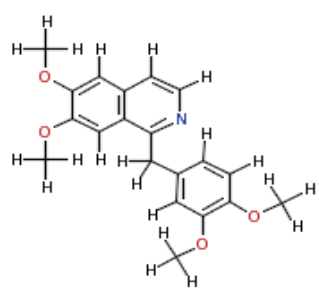
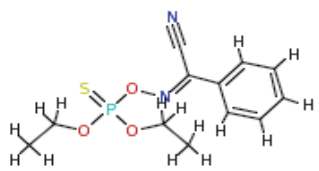
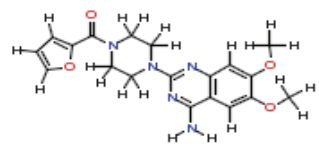
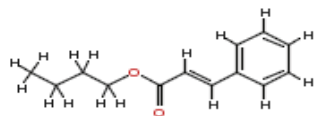
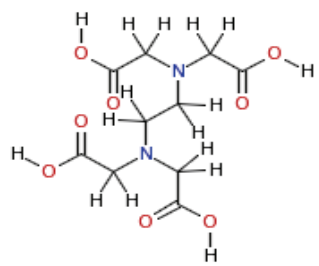
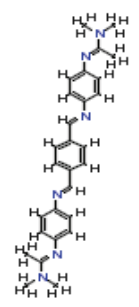
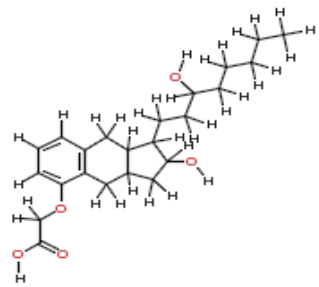
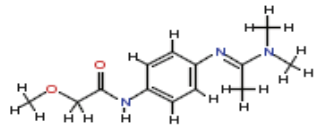
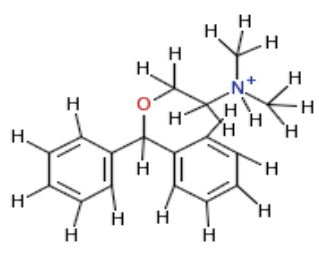


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	7969		DB00968
	161362		DB00831
	3610		DB01207
	19666		DB00582
	708857		DB00384

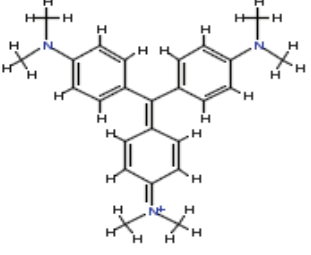
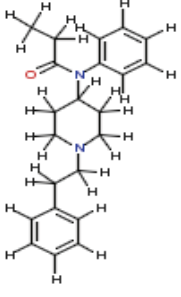
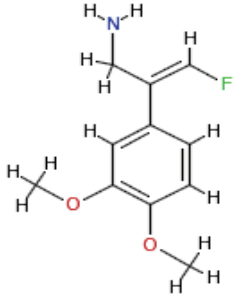
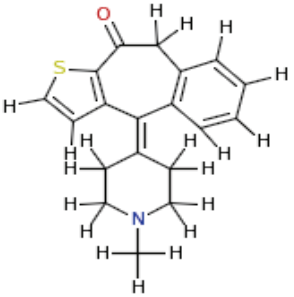
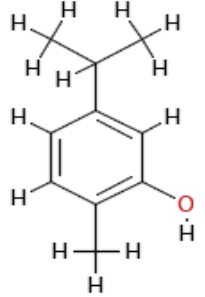
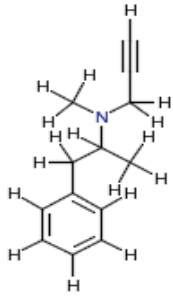
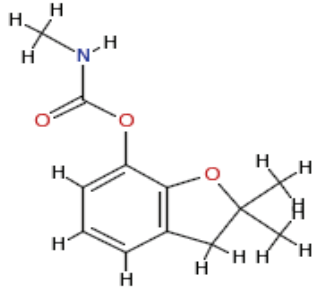
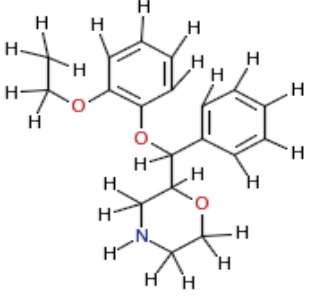
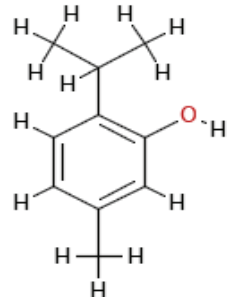
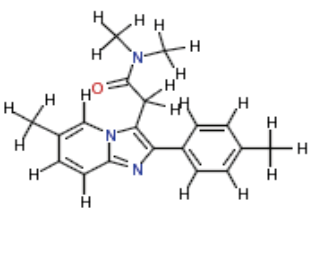


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	6433951		DB00604
	5281087		DB01299
	66991		DB01581
	6199		DB00576
	170351		DB00469

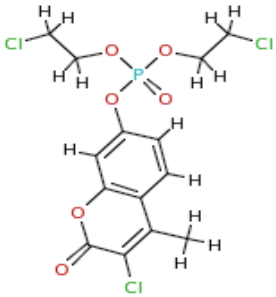
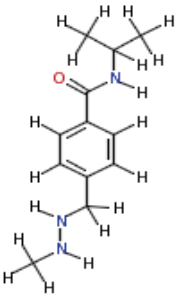
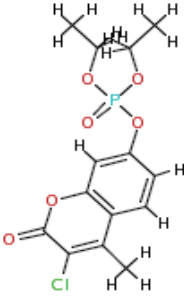
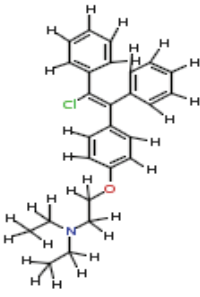
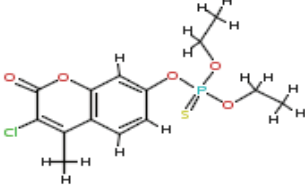
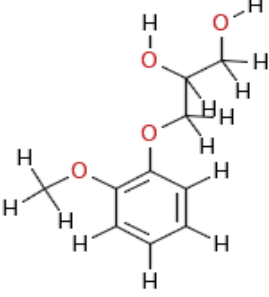
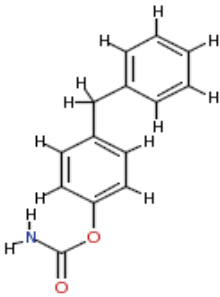
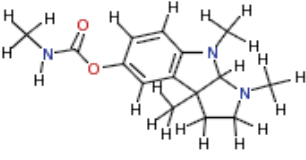
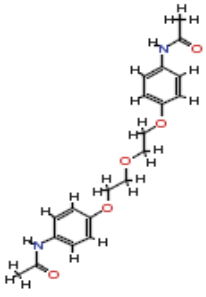
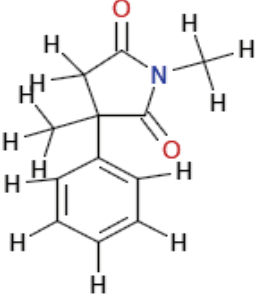


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	498092		DB01113
	9570290		DB00457
	5273465		DB00974
	3086564		DB00374
	39521		DB00985

