

Chemoinformatics based investigation of plant metabolites for their medicinal and crop protection values

by

**Divya Karade
10BB15J26034**

A thesis submitted to the
Academy of Scientific & Innovative Research

for the award of the degree of
DOCTOR OF PHILOSOPHY

in
SCIENCE

Under the supervision of

Dr. M. Karthikeyan (Supervisor)

Dr. Narendra Y. Kadoo (Co-supervisor)



CSIR - National Chemical Laboratory, Pune



Academy of Scientific and Innovative Research
AcSIR Headquarters, CSIR-HRDC Campus
Sector 19, Kamla Nehru Nagar,
Ghaziabad, U.P. – 201 002, India

May 2021

Certificate

This is to certify that the work incorporated in this Ph.D. thesis entitled, "Chemoinformatics based investigation of plant metabolites for their medicinal and crop protection values", submitted by Divya Karade to the Academy of Scientific and Innovative Research (AcSIR) in fulfillment of the requirements for the award of the Degree of Doctor of Philosophy in Science, embodies original research work carried out by the student. We, further certify that this work has not been submitted to any other University or Institution in part or full for the award of any degree or diploma. Research material(s) obtained from other source(s) and used in this research work has/have been duly acknowledged in the thesis. Image(s), illustration(s), Figure(s), table(s) etc., used in the thesis from other source(s), have also been duly cited and acknowledged.



(Signature of Student)

Divya Karade
04-05-2021



(Signature of Co-Supervisor)

Dr. Narendra Kadoo
04-05-2021



(Signature of Supervisor)

Dr. M. Karthikeyan
04-05-2021

Statements of Academic Integrity

I, Divya Karade, a Ph.D. student of the Academy of Scientific and Innovative Research (AcSIR) with Registration No. 10BB15J26034, hereby undertake that, the thesis entitled “Chemoinformatics based investigation of plant metabolites for their medicinal and crop protection values” has been prepared by me and that the document reports original work carried out by me and is free of any plagiarism in compliance with the UGC Regulations on “*Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions (2018)*” and the CSIR Guidelines for “*Ethics in Research and in Governance (2020)*”.



Signature of the Student

Date: 04-05-2021

Place: Pune

It is hereby certified that the work done by the student, under my/our supervision, is plagiarism-free in accordance with the UGC Regulations on “*Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions (2018)*” and the CSIR Guidelines for “*Ethics in Research and in Governance (2020)*”



Signature of the Co-supervisor

Name : Dr. Narendra Kadoo

Date : 04-05-2021

Place : Pune



Signature of the Supervisor

Name : Dr. M. Karthikeyan

Date : 04-05-2021

Place : Pune

Dedicated

To wonderful readers...

Acknowledgments

It is my pleasure to convey my gratitude and acknowledge all who have contributed to my Ph.D. thesis research and writing this dissertation. First of all, I owe sincere appreciation to my research advisor Dr. M. Karthikeyan for his constructive suggestions and criticism, which helped me to focus on my research objectives. He has molded me into an independent researcher by giving me all the intellectual freedom to put my thoughts into designing and conducting the experiments. Without his guidance and persistent help, this dissertation would not have been possible.

I would like to convey my sincere gratitude to my research co-advisor, Dr. Narendra Kadoo, for his supervision, patience, motivation and guidance from the very early stage of this research. His enthusiasm for teaching as well as learning new computational tools has greatly influenced me. He helped me unconditionally in all aspects and I had the privilege of learning many things from him.

I want to thank my DAC members, Dr. Anu Raghunathan, Dr. Dhiman Sarkar and Dr. Dhanasekaran Shanmugam for their timely feedback and assessment of the progress of my doctoral research. I wish to thank the Director, CSIR-NCL, Head of Department, Biochemical Sciences Division, and former heads for providing the infrastructure and lab facilities.

I am thankful to Dr. Santha Kumari for providing me the central facilities of CSIR-National Chemical Laboratory for metabolite data acquisition. I am also thankful to Mr. Yogesh Mahajan for helping me in performing field studies with soybean crop. I thank Dr Renu Vyas, Head, MIT School of Bioengineering Sciences & Research, for painstakingly proofreading my thesis despite her busy schedule.

I enjoyed the company of my friends and labmates- Samir, Sanjeev, Rashid, Shrilatha, Tapos, Vijay, Nilofer, Amay, Sanket, Rakhi, Deepthi, Ruchita, Amol K., Deepika, Pranjali, Sucheta, Swapnil, Tejas, Uma, Vaishnavi, Bhakti, Sonal, Gouri, Monika, Sagar, Santosh, and other lab members. I would like to thank all the supporting staff, including Jagtap kaka. A special thanks to my special friend, Mr. Ajit.

My sincere thanks also go to my M.Sc. guide and Retd. Head of the Biotechnology Dept., Dr. S. Sivaramakrishnan and many other teachers, who instilled in me an interest in biotechnology research and motivated me to undertake doctoral research.

I feel happy to express my thanks to Prof. Ashwini Kumar Nangia, former Director, CSIR-NCL, for providing me with the research facilities. I must thank the office staff of the DIRC, Biochemical Sciences Division and Students Academic office, who were always ready to help whenever required. I thank CSIR-UGC, India for supporting my research work through research fellowships.

Last but not the least, I would like to pay high regard to my family. There are no words to acknowledge my family for their constant support and encouragement. I owe my deepest gratitude to my brother, Dr. Vikas Karade, for giving me freedom and supporting me all the way, Momi, Papa for their support and love which made me reach this level.

Besides this, several people have knowingly and unknowingly helped me in the successful completion of this project. I thank all of them for their assistance.



Divya Karade

Table of Contents

Table of Contents	i
List of Figures	iv
List of Tables	vii
Abbreviations	ix
Synopsis Report	xi
Introduction	xi
Statement of the problem	xii
Methodology	xii
Results	xiii
Summary and future directions	xvi
Chapter 1: Introduction and Review of Literature	1
1.1 A brief overview of chemoinformatics and metabolomics	1
1.2 Technological trends in plant metabolomics.....	1
1.2.1 Experimental approach	5
1.2.2 Computational approach.....	6
1.3 Classification of phytochemicals in plants.....	11
1.3.1 Phenolic compounds.....	12
1.3.2 Phytosterols	16
1.3.3 Phytates.....	16
1.3.4 Nitrogenous compounds	17
1.4 Role of secondary metabolites as bioactive compounds.....	18
1.4.1 Role of bioactive compounds from food crops	18
1.4.2 Role of bioactive compounds from medicinal plants	19
1.4.3 Network analysis of bioactive compounds	20
1.5 Role of secondary metabolites in plant defense	27
1.6 Conservation of endangered species of valuable plants and trees	30
1.7 Genesis of the thesis.....	30
Chapter 2: Design of Novel Drug-like Molecules using Informatics Rich Secondary Metabolites Analysis of Indian Medicinal and Aromatic Plants	44
2.1 Introduction	44
2.2 Materials and methods	50

2.2.1 Data Collection	50
2.2.2 Computational protocol	52
2.2.3 Softwares and Databases	56
2.3 Results and Discussion.....	57
2.3.1 Chemoinformatics Analysis Based on Scientific Literature Mining.....	59
2.3.2 Scaffold Drug Network of the Indian Medicinal and Aromatic Plant Species	68
2.3.3 Screening of the Virtual Library.....	77
2.3.4 Cluster Analysis of Virtual Library	80
2.3.5 Applications of the Virtual Library	83
2.3.6 DoMINE	85
2.4 Conclusions	88
Chapter 3: Bridging In-Silico and Experimental: Chemoinformatics Analysis for Mass Spectrometry-Based Metabolomics study of Soybean	103
3.1 Introduction	103
3.2 Materials and methods	106
3.2.1 Chemoinformatics analysis.....	106
3.2.2 Metabolomics analysis	109
3.3 Results and Discussion.....	116
3.3.1 Chemoinformatics analysis of soybean phytochemicals	116
3.3.2 Metabolomics profiling	122
3.3.3 Soybean scaffold drug network	148
3.3.4 Development of a virtual library and virtual screening.....	157
3.4 Conclusions	163
Chapter 4: Chemoinformatics Investigation on Chemical Defense in Plants.....	175
4.1 Introduction	175
4.1.1 Role of secondary metabolites.....	177
4.2 Materials and Methods	181
4.3 Results and Discussion.....	183
4.3.1 Chemoinformatics analysis.....	183
4.3.2 Scaffold molecule network.....	195
4.3.3 Virtual Library	202
4.4 Conclusions	207
Chapter 5: Summary and Future Directions	221

5.1 Summary	221
5.2 Future directions.....	223
Bibliography	226
Supplementary Data.....	247
Abstract.....	255
Details of the Publications	256

List of Figures

Figure 1.1: Schematic view of biosynthesis of secondary metabolites in plants.....	3
Figure 1.2: Plant metabolomics workflow.....	5
Figure 1.3: Chemoinformatics methods in designing of novel molecules from organic metabolites of plants (VL- Virtual Library, PDL- Progressive Drug Like score, PLL- Progressive Lead Like score, DLF- Drug Like Failure, LLF- Lead Like Failure, TPC – Toxicophoric, Pharmacophoric and Chemophoric scores).....	8
Figure 1.4: Network containing food crops and Ayurvedic medicinal plants with their respective bioactive compounds targeting various proteins involved in respective diseases. (Nodes = 53, edges = 138; Black edges: Interactions/ hidden relationships, Color edges: Respective pathways)	22
Figure 1.5: Hypothesis for designing of novel drug-like, lead-like, and pesticide-like molecules from natural plant resources	31
Figure 2.1: Workflow highlighting the steps of extracting drug-like molecules from medicinal and aromatic plants (TPC = Toxicophoric, Pharmacophoric, and Chemophoric, PDL= Progressive Drug Like, PLL= Progressive Lead Like, DLF= Drug Like Failure, LLF= Lead Like Failure features as generated in ChemScreener program, PBC= Plant-Based Clustering).....	53
Figure 2.2: Schematic view of 16 medicinal properties of 104 Indian medicinal and aromatic plants (* List of all other medicinal plants with their medicinal properties is provided in Supplementary Table S2.4).	59
Figure 2.3: Distribution of the number of PubMed publications for the top 10 families of the medicinal and aromatic plants (as of March 2021)	60
Figure 2.4: Distribution of the number of PubMed publications for the top 10 medicinal and aromatic plants (as of March 2021).....	61
Figure 2.5: The 2D PCA plot representing the molecular diversity of natural products from Indian medicinal plants (Supplementary Table S2.5).....	63
Figure 2.6: Indian medicinal aromatic plant molecules, drug molecules, and scaffolds merged network as depicted in organic and edge-weighted spring embedded layout (for selected nodes only) in Cytoscape. Nodes = 4623, edges = 6216 (Nodes:	

Molecules, scaffolds, plants and plant families; Edges: Interactions/ hidden relationships).....	70
Figure 2.7: Dendrograms for toxicophoric (a), pharmacophoric (b), and chemophoric (c) fingerprints of virtual library molecules based on plant-based clustering (Distance/Similarity Measure = Euclidean Distance, Cluster Method = Nearest Neighbor) (Please refer to Supplementary Table S2.7.9 for Distance matrix (Euclidean distance))	81
Figure 2.8: The DoMINE cheminformatics toolkit. A. DoMINE showing the medicinal plant species <i>Abrus precatorius</i> with its therapeutic properties. B. DoMINE showing virtual molecules built from Indian medicinal plant molecules with TPC scores.....	87
Figure 3.1: An overview of the analytical steps deployed in the present study.....	108
Figure 3.2: Workflow highlighting UHPLC-MS/MS data analysis grouped into three categories: preprocessing, metabolite identification and statistical analysis	114
Figure 3.3: Histograms depicting descriptor ranges of soybean small molecules and approved drug	121
Figure 3.4: Cloud plot generated by XCMS for positive ion mode.....	123
Figure 3.5: Cloud plot generated by XCMS for negative ion mode	123
Figure 3.6: Venn diagram showing the differences between soybean seed and leaf annotated mass features of small molecules identified using two methods: ProbMetab (for probability score: 1) and PUTMEDID LC-MS in Taverna workflow (up to 2 ppm error)	134
Figure 3.7: Statistical analysis by ANOVA for +ve and -ve polar molecules. PCA plots (A.) +ve (B.) -ve; HCA (C.) +ve (D.) -ve; PLS-DA loadings for top 15 important features of differentially co accumulated metabolites in soybean seed and leaf (E.) +ve (F.) -ve. L1: Leaf sample of variety NRC119, L2: Leaf sample of variety JS335, L3: Leaf sample of variety JS7105, L4: Leaf sample of variety JS9305, S1: Seed sample of variety NRC119, S2: Seed sample of variety JS335, S3: Seed sample of variety JS7105, and S4: Seed sample of variety JS9305.....	137

Figure 3.8: Metabolic pathway network with the list of pathway names and the number of molecules involved in it for four varieties of soybean retrieved from KEGG soybean pathways	142
Figure 3.9: Metabolic accumulation in soybean varieties according to KEGG pathways. The heat maps were drawn using the R package ggplot2, and the green-red color represents the transformed raw data of soybean metabolites with significant differences among four sample varieties. Green and red colors indicate an increase and a decrease in metabolite levels, respectively. Categories represent the type of metabolic pathways. Category 1: Flavonoids, Category 2: Terpenoids and Category 3: Others.....	144
Figure 3.10: Soybean small molecules, drug molecules, and scaffold merged network as depicted in an organic layout in Cytoscape. Nodes = 10670 edges = 11482 (Nodes: Molecules; Edges: Interactions/ hidden relationships)	151
Figure 4.1: Schematic view of biosynthesis of secondary metabolites for plant defense	176
Figure 4.2: Examples of 2D structures of alkaloids.....	178
Figure 4.3: Examples of 2D structures of phenolic compounds.....	179
Figure 4.4: Examples of 2D structures of terpenoids	180
Figure 4.5: An overview of the steps deployed in the present methodology.....	182
Figure 4.6: Allelopathic interactions between a plant and the pathogen, insect, animals, and others (UV, weeds, herbicide, mechanical damage, etc.).....	184
Figure 4.7: Scaffold molecule network between the allelochemicals and pesticides generated using Cytoscape (Nodes- Molecules, scaffolds: 2306, Edges - Interactions/ hidden relationships: 2350).....	196
Figure 4.8: The 2D PCA plot representing the molecular diversity of bio-pesticides (n=39) in chemical space of allelochemicals	201
Figure 4.9: Comparison of TPC proportion model graph between allelochemical specific virtual library molecules and pesticide molecules.....	203

List of Tables

Table 1.1: Rule-based filters for drugs and pesticides.....	11
Table 1.2: Examples of 2D chemical structures of bioactive compounds with their phytochemical classification and bioactivities.....	14
Table 1.3: Bioactive compounds identified from food crops and medicinal plants with their protein targets involved in respective diseases.....	23
Table 1.4: Examples of allelochemicals induced in plants resistant to pathogens, animals, and insects	28
Table 2.1: Comparison of DoMINE with other databases on Indian medicinal plants and their metabolites for drug development	47
Table 2.2: List of medicinal plant databases.....	51
Table 2.3: A representative list of Indian medicinal and aromatic plants with their therapeutic categories and PubMed counts (as of March 2021) (* For more therapeutic categories, please refer to Supplementary Table S2.1).....	54
Table 2.4: Representative scaffolds from 104 aromatic and medicinal plant molecules representing 16 medicinal properties (* For more scaffolds, please refer to Supplementary Table S2.1).....	64
Table 2.5: Similar scaffolds identified from 104 Indian medicinal and aromatic plants molecules and drug molecules in supra network with their therapeutic category information (n = 8) (Sc: Scaffold)	72
Table 2.6: Virtual Library (VL) with PDL, PLL and P, T, C, DLF, and LLF scores (for selected n = 6 molecules of each cluster from Plant-based clustering method) ...	78
Table 3.1: Morpho-physiological characteristics of the four soybean varieties used for the study.....	110
Table 3.2: Small organic molecules (n=20) validated through tandem mass spectrometry by performing in silico fragmentation approach (CFM-ID) using putatively annotated and identified molecules in UHPLC-MS experiments having up to 2 ppm error and highest probability score i.e., 1 for soybean samples in positive and negative polar modes	125

Table 3.3: Common scaffolds identified in soybean small molecules, drug molecules, and scaffold merged network with information about their therapeutic categories (n=10) (Sc: Scaffold)	152
Table 3.4: Virtual library novel molecules with their molecular weight, TPC and PDL, PLL scores (n= 10) [Notes: PDL: Progressive Drug-Like, PDL: Progressive Lead-Like, T: Toxicophore, P: Pharmacophore, C: Chemophor.].....	159
Table 4.1: Examples of allelochemicals induced in plants and imparting resistance to pathogens, animals, and insects (For more examples, please refer to Supplementary Table S1.1).....	187
Table 4.2: Allelochemicals induced in plants in response to attack by pathogens, animals, and insects, with their phytochemical class and structures (n=10) (For more examples, please refer to Supplementary Table S4.1.3).....	192
Table 4.3: Similar scaffolds (n=5) identified from 280 allelochemicals and 1985 unique pesticides by scaffold molecule network	197
Table 4.4: Organic allelochemicals already in the market as biopesticides (n= 5; selected molecules) (For more examples, please refer Supplementary Table S4.1.7)	199
Table 4.5: Examples of virtual library novel pesticide-like molecules with their TPC scores and LC50 and Chronic values for fish and daphnids	205

Abbreviations

AHEAD	Asian Health, Environmental and Allied Databases
ANOVA	Analysis of variance
CFM-ID	Competitive Fragmentation Modeling for Metabolite Identification
DIMBOA	2, 4-dihydroxy-7-methoxy-1, 4-benzoxazin-3-one
DLF	Drug-like failure
DoMINE	Database of Medicinally Important Natural products from plantaE
EC-50	Half maximal effective concentration
ECOSAR	Ecological Structure Activity Relationships
EIC	Extracted ion chromatogram
ESI(-)	Electrospray ionization negative polar mode
ESI(+)	Electrospray ionization positive polar mode
FDA	Food and Drug Administration
GC-MS	Gas Chromatography coupled with Mass Spectrometry
GPCRs	G protein-coupled receptors
HCA	Hierarchical cluster analysis
HESI	Heated electrospray ionization source
HMDB	Human metabolome database
HPLC-MS	High Performance Liquid Chromatography coupled with Mass Spectrometry
IMPD	Indian Medicinal Plants Database
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC-50	Half maximal lethal concentration
LC-MS	Liquid Chromatography coupled with Mass Spectrometry
LLF	Lead-like failure
logP o/w	Log of octanol/ water partition coefficient
logS	Log solubility in water
MOE	Molecular Operating Environment
NCBI	National Center for Biotechnology Information
NMR	Nuclear Magnetic Resonance
PAN	Pesticide Action Network Database
PCA	Principal component analysis

PDL	Progressive drug-like
PLL	Progressive lead-like
PLS-DA	Partial least squares discriminant analysis
PPDB	Pesticide Properties Database
QSAR	Quantitative Structure Activity Relationship
SAR	Structure Activity Relationship
SDF	Structure data file
SMILES	Simplified Molecular Input Line Entry System
SoyChemDB	Soybean Chemoinformatics DataBase
SoyKB	Soybean Knowledge Base
TIM	Traditional Indian Medicine
TPC	Toxicophoric, Pharmacophoric and Chemophoric
UHPLC-MS	Ultra High-Performance Liquid Chromatography coupled with Mass Spectrometry
VIP	Variable's Importance in Projection

Synopsis Report

Introduction

Secondary metabolites produced by plants are valuable for their essential roles in food, medicine, and agrochemicals. Several of them have positive effects on health, such as reducing the risks of many chronic diseases like cardiovascular diseases, diabetes and cancer. They also show a wide range of pest control activities and have long been used to produce pesticides. However, the information about the classification of secondary metabolites and their known protein targets in human diseases is scattered in many publications. Furthermore, large data of published findings is available, which needs to be analyzed and managed for the best use and efficacy of the available products. Chemoinformatics techniques can use the vast chemical and bioactivity experimental data of various compounds and convert it into valuable knowledge for drug or lead design. Chemoinformatics tools and other *in-silico* drug designing software play an important role in designing novel drugs with no or fewer side effects and other drug interactions. We performed chemoinformatic analysis for ligand-based drug designing from secondary metabolites of plants. The analysis revealed that several bioactive compounds could serve as scaffolds for developing novel drugs, which can be analyzed further by experimental methods. Lipinski's "Rule of Five" approach was quickly adopted in the field of agrochemical discovery and led to the establishment of rules for pesticide-likeness. The challenges identified in this study will serve as a useful reference for future intensive research in drug and pesticide discovery.

Statement of the problem

1. To develop a computational protocol and a toolkit for generating novel potential drug candidates from bioactive molecules of Indian medicinal and aromatic plants through a chemoinformatics approach
2. To design drug-like and lead-like molecules based on chemoinformatics and UHPLC-MS/MS analysis of secondary metabolites of soybean
3. To design plant-defense specific novel molecules with pesticidal properties.

Methodology

Secondary metabolites obtained from food crops and medicinal plants having specific activities against human diseases were used to build a focused virtual library of novel molecules by extracting scaffolds and functional groups. For this purpose, we text-mined the literature related to Indian medicinal plant species and food crops like soybean for identifying chemical names of the molecules associated with each plant species. Chemical names of the extracted plant molecules were converted into SMILES (Simplified Molecular Input Line Entry System) strings and screened for 5-6 membered rings containing molecules up to 1000 molecular weight. We extracted molecular scaffolds from these molecular structures and used diverse scaffolds to build a focused virtual library. A representative virtual library of novel molecules was generated and prioritized further by virtual screening methods. The novel molecules were prioritized by progressive drug-like (PDL), progressive lead-like (PLL), drug-like failure (DLF), lead-like failure (LLF), and Toxicophoric, Pharmacophoric, and Chemophoric (TPC) scores. Molecules having good scores can be used for further analysis through molecular docking and molecular dynamics techniques etc. The same computational protocols were followed for designing pesticide-like molecules from secondary metabolites involved in plant defense, also known as allelochemicals.

In the future, the results of this project can lead to the development of efficient and more target specific drugs and pesticides from readily available plant sources in minimum time.

Results

The results of the present work are divided into three sections based on the objectives are:-

1. Design of Novel Drug-like Molecules using Informatics Rich Secondary Metabolites Analysis of Indian Medicinal and Aromatic Plants

Several medicinal plants are being used in Indian medicine systems from ancient times. However, in most cases, the specific molecules or the active ingredients responsible for the medicinal or therapeutic properties are not yet known. The objective of this study was to develop a computational protocol as well as a tool for generating novel potential drug candidates from the bioactive molecules of Indian medicinal and aromatic plants through the chemoinformatics approach. We employed chemoinformatics approaches to *in-silico* screened metabolites from 104 Indian medicinal and aromatic plants and designed novel drug-like bioactive molecules. For this purpose, 1665 ring-containing molecules were identified by text mining of literature related to the medicinal plant species, which were later used to extract 209 molecular scaffolds for building a focused virtual library. Virtual screening was performed with cluster analysis to predict drug-like and lead-like molecules from these plant molecules in the context of drug discovery.

The predicted drug-like and lead-like molecules were evaluated using chemoinformatics approaches and statistical parameters, and only the most significant molecules were proposed as the candidate molecules to develop new drugs. A supra

network of molecules and scaffolds identifying the relationships between the plant molecules and drugs was developed. Cluster analysis of virtual library molecules showed that the novel molecules had more pharmacophoric properties than toxicophoric and chemophoric properties. These predicted molecules need to be subjected to biological screening to identify potential molecules for drug discovery research. We also developed a Java-based open-source toolkit-cum-database called DoMINE (Database of Medicinally Important Natural products from plantaE) to advance the natural product-based drug discovery through chemoinformatics approaches. This study will be useful in developing new drug molecules from the known medicinal plant molecules. We hope that this work will encourage experimental organic chemists to synthesize these molecules based on the predicted values.

2. Bridging *In-Silico* and Experimental: Chemoinformatics Investigation for Mass Spectrometry-Based Metabolomics Study of Soybean

Soybean (*Glycine max* L. Merr.) is a globally important legume crop and contains various small organic molecules that are valuable sources for drug development. This study intended to identify, analyze and design a virtual library of prioritized novel and promising drug-like molecules based on the analysis of secondary metabolites of soybean using chemoinformatics and untargeted mass spectrometry (UHPLC-MS/MS) approaches. In this study, we performed chemoinformatics analysis of previously reported and unreported secondary metabolites from four soybean varieties. The secondary metabolites were identified by UHPLC-MS/MS analysis and text mining, and a virtual library of novel molecules was generated. The metabolomics data were analyzed using machine learning-based quantitative and qualitative methods for identifying putative metabolites by spectral matching and

multivariate statistical analysis. A representative virtual library of novel molecules was generated and prioritized further by virtual screening methods. We detected 6628 annotated mass features for small molecules that have not been reported in soybean before, in addition to 443 mass features of molecules that were previously reported in the literature. Tandem mass spectrometry (MS/MS) confirmed the presence of 14 new and six previously reported soybean molecules. We found high molecular diversity in seed and leaf tissues of four soybean varieties (NRC-119, JS-335, JS-7105, and JS-9305). We identified 25 common scaffolds and 231 molecules through scaffold-molecule networks between soybean molecules and known drugs. Five representative scaffolds were used to build a focused virtual library of novel molecules (n= 1225), which were virtually screened to obtain potential drug-like candidates (n= 815) for further studies. We developed a novel virtual library of molecules with drug-like and lead-like properties for further drug discovery-related studies. This study suggests that a combinatorial approach employing high-throughput metabolomics and chemoinformatics methods can efficiently identify new drug-like and lead-like candidates from plant metabolites.

3. Chemoinformatics Investigation on Chemical Defense in Plants

Chemical defense against predation has been studied for a long time. Plants produce many secondary metabolites called allelochemicals to protect themselves against herbivores, pests and pathogens. In this study, we performed chemoinformatics investigations to build combinatorial libraries of allelochemicals that were then quantitatively evaluated for their pesticide properties. We identified five common scaffolds and 15 common molecules through scaffold-molecule networks between allelochemicals and pesticides. Scaffolds (74) were extracted from allelochemicals used for building a focused virtual library of novel molecules (380). We propose new

virtual molecules with pesticide-likeness properties according to rules published by Hao *et al.* (2011) (Hao, Dong and Yang 2011) for further agrochemical studies. Their LC-50 and EC-50 values for daphnid, mysids, algae, and fishes like Fathead minnow, etc., were predicted by ECOSAR- QSAR methods. These values indicate their lethal concentration for lower aquatic organisms, which will be used for their pesticide activities. This study shows that a combinatorial approach employing QSAR studies provides a novel perspective to the future directions for pesticides of natural origin.

Summary and future directions

Considering the demand for organic production of food and drugs, novel and innovative approaches are required. Natural products derived from plants are emerging as valuable alternatives for human needs and rescuing from the bio-apocalypse. In the present study, we report a novel computational protocol and a tool to access the database with structural information, plant information, and traditional therapeutic use and generate scaffolds to perform *in-silico* based combinatorial synthesis of the virtual library from Indian medicinal plant molecules. This is a simple, fast, and cost-effective computational protocol. Soybean plant was chosen as a case study where we successfully bridged the gap between chemoinformatics and experimental mass spectrometric approaches to identify and screen drug-like and lead-like compounds. We also demonstrated pesticide-like activities of allelochemicals and virtual novel molecules designed from them with pesticide-likeness and ECOSAR- QSAR methods. The outcomes from this study will serve as the foundation for the development of drugs and pesticide molecules by the inclusion of novel methods of rational drug designing and machine learning in screening molecules against multiple targets.

CHAPTER 1

INTRODUCTION AND REVIEW OF LITERATURE

Chapter 1: Introduction and Review of Literature

1.1 A brief overview of chemoinformatics and metabolomics

Metabolomics is an emerging and valuable technology that involves a comprehensive analysis of metabolites and their interactions in biological systems (Clish 2015). The study of human diseases is one of the most important applications of metabolomics. Plant secondary metabolites have been shown to possess various biological effects, such as disease prevention and treatment (Fakhri et al. 2020, Leicach and Chludil 2014). Secondary metabolites from plants provide lead molecules for the development of drugs (Verpoorte 1998). Besides, plant secondary metabolites have been earlier used in agriculture to protect crops from pests and are currently contributing to integrated pest management (Bennett and Wallsgrove 1994). Metabolomics relies on analytical chemistry techniques and technology platforms such as mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy, which help to identify and isolate a range of metabolites (Ren et al. 2018). The implementation of data mining techniques and chemoinformatics approaches facilitate the easy use of metabolomic platforms. Cheminformatics is a field of information science that focuses on storing, indexing, analyzing, and applying information about chemical compounds (Wishart 2007). This chapter discusses the role of plant secondary metabolites in drug development, plant defense, and ultimately how metabolic data enables generation of novel molecules through the chemoinformatics approach.

1.2 Technological trends in plant metabolomics

Natural plant products can serve as important, economical, and viable sources to develop drugs. The metabolites produced by plants, also called phytochemicals, can have pharmacological effects on humans and animals (Hussein and El-Anssary 2018).

Many of these compounds are produced in plants via secondary metabolism and are not essential for their survival. However, these compounds often play important role in plant defense against herbivory and interaction with other species. Many of these compounds are also used as medicines, flavorings, etc. Secondary metabolites vary in their chemical structures and functions and are grouped into classes such as flavonoids, terpenoids, alkaloids, etc. (Kris-Etherton et al. 2002). They are formed as by-products or intermediates of primary metabolism (**Figure 1.1**) and are usually produced in small amounts. Several of them have positive effects on health, such as reducing the risks of many chronic diseases such as cardiovascular diseases, diabetes, and cancer (Bahmani et al. 2014, Shin et al. 2018). However, the information about the secondary metabolites or bioactive compounds and their known protein targets in human diseases is scattered in many publications. Chemoinformatics techniques can use the vast chemical and bioactivity experimental data of various compounds and convert it into the knowledge useful for drug or lead design.

Chemoinformatics tools and other *in-silico* drug designing software can help in efficiently designing novel drugs with no or fewer side effects and other drug interactions. They can also identify drug targets and predict novel drugs (Wadood et al. 2013). Several tools and software for drug designing and visualization are available, which can be employed to develop drugs based on natural products. The cost of inventing a new drug through conventional approaches is increasing day by day. However, using chemoinformatics tools makes it possible to quickly develop a new drug at a much-reduced cost (Xu and Hagler 2002, Martinez-Mayorga et al. 2020). Hence, pharmaceutical industries are increasingly employing chemoinformatics tools to analyze the vast amount of experimental data available in the public domain and repurposing the already known drugs.

molecules having similar pharmacological effects. QSAR / 3D-QSAR models can be used to screen a chemical library for potential drug leads (Kubinyi 1993).

The availability of a curated library of information about plants, their related natural products, and a repository of their chemical structures can aid in the identification of new drugs. In this regard, significant progress has been made in the development of natural product databases such as Nutrichem (Jensen et al. 2015), Phytochemica (Pathania, Ramakrishnan and Bagler 2015), TCM-Mesh (Zhang et al. 2017) and COCONUT online (Sorokina et al. 2021), which can aid in the virtual screening of prospective drug compounds or the investigation of plant-disease associations. However, in terms of traditional Indian medicine, there have been limited initiatives to create online databases that cover Indian medicinal plants, phytochemicals, and therapeutic applications. In 2011, an *in silico* library of natural products from Ayurvedic medicines was developed with structural information, plant origin, and traditional therapeutic uses of the natural products (Polur et al. 2011). In this study also, the chemical structures of compounds identified from Traditional Indian Medicine (TIM) were compared with drugs from DrugBank and a structural similarity network was constructed. This was achieved by matching the traditional medicinal uses of the plants with the medicinal use of the drugs that are structurally similar to the plant components. In this way, novel natural leads were identified from medicinal plants used to prepare Ayurvedic medicines. The Phytochemica (Pathania et al. 2015) database gathered information on five Indian medicinal plants and their 963 phytochemicals and gave chemical structures and pharmacological effects of the phytochemicals inside its database. The IMPPAT (Mohanraj et al. 2018), a curated database, comprises data on 1742 Indian medicinal plants, 9596 phytochemicals, and

1124 therapeutic uses, with 27074 plant-phytochemical correlations and 11514 plant-therapeutic associations.

1.2.1 Experimental approach

New techniques associated with genomics, transcriptomics, proteomics and metabolomics have been used to depict the pharmacological mechanisms of Ayurvedic medicines. For metabolomics studies, sophisticated analytical spectroscopic and chromatographic techniques coupled with mass spectrometry have been applied as it is difficult to identify each of the metabolites in biological samples (Teo et al. 2011). Gas Chromatography coupled with Mass Spectrometry (GC-MS) and Nuclear Magnetic Resonance (NMR) is used for primary metabolite profiling in plants (Bhalla, Narasimhan and Swarup 2005, Kumar 2015) (**Figure 1.2**). NMR is currently the most powerful tool available for organic structure determination (Mannina, Sobolev and Capitani 2012). High-Performance Liquid Chromatography (HPLC) is a type of liquid chromatographic technique that can identify, quantify and purify the individual components of a mixture of compounds.

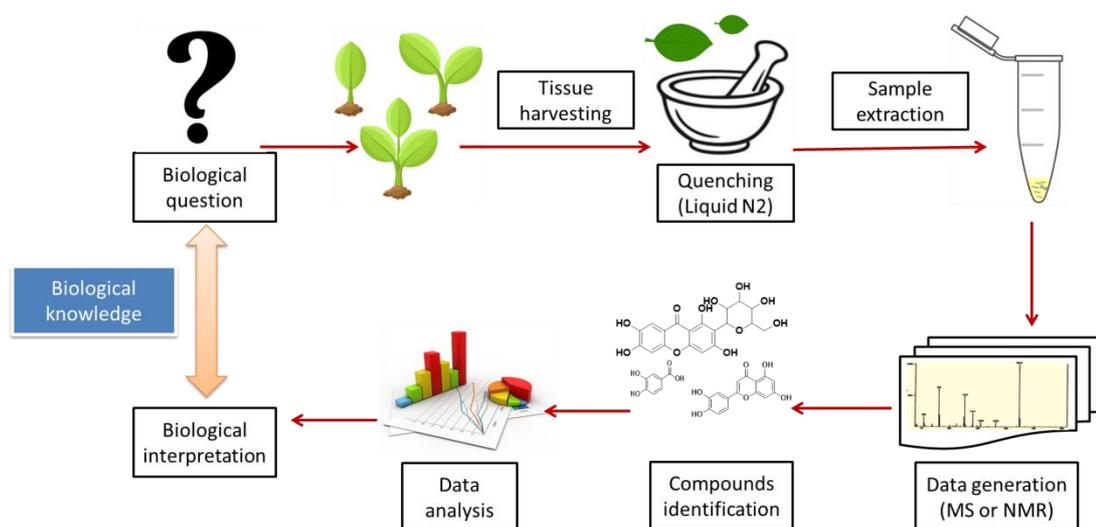


Figure 1.2: Plant metabolomics workflow

High-Performance Liquid Chromatography coupled with Mass Spectrometry (HPLC-MS) and NMR are usually used for secondary metabolite profiling (Yang et al. 2012). Gas or liquid chromatography is used to separate molecules, whereas Mass Spectrometry is used to identify small molecules. In this way, when both the technologies are combined, a powerful analytical tool is developed (Marney et al. 2014). The speed, efficiency, sensitivity, and ease of operation of HPLC are superior to Liquid Chromatography (LC) (Manayi, Vazirian and Saeidnia 2015, Zare et al. 2014). In a previous study, NMR and GC/LC-MS were used to investigate the responses of *Arabidopsis thaliana* to various environmental stressors such as heat, freezing, drought and salinity, etc. (Tian, Lam and Shui 2016). Such experimental studies help acquire knowledge on the metabolic profiles of the specific plants and genetic and biochemical mechanisms related to plant growth, development and stress responses, etc.

1.2.2 Computational approach

Computation is an easy and efficient method to solve various problems in computational chemistry and biology. Traditional methods of screening plants and extracting compounds following bioassays are time-consuming and tedious processes. Through computational approaches, the nature of chemicals can be predicted based on chemical structures. These chemical data are then linked to pharmacological profiles and system biology functional data by constructing networks (Tao et al. 2013, Zhao, Jiang and Zhang 2010). The selected bioactive compounds can be further processed with QSAR and docking computational methods to evaluate the potential drug-like or lead-like molecules (Aguiar-Pulido et al. 2013). Applying docking algorithms to such models can help predict the structures of small and large target protein molecules in a simulated system of verifiable thermodynamic properties. The analysis and

interpretation of such models can help the process of drug designing (Meetei et al. 2016).

A large amount of experimental metabolomics data is available in the literature, and various databases such as the human metabolome database, BioMagResBank (BMRB- metabolomics), BiGG (database of biochemical, genetic, and genomic metabolic network reconstructions), Fiehn metabolome database, Golm metabolome database, etc., which need to be analyzed and interpreted. With the help of chemoinformatics, this information can be processed in a short time to contribute to drug design and development (Lusher et al. 2011). In chemoinformatics, lead identification and optimization play an important role in the discovery of new drugs from existing compounds (Jorgensen 2009). Lead optimization aims to enhance the most promising compounds to improve efficiency, reduce toxicity or increase absorption. Many lead discovery technologies overlap with lead optimization as researchers attempt to incorporate the best drug characteristics early in the process. New lead-like and drug-like molecules can be generated by structural analogy with existing drugs. Through chemoinformatics, it is easy to search and identify molecules having toxicophore, pharmacophore and chemophore (TPC) properties (Karthikeyan and Vyas 2015) (**Figure 1.3**).

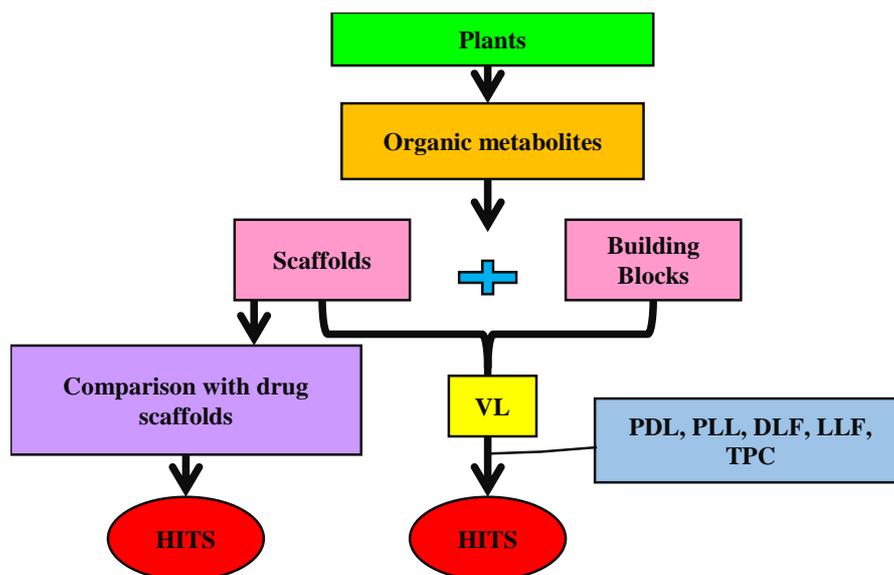


Figure 1.3: Cheminformatics methods in designing of novel molecules from organic metabolites of plants (VL- Virtual Library, PDL- Progressive Drug Like score, PLL- Progressive Lead Like score, DLF- Drug Like Failure, LLF- Lead Like Failure, TPC – Toxicophoric, Pharmacophoric and Chemophoric scores)

In medicinal chemistry, the molecules identified as ‘Pharmacophore’ possess molecular substructure responsible for pharmacological interactions (Karthikeyan and Vyas 2014). ‘Toxicophore’ refers to the substructural groups, which are toxic, such as azides, diazo structures, triazenes, aromatic azo moieties, aromatic hydroxylamines, aliphatic halides, etc. and which cannot be used as potential drugs. Whereas ‘Chemophores’ are those substructural groups that are either too reactive, inert or synthetically inaccessible. Similarly, molecules with drug-like properties were scored with progressive drug-likeness (PDL) and progressive lead-likeness (PLL). Whereas, the molecules that fail to possess drug-like properties were scored with drug-like failure (DLF) and lead-like failure (LLF).

With the molecules prioritized by virtual screening, scaffolds are generated or extracted based on which the virtual library can be synthesized after attachment of linker atoms and functional groups (Polur et al. 2011). Scaffolds are the basic core of the molecule without functional groups attached to it (Karthikeyan and Vyas 2014). The structure of a scaffold might be the same for more than one molecule. These scaffolds are used to enumerate a virtual library by supplying linkers and functional groups. A virtual library is an important tool in drug discovery. A virtual library is a combinatorial library of chemical compounds generated from multiple combinations of functional groups with scaffolds (Van Drie and Lajiness 1998). These libraries are analyzed and screened by virtual screening methods. Virtual screening is a computational technique used in drug discovery to search libraries of small molecules to identify those structures, which are drug-like or lead-like (Lionta et al. 2014, Sheppard and MacRitchie 2013). Such computational techniques are important for lead-specific drug designing.

QSAR based virtual screening methods are one of the computational methods that help find potential leads with different scaffolds from a chemical library (Neves et al. 2018). QSAR or Quantitative Structure-Activity Relationships is a statistical method that complements molecular modeling (Aguilar-Pulido et al. 2013, Roy and Mitra 2011). In QSAR analysis, the biological activity of a set of molecules is measured using statistical methods based on their structures. QSAR models first summarize the supposed relationship between chemical structures and biological activities in a virtual library of compounds. Based on the analysis, QSAR models can then predict the activities of new compounds. The molecular descriptions in the QSAR study are expressed in numerical values known as molecular descriptors. Molecular descriptors are divided into two classes, *viz.* 2D and 3D descriptors (Garro

Martinez et al. 2015, Roy and Das 2014). 2D descriptors use atoms and connection information of molecules for calculation like molecular weight, Log of octanol/ water partition coefficient ($\log P_{o/w}$), etc. Whereas, 3D descriptors are based on 3D coordinate information of each molecule like the radius of gyration (rgyr), Van der Waals surface area (VSA), etc.

In computational chemistry, small molecule modeling, protein modeling, and QSAR are interrelated and synergistically help in the progress of drug discovery and designing. By integrating the results of various methods for calculating drug-likeness for a particular set of molecules, determination and screening of drug-like and lead-likeness can be performed. Such selection criteria are based on the Lipinski's "rule of five" (Benkendorff 2013, Zhong et al. 2013), derived from a comprehensive study of orally active drugs. These rules state that hydrogen bond donors should be less than five, hydrogen bond acceptors should be more than ten, the molecular mass should be less than 500 Daltons, and the octanol-water partition coefficient i.e., $\log P$ should not be greater than five.

The Lipinski's "rule of five" approach was quickly adopted in the field of agrochemical discovery and led to the establishment of rules for pesticide-likeness (Hao et al. 2011) (**Table 1.1**). The authors described easy-to-implement and straightforward rules for pesticide-likeness by including molecular weight (MW), lipophilicity (expressed as $\log P$), number of H-bond acceptors (HBA), and donors (HBD), number of rotatable bonds (RB), and number of aromatic bounds. In contrast, the traditional method of discovering pesticides is based on the synthesis of large numbers of compounds and mass screening which is an expensive and time-consuming method (Das 2016). We believe that a similar treatment of pesticide-likeness will support the agrochemical discovery sector.