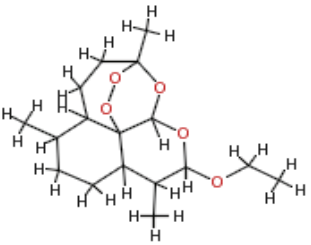
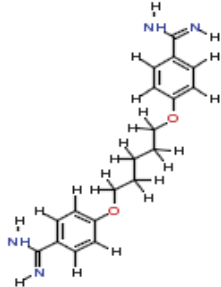
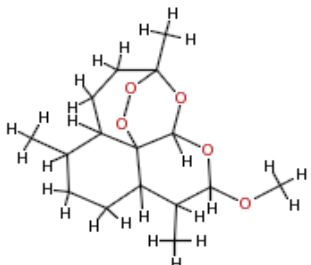
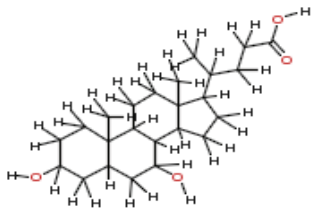
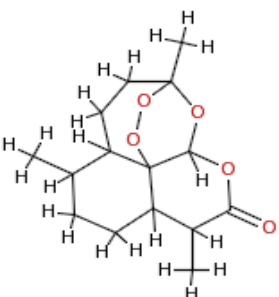
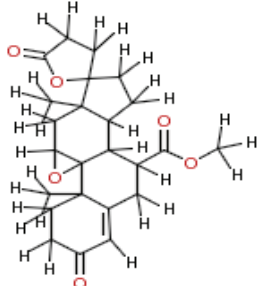
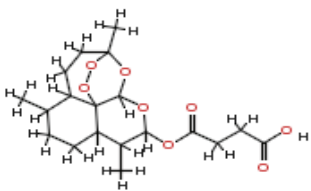


ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	11057		DB00813
	6438383		DB00920
	10364		DB01037
	2566		DB00234
	6989		DB00425



ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	9454		DB01168
	9453		DB00882
	2871		DB00874
	7572		DB00981
	37384		DB05246



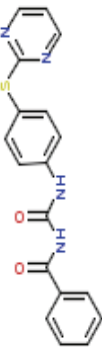
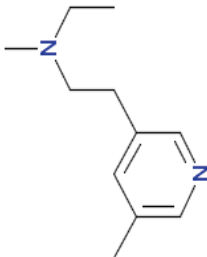
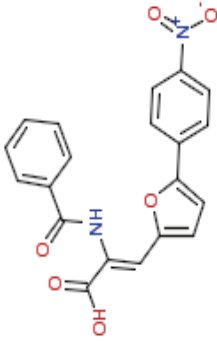
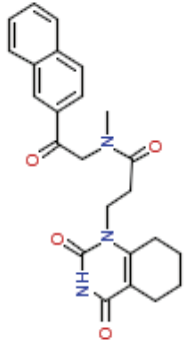
ACTIVE COMPOUNDS STRUCTURES	PUBCHEM ID	INACTIVE COMPOUNDS STRUCTURES	DRUGBANK ID
	3000469		DB00738
	20054835		DB01586
	68827		DB00700
	5464098		



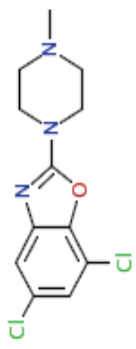
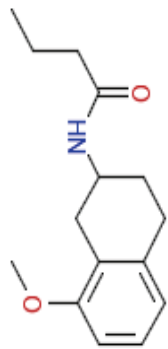
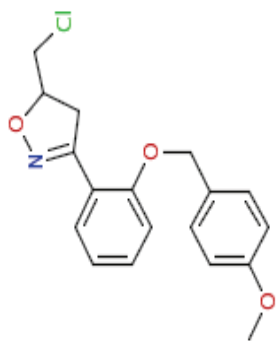
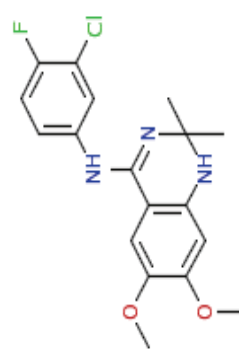
Additional File 2

In silico **approach to screen compounds active against parasitic nematodes of major socio-economic importance** by *Varun Khanna and Shoba Ranganathan*

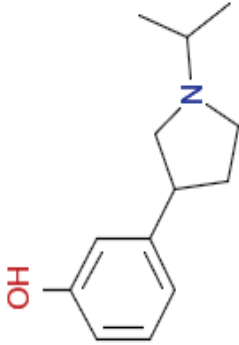
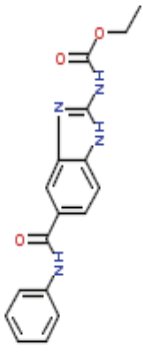

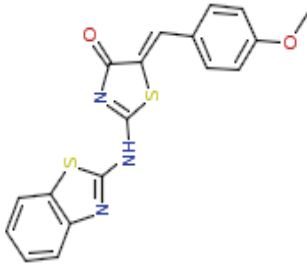
Table S2: Predicted compounds with AlogP, molecular weight and SMILES information.

S.No.	Structures	AlogP	Mol. Wt.	Canonical SMILES representation
1		3.26	350.39	<chem>O=C(NC(=O)c1ccccc1)Nc2ccc(Sc3ncccn3)cc2</chem>
2		1.91	178.27	<chem>CCN(C)CCc1cncc(C)c1</chem>
3		3.52	378.33	<chem>OC(=O)C(=C)c1oc(cc1)c2ccc(cc2)[N+](=O)[O-]NC(=O)c3ccccc3</chem>
4		2.44	419.47	<chem>CN(CC(=O)c1ccc2ccccc2c1)C(=O)CCN3C(=O)NC(=O)C4=C3CCCCC4</chem>

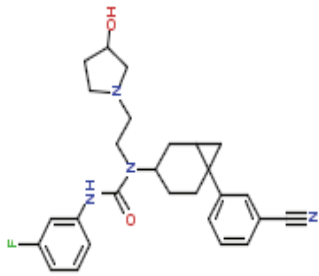
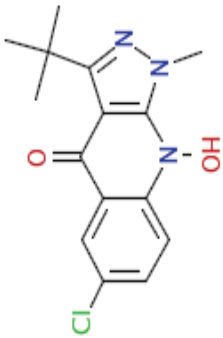
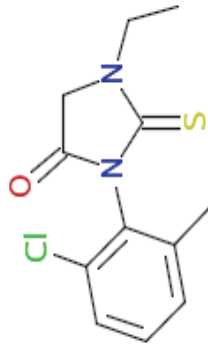
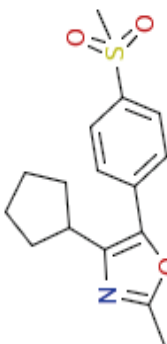


S.No.	Structures	ALogP	Mol. Wt.	Canonical SMILES representation
5		2.73	286.16	<chem>CN1CCN(CC1)c2oc3c(Cl)cc(Cl)cc3n2</chem>
6		2.94	247.33	<chem>CCCC(=O)NC1CCc2ccccc(OC)c2C1</chem>
7		3.86	331.79	<chem>COc1ccc(COc2ccccc2C3=NOC(CI)C3)cc1</chem>
8		3.44	363.81	<chem>COc1cc2NC(C)(C)N=C(Nc3ccc(F)c(Cl)c3)c2cc1OC</chem>

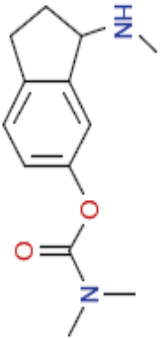
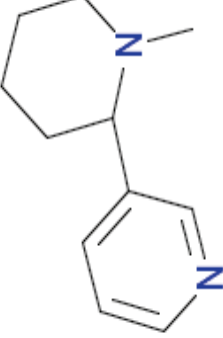

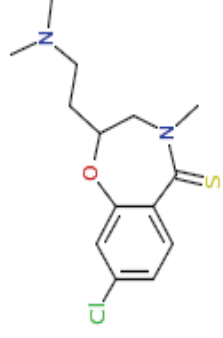
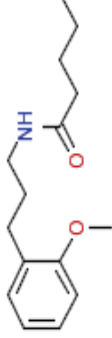


S.No.	Structures	ALogP	Mol. Wt.	Canonical SMILES representation
9		2.62	205.30	<chem>CC(C)N1CCCC(C1)c2ccccc(O)c2</chem>
10		2.78	324.33	<chem>CCOC(=O)Nc1nc2cc(ccc2[nH]1)C(=O)Nc3ccccc3</chem>
11		0.77	194.27	<chem>CN(CC#CCN1CCCC1)C(=O)C</chem>
12		4.39	367.44	<chem>COC1ccc(\C=C\2/SC(=NC2=O)Nc3nc4ccccc4s3)cc1</chem>