Table 1.1: Rule-based filters for drugs and pesticides

Molecular descriptors	Lipinski's Rule (Drugs)	Hao's Rule (Pesticides)
MW (Da)	≤ 500	≤ 435
logP	≤ 5	≤ 6
HBD	≤ 5	≤ 2
HBA	≤ 10	≤ 6
RB	≤ 3 (for lead-likeness)	≤ 9
Aromatic bonds	-	≤ 17

This chapter reviews the classification of phytochemicals, their occurrences in plants, biological activities in human health regarding nutrition, and their roles in drug and pesticide discovery. The benefits of integrating the knowledge of medicinal plants and food crops in chemoinformatics-driven drug and pesticide discovery have been discussed. This can be done by predicting the biological activity of novel natural compounds from these plants following the structural similarity principle. Through this analysis, we found that various compounds have novel drug-like and lead-like properties, and at least some of them can be used to develop new drugs. In addition to identifying novel bioactive molecules, the chemoinformatics-based natural product protein target network that we developed could also be helpful in drug repurposing.

1.3 Classification of phytochemicals in plants

Phytochemicals are classified into four major groups based on the major chemical constituent as phenolics, phytosterols, phytate, and nitrogenous compounds.

1.3.1 Phenolic compounds

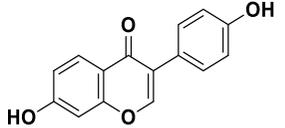
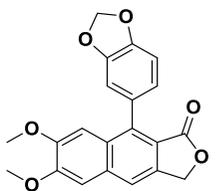
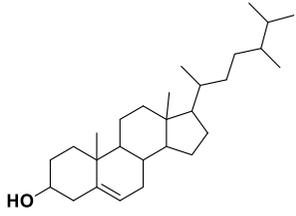
a. Polyphenols

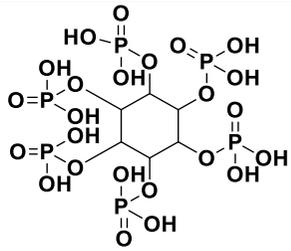
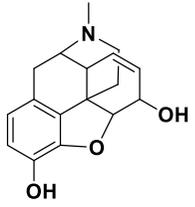
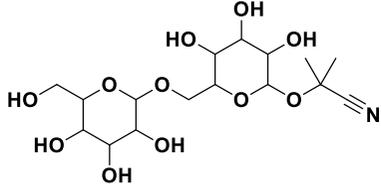
Polyphenols are the biggest group of phytochemicals with phenolic structural features present in virtually all food plants. More than 8000 polyphenols and 4000 flavonoids have been identified (Patel et al. 2018, Mutha, Tatiya and Surana 2021). Polyphenols are characterized based on their chemical structures of the aglycones such as (a) phenolic acids - benzoic acid and cinnamic acid derivatives based on C1–C6 and C3–C6 backbones, (b) flavonoids - isoflavones, flavonols, etc. (c) anthocyanidins - cyanidin, delphinidin and pelargonidin, etc. (d) polyphenolic amides - capsaicinoids, avenanthramides, etc. and (e) other polyphenols - resveratrol, ellagic acid, curcumin, etc. Plants produce polyphenols to protect themselves from other organisms. They also play an important role in maintaining human health due to their antioxidant and anti-inflammatory activities (Hussain and Tan 2016, Saric and Sivamani 2016). The ongoing research on polyphenols has opened up a promising field for drug development and treatment of various cancers. Extensive reviews have discussed the anticancer (Abdal Dayem et al. 2016, Niedzwiecki et al. 2016, Zhou et al. 2016, Amararathna, Johnston and Rupasinghe 2016, Moga et al. 2016), anticardiac (Hussain and Tan 2016) and antidiabetic (Coe and Ryan 2016) properties of polyphenols including their protective effects against neurodegenerative (Cirimi et al. 2016b, Caruana, Cauchi and Vassallo 2016) and neurodevelopmental disorders (Vacca et al. 2016). For example, daidzein (**Table 1.2, entry 1**) from soybean has anticancer activity against various types of cancers such as prostate, breast cancers, etc. (De Lemos 2001, Adjakly et al. 2013).

b. Phytoestrogen

Phytoestrogen is diphenolic nonsteroidal compound derived from plants or their seeds. They have estrogen-like properties due to similarity in chemical structure with estrogen (Yuan et al. 2014). Phytoestrogens are divided into three classes: isoflavones, coumestans, and lignans (Murkies, Wilcox and Davis 1998). Phytoestrogens exhibit physiological effects in humans and protect against diseases such as cardiovascular diseases (Ishimi 2015), various types of cancers (Lee, Hwang and Choi 2016a), menstrual problems (Rietjens, Louisse and Beekmann 2016, Sobenin, Myasoedova and Orekhov 2016, Dittfeld et al. 2015), osteoporosis (Ishimi 2015) and antimicrobial diseases for which many studies have been performed (Rishi 2002). Lignans such as Justicidin B (**Table 1.2, entry 2**) isolated from *Justicia*, *Phyllanthus*, *Haplophyllum* and *Linum* species act as antiprotozoal agents against *Trypanosoma brucei* for the treatment of tropical diseases (Hemmati and Seradj 2016).

Table 1.2: Examples of 2D chemical structures of bioactive compounds with their phytochemical classification and bioactivities

Sr. No.	Phytochemicals	Classification	Plant Source	2D structure	Bioactivity	Ref.
1	Daidzein	Polyphenols: flavonoids	Soybean		Anticancer	(Mahmoud, Yang and Bosland 2014)
2	Justicidin B	Phytoestrogen: Lignan	<i>Justicia</i> , <i>Phyllanthus</i> , <i>Haplophyllum</i> and <i>Linum</i> species		Antiplatelet, anti-inflammatory, antiprotozoal agent	(Hemmati and Seradj 2016)
3	Campesterol	Phytosterols	Vegetable oils, nuts, seeds, cereals		Anticardiac	(Genser et al. 2012)

Sr. No.	Phytochemicals	Classification	Plant Source	2D structure	Bioactivity	Ref.
4	Phytic acid	Phytate	Cereals, legumes, oilseeds and nuts		Antioxidant	(Graf and Eaton 1990, Graf, Empson and Eaton 1987)
5	Morphine	Nitrogenous compounds: alkaloid	Opium poppy		Analgesic	(Ghelardini, Di Cesare Mannelli and Bianchi 2015)
6	Linustatin	Nitrogenous compounds: cyanogenic glycosides	<i>Linum usitatissimum</i> or Flax		Chemical defense against predators	(Niedzwiedz-Siegień 1998)

1.3.2 Phytosterols

Phytosterols are the bioactive compounds present in all plants and foods in varying concentrations. They perform similar functions in plants as cholesterol in animals due to the similarity in structures, except the side chains in phytosterols containing additional double bonds and methyl and/or ethyl groups (Gylling and Simonen 2015). Phytosterols exist as free sterol, sterol esters, sterol glycosides, acylsterol glycosides, etc. These are not found in animals hence referred to as plant sterols. Beta-sitosterol, stigmasterol, campesterol (**Table 1.2, entry 3**) are the major sterols present in higher plants and foods. The enzymes involved in the synthesis of phytosterols are 3-hydroxy-3-methylglutaryl-CoA reductase, C24-sterol methyltransferase, and C22-sterol desaturase (Valitova, Sulkarnayeva and Minibayeva 2016). They are present in vegetable foods such as vegetable oils, nuts, seeds, cereals, etc. (Piironen and Lampi 2004, Klingberg et al. 2008). Primarily, phytosterols are known to reduce cholesterol absorption in the intestine and maintain the serum level of cholesterol. Phytosterol supplements are also prescribed in patients to reduce LDL- cholesterol (Ostlund Jr 2004). Hence, natural dietary phytosterols help reduce the risk of cardiovascular diseases (Racette et al. 2015). Phytosterols are also beneficial in reducing various forms of cancer such as breast, lung, prostate, stomach, etc. (Ramprasath and Awad 2015).

1.3.3 Phytates

Phytic acid (**Table 1.2, entry 4**) is also known as myoinositol 1,2,3,4,5,6-hexakis dihydrogen phosphate (Gupta, Gangoliya and Singh 2015). There are two kinds of phytates i.e., 3-phytase and 6-phytase, based on the first phosphate hydrolyzation. Phytate occurs in cereals (0.50%- 1.89%), legumes (0.40% - 2.06%), oilseeds (2.00%

– 5.20%) and nuts as the primary storage form of phosphorus (Martinez Dominguez, Ibanez Gomez and Rincon Leon 2002, Harland and Oberleas 1987, Reddy, Sathe and Salunkhe 1982). They have the potential to reduce mineral absorption as they are potent chelators of cations such as iron, zinc, magnesium, and calcium (Sparvoli and Cominelli 2015, Zhou and Erdman 1995). Hence, they are beneficial for preventing calcification and stone disease (Grases and Costa-Bauza 1999) and lowering blood glucose and lipids (Katayama 1999). In addition to anti-nutritional properties, phytate also has beneficial effects such as antioxidative and anti-carcinogenic effects (Nawrocka-Musial and Latocha 2012, Shamsuddin and Vucenik 1999, Shamsuddin, Vucenik and Cole 1997).

1.3.4 Nitrogenous compounds

a. Alkaloids

Alkaloids are heterocyclic nitrogenous compounds present in some plants to protect themselves from predators due to their potent toxic activity and bitter taste (Zenk and Juenger 2007). Alkaloids are produced in various forms in specific plants. For example, tropane alkaloids are present in *Solanaceae* (nightshade family) (Pigatto et al. 2015), such as *Atropa belladonna* (deadly nightshade), *Datura* spp. (thorn apples), etc., pyrrolizidine alkaloids in *Asteraceae* (daisy family) such as *Senecio* spp. (Ragworts) (Stegelmeier 2011) and in *Boraginaceae* (borage family) (Stegelmeier 2011), isoquinoline alkaloids in *Papaveraceae* (poppy family) (Opletal et al. 2014) and *Berberidaceae* (barberry family) (Alamzeb et al. 2015), methylxanthine alkaloids in *Coffea arabica* (coffee) (Ashihara 2006), *Theobroma cacao* (cocoa) (Sugimoto et al. 2014) and pseudoalkaloids produced by *Taxaceae* (yew family) (Hou et al. 2014). Alkaloids show gastroprotective and anti-ulcer activities (De Sousa Falcao et al.

2008) and are known to act as an anti-platelet agent used for treating various diseases such as malaria, diabetics, cancer, cardiac dysfunction, etc. (De Sousa Falcao et al. 2008). The morphine alkaloid (**Table 1.2, entry 5**) synthesized in opium poppy is used as a painkiller (Ghelardini et al. 2015). β -carboline alkaloids are known to reduce the growth of *Trypanosoma cruzi*, the organism responsible for Chagas' disease (Cavin, Krassner and Rodriguez 1987).

b. Cyanogenic glucosides/glucoisnolates

Cyanogenic glucosides and glucoisnolates are very effective compounds in the chemical defense of plants from grazing animals and other predators (Zagrobelyny and Møller 2011). Because after consuming the plant parts containing these compounds, poisonous HCN and aldehydes are released in the body, which can even cause the animal's death (Cavin et al. 1987). For example, cyanogenic glucosides present in *Prunus* spp. lead to poisoning in livestock and insects (Patton et al. 1997). However, some of the plants containing cyanogen compounds are used as staple foods as well as for medicinal treatment. For example, flax seeds containing cyanogenic compounds such as diglucoside linustatin (**Table 1.2, entry 6**) and neolinustatin are seen as causing no health hazard (Oomah, Mazza and Kenaschuk 1992, Stijve and De Meijer 1999). Whereas, cyanogenic glycosides extracted from *Prunus persica* seeds have been shown to produce antitumor promoting activities (Fukuda et al. 2003).

1.4 Role of secondary metabolites as bioactive compounds

1.4.1 Role of bioactive compounds from food crops

Bioactive compounds in food are present in small amounts and are considered as extra-nutritional constituents. These compounds have specific functions in our body to maintain health and prevent diseases (Abuajah, Ogbonna and Osuji 2015). For

example, lycopene present in tomato and its products help to reduce prostate cancer (Giovannucci et al. 2002), isoflavones such as daidzein, genistein, etc. present in soybean are beneficial in menstrual-related problems, cardiovascular diseases, cancers and help in lowering cholesterol (Brouns 2002). Fruits and vegetables have great potential in cancer prevention because of their phytochemical content. Citrus fruits contain flavonoids, which may act as anticancer drugs (Cirimi et al. 2016a). The food contains diverse kinds of bioactive compounds, which influence human homeostasis and are responsible for various diseases.

Phytochemicals have high therapeutic potential with a great diversity of chemical structures as depicted in **Table 1.2**. Chemoinformatics tools can help understand biological and chemical aspects of pharmacological actions of phytomedicines (Lawless et al. 2016). The scaffold and functional groups of bioactive molecules play a significant role in drug discovery by providing it a particular therapeutic property (Kumar et al. 2013, Sravanthi and Manju 2016). Scaffolds extracted from bioactive compounds of plants can be used for building a focused virtual library of molecules with drug-like and lead-like properties (Karthikeyan et al. 2015a). Likewise, various food crops with medicinal properties due to their phytochemical content can be used to further studies in drug discovery.

1.4.2 Role of bioactive compounds from medicinal plants

Indian medicinal plant species are used for treating various diseases as described in ancient literature of Ayurveda and Siddha (Patwardhan, Vaidya and Chorghade 2004). Herbal medicines are considered to be less toxic and possessing fewer side effects than chemically synthesized drugs. The prevalence of natural products derived medicinal properties is due to the evolution of bioactive compounds in medicinal plants. For example, ajmaline isolated from the roots of *Rauwolfia serpentina* comes

under FDA approved drug as an antiarrhythmic agent (Makarevich et al. 1979), Cannabidiol identified in *Cannabis sativa* used for the treatment of Dravet syndrome was approved for marketing under the brand name Epidiolex in the US (Saade and Joshi 2015). Colchicine is a major alkaloid present in *Colchicum* species. It is used to treat familial Mediterranean fever as an approved drug (Zemer et al. 1991).

Bioactive molecules identified from medicinally important plants used in Ayurvedic preparations provide a vast range of chemical structures (Mishra and Tiwari 2011). Their scaffolds compared with drug molecules resulted in the natural products-based drug discovery (Polur et al. 2011). The abundant scaffold diversity in medicinal plants is continuously used for purposeful drug designing in a biologically friendly way (Cragg and Newman 2013). The current status of natural product databases indicates that there is a need to exploit the knowledge of traditional therapeutics. Chemoinformatics and molecular approaches not only help to consolidate the experimental data but also help to make use of natural products easier than before (Harvey 2000).

1.4.3 Network analysis of bioactive compounds

To understand the biological and chemical aspects of pharmacological actions of phytochemicals, we performed a network analysis of phytochemicals, their respective protein targets, and pathways involved in various chronic diseases (**Figure 1.4**). For this purpose, we selected five food crops and five Indian medicinal plants mentioned in the Ayurveda. **Table 1.3** describes the functions of bioactive compounds identified from food crops and medicinal plants, having therapeutic potential by targeting various proteins. Some bioactive compounds are present among multiple plants (**Supplementary Tables S1.1, S1.2, S1.3**). For example, betulinic acid is present in

food crops such as soybean (*Glycine max*) and Jamun (*Syzygium cumini*), as well as in medicinal plants such as Chaulmoogra (*Hydnocarpus wightianus*).

For this study, protein targets for bioactive compounds were identified from the literature. The biological pathways of the protein targets involved in respective diseases were also retrieved from the literature. **Figure 1.4** shows the medicinal effects of multi-targeting bioactive molecules. Most of the bioactive compounds were found to target multiple proteins. For example, betulinic acid acts on diacylglycerol acyltransferase, nitric oxide synthase, 5-alpha reductase, glycogen phosphorylase, DNA polymerase-beta, NADPH oxidase, and LXR-alpha. Therefore, bioactive compounds acting on multiple targets would be beneficial to treat more than one disease using the same compound/drug. However, on the contrary, this can also cause side-effects. Hence, both these aspects must be carefully examined before using a bioactive compound having multiple targets. A single protein target might be involved in several biological pathways. For example, NADPH oxidase is involved in leukocyte-endothelial migration, osteoclast differentiation, ROS generation and oxidative stress pathways (Fayaz, Kumar and Rajanikant 2014). The network analysis shows that all the selected protein targets for phytochemicals are involved in diabetes-related pathways. Whereas, most of them are also involved in cancer-related pathways. Apart from this, many proteins are also involved in other chronic diseases such as Parkinson's disease, cardiovascular disease, hepatitis, etc.

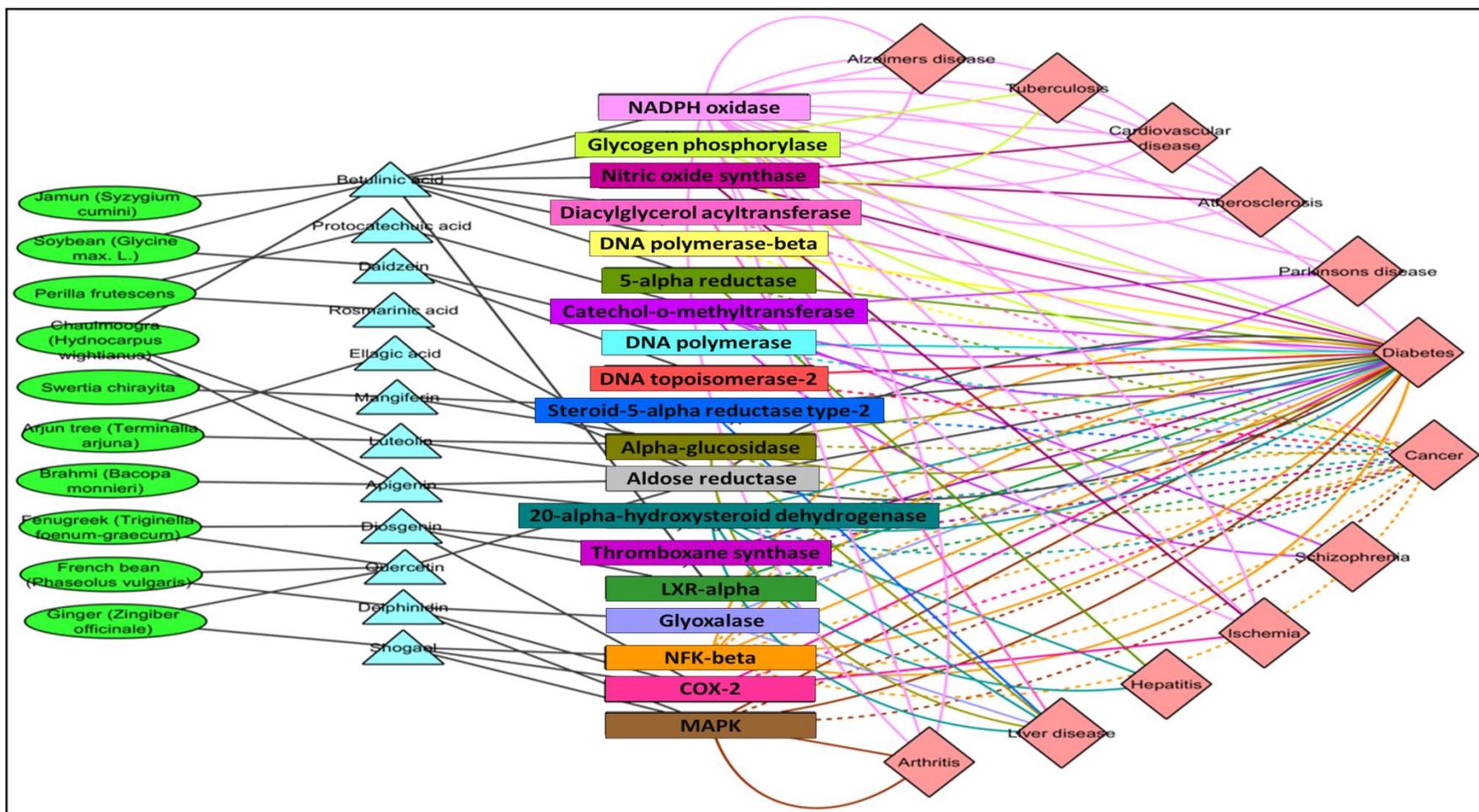
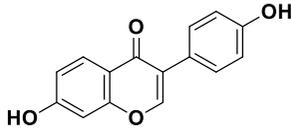
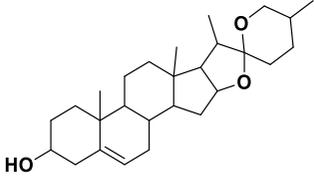
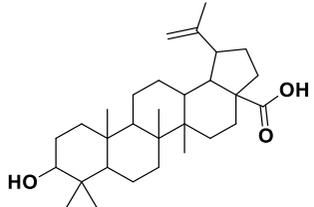
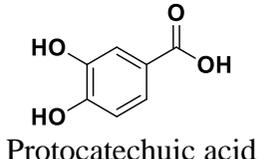
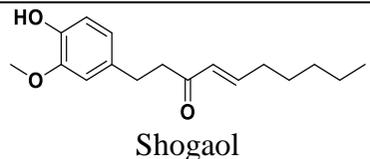
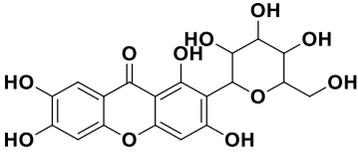
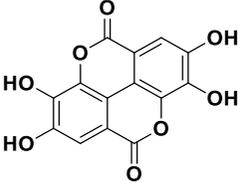
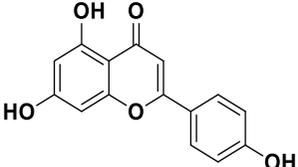


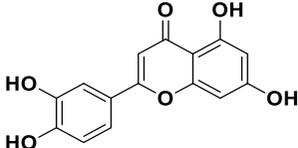
Figure 1.4: Network containing food crops and Ayurvedic medicinal plants with their respective bioactive compounds targeting various proteins involved in respective diseases. (Nodes = 53, edges = 138; Black edges: Interactions/ hidden relationships, Color edges: Respective pathways)

Table 1.3: Bioactive compounds identified from food crops and medicinal plants with their protein targets involved in respective diseases.

Sr. no.	Plant Source	Bioactive compounds	Ref.	Protein target	Pathway	Disease	Ref
Food crops							
1	Soybean (<i>Glycine max.</i> L.)	 Daidzein	(Kalaiselvan et al. 2010)	Glycogen phosphorylase, DNA topoisomerase-2	Glycogen degradation, Cell cycle, and DNA replication,	Diabetes, Cancer	(Matsumura et al. 2005)
2	Fenugreek (<i>Trigonella foenum-graecum</i>)	 Diosgenin	(Fuller and Stephens 2015)	Aldose reductase, LXR-alpha	Galactose metabolism, PPAR signaling pathway	Diabetes, Ischemia	(Makishima, Takahashi and Kawada 2010)
3	Jamun (<i>Syzygium cumini</i>)	 Betulinic acid	(Ramteke, Kurrey and Kar 2015)	Diacylglycerol acyltransferase, NADPH oxidase	Glycerolipid metabolism, Leukocyte endothelial migration	Diabetes, Liver disease	(Chung et al. 2006)

Sr. no.	Plant Source	Bioactive compounds	Ref.	Protein target	Pathway	Disease	Ref
4	French bean (<i>Phaseolus vulgaris</i>)	 Delphinidin	(Lin et al. 2008)	MAPK	Stat-3 signaling pathway, Aldosterone regulated sodium reabsorption Galactose metabolism,	Arthritis, Liver disease	(Oak et al. 2006)
5	<i>Perilla frutescens</i>	 Protocatechuic acid	(Speijers et al. 2010)	Aldose reductase, Catechol-o-methyltransferase	steroid hormone biosynthesis	Schizophrenia, Parkinson's disease	(Bonifácio et al. 2007, Woodard et al. 1980)
<u>Ayurvedic medicinal plants</u>							
6	Ginger (<i>Zingiber officinale</i>)	 Shogaol	(Chen et al. 2013)	MAPK, NFK-beta	Osteoclast differentiation, apoptosis	Cancer, Arthritis	(Kim et al. 2015)

Sr. no.	Plant Source	Bioactive compounds	Ref.	Protein target	Pathway	Disease	Ref
7	<i>Swertia chirayita</i>	 <p>Mangiferin</p>	(Mahendran et al. 2014)	Alpha-glucosidase, Steroid-5-alpha reductase type-2	N-Glycan biosynthesis, sucrose metabolic process	Diabetes, Liver disease	(Yoshikawa et al. 2001)
8	Arjun tree (<i>Terminalia arjuna</i>)	 <p>Ellagic acid</p>	(Kaur, Grover and Kumar 1997)	Aldose reductase, alpha- glucosidase	Fructose and mannose metabolism, sucrose metabolic process	Diabetes, Cancer	(Benalla, Bellahcen and Bnouham 2010)
9	Brahmi (<i>Bacopa monnieri</i>)	 <p>Apigenin</p>	(Umbelliferae)	Aldose reductase, 20-alpha-hydroxysteroid dehydrogenase	Galactose metabolism, Fructose and mannose metabolism	Hepatitis, Diabetes	(Qiang et al. 2012)

Sr. no.	Plant Source	Bioactive compounds	Ref.	Protein target	Pathway	Disease	Ref
10	Chaulmoogra (<i>Hydnocarpus wightianus</i>)	 <p>Luteolin</p>	(Sahoo et al. 2014)	5-alpha reductase, DNA polymerase-beta, NADPH oxidase	Steroid hormone biosynthesis, Purine metabolism, Osteoclast differentiation	Tuberculosis, Atherosclerosis	(Zainal et al. 2014)

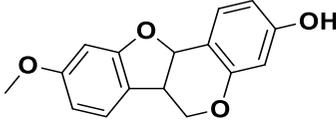
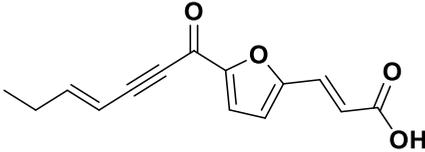
*(For more information, please refer to Supplementary Tables S1.1, S1.2, S1.3.)

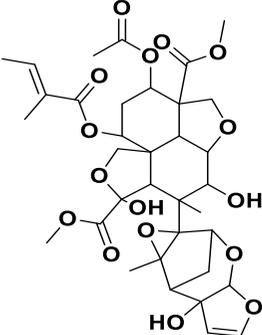
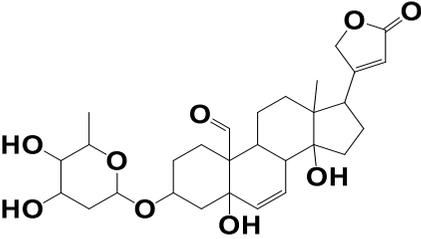
1.5 Role of secondary metabolites in plant defense

Plants are sessile organisms, and millions of insect species, microbes, animals, etc., depend on them for their survival. While in many cases, such relationships are mutually beneficial, in several cases, the plants get adversely affected. Thus, plants have evolved defense mechanisms to counter predators and parasites. However, these organisms, in turn, have developed molecular tools to overcome plant defense. During such co-evolution, plants have evolved a specific defense system against herbivores. Two kinds of defense mechanisms are present in plants. One is by structural defense mechanisms such as thorns, thick bark, etc., making it difficult for the pests to attack the plants. While, the second is by biochemical defense mechanisms in which plants produce toxic chemicals, pathogen degrading enzymes, etc.

Allelochemicals are the secondary metabolites produced by plants to defend themselves against herbivores, insects, pests etc. (Putnam 1988) (**Table 1.4**). Various allelochemicals show a wide range of pest control activities and have long been used in the production of pesticides (Koul and Walia 2009). For example, azadirachtin from neem tree is used in commercial pesticide named Azamax against various pests and insects (Nisbet 2000). Similarly, sanguinarine from *Chelidonium majus* possesses strong insecticidal activity against *Lymantria dispar* (moth) larvae (Zou et al. 2019). Likewise, the powder and extract from the dried flowers of the pyrethrum daisy, *Chrysanthemum cinerariaefolium* contain various pyrethrins and are used as insecticides to control weevils, beetles, grain borers, mealworms etc. (Gallo et al. 2017). These observations show that allelochemicals play an important role in pest control and, therefore, in increasing crop production.

Table 1.4: Examples of allelochemicals induced in plants resistant to pathogens, animals, and insects

Sr. no.	Plant Source	2D structure of Allelochemicals	Pathogens or Animals or Insects or Others	Refs
1	Alfa alfa (<i>Medicago sativa</i>)	 <p data-bbox="763 655 943 687">1. Medicarpin</p>	Fungus: <i>Phytophthora megasperma</i> , <i>Phoma medicaginis</i> , <i>Nectria haematococca</i> , <i>Colletotrichum trifolii</i>	(Blount, Dixon and Paiva 1992)
2	Arabidopsis (<i>Arabidopsis thaliana</i>)	 <p data-bbox="781 951 920 983">Camalexin</p>	Gram-negative bacteria: <i>Pseudomonas syringae</i> ; Fungus: <i>Alternaria brassicicola</i> , <i>Botrytis cinerea</i>	(Ahuja, Kissen and Bones 2012)
3	Broad bean (<i>Vicia faba</i>)	 <p data-bbox="792 1294 909 1326">Wyerone</p>	Fungus: <i>Botrytis cinerea</i> , <i>B. fabae</i> , <i>B. allii</i>	(Letcher et al. 1970)

Sr. no.	Plant Source	2D structure of Allelochemicals	Pathogens or Animals or Insects or Others	Refs
4	Neem (<i>Azadirachta indica</i>)	 <p data-bbox="770 807 931 836">Azadirachtin</p>	<p>Insects: Mosquitoes: <i>Anopheles</i> sp., tobacco hornworm (<i>Manduca sexta</i>) in tobacco, fall armyworm (<i>Spodoptera frugiperda</i>) in cotton seedling</p>	<p>(Maia and Moore 2011, Senthil-Nathan 2013, Raffa 1987)</p>
5	Milkweed (<i>Asclepias syriaca</i> L.)	 <p data-bbox="770 1163 931 1192">Cardenolide</p>	<p>Insects: Butterflies (Danaini), bees, wasps, beetles, moths, and true bugs</p>	<p>(Singh and Rastogi 1970)</p>

1.6 Conservation of endangered species of valuable plants and trees

Several medicinally and economically important plants have become endangered due to over-harvesting and destruction of their natural habitats. Urbanization and the unrestricted extraction of valuable plants from the wild leads to overexploitation of natural resources (Najar and Agnihotri 2012). Several medicinal plant species, such as *Picrorhiza kurroa*, a perennial herb, once plentiful in the valleys of Kullu, have now become depleted because of habitat degradation, overexploitation, and loss of natural regeneration (Rawat 2008). Similarly, there are a large number of other plants and trees, which are beneficial for humankind's survival and well-being but are not being properly conserved and serious efforts need to be taken for their conservation. Some of the plant conservation approaches are *in situ* and *ex situ* conservation. For long-term conservation of all our natural resources, sustainable living is a viable solution by integrated management of human diseases (International Union for Conservation of Nature and Ecosystems 2006).

1.7 Genesis of the thesis

Based on literature reports spanning over the last few decades, we can most certainly propose that the scaffolds and functional groups of metabolites present in food crops and medicinal plants have a modulatory role in designing novel compounds (**Figure 1.5**).

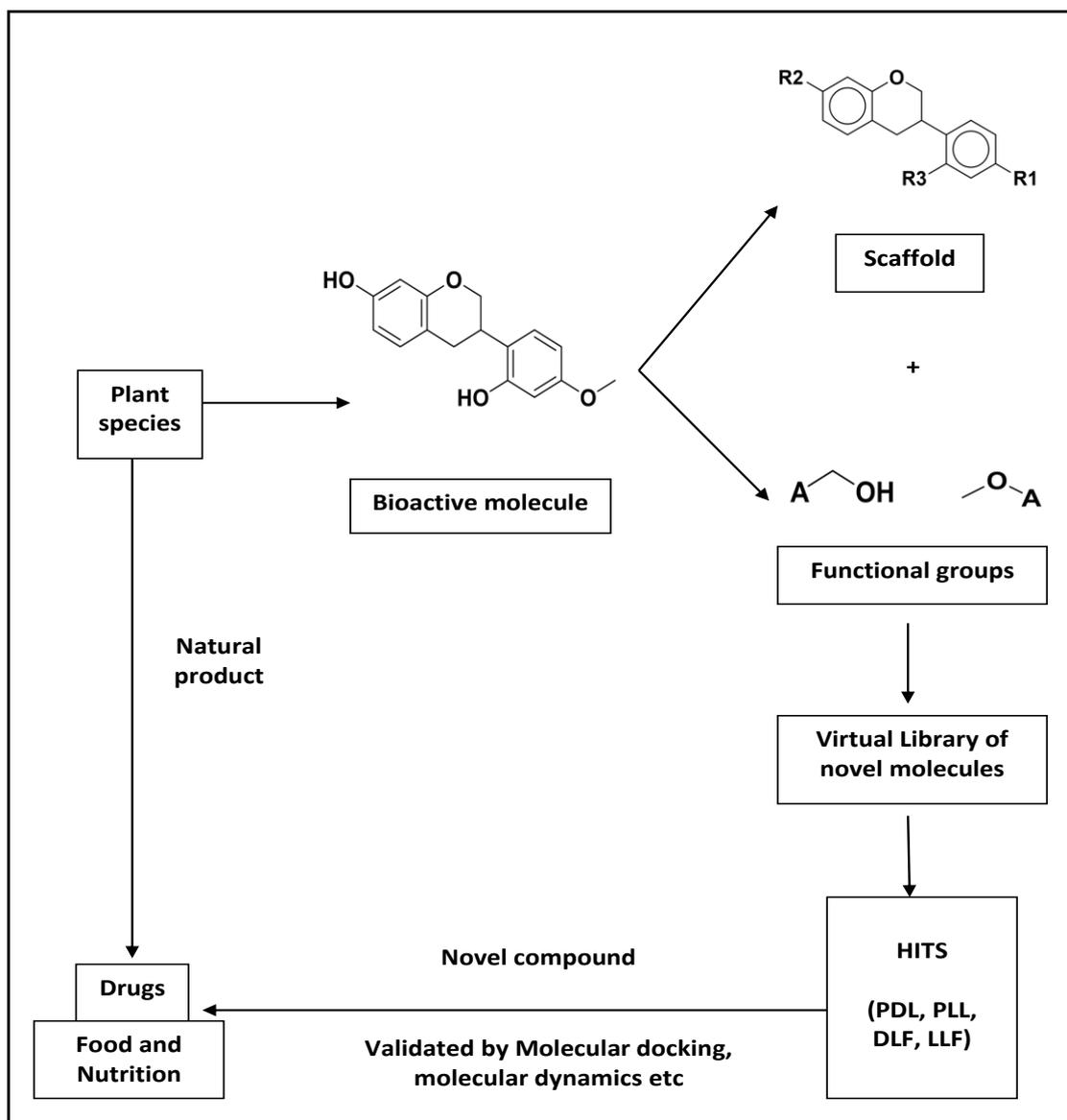


Figure 1.5: Hypothesis for designing of novel drug-like, lead-like, and pesticide-like molecules from natural plant resources

Further, the rich reservoir of bioactive molecules obtained from such plants having specific activities against human diseases can be used to build a focused virtual library of novel molecules. The novel molecules can be prioritized by progressive drug-like (PDL), progressive lead-like (PLL), drug-like failure (DLF), lead-like failure (LLF), and Toxicophoric, Pharmacophoric, and Chemophoric (TPC)

scores computed using inhouse developed ChemScreener software (Karthikeyan and Vyas 2015). This program shows the number of chemophore, toxicophore, and pharmacophore matches, which may be used to fine-tune the library created. Molecules having good scores (high pharmacophoric and low toxicophoric scores and chemophoric scores lower than pharmacophoric scores) can be used for further analysis through molecular docking and molecular dynamics. In the future, the results of this work can lead to the development of efficient and more targeted drugs and pesticides from readily available plant sources in minimum time. Indian medicinal plants are the biggest source of bioactive molecules against various diseases. As several medicinal Ayurvedic plants are not readily available and are endangered, they need to be conserved for our future well-being.

Several bioactive molecules are produced by multiple plant species, including crop plants. Hence, such plants could be used to extract bioactive compounds, saving the endangered medicinal plants. Moreover, some scaffolds are found to be shared among multiple bioactive molecules. Therefore, proper use of valuable and bioactive compounds can lead to the development of novel compounds like drugs. Considering the demand for organic production of food and drugs, novel and innovative approaches are required. Natural products derived from plants are emerging as valuable alternatives to human needs and rescuing from the bio-apocalypse. This study highlights the roles of bioactive compounds in food and medicine, including the conservation of beneficial plants. Network analysis of bioactive compounds from food crops and medicinal plants mentioned in Ayurveda and their protein targets involved in various disease-related pathways was performed. The results obtained from this study could be useful in constructive experiments for preparing polyherbal formulations against particular diseases.

In conclusion, bioactive compound scaffolds are the key to optimize chemical diversity in drugs. Furthermore, large data of published findings are available, which needs to be analyzed and managed for the best use and efficacy of the available products. This study provides the information to plan research and fill that void.

The present thesis has been organized in the following manner:

Chapter 1: Introduction and review of the literature (this Chapter)

Chapter 2: Design of Novel Drug-like Molecules using Informatics Rich Secondary Metabolites Analysis of Indian Medicinal and Aromatic Plants

Chapter 3: Bridging *in-silico* and experimental: Chemoinformatics Analysis for Mass Spectrometry-Based Metabolomics Study of Soybean

Chapter 4: Chemoinformatics Investigation on Chemical Defense in Plants

Chapter 5: Summary and future directions

Bibliography

CHAPTER 2

**DESIGN OF NOVEL DRUG-LIKE
MOLECULES USING INFORMATICS
RICH SECONDARY METABOLITES
ANALYSIS
OF INDIAN MEDICINAL AND
AROMATIC PLANTS**

Chapter 2: Design of Novel Drug-like Molecules using Informatics Rich Secondary Metabolites Analysis of Indian Medicinal and Aromatic Plants

2.1 Introduction

A large amount of experimental data related to biological and chemical researches is available in the public domain. These data can be of immense use in drug discovery (Yoo et al. 2012, Kostoff 2005). With the help of various chemoinformatics tools, these data can be mined and analyzed to discover candidate biomolecules. Chemoinformatics techniques can help in an accurate prediction of bioactive molecules and their activities, making drug discovery quicker and precise (Bellis et al. 2011). This includes mining of molecules from scientific literature, predicting their structures and functions, as well as those of their targets, ligand-target screening, building compound libraries, predicting the drug-like and lead-like molecules, etc.

The Indian subcontinent has rich plant biodiversity, and several plants are medicinally and economically important (Samal 2015). There are about 17,000 species of flowering plants in India, of which 7,500 are medicinally important (Kala, Dhyani and Sajwan 2006). Of these, about 1,300 plant species are aromatic (Shiva 1998). These plants produce and exude aromatic substances such as essential oils, which are used in cooking, cosmetics, as well as for making perfumes and in pharmaceutical industries (Chauhan 1999). The plant families such as *Lauraceae*, *Umbelliferae*, *Myrtaceae*, and *Labiatae* have several aromatic as well as medicinal species (Wojdylo, Oszmianski and Czemerz 2007). Many of these aromatic and medicinal plants have been screened for biological activities, and several bioactive

compounds have been isolated from them (Raut and Karuppayil 2014). However, a large number of medicinally relevant compounds are yet to be discovered due to the enormous structural diversity and biological activities of plant-derived compounds. The increased demand for plant-based drugs has resulted in the over-exploitation of the medicinally important species from their native habitats (Verma et al. 2012).

Over the past six decades, the Council of Scientific and Industrial Research (CSIR) has contributed to the essential oil-based aroma industry and medicinal plants-driven drug discovery in India (Baruati and Gogoi 2020). The “CSIR-Aroma and Phytopharmaceutical Mission” was conceptualized in 2016 to bring a decisive and transformative change in the current rural economy, market dynamics and growth opportunity through research on aromatic plants, in which end-to-end technology and value addition solutions will be provided across the country at a sizable scale.

Comprehensive knowledge of the metabolic profile of plants is essential for assessing their medicinal values. Chemoinformatics techniques can help in making a sense of the vast chemical and bioactivity data and convert it into knowledge useful in drug discovery. Previous studies in our lab identified medicinally important molecules from marine organisms (Karthikeyan and Vyas 2015). In the present work, we have developed a Java-based database cum toolkit for collecting the data regarding Indian medicinal & aromatic plants and associated molecules, their structural information, plant information, and traditional therapeutic use for comparing with drug molecules and for virtual library generation. There are some important recent efforts for building online databases (Mohanraj et al. 2018, Pathania, Ramakrishnan and Bagler 2015, Polur et al. 2011) about Indian medicinal plants, their phytochemicals, and therapeutic uses. Polur *et al.* (2011) compiled the information on Ayurvedic plants with their phytochemical and therapeutic properties and also studied the structural similarity of

phytochemicals with drugs from DrugBank, which is an important step in drug discovery. Subsequently, the Phytochemica database (Pathania et al. 2015) provided 963 phytochemicals derived from five Indian medicinal plants with chemical structures and pharmacological properties. The IMPPAT (Mohanraj et al. 2018) database recently compiled information on 1742 Indian medicinal plants and their 9596 phytochemicals and 1124 therapeutic uses.

The features of some of the previously reported databases have been compared (**Table 2.1**) with DoMINE (**D**atabase **o**f **M**edicinally **I**mportant **N**atural products from **p**lanta**E**) that we have developed. It was found that the previous databases are limited to categorizing and classifying the medicinal plants, their phytochemicals and therapeutic properties. However, building a virtual library of novel molecules from known medicinal molecules and virtual screening of those novel molecules also plays a pivotal role in drug development provided by DoMINE. It also includes the catalogue of Indian medicinal plants, their phytochemicals, therapeutic properties and scaffold similarity comparison with approved drugs by the generation of physicochemical descriptors. For this purpose, we text mined the literature related to medicinal plant species and identified ring containing molecules (n = 1665) associated with each plant species. We extracted molecular scaffolds (n = 209) from these molecules and used diverse scaffolds to build a focused virtual library. Using chemoinformatics approaches, we predicted drug-like and lead-like molecules from these medicinal plant molecules to elucidate the molecular basis of therapeutic indications of Indian medicinal and aromatic plants.

Table 2.1: Comparison of DoMINE with other databases on Indian medicinal plants and their metabolites for drug development

Database	DoMINE	IMPATT	Phytochemica	Polur <i>et al.</i> (2011)
<u>Input Data</u>				
Number of Indian medicinal plants	104	1742	5	295
Number of text mined metabolites	3459	9596	963	1829
Number of therapeutic properties/ uses	16	1100	Nil	Nil
Number of Scaffolds and functional groups extracted from plant metabolites	209 and 97	Nil	Nil	Nil
Number of drugs from DrugBank	2334	2069	Nil	4887
	(FDA Approved drugs)	(FDA Approved drugs)		
Number of scaffolds and functional groups extracted from drug molecules	306 and 291	Nil	Nil	Nil

Database	DoMINE	IMPATT	Phytochemica	Polur <i>et al.</i> (2011)
Images of plants with their parts	Yes	Nil	Nil	Nil
<u>Interconnections</u>				
Plant - Family	Yes	Nil	Nil	Nil
Plant- metabolites	Yes	Yes	Yes	Yes
Plant- therapeutic properties	Yes	Yes	Nil	Yes
Metabolite- Scaffolds and Functional groups	Yes	Nil	Nil	Nil
Drugs- Scaffolds and Functional groups	Yes	Nil	Nil	Nil
Structural similarity search between plants and drugs (molecules, scaffolds based)	Yes (molecules and scaffolds based)	Yes (molecules based)	Yes (molecules based)	Yes (molecules based)
<u>Other features</u>				
Web interface / GUI application	Yes	Yes	Yes	Nil

Database	DoMINE	IMPATT	Phytochemica	Polur <i>et al.</i> (2011)
Chemical structure representation (2D and 3D)	Yes	Yes	Yes (3D)	Nil
Downloadable structure file formats	Yes	Yes	Yes	Nil
Physiochemical properties (2D)	Yes	Yes	Yes	Nil
Similarity search (sub, similar, super, formula)	Yes	Nil	Nil	Nil
Scaffold and functional groups extraction	Yes	Nil	Nil	Nil
Virtual library generation	Yes	Nil	Nil	Nil
Virtual screening (Prediction of chemical properties- TPC, PDL, PLL, DLF, LLF scores)	Yes	Nil	Nil	Nil

2.2 Materials and methods

2.2.1 Data Collection

The databases from which the plant data were collected include the Indian Medicinal Plants Database (IMPD) (<http://www.medicinalplants.in/>) and “The Wealth of India” (Bhat 1997) (**Table 2.2**). The IMPD contains information about the plants used in Ayurveda (2559), Siddha (2267), Unani (1049) and Homeopathy (460). From these four systems of medicine, the names of common medicinal and aromatic plants (n = 104) were identified, which also included the “CSIR-Aroma Mission 2016” based medicinal plants of India. “The Wealth of India” consists of a bibliographic full-text database of Indian medicinal and aromatic plants (Bhat 1997). The database covers over 5,000 plant species belonging to about 1,800 plant genera. All the available information about the medicinal plants was manually extracted from these databases. Sowa rigpa and folk plants were not included in this study as the information about them is passed on from one generation to the next through word of mouth. Similarly, FDA-approved drugs were obtained from DrugBank (Wishart et al. 2017) for the structural comparison with plant molecules.