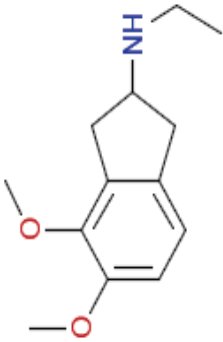
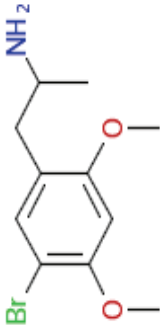
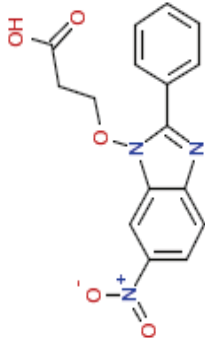
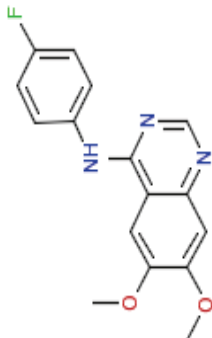


S.No.	Structures	ALogP	Mol. Wt.	Canonical SMILES representation
13		3.52	462.56	<chem>OC1CCN(CCN(C2CCCC3(CC3C2)c4cccc(c4)C#N)C(=O)Nc5cccc(F)c5)C1</chem>
14		4.23	305.76	<chem>Cn1nc(c2C(=O)c3cc(Cl)ccc3N(O)c12)C(C)(C)C</chem>
15		3.75	268.76	<chem>CCN1CC(=O)N(C1=S)c2c(C)cccc2Cl</chem>
16		3.28	305.39	<chem>Cc1oc(c2ccc(cc2)S(=O)(=O)C)c(n1)C3CCCCC3</chem>

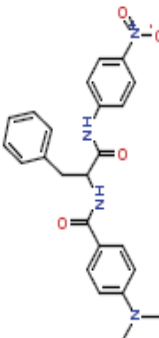
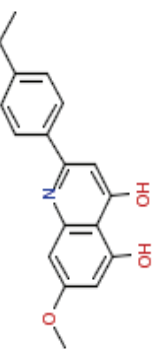
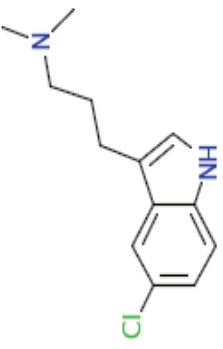
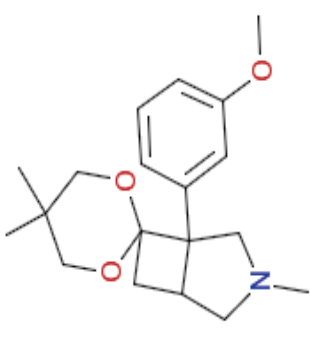


S.No.	Structures	ALogP	Mol. Wt.	Canonical SMILES representation
17		1.90	234.29	<chem>CNC1CCc2ccc(OC(=O)N(C)C)cc12</chem>
18		1.70	176.26	<chem>CN1CCCCC1c2ccccc2</chem>
19		1.63	182.26	<chem>CCCCOCCc1c[nH]cn1</chem>
20		3.36	298.83	<chem>CN(C)CCC1CN(C)C(=S)c2ccc(Cl)cc2O1</chem>
21		3.30	249.35	<chem>CCCCC(=O)NCCCC1CCCCC1OC</chem>

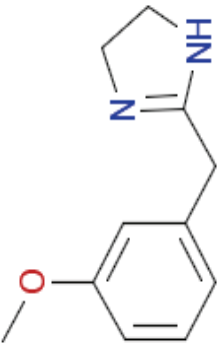
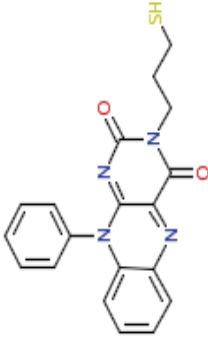
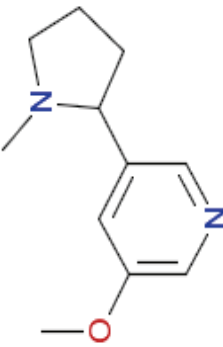
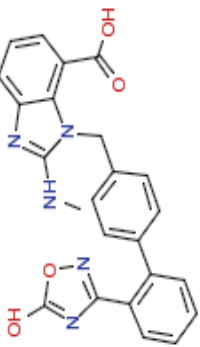


S.No.	Structures	ALogP	Mol. Wt.	Canonical SMILES representation
22		2.11	221.30	<chem>CCNC1Cc2ccc(OC)c(OC)c2C1</chem>
23		2.35	274.15	<chem>COC1cc(OC)c(CC(C)N)cc1Br</chem>
24		0.81	327.29	<chem>OC(=O)CCOn1c(nc2ccc(cc12)[N+](=O)[O-])c3ccccc3</chem>
25		3.01	299.30	<chem>COC1cc2ncnc(Nc3ccc(F)cc3)c2cc1OC</chem>

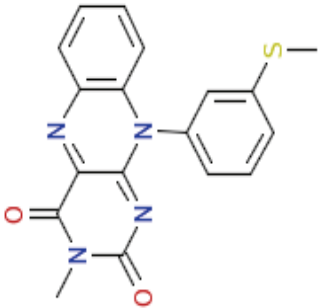
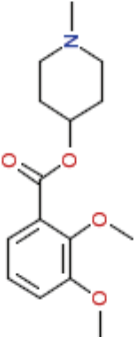
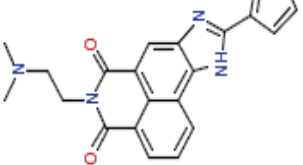
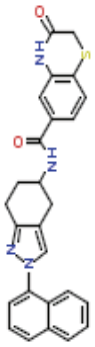


S.No.	Structures	ALogP	Mol. Wt.	Canonical SMILES representation
26		3.77	432.47	<chem>CN(C)c1ccc(cc1)C(=O)NC(Cc2ccccc2)C(=O)Nc3ccc(cc3)[N+](=O)[O-]</chem>
27		4.41	295.33	<chem>CCc1ccc(cc1)c2cc(O)c3c(O)cc(OC)cc3n2</chem>
28		3.64	236.74	<chem>CN(C)CCCC1c[nH]c2cccc(Cl)cc12</chem>
29		2.14	317.42	<chem>COC1CCCC(c1)C23CN(C)CC2CC34OCC(C)(C)CO4</chem>

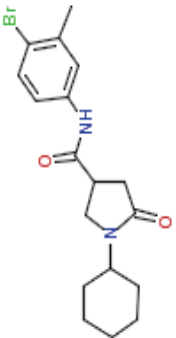
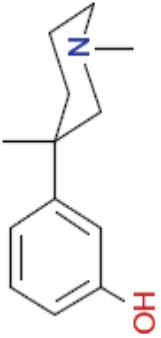
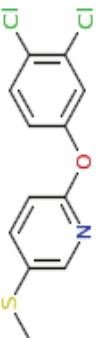
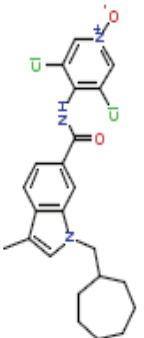
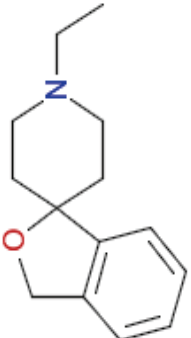


S.No.	Structures	ALogP	Mol. Wt.	Canonical SMILES representation
30		1.14	190.24	<chem>COC1CCC(CC2=NCCN2)C1</chem>
31		3.12	364.42	<chem>SCCCN1C(=O)N=C2N(c3cccc3)c4cccc4N=C2C1=O</chem>
32		1.23	192.26	<chem>COC1CNCC(c1)C2CCCN2C</chem>
33		3.85	441.44	<chem>CNc1nc2cccc(C(=O)O)c2n1Cc3cccc(cc3)c4cccc4c5noc(O)n5</chem>

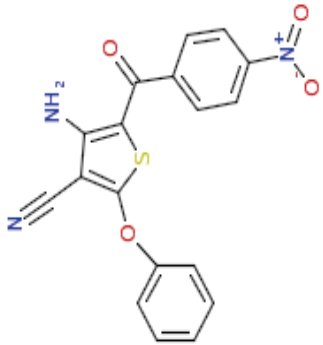
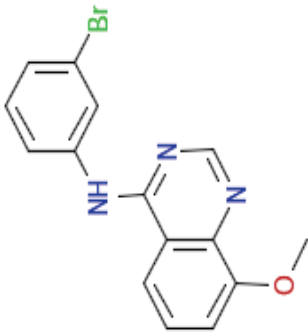


S.No.	Structures	ALogP	Mol. Wt.	Canonical SMILES representation
34		3.17	350.39	<chem>CSc1cccc(c1)N2C3=NC(=O)N(C)C(=O)C3=Nc4ccccc24</chem>
35		1.76	279.33	<chem>COC1CCCC(C(=O)OC2CCN(C)CC2)c1OC</chem>
36		2.45	374.39	<chem>CN(C)CCN1C(=O)c2cccc3c4[nH]c(nc4cc(C1=O)c23)c5ccccc5</chem>
37		3.99	454.54	<chem>O=C(NC1CCc2n(cc2C1)c3cccc4ccccc34)c5ccc6SCC(=O)Nc6cc5</chem>

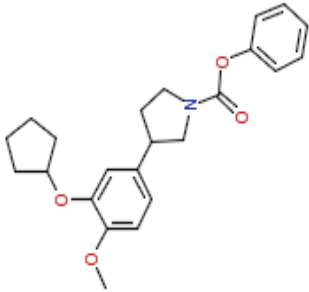


S.No.	Structures	ALogP	Mol. Wt.	Canonical SMILES representation
38		3.40	379.29	<chem>Cc1cc(NC(=O)C2CN(C3CCCCC3)C(=O)C2)ccc1Br</chem>
39		2.62	205.30	<chem>CN1CCCC(C)(C1)c2cccc(O)c2</chem>
40		4.65	286.18	<chem>CSc1ccc(Oc2ccc(Cl)c(Cl)c2)nc1</chem>
41		5.58	446.37	<chem>Cc1cn(CC2CCCCC2)c3cc(ccc13)C(=O)Nc4c(Cl)c[n+](O-)]cc4Cl</chem>
42		1.70	217.31	<chem>CCN1CCC2(CC1)OCc3cccccc23</chem>



S.No.	Structures	ALogP	Mol. Wt.	Canonical SMILES representation
43		3.90	365.36	<chem>Nc1c(C#N)c(Oc2ccccc2)sc1C(=O)c3ccc(cc3)[N+](=O)[O-]</chem>
44		3.57	330.18	<chem>COc1cccc2c(Nc3ccccc3Br)ncnc12</chem>



S.No.	Structures	ALogP	Mol. Wt.	Canonical SMILES representation
45		4.98	381.46	<chem>COc1ccc(cc1OC2CCCC2)C3CCN(C3)C(=O)Oc4ccccc4</chem>



## 6.2. Conclusion

Computational screening can significantly reduce the effort involved in discovering novel lead molecules. The objective of any virtual screening task is to identify the potentially active compounds for experimental validations. In this manuscript we employed support vector machines to discover new lead molecules that are active against parasitic nematodes.

Our active and inactive datasets contain 148 and 147 compounds, respectively. We found that despite the almost equal number of compounds in both the datasets, the number of unique scaffolds found in the inactive set is twice the number found in the active set. This result is consistent with the fact that compounds in the inactive set were collated from pharmacologically diverse backgrounds, resulting in larger diversity. Further, we noted that five of top ten scaffolds are shared between the two datasets.

During the course of this study we were able to identify 45 compounds from the prediction set with potential anthelmintic activity. Many of the predicted compounds contain the piperazine-like substructure, suggesting its importance in anthelmintic activity. From a thorough literature survey, we were able to identify five major targets in helminths and related organisms *viz.*  $\gamma$ -aminobutyric acid receptor, glutamate-gated chloride channels, glutathione S-transferase,  $\beta$ -tubulin and nicotinic acetylcholine receptor. These compounds are starting points for further investigation by high-throughput docking to these target receptors, followed by experimental validation, to confirm the activity of the predicted molecules.



## Chapter 7: Conclusions and future directions

### 7.1 Summary

This thesis is divided into seven chapters. Chapter 1 starts with a brief introduction of chemoinformatics and its role in drug discovery followed by an extensive literature survey on public, small molecule databases. Furthermore, we have reviewed methods available in chemoinformatics to design and optimize virtual combinatorial lead libraries. In addition, we have systematically reviewed studies over the past few years that were attempted to evaluate the suitability of chemical compounds as potential lead candidates, with specific pharmacokinetic properties. Chapter 2 lists the publications included in this thesis and the respective chapters they are included in, as a table for cross reference purposes.

In Chapter 3, we describe CMKb database to store, preserve and disseminate the traditional Australian Aboriginal medicinal plant knowledge. The bioactive compounds in these traditional medicinal plants have been curated and visualized using the chemoinformatics module that I have developed, as a part of our effort to help preserve customary medicinal knowledge and to integrate chemoinformatics with the customary medicinal knowledge.

Results from the analysis of physicochemical properties and functional groups of current drugs, human metabolites and toxics are described in Chapter 4. The effect of clustering on physicochemical property analysis was further analysed. It was established that although the physicochemical property space occupied by all the groups was distinct, however, currently used drugs are akin to toxic compounds than metabolites in physicochemical properties distribution. This result was in accordance with high attrition rates in drug discovery projects. We also noted that an empirical rule like Ro5 does not explicitly take toxicity information into account.

In Chapter 5, we employed a multi-criteria approach to compare various biologically relevant compounds freely available in public datasets. The compounds compared were obtained from drugs, human metabolites, toxics, natural products and screening lead libraries. We confirmed our earlier results of physicochemical analysis in this study and supplemented it by studying scaffolds frequencies and fragment co-occurrences (using



association analysis) in these datasets. For the first time to the best of our knowledge, association analysis was employed in chemoinformatics, to examine the co-occurrence of the fragments. We found that metabolites are scarcely distributed in the chemical space while drugs and natural products are quite diverse. Further, we identified few scaffolds that are present in metabolites or natural products with a close counterpart in drugs but are missing in screening libraries. Hence, we concluded that scaffold space of metabolites and natural products could provide interesting leads. We also noted that the ChEMBL database contains a large number of drug-like scaffolds, along with significant overlap to metabolites, making it a good source for novel leads.

In Chapter 6, we discuss a virtual screening application using a machine learning algorithm called the support vector machine. Compounds active against parasitic nematodes were screened from the ChEMBL dataset, after training the algorithm with actives and inactive compounds collected from literature.

Chapter 7 highlights the innovations, significance and contributions of this thesis and draws conclusions from the scaffold analysis, fragment co-occurrences analysis. This chapter also discusses future directions. The work presented in this thesis has been published as a set of book chapters and journal articles.

## **7.2 Conclusions**

This thesis reports a series of novel work in the field of chemoinformatics. From the reviews undertaken and the studies carried out, the following overall conclusions can be arrived at:

1. Empirical rules like Ro5 can be improved by including a toxicity parameter because toxicity is the main cause of attrition of compounds during a drug discovery program. Current drugs are more similar to toxics than human metabolites in physicochemical properties this might explain the withdrawal of many drugs at the later stages of drug discovery pipeline and even sometimes after the launch of a drug in the market. We identified certain functional groups that are mostly founds in toxics, this could serve as an efficient filter during lead screening (Chapter 4).
2. The low diversity of human metabolites limits their usage in lead library design. However, we identified some scaffolds and fragments in metabolites that are missing in



currently used lead compound datasets. The identified scaffolds and fragments could provide useful leads during drug discovery. Similarly we found potentially useful scaffolds and fragments in natural product dataset. Since metabolites and NPs are already optimized by millions of years of evolution to bind to at least one protein in the biosphere therefore, it is highly probable that libraries designed based on the scaffolds and fragments occurring in metabolite and NP space will result in molecules with better ADMET properties.

3. An innovative, systematic and stepwise application of association analysis (Chapter 5) for the investigation of the co-occurring molecular fragments in biologically relevant compounds has demonstrated its utility. This provides new insights to the inter-dependency of fragments, suggesting that some fragments tend to co-occur together more often in a particular compound dataset while some other fragments tend to avoid each other. The information regarding the fragment inter-dependencies could be very useful while designing combinatorial lead libraries by avoiding the combinations found in toxic and synthetically unfeasible compounds.
4. Further, in Chapter 5, we concluded that current drugs and metabolites share 7.0% of the total non-redundant scaffolds, i.e. over 42% of the metabolite scaffolds are present in drugs, whereas only 23% of the metabolite scaffolds are shared between leads and metabolites. This shows that although drugs and metabolites share many scaffolds, they largely differ in topological fragment space.