Table 2.2: List of medicinal plant databases

Sr. no.	Database	URL	Number of Plants#
1.	Indian Medicinal Plants Database	http://www.medicinalplants.in/	
	Ayurveda	http://www.medicinalplants.in/showfullist/ayurveda	2559
	Siddha	http://www.medicinalplants.in/showfullist/siddha	2267
	Unani	http://www.medicinalplants.in/showfullist/unani	1049
	Homeopathy	http://www.medicinalplants.in/showfullist/homeopathy	460
2.	The Wealth of India (1948-1992) (Bhat 1997)	http://www.niscair.res.in/includes/images/wealthofindia/woi-article.pdf	>5000

#: No. of plants as of March 2021

2.2.2 Computational protocol

The computational protocol followed in the present study is presented in **Figure 2.1** and **Supplementary File S2.1**. Being quick and cost-effective, this methodology can evaluate a large number of novel and potential drug candidates against diseases for further experimental studies. We used the traditional therapeutic properties of plants to establish the predicted bioactivities of their natural molecules. The details of bioactivity studies already performed on the Indian medicinal plants are listed in **Supplementary Table S2.1.1**. **Supplementary Table S2.1.2** shows the classification of Indian medicinal plants according to the Indian systems of medicines, *i.e.*, Ayurveda, Siddha, Unani, and Homeopathy. The molecules text mined from these plants (Keywords: Name of Medicinal plants through PubMed literature, Dec. 2019) were used to identify particular scaffolds or drug-like or lead-like compounds with expected bioactivity.

Metabolite names (n = 3459) were extracted from PubMed abstracts of each medicinal plant through text mining using PubTator (Wei, Kao and Lu 2013) (**Supplementary Table S2.2**). Every text mined molecule was manually verified, and its presence in their respective medicinal plant was confirmed by reading the concerned PubMed abstract(s). Chemical names of the extracted plant molecules were converted into SMILES (Simplified Molecular Input Line Entry System) strings and screened for 5-6 membered ring-containing molecules up to 1000 molecular weight. Similarly, a representative list of Indian medicinal and aromatic plants with their therapeutic categories and PubMed counts is presented in **Table 2.3**. Scaffolds (n = 209) and functional groups (n = 97) (**Supplementary Table S2.3.1**) were generated from the ring containing molecules (n = 1665) employing an in-house developed program ChemScreener (Karthikeyan, Pandit and Vyas 2015b,

Karthikeyan and Vyas 2014). Similarly, scaffolds (n = 306) and functional groups (n = 291) were also extracted from the approved drug molecules (n = 2354) (Supplementary Table S2.3.2).

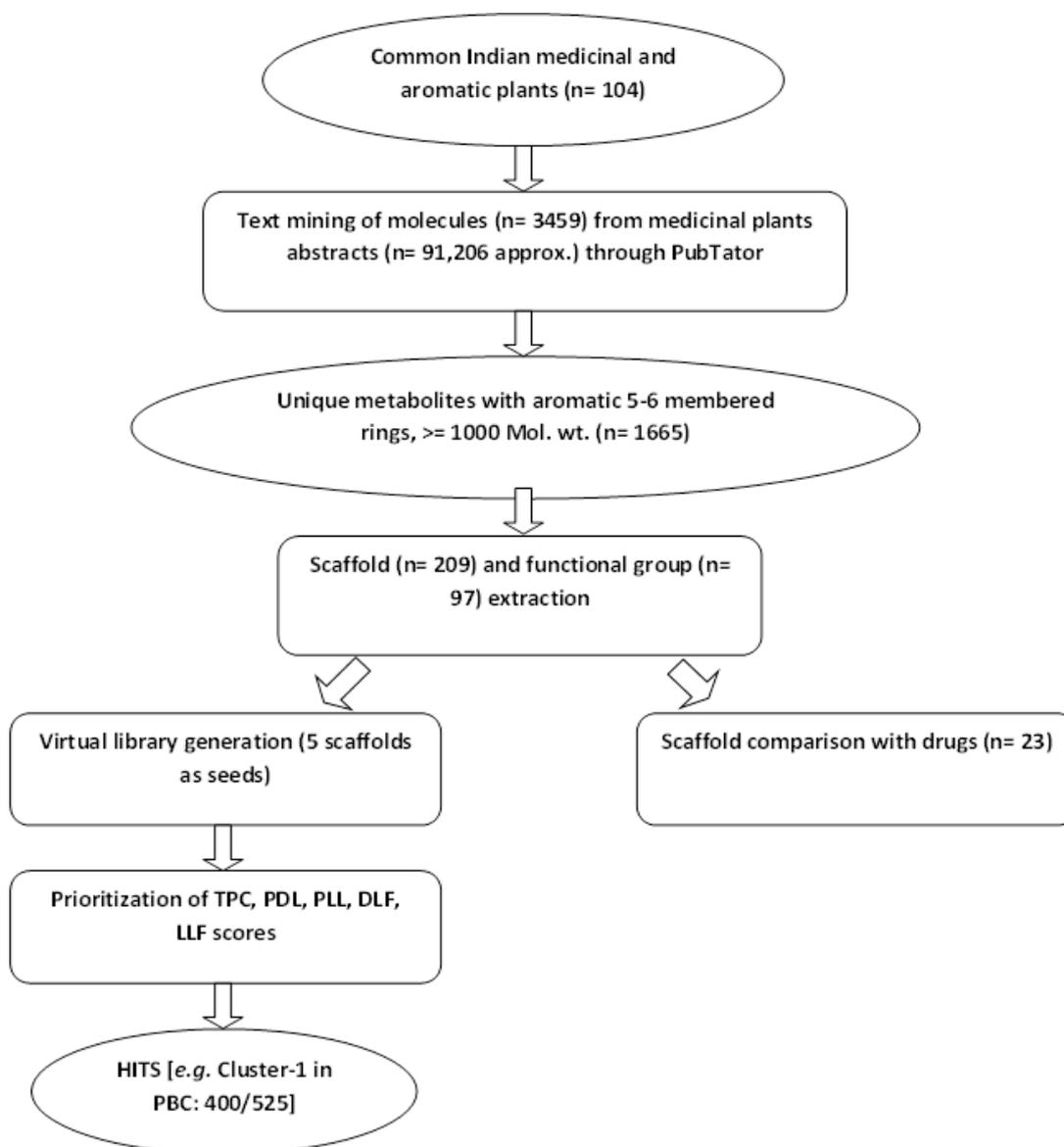


Figure 2.1: Workflow highlighting the steps of extracting drug-like molecules from medicinal and aromatic plants (TPC = Toxicophoric, Pharmacophoric, and Chemophoric, PDL= Progressive Drug Like, PLL= Progressive Lead Like, DLF= Drug Like Failure, LLF= Lead Like Failure features as generated in ChemScreener program, PBC= Plant-Based Clustering)

Table 2.3: A representative list of Indian medicinal and aromatic plants with their therapeutic categories and PubMed counts (as of March 2021)
 (* For more therapeutic categories, please refer to Supplementary Table S2.1)

Sr. no.	Plant Species	Common Name	Therapeutic Properties	Number of Therapeutic Properties	Number of PubMed Publications
1	<i>Valeriana jatamansi</i>	Garden valerian	Carminative or laxative, stimulant, hypnotic, analgesic, diuretic, diaphoretic or antipyretic, anti-inflammatory, neuroprotective	8	101
2	<i>Centella asiatica</i>	Gotu kola	Anti-inflammatory, anti-diabetic, neuroprotective, anti-cardiac, analgesic, anticancer, antimicrobial, carminative or laxative	8	899
3	<i>Semecarpus anacardium</i>	Marking nut	Anti-inflammatory, antimicrobial, stimulant, anticancer, anti-cardiac, expectorant, carminative or laxative	7	141
4	<i>Allium sativum</i>	Garlic	Diaphoretic or antipyretic, expectorant, anticancer, diuretic, anti-cardiac, stimulant, anti-diabetic	7	7088

Sr. no.	Plant Species	Common Name	Therapeutic Properties	Number of Therapeutic Properties	Number of PubMed Publications
5	<i>Cymbopogon winterianus</i>	Java citronella	Anticancer, stimulant, analgesic, carminative or laxative, antimicrobial, expectorant	6	56
6	<i>Santalum album</i>	Indian sandalwood	Anti-cardiac, antimicrobial, diuretic, diaphoretic or antipyretic, expectorant, anti-inflammatory	6	138
7	<i>Cymbopogon martini</i>	Palmarosa or Indian geranium	Carminative or laxative, antimicrobial, stimulant, expectorant, anti-cancer	5	25
8	<i>Cannabis sativa</i>	Bhang or Marijuana	Hallucinogenic, hypnotic, anti-inflammatory, analgesic, anti-cancer	5	24538

All the scaffolds and molecules extracted from the Indian medicinal and aromatic plants related to scientific literature were compared with the drug scaffolds by generating a network. Five random scaffolds of molecules identified from the medicinal plants were selected to form a cluster. Six clusters of unique scaffolds were built. These scaffolds were used as seeds for generating a virtual library using ChemScreener. The scaffolds in the virtual library were annotated using Toxicophoric, Pharmacophoric and Chemophoric (TPC) scores and progressive drug-like (PDL), progressive lead like (PLL), drug-like failure (DLF), and lead-like failure (LLF) scores using ChemScreener to get a focused set of novel and virtual bioactive molecules.

2.2.3 Softwares and Databases

A total of 91,206 PubMed abstracts related to the 104 Indian medicinal and aromatic plant species were downloaded (Sood and Ghosh 2006). We used PubTator (Wei et al. 2013), a web-based text mining tool from NCBI, to extract chemical entities (n = 3459) from PubMed literature using PMID numbers. Structures of approved drugs were obtained from DrugBank (Wishart et al. 2017). The chemical names of the molecules were converted into SMILES by using the JChem-Base ChemAxon tool (Weber 2008). All the data were converted to SDF format for easy access in Molecular Operating Environment (MOE) used for descriptor generation and principal component analysis (PCA) (Chemical Computing Group 2008). ChemScreener was used for generating scaffolds and a virtual library with TPC, PDL, PLL, DLF, and LLF scores. Cytoscape 3.7.1 (Ross 2010) was used to generate networks of Indian medicinal and aromatic plants, medicinal plant families, their text mined ring containing molecules and their scaffolds, drug molecules, and their scaffolds. We extracted scaffolds and functional groups from the text-mined

molecules from literature and built a focused virtual library of novel molecules. StatistiXL 1.8 (Robert and Wither 2007) was used to generate dendrograms from clusters with Euclidean distance matrix calculation for plant-based clustering method; whereas, LibMCS 6.1.0 provided by ChemAxon (Zloh et al. 2017), was used for clustering all the molecules based on their maximum common substructures in a hierarchical manner.

2.3 Results and Discussion

This study aimed to utilize traditional knowledge about the Indian medicinal plants to identify natural bioactive molecules to design novel drugs. From the various medicinal systems (Ayurveda, Siddha, Unani, and Homeopathy) and databases (IMPD and “The Wealth of India”) of Indian medicinal plants, common medicinal and aromatic plants were identified. These Indian medicinal plants are traditionally used in Ayurvedic medicines in polyherbal formulations prescribed mainly for healthy living rather than the treatment of diseases (Gogte 2000, Parasuraman, Thing and Dhanaraj 2014). We categorized the 104 Indian medicinal and aromatic plants into 16 selected generally studied therapeutic properties (**Figure 2.2**). Among them, 15 medicinal plants, including *Atropa belladonna*, *Aconitum ferox*, also produce high levels of toxic compounds like atropine and scopolamine (Glatstein et al. 2014, Panda and Debnath 2010). These are the main toxic tropane alkaloids that were quantitatively determined in *Atropa belladonna*, *Datura stramonium*, and other species of the *Solanaceae* family (Boros et al. 2010). Atropine and scopolamine have anticholinergic properties and have legitimate medical applications in very low doses. Seven plant species, including *Solanum virginianum*, *Ricinus communis*, were found to be mildly poisonous as some of their parts contain toxic metabolites, while other plant parts can be used for medicinal purposes (Khan et al. 2014, Scarpa and Guerci

1982). LC-MS methods were used to identify the toxic alkaloids in *Ricinus communis*, 3-carbonitrile-4-methoxy-N-methyl-2-pyridone (ricinine), and its carboxylic acid derivative, 3-carboxy-4-methoxy-N-methyl-2-pyridone (Wachira et al. 2014). Ricinine is one of the main compounds obtained from *Ricinus communis*, which shows insecticidal effects against leaf-cutting ant (*Atta sexdens rubropilosa*) (De Melo Casal et al. 2009), *Spodoptera frugiperda*, etc. (Ramos-López et al. 2010).

As expected, several medicinal and aromatic plants had multiple medicinal properties. For example, *Centella asiatica* is reported to have at least eight therapeutic properties such as anti-inflammatory, anti-diabetic, neuroprotective, anti-cardiac, analgesic, anticancer, antimicrobial, carminative, and laxative (Duarte and Rai 2015). LC-MS fingerprint analysis of *Centella asiatica* extracts revealed asiatic acid and madecassic acid as the dominant components (Jiang et al. 2016). The combination of asiatic acid and madecassic acid shows an effective means to intervene in neurodegenerative diseases in which neurotrophin deficiency is involved. *Allium sativum* has been reported for seven therapeutic properties such as diaphoretic or antipyretic, expectorant, anticancer, diuretic, anti-cardiac, stimulant and anti-diabetic (Augusti 1996). Polyphenolic compounds and sterol were identified from *Allium* species using the HPLC-UV-MS method (Vlase et al. 2013, Martins, Petropoulos and Ferreira 2016). Allicin present in garlic (*Allium sativum*) is responsible for its typical flavor and has antimicrobial properties (Pacirc et al. 2010). Among all the medicinal properties studied in this work, most of the medicinal plants possess biological activities like anti-inflammatory, anti-microbial, anti-cancer, carminative, and laxative.

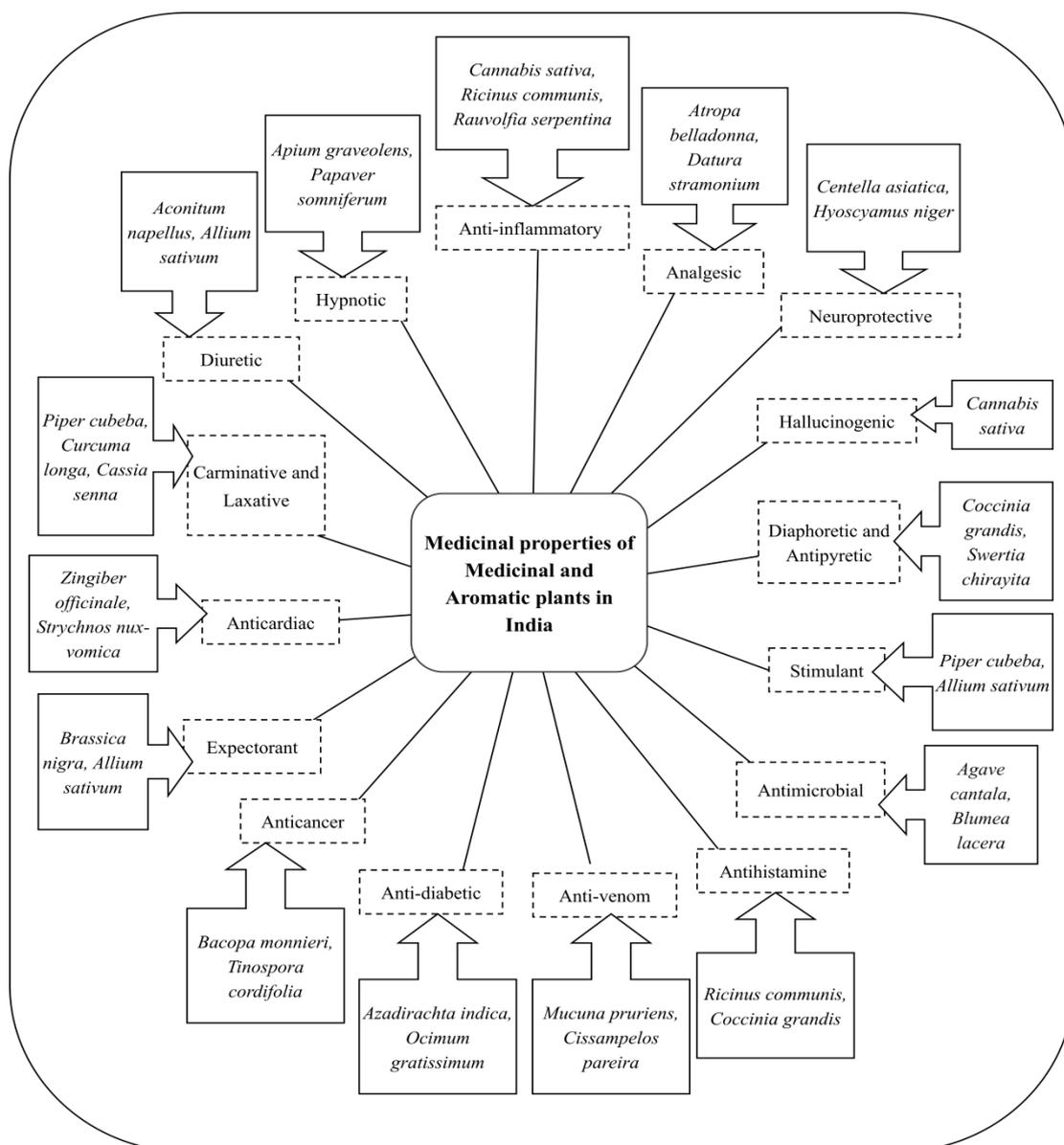


Figure 2.2: Schematic view of 16 medicinal properties of 104 Indian medicinal and aromatic plants (* List of all other medicinal plants with their medicinal properties is provided in Supplementary Table S2.4).

2.3.1 Chemoinformatics Analysis Based on Scientific Literature Mining

Text mining was performed to identify plant molecules from PubMed literature, citing Indian medicinal and aromatic plants. The scientific trend in publications dealing with botanical families of promising medicinal plant species is presented in **Figure 2.3** and **Supplementary Table S2.4.1**. The families with the highest number of records were

Poaceae, *Brassicaceae*, *Fabaceae*, *Solanaceae*, and *Asteraceae*. The *Poaceae* family was the most studied with 13,254 publications whereas, only 179 publications were reported for the *Caesalpiniaceae* family. This is because the number of plants found under the *Poaceae* family is more and readily available, which is the opposite in the case of the *Caesalpiniaceae* family. Similarly, many other Indian medicinal plants are yet to be scientifically explored, such as *Hygrophila schulli* with only two publications, *Agave cantala* with three publications, *Anamirta cocculus* with three publications, etc. Hence, very few or no molecules could be mined from the publications mentioning these plants.

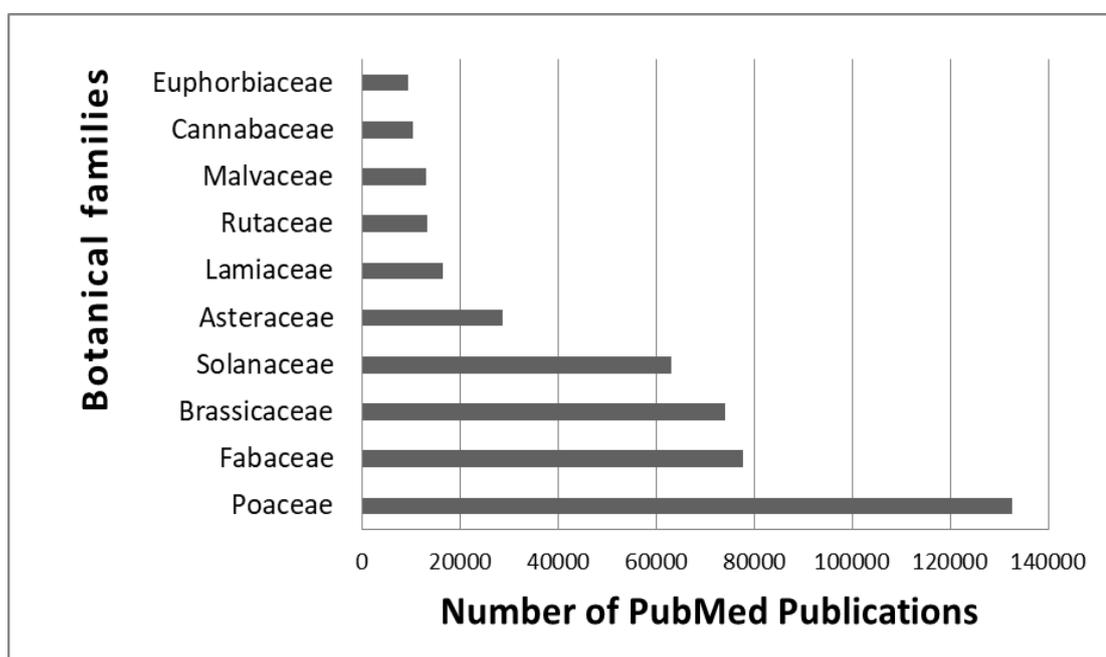


Figure 2.3: Distribution of the number of PubMed publications for the top 10 families of the medicinal and aromatic plants (as of March 2021)

All the plants were categorized based on their therapeutic properties (**Supplementary Table S2.4.2**). The top ten promising plants were selected from them, which are reported to possess the maximum number of therapeutic properties. The number of publications obtained by using the scientific names of these promising

plants as keywords is presented in **Figure 2.4** and **Supplementary Table S2.2**. It was found that the most studied plants based on bioactivity assessments were *Cannabis sativa*, *Piper cubeba*, and *Allium sativum* with 24538, 8820, and 7088 publications, respectively. Whereas, the least studied plants based on the biological evaluation were *Cymbopogon khasianus* and *Hygrophila schulli*, with only 2 and 3 publications, respectively.

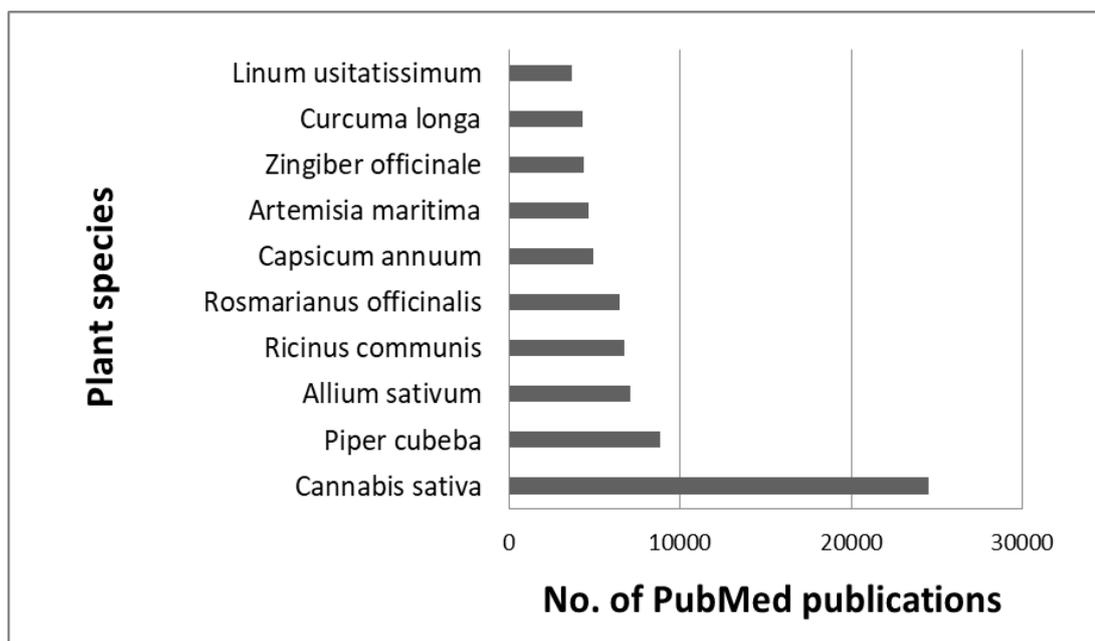


Figure 2.4: Distribution of the number of PubMed publications for the top 10 medicinal and aromatic plants (as of March 2021)

The molecules extracted through text mining for medicinal plants have pharmacological effects. For example, 6-gingerol, and 6-paradol extracted from ginger (*Zingiber officinale*) have anticancer activities (Mishra, Kumar and Kumar 2012), withaferin A from *Withania somnifera* exhibits anti-arthritic and anti-inflammatory activities (Uddin et al. 2012), etc. The text mined data from the literature related to Indian medicinal plants shows that there is still a lot of hidden potential in the Indian medicinal and aromatic plants, which needs to be explored. We

believe that this hidden treasure can be explored and untangled by applying various chemoinformatics tools and methods.

A 2D principal component analysis (PCA) of all the molecules was performed to study their distribution in the chemical space. The analysis depicts the diverse nature of the chosen molecules due to their different chemical structures. The PCA was performed by generating molecular descriptors such as the number of hydrogen bond acceptor atoms, number of H-bond donor atoms, Lipinski druglike test, log solubility in water, number of rings, molecular weight, Weiner path number, number of rotatable bonds, etc. for all the molecules (**Supplementary Table S2.5**). The molecules having more unique features in their chemical structures occupied separate regions in the plot (**Figure 2.5**). Some of the representative molecules of Indian medicinal plants were randomly picked up, such as curcuminoid D, turpethoside B, bisgingerdione A, etc., and are highlighted in the 2D PCA plot figure. It was found that most of the outlier molecules in the PCA plot figure had complex polycyclic structures. Scaffolds were extracted from all these unique molecules. A few representative scaffolds are presented in **Table 2.4** with their 2D structures, therapeutic properties, molecule names, and source of plant species. This table clearly shows that the scaffolds extracted from the bioactive molecules obtained from the medicinal plants can be used to synthesize drug-like and lead-like molecules. It was found that Scaffold ID 2 was common between carvacrol and crocin from the plants *Majorana hortensis* and *Crocus sativus*, representing anti-inflammatory and diaphoretic or antipyretic properties, respectively. It has been reported that the chemical nature of a molecule is associated with its chemical structure (Harikarnpakdee and Chuchote 2018). Here, the chemical structure of the molecule is represented by both scaffolds and functional groups. Hence, different therapeutic properties are not only based on scaffolds but also on their linked functional groups.

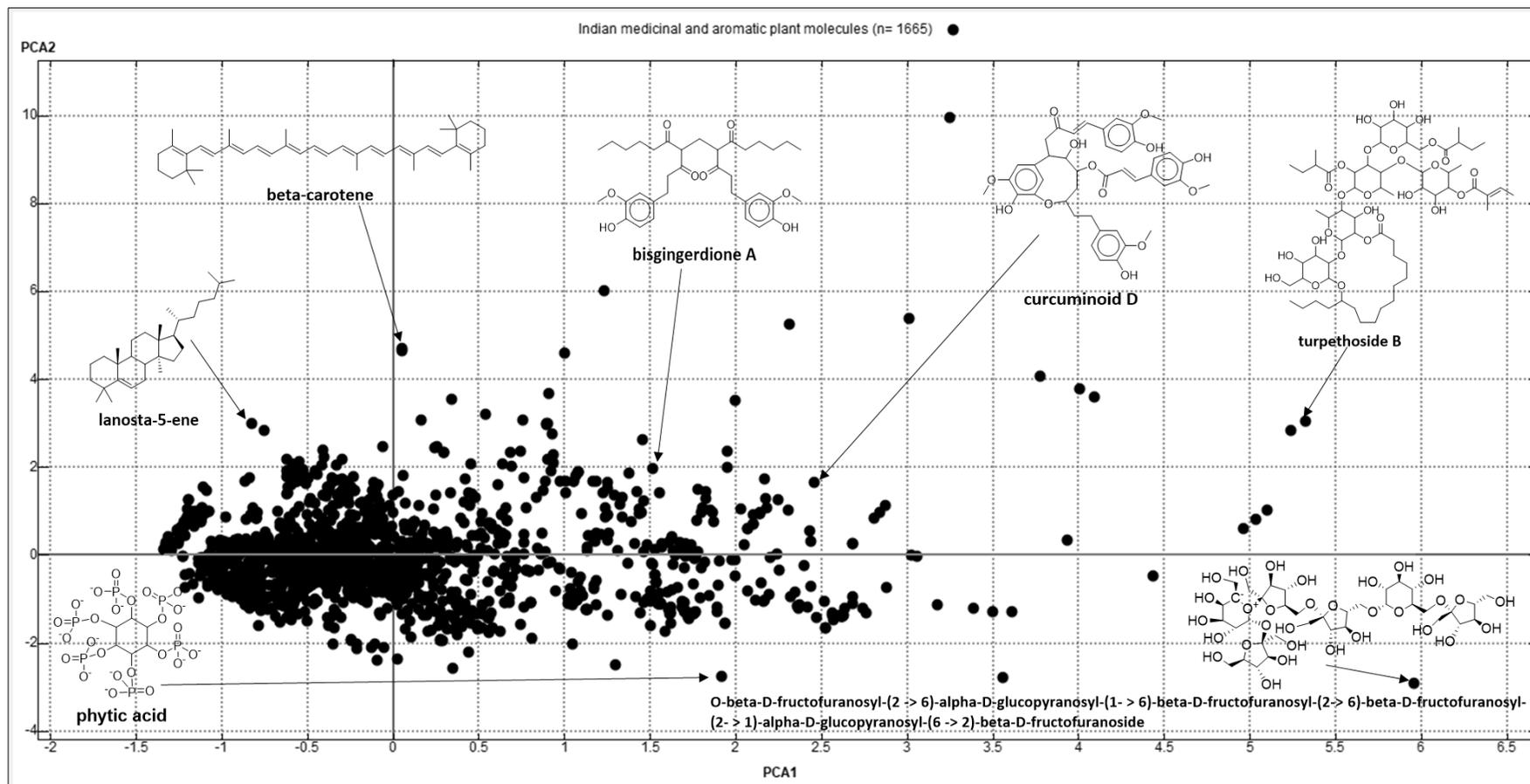
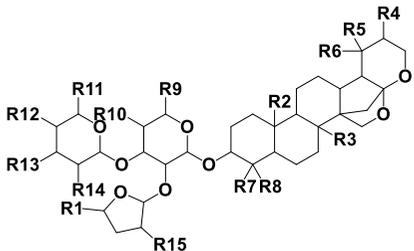
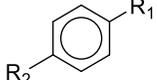
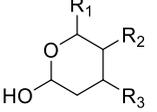
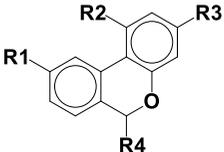
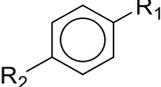
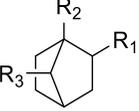
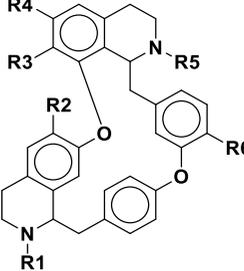
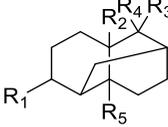
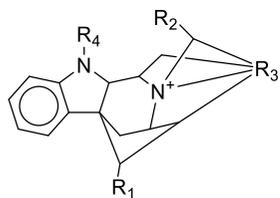


Figure 2.5: The 2D PCA plot representing the molecular diversity of natural products from Indian medicinal plants (Supplementary Table S2.5)

Table 2.4: Representative scaffolds from 104 aromatic and medicinal plant molecules representing 16 medicinal properties (* For more scaffolds, please refer to Supplementary Table S2.1)

 <p>Scaffold ID: 1</p> <p>Molecule Name: Bacopaside II</p> <p>Plant source: <i>Bacopa monnieri</i></p> <p>Therapeutic category: Antihistamine</p>	 <p>Scaffold ID: 2</p> <p>Molecule Name: Carvacrol</p> <p>Plant source: <i>Majorana hortensis</i></p> <p>Therapeutic category: Anti-inflammatory</p>	 <p>Scaffold ID: 3</p> <p>Molecule Name: Lotaustralin</p> <p>Plant source: <i>Linum usitatissimum</i></p> <p>Therapeutic category: carminative or Laxative</p>	 <p>Scaffold ID: 4</p> <p>Molecule Name: Cannabinol</p> <p>Plant source: <i>Cannabis sativa</i></p> <p>Therapeutic category: Hallucinogenic</p>
--	---	---	--

 <p>Scaffold ID: 2</p> <p>Molecule Name: Crocin</p> <p>Plant source: <i>Crocus sativus</i></p> <p>Therapeutic category: Diaphoretic or Antipyretic</p>	 <p>Scaffold ID: 5</p> <p>Molecule Name: Borneol</p> <p>Plant source: <i>Artemisia nilagirica</i></p> <p>Therapeutic category: antimicrobial</p>	 <p>Scaffold ID: 6</p> <p>Molecule Name: tetrandrine</p> <p>Plant source: <i>Centella asiatica</i></p> <p>Therapeutic category: neuroprotective</p>	 <p>Scaffold ID: 7</p> <p>Molecule Name: patchoulol</p> <p>Plant source: <i>Pogostemon cablin</i></p> <p>Therapeutic category: analgesic</p>
---	---	--	---

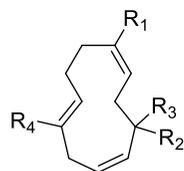


Scaffold ID: 8

Molecule Name: Ajmaline

Plant source: *Rauvolfia serpentina*

Therapeutic category: Anti-venom

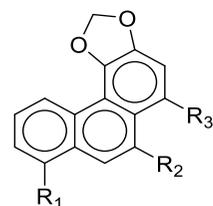


Scaffold ID: 9

Molecule Name: alpha-humulene

Plant source: *Ocimum gratissimum*

Therapeutic category: Anticancer



Scaffold ID: 10

Molecule Name: Aristolochic acid

Plant source: *Asarum europaeum*

Therapeutic category: Stimulant



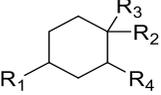
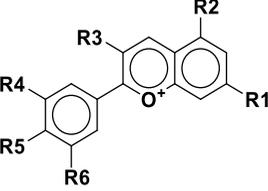
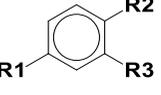
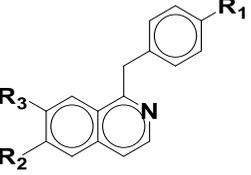
Scaffold ID: 11

Molecule Name: Santalene

Plant source: *Santalum album*

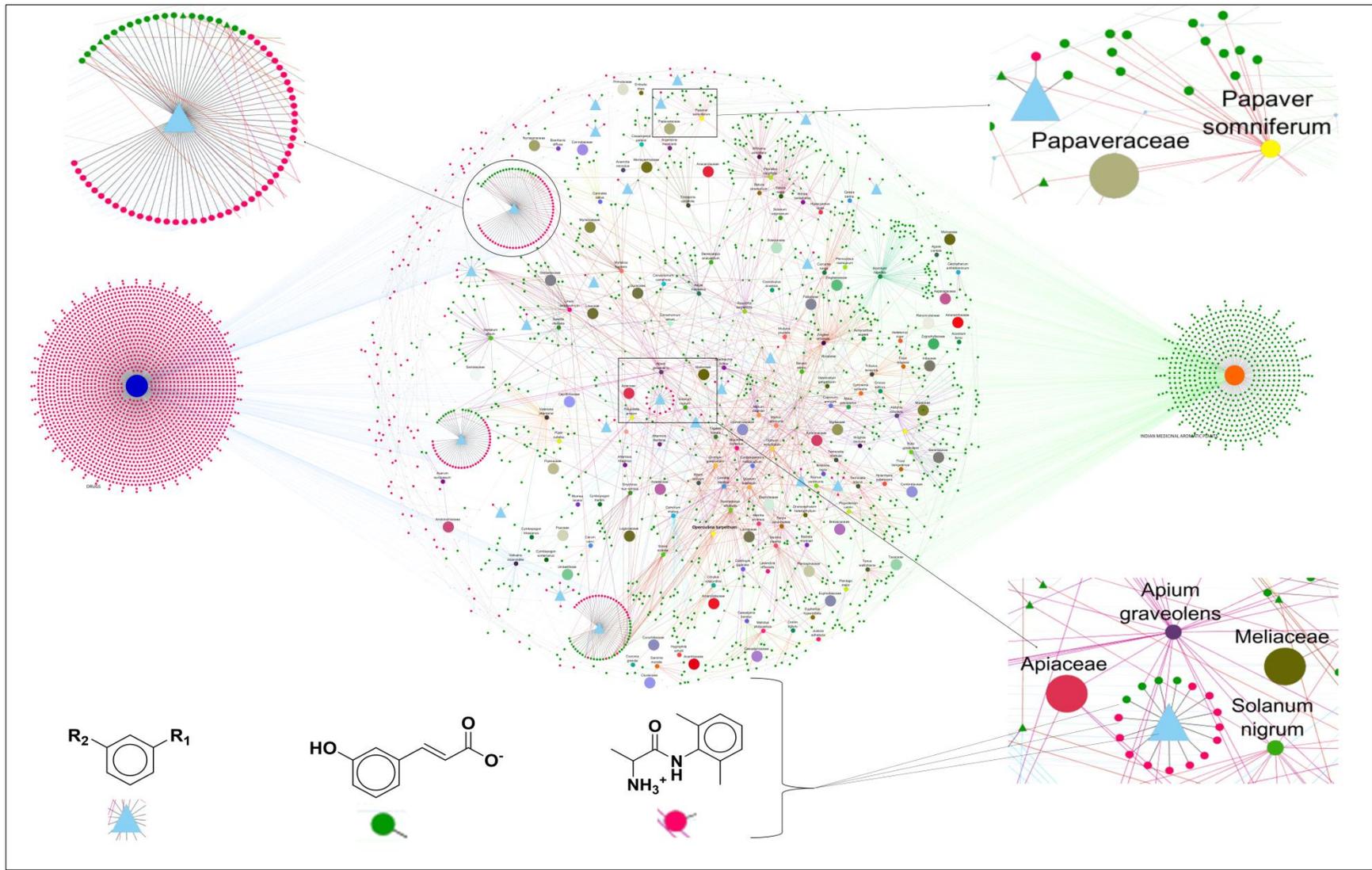
Therapeutic category:

Expectorant

 <p>Scaffold ID: 12</p> <p>Molecule Name: Beta- elemene</p> <p>Plant source: <i>Ocimum gratissimum</i></p> <p>Therapeutic category: anti-diabetic</p>	 <p>Scaffold ID: 13</p> <p>Molecule Name: delphinidin</p> <p>Plant source: <i>Crocus sativus</i></p> <p>Therapeutic category: Anticardiac</p>	 <p>Scaffold ID: 14</p> <p>Molecule Name: Ferulic acid</p> <p>Plant source: <i>Linum usitatissimum</i></p> <p>Therapeutic category: Diuretic</p>	 <p>Scaffold ID: 15</p> <p>Molecule Name: Papaverine</p> <p>Plant source: <i>Papaver somniferum</i></p> <p>Therapeutic category: Hypnotic</p>
--	---	---	--

2.3.2 Scaffold Drug Network of the Indian Medicinal and Aromatic Plant Species

A large network depicting relationships among the scaffolds obtained from medicinal plant species and approved drug molecules was constructed (**Figure 2.6**). The structural similarity of plant molecules was matched with the drug molecules based on their scaffolds to infer the similarity between traditional medicinal use of the plants with the medicinal use of the drugs (Lagunin et al. 2014). In this network, scaffolds act as linkers between molecules from medicinal plants and known drugs. Seven networks (Drug molecules, Drugs, and their Scaffolds, Medicinal plants, Medicinal plant families, Medicinal Plants and their families, Medicinal plant small molecules, and Medicinal plant small molecules and their scaffolds) were constructed to study the inter-relationship between scaffolds and molecules of the Indian medicinal plant species and the drugs. All the scaffold networks were then merged to create a supra-network containing 4623 nodes and 6216 edges. The network analysis of the topological features computed for the merged network showed an average number of neighbors with 2.689 and characteristic path length with 3.173 scores depicting the maximum connectivity of all molecules and their common scaffolds. The set of neighbors of a particular node 'n' is known as its neighborhood. The size of n's neighborhood is given by 'kn', which is its connectedness. The average number of neighbors reflects a node's average connectedness in the network. The predicted distance between two linked nodes is given by the average shortest path length, also known as the characteristic path length.



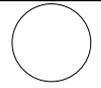
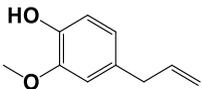
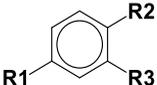
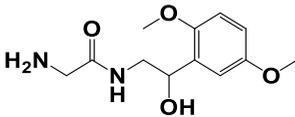
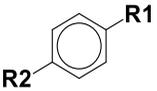
	Indian medicinal aromatic plants molecules (n = 1665) [Node size = 100]
	Drug molecules (n = 2354) [Node size = 100]
	Scaffolds [Node size = 35]
	Indian medicinal aromatic plants families (n = 47) [Node size = 500 with 47 different colors for each family]
	Indian medicinal aromatic plants (n = 104) [Node size = 200 with 104 different colors for each plant]
	Common scaffolds between Indian medicinal aromatic plants molecules and drugs: (n = 23) [Node size = 500]
	Common molecules among Indian medicinal aromatic plants: (n = 169) [Node size = 100]

Figure 2.6: Indian medicinal aromatic plant molecules, drug molecules, and scaffolds merged network as depicted in organic and edge-weighted spring embedded layout (for selected nodes only) in Cytoscape. Nodes = 4623, edges = 6216 (Nodes: Molecules, scaffolds, plants and plant families; Edges: Family-Plant-Molecule-Scaffold Interactions/ hidden relationships)

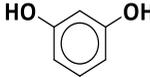
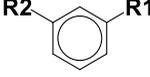
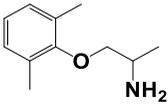
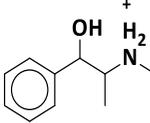
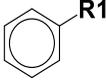
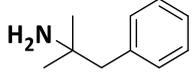
The scaffolds obtained from plant molecules and drugs were compared to obtain common scaffolds among them (**Table 2.5** and **Supplementary Table S2.6**). Some molecules were identified from multiple plant sources; e.g. eugenol has been reported from several plant species such as *Cinnamomum verum*, *Cymbopogon martini*, *Linum usitatissimum*, *Ocimum basilicum*, *Ocimum gratissimum*, *Ocimum tenuiflorum*, *Pogostemon cablin*, *Majorana hortensis*, *Mucuna pruriens*, *Myristica fragrans*, etc. (Dardouri 2019, Ross 2010, Khan and Ahmad 2011, Harikarnpakdee and Chuchote 2018, Swamy and Sinniah 2015, Padalia et al. 2014, Du et al. 2014). Twenty-three scaffolds were common between Indian medicinal & aromatic plant molecules and drugs, e.g., eugenol from *Ocimum basilicum*, L-alpha-curcumene from *Curcuma longa*, secoisolariciresinol from *Linum usitatissimum*, have similar scaffolds to the drugs Midodrine, Esmolol, and Masoprocol, respectively. The scaffold ID 14 of eugenol was similar to that of Midodrine, which is a vasoconstrictor agent (Thulesius, Gjöres and Berlin 1979). The same scaffold was shared among multiple plant molecules such as vanillin, thymol, carvacrol, etc. Likewise, scaffold ID 17 of L-alpha-curcumene showed similarity with the scaffold extracted from Esmolol, a cardio-selective beta-1 receptor blocker (Lee et al. 2016b). It was found that 182 plant molecules and 227 drugs shared common scaffolds (n = 23). The medicinal plants (n = 90 out of 104) which contain these 182 molecules belong to a specific group of families (n = 42), whereas 169 molecules shared among all the Indian medicinal plants were identified.