### **7.3 Innovations**

This thesis highlights original findings from the application of chemoinformatic tools to the study of the chemical property space occupied by bioactive molecules, and its significance in drug discovery and development. Several novel aspects are presented in this thesis. This is, to the best of my knowledge, the first study of its kind where the application of association rule mining, to find frequently occurring molecular substructural patterns in bioactive compounds, has been applied. This is a robust method for HTS screening of biologically interesting chemical fragments and the information on co-occurring fragments could also be extremely useful while designing lead libraries. Besides this, many other innovations, such as the CMKb chemoinformatics module developed as a part of this work, can serve as a platform to visualise and annotate chemical data and would be a useful resource for drug discovery. Further, the inclusion of human metabolites during the



analysis, has given insights into the metabolite space. Also, for the first time to the best of our knowledge, we have compared the distribution of physicochemical properties in clustered and unclustered datasets, recommending the use of clustered datasets.

## 7.4 Significance and contributions

This work reverberates with inherent importance. Among many other significances and contributions of this thesis, a few critical ones are listed below:

1. This research project proposes a prototype (multi-disciplinary CMKb database) for the conservation of customary medicinal plant knowledge and associated chemical information. (Chapter 3)
2. It offers compelling insights into physicochemical properties of current drugs, human metabolites and toxics. (Chapter 4)
3. It lists new functional groups that could serve as filters to remove toxic compounds in order to improve *hit* rates in virtual screening (Chapter 4).
4. It outlines the rationale behind the usage of metabolites and natural products scaffolds or fragments in lead designing tasks. (Chapter 5).
5. It outlines the argument for the through screening of metabolites and natural products scaffolds and fragments space for designing novel lead libraries with better ADMET properties (Chapter 5).
6. It describes an application of a robust, statistical, pattern finding technique called association analysis to help understand the co-occurrences and inter-dependencies of molecular fragments in biologically relevant molecules (Chapter 5).
7. Ligand-based VS has been successfully employed in drug discovery programs, however, it still remains an unproven approach for discovering antiparasitic drugs. This study helps to prove the validity of ligand-based VS in discovering novel leads and predicts compounds active against parasitic nematodes (Chapter 6).



8. It assists in library design, VS of active compounds and hence cuts down the time and effort involved in classical drug design.

## 7.5 Future directions

The studies presented in this thesis could lead to advancements in many directions for better understanding of chemical space occupied by the compounds of biological importance.

The methodology described in Chapter 5 to analyse the co-occurrence of fragments could be automated for high-throughput identification of strong, moderate or weak correlations among the fragments. The above approach combined with scaffold identification and optimization methodology can be developed into a fully automated ligand-based virtual screening tool. This fully automated VS tool can then be implemented as a web-server that provides service to the scientific community, especially to chemical biologists and computational chemists. The chemoinformatics module described in Chapter 3 could be used a prototype to develop a new database module or can be extended to store fragments relevant to drug-design along with their frequencies and co-occurrences information which can then be used to find bioisosteric replacements in future work. Further, the search component of the chemoinformatics module described in Chapter 3 could be upgraded with the ability to search substructures and similar molecules.

The analysis done in Chapter 6 has revealed a series of interesting compounds that could potentially be anthelmintic in nature. Experimental validation of the predicted compounds is next step in the overall analysis. Although the current methodology focuses on the use of existing machine learning approaches for screening novel compounds and data analysis, this work can be extended to structure-based approaches such as high-throughput docking methodology for screening novel compounds active against parasitic nematodes where the experimental 3D structures of the biological target is known.



## References:

1. Collier R: **Drug development cost estimates hard to swallow.** *Canadian Medical Association Journal* 2009, **180**(3):279-280.
2. Dobson CM: **Chemical space and biology.** *Nature* 2004, **432**(7019):824-828.
3. Harper G, Pickett SD, Green DV: **Design of a compound screening collection for use in high throughput screening.** *Combinatorial Chemistry & High Throughput Screening* 2004, **7**(1):63-70.
4. Schuster D, Laggner C, Langer T: **Why drugs fail--a study on side effects in new chemical entities.** *Current Pharmaceutical Design* 2005, **11**(27):3545-3559.
5. Ranganathan S: **Towards a career in bioinformatics.** *BMC Bioinformatics* 2009, **10 Suppl 15**:S1.
6. Chen WL: **Chemoinformatics: Past, Present, and Future.** *Journal of Chemical Information and Modeling* 2006, **46**(6):2230-2255.
7. Engel T: **Basic Overview of Chemoinformatics.** *Journal of Chemical Information and Modeling* 2006, **46**(6):2267-2277.
8. Varnek A, Baskin II: **Chemoinformatics as a Theoretical Chemistry Discipline.** *Molecular Informatics* 2011, **30**(1):20-32.
9. Willett P: **From chemical documentation to chemoinformatics: 50 years of chemical information science.** *Journal of Information Science* 2008, **34**(4):477-499.
10. Hann M, Green R: **Chemoinformatics -- a new name for an old problem?** *Current Opinion in Chemical Biology* 1999, **3**(4):379-383.
11. Stahura FL, Bajorath J: **Bio- and chemo-informatics beyond data management: crucial challenges and future opportunities.** *Drug Discovery Today* 2002, **7**(11):S41-S47.
12. Xing W-L, Cheng J, Shi L, Su Z, Xie A, Liao C, Qiao W, Zhang D, Shan S, Pan D *et al*: **An Integrated Biochemoinformatics System for Drug Discovery.** In: *Frontiers in Biochip Technology*. Springer US; 2006: 191-206.
13. Martin YC, Kofron JL, Traphagen LM: **Do structurally similar molecules have similar biological activity?** *Journal of Medicinal Chemistry* 2002, **45**(19):4350-4358.
14. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE: **Neighborhood Behavior: A Useful Concept for Validation of Molecular Diversity Descriptors.** *Journal of Medicinal Chemistry* 1996, **39**(16):3049-3059.



15. Martin YC, Kofron JL, Traphagen LM: **Do Structurally Similar Molecules Have Similar Biological Activity?** *Journal of Medicinal Chemistry* 2002, **45**(19):4350-4358.
16. Maldonado AG, Doucet JP, Petitjean M, Fan BT: **Molecular similarity and diversity in chemoinformatics: from theory to applications.** *Molecular Diversity* 2006, **10**(1):39-79.
17. Bath PA, Morris CA, Willett P: **Effect of standardization on fragment-based measures of structural similarity.** In.: Wiley-Blackwell; 1993.
18. Chen X, Reynolds CH: **Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients.** *Journal of Chemical Information and Computer Sciences* 2002, **42**(6):1407-1414.
19. Brown RD: **Descriptors for diversity analysis.** *Perspectives in Drug Discovery and Design* 1997, **7/8**:31-49.
20. Xue L, Bajorath J: **Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening.** *Combinatorial Chemistry & High Throughput Screening* 2000, **3**:363-372.
21. Willett P: **Chemoinformatics - similarity and diversity in chemical libraries.** *Current Opinion in Biotechnology* 2000, **11**(1):85-88.
22. Willett P, Barnard JM, Downs GM: **Chemical Similarity Searching.** *Journal of Chemical Information and Computer Sciences* 1998, **38**(6):983-996.
23. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E: **The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics.** *Journal of Chemical Information and Computer Sciences* 2003, **43**(2):493-500.
24. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner Jr, Willighagen EL: **The Blue Obelisk Interoperability in Chemical Informatics.** *Journal of Chemical Information and Modeling* 2006, **46**(3):991-998.
25. Liu K, Feng J, Young SS: **PowerMV: A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation.** *Journal of Chemical Information and Modeling* 2005, **45**(2):515-522.
26. **Pipeline Pilot** [<http://accelrys.com/>]
27. **Chemical Computing Group** [<http://www.chemcomp.com/index.htm>]



28. Spjuth O, Alvarsson J, Berg A, Eklund M, Kuhn S, Masak C, Torrance G, Wagener J, Willighagen EL, Steinbeck C *et al*: **Bioclipse 2: a scriptable integration platform for the life sciences**. *BMC Bioinformatics* 2009, **10**:397.
29. **Codessa Pro software** [[www.codessa-pro.com](http://www.codessa-pro.com)]
30. Koniver DA, Wiswesser WJ, Usdin E: **Wiswesser Line Notation: Simplified Techniques for Converting Chemical Structures to WLN**. *Science* 1972, **176**(4042):1437-1439.
31. Barnard John M, Jochum Clemens J, Welford Stephen M: **A Universal Structure/Substructure Representation for PC-Host Communication**. In: *Chemical Structure Information Systems*. vol. 400: American Chemical Society; 1989: 76-81.
32. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules**. *Journal of Chemical Information and Computer Sciences* 1988, **28**(1):31-36.
33. Homer RW, Swanson J, Jilek RJ, Hurst T, Clark RD: **SYBYL Line Notation (SLN): A Single Notation To Represent Chemical Structures, Queries, Reactions, and Virtual Libraries**. *Journal of Chemical Information and Modeling* 2008, **48**(12):2294-2307.
34. **Daylight Chemical Information Systems Inc.** [<http://www.daylight.com/>]
35. Kier LB, Hall LH: **Derivation and significance of valence molecular connectivity**. *Journal of Pharmaceutical Sciences* 1981, **70**(6):583-589.
36. Hall LH, Kier LB: **The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling**: John Wiley & Sons, Inc.; 2007.
37. Estrada E, Uriarte E: **Recent advances on the role of topological indices in drug discovery research**. *Current Medicinal Chemistry* 2001, **8**(13):1573-1588.
38. Flower DR: **On the Properties of Bit String-Based Measures of Chemical Similarity**. *Journal of Chemical Information and Computer Sciences* 1998, **38**(3):379-386.
39. Wang Y, Bajorath J: **Bit silencing in fingerprints enables the derivation of compound class-directed similarity metrics**. *Journal of Chemical Information and Modeling* 2008, **48**(9):1754-1759.
40. Barnard JM, Downs GM: **Chemical fragment generation and clustering software**. *Journal of Chemical Information and Computer Sciences* 1997, **37**(1):141-142.
41. **Tripos Inc.** [<http://www.tripos.com>]



42. Rogers D, Hahn M: **Extended-connectivity fingerprints**. *Journal of Chemical Information and Modeling* 2010, **50**(5):742-754.
43. Cramer RD, Patterson DE, Bunce JD: **Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins**. *Journal of the American Chemical Society* 1988, **110**(18):5959-5967.
44. Lemmen C, Lengauer T: **Computational methods for the structural alignment of molecules**. *Journal of Computer-Aided Molecular Design* 2000, **14**(3):215-232.
45. Klebe G, Abraham U, Mietzner T: **Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity**. *Journal of Medicinal Chemistry* 1994, **37**(24):4130-4146.
46. Cruciani C, Crivori P, Carrupt PA, Testa B: **Molecular fields in quantitative structure-permeation relationships: the VolSurf approach**. *Journal of Molecular Structure-Theochem* 2000, **503**(1-2):17-30.
47. Pastor M, Cruciani G, McLay I, Pickett S, Clementi S: **GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors**. *Journal of Medicinal Chemistry* 2000, **43**(17):3233-3243.
48. Nettles JH, Jenkins JL, Bender A, Deng Z, Davies JW, Glick M: **Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors**. *Journal of Medicinal Chemistry* 2006, **49**(23):6802-6810.
49. Brown RD, Martin YC: **The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding**. *Journal of Chemical Information and Computer Sciences* 1997, **37**(1):1-9.
50. Schuffenhauer A, Gillet VJ, Willett P: **Similarity Searching in Files of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-Dimensional Fingerprints and Molecular Field Descriptors**. *Journal of Chemical Information and Computer Sciences* 1999, **40**(2):295-307.
51. Makara GM: **Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity**. *Journal of Medicinal Chemistry* 2001, **44**(22):3563-3571.
52. Andrews KM, Cramer RD: **Toward General Methods of Targeted Library Design: Topomer Shape Similarity Searching with Diverse Structures as Queries**. *Journal of Medicinal Chemistry* 2000, **43**(9):1723-1740.



53. Feng J, Lurati L, Ouyang H, Robinson T, Wang Y, Yuan S, Young SS: **Predictive toxicology: benchmarking molecular descriptors and statistical methods.** *Journal of Chemical Information and Computer Sciences* 2003, **43**(5):1463-1470.
54. Gorse D, Rees A, Kaczorek M, Lahana R: **Molecular diversity and its analysis.** *Drug Discovery Today* 1999, **4**(6):257-264.
55. Kohavi R, John GH: **Wrappers for feature subset selection.** *Artificial Intelligence* 1997, **97**(1-2):273-324.
56. Pudil P, Ferri FJ, Novovicova J, Kittler J: **Floating search methods for feature selection with nonmonotonic criterion functions.** In: *Pattern Recognition, 1994 Vol 2 - Conference B: Computer Vision & Image Processing, Proceedings of the 12th IAPR International Conference on: 9-13 Oct 1994*; 1994: 279-283 vol. 272.
57. Shanno DF: **Conditioning of Quasi-Newton Methods for Function Minimization.** *Mathematics of Computation* 1970, **24**(111):647-656.
58. Nelder JA, Mead R: **A Simplex Method for Function Minimization.** *The Computer Journal* 1965, **7**(4):308-313.
59. Goldberg DE: **Genetic Algorithms.** Addison-Wesley Professional, New York, 1989.
60. Kirkpatrick S, Gelatt CD, Jr., Vecchi MP: **Optimization by simulated annealing.** *Science* 1983, **220**(4598):671-680.
61. Lin W-Q, Jiang J-H, Shen Q, Shen G-L, Yu R-Q: **Optimized Block-wise Variable Combination by Particle Swarm Optimization for Partial Least Squares Modeling in Quantitative Structure Activity Relationship Studies.** *Journal of Chemical Information and Modeling* 2005, **45**(2):486-493.
62. Rogers D, Hopfinger AJ: **Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships.** *Journal of Chemical Information and Computer Sciences* 1994, **34**(4):854-866.
63. Ghosh P, Bagchi MC: **QSAR Modeling for Quinoxaline Derivatives using Genetic Algorithm and Simulated Annealing Based Feature Selection.** *Current Medicinal Chemistry* 2009, **16**(30):4032-4048.
64. Hansen L, Lee EA, Hestir K, Williams LT, Farrelly D: **Controlling Feature Selection in Random Forests of Decision Trees Using a Genetic Algorithm: Classification of Class I MHC Peptides.** *Combinatorial Chemistry & High Throughput Screening* 2009, **12**(5):514-519.



65. Černý V: **Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm.** *Journal of Optimization Theory and Applications* 1985, **45**(1):41-51.
66. Guha R, Jurs PC: **Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors.** *Journal of Chemical Information and Computer Sciences* 2004, **44**(6):2179-2189.
67. Lin T-H, Li H-T, Tsai K-C: **Implementing the Fisher's Discriminant Ratio in a k-Means Clustering Algorithm for Feature Selection and Data Set Trimming.** *Journal of Chemical Information and Computer Sciences* 2003, **44**(1):76-87.
68. Tanimoto TT: **An elementary mathematical theory of classification and prediction by T.T. Tanimoto.** New York: International Business Machines Corporation; 1958.
69. Fligner MA, Verducci JS, Blower PE: **A Modification of the Jaccard Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings.** *Technometrics* 2002, **44**(2):110-119.
70. Russel PF, Rao TR: **On habitat and association of species of anopheline larvae in south-eastern Madras** *Journal of Malerial Institute of India* 1940, **3**:154-178.
71. Ochiai A: **Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions** *Bulletins of the Japanese Society for Scientific Fisheries* 1957, **22**:526-530.
72. Fallaw WC: **A Test of the Simpson Coefficient and Other Binary Coefficients of Faunal Similarity.** *Journal of Paleontology* 1979, **53**(4):1029-1034.
73. Forbes SA: **On the local distribution of certain Illinois fishes.** . *Bulletin of the Illinosis State Laboratory of Natural History* 1907, **7**:8.
74. Dice LR: **Measures of the Amount of Ecologic Association Between Species.** *Ecology* 1945, **26**(3):297-302.
75. Dennis SF: **The construction of a thesaurus automatically from a sample of text.** In: *Statistical association techniques for mechanized documentation: Symposium proceedings National Bureau of Standards* 1965; 1965: 269.
76. Ellis D, Furner-Hines J, Willett P: **Measuring the degree of similarity between objects in text retrieval systems.** *Perspectives in Information Management* 1994, **3**:128– 149.