Thus, this analysis revealed that the metabolites from the Indian medicinal plant species possess properties or bioactivities similar to that of known drugs, based on their common scaffold structures, as structural descriptors encode activity. The inferred similarity in bioactivities of these molecules makes the concerned plant species prospective candidates for the future development of new drugs. By analyzing the molecule scaffold network, we propose employing such an ethnopharmacological approach to identify lead molecules from plants and use them to develop novel drugs.

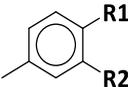
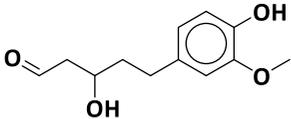
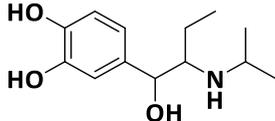
Table 2.5: Similar scaffolds identified from 104 Indian medicinal and aromatic plants molecules and drug molecules in supra network with their therapeutic category information (n = 8) (Sc: Scaffold)

Sr. No	Plant source	Plant molecules	Similar scaffold	Drug
1.	<i>Cinnamomum verum</i> , <i>Cymbopogon martini</i> , <i>Linum usitatissimum</i> , <i>Ocimum basilicum</i> , <i>Ocimum gratissimum</i> , <i>Ocimum tenuiflorum</i> , <i>Pogostemon cablin</i> , <i>Majorana hortensis</i> , <i>Mucuna pruriens</i> , <i>Myristica fragrans</i>	Eugenol  Sc ID: 14		Midodrine: Vasoconstrictor agent  DrugBank ID: DB00211 Sc ID: 2
2.	<i>Curcuma longa</i> , <i>Croton tiglium</i>	L-alpha-Curcumene 	Esmolol: Cardioselective beta1 receptor blocker	

Sr. No	Plant source	Plant molecules	Similar scaffold	Drug
		Sc ID: 17		DrugBank ID: DB00187
				Sc ID: 3
3.	<i>Linum usitatissimum</i>	Secoisolariciresinol		Masoprolol: Antineoplastic agent, antioxidant, a cyclooxygenase inhibitor, lipoxygenase inhibitor
		Sc ID: 18		DrugBank ID: DB00179
				Sc ID: 4

Sr. No	Plant source	Plant molecules	Similar scaffold	Drug
4.	<i>Apium graveolens</i>	Resorcinol  Sc ID: 19		Mexiletine: Antiarrhythmic agent  DrugBank ID: DB00379 Sc ID: 5
5.	<i>Strychnos nux-vomica</i>	Ephedrin  Sc ID: 20		Phentermine: Sympathomimetic amine anorectic agent  DrugBank ID: DB00191, Sc ID: 6
6.	<i>Brassica nigra</i>	Brassinin		L-tryptophan: Anti-depressive agent, dietary supplement

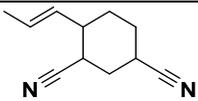
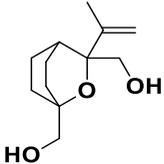
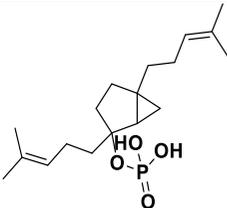
Sr. No	Plant source	Plant molecules	Similar scaffold	Drug
		Sc ID: 21		DrugBank ID: DB00150
				Sc ID: 7
7.	<i>Chrysopogon zizanioides</i>	Prometryn		Altretamine: antineoplastic
		Sc ID: 23		DrugBank ID: DB00488
				Sc ID: 9

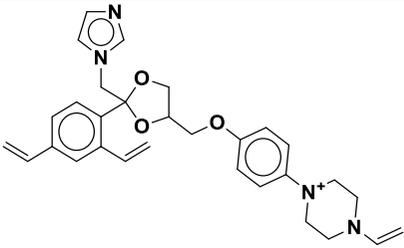
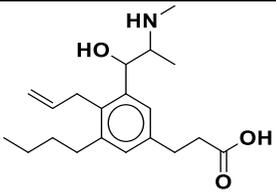
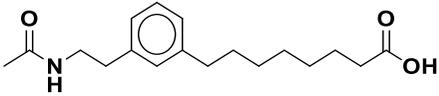
Sr. No	Plant source	Plant molecules	Similar scaffold	Drug
8.	<i>Zingiber officinale</i>	3-Hydroxy-5-(4-hydroxy-3-methoxyphenyl)pentanal		Isoetarine: fast-acting bronchodilator
				DrugBank ID: DB00221
		Sc ID: 24		Sc ID: 10

2.3.3 Screening of the Virtual Library

Large combinatorial virtual libraries have previously been developed to emphasize the chemical space of molecules as drugs by analyzing structure-activity relationships (Ghose, Viswanadhan and Wendoloski 1999). Hence, to obtain insights into the distribution of molecular data, we created clusters of virtual library molecules and studied the characteristics of each cluster. Cluster analysis was performed by two methods: (i) plant-based clustering and (ii) scaffold or fragment-based clustering. In each method, six clusters of novel virtual molecules were generated using unique but randomly selected scaffolds and functional groups (Singh et al. 2009). The molecules of the focused virtual library were prioritized based on standard properties such as toxicophoric, pharmacophoric and chemophoric (TPC), progressive drug-like (PDL), progressive lead-like (PLL), drug-like failure (DLF), and lead-like failure (LLF) (**Table 2.6**). The molecules which scored less T, DLF, LLF scores and more P score with stable and non-reactive C score with their respective PDL and PLL scores were selected for future drug development. Thus, the scaffolds were further subjected to an *in-silico* enumeration for the generation of focused virtual libraries to design novel molecules from known molecules using the already well-established algorithms and methods (Karthikeyan and Vyas 2014).

Table 2.6: Virtual Library (VL) with PDL, PLL and P, T, C, DLF, and LLF scores (for selected n = 6 molecules of each cluster from Plant-based clustering method)

Cluster	2D Structure of VL molecule	PDL ^a	PLL ^b	DLF	LLF	P	T	C
1		0.833	1.000	0	1	17	11	8
2		0.166	0.817	0	1	36	14	12
3		0.555	1.908	0	2	42	39	13

Cluster	2D Structure of VL molecule	PDL ^a	PLL ^b	DLF	LLF	P	T	C
4		0.544	2.129	0	2	56	32	19
5		0.500	1.477	0	1	44	20	12
6		0.278	1.000	0	1	42	19	11

^a: Progressive drug-like score; ^b: Progressive lead-like score; DLF: Drug Like Failure; LLF: Lead Like Failure; P: Pharmacophoric score; T: Toxicophoric score; C: Chemophoric score. For more PDL, PLL and P, T, C, DLF and LLF scores, please refer Table S6 and S7.

2.3.4 Cluster Analysis of Virtual Library

a. Plant-Based Clustering

In plant-based clustering, six clusters were constructed based on unique therapeutic properties of medicinal plants, i.e., analgesic, carminative/ laxative, anti-inflammatory, antimicrobial, diaphoretic/ antipyretic, and expectorant (**Supplementary Table S2.7.1**). Each cluster of novel virtual bioactive molecules was built with their extracted three to five scaffolds and functional groups. Hence, six clusters of virtual libraries were built representing each therapeutic property. The TPC fingerprints were generated for each virtual library (**Supplementary Table S2.7.2 – S2.7.7**) in the form of binary data (**Supplementary Table S2.7.8**). The total sum was calculated for all the molecular fingerprints belonging to their respective scores data, such as pharmacophore (n = 297), toxicophore (n = 209), and chemophore (n = 124) for all the clusters. Dendrograms were generated based on the sum calculation of fingerprints of all clusters for TPC scores by calculating the Euclidean distance matrix (**Figure 2.7** and **Supplementary Table S2.7.9**).

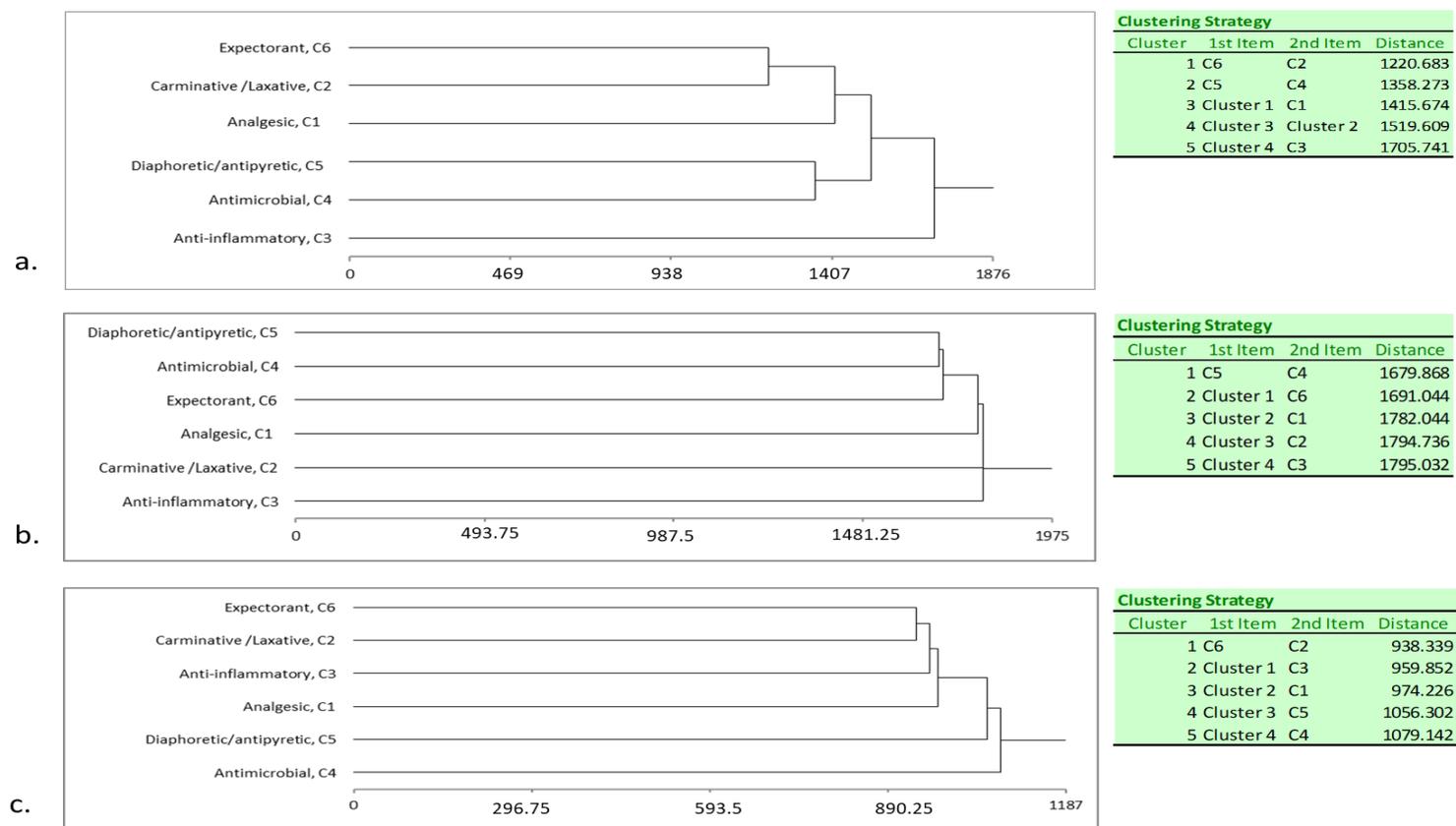


Figure 2.7: Dendrograms for toxicophoric (a), pharmacophoric (b), and chemophoric (c) fingerprints of virtual library molecules based on plant-based clustering (Distance/Similarity Measure = Euclidean Distance, Cluster Method = Nearest Neighbor) (Please refer to Supplementary Table S2.7.9 for Distance matrix (Euclidean distance))

The dendrograms for toxicophoric scores (**Figure 2.7a**) as well as chemophoric scores (**Figure 2.7c**) showed that cluster 2 and cluster 6 were the nearest neighbors, which means that the molecules of clusters 2 and 6 contain similar kind of chemical fragments, which make them more toxicophoric and chemophoric than pharmacophoric. The dendrogram for pharmacophoric scores (**Figure 2.7b**) showed that cluster 4 and cluster 5 had more similar molecules, which might be due to the similar kind of chemical fragments like scaffolds and functional groups in them. However, cluster 1 and cluster 3 contain several classes of molecules that are toxicophoric, chemophoric, and pharmacophoric substructures. The total sum of the fingerprint scores was converted to binary form, and absolute differences were calculated between each cluster of similar properties. The final result showed that all the molecules from the virtual library of the Indian medicinal plants had more pharmacophoric (n = 297) and less toxicophoric (n = 209) and chemophoric (n = 124) fingerprints.

b. Scaffold or Fragment-Based Clustering

In this method, bioactive molecules were divided into clusters (n = 122) based on their maximum common substructures. Among them, the top five clusters with the highest number of molecules were selected and six clusters were built with three to five scaffolds and functional groups (**Supplementary Table S2.8.1**) for generating the virtual library (**Supplementary Table S2.8.2- S2.8.7**). These molecules belong to various medicinal plants having different therapeutic properties. Further screening analysis is the same as mentioned in “Plant-based clustering”. The result showed that all the novel molecules from the virtual library of Indian medicinal plants had more pharmacophoric (n = 250) and less toxicophoric (n = 193) and chemophoric (n = 104) fingerprints (**Supplementary Table S2.8.8**).

This analysis revealed that the differences in the chemical nature of the molecules were due to their different fragment structures, which involved scaffolds and functional groups represented in the form of fingerprints. The TPC sum ratio in plant-based clustering was 1.69 : 2.40 : 1.00, whereas, for scaffold /fragment-based clustering, it was 1.86 : 2.40 : 1.00, which was slightly varied for the toxicophoric property. These results reveal those novel molecules generated from the bioactive molecules of plants having the same therapeutic properties show less toxicity than the molecules selected based on similar scaffolds. For example, if the virtual library is generated by combining molecules from plants with different therapeutic categories such as expectorant, antidiabetic, anticancer, etc., the toxicity of novel molecules would increase despite having a similar scaffold structure. Alternatively, if the virtual library is generated using the molecules identified from plants having similar therapeutic properties, the chances of getting toxic drug-like molecules would decrease. It has been suggested that natural templates for generating virtual libraries show greater biological relevance by specific distribution properties of natural compounds (Lee and Schneider 2001a). Therefore, it is necessary to carefully select the molecular scaffolds from the plants having similar therapeutic properties for designing novel drug-like molecules.

2.3.5 Applications of the Virtual Library

Using the classical drug discovery methods, it usually takes over ten years for a new drug from initial discovery to come into the market (Torjesen 2015). Further, the estimated cost of research and development of a successful drug is approximately 2.8 million USD, including the cost of thousands to millions of compounds that did not succeed (DiMasi, Grabowski and Hansen 2016). Thus, the probability of clinical success decreases to a very low value. Comparatively, modern drug discovery

employs virtual libraries and virtual screening to identify new potential drug candidates (Gordon et al. 1994). Using various computational tools, possibly an infinite number of diverse and novel molecular structures can be designed and screened virtually in a very short time (Blondelle, Perez-Paya and Houghten 1996). Several tools are now available to generate virtual libraries from molecules, including open-source (Truchon and Bayly 2006, Truszkowski et al. 2011, Schuller, Hahnke and Schneider 2007) and commercial software (Stevenson and Mulready 2003, Kochev 2017, Sud, Fahy and Subramaniam 2012, Buntrock 2002, Liao et al. 2005, Feuston et al. 2005, Leach et al. 1999, Yasri et al. 2004) using a multi-step process to enumerate the virtual library using scaffolds and functional groups extracted from molecules by combinatorial means.

Scaffold hopping or lead hopping is an alternate method for discovering structurally novel compounds by modifying the central core structure of the molecule (Sun, Tawa and Wallqvist 2012, Martin and Muchmore 2009). *In silico* screening based on various chemo- and bioinformatics approaches can be performed to identify potentially useful molecules from virtual libraries, which can be chemically synthesized and evaluated as potential drug candidates. Diverse molecular scaffolds provide diverse chemical accessibility by increasing the possibilities for drug and lead structure optimization (Abel et al. 2002). Such combinatorial techniques provide new possibilities to pharmaceutical industries by providing an enormous number of molecules for *in silico* screening, increasing the potential of discovering new drug candidates quickly, instead of starting with only a few chemically synthesized or natural bioactive compounds for screening. Hence, the *in-silico* enumerated focused virtual library that we developed using the open-source software “ChemScreener”, integrated with “Scaffolder” (a program to identify the biologically relevant

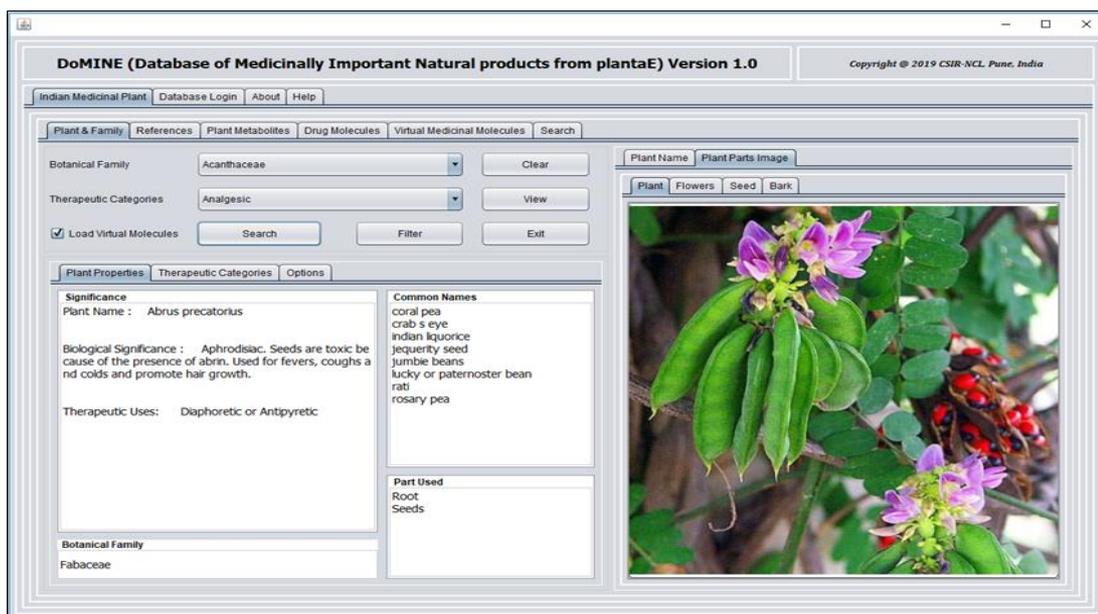
molecular scaffolds from the list of secondary metabolites identified and listed from Indian Medicinal Plants), can serve as a powerful resource to screen the molecules as candidate drugs for diverse diseases (Karthikeyan and Vyas 2015). This can help pharmaceutical industries identify new potential drugs from the Indian medicinal and aromatic plants quickly and efficiently and reduce the time required to bring a new drug into the market.

2.3.6 DoMINE

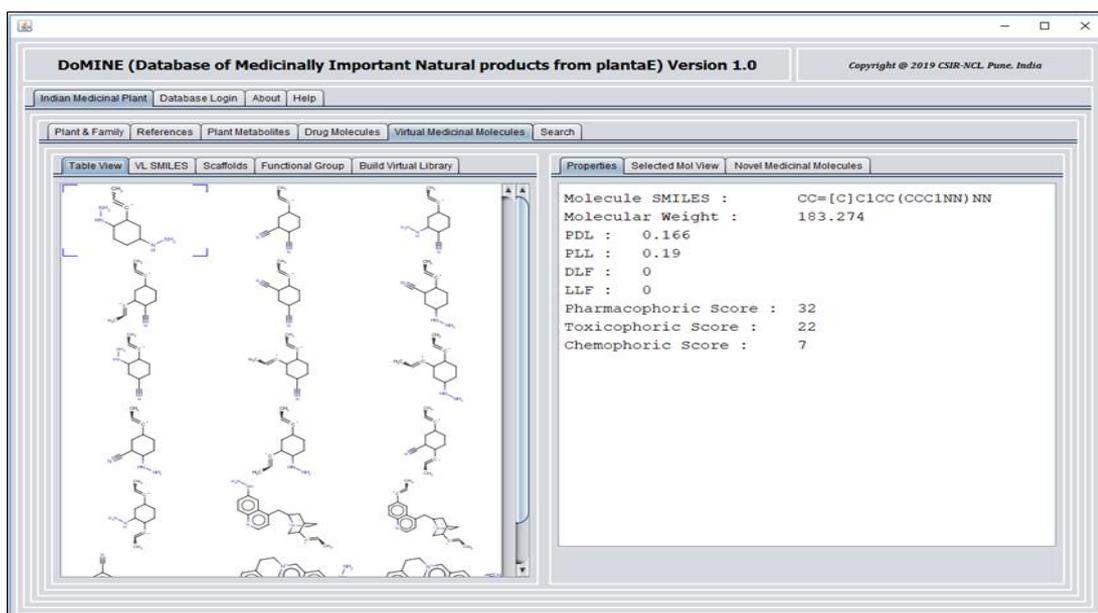
A chemoinformatics-oriented database of Indian medicinal plants and their natural metabolites is the single most productive source of leads for drug development. This encouraged us to embark on the project of identifying and cataloguing natural compounds reported in the literature, thereby connecting species, molecules, and diseases. A vast amount of data about the species, chemicals, and drugs was collected, filtered, and used to extract the relevant information from up-to-date literature. Several chemoinformatics tools, databases, and techniques have previously been employed to achieve similar results. However, they do not include designing novel virtual molecules from the known Indian medicinal plant molecules (Polur et al. 2011, Pathania et al. 2015, Mohanraj et al. 2018). Hence, we developed a chemoinformatics open-source toolkit DoMINE (**D**atabase of **M**edicinally **I**mportant **N**atural products from **p**lanta**E**), using Java. It can be used to build and access the Indian medicinal plant database created in this study and generate a scaffold and virtual library (**Figure 2.8**). The program comprises a curated database of chemical information gathered from various Indian traditional plants. The database consists of the data relating to the plant species, chemical compounds, their molecular properties, scaffolds, drug molecules, diseases, biological significance, therapeutic uses, plant images (plant,

flower, seed and bark) and relevant PubMed references. DoMINE allows the user to access the relevant data fields easily by writing a structured query.

A huge network was built showing the relationships between the plant species, chemical compounds, scaffolds, drugs, disease and therapeutic use. The database is compatible with chemoinformatics oriented sub-structure, exact structure and similar-structure queries to retrieve the details of chemical structure relevant to the particular medicinal plants with active chemical ingredients along with their therapeutic importance. The program also supports adding details of new plant species, as well as updating the existing details. Currently, we have constructed the database with 104 Indian medicinal plants possessing therapeutic properties that were used for scaffold generation and building novel bioactive molecules and subsequently predicting TPC scores for prioritizing molecules in the virtual screening process.



A.



B.

Figure 2.8: The DoMINE cheminformatics toolkit. A. DoMINE showing the medicinal plant species *Abrus precatorius* with its therapeutic properties. B. DoMINE showing virtual molecules built from Indian medicinal plant molecules with TPC scores.

2.4 Conclusions

In the present study, we developed a simple, quick, and cost-effective computational protocol for generating novel potential drug candidates from the bioactive molecules of Indian medicinal and aromatic plants through a chemoinformatics approach. We also developed the DoMINE toolkit for the advancement of natural product-based drug discovery through chemoinformatics approaches. This study will be useful in developing new drug molecules from the known medicinal plant molecules. Hence, this work will encourage experimental organic chemists to synthesize these molecules prioritized based on the predicted values. These synthesized molecules need to be subjected to biological screening to identify potential molecules for drug discovery research.

CHAPTER 3

**BRIDGING IN-SILICO AND
EXPERIMENTAL:**

**CHEMOINFORMATICS
INVESTIGATION**

**FOR MASS SPECTROMETRY-BASED
METABOLOMICS STUDY OF
SOYBEAN**

Chapter 3: Bridging In-Silico and Experimental: Chemoinformatics Analysis for Mass Spectrometry-Based Metabolomics study of Soybean

3.1 Introduction

Soybean (*Glycine max* L. Merr.) has global importance and is one of the most widely cultivated legumes worldwide. It is rich in seed protein (~36%) and oil content (~20%) (<https://en.wikipedia.org/wiki/Soybean>) and is used for both human and animal consumption as well as for industrial purposes. It is also used in folk, Chinese, and modern medicine (<http://envis.frlht.org/implad>). Considering its significance, the Indian Council of Agricultural Research (ICAR) established an All India Coordinated Research Project on Soybean in 1967, with its headquarters initially in New Delhi and subsequently in Pantnagar. ICAR-Indian Institute of Soybean Research was established at Indore (M.P.), which is a premier soybean research institute that has developed and maintained various soybean germplasm accessions for high oil content, high oleic acid content, high protein, bold seeds, good germinability, rust resistance, YMV tolerance, and other characteristics.

The major secondary metabolites in soybean are: (i) phytic acid (1.0 - 2.2%), (ii) sterols (0.23-0.46%), (iii) saponins (0.17-6.16%), (iv) isoflavones (0.10-0.30%), and (v) lignans (0.02%), which play diverse and indispensable roles in plant development, reproduction, defense, etc. (Kang et al. 2010). Soybean has a favorable nutrient profile for heart health, decreasing the negative effects of menopause and reducing the risks for cancer, paralysis, diabetes, kidney diseases, allergies (Choudhary and Tran 2011), etc. Genistein and daidzein are the most potent

antioxidants among the soy isoflavones. Genistein plays a vital role in immunity and inhibits allergic inflammatory responses (Sakai and Kogiso 2008). Likewise, several other soybean metabolites such as isoflavones, terpenoids, alkaloids, etc., have therapeutic properties against several chronic diseases such as cancer, cardiovascular, obesity, and osteoporosis (Munro et al. 2003, Barnes et al. 1996, Balandrin, Kinghorn and Farnsworth 1993).

Developing drugs from plant-derived compounds is being practiced for centuries (Potterat and Hamburger 2008). However, in modern drug discovery, *in-silico* pharmacology tools such as building virtual libraries from natural products, virtual screening, predicting properties based on 2D and 3D molecular structures, etc., are routinely employed (Rollinger, Stuppner and Langer 2008, Lavecchia and Di Giovanni 2013, Cheng et al. 2012). These tools can virtually screen a large number of molecules in a short time and are highly efficient in identifying potential drugs; saving time, energy, and cost. They can effectively filter out the molecules not having drug-like or lead-like properties and can provide precise candidates to the pharmaceutical industry (Reddy et al. 2007). The integration of experimental and computational technologies is a powerful tool for drug development (Yu and Adedoyin 2003, Chaturvedi, Decker and Odinecs 2001). Mass spectrometry, coupled with the liquid chromatography (LC-MS) method of metabolite screening, plays an essential role in drug development (Rossi and Sinz 2001). However, there are very few reports on LC-MS analyses in soybean samples to identify the medicinally important compounds. Most of the reports are focused on soybean root and seed tissue samples for targeted mass spectrometric quantification (Wu et al. 2008, Gu et al. 2017, Brechenmacher et al. 2010, Griffith and Collison 2001).

To our knowledge, this is the first study integrating chemoinformatics with LC-MS analysis of metabolites for drug designing and development. A comprehensive understanding of the metabolite profile of plants is essential for assessing their medicinal values. We have previously identified medicinally important molecules from Indian medicinal and aromatic plants (Karade et al. 2020). Several databases related to soybean have been previously reported, which include SoyMetDB (Joshi et al. 2010), SoyCyc (<http://soybase.org/soycyc>), and Soybean Knowledge Base (SoyKB) (Joshi et al. 2014), which were developed explicitly for genomics, transcriptomics, proteomics and metabolomics data analyses. This inspired us to study and analyze soybean metabolomics data for drug discovery research. This chapter focuses on designing drug-like and lead-like molecules based on chemoinformatics and UHPLC-MS/MS analysis of secondary metabolites of soybean.

For this purpose, soybean small molecules (n=1622) from SoyKB, SoyCyc, as well as those text mined from PubMed, and the FDA-approved drugs (n=2354) from DrugBank (Wishart et al. 2017), were used to extract corresponding molecular scaffolds. UHPLC-MS/MS analyses were performed to detect and validate the known and unknown molecules from four soybean varieties (NRC-119, JS-335, JS-7105, and JS-9305). Scaffolds obtained from drugs and previously reported soybean small molecules were compared by combining annotated mass features of small organic molecules detected by UHPLC-MS/MS analyses (n = 7185; including reported and unreported molecules) to reveal common scaffolds among them. Later, scaffolds of the previously reported molecules were supplied with linkers and functional groups to enumerate diverse virtual libraries. The new virtual molecules were prioritized by annotation with drug-like and lead-like scores, and (n = 523) potential drug-like

molecules were identified. In summary, this research will be useful in the discovery of natural products-based drugs.

3.2 Materials and methods

In the present study, soybean was considered a test case. We implemented chemoinformatics and UHPLC-MS/MS-assisted approaches to identify and analyze small organic molecules to design a virtual library of the novel drug-like and lead-like molecules. **Figure 3.1** presents an overview of the steps deployed in the present methodology.

3.2.1 Chemoinformatics analysis

a. Data collection

Soybean is endowed with several diverse forms of phytochemicals having pharmaceutical properties (Salim, Chin and Kinghorn 2008, Nuraini, Rahayu and Rifai 2019, Shlyankevich 1995). A list of soybean small molecules (n=1622) was prepared from the databases such as SoyKb ([SoyKB: Soybean Knowledge Base - Metabolite Search](#) -), SoyCyc (<https://pmn.plantcyc.org/SOY/class-tree?object=Compounds#>) and by text mining (keywords used: ‘soybean’, ‘*Glycine max*’; Mar 2020) the PubMed literature (<https://www.ncbi.nlm.nih.gov/pubmed/>) (**Supplementary Table S3.1.1**). Structural information about FDA-approved drugs (n=2354) was downloaded from DrugBank (Law et al. 2014) (**Supplementary Table S3.1.2**).

b. Chemoinformatics tools

The data about soybean-related small molecules were extracted by text mining the PubMed abstracts (<https://www.ncbi.nlm.nih.gov/pubmed/>) (n = 47,541) using PubTator (Wei et al. 2013) (a web-based text mining tool for recognizing the term as chemical). The chemical names text mined from PubMed literature were manually verified against their PMID numbers for their presence in soybean. Chemical names

were converted into SMILES (Simplified Molecular Input Line Entry System) strings using the JChem-Base ChemAxon tool (Weber 2008) and screened for 5-6 membered rings containing molecules up to 1000 molecular weight. All the data were converted to SDF format for easy access in Molecular Operating Environment (MOE) (Chemical Computing Group 2008) and the in-house developed ChemScreener program (Karthikeyan et al. 2015b, Karthikeyan and Vyas 2014). Descriptor generation and analysis were performed using MOE, while ChemScreener was used to generate scaffolds and functional groups that were further enumerated with drug-like and lead-like properties. Cytoscape (Shannon et al. 2003) was used to view and analyze the network of the soybean molecules, drugs, and their scaffolds, respectively.

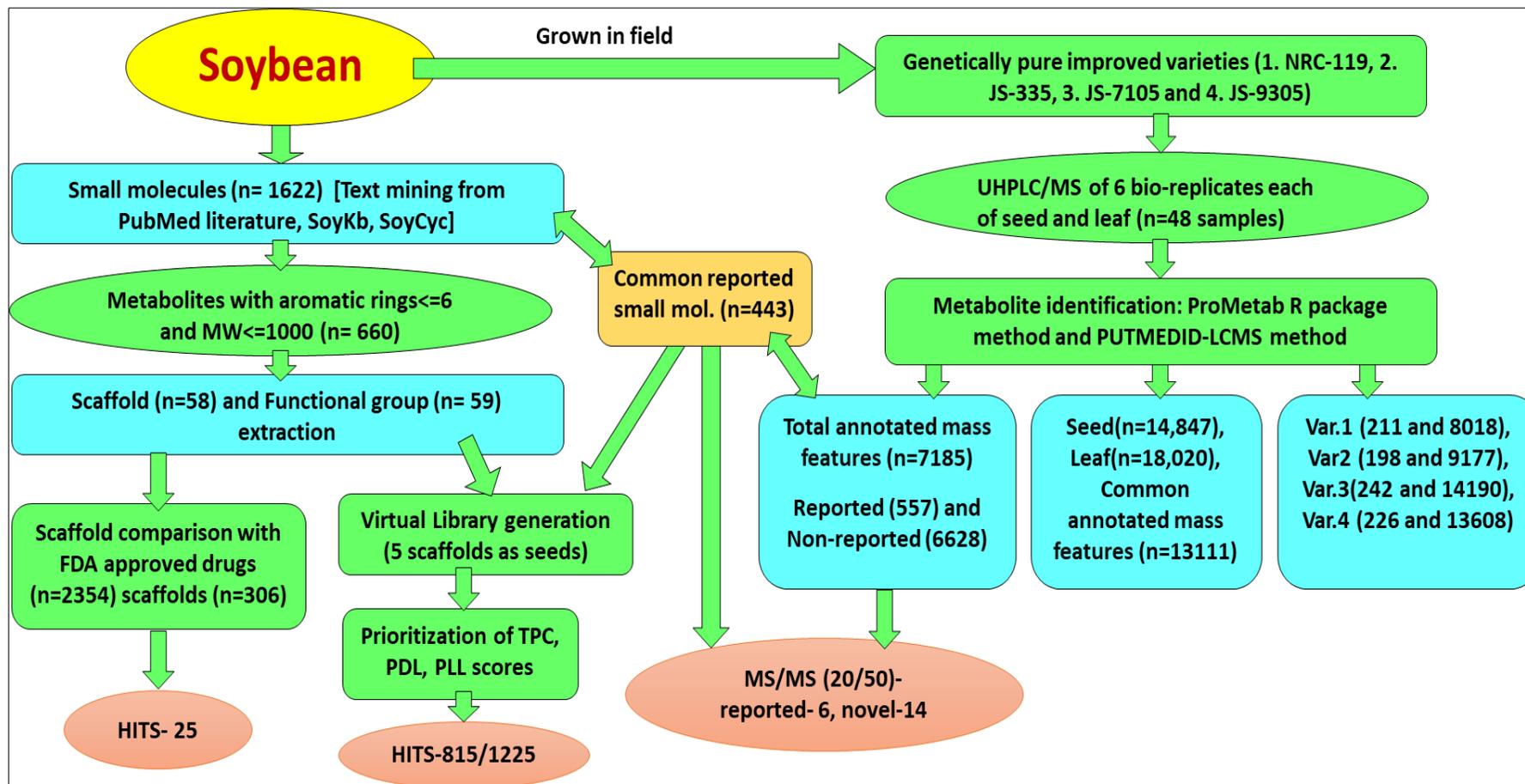


Figure 3.1: An overview of the analytical steps deployed in the present study

3.2.2 Metabolomics analysis

a. Plant material

Four soybean varieties (NRC-119, JS-335, JS-7105, and JS-9305) from the germplasm collection of the Indian Institute of Soybean Research (IISR), Indore, India, were used for this study. These varieties are resistant to lodging, shattering, stem fly, bold-seeds, and good germinability. Detailed information about the varieties is presented in **Table 3.1**. The four varieties were sown in a randomized block design with three replications during June 2017 in plots at CSIR-National Chemical Laboratory, Pune, India. A spacing of 30 cm between rows and 5 cm between plants was maintained, and two seeds per hill were sown. Leaf samples were collected before the flowering stage, while the seeds were harvested at maturity, from two randomly selected plants of each variety in each replication. Both leaf and seed samples were ground in liquid nitrogen comprising six biological replicates for each of the four soybean varieties. Metabolites from 100 mg of the crushed seed and leaf tissue were extracted with 350 ml of 70% ice-cold methanol followed by sonication for 20 min and centrifugation at 4°C at 10,000 rpm for 20 min. The supernatant was syringe filtered with a 0.22 µm nylon filter (Chromatopak, India) and stored at -80°C until further use.

Table 3.1: Morpho-physiological characteristics of the four soybean varieties used for the study

Sr No	Character	NRC119	JS335	JS7105	JS9305
1	Growth type	Semi-determinate	Semi-determinate	Determinate	Semi-determinate
2	Flower color	White	Purple	Purple	Purple
3	Seed size	Bold	Small	Small	Small
4	Seed color	Yellow	Yellow	Yellow	Yellow
5	Hilum	Light black to Black	Black	Light black to Black	Black
6	Seed longevity	-N.A.-	High	Poor	High
7	Resistance to lodging	-N.A.-	No	Yes	No
8	Resistance to shattering	Yes	Yes	No	Yes
9	Tolerant to stem fly	Yes	Yes	Yes	Yes
10	Resistance to bacterial pustule	Yes	Yes	Yes	Yes
11	Resistant to Yellow Mosaic Virus (YMV)	Yes	No	Yes	Yes
12	Resistance to Myrothecium Leaf Spot	Highly resistant	Moderately resistant	Moderately susceptible	Highly resistant

-N.A.-: Data not available

b. UHPLC-MS/MS profiling and analysis

The Accela™ ultra-high-performance liquid chromatography (UHPLC) system (Thermo Fisher Scientific, USA) was operated using Xcalibur Ver. 2.0 (Thermo

Fisher Scientific, USA) software platform. It was coupled online via a heated electrospray ionization source (HESI) with a Q-Exactive-Orbitrap mass spectrometer, which was employed for non-targeted metabolomics profiling. The sample injection volume was 3 μl , and the metabolites were profiled using a reverse-phase UHPLC C18 column: 2.1 \times 150 mm, 1.9 μm i.d., and Accela 1250 pump. The column oven temperature was set at 40°C, and the sample manager was maintained at 4°C. The mobile phase consisted of solvent A (water containing 0.1% formic acid) and B (acetonitrile (ACN) containing 0.1% formic acid) employed both in electrospray ionization positive (ESI(+)) and negative (ESI(-)) polar modes. The flow rate was 350 $\mu\text{l}/\text{min}$ with a linear gradient elution over 15 min. From the start to 0.3 min, eluent A was held at 2%, linearly increased to 45% till 10 min and then to 98% in 13 min. Subsequently, eluent B was returned to 2% in 13 min and held for an additional 1.3 min before returning to the initial conditions. The sample sequence was random. In the ESI(+) and ESI(-) modes, the MS spray voltage was 4.2 and 3.6 kV, respectively. The capillary temperature was set at 320°C with the sheath gas at 45 arbitrary units and the aux gas at 12 arbitrary units. The tube lens was set to 50 V, and the mass spectra were recorded over the range 81.034-999.5043 m/z. The resolution of the Orbitrap was set at 70,000. The tandem mass spectrometry (MS/MS or MS²) data were collected with the collision energy between 10 and 35 eV.

c. Data processing and analysis

Metabolomics data handling tasks were divided into two steps, i.e., data processing and data analysis (Katajamaa and Oresic 2007). LC-MS raw data files (n=48; 4 soybean varieties \times 6 replicates \times 2 tissues) were converted to .mzxml formats using the MSConvert module of ProteoWizard Ver. 3.0.10922 (Holman, Tabb and Mallick 2014). The files were then analyzed using two methods (**Figure 3.2**): (i) XCMS

online used with PUTMEDID-LCMS tool (Brown et al. 2011, Tautenhahn et al. 2012) and (ii) ProbMetab (Silva et al. 2014), an R package. XCMS online is a web-based platform for processing untargeted metabolomics data. PUTMEDID-LCMS is a tool operating in Taverna workflow to identify metabolites from accurate molecular mass data acquired in LC-MS studies (<http://www.mcisb.org/resources/putmedid.html>). ProbMetab is an R package based on the Naive Bayes machine learning method for probabilistic annotation of compounds. ProbMetab was also used to incorporate the information about possible biochemical pathways of identified molecules from LC-MS experiments by comparing with the Kyoto Encyclopedia of Genes and Genomes (KEGG) soybean pathway. Statistical analysis was performed using MetaboAnalyst 4.0 (Chong, Wishart and Xia 2019) online, which is a suite of tools for metabolomics analysis of MS data.

In the first method of preprocessing and identification of metabolites, the processed raw data of the positive and negative ionization modes were uploaded in XCMS Online (<https://xcmsonline.scripps.edu/>) for calculating ANOVA. The analysis was conducted using the following parameters in a custom-designed R script: (i) General parameters: polarity = positive/ negative, retention time format = minutes (for statistical analysis in MetaboAnalyst) or seconds (for putative metabolic identification in PUTMEDID Taverna workflow); (ii) Feature detection: centWave method, ppm = 2.5 min. and max peak width = 5 and 20, S/N threshold = 10, mzdifff = 0.01, integration method = 1, prefilter peaks = 3, prefilter intensity = 5000, Noise filter = 1000; (iii) Retention time correction: Obiwarp method, profStep = 1; (iv) Alignment: mzwid = 0.015, minfrac = 0.5, bw = 5, max = 100, minsamp = 1; (v) Statistics: statistical test = ANOVA parametric test, p-value threshold = 0.05, fold

change threshold = 1.5; (vi) Annotation: search for = isotopes and adducts, m/z absolute error = 0.015, ppm error = 5; (vii) Identification: ppm tolerance = 2, adducts (+ve) = [M+H]⁺, [M+Na]⁺, sample biosource = soybean (biocyc), pathway ppm deviation = 5; and (viii) Visualization: EIC width = 100. The obtained peak lists were normalized for multivariate statistical analysis using MetaboAnalyst. ANOVA, principal component analysis (PCA), and partial least squares discriminant analysis (PLS-DA) applied after Pareto scaling were evaluated for sample discrimination in data scaling for normalization.