77. Khalifa AA, Haranczyk M, Holliday J: **Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection.** *Journal of Chemical Information and Modeling* 2009, **49**(5):1193-1201.
78. Lajiness M, Watson I: **Dissimilarity-based approaches to compound acquisition.** *Current Opinion in Chemical Biology* 2008, **12**(3):366-371.
79. Haranczyk M, Holliday J: **Comparison of Similarity Coefficients for Clustering and Compound Selection.** *Journal of Chemical Information and Modeling* 2008, **48**(3):498-508.
80. Whittle M, Willett P, Klaffke W, van Noort P: **Evaluation of Similarity Measures for Searching the Dictionary of Natural Products Database.** *Journal of Chemical Information and Computer Sciences* 2003, **43**(2):449-457.
81. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Deliv Rev* 2001, **46**(1-3):3-26.
82. Leeson PD, Davis AM: **Time-Related Differences in the Physical Property Profiles of Oral Drugs.** *Journal of Medicinal Chemistry* 2004, **47**(25):6338-6348.
83. Ghose AK, Viswanadhan VN, Wendoloski JJ: **A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases.** *Journal of Combinatorial Chemistry* 1999, **1**(1):55-68.
84. Hann MM, Oprea TI: **Pursuing the leadlikeness concept in pharmaceutical research.** *Current Opinion in Chemical Biology* 2004, **8**(3):255-263.
85. Ertl P, Roggo S, Schuffenhauer A: **Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries.** *Journal of Chemical Information and Modeling* 2007, **48**(1):68-74.
86. Eckert H, Bajorath Jr: **Exploring Peptide-likeness of Active Molecules Using 2D Fingerprint Methods.** *Journal of Chemical Information and Modeling* 2007, **47**(4):1366-1378.
87. Gupta S, de Sousa A: **Comparing the chemical spaces of metabolites and available chemicals: models of metabolite-likeness.** *Molecular Diversity* 2007, **11**(1):23-36.
88. Dobson P, Patel Y, Kell D: **'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries.** *Drug Discovery Today* 2009, **14**(1-2):31-40.



89. Congreve M, Carr R, Murray C, Jhoti H: **'A Rule of Three' for fragment-based lead discovery?** *Drug Discovery Today* 2003, **8**(19):876-877.
90. Bemis GW, Murcko MA: **The Properties of Known Drugs. 1. Molecular Frameworks.** *Journal of Medicinal Chemistry* 1996, **39**(15):2887-2893.
91. Medina-Franco JL, Martínez-Mayorga K, Bender A, Scior T: **Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure.** *QSAR & Combinatorial Science* 2009, **28**(11-12):1551-1560.
92. Singh N, Guha R, Giulianotti MA, Pinilla C, Houghten RA, Medina-Franco JL: **Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository.** *Journal of Chemical Information and Modeling* 2009, **49**(4):1010-1024.
93. de Kloe G, Bailey D, Leurs R, de Esch I: **Transforming fragments into candidates: small becomes big in medicinal chemistry.** *Drug Discovery Today* 2009, **14**(13-14):630-646.
94. Chessari G, Woodhead A: **From fragment to clinical candidate: a historical perspective.** *Drug Discovery Today* 2009, **14**(13-14):668-675.
95. Boyd SM, de Kloe GE: **Fragment library design: efficiently hunting drugs in chemical space.** *Drug Discovery Today: Technologies* 2010, **7**(3):e173-e180.
96. Krueger BA, Dietrich A, Baringhaus KH, Schneider G: **Scaffold-Hopping Potential of Fragment-Based De Novo Design: The Chances and Limits of Variation.** *Combinatorial Chemistry & High Throughput Screening* 2009, **12**:383-396.
97. Horton DA, Bourne GT, Smythe ML: **The Combinatorial Synthesis of Bicyclic Privileged Structures or Privileged Substructures.** *Chemical Reviews* 2003, **103**(3):893-930.
98. Lewell XQ, Judd DB, Watson SP, Hann MM: **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* 1998, **38**(3):511-522.
99. Graham DJ, Malarkey C, Schulmerich MV: **Information Content in Organic Molecules:Quantification and Statistical Structure via Brownian Processing.** *Journal of Chemical Information and Computer Sciences* 2004, **44**(5):1601-1611.
100. Wang J, Hou T: **Drug and Drug Candidate Building Block Analysis.** *Journal of Chemical Information and Modeling* 2009, **50**(1):55-67.



101. Hu Y, Bajorath Jr: **Molecular Scaffolds with High Propensity to Form Multi-Target Activity Cliffs**. *Journal of Chemical Information and Modeling* 2010, **50**(4):500-510.
102. Blower PE, Cross KP: **Decision tree methods in pharmaceutical research**. *Current Topics in Medicinal Chemistry* 2006, **6**(1):31-39.
103. Salzberg SL: **C4.5: Programs for Machine Learning by J. Ross Quinlan**. **Morgan Kaufmann Publishers, Inc., 1993**. *Machine Learning* 1994, **16**(3):235-240-240.
104. Wagener M, van Geerestein VJ: **Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features**. *Journal of Chemical Information and Computer Sciences* 2000, **40**(2):280-292.
105. Rusinko A, Farnen MW, Lambert CG, Brown PL, Young SS: **Analysis of a large structure/biological activity data set using recursive partitioning**. *Journal of Chemical Information and Computer Sciences* 1999, **39**(6):1017-1026.
106. **Self-Organizing Maps**: Springer-Verlag New York, Inc.; 2001.
107. Schneider G, Wrede P: **Artificial neural networks for computer-based molecular design**. *Progress in Biophysics and Molecular Biology* 1998, **70**(3):175-222.
108. Devillers J: **Neural Networks in QSAR and Drug Design**. Orlando, FL, USA. : Academic Press, Inc.; 1996.
109. Kubinyi H, Folkers G, Martin YC, Anzali S, Gasteiger J, Holzgrabe U, Polanski J, Sadowski J, Wagener ATM: **The Use of Self-organizing Neural Networks in Drug Design**. In: *3D QSAR in Drug Design*. vol. 2: Springer Netherlands; 2002: 273-299-299.
110. Ajay, Walters WP, Murcko MA: **Can We Learn To Distinguish between Drug-like and Nondrug-like Molecules?** *Journal of Medicinal Chemistry* 1998, **41**(18):3314-3324.
111. Cortes C, Vapnik V: **Support-Vector Networks**. *Machine Learning* 1995, **20**(3):273-297.
112. Burges C: **A Tutorial on Support Vector Machines for Pattern Recognition**. *Data Mining and Knowledge Discovery* 1998, **2**(2):121-167.
113. Liu XH, Ma XH, Tan CY, Jiang YY, Go ML, Low BC, Chen YZ: **Virtual Screening of Abl Inhibitors from Large Compound Libraries by Support Vector Machines**. *Journal of Chemical Information and Modeling* 2009, **49**(9):2101-2110.