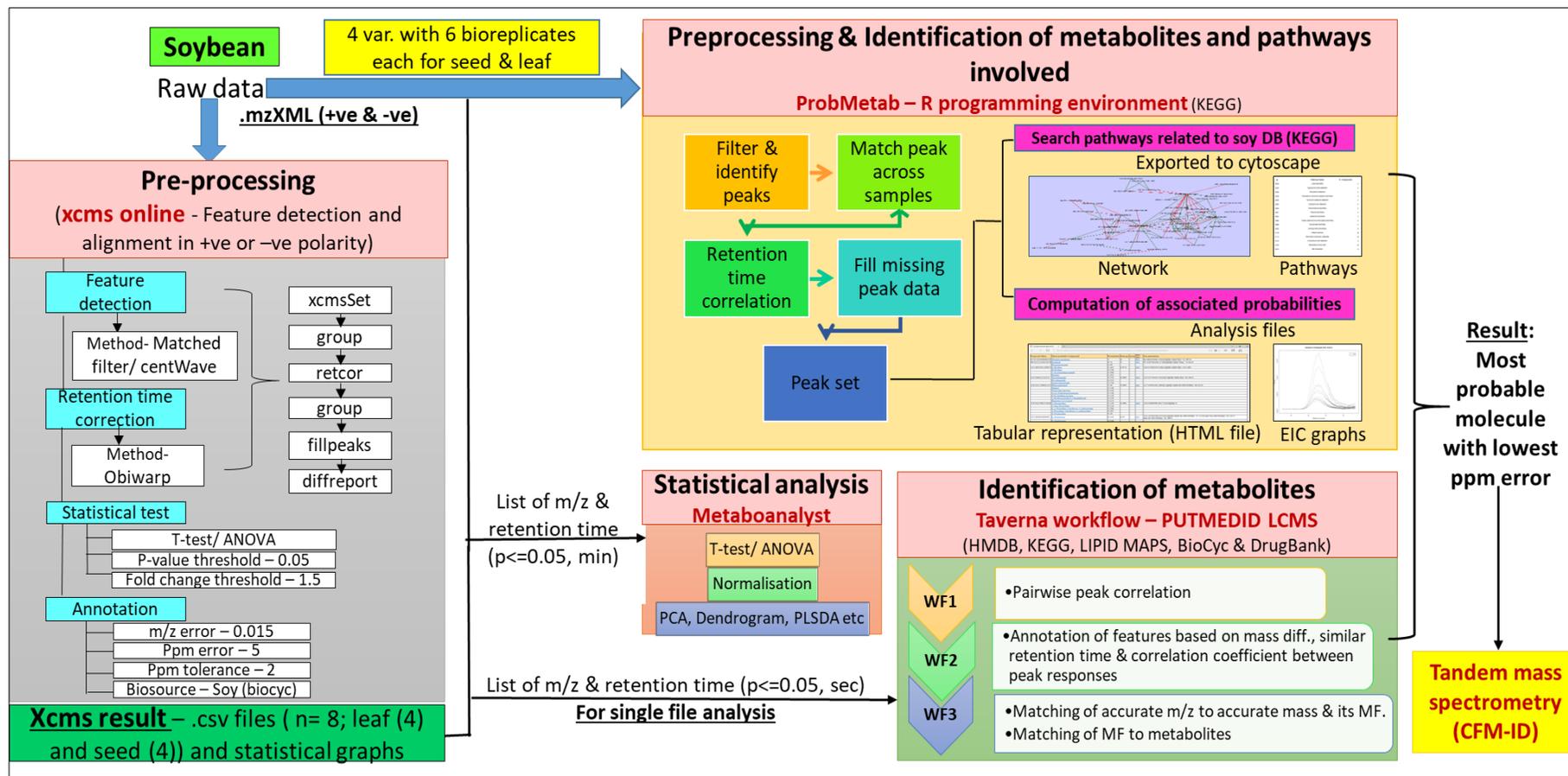


Figure 3.2: Workflow highlighting UHPLC-MS/MS data analysis grouped into three categories: preprocessing, metabolite identification and statistical analysis

The LC-MS data were putatively feature annotated using PUTMEDID operated in the Taverna workflow environment. The accurate mass for each peak of the experimentally determined matching mass was assigned by a single or multiple molecular formulae. The LC-MS metabolites matching the molecular formula in the Manchester Metabolomics Database (Brown et al. 2009) were compared with a mass error of less than ± 5 ppm. In the second method of spectral annotation, ProbMetab was used for probabilistic annotation of compounds. It is based on matching spectra or exact masses from unknown compounds against the spectral data deposited in the KEGG database related to soybean. The annotation in a network is displayed in a visualization scheme exported to Cytoscape Ver. 2.8.1, wherein observed mass peaks are connected if their candidate metabolites are substrate/product of known biochemical reactions.

Tandem mass spectrometry was also performed to confirm the presence of the annotated metabolites in the soybean samples using Competitive Fragmentation Modeling for Metabolite Identification (CFM-ID) (Allen et al. 2014) webserver. The MS/MS sample was prepared by pooling 10 μ L of the sample solution, each from the 48 samples of soybean varieties. Input data of small organic molecules (n=50) for MS/MS run was prepared according to their intensity, probability score, and ppm error analyzed by ProbMetab and PUTMEDID-LCMS methods. Fragmented molecules were identified from KEGG and HMDB (using exact mass and MS/MS fragmentation patterns) databases using CFM-ID under the compound identification section. Venn diagrams were created using Venny Ver. 2.1 (Oliveros 2015) and the heat maps were drawn using the ggplot2 R package (Wickham 2011).

3.3 Results and Discussion

3.3.1 Chemoinformatics analysis of soybean phytochemicals

In this study, we performed chemoinformatics analysis to identify drug-like and lead-like molecules from soybean and develop a virtual library of prioritized novel and promising drug candidate molecules. For this purpose, molecular structures of soybean plant metabolites and drugs were identified and downloaded from databases such as SoyKb, SoyCyc, DrugBank, as well as through a literature survey by text mining. We also performed mass spectrometry analysis of four Indian varieties of soybean to confirm the presence of the molecules in soybean seed and leaf tissues. LC-MS data were also used for multivariate analysis.

a. Descriptor space of soybean small organic molecules

The chemical space of the soybean metabolites was analyzed by computing the descriptors implemented in MOE for the soybean small molecules and drugs. Chemical names of the extracted soybean molecules and drugs were converted into SMILES strings and screened for 5-6 membered rings containing molecules up to 1000 molecular weight. A total of 186 2D descriptors were computed for all the soybean small molecules and drugs (**Supplementary Table S3.1.3, S3.1.4**). All these descriptors follow the “Lipinski’s rule of five” (Lipinski et al. 1997), which assesses the biological activities of orally active drugs. The 2D calculated descriptors encode physical properties such as molecular weight, molecular mass density, log S and log P, number of rings, number of rotatable bonds, and the number of hydrogen bond donors and acceptors. The ranges of descriptors for soybean small molecules were computed and compared concerning the generally accepted ranges for the properties of drug-like and lead-like molecules. The histograms in **Figure 3.3** illustrate the

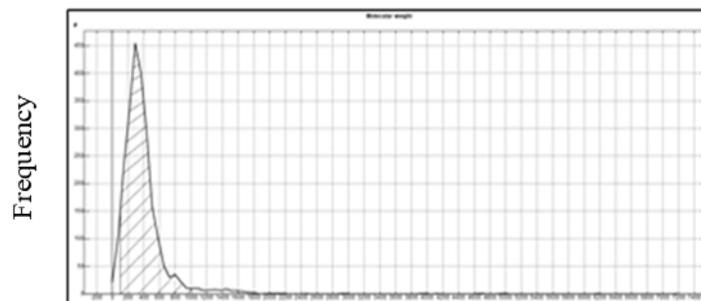
respective ranges obtained for the soybean small molecules and drug molecules. The complexity and diversity were also assessed in the molecules from soybean that possesses drug-like and lead-like features.

The molecular weight of the identified soybean small molecules ranged from 100 – 660. The generally acceptable range for drugs is 100 – 1000, indicating that several soybean molecules possessing medicinal properties might be suitable as drug molecules. The Oprea (Oprea 2000) number of rotatable bonds for soybean phytochemicals ranged from 1 to 16, which falls between the drug range, i.e., 1 to 20. The diversity and complexity of the soybean phytochemicals were ascertained by computing parameters such as molecular mass density (0.66 – 1), number of hydrogen bond acceptors (1 – 10), number of hydrogen bond donor (1 – 7), the log of the aqueous solubility (mol/l), i.e., log S (-5 – 1), the log of the octanol/ water partition coefficient, i.e., log P, (-2 – 8.5), number of rings (1 – 6), etc., which were closer to the corresponding drug range values. For almost all the drug molecules, the number of rings was between 1 and 8. It was found that for hydrogen bond donors, almost all the soybean phytochemicals were located in the acceptable range of drug molecules. These descriptors emphasize the fact that soybean molecules are complex and diverse and possess drug-like and lead-like features, which can be further fine-tuned to develop potential drugs. All these filtered soybean small molecules (n=660) were used to extract scaffolds (n=58) and functional groups (n=59) using ChemScreener (**Supplementary Table S3.1.5**). Scaffolds (n= 306) and functional groups (n= 291) were also extracted from all the approved drug molecules (**Supplementary Table S3.1.6**).

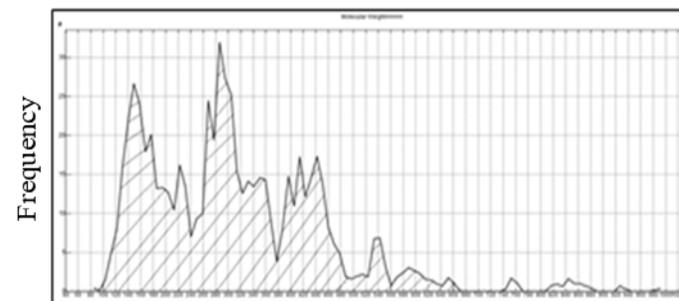
Drugs

Soybean

1. Molecular Weight (Atomic mass unit with atomic weights)

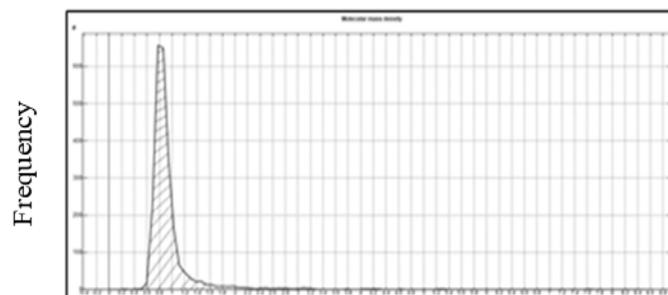


Molecular weight: 100 - 1000

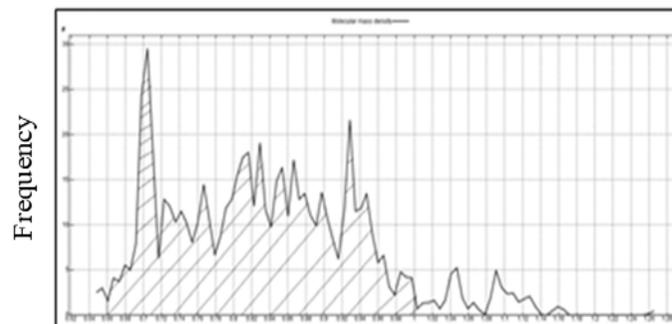


Molecular weight: 100- 660

2. Molecular mass density ($\text{amu}/\text{\AA}^3$)



Density: 0.6 - 1.6

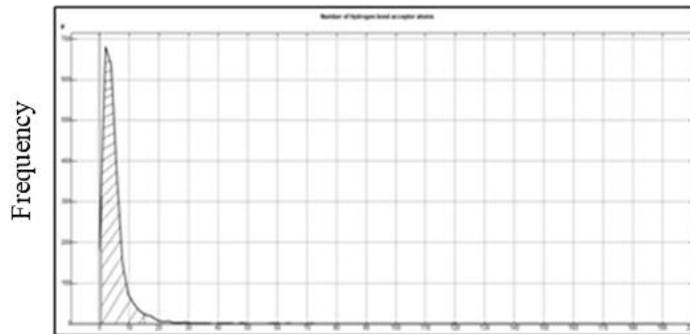


Density: 0.66- 1

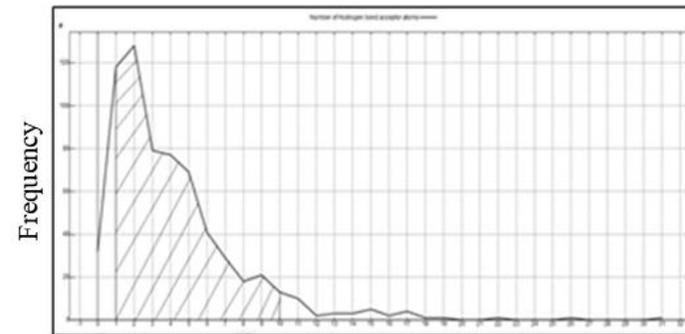
Drugs

Soybean

3. Number of Hydrogen bond acceptor atoms

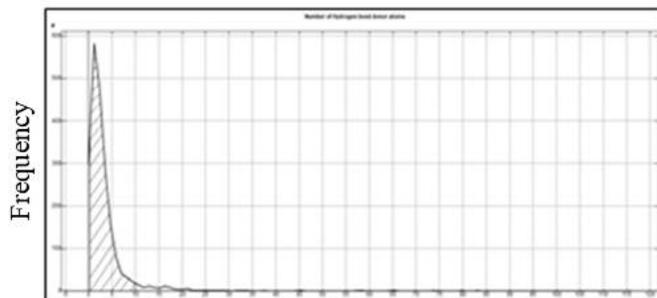


Hydrogen bond acceptors: 1 - 15

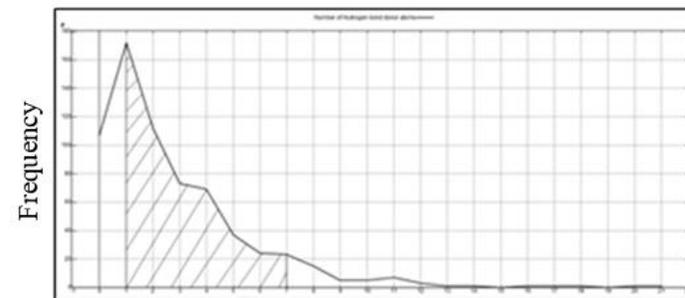


Hydrogen bond acceptors: 1 - 10

4. Number of Hydrogen bond donor atoms



Hydrogen bond donors : 1 - 10

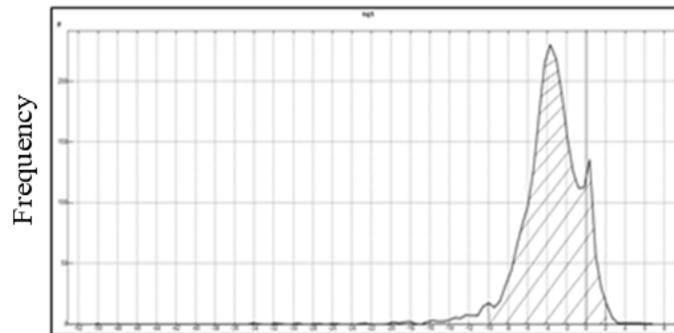


Hydrogen bond donors : 1 - 7

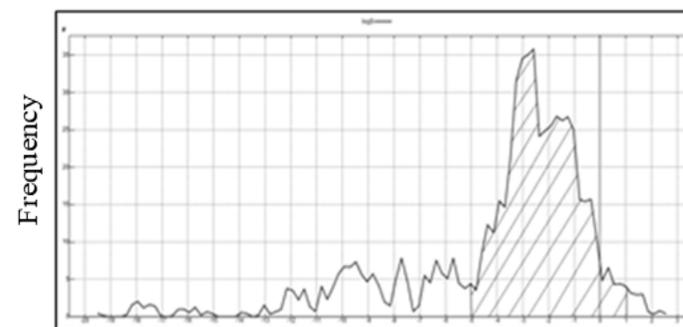
Drugs

Soybean

5. Log of the aqueous solubility; Log S (mol/l)

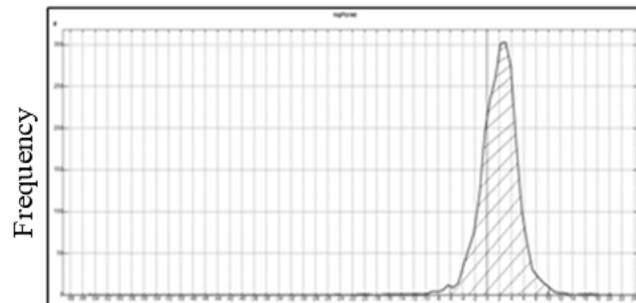


Log S: -10 - 2

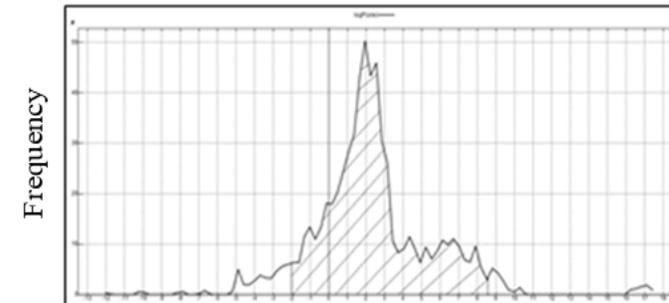


Log S: -5 - 1

6. Log of the octanol/ water partition coefficient; Log P (o/w)



Log P: -6 - 10



Log P: -2 - 8.5

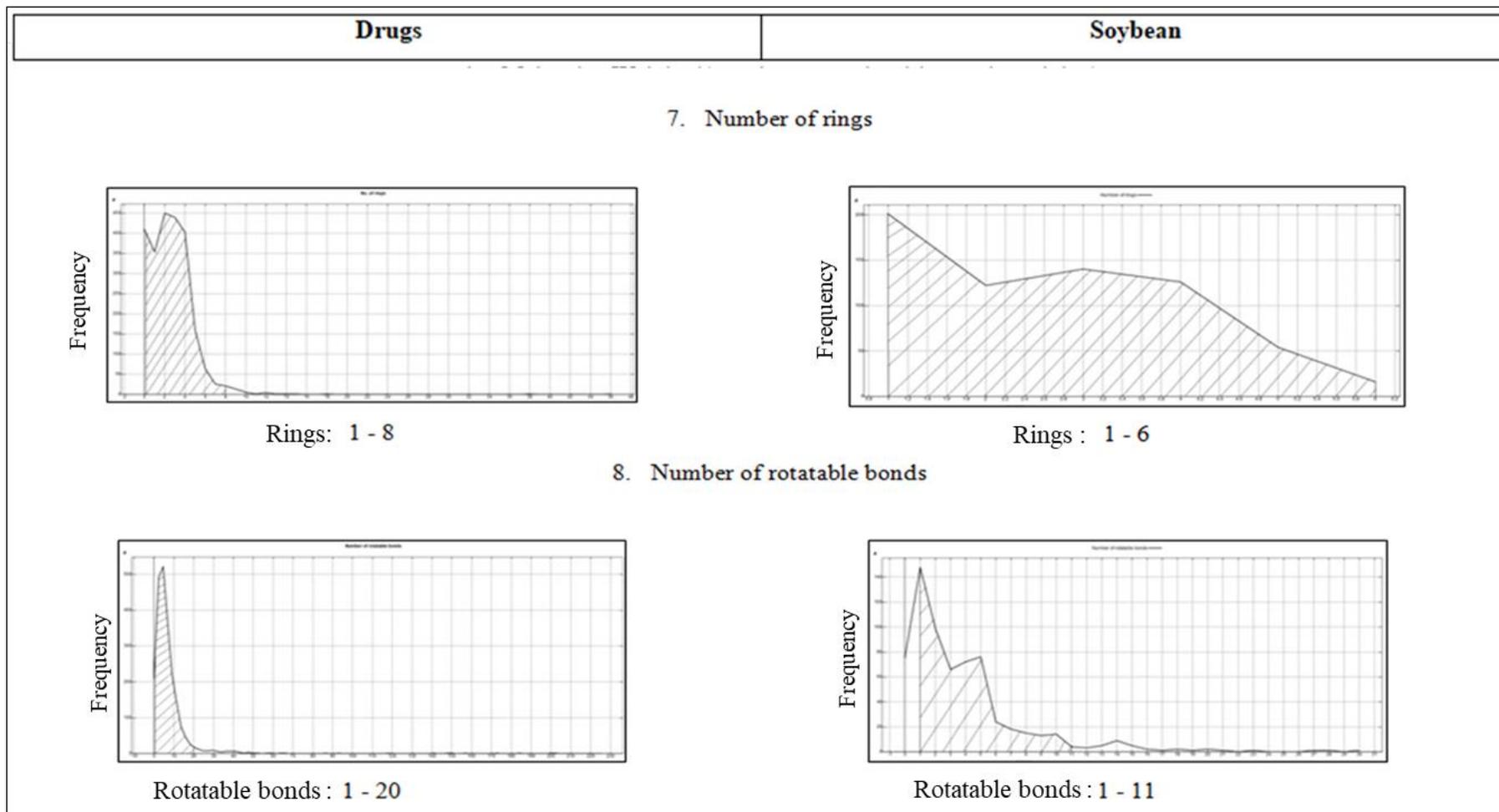


Figure 3.3: Histograms depicting descriptor ranges of soybean small molecules and approved drug

3.3.2 Metabolomics profiling

Chemical class identification and metabolic profiling using non-targeted LC-MS-based analysis are well-established. The untargeted UHPLC-MS analysis was performed to identify the metabolites in seed and leaf tissues of soybean varieties. All the data were processed systematically to MS Level 2 identification of metabolites (Acevska et al. 2015). The data were preprocessed using XCMS, which included feature detection, retention time correction, and alignment of metabolites. The XCMS provides a table of m/z values, retention time, p-value, and folds change for each feature, along with the integrated feature intensities from all aligned samples (**Supplementary Table S3.2**). The cloud plot generated by XCMS for positive ion mode showed 8810 significant features with p-value ≤ 0.05 (**Figure 3.4**), and for negative ion mode, 7488 significant features with p-value ≤ 0.05 (**Figure 3.5**). These results of m/z values and retention time were used as inputs for quantitative and qualitative data analysis. For quantitative analysis, ProbMetab and PUTMEDID-LCMS methods were used to identify the putative metabolites by spectral matching. For qualitative statistical analysis, multivariate approaches were used via MetaboAnalyst online tools.

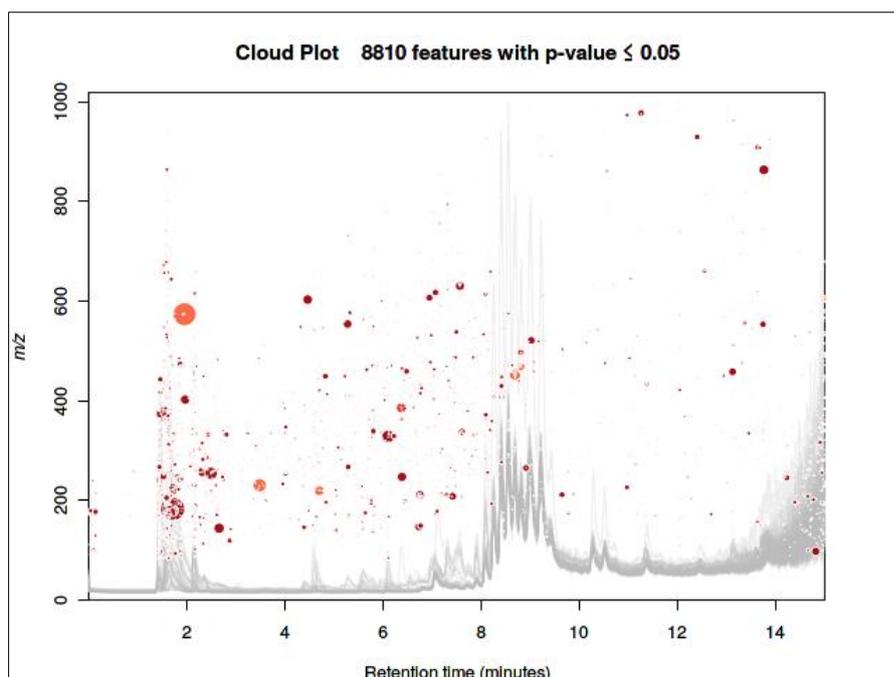


Figure 3.4: Cloud plot generated by XCMS for positive ion mode

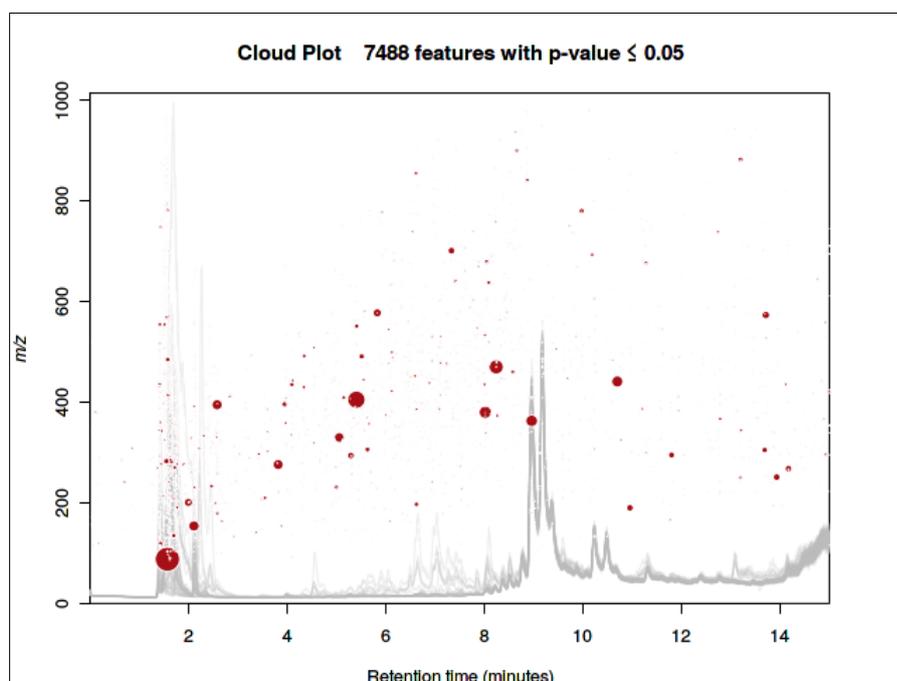
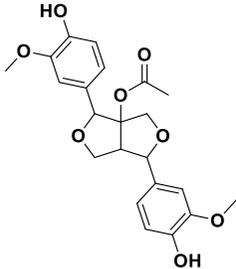
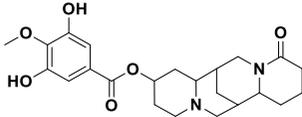
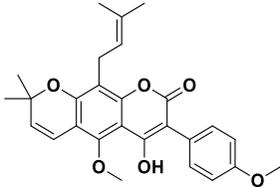
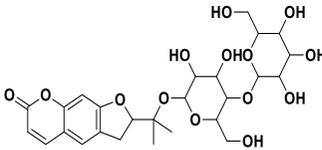


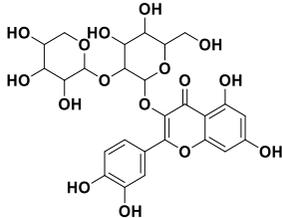
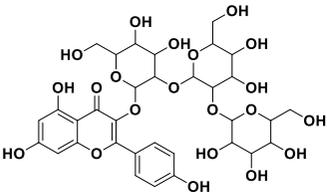
Figure 3.5: Cloud plot generated by XCMS for negative ion mode

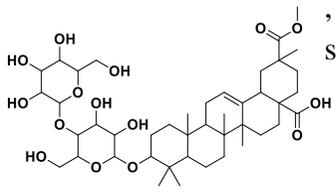
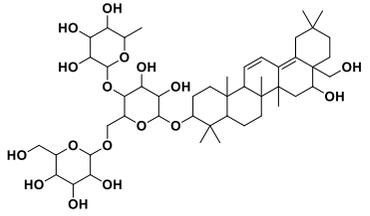
The soybean metabolites were identified using two methods, PUTMEDID-LCMS workflow operating in the Taverna environment and ProbMetab (**Supplementary Table S3.3.1**). The metabolites were identified by PUTMEDID-LCMS workflow operated in the Taverna environment by correlation analysis after annotating adducts, isotopes, dimers and fragments. However, isomers were detected with the same mass and retention time and hence they were difficult to be differentiated. Therefore, all the identified isomers have been listed in the analysis report. ProbMetab has the advantage of combining positive and negative polar mode LC-MS data files, which cannot be done by PUTMEDID-LCMS workflow. Hence, single result files were generated for seed and leaf tissues of the four soybean varieties. The molecules were prioritized based on ppm error identified from PUTMEDID-LCMS workflow and probability scores from ProbMetab. **Table 3.2** depicts the metabolites identified from soybean seed and leaf tissues by PUTMEDID-LCMS and ProbMetab methods in positive and negative ion modes. UHPLC-MS analyses confirmed the presence of small molecules identified by text mining and soybean databases in the four soybean varieties. A Venn diagram was created to depict differences between the numbers of detected metabolites in soybean seed and leaf tissues as identified by ProbMetab (for probability score: 1) and PUTMEDID-LCMS (up to 2 ppm error) (**Figure 3.6**). Leaf (n= 18,020; 24.8%) had high metabolic content than seeds (n=14,847; 8.8%), and 66.5% (n=13,111) of the metabolites were common between the leaf and seed tissues (**Supplementary Table S3.3.2**).

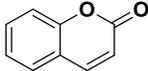
Table 3.2: Small organic molecules (n=20) validated through tandem mass spectrometry by performing in silico fragmentation approach (CFM-ID) using putatively annotated and identified molecules in UHPLC-MS experiments having up to 2 ppm error and highest probability score i.e., 1 for soybean samples in positive and negative polar modes

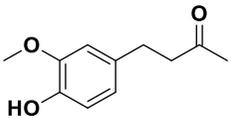
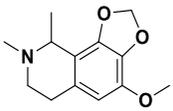
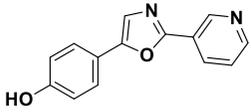
Sr. no.	m/z	Retention Time (min.)	Molecular Formula	Adduct	ppm error	Probability score	Molecule name	Structure	Sample Tissue	Polarity	Reported/ Unreported
1	416.1499	7.45	C ₂₂ H ₂₄ O ₈	[M] ⁻ [M+Cl] ⁻ [M+HCOO] ⁻	0.32	1	1-Acetoxyypinoresinol		Leaf, seed	Negative	Non-reported for soybean (Reported in <i>Olea europaea</i> Linne) (Kadowaki et al. 2003)
2	430.2102	11.79	C ₂₃ H ₃₀ N ₂ O ₆	[M] ⁻ [M+HCOOH] ⁻	0.00	1	Cinegalline		Leaf, seed	Positive	Non-reported for soybean (Reported in <i>Genista cinerea</i>) (Faugeras and Paris 1968)

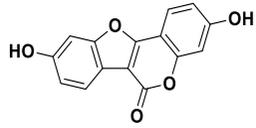
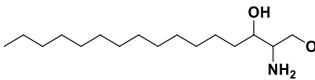
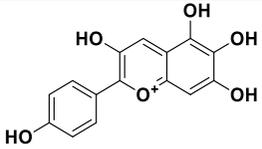
Sr. no.	m/z	Retention Time (min.)	Molecular Formula	Adduct	ppm error	Probability score	Molecule name	Structure	Sample Tissue	Polarity	Reported/ Unreported
3	448.19	11.97	C ₂₇ H ₂₈ O ₆	[M]- [M-H]-	2.72	1	Lonchocarpenin		Leaf, seed	Positive, Negative	Non-reported for soybean (Reported in <i>Millettia richardiana</i>) (Rajemiarimiraho et al. 2013)
4	570.1944	9.41, 9.45	C ₂₆ H ₃₄ O ₁₄	[M]- [M+Cl]-	-	1	Decuroside III		Leaf, seed	Negative	Non-reported for soybean (Reported in <i>Peucedanum decursivum Maxim.</i>) (Matano et al. 1986)

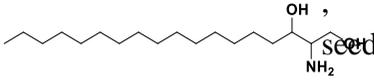
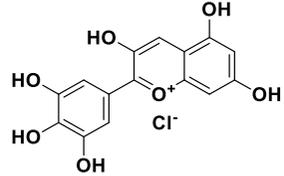
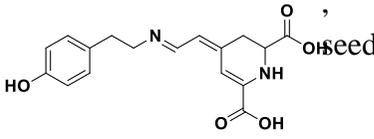
Sr. no.	m/z	Retention Time (min.)	Molecular Formula	Adduct	ppm error	Probability score	Molecule name	Structure	Sample Tissue	Polarity	Reported/ Unreported
5	596.1386	7.94, 8.35	C ₂₆ H ₂₈ O ₁₆	[M] ⁻ [M+Na+HCOONa] ⁻	0.19	1	Quercetin 3-O-[beta-D-xylosyl-(1->2)-beta-D-glucoside]		Seed	Negative	Non-reported for soybean (Reported in <i>Eucommia ulmoides</i>) (Yang et al. 2014)
6	772.2056	7.39, 7.54	C ₃₃ H ₄₀ O ₂₁	[M] ⁻	-	1	Kaempferol 3-sophorotrioside		Leaf	Positive, Negative	Non-reported for soybean (Reported in <i>Ficus carica</i> L.) (Nadeem and Zeb 2018)

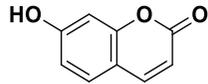
Sr. no.	m/z	Retention Time (min.)	Molecular Formula	Adduct	ppm error	Probability score	Molecule name	Structure	Sample Tissue	Polarity	Reported/ Unreported
7	824.4615	7.95, 8.09	C ₄₃ H ₆₈ O ₁₅	[M]-	-	1	Yiamoloside B		Leaf, seed	Positive, Negative	Non-reported for soybean (Reported in <i>Phytolacca octandra</i>) (Moreno and Rodriguez 1981)
8	926.52	12.44, 12.45	C ₄₈ H ₇₈ O ₁₇	[M]- [M-H]-	0.05	1	Saikosaponin BK1		Leaf	Negative	Non-reported for soybean (Reported in <i>Bupleurum kunmingense</i> Y.) (Luo et al. 1987)

Sr. no.	m/z	Retention Time (min.)	Molecular Formula	Adduct	ppm error	Probability score	Molecule name	Structure	Sample Tissue	Polarity	Reported/ Unreported
9	85.09	2.04, 2.08	C ₅ H ₁₁ N	[M] ⁺ [M+K+HCOOH] ⁺ [M+Na+HCOOH] ⁺ [M+HCOONa] ⁺ [M+H+HCOOH] ⁺	0.70	1	Piperidine		Leaf, seed	Positive	Reported in soybean (Luo et al. 1987)
10	146.0368	6.92, 6.94	C ₉ H ₆ O ₂	[M] ⁺ [M+H] ⁺	0.03	1	Coumarin		Leaf, seed	Positive	Reported in soybean (Beyer et al. 2019)

Sr. no.	m/z	Retention Time (min.)	Molecular Formula	Adduct	ppm error	Probability score	Molecule name	Structure	Sample Tissue	Polarity	Reported/ Unreported
11	194.0942	13.22	C ₁₁ H ₁₄ O ₃	[M] ⁺	-	1	Zingerone		Leaf, seed	Positive, Negative	Non-reported for soybean (Reported in Ginger) (Monge, Scheline and Solheim 1976)
12	235.1205	2.59, 2.63	C ₁₃ H ₁₇ NO ₃	[M+K] ⁺ [M+Na] ⁺	0.91	1	Lophophorine		Leaf, seed	Positive	Non-reported for soybean (Reported in Lophophora) (Bruhn et al. 1978)
13	238.0738	7.10, 7.11	C ₁₄ H ₁₀ N ₂ O ₂	[M] ⁻	-	1	Halfordinol		Leaf, seed	Negative	Non-reported for soybean (Reported in <i>Aeglopsis Chevalieri</i> Swing.) (Dreyer 1968)

Sr. no.	m/z	Retention Time (min.)	Molecular Formula	Adduct	ppm error	Probability score	Molecule name	Structure	Sample Tissue	Polarity	Reported/ Unreported
14	268.0361	3.15	C ₁₅ H ₈ O ₅	[M+2Na] ²⁺ [M+NH ₃] ⁺	0.11	1	Coumestrol		Leaf, seed	Negative	Reported in soybean (Hutabarat, Greenfield and Mulholland 2000)
15	273.2663	12.73	C ₁₆ H ₃₅ NO ₂	[M] ⁺ [M+H] ⁺	0.73	1	Hexadecaspinganine		Leaf, seed	Positive	Non-reported for soybean (Reported in <i>Manduca sexta</i>) (Abeytunga et al. 2008)
16	287.0548	6.51, 6.52	C ₁₅ H ₁₀ O ₆	[M+H] ⁺	0.76	-	Aurantidin		Leaf, seed	Positive	Non-reported for soybean (Reported in <i>Impatiens aurantiaca</i>) (Iwashina 2000)

Sr. no.	m/z	Retention Time (min.)	Molecular Formula	Adduct	ppm error	Probability score	Molecule name	Structure	Sample Tissue	Polarity	Reported/ Unreported
17	301.2977	13.53	C ₁₈ H ₃₉ NO ₂	[M] ⁺ [M+H] ⁺	1.70	1	Sphinganine		Leaf	Positive	Reported in soybean (Ohnishi and Fujino 1982)
18	303.05	7.93, 7.95	C ₁₅ H ₁₀ O ₇	[M+H] ⁺	0.79	-	Delphinidin		Leaf	Positive	Reported in soybean (Lee et al. 2017)
19	330.1211	8.14, 8.16	C ₁₇ H ₁₈ N ₂ O ₅	[M] ⁺ [2M+Na] ⁺ [M+Na] ⁺	0.17	1	Miraxanthin-III		Leaf	Positive, Negative	Non-reported for soybean (Reported in <i>Beta vulgaris</i> L.) (Kugler, Stintzing and Carle 2004)

Sr. no.	m/z	Retention Time (min.)	Molecular Formula	Adduct	ppm error	Probability score	Molecule name	Structure	Sample Tissue	Polarity	Reported/ Unreported
20	162.0316	5.42	C ₉ H ₆ O ₃	[M] ⁺ [M+H] ⁺	0.83	1	Umbelliferone		Leaf, seed	Positive	Reported in soybean (Dardanelli et al. 2010)

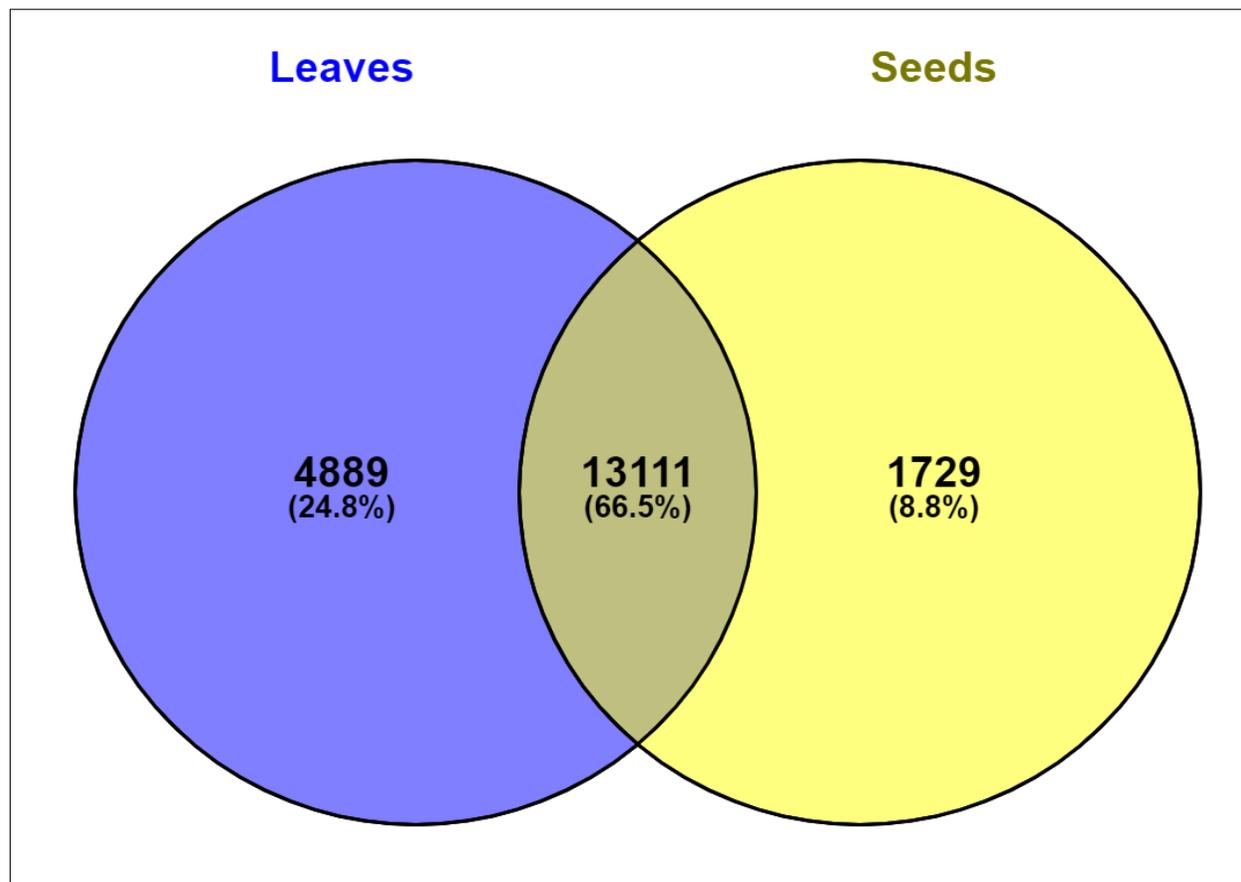


Figure 3.6: Venn diagram showing the differences between soybean seed and leaf annotated mass features of small molecules identified using two methods: ProbMetab (for probability score: 1) and PUTMEDID LC-MS in Taverna workflow (up to 2 ppm error)

a. Multivariate analysis

In the present study, we used chemometric approaches to investigate the intervarietal differences in soybean. The LC-MS data analysis was combined with multivariate chemometric techniques such as PCA and PLS-DA. These provide a powerful solution for investigating and comparing the metabolism of various species or varieties, which can be employed in drug discovery and drug development (Lee, Kim and Yoo 2011, Acevska et al. 2015). Unlike univariate statistics (ANOVA), multivariate statistics like PCA score plots reveal grouping between different samples of soybean varieties. PCA is an unsupervised method that best explains the variance of a data set (X) without referring to class labels (Y).

In contrast, PLS is a supervised method that uses multivariate regression methods to extract information that can predict class labels (Y) through a linear combination of data set variables (X). The Variable Importance in Projection (VIP) scores estimate the importance of each variable in the projection used in a PLS model and are often employed for variable selection. The VIP score plot shows how the variables or metabolites contribute to the variations among the samples. This finding is consistent with the results of the PCA. Hence, PCA was applied without considering the correlation between dependent and independent variables. In contrast, PLS-DA is applied based on correlations. A dendrogram was generated to study the hierarchical cluster analysis (HCA).

The peak lists obtained from XCMS were normalized using MetaboAnalyst to a constant sum method, which is a commonly used metabolomic normalization method. After Pareto scaling of the data, PCA was performed to reveal any variations due to outliers in all dataset samples. The PCA 2D score plot of all the samples is presented in **Figure 3.7**. The score plot indicated that the two principal components

(PC1 and PC2) cumulatively accounted for 42.3% (26% and 17.7%) of the total variance of the dataset for ESI(+) mode and 31% (21.9% and 9.1%) of the total variance of the dataset for ESI(-) modes, respectively (**Figure 3.7A and 3.7B**). The score plot reveals a two-dimensional representation of 48 samples (i.e., six biological replicates each of seed and leaves of four different varieties) for ESI(+) and ESI(-) modes. The PC score plots showed a distinctive and isolated cluster of leaf and seed samples for the variety JS-7105 among the eight samples, indicating characteristic intervarietal variations of metabolites in positive polar mode. The leaf samples of JS-7105 formed a different cluster, whereas all other samples were interacting with each other due to the presence of similar metabolites.

However, in the negative polar mode, it was observed that the outlying grouping of seed and leaf samples of JS-9305 indicated variations among all other samples in the dataset. For further analysis, HCA plots clearly show the segregation of leaf and seed samples of JS-7105 apart from each other due to the presence of different metabolites in the positive mode (**Figure 3.7 C**). Similarly, the leaf and seed samples of JS-9305 formed a cluster, except for one sample of leaf and seed of NRC-119 for negative polar mode (**Figure 3.7 D**). It also reveals that leaf samples of JS-7105 formed a cluster due to the presence of specific and common metabolites in the sample and different from all other soybean varieties.

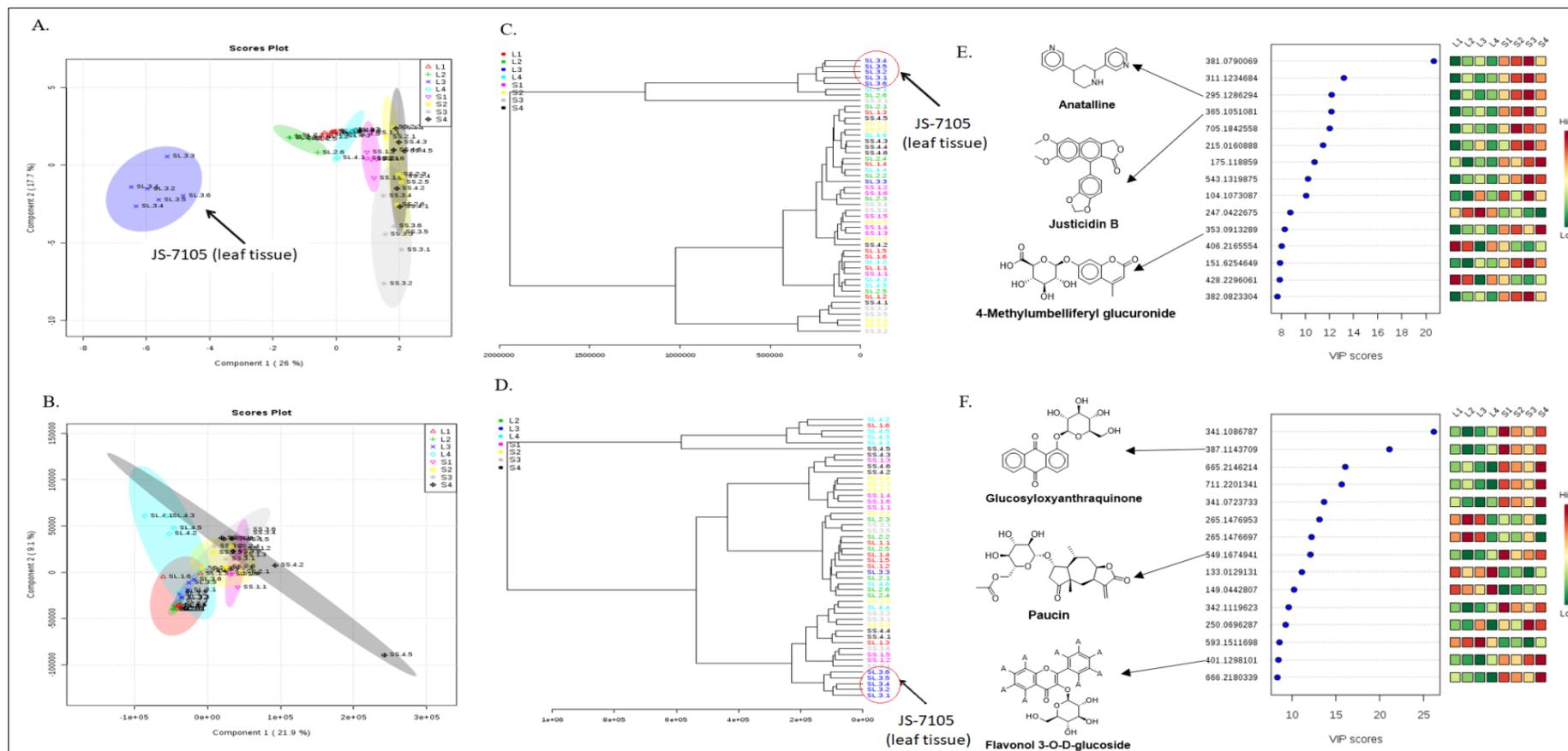
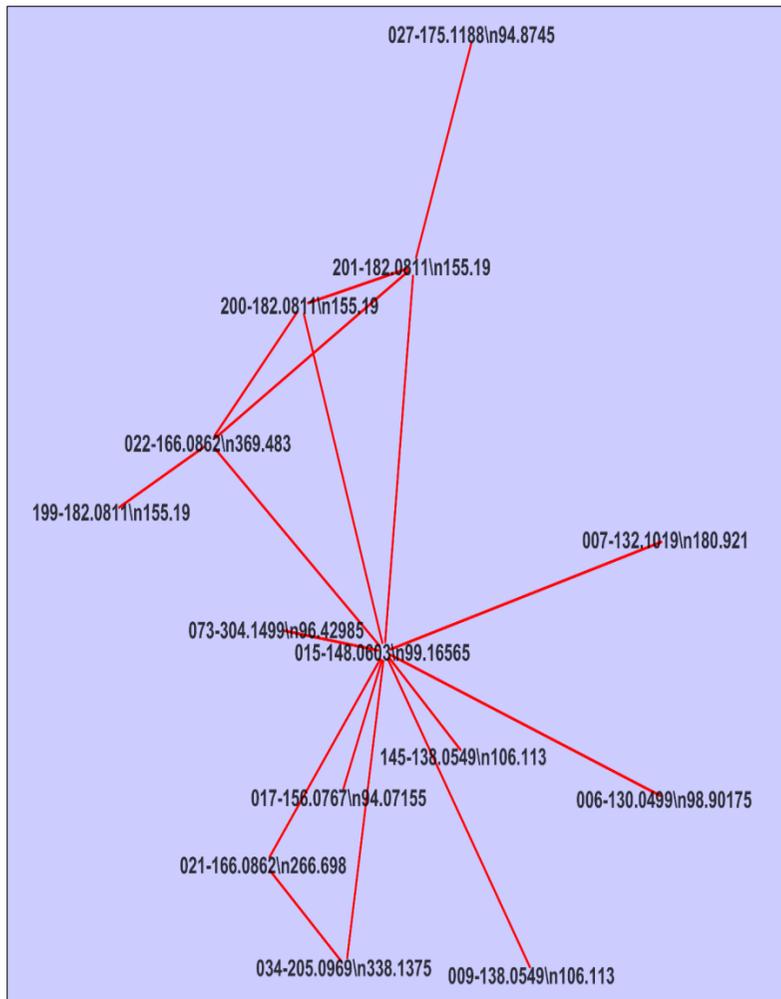


Figure 3.7: Statistical analysis by ANOVA for +ve and -ve polar molecules. PCA plots (A.) +ve (B.) -ve; HCA (C.) +ve (D.) -ve; PLS-DA loadings for top 15 important features of differentially co-accumulated metabolites in soybean seed and leaf (E.) +ve (F.) -ve. L1: Leaf sample of variety NRC119, L2: Leaf sample of variety JS335, L3: Leaf sample of variety JS7105, L4: Leaf sample of variety JS9305, S1: Seed sample of variety NRC119, S2: Seed sample of variety JS335, S3: Seed sample of variety JS7105, and S4: Seed sample of variety JS9305.

The PLS-DA model was constructed and validated using MetaboAnalyst. The 15 most significant variables were selected according to the VIP scores after the Pareto scaling for sample discrimination and to identify the features essential for group classification (**Figure 3.7 E and 3.7 F**). It is also provided with the patterns of change for those variables, i.e., the concentration of the metabolites in arbitrary units in the seed and leaf data sets of the four varieties. This shows the differences between molecular features for their content in the four varieties of sample tissues. The resonances corresponding to those variables are attributable to metabolites whose levels were significantly different either in seed or leaf among the four soybean varieties. The results of the VIP plot considerably contributed to the characterization of the metabolic profile of each variety based on the PC scores. These results suggest that LC-MS-based multivariate analytical approaches are useful for the evaluation of potential intervarietal comparison.

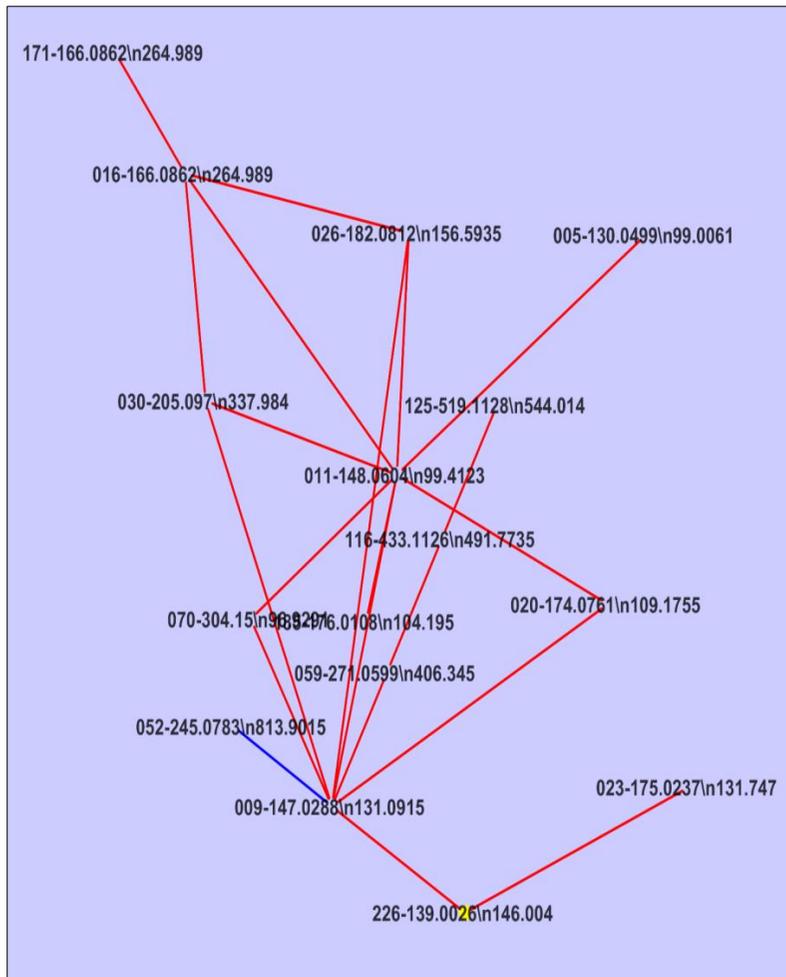
b. Metabolic accumulation in soybean varieties according to KEGG pathways

Metabolomics plays an essential role in unraveling the mechanistic changes between species or varieties, whether due to genetics or environmental conditions by providing information about the main pathways and metabolites for phenotypes and genotypes of interest (Fiehn 2002, Schauer and Fernie 2006). In this study, the comprehensive analysis and identification of small organic molecules in four soybean varieties were performed using ProbMetab. This includes the information from peak ranking, probabilistic annotation of compounds, and associates with metabolic pathways retrieved from the KEGG database (**Supplementary Table S3.3.3, Figure 3.8**).



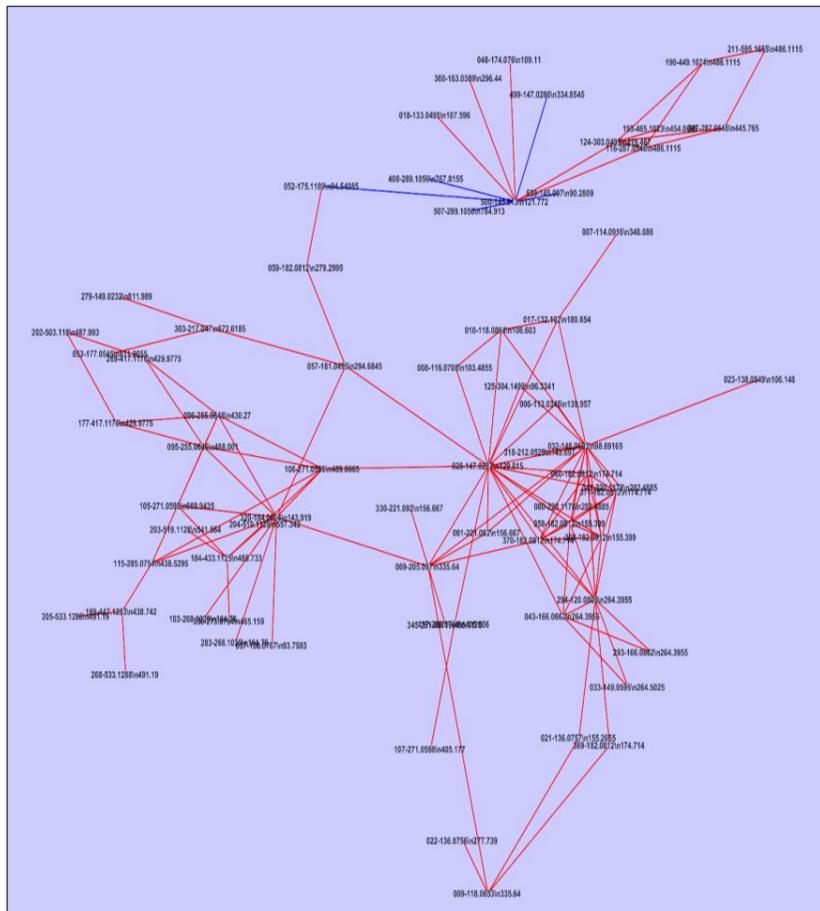
Id	Pathway Name	N. Compounds
00300	Lysine biosynthesis	5
00330	Arginine and proline metabolism	6
00360	Phenylalanine metabolism	6
00400	Phenylalanine, tyrosine and tryptophan biosynthesis	9
00430	Taurine and hypotaurine metabolism	5
00460	Cyanoamino acid metabolism	6
00940	Phenylpropanoid biosynthesis	5
00941	Flavonoid biosynthesis	5
00943	Isoflavonoid biosynthesis	11
00960	Tropane, piperidine and pyridine alkaloid biosynthesis	8
00966	Glucosinolate biosynthesis	6
00970	Aminoacyl-tRNA biosynthesis	13
01100	Metabolic pathways	29
01110	Biosynthesis of secondary metabolites	29
01210	2-Oxocarboxylic acid metabolism	12
01230	Biosynthesis of amino acids	20
02010	ABC transporters	13

a. Variety 1 – NRC119



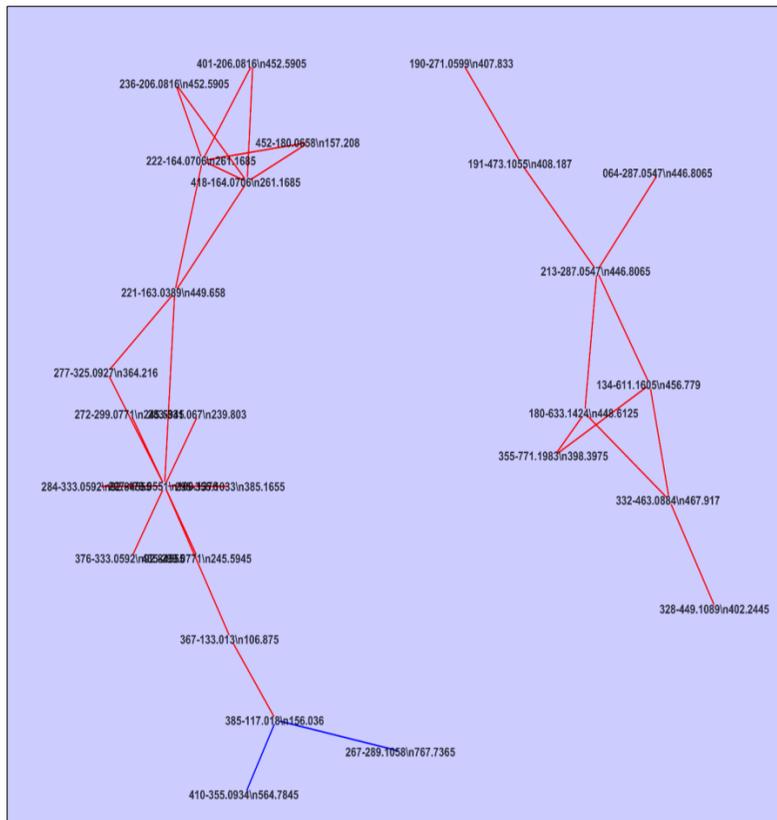
Id	Pathway Name	N. Compounds
00040	Pentose and glucuronate Interconversions	5
00053	Ascorbate and aldarate metabolism	8
00300	Lysine biosynthesis	5
00430	Taurine and hypotaurine metabolism	5
00630	Glyoxylate and dicarboxylate metabolism	6
00941	Flavonoid biosynthesis	5
00943	Isoflavonoid biosynthesis	8
01100	Metabolic pathways	34
01110	Biosynthesis of secondary metabolites	21
01200	Carbon metabolism	5
01210	2-Oxocarboxylic acid metabolism	9
01230	Biosynthesis of amino acids	12
02010	ABC transporters	6

b. Variety 2 – JS335



Id	Pathway Name	N. Compounds
00040	Pentose and glucuronate interconversions	5
00053	Ascorbate and aldarate metabolism	7
00130	Ubiquinone and other terpenoid-quinone biosynthesis	10
00220	Arginine biosynthesis	6
00230	Purine metabolism	7
00261	Monobactam biosynthesis	8
00300	Lysine biosynthesis	6
00350	Tyrosine metabolism	8
00360	Phenylalanine metabolism	16
00380	Tryptophan metabolism	6
00400	Phenylalanine, tyrosine and tryptophan biosynthesis	13
00430	Taurine and hypotaurine metabolism	5
00460	Cyanoamino acid metabolism	14
00660	C5-Branched dibasic acid metabolism	5
00730	Thiamine metabolism	5
00770	Pantothenate and CoA biosynthesis	5
00940	Phenylpropanoid biosynthesis	13
00941	Flavonoid biosynthesis	13
00943	Isoflavonoid biosynthesis	17
00944	Flavone and flavonol biosynthesis	11
00950	Isoquinoline alkaloid biosynthesis	8
00960	Tropane, piperidine and pyridine alkaloid biosynthesis	5
00965	Betalain biosynthesis	5
00966	Glucosinolate biosynthesis	11
00970	Aminoacyl-tRNA biosynthesis	15
01100	Metabolic pathways	68
01110	Biosynthesis of secondary metabolites	51
01200	Carbon metabolism	5
01210	2-Oxocarboxylic acid metabolism	22
01230	Biosynthesis of amino acids	28
02010	ABC transporters	12

c. Variety 3 – JS7105



Id	Pathway Name	N. Compounds
00040	Pentose and glucuronate interconversions	5
00052	Galactose metabolism	7
00053	Ascorbate and aldarate metabolism	9
00360	Phenylalanine metabolism	7
00400	Phenylalanine, tyrosine and tryptophan biosynthesis	6
00460	Cyanoamino acid metabolism	5
00630	Glyoxylate and dicarboxylate metabolism	5
00940	Phenylpropanoid biosynthesis	5
00941	Flavonoid biosynthesis	12
00943	Isoflavonoid biosynthesis	12
00944	Flavone and flavonol biosynthesis	11
00970	Aminoacyl-tRNA biosynthesis	6
01100	Metabolic pathways	42
01110	Biosynthesis of secondary metabolites	35
01200	Carbon metabolism	8
01210	2-Oxocarboxylic acid metabolism	9
01230	Biosynthesis of amino acids	12
02010	ABC transporters	9

d. Variety 4 – JS930

Figure 3.8: Metabolic pathway network with the list of pathway names and the number of molecules involved in it for four varieties of soybean retrieved from KEGG soybean pathways

The workflow presented in **Figure 3.2** for ProbMetab involves preprocessing, identification of metabolites from UHPLC-MS data, and identifying the metabolic pathways for these metabolites by merging them with previously known soybean pathways in KEGG. By doing this, small organic molecules were annotated and identified by probabilistic ranking using the Bayesian method. The associated metabolic pathways were retrieved from KEGG soybean pathways for all the four soybean varieties in the form of a Cytoscape network. The top ten pathways were selected from the soybean pathways extracted for the four varieties based on the medicinally important and drug-like metabolites involved in these pathways (Verpoorte 1998, Julsing et al. 2006).

Further, these ten pathways were divided into three categories based on the similarities of type of phytochemical pathways; *viz.* Category 1: Flavonoids- Flavonoid biosynthesis, Flavone and flavonol biosynthesis, Isoflavonoid biosynthesis; Category 2: Terpenoid alkaloids- Ubiquinone and other terpenoid-quinone biosynthesis, Isoquinoline alkaloid biosynthesis, Tropane, piperidine, and pyridine alkaloid biosynthesis; and, Category 3: Others- Phenylpropanoid biosynthesis, Betalain biosynthesis, Pentose and glucuronate biosynthesis, 2-Oxocarboxylic acid metabolism. Based on this, a heatmap of soybean seed and leaf metabolites identified through UHPLC-MS analysis for the ten selected biochemical pathways was generated for the four varieties of soybean (**Figure 3.9**).

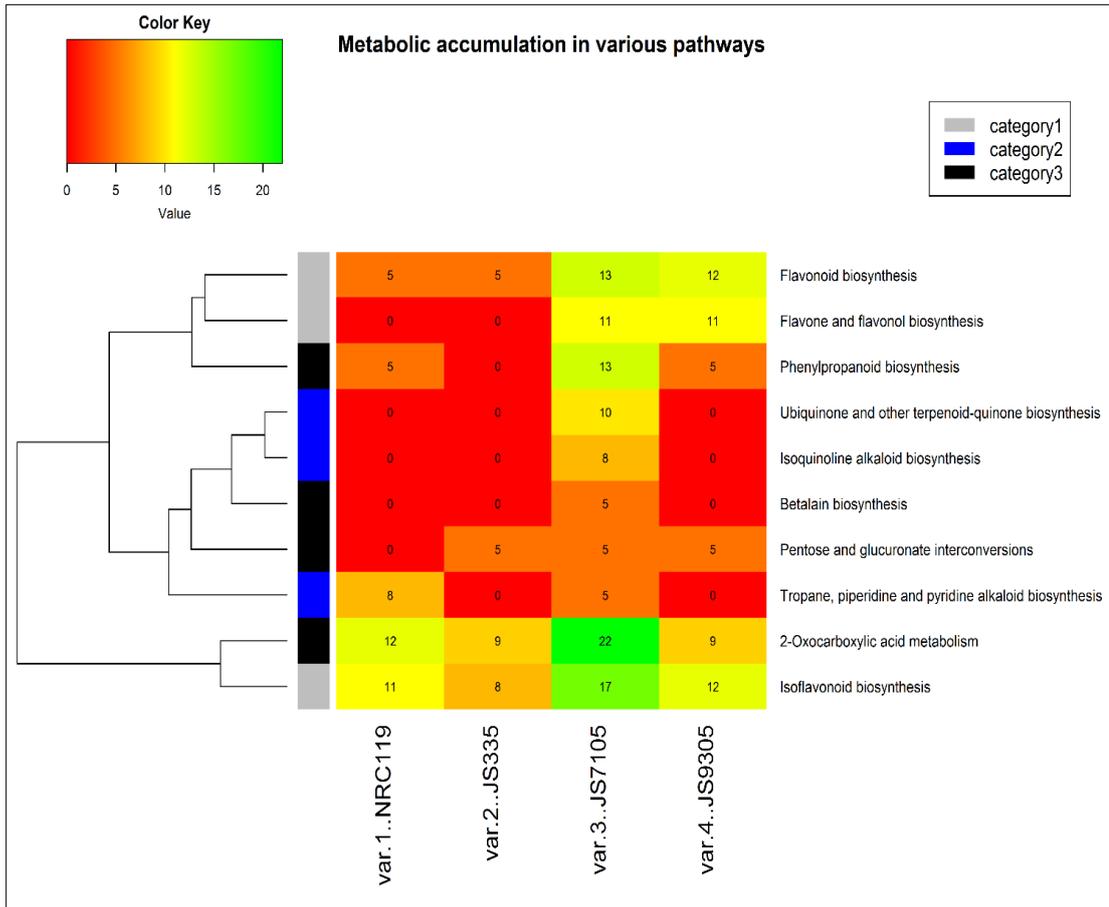


Figure 3.9: Metabolic accumulation in soybean varieties according to KEGG pathways. The heat maps were drawn using the R package ggplot2, and the green-red color represents the transformed raw data of soybean metabolites with significant differences among four sample varieties. Green and red colors indicate an increase and a decrease in metabolite levels, respectively. Categories represent the type of metabolic pathways. Category 1: Flavonoids, Category 2: Terpenoids and Category 3: Others.

The heat maps show that the soybean variety JS-7105 had the highest metabolic accumulation and the variety JS-335 had the lowest metabolic accumulation in the selected pathways according to the KEGG soybean pathway. These outcomes confirm the results of PCA plots for multivariate statistical analysis.

Metabolic pathway networks with the list of pathway names and the number of molecules involved were also retrieved for the other three varieties (NRC-119, JS-335, and JS-9305) of soybean from KEGG soybean pathways. The remaining data of the pathway network and the list of small organic molecules with probability scores for the four soybean varieties are presented in **Supplementary Table S3.3.3 and Figure 3.8**. It was observed that the metabolic pathways obtained for JS-7105 were more diverse than those for the other three varieties due to higher metabolic accumulation, which is visible in the heatmap. **Figure 3.8d** shows the metabolic pathway network with a list of pathway names and the number of molecules involved in it for the soybean variety JS-7105 retrieved from KEGG soybean pathways. The heatmap also shows the clustering of metabolic pathways in all four soybean varieties according to their metabolite content. The number of metabolites extracted under each pathway for each soybean variety is limited as the metabolic data for soybean extracted from the KEGG database is sparse. Hence, more mass spectrometric analysis-based research is required in soybean.

c. Tandem mass spectrometry

Tandem mass spectrometry or MS/MS or MS² is a method to screen the molecules based on their fragmented molecular masses in an untargeted metabolomics experiment (McLafferty 1981). Here, we used tandem mass spectrometry to screen and validate the small molecules identified putatively using ProbMetab and PUTMEDID-LCMS methods. The input data of metabolites for MS/MS run were

prepared according to their high intensity, high probability score, and low ppm error analyzed using ProbMetab and PUTMEDID-LCMS. ProbMetab was used to annotate the molecules by combining mass spectral peaks of negative and positive modes for leaf and seed samples of the four soybean varieties. ProbMetab can also be used to annotate the molecules by combining the number of sample spectra files, which in this case is of a combination of seed and leaf spectral files in negative and positive polar modes of the four soybean varieties (**Supplementary Table S3.3.3**); whereas, PUTMEDID-LCMS method is limited to analyze a single file at a time (**Supplementary Table S3.3.1**).

Later, all the small molecules obtained using ProMetab and PUTMEDID-LCMS were filtered with a probability score of 1 and ppm error up to 2, respectively. Thus, 211 and 8018 annotated mass features of small molecules were identified in the variety NRC119; 198 and 9177 in variety JS335; 242 and 14190 in variety JS7105; and 226 and 13608 in variety JS9305, using ProMetab and PUTMEDID-LCMS, respectively. The total number of putatively annotated molecules obtained after removing duplicate molecules for all the four soybean varieties by combining these two methods with the highest probability score i.e., 1 and up to 2 ppm error was 7185, of which 557 molecules were known to be present in soybean (**Supplementary Table S3.4.1**). Among these known molecules, 443 molecules were previously reported (Text mined + SoyCyc + SoyKB), while the remaining were identified and extracted from KEGG soybean pathways using the ProMetab R package. From these, 50 molecules containing both known and unknown molecules were selected for performing tandem mass spectrometry.

The CFM-ID webservice was employed to confirm the structures of small molecules (n=20) by searching against all the databases present in the webservice

(using exact mass and MS/MS fragmentation patterns). However, confirmed structures could be obtained only from the KEGG and HMDB databases, most of them being identified from KEGG. The query spectra in the form of a list of m/z and intensity, parent ion mass and adduct type (neutral for both positive and negative polar molecules) were used as inputs for comparing with the pre-trained model to select the most likely candidate compound from the databases (KEGG or HMDB). The top 10 score ranking lists were considered for molecule validation according to the Jaccard scoring function rule. Thus, we identified and validated previously reported six small molecules (piperidine (Arai et al. 1966), coumarin (Colpas et al. 2003), coumestrol (Hutabarat et al. 2000), sphinganine (Ahn and Schroeder 2002), delphinidin (Buzzell, Buttery and MacTavish 1987) and umbelliferone (Rao and Cooper 1995)) in soybean (**Table 3.2, Supplementary Table S3.4.2, and S3.4.3**).

We also detected and validated 14 new molecules (lophophorine, halfordinol, zingerone, 1-acetoxypinoresinol, saikosaponin BK1, lonchocarpenin, decucoside III, yiamoloside B, miraxanthin-III, quercetin 3-O-[beta-D-xylosyl-(1->2)-beta-D-glucoside], cinegalline, hexadecasphinganine, kaempferol 3-sophorotrioside and aurantidin) for the first time in soybean. Most of these molecules were detected in both the seed and leaf tissue samples of soybean varieties, except one molecule (quercetin 3-O-[beta-D-xylosyl-(1->2)-beta-D-glucoside]) detected only in seed tissues and three molecules (kaempferol 3-sophorotrioside, delphinidin and saikosaponin BK1) that were detected only in leaf tissues.

Many of these molecules have been previously analyzed for their therapeutic properties and their beneficial effects on human health. As an example, the lignan, 1-acetoxypinoresinol (**Table 3.2, entry 1**) is a derivative of pinoresinol. It is effective in leukemia treatment and inhibits P-glycoprotein transporter-mediated *MDR-1* gene

(Gonzalez et al. 2017). Its beneficial effect on fatty acid synthase (FASN) expression in human breast epithelial cell lines to induce anti-cancer effects has also been reported (Menendez et al. 2008). Likewise, lonchocarpinin (**Table 3.2, entry 3**) is a hydroxycoumarin reported for its antiprotozoal activities (Rajemiarimiraho et al. 2014). Similarly, piperidine (**Table 3.2, entry 9**), which acts as a renin inhibitor, is suggested as the most promising drug candidate in treating chronic renal failure (Märki et al. 2001). Thus, we used the putatively annotated small molecules detected with LC-MS, to perform an accurate mass search via an *in-silico* fragmentation approach, which provided a list of the candidate molecules for drug development.

3.3.3 Soybean scaffold drug network

A network was constructed to visualize the relationships among the previously reported soybean small organic molecules (Text mined + SoyCyc + SoyKB), FDA approved drugs, and soybean small organic molecules detected by UHPLC-MS/MS in the four soybean varieties to identify common scaffolds that link the soybean molecules to drugs (**Figure 3.10, Supplementary Table S3.5**). The final network consists of four merged networks of previously reported soybean small molecules (1622), drug molecules (2354), previously reported soybean small molecules identified by UHPLC-MS (557), and the previously unreported annotated mass features for soybean small molecules identified by UHPLC-MS (6628), to study the inter-relationship between scaffolds and molecules. All the networks were merged to create a supra-network containing 10670 nodes and 11482 edges. The network analysis of the topological features computed for the network showed an average number of neighbors with 2.152 and characteristic path length with 3.477 scores depicting the maximum connectivity of all molecules and their common scaffolds.

All the drugs and soybean molecules were compared to reveal common scaffolds (n=25) among them (**Table 3.3**). From them, drug molecules with common scaffolds between drug and soybean molecules were n=49, and soybean molecules with common scaffolds between drug and soybean molecules were n=48. Three soybean molecules (nicotinate, pseudoxynicotine, and quinolinate) had similar scaffolds with three drugs (DB13882, DB12911, and DB09220), while 13 soybean metabolites scaffolds were also similar to 12 drugs scaffolds. For example, the scaffold with ID: 15 of sinapaldehyde possesses a scaffold similar to chloroxylenol, an antiseptic and disinfectant agent. Likewise, Sc ID: 01 of myoinositol shows similarities with scaffold (ID 23) extracted from lindane, an anti-scabies agent. The soybean molecules, indol-3-yl) acetate, indole-3-butanoate, and methyl (indol-3-yl) acetate had a similar scaffold (ID- 48) with L-tryptophan (DB ID: DB00150), which is an anti-depressant and dietary supplement.

This analysis reveals that the soybean molecules identified in this study possess properties or bioactivities similar to the commercially available drugs based on their common scaffold structures, as structural descriptors encode activity. Similar bioactivities of these phytochemicals and metabolites make soybean a prospective candidate for further use in drug discovery. Similarly, eight soybean molecules were found to have similar scaffolds (ID-25) to sixteen drug molecules, which were antihypertensive agents, anti-arrhythmia agents, etc. Likewise, 15 soybean molecules with scaffold ID-23 had a similar scaffold to nine drugs and four soybean molecules (SC ID-38) had a similar scaffold to six drug molecules, which are alpha-1 adrenergic receptors agonist.

The total number of common molecules between soybean molecules and drugs was n=231, as shown in the network. Among them, the common molecules between

drugs and soybean (reported molecules identified by UHPLC-MS) were $n=3$, common molecules between drugs and soybean (reported molecules identified by UHPLC-MS + Text mined + SoyCyc + SoyKB) were $n=26$, common molecules between unreported and drugs are $n=155$ and common molecules between soybean and drugs are $n=47$. For example, Berberine is a previously reported molecule from soybean, which is also an antidiarrheal and antifungal drug (DB04115). Similarly, the previously reported soybean molecule Papaverine also acts as a muscle relaxant with drug bank ID: DB01113.

It was found that 443 molecules were common between soybean molecules, which have been previously reported (Text mined + SoyCyc + SoyKB), and molecules identified from soybean leaf and seed tissue samples identified through untargeted UHPLC-MS. The remaining reported soybean molecules were identified from the KEGG database for soybean pathways while analyzing the UHPLC-MS raw data using ProbMetab. Thus, out of 1622 previously reported molecules, we identified 443 molecules in soybean varieties through untargeted UHPLC-MS experiments. From them, six soybean molecules were validated through tandem mass spectrometry. Similarly, 14 molecules that were not previously reported in soybean were also validated by tandem mass spectrometry. This comparison shows that common molecules between soybean small molecules and drugs have drug-like or lead-like properties and could be developed as drugs after further research.

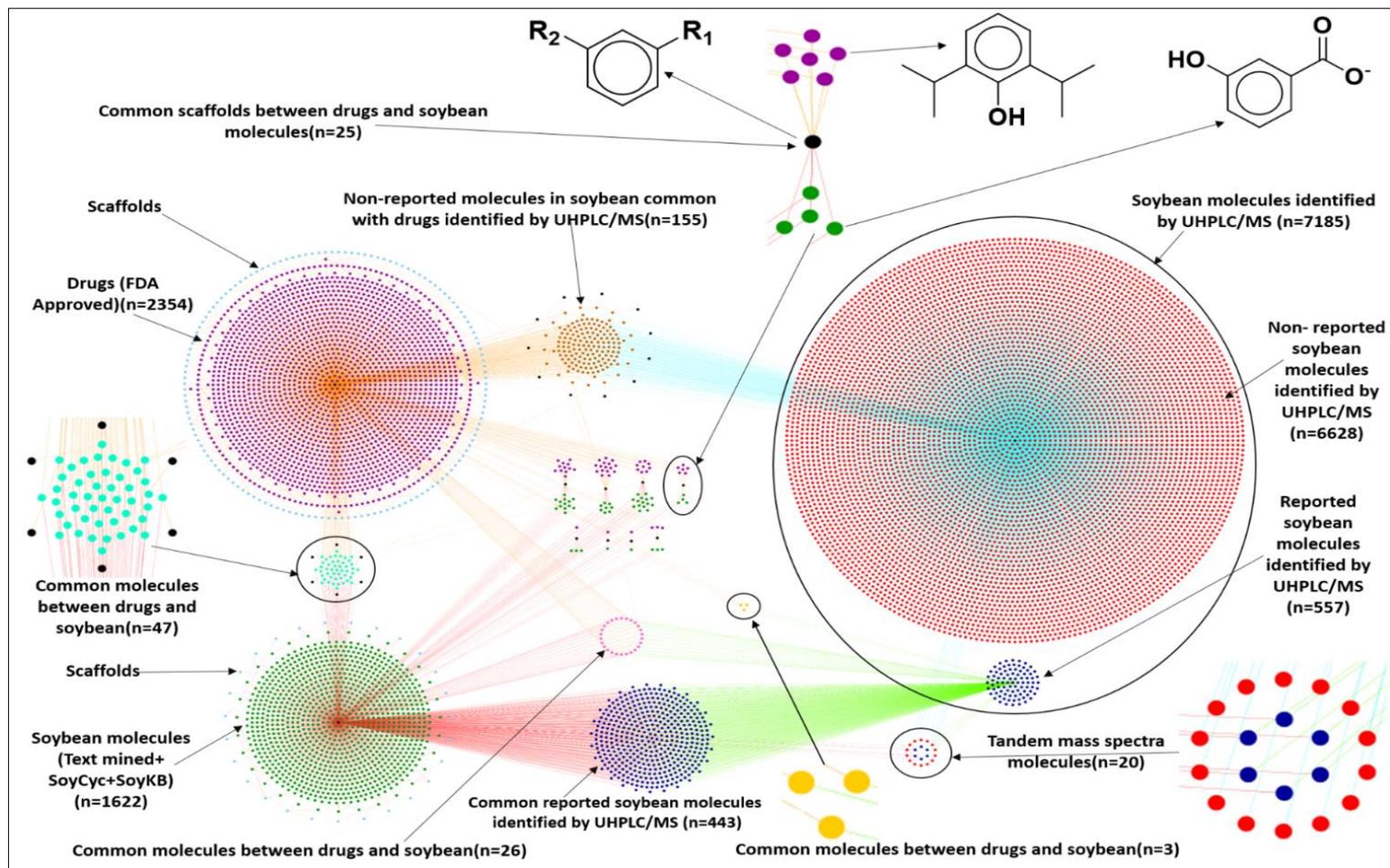
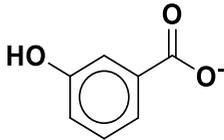
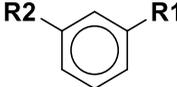
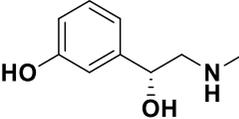
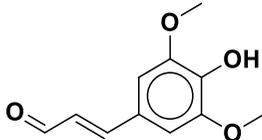
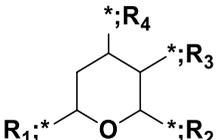
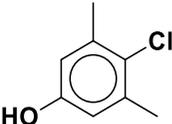
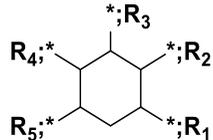
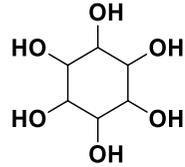
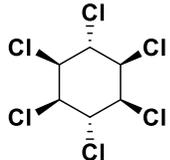
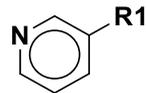
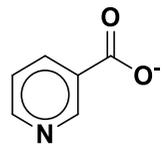
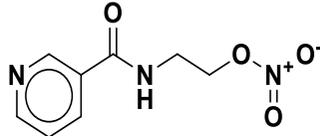
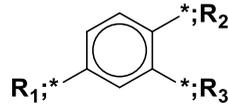
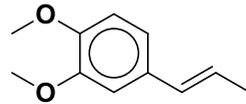
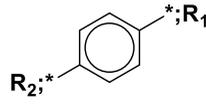
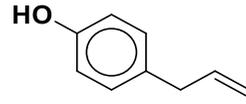
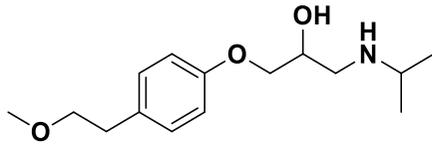


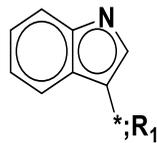
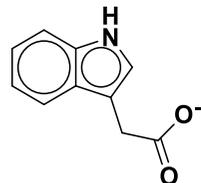
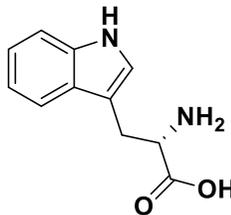
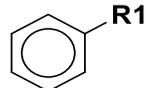
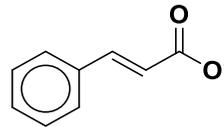
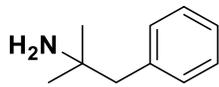
Figure 3.10: Soybean small molecules, drug molecules, and scaffold merged network as depicted in an organic layout in Cytoscape. Nodes = 10670 edges = 11482 (Nodes: Molecules; Edges: Interactions/ hidden relationships)

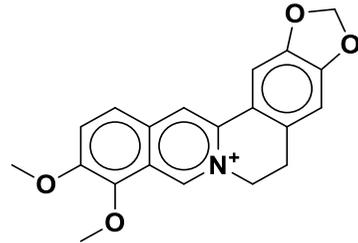
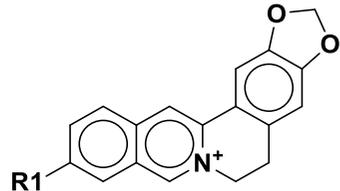
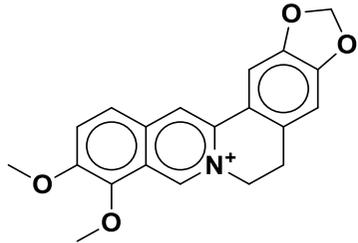
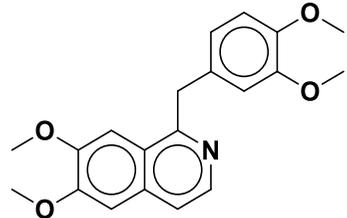
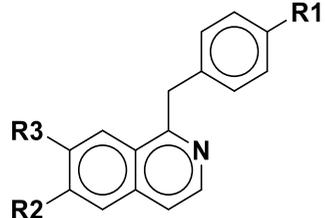
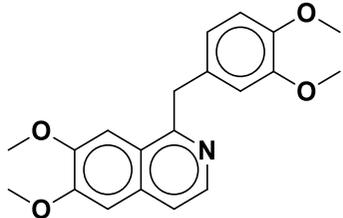
Table 3.3: Common scaffolds identified in soybean small molecules, drug molecules, and scaffold merged network with information about their therapeutic categories (n=10) (Sc: Scaffold)

Sr. No.	Chemical Class (a)	Scaffold (b)	Drug (c)
1.	3-hydroxybenzoate  Sc ID: 38		Phenylephrine: alpha-1 adrenergic receptor agonist  Sc ID: 05 DB ID: DB00179
2.	Sinapaldehyde  Sc ID: 15		Chloroxylenol: antiseptic and disinfectant agent  Sc ID: 31 DB ID: DB11121

Sr. No.	Chemical Class	Scaffold	Drug
	(a)	(b)	(c)
3.	Myo-inositol		Lindane: Antiscabies agent
			
	Sc ID: 01		Sc ID: 23 DB ID: DB00431
4.	Quinolinate		Nicorandil: An orally efficacious vasodilatory drug and antianginal
			
	Sc ID: 45		Sc ID: 121 DB ID: DB00198

Sr. No.	Chemical Class	Scaffold	Drug
(a)	(b)	(c)	
5.	Isomethyleugenol		Midodrine: Vasoconstrictor agent
			
	Sc ID: 23		Sc ID: 02
			DB ID: DB00211
6.	Chavicol		Metoprolol: Antihypertensive agent, anti-arrhythmia agent
			
	Sc ID: 25		Sc ID: 26
			DB ID: DB00264

Sr. No.	Chemical Class	Scaffold	Drug
	(a)	(b)	(c)
7.	(Indol-3-yl)acetate		L-tryptophan: Anti-depressive agent, dietary supplement
			
	Sc ID: 48		Sc ID: 07 DB ID: DB00150
8.	Trans-cinnamate		Phentermine: Sympathomimetic amine anorectic agent
			
	Sc ID: 42		Sc ID: 6 DB ID: DB00191

Sr. No.	Chemical Class	Scaffold	Drug
(a)	(b)	(c)	
9.	Berberine		Berberine: antidiarrheal, antifungal
			
	Sc ID: 36		Sc ID: 165 DB ID: DB04115
10.	Papaverine		Papaverine: muscle relaxant
			
	Sc ID: 34		Sc ID: 129 DB ID: DB01113

3.3.4 Development of a virtual library and virtual screening

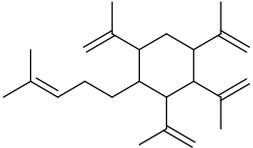
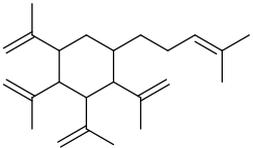
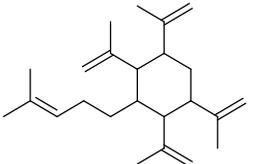
Virtual libraries were generated for the soybean small organic molecules, which were previously reported and which were also detected using UHPLC-MS (**Table 3.4, Supplementary Table 3.6**). For this, five previously reported soybean molecules were selected from diverse groups of phytochemical classifications according to the KEGG phytochemical classification (https://www.genome.jp/kegg-bin/get_htext) such as phenylpropanoid-anhydrosecoisolariciresinol (Acevska et al. 2015, Kang et al. 2010), alkaloid- 1-methyl-beta-carboline or harmane (Adachi et al. 1991, Heshmati, Nasehi and Zarrindast 2013), terpenoid- campesterol (Rozanski 1966, Yamaya et al. 2007), flavonoid-medicarpin (Stafford 1997, Silva et al. 2020) and others- phytic acid or phytate (Han 1988). All of these molecules have pharmacological significance, as reported before.