114. Jorissen RN, Gilson MK: **Virtual Screening of Molecular Databases Using a Support Vector Machine.** *Journal of Chemical Information and Modeling* 2005, **45**(3):549-561.
115. Yap CW, Chen YZ: **Prediction of Cytochrome P450 3A4, 2D6, and 2C9 Inhibitors and Substrates by Using Support Vector Machines.** *Journal of Chemical Information and Modeling* 2005, **45**(4):982-992.
116. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV: **Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions.** *Journal of Chemical Information and Computer Sciences* 2003, **43**(6):2048-2056.
117. Warmuth MK, Liao J, Rätsch G, Mathieson M, Putta S, Lemmen C: **Active Learning with Support Vector Machines in the Drug Discovery Processes.** *Journal of Chemical Information and Computer Sciences* 2003, **43**(2):667-673.
118. Burbidge R, Trotter M, Buxton B, Holden S: **Drug design by machine learning: support vector machines for pharmaceutical data analysis.** *Computers & Chemistry* 2001, **26**(1):5-14.
119. Meyer D, Leisch F, Hornik K: **The support vector machine under test.** *Neurocomputing* 2003, **55**(1-2):169-186.
120. Agrawal R, Imielinski T, Swami AN: **Mining association rules between sets of items in large databases.** In: *SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data* vol. 22. New York, USA; 1993: 207-216.
121. Atluri G, Gupta R, Fang G, Pandey G, Steinbach M, Kumar V: **Association Analysis Techniques for Bioinformatics Problems.** *Bioinformatics and Computational Biology, Proceedings* 2009, **5462**:1-13.
122. Kuramochi M, Karypis G: **An efficient algorithm for discovering frequent subgraphs.** *Knowledge and Data Engineering, IEEE Transactions on* 2004, **16**(9):1038-1051.
123. He Y: **Application Research of Association Analysis in Business Intelligence.** In: *2009*; 2009: 304-307.
124. Darawaty IS, Syarah S, Nugroho AS, Ayuningtyas F, Istanto F, Prasetyo B, Uliniansyah MT, Gunawan M, Desiani, Jarin A, Handoko D: **Intelligent Searching Using Association Analysis for law Documents of Indonesian Government.** In: *Proc. of 2nd International Conference on Advances in Computer, Control & Telecommunication Technologies (ACT 2010)*; 2010: 122-124.



125. Furukawa T, Ishizuka M, Matsuo Y, Ohmukai I, Uchiyama K: **Analyzing reading behavior by blog mining**. In: *AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence 2007*; 2007.
126. Baldi P: **Chemoinformatics, drug design, and systems biology**. *Genome Informatics* 2005, **16**(2):281-285.
127. Chen J, Linstead E, Swamidass J, Wang D, Baldi P: **ChemDB update full-text search and virtual chemical space**. *Bioinformatics* 2007, **23**(17):2348-2351.
128. Oprea T, Tropsha A: **Target, chemical and bioactivity databases - integration is key**. *Drug Discovery Today: Technologies* 2006, **3**(4):357-365.
129. Irwin J, Shoichet B: **ZINC--a free database of commercially available compounds for virtual screening**. *Journal of Chemical Information and Modeling* 2005, **45**(1):177-182.
130. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH: **PubChem: a public information system for analyzing bioactivities of small molecules**. *Nucleic Acids Research* 2009, **37**(suppl 2):W623-W633.
131. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M: **ChEBI: a database and ontology for chemical entities of biological interest**. *Nucleic Acids Research* 2008, **36**(suppl 1):D344-D350.
132. Seiler KP, George G, Happ MP, Bodycombe N, Carrinski H, Norton S, Brudz S, Sullivan J, Muhlich J, Serrano M *et al*: **ChemBank: a small-molecule screening and cheminformatics resource database**. *Nucleic Acids Research* 2008, **36**(Database issue): D351-D359.
133. Tomasulo P: **ChemIDplus-super source for chemical and drug information**. *Medical Reference Services Quarterly* 2002, **21**(1):53-59.
134. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T *et al*: **KEGG for linking genomes to life and the environment**. *Nucleic Acids Research* 2008, **36**(Database issue):D480-484.
135. Huang N, Shoichet BK, Irwin JJ: **Benchmarking Sets for Molecular Docking**. *Journal of Medicinal Chemistry* 2006, **49**(23):6789-6801.
136. von Grotthuss M, Koczyk G, Pas J, Wyrwicz LS, Rychlewski L: **Ligand.Info Small-Molecule Meta-Database**. *Combinatorial Chemistry & High Throughput Screening* 2004, **7**(8):757-761.



137. Milne GWA, Nicklaus MC, Driscoll JS, Wang S, Zaharevitz D: **National Cancer Institute Drug Information System 3D Database**. *Journal of Chemical Information and Computer Sciences* 1994, **34**(5):1219-1224.
138. Schmidt U, Struck S, Gruening B, Hossbach J, Jaeger IS, Parol R, Lindequist U, Teuscher E, Preissner R: **SuperToxic: a comprehensive database of toxic compounds**. *Nucleic Acids Research* 2009, **37**(suppl 1):D295-D299.
139. Richard AM, Williams CR: **Distributed structure-searchable toxicity (DSSTox) public database network: a proposal**. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 2002, **499**(1):27-52.
140. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic Acids Research* 2008, **36**(suppl 1):D901-D906.
141. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S *et al*: **HMDB: a knowledgebase for the human metabolome**. *Nucleic Acids Research* 2009, **37**(suppl 1):D603-D610.
142. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK: **BindingDB: a web-accessible database of experimentally determined protein ligand binding affinities**. *Nucleic Acids Research* 2006, **35**(suppl 1):D198-D201.
143. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, Huang L, Guo Y, Han L, Zheng C *et al*: **Update of TTD: Therapeutic Target Database**. *Nucleic Acids Research* 2010, **38**(suppl 1):D787-D791.
144. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R *et al*: **The Universal Protein Resource (UniProt): an expanding universe of protein information**. *Nucleic Acids Research* 2006, **34**(Database issue):D187-D191.
145. Gunther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiss A, Jensen LJ *et al*: **SuperTarget and Matador: resources for exploring drug-target relationships**. *Nucleic Acids Research* 2008, **36**(Database issue):D919-922.
146. Overington J: **ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI)**. Interview by Wendy A. Warr. *Journal of Computer Aided Molecular Design* 2009, **23**(4):195-198.
147. Gong L, Owen R, Gor W, Altman R, Klein T: **PharmGKB: an integrated resource of pharmacogenomic data and knowledge**. *Current protocols in*



- bioinformatics / editorial board, Andreas D Baxevanis et al. 2008, Chapter 14, Unit 14.7.*
148. Heinrich M, Lew M, Hung-Wen L: **Ethnopharmacology and Drug Discovery.** In: *Comprehensive Natural Products II.* Oxford: Elsevier; 2010: 351-381.
  149. Verpoorte R, Lew M, Hung-Wen L: **Overview and Introduction.** In: *Comprehensive Natural Products II.* Oxford: Elsevier; 2010: 1-4.
  150. Wei HY, Tsai KC, Lin TH: **Modeling ligand-receptor interaction for some MHC class II HLA-DR4 peptide mimetic inhibitors using several molecular docking and 3D QSAR techniques.** *Journal of Chemical Information and Modeling* 2005, **45**(5):1343-1351.
  151. Fabricant DS, Farnsworth NR: **The value of plants used in traditional medicine for drug discovery.** *Environmental Health Perspectives* 2001, **109** Suppl 1:69-75.
  152. Khanna V, Ranganathan S: **Physicochemical property space distribution among human metabolites, drugs and toxins.** *BMC Bioinformatics* 2009, **10** Suppl **15**:S10.
  153. Lacey E: **Mode of action of benzimidazoles.** *Parasitology Today* 1990, **6**(4):112-115.
  154. Cully DF, Vassilatis DK, Liu KK, Paress PS, Van der Ploeg LHT, Schaeffer JM, Arena JP: **Cloning of an avermectin-sensitive glutamate-gated chloride channel from *Caenorhabditis elegans*.** *Nature* 1994, **371**(6499):707-711.
  155. Dye-Holden L, Walker RJ: **Avermectin and avermectin derivatives are antagonists at the 4-aminobutyric acid (GABA) receptor on the somatic muscle cells of *Ascaris*; is this the site of anthelmintic action?** *Parasitology* 1990, **101**(02):265-271.
  156. Bokisch AJ, Walker RJ: **The action of avermectin (MK 936) on identified central neurones from *Helix* and its interaction with acetylcholine and gamma-aminobutyric acid (GABA) responses.** *Comparative Biochemistry and Physiology Part C: Comparative Pharmacology* 1986, **84**(1):119-125.
  157. Martin RJ, Verma S, Levandoski M, Clark CL, Qian H, Stewart M, Robertson AP: **Drug resistance and neurotransmitter receptors of nematodes: recent studies on the mode of action of levamisole.** *Parasitology* 2005, **131** Suppl:S71-84.
  158. Geerts S, Gryseels B: **Drug resistance in human helminths: current situation and lessons from livestock.** *Clinical Microbiology Reviews* 2000, **13**(2):207-222.
  159. James CE, Hudson AL, Davey MW: **Drug resistance mechanisms in helminths: is it survival of the fittest?** *Trends in Parasitology* 2009, **25**(7):328-335.