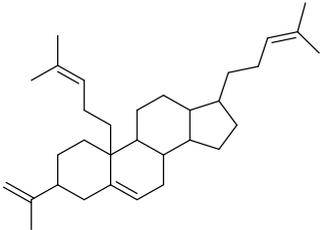
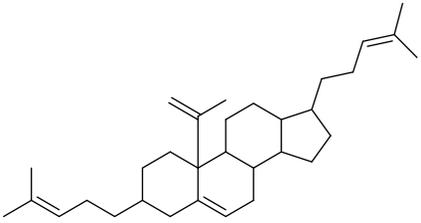
Anhydrosecoisolariciresinol has previously been shown to have anti-HIV-1 activity *in vitro* (Shang et al. 2013). Campesterol, a plant sterol, on the other hand, has been reported to decrease cholesterol and shows anticarcinogenic properties (Choi et al. 2007). Harmane, an alkaloid, was also been described as an anti-HIV drug and a reversible inhibitor of monoamine oxidase A. (Glover et al. 1982). Medicarpin is a chemopreventive agent, causing apoptosis and increasing the cytotoxicity of chemotherapeutic medicines in multidrug-resistant cancer cells (Gatouillat et al. 2015). Phytic acid is also known as Myo-inositol hexakisphosphate. It acts as a hypocalcemic agent by removing traces of heavy metal ions and hence helps prevent over-mineralization of joints, blood vessels, and other parts of the body, which is common in older persons (Kapral et al. 2012).

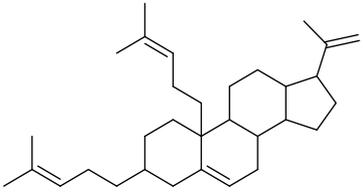
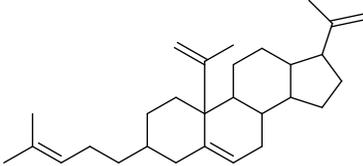
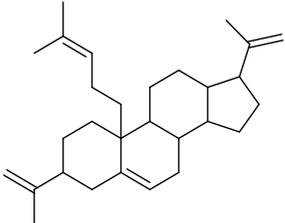
The scaffolds and functional groups extracted from these medicinally active molecules were used as seeds for virtual library generation. The virtual molecules were prioritized based on their “Progressive Drug-Like” (PDL), “Progressive Lead-Like” (PLL), toxicophore, pharmacophore, and chemophore (TPC) scores. The PDL and PLL scores were computed based on the progressive limits on physicochemical properties. The TPC scores are based on the molecular fingerprint pattern recognition module. A virtual molecule is considered to be a good drug-like and lead-like molecule if it has more pharmacophoric than toxicophoric scores. Chemophoric scores indicate reactivity with other molecules, which should be lower than pharmacophoric scores.

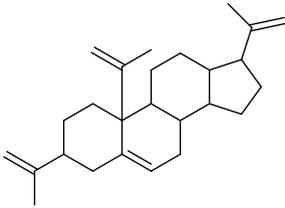
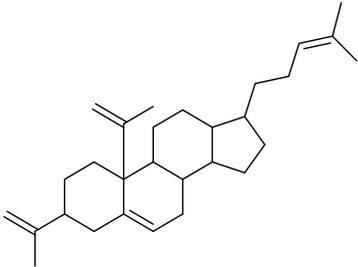
Previous research has found that virtual molecules derived from natural product molecules exhibit the same bioactivity as the natural product molecules due to the inheritance of identical scaffolds and functional groups (Lee and Schneider 2001b, Rollinger et al. 2008). The novel bioactive molecules are predicted to have anti-microbial, anti-cardiac, anti-cancer, anti-diabetic, etc., properties, based on the scaffold similarities with known drugs having therapeutic properties. Thus, from only five soybean molecules, 1225 virtual novel molecules were generated. Likewise, a large number of virtual molecules could be generated from the remaining soybean reported molecules and virtually screened for various parameters to select candidate molecules for drug development, which is highly efficient in developing new drugs.

Table 3.4: Virtual library novel molecules with their molecular weight, TPC and PDL, PLL scores (n= 10) [Notes: PDL: Progressive Drug-Like, PDL: Progressive Lead-Like, T: Toxicophore, P: Pharmacophore, C: Chemophor.]

Sr. No.	VL molecule structure	Molecular weight	PDL	PLL	T	P	C
1		326.559	1.278	2.03	1	21	6
2		326.559	1.278	2.03	1	20	6
3		326.559	1.278	2.03	1	20	6

Sr. No.	VL molecule structure	Molecular weight	PDL	PLL	T	P	C
4		434.739	0.845	2.775	1	26	8
5		434.739	0.512	2.593	1	26	8

Sr. No.	VL molecule structure	Molecular weight	PDL	PLL	T	P	C
6		434.739	0.345	1.799	1	26	8
7		392.66	0.086	1.115	1	26	8
8		392.66	0.086	1.115	1	26	8

Sr. No.	VL molecule structure	Molecular weight	PDL	PLL	T	P	C
9		350.58	-0.007	0.622	1	22	8
10		392.66	0.086	1.536	1	26	8

3.4 Conclusions

In this study, we successfully employed the chemoinformatics and experimental mass spectrometric approaches to identify and screen drug-like and lead-like compounds from soybean. This study suggests a combinatorial approach employing high-throughput metabolomics and chemoinformatics methods to efficiently identify new drug-like plant metabolites for targeted drug development. These molecules can be subsequently purified and experimentally evaluated for drug discovery research.

CHAPTER 4

CHEMOINFORMATICS INVESTIGATION ON CHEMICAL DEFENSE IN PLANTS

Chapter 4: Chemoinformatics Investigation on Chemical

Defense in Plants

4.1 Introduction

Plants are sessile organisms and provide food, energy and shelter to insects, animals, birds, and other organisms, who depend on them for their lives. While in most cases, this relationship between plants and other organisms is mutually beneficial, in some cases, the plants become victims of parasitism or predation. Hence, during evolution, they have evolved a specific defense system against herbivores. Plants have two kinds of defense mechanisms. The first line of defense, called a constitutive defense system, is offered by physical or mechanical barriers, such as thorns, waxy epidermal cuticles, and bark. The second class of defense systems involves chemical warfare, which includes the production of toxic chemicals, pathogen-degrading enzymes, etc. This is known as the inducible defense system, and is activated only when the plants are attacked by herbivores or pathogens (Wu and Baldwin 2010). Plants produce secondary metabolites, known as allelochemicals, to defend themselves against herbivores, pests and pathogens, as well as abiotic stresses. Their role may involve deterrence/anti-feeding activity, toxicity, or precursors to physical defense systems like interference in plant growth (Zavala, Nability and DeLucia 2013). This biological phenomenon of eliciting secondary metabolites or allelochemicals also known as allelopathy (Latif, Chiapusio and Weston 2017).

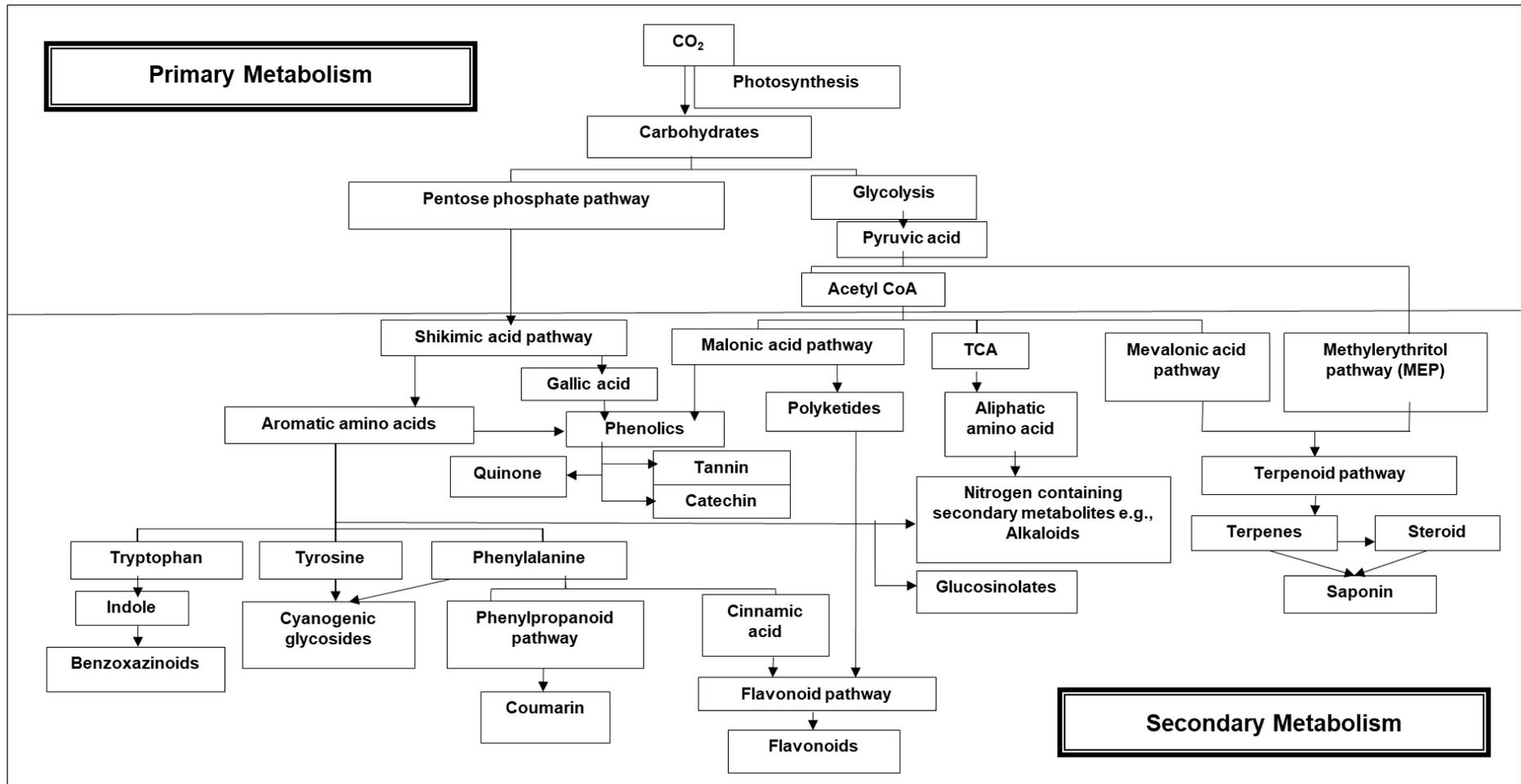


Figure 4.1: Schematic view of biosynthesis of secondary metabolites for plant defense

4.1.1 Role of secondary metabolites

Secondary metabolites are formed from the by-products or intermediates of primary metabolism (**Figure 4.1**). They are the diverse organic compounds that are not directly involved in the normal growth, development, or reproduction of an individual. However, they facilitate ecological interactions, which might provide a selective advantage to the organism by increasing its survivability or fecundity. Secondary metabolites protect primary metabolism by deterring herbivores, reduce tissue loss and avoid infection by microbial pathogens. They also attract pollinators, seed-dispersing birds and animals and also play roles in plant-plant competition. Three groups of secondary metabolites are involved in plant defense (Freeman and Beattie 2008): i) Nitrogen-containing secondary products, ii) Phenolic compounds, and iii) Terpenes.

a. Nitrogen-containing secondary metabolites

Nitrogen-containing secondary metabolites like alkaloids are synthesized primarily from amino acids. They are present in less amount in plants as compared to phenolics and terpenoids. Nevertheless, they are important because of their bioactivity as drugs and toxins. These are synthesized from aliphatic amino acids via the TCA cycle and aromatic amino acids via shikimic acid pathways. Alkaloids are low molecular weight, bitter-tasting nitrogenous compounds synthesized from lysine, tyrosine and tryptophan and are alkaline. Some alkaloids produced by plants are caffeine, cocaine, morphine, nicotine etc. Most of them are utilized by humans as pharmaceuticals, stimulants or narcotics. (**Figure 4.2**).

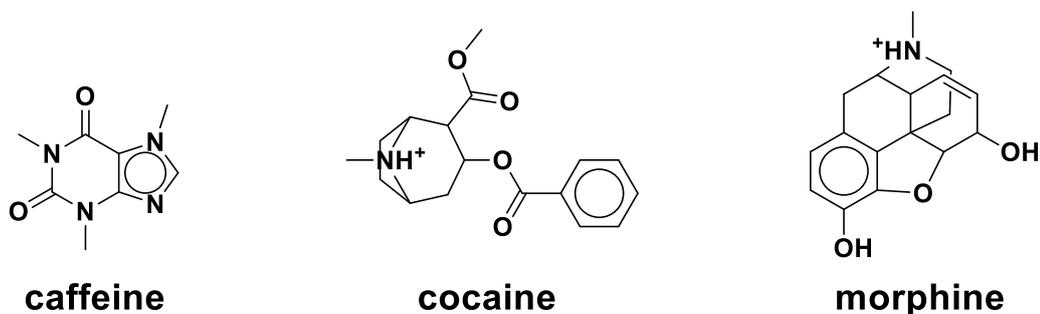


Figure 4.2: Examples of 2D structures of alkaloids

b. Phenolic compounds

Plants contain a diverse form of secondary metabolites containing one or more phenol groups called phenolics. Phenolics are aromatic compounds formed via the shikimic acid pathway or the malonic acid pathway. Many of them function as defense compounds against herbivores and pathogens. Others function in attracting pollinators and fruit dispersers. Among the phenolic compounds, isoflavonoids and tannins play major roles in defense against herbivores and pathogens. Lignin provides mechanical support to plants. Anthocyanins are colored flavonoids that attract pollinators and fruit dispersers and help the plants in propagation (**Figure 4.3**). Flavonoids like quercetin and luteolin protect plants by absorbing harmful UV radiation for their proper development and growth. Phenylalanine plays an important regulatory role in the formation of many phenolic compounds. Phenylalanine production in plants increases during environmental stresses such as nutrient deficiency, low temperature, low light intensity, fungal infection, etc. Some phenolic compounds are activated by light called phototoxic phenolic compounds like umbelliferone, salicylic acid, etc. Their production level may increase up to 100-fold in stressed plants. These phototoxic phenolic compounds may cause skin rashes in humans.

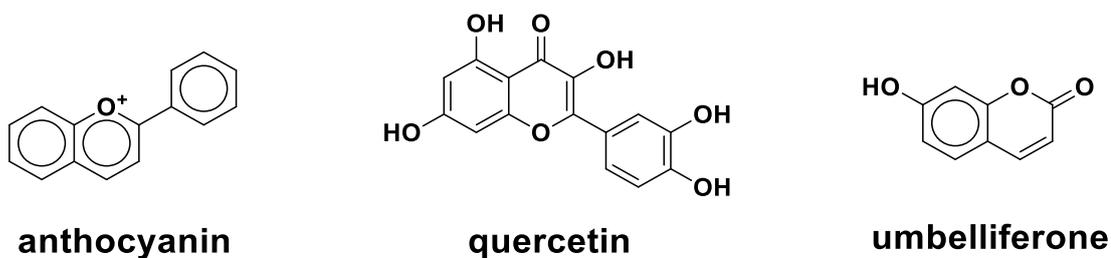


Figure 4.3: Examples of 2D structures of phenolic compounds

c. Terpenes

Terpenes represent the largest class of secondary metabolites. They are produced from the mevalonic acid pathway, which begins with acetyl CoA or primary intermediates of glycolysis. Terpenes function as herbivore deterrents. They can be produced in response to herbivore feeding and attract predatory insects and parasites. Terpenoids are also called isoprenoids and are classified according to their number of constituent isoprene units such as monoterpenoids- two isoprene units, diterpenoids- four isoprene units, triterpenoids- six units, and so on. Monoterpenes and sesquiterpenes are commonly found in glandular hairs on the plant surface and are the primary components of essential oils, which are volatile compounds (Paré and Tumlinson 1999). Many essential oils function as toxins and protect plants against insects, fungal or bacterial attacks. Few examples of plants producing terpenoids are basil (*Ocimum basilicum* L.) produces anethole, ocemene; rosemary (*Salvia rosmarinus*) - rosmarinic acid; thyme (*Thymus vulgaris*) - α -terpinene, carvacrol, thymol, p-cymene, linalool, geraniol, terpineol; black pepper (*Piper nigrum*) – Capsidiol; cinnamon (*Cinnamomum camphora*) - Camphor, α -terpineol, linalool, etc. (**Figure 4.4**).

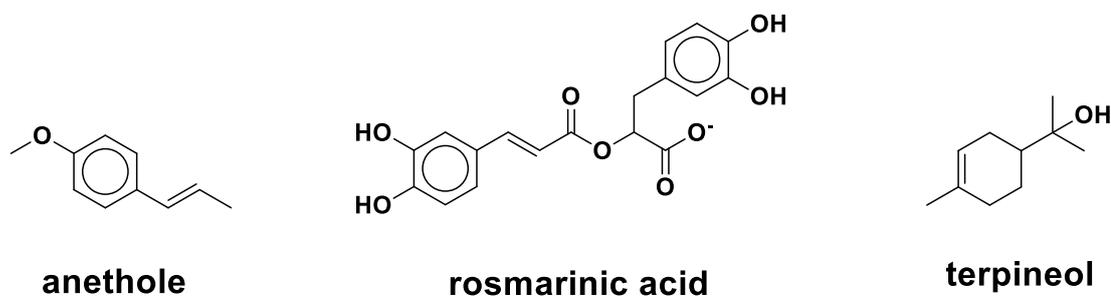


Figure 4.4: Examples of 2D structures of terpenoids

The study of plant secondary metabolites has many practical applications. These bioactive molecules can help in developing lead-like or drug-like compounds used in manufacturing medicines. These molecules can also be used in developing natural environment-friendly pesticides, insecticides, etc., which ultimately reduce the need for certain costly and potentially harmful pesticides. Agrochemicals such as pesticides, insecticides, herbicides, fungicides, etc., play a significant role in controlling agricultural ecosystems. Transgenic plant technology is a method for adapting plant defense with new molecule genes (Gatehouse et al. 1993). When transferred to *Physcomitrella patens* from rice and the moss *Hypnum plumaeforme* through transformation, momilactones genes show similar transcriptional responses to the stresses (Okada et al. 2016). Similarly, when a gene involved in the biosynthesis of the molecule responsible for resistance to a particular disease, or bacterial or fungal infection, is transferred to a susceptible plant, it can produce the molecule and make the plant resistant to biotic stress. Thus, this approach offers an alternative to pesticides and chemicals (Langenbach et al. 2016). However, several concerns are associated with transgenic technology like genetic contamination, horizontal transfer of transgene to other microorganisms etc.

Hence, there is considerable interest in developing new compounds or methods to control pathogens, animals, and insects that are harmful to plant or crop growth. Advancements in combinatorial chemistry and high-throughput screening allow the discovery of such allelochemicals (drugs, pesticides, etc.) with desirable properties in large chemical spaces. In the present study, we have designed allelochemicals specific novel virtual molecules with pesticidal properties and quantitatively associated them with pesticide-likeness by estimating their LC₅₀ and EC₅₀ values for lower aquatic organisms.

4.2 Materials and Methods

The workflow deployed in the present study is presented in **Figure 4.5**. The literature related to the chemical defense of plants from pathogens, animals and insects was collected using Google Scholar and PubMed searches. A list of chemical names of allelochemicals involved in the chemical defense of plants was prepared manually from the literature. In this way, allelochemicals (n=280) were identified from plants (n=162) related to chemical defense (**Supplementary Table S4.1.1**). A list of chemical names of pesticide molecules was collected from Pesticide Properties Database (PPDB) (Lewis et al. 2006) and PAN Pesticides Database (http://www.ipacv.ro/proiecte/risk/files/pan_pesticides_database.htm). All the chemical names were converted to SMILES format by the ChemAxon program of JChem (Csizmadia 2000). Pesticide molecules were imported and processed in Molecular Operating Environment (MOE, Ver. 2010.10) (Chemical Computing Group 2008) to get unique (Molecular rings = 5 to 6, Mol. weight <=1000) molecules (n = 1985) (**Supplementary Table S4.1.2**).

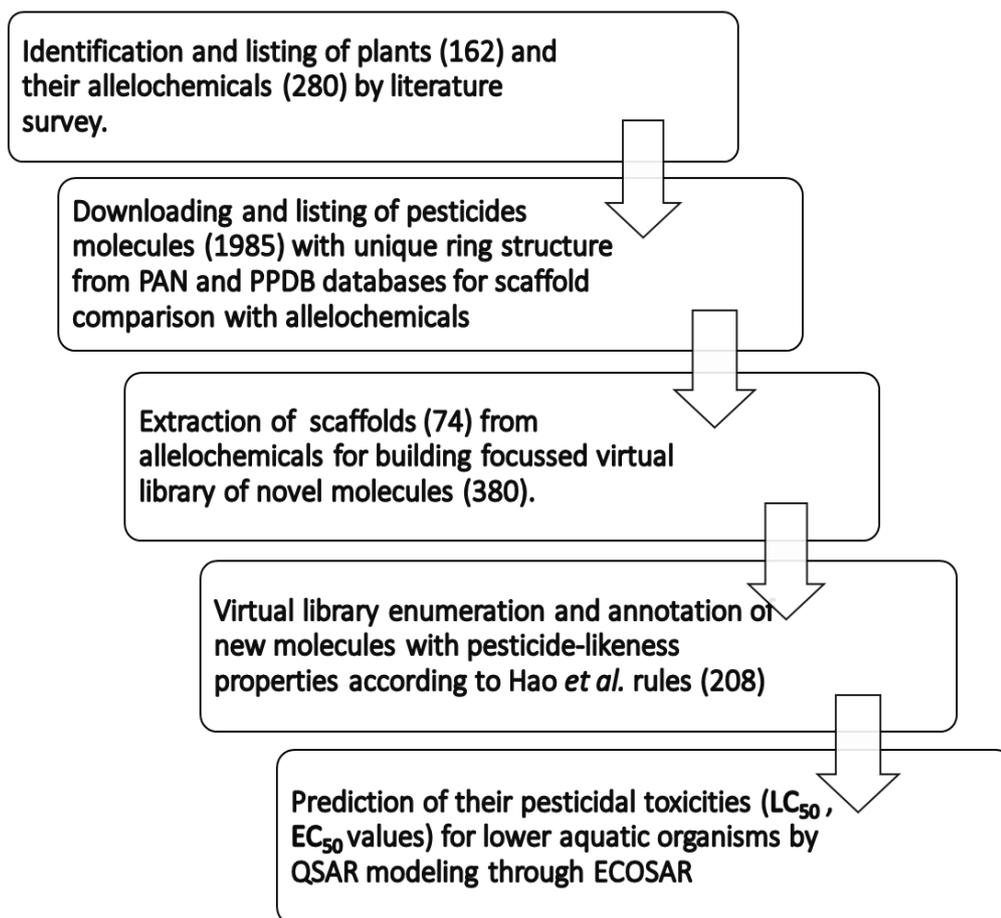


Figure 4.5: An overview of the steps deployed in the present methodology

Scaffold molecule networks between pesticides and allelochemicals were generated using Cytoscape Ver. 3.7.1 (Su et al. 2014). ChemScreener (Karthikeyan et al. 2015b), an in-house developed program, was employed to extract corresponding molecular scaffolds and functional groups from allelochemicals to build a focused virtual library of novel molecules. Allelochemicals and novel virtual molecules were filtered according to (Hao et al. 2011) rules for their pesticide-likeness after generating their 2D descriptors in MOE. The pesticide-likeness rules include molecular weight (MW) ≤ 435 , hydrophobicity (LogP) ≤ 6 , number of H-bond acceptors (HBA) ≤ 6 and donors (HBD) ≤ 2 , number of rotatable bonds (RB) ≤ 9 , and

number of aromatic bounds ≤ 17 . To support the beneficial properties of all novel molecules, their toxicophoric, pharmacophoric, and chemophoric (TPC) scores were also computed by ChemScreener. The toxicity of the molecules to aquatic organisms, such as fish, aquatic invertebrates, and aquatic plants, was estimated using Ecological Structure-Activity Relationships (ECOSAR) Class program (ECOSAR Ver. 2.0) (Mayo-Bean et al. 2017).

4.3 Results and Discussion

The objective of the present study was to design new virtual pesticide-like molecules from the natural bioactive compounds, i.e., allelochemicals involved in chemical defense in plants, that can potentially act as biopesticides. For this, we made use of the information published in the literature for plants related to chemical defense from pathogens, animals, and insects. Subsequently, we developed an *in silico* virtual library of molecules from allelochemicals. For the first time, such methodology was employed based on structural similarity pairing with pesticide molecules and bioactive compounds.

4.3.1 Chemoinformatics analysis

Figure 4.6 provides an overview of the chemical defense of plants by their allelochemicals content to various damage-causing agents, which has been extracted from the published literature. The figure depicts how plants defend themselves from various damages caused by pathogens, insects, animals, and other damages caused by UV and other ionizing radiations, weeds, herbicides, mechanical damages, etc.

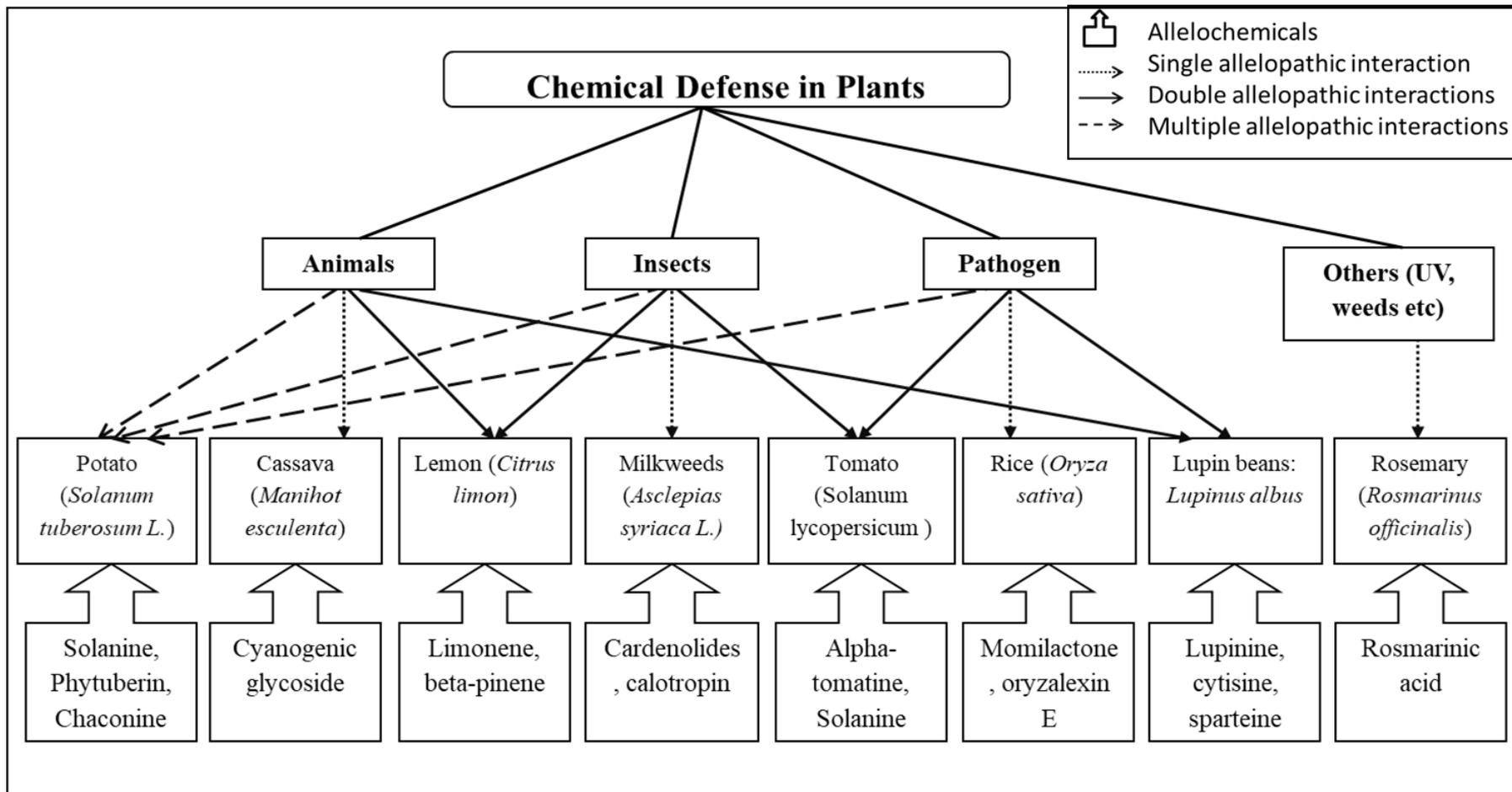


Figure 4.6: Allelopathic interactions between a plant and the pathogen, insect, animals, and others (UV, weeds, herbicide, mechanical damage, etc.).

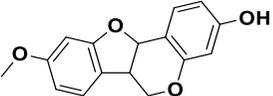
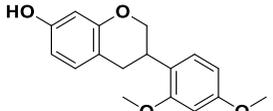
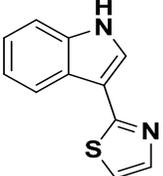
Some plants also defend themselves from the attack of multiple damages-causing agents at a time through multiple allelopathic interactions. For example, potato (*Solanum tuberosum* L.) contains solanine, rishitin, lubimin, chaconine, and phytuberin allelochemicals in its roots, leaves, and tops, which can defend it from herbivorous insects such as Colorado potato beetle (*Leptinotarsa decemlineata*) and livestock animals, pathogens such as *Phytophthora infestans*, etc. (Osman and Moreau 1985, Felton, Workman and Duffey 1992). Some plants are also able to defend themselves in response to other physical damaging agents, such as opium poppy (*Papaver somniferum*) induces morphine in response to mechanical damage (Sánchez-Campillo et al. 2009), rosemary (*Rosmarinus officinalis*) induces rosmarinic acid, a photoprotective agent, on exposure to UV and other ionizing radiations, etc. (Morimoto et al. 2001).

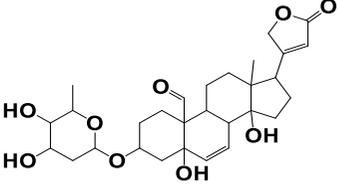
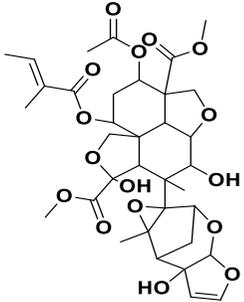
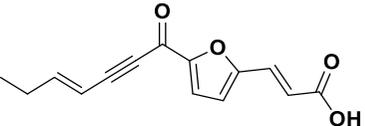
Most compounds are selectively found in few plants, such as cardenolides found in *Nerium indicum*, *Cynanchum* genus, *Funastrum clausum*, *Gomphocarpus* genus, *Marsdenia* genus, *Matelea maritima*, *Sarcostemma* genus, *Asclepias* genus, etc. Some allelochemicals offer resistance to multiple agents by multiple allelopathic interactions. These include DIMBOA (2, 4-dihydroxy-7-methoxy-1, 4-benzoxazin-3-one) in wheat (*Triticum* spp.), which provides resistance to insects (European corn borer (*Ostrinia nubilalis*), maize plant louse (*Rhopalosiphum maydis*), and stalk rot (*Diplodia maydis*)), fungus (Northern corn leaf blight (*Helminthosporium turcicum*)) and herbicide (Atrazine), etc. Whereas, oleandrin in oleander (*Nerium indicum* Mill.) provides resistance to animals (*Lymnaea acuminata* snails, predatory fish *Channa punctatus*) and nematodes (*Bursaphelenchus xylophilus*, *Panagrellus redivivus*, and *Caenorhabditis elegans*) as well. **Table 4.1** lists 2D structures of several

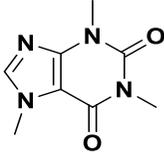
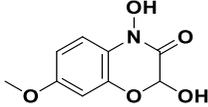
allelochemicals induced in plants, along with their multiple resistant actions against pathogens, animals, insects, and others like herbicides, UV, etc.

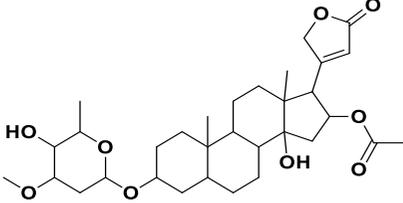
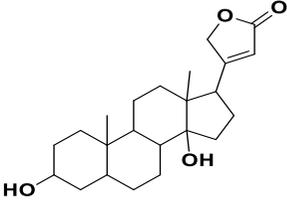
Several organic molecules have been well-studied, and their biological activities have already been explained and documented in the scientific literature. All the 280 ring-containing allelochemicals identified from 162 plants in this study were classified into ten phytochemical classes such as flavonoids, alkaloids, amino acid-related compounds, fatty acids-related compounds, indoles, phenylpropanoids, skimate/ acetate - malonate pathway derived compounds, tannins, terpenoids and others (**Supplementary Table S4.1.3**). It was found that among the 280 unique molecules, most of the molecules were terpenoids (104), followed by flavonoids (46), alkaloids (32), phenylpropanoids (30), etc. The remaining molecules belong to fatty acid-related compounds, amino acid-related compounds, tannins, indoles, etc. Some examples for terpenoid molecules related to chemical defense in plants were rosmarinic acid, avenalumin, brassinolide, etc.; for flavonoids were wightone, arachidin, sakuranetin, pisatin, etc. **Table 4.2** provides few examples of allelochemicals induced in plants in response to attack by pathogens, animals and insects, with their ten different phytochemical classes and their 2D structures. Several allelochemicals were also identified in multiple plants such as cardenolides are present in *Asclepias* genus (milkweeds), *Cynanchum louiseae*, *Funastrum clausum*, *Gomphocarpus cancellatus*, *Marsdenia macrophylla*, *Matelea maritima*, *Sarcostemma acidum*, *Telosma cordata*, *Tylophora indica*, *Vincetoxicum hirundinaria*, etc. (Singh and Rastogi 1970).

Table 4.1: Examples of allelochemicals induced in plants and imparting resistance to pathogens, animals, and insects (For more examples, please refer to Supplementary Table S1.1)

Sr. no.	Plant	Allelochemicals with their 2D structures	Pathogen, animal, insect, or other stress	References
1	Alfalfa (<i>Medicago sativa</i>)	 <p>1. Medicarpin</p>  <p>2. Sativan</p>	Fungus: <i>Phytophthora megasperma</i> , <i>Phoma medicaginis</i> , <i>Nectria haematococca</i> , <i>Colletotrichum trifolii</i>	(Blount et al. 1992)
2	Arabidopsis (<i>Arabidopsis thaliana</i>)	 <p>Camalexin</p>	Gram-negative bacteria: <i>Pseudomonas syringae</i> ; Fungus: <i>Alternaria brassicicola</i> , <i>Botrytis cinerea</i>	(Ahuja et al. 2012)

Sr. no.	Plant	Allelochemicals with their 2D structures	Pathogen, animal, insect, or other stress	References
3	Milkweed (<i>Asclepias syriaca</i> L.)	 <p style="text-align: center;">Cardenolide</p>	Insects: Butterflies (Danaini), bees, wasps, beetles, moths, and true bugs	(Singh and Rastogi 1970)
4	Neem (<i>Azadirachta indica</i>)	 <p style="text-align: center;">Azadirachtin</p>	Insects: Mosquitoes: <i>Anopheles</i> spp., tobacco hornworm (<i>Manduca sexta</i>) in tobacco, fall armyworm (<i>Spodoptera frugiperda</i>) on cotton seedlings	(Maia and Moore 2011, Senthil-Nathan 2013, Raffa 1987, Sengottayan 2013)
5	Broad bean (<i>Vicia faba</i>)	 <p style="text-align: center;">Wyerone</p>	Fungus: <i>Botrytis cinerea</i> , <i>B. fabae</i> , <i>B. allii</i>	(Letcher et al. 1970)

Sr. no.	Plant	Allelochemicals with their 2D structures	Pathogen, animal, insect, or other stress	References
6	Coffee (<i>Coffea arabica</i>), Tea (<i>Camellia sinensis</i>), and Cocoa (<i>Theobroma cacao</i>)	 <p style="text-align: center;">Caffeine</p>	Gram-positive bacteria: <i>Staphylococcus aureus</i> , <i>Bacillus cereus</i> ; Gram-negative bacteria: <i>Escherichia coli</i> , <i>Klebsiella pneumoniae</i> ; Insects: Honey bee (<i>Apis mellifera</i>), Tobacco hornworm (<i>Manduca sexta</i>)	(Sledz et al. 2015)
7	Wheat (<i>Triticum</i> spp.)	 <p style="text-align: center;">DIMBOA (2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one)</p>	Insects: European corn borer (<i>Ostrinia nubilalis</i>), maize plant louse (<i>Rhopalosiphum maydis</i>), and stalk rot (<i>Diplodia maydis</i>); Fungus: Northern corn leaf blight (<i>Helminthosporium turcicum</i>); Herbicide: Atrazine	(Niemeyer 1988)

Sr. no.	Plant	Allelochemicals with their 2D structures	Pathogen, animal, insect, or other stress	References
8	Oleander (<i>Nerium indicum</i> Mill.)	 <p>1. Oleandrin</p>  <p>2. Uzarigenin</p>	Animals: <i>Lymnaea acuminata</i> snails, predatory fish (<i>Channa punctatus</i>); Nematodes: <i>Bursaphelenchus xylophilus</i> , <i>Panagrellus redivivus</i> , and <i>Caenorhabditis elegans</i>	(Wang et al. 2009)

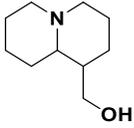
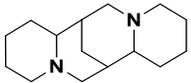
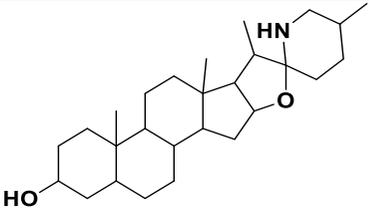
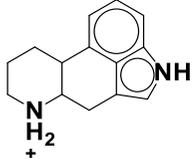
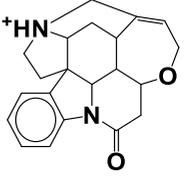
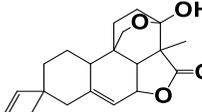
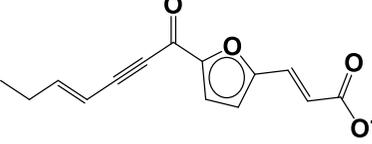
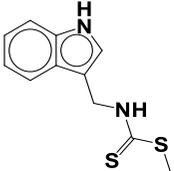
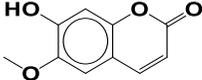
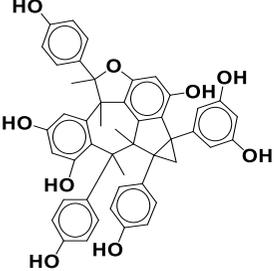
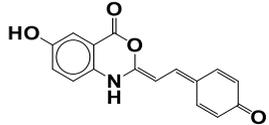
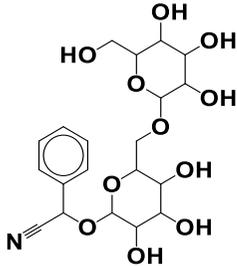
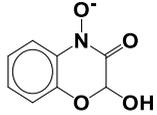
Sr. no.	Plant	Allelochemicals with their 2D structures	Pathogen, animal, insect, or other stress	References
9	Lupinus genus: <i>L. albus</i> , <i>L. mutabilis</i> , <i>L. luteus</i> , <i>L. albus</i> , <i>L. angustifolius</i>	 <p>1. Lupinine</p>  <p>2. Sparteine</p>	Animals: Herbivores (arthropods, vertebrates)- mouse, rats, etc.; Bacteria: <i>Serratia marcescens</i> , <i>Bacillus megaterium</i> ; Fungus: <i>Alternaria porri</i> , <i>Piricularia oryzae</i>	(Wink 1988)
10	Tomato (<i>Solanum lycopersicum</i>)	 <p>tomatidine</p>	Insects: Colorado beetle and snails; Fungus: <i>Cladosporium fulvum</i> ; Bacteria: <i>E. coli</i> and <i>Staphylococcus aureus</i> ; Weed: hemp sesbania	(Osbourn 1996, Hoagland 2009)

Table 4.2: Allelochemicals induced in plants in response to attack by pathogens, animals, and insects, with their phytochemical class and structures (n=10) (For more examples, please refer to Supplementary Table S4.1.3)

Sr. No.	Plant	Allelochemicals	Classification	2D Structure
1	Yellow-Throated Morning Glory (<i>Ipomoea parasitica</i>)	Ergolines	Flavonoids	
2	Strychnine tree (<i>Strychnos nux-vomica</i>)	Strychnine	Alkaloids	
3	Rice (<i>Oryza sativa</i>)	Momilactone B	Terpenoids	
4	Broad bean (<i>Vicia faba</i>)	Wyerone acid	Fatty acids related compounds	

Sr. No.	Plant	Allelochemicals	Classification	2D Structure
5	Rape Mustard (<i>Brassica rapa</i>), <i>B. juncea</i>	Brassinin	Indole	
6	Tobacco (<i>Nicotiana tabacum</i>), <i>Helianthus annuus</i> , <i>Platanus acerifolia</i>	Scopoletin	Phenylpropanoids	
7	Moutan or Chinese tree peony (<i>Paeonia suffruticosa</i>)	Suffruticosol A	Skimate / acetate-malonate pathway derived compounds	

Sr. No.	Plant	Allelochemicals	Classification	2D Structure
8	Oat (<i>Avena sativa</i>)	Avenalumin	Others: benzoxazines	 <p>The structure shows a benzoxazine ring system with a hydroxyl group at the 6-position and a 4-hydroxyphenyl group at the 4-position.</p>
9	Almonds, apricot, cherries, and peaches	Amygdalin	Amino acid-related compounds	 <p>The structure shows a cyanohydrin derivative of glucose, specifically a 1,6-diphosphate derivative with a cyanide group at the 1-position.</p>
10	Maize (<i>Zea mays</i>)	2,4-dihydroxy-1,4-benzoxazin-3-one (DIBOA)	Tannin	 <p>The structure shows a benzoxazinone ring system with hydroxyl groups at the 2 and 4 positions.</p>

4.3.2 Scaffold molecule network

Scaffolds and building blocks extracted from the 280 organic allelochemicals and 1985 unique pesticides, were 74, 33, and 56, 62, respectively (**Supplementary Tables S4.1.4, S4.1.5**). A network was constructed to visualize the inter-relationship between scaffolds and molecules of allelochemicals and pesticides containing 2306 nodes and 2350 edges where nodes act as molecules and scaffolds and edges are their interactions (**Figure 4.7, Supplementary Table S4.1.6**). The network analysis of the topological features computed for the network showed an average number of neighbors with 2.038 and characteristic path length with 2.511 scores depicting the maximum connectivity of all molecules and their common scaffolds. It was identified that five scaffolds and 15 molecules were common between allelochemicals and pesticide molecules. **Table 4.3** shows the common scaffolds between allelochemicals and pesticide molecules. This comparison shows that the common molecules can be developed as pesticides after further research. However, it was also found that among the 280 organic allelochemicals, 39 molecules were already used as biopesticides (**Table 4.4; Supplementary Table S4.1.7**). The molecular diversity of these biopesticides in the chemical space of allelochemicals is presented in **Figure 4.8**. It was found that the molecules with more unique features in their chemical structures occupied separate regions in the plot. Some of the representative biopesticides and allelochemicals were randomly picked up, such as azadirachtin, avenacoside B, tannic acid, etc., and are highlighted in the 2D PCA plot figure. Scaffolds were extracted from all these molecules. It was found that most of the outlier molecules in the PCA plot figure had complex polycyclic structures.

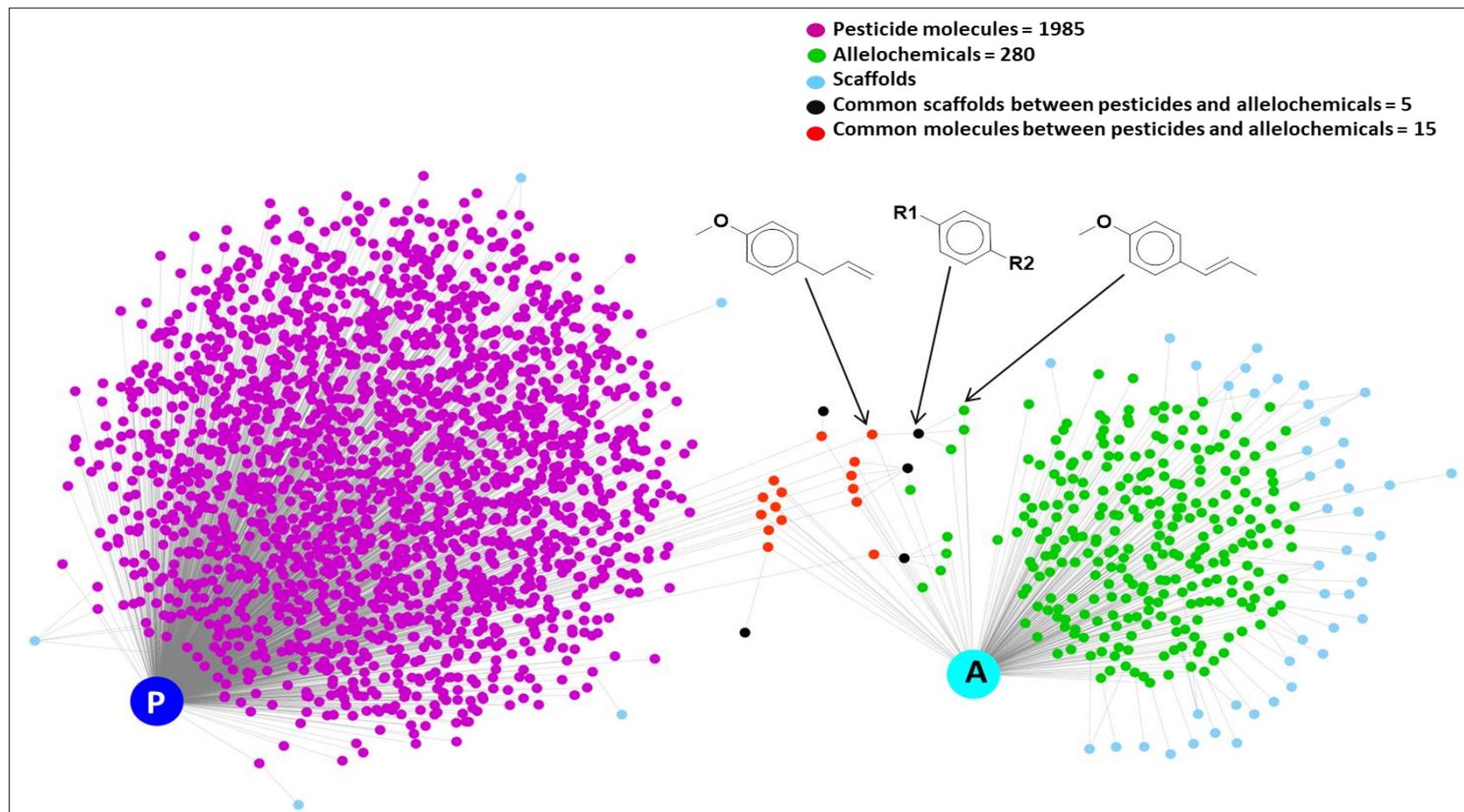
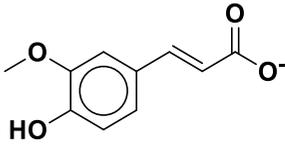
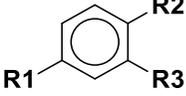
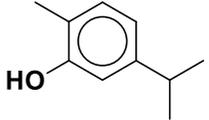
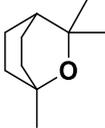
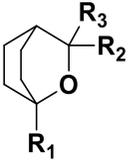
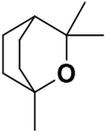


Figure 4.7: Scaffold molecule network between the allelochemicals and pesticides generated using Cytoscape (Nodes- Molecules, scaffolds: 2306, Edges – Molecule-scaffold Interactions/ hidden relationships: 2350)

Table 4.3: Similar scaffolds (n=5) identified from 280 allelochemicals and 1985 unique pesticides by scaffold molecule network

Sr. No	Allelochemicals	Similar Scaffold	Pesticide
1	<p>Plant Source: Sorghum, Soybean (Einhellig and Eckrich 1984, Dos Santos, Ferrarese and Ferrarese-Filho 2008)</p>  <p>Ferulic acid</p>		 <p>2-para-cymenol</p>
2	<p>Plant Source: <i>Ruta graveolens</i> (De Feo, De Simone and Senatore 2002)</p>  <p>1,8-cineole</p>		 <p>1,8-cineole</p>

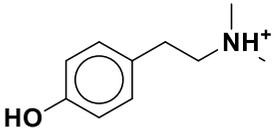
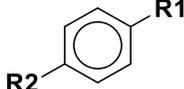
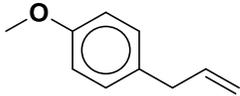
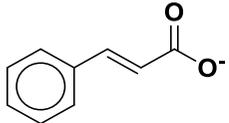
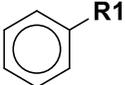
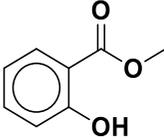
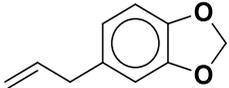
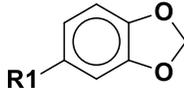
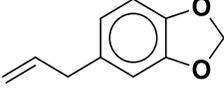
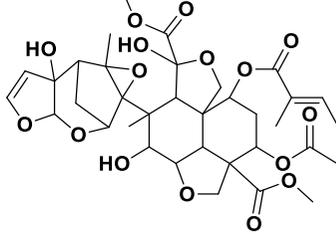
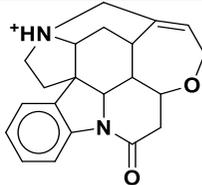
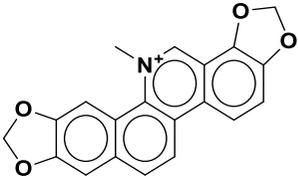
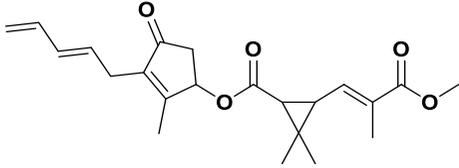
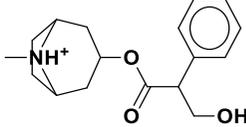
Sr. No	Allelochemicals	Similar Scaffold	Pesticide
3	Plant Source: Barley (Chaniago, Lovett and Roberts 2011)  Hordenine		 4-allylanisole
4	Plant Source: Lettuce (Li et al. 1993)  trans-cinnamic acid		 Methyl salicylate
5	Plant Source: <i>P. polyxenes</i> (Wen, Berenbaum and Schuler 2006)  Safrole		 Safrole

Table 4.4: Organic allelochemicals already in the market as biopesticides (n= 5; selected molecules) (For more examples, please refer Supplementary Table S4.1.7)

Sr no.	Source	Allelochemicals	2D structure
1.	<i>Azadirachta indica</i> (Neem)	Azadirachtin	
2	<i>Strychnos nux-vomica</i> L. (Strychnine tree)	Strychnine	
3	<i>Papaver somniferum</i> (Opium poppy)	Sanguinarine	

Sr no.	Source	Allelochemicals	2D structure
4	<i>Chrysanthemum</i> (Mums or Chrysanthus)	Pyrethrum	 <p>The image shows the 2D chemical structure of Pyrethrum, a natural insecticide. It consists of a pyrethric acid moiety (a cyclohexenone ring with a methyl group and a propenyl side chain) linked via an ester bond to a chrysanthemic acid moiety (a cyclopropane ring with two methyl groups and a propenyl side chain).</p>
5	<i>Atropa belladonna</i> (Belladonna or Deadly nightshade)	Atropine	 <p>The image shows the 2D chemical structure of Atropine. It features a tropane bicyclic ring system (8-azabicyclo[3.2.1]octane) with a protonated nitrogen atom (NH⁺). This ring is connected via an ester linkage to a side chain that includes a phenyl ring and a hydroxyl group (-OH).</p>

● Biopesticide ● Allelochemicals

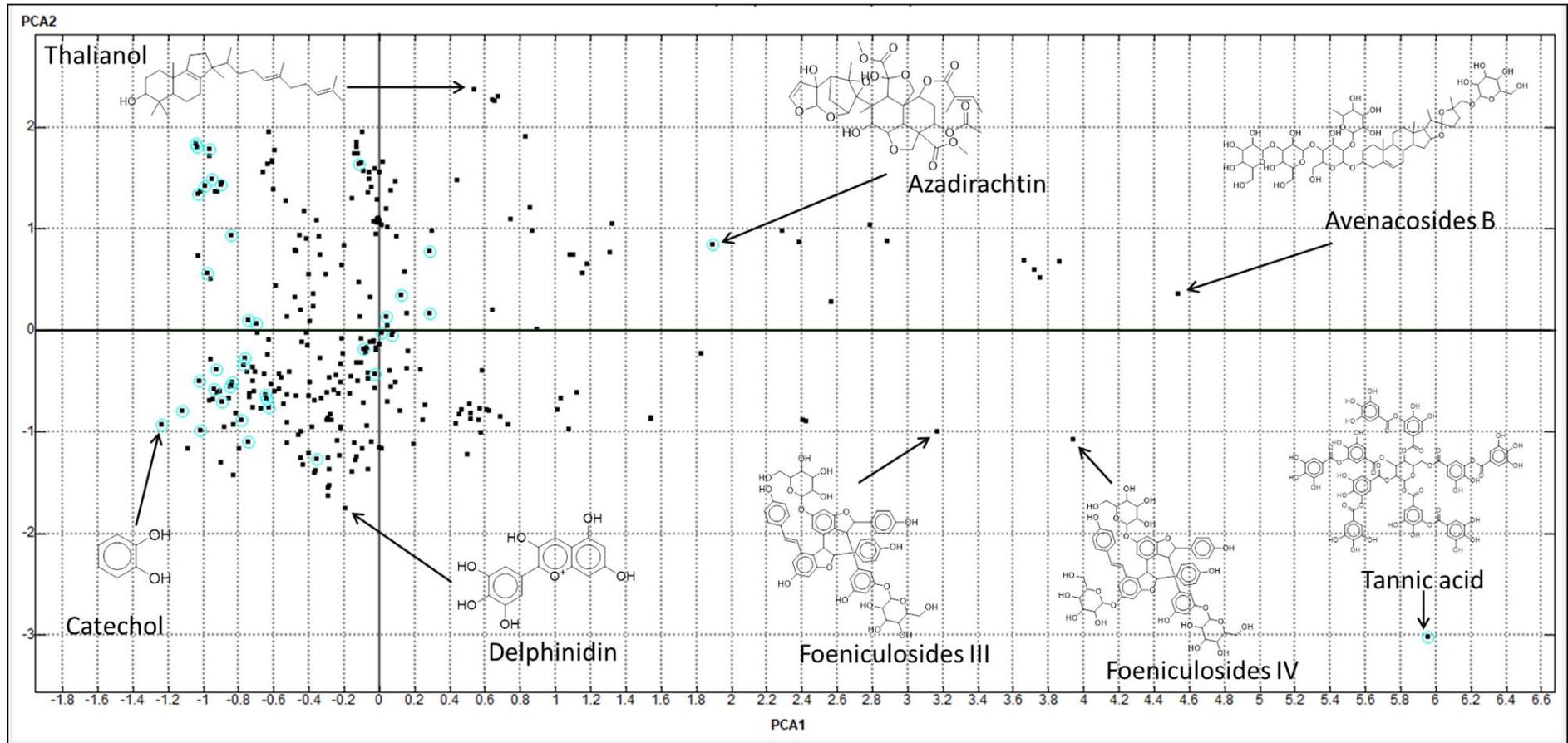


Figure 4.8: The 2D PCA plot representing the molecular diversity of bio-pesticides (n=39) in chemical space of allelochemicals

4.3.3 Virtual Library

A knowledge-based approach for designing combinatorial libraries and virtual screening helps to explore the chemical spaces for molecules with desirable properties. In this study, we generated a virtual library of n=380 novel molecules from five selected scaffolds (having up to three functional groups) and building blocks extracted from allelochemicals (**Supplementary Table S4.2.1**). It was emphasized that the pesticide molecules of natural origin should be environment friendly. Therefore, to screen them, we used two methods, 1. Virtual screening by molecular scoring and 2. Virtual screening by QSAR estimation of pesticidal toxicities.

a. Virtual screening by molecular scoring

For virtual screening of novel virtual library molecules, we have calculated their PDL (Progressive Drug Like), PLL (Progressive Lead Like), and TPC (Toxicophoric, Pharmacophoric, and Chemophoric) scores for all the virtual library molecules and pesticides as well. **Figure 4.9** shows that the new molecules designed from allelochemicals showed more pharmacophoric features and less toxicophoric and chemophoric features (**Supplementary Table S4.2.2, S4.2.3**). The computed TPC scores show that these new molecules are neither toxic nor chemophoric or more reactive instead of having drug-like characteristics. For both the model graphs above, T, P, and C scores increased and decreased in the same proportions. Therefore, these values indicate that the allelochemicals specific virtual library molecules will probably act as potential environment-friendly pesticides. They were then screened with pesticide-like molecules according to the rules defined by (Hao et al. 2011) after generating their descriptor values. The screened pesticide-like novel virtual molecules (n= 208/380) were listed in **Supplementary Table S4.2.4**.

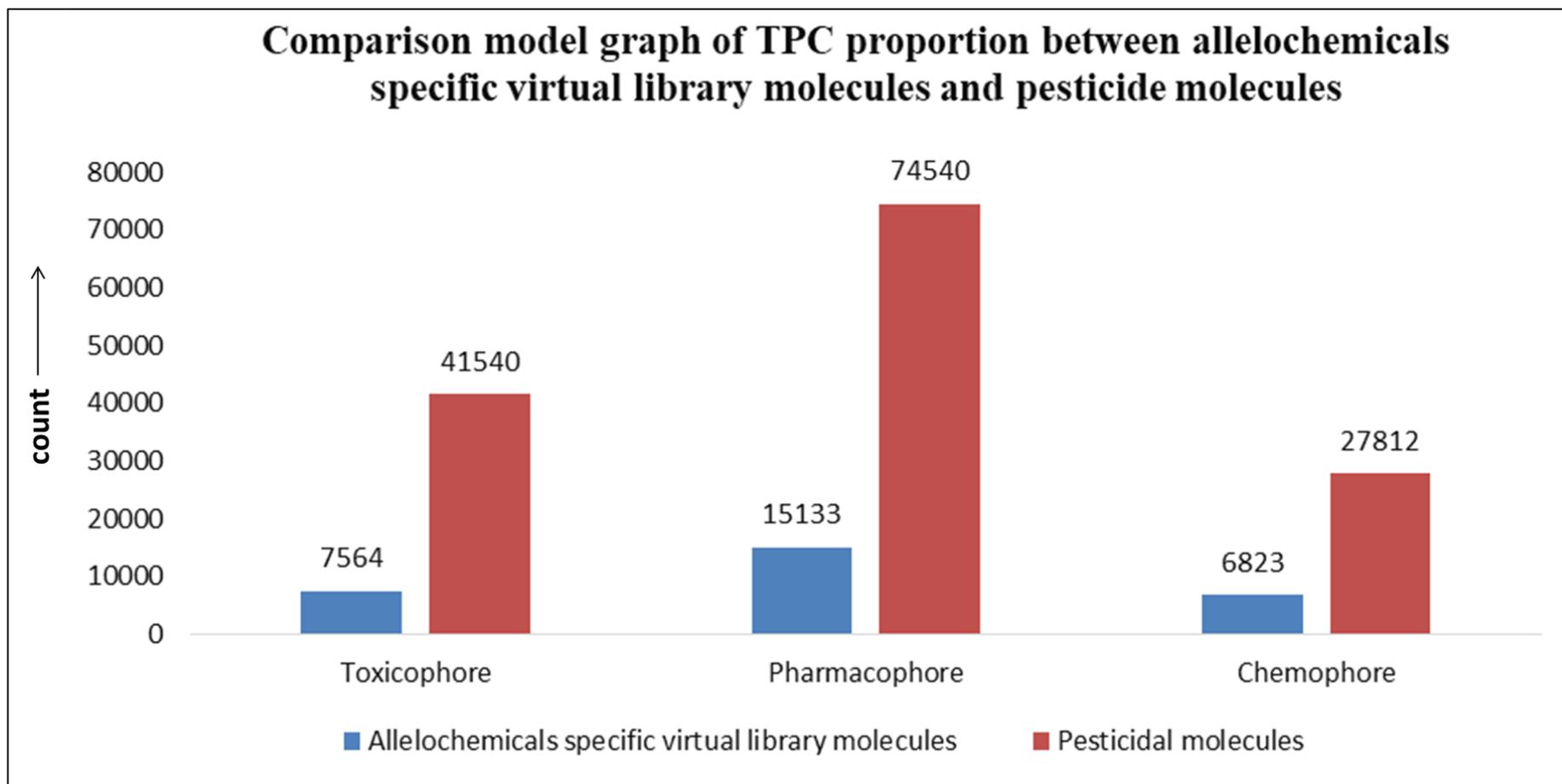
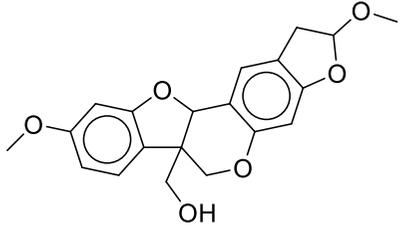
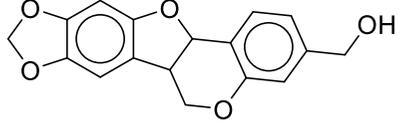


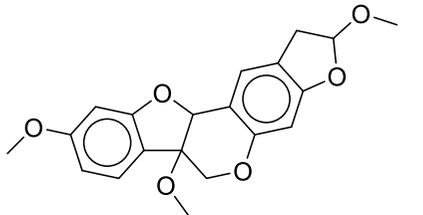
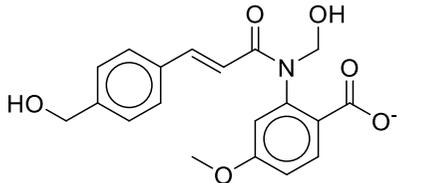
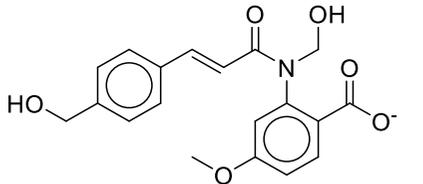
Figure 4.9: Comparison of TPC proportion model graph between allelochemical specific virtual library molecules and pesticide molecules.

b. Virtual screening by QSAR estimation of pesticidal toxicities

In this method, we used a QSAR based method to predict pesticidal activities of novel virtual molecules generated from allelochemicals for lower aquatic organisms. This method is based on the analysis of molecular structures of known molecules having experimentally measured activities. Before performing QSAR, descriptors were generated for the new virtual library molecules to screen them with pesticide-likeness rules. For predicting the aquatic toxicities of screened pesticide-like novel virtual molecules (n=208), we performed QSAR analysis using ECOSAR, a computerized predictive system that estimates aquatic toxicities against fish, aquatic invertebrates, and aquatic plants using structure-activity relationships. With this, we evaluated pesticide-like allelochemicals and novel virtual molecules for their chemical toxicity towards aquatic organisms (**Supplementary Table S4.2.5**). Depending on the species (fish, daphnid or algae), acute aquatic toxicity was calculated as LC₅₀ or EC₅₀ (mg/L). Chronic aquatic toxicity estimate was expressed in ChV (in mg/L). Multiple QSAR classes were generated for molecules with multiple functional groups (esters, ketones, etc.), with different LC₅₀, EC₅₀, and ChVs for each QSAR class. A lower value indicates a higher level of toxicity (Di Toro, McGrath and Stubblefield 2007, Tisler and Zagorc-Koncan 1997). We screened n=169/208 pesticide-like molecules with pesticidal toxic activities (LC₅₀, EC₅₀, ChV) ≥ 1 for lower aquatic organisms calculated from ECOSAR (**Table 4.5, Supplementary Table S4.2.6**). These estimated values of screened virtual library molecules can be used for further assessment for developing new pesticides.

Table 4.5: Examples of virtual library novel pesticide-like molecules with their TPC scores and LC50 and Chronic values for fish and daphnids

Sr. no.	VL molecule structure	Toxicophores	Pharmacophores	Chemophores	ECOSAR Class	Fish LC ₅₀ - 96h (mg/L)	Daphnid LC ₅₀ - 48h (mg/L)	Fish Chronic value (ChV)
1.		16	36	16	Neutral Organics	41.39	25.65	4.48
2.		14	28	14	Benzyl Alcohols	16.43	13.93	1.74

Sr. no.	VL molecule structure	Toxicophores	Pharmacophores	Chemophores	ECOSAR Class	Fish LC ₅₀ - 96h (mg/L)	Daphnid LC ₅₀ - 48h (mg/L)	Fish Chronic value (ChV)
3.		15	31	15	Neutral Organics	26.91	16.99	2.98
4.		31	45	23	Benzyl Alcohols	1304.95	1028.12	113.38
5.		31	45	23	Acrylamides	105.95	285.73	5.21

4.4 Conclusions

In this study, we developed a computational protocol to identify allelochemicals with pesticide-like properties through the chemoinformatics approach. We demonstrated their pesticidal activities by screening them with pesticide-likeness and QSAR modeling. Further modeling studies with a large amount of experimental data are needed to increase the predictive ability of the molecules with pesticidal properties. The identified allelochemicals with pesticidal properties need to be isolated from the respective plant for further experimental validation and synthesized as potential biocontrol agents. These natural or nature-identical pesticides will be cost-effective as well as environment friendly.

CHAPTER 5

SUMMARY AND FUTURE DIRECTIONS

Chapter 5: Summary and Future Directions

5.1 Summary

Bioactive compounds produced by plants are valuable as food, medicine, and biopesticides. While several medicinally important plants have been studied for their bioactive properties, many more are yet to be explored. In this study, we used chemoinformatics methods to design novel drug-like and lead-like molecules inspired by organic metabolites of Indian medicinal and aromatic plants and food crops such as soybean. The names of organic metabolites were identified and selected from literature through text mining. Common scaffolds were identified between organic metabolites and drug molecules that show drug- or lead-like properties based on the computed 2D descriptors such as molecular weight, number of hydrogen bond donor atoms, number of rotatable bonds, number of rings, etc. Integrated knowledge of chemical scaffold composition and similarity with drug molecules has the potential to reveal prospective metabolite relationships with biological activities. With the help of in-house developed chemoinformatics tools, diverse scaffolds were used for building a focused virtual library. Further, the new virtual molecules annotated with desired properties can serve as readymade libraries for experimental screening in the context of drug discovery. Through network analysis, protein targets involved in various chronic disease-related pathways for the bioactive compounds present in food crops and Indian medicinal plants were also depicted.

To corroborate the results, we also performed UHPLC-MS/MS experiments of leaf and seed tissues of four soybean varieties grown in a plot at the Biochemical Sciences Division in CSIR-NCL, Pune. A general sample extraction procedure was performed with 70% methanol, and untargeted LC-MS metabolomics analysis was

followed. The raw data were processed using various tools like XCMS Online, PUTMEDID-LCMS workflow in the taverna environment, ProbMetab in R environment, etc. We detected 6628 annotated mass features and identified 443 small molecules out of 1622 previously reported molecules in soybean through untargeted UHPLC/MS experiments with these analyses. We also identified 14 new soybean molecules that were not reported in soybean before and 06 previously reported soybean molecules confirmed by Tandem mass spectrometry (MS/MS). The methodology allows efficient detection and annotation of a large number of small organic molecules by merging previously known biosynthetic pathways from KEGG to that of the plant species by compound substructure sharing. Untargeted metabolomics revealed the global picture of the metabolite composition of leaf and seeds in the four soybean varieties, which included different classes of metabolites. This will help to understand the significant correlation between varieties and tissues that will make better decisions to select varieties and tissues containing high amounts of the target metabolite for drug development.

Integrated knowledge of chemical scaffold composition and similarity with drug molecules reveals prospective metabolite relationship with biological activities. The scaffold network analysis revealed 14 scaffolds and 73 common molecules between previously reported soybean molecules and approved drugs. Similarly, 184 common soybean molecules were identified from the network of soybean molecules identified by UHPLC/MS experiments. A virtual library of novel molecules (n=1225) was generated from previously reported soybean molecules (n=5), as an example to establish the systematic approaches for drug designing. Further, the new virtual molecules annotated with desired properties (TPC, PDL, and PLL scores) can serve as readymade libraries for experimental screening.

This study demonstrates the application of chemoinformatics tools and UHPLC-MS/MS metabolomics to reveal both the previously known and novel molecules. These molecules can be analyzed *in silico* and those with desired properties (such as drug-like, lead-like, pesticide-like, etc.) can be chosen for subsequent targeted isolation. Thus, the study would enable the identification and isolation of the desired bioactive molecules from plants. Similarly, we have also designed allelochemicals-specific environment-friendly novel molecules inspired by pesticidal activities through the chemoinformatics approach. Further, we also developed a chemoinformatics open-source toolkit DoMINE (**D**atabase of **M**edicinally **I**mportant **N**atural products from **p**lanta**E**) using Java. It can be used to build and access the Indian medicinal plant, soybean, and pesticide-inspired allelochemical molecular database created in this study, as well as to generate scaffold and virtual library.

5.2 Future directions

The major aim of this study was to design novel molecules (drug-like, lead-like, and pesticide-like) based on the metabolomics of Indian medicinal and aromatic plants, food crops i.e., soybean, and plants involved in chemical defense through chemoinformatics approach. This study has predicted several potential novel molecules concerning drugs and pesticides, which can be investigated in detail in the future. Thus, the study can have the following future directions:

1. Isolation of target molecules from plants and their confirmation through wet-lab analyses
2. Performing bioactivity studies to confirm their bioactive potentials
3. Employing the active molecules towards synthesis of drugs, pesticides, etc. and their efficacy studies
4. Investigating other sources of natural products for drug development.

BIBLIOGRAPHY

Bibliography

- Abdal Dayem, A., H. Y. Choi, G. M. Yang, K. Kim, S. K. Saha & S. G. Cho (2016) The Anti-Cancer Effect of Polyphenols against Breast Cancer and Cancer Stem Cells: Molecular Mechanisms. *Nutrients*, 8, 581.
- Abel, U., C. Koch, M. Speitling & F. G. Hansske (2002) Modern methods to produce natural-product libraries. *Current Opinion in Chemical Biology*, 6, 453-458.
- Abeytunga, D. T. U., L. Oland, A. Somogyi & R. Polt (2008) Structural studies on the neutral glycosphingolipids of *Manduca sexta*. *Bioorganic Chemistry*, 36, 70-76.
- Abuajah, C. I., A. C. Ogbonna & C. M. Osuji (2015) Functional components and medicinal properties of food: a review. *Journal of Food Science and Technology*, 52, 2522-2529.
- Acevska, J., G. Stefkov, I. Cvetkovikj, R. Petkovska, S. Kulevanova, J. Cho & A. Dimitrovska (2015) Fingerprinting of morphine using chromatographic purity profiling and multivariate data analysis. *Journal of Pharmaceutical and Biomedical Analysis*, 109, 18-27.
- Adachi, J., Y. Mizoi, T. Naito, K. Yamamoto, S. Fujiwara & I. Ninomiya (1991) Determination of β -carbolines in foodstuffs by high-performance liquid chromatography and high-performance liquid chromatography mass spectrometry. *Journal of Chromatography A*, 538, 331-339.
- Adjakly, M., M. Ngollo, J. P. Boiteux, Y. J. Bignon, L. Guy & D. Bernard-Gallon (2013) Genistein and daidzein: different molecular effects on prostate cancer. *Anticancer Res*, 33, 39-44.
- Aguiar-Pulido, V., M. Gestal, M. Cruz-Monteagudo, J. R. Rabunal, J. Dorado & C. R. Munteanu (2013) Evolutionary computation and QSAR research. *Curr Comput Aided Drug Des*, 9, 206-25.
- Ahn, E. & J. Schroeder (2002) Bioactive sphingolipids are constituents of soy and dairy products. *Journal of food science*, 67, 522-524.
- Ahuja, I., R. Kissen & A. M. Bones (2012) Phytoalexins in defense against pathogens. *Trends Plant Sci*, 17, 73-90.
- Alamzeb, M., M. R. Khan, R. Mamoon Ur, S. Ali & A. A. Khan (2015) A new isoquinoline alkaloid with anti-microbial properties from *Berberis jaeschkeana* Schneid. var. *jaeschkeana*. *Nat Prod Res*, 29, 692-7.
- Allen, F., A. Pon, M. Wilson, R. Greiner & D. Wishart (2014) CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic acids research*, 42, W94-W99 (Accessed: 12-01-2021).
- Amararathna, M., M. R. Johnston & H. P. Rupasinghe (2016) Plant Polyphenols as Chemopreventive Agents for Lung Cancer. *Int J Mol Sci*, 17, 1352.
- Arai, S., H. Suzuki, M. Fujimaki & Y. Sakurai (1966) Studies on Flavor Components in Soybean: Part III. Volatile Fatty Acids and Volatile Amines. *Agricultural and Biological Chemistry*, 30, 863-868.
- Ashihara, H. (2006) Metabolism of alkaloids in coffee plants. *Brazilian Journal of Plant Physiology*, 18, 1-8.
- Augusti, K. (1996) Therapeutic values of onion (*Allium cepa* L.) and garlic (*Allium sativum* L.). *Indian Journal of Experimental Biology*, 34, 634-640.
- Bahmani, M., H. Golshahi, K. Saki, M. Rafieian-Kopaei, B. Delfan & T. Mohammadi (2014) Medicinal plants and secondary metabolites for diabetes mellitus control. *Asian Pacific Journal of Tropical Disease*, 4, S687-S692.

- Balandrin, M. F., A. D. Kinghorn & N. R. Farnsworth. 1993. Plant-derived natural products in drug discovery and development: an overview. 2-12.
- Barnes, S., J. Sfakianos, L. Coward & M. Kirk. 1996. Soy isoflavonoids and cancer prevention. In *Dietary Phytochemicals in Cancer Prevention and Treatment*, 87-100. Springer.
- Baruati, D. & S. Gogoi (2020) The Aroma Mission-For Boosting Medicinal Aromatic Plant Cultivation in India. *Agriculture and Food E-Newsletter*.
- Bellis, L. J., R. Akhtar, B. Al-Lazikani, F. Atkinson, A. P. Bento, J. Chambers, M. Davies, A. Gaulton, A. Hersey & K. Ikeda (2011) Collation and data-mining of literature bioactivity data for drug discovery. *Biochemical Society Transactions*, 39, 1365-1370.
- Benalla, W., S. Bellahcen & M. Bnouham (2010) Antidiabetic medicinal plants as a source of alpha glucosidase inhibitors. *Current diabetes reviews*, 6, 247-254.
- Benkendorff, K. (2013) Natural product research in the Australian marine invertebrate *Dicathais orbita*. *Mar Drugs*, 11, 1370-98.
- Bennett, R. N. & R. M. Wallsgrove (1994) Secondary metabolites in plant defence mechanisms. *New phytologist*, 127, 617-633.
- Beyer, S. F., A. Beesley, P. F. Rohmann, H. Schultheiss, U. Conrath & C. J. Langenbach (2019) The Arabidopsis non-host defence-associated coumarin scopoletin protects soybean from Asian soybean rust. *The Plant Journal*, 99, 397-413.
- Bhalla, R., K. Narasimhan & S. Swarup (2005) Metabolomics and its role in understanding cellular responses in plants. *Plant Cell Rep*, 24, 562-71.
- Bhat, K. (1997) Medicinal and plant information databases. *Medicinal Plants for Forests Conservation and Health Care*.
- Blondelle, S. E., E. Perez-Paya & R. A. Houghten (1996) Synthetic combinatorial libraries: novel discovery strategy for identification of antimicrobial agents. *Antimicrobial Agents and Chemotherapy*, 40, 1067.
- Blount, J. W., R. A. Dixon & N. L. Paiva (1992) Stress responses in alfalfa (*Medicago sativa* L.) XVI. Antifungal activity of medicarpin and its biosynthetic precursors; implications for the genetic manipulation of stress metabolites. *Physiological and Molecular Plant Pathology*, 41, 333-349.
- Bonifácio, M. J., P. N. Palma, L. Almeida & P. Soares-da-Silva (2007) Catechol-O-methyltransferase and Its Inhibitors in Parkinson's Disease. *CNS drug reviews*, 13, 352-379.
- Boros, B., A. Farkas, S. Jakabová, I. Bacskay, F. Kilár & A. Felinger (2010) LC-MS quantitative determination of atropine and scopolamine in the floral nectar of *Datura* species. *Chromatographia*, 71, 43-49.
- Brechenmacher, L., Z. Lei, M. Libault, S. Findley, M. Sugawara, M. J. Sadowsky, L. W. Sumner & G. Stacey (2010) Soybean metabolites regulated in root hairs in response to the symbiotic bacterium *Bradyrhizobium japonicum*. *Plant Physiology*, 153, 1808-1822.
- Brouns, F. (2002) Soya isoflavones: a new and promising ingredient for the health foods sector. *Food Research International*, 35, 187-193.
- Brown, M., W. B. Dunn, P. Dobson, Y. Patel, C. Winder, S. Francis-McIntyre, P. Begley, K. Carroll, D. Broadhurst & A. Tseng (2009) Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *Analyst*, 134, 1322-1332.
- Brown, M., D. C. Wedge, R. Goodacre, D. B. Kell, P. N. Baker, L. C. Kenny, M. A. Mamas, L. Neyses & W. B. Dunn (2011) Automated workflows for accurate

- mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics*, 27, 1108-1112.
- Bruhn, J. G., J.-E. Lindgren, B. Holmstedt & J. M. Adovasio (1978) Peyote alkaloids: identification in a prehistoric specimen of *Lophophora* from Coahuila, Mexico. *Science*, 199, 1437-1438.
- Buntrock, R. E. (2002) ChemOffice Ultra 7.0. *Journal of Chemical Information and Computer Sciences*, 42, 1505-1506.
- Buzzell, R., B. Buttery & D. MacTavish (1987) Biochemical genetics of black pigmentation of soybean seed. *Journal of Heredity*, 78, 53-54.
- Caruana, M., R. Cauchi & N. Vassallo (2016) Putative Role of Red Wine Polyphenols against Brain Pathology in Alzheimer's and Parkinson's Disease. *Front Nutr*, 3, 31.
- Cavin, J. C., S. M. Krassner & E. Rodriguez (1987) Plant-derived alkaloids active against *Trypanosoma cruzi*. *J Ethnopharmacol*, 19, 89-94.
- Chaniago, I., J. Lovett & J. Roberts (2011) Barley allelochemicals of gramine and hordenine: their effects on broiler chickens. *Animal Production*, 13.
- Chaturvedi, P. R., C. J. Decker & A. Odinecs (2001) Prediction of pharmacokinetic properties using experimental approaches during early drug discovery. *Current Opinion in Chemical Biology*, 5, 452-463.
- Chauhan, N. S. 1999. *Medicinal and aromatic plants of Himachal Pradesh*. Indus publishing.
- Chemical Computing Group, M. (2008) Molecular Operating Environment. *Chemical Computing Group Montreal, Quebec, Canada*.
- Chen, H., D. N. Soroka, Y. Zhu, Y. Hu, X. Chen & S. Sang (2013) Cysteine-conjugated metabolite of ginger component [6]-shogaol serves as a carrier of [6]-shogaol in cancer cells and in mice. *Chemical research in toxicology*, 26, 976-985.
- Cheng, T., Q. Li, Z. Zhou, Y. Wang & S. H. Bryant (2012) Structure-based virtual screening for drug discovery: a problem-centric review. *The AAPS journal*, 14, 133-141.
- Choi, J. M., E. O. Lee, H. J. Lee, K. H. Kim, K. S. Ahn, B. S. Shim, N. I. Kim, M. C. Song, N. I. Baek & S. H. Kim (2007) Identification of campesterol from *Chrysanthemum coronarium* L. and its antiangiogenic activities. *Phytotherapy Research*, 21, 954-959.
- Chong, J., D. S. Wishart & J. Xia (2019) Using metaboanalyst 4.0 for comprehensive and integrative metabolomics data analysis. *Current Protocols in Bioinformatics*, 68, (Accessed: 12-01-2021).
- Choudhary, P. & S. Tran (2011) Phytosterols: perspectives in human nutrition and clinical therapy. *Current Medicinal Chemistry*, 18, 4557-4567.
- Chung, M. Y., M. C. Rho, S. W. Lee, H. R. Park, K. Kim, I. A. Lee, D. H. Kim, K. H. Jeune, H. S. Lee & Y. K. Kim (2006) Inhibition of diacylglycerol acyltransferase by betulinic acid from *Alnus hirsuta*. *Planta Med*, 72, 267-9.
- Cirmi, S., N. Ferlazzo, G. E. Lombardo, A. Maugeri, G. Calapai, S. Gangemi & M. Navarra (2016a) Chemopreventive agents and inhibitors of cancer hallmarks: may Citrus offer new perspectives? *Nutrients*, 8, 698.
- Cirmi, S., N. Ferlazzo, G. E. Lombardo, E. Ventura-Spagnolo, S. Gangemi, G. Calapai & M. Navarra (2016b) Neurodegenerative Diseases: Might Citrus Flavonoids Play a Protective Role? *Molecules*, 21, 1312.
- Clish, C. B. (2015) Metabolomics: an emerging but powerful tool for precision medicine. *Molecular Case Studies*, 1, a000588.

- Coe, S. & L. Ryan (2016) Impact of polyphenol-rich sources on acute postprandial glycaemia: a systematic review. *J Nutr Sci*, 5, e24.
- Colpas, F. T., E. O. Ono, J. D. Rodrigues & J. R. d. S. Passos (2003) Effects of some phenolic compounds on soybean seed germination and on seed-borne fungi. *Brazilian Archives of Biology and Technology*, 46, 155-161.
- Cragg, G. M. & D. J. Newman (2013) Natural products: a continuing source of novel drug leads. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1830, 3670-3695.
- Csizmadia, F. (2000) JChem: Java applets and modules supporting chemical database handling from web browsers. *Journal of Chemical Information and Computer Sciences*, 40, 323-324.
- Dardanelli, M. S., H. Manyani, S. González-Barroso, M. A. Rodríguez-Carvajal, A. M. Gil-Serrano, M. R. Espuny, F. J. López-Baena, R. A. Bellogín, M. Megías & F. J. Ollero (2010) Effect of the presence of the plant growth promoting rhizobacterium (PGPR) *Chryseobacterium balustinum* Aur9 and salt stress in the pattern of flavonoids exuded by soybean roots. *Plant and soil*, 328, 483-493.
- Dardouri, T., et al. (2019) Repellence of *Myzus persicae* (Sulzer): evidence of two modes of action of volatiles from selected living aromatic plants. *Pest Management Science*, 75, 1571-1584.
- Das, S. K. (2016) Screening of Bioactive Compounds for Development of New Pesticides: A Mini Review. *Universal Journal of Agricultural Research*, 4, 15-20.
- De Feo, V., F. De Simone & F. Senatore (2002) Potential allelochemicals from the essential oil of *Ruta graveolens*. *Phytochemistry*, 61, 573-578.
- De Lemos, M. L. (2001) Effects of soy phytoestrogens genistein and daidzein on breast cancer growth. *Annals of Pharmacotherapy*, 35, 1118-1121.
- De Melo Casal, C., J. R. Batalhão, V. de Cássia Domingues, O. C. Bueno, E. Rodrigues Filho, M. R. Forim, M. F. G. F. da Silva, P. C. Vieira & J. B. Fernandes (2009) High-speed counter-current chromatographic isolation of ricinine, an insecticide from *Ricinus communis*. *Journal of Chromatography A*, 1216, 4290-4294.
- De Sousa Falcao, H., J. A. Leite, J. M. Barbosa-Filho, P. F. de Athayde-Filho, M. C. de Oliveira Chaves, M. D. Moura, A. L. Ferreira, A. B. de Almeida, A. R. Souza-Brito, M. de Fatima Formiga Melo Diniz & L. M. Batista (2008) Gastric and duodenal antiulcer activity of alkaloids: a review. *Molecules*, 13, 3198-223.
- Di Toro, D. M., J. A. McGrath & W. A. Stubblefield (2007) Predicting the toxicity of neat and weathered crude oil: Toxic potential and the toxicity of saturated mixtures. *Environmental Toxicology and Chemistry: An International Journal*, 26, 24-36.
- DiMasi, J. A., H. G. Grabowski & R. W. Hansen (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. *Journal of Health Economics*, 47, 20-33.
- Dittfeld, A., A. Koszowska, A. P. Bronczyk, J. Nowak, K. Gwizdek & B. Zubelewicz-Szkodzinska (2015) Phytoestrogens--whether can they be an alternative to hormone replacement therapy for women during menopause period? *Wiad Lek*, 68, 163-7.

- Dos Santos, W. D., M. Ferrarese & O. Ferrarese-Filho (2008) Ferulic acid: an allelochemical troublemaker. *Functional Plant Science and Biotechnology*, 2, 47-55.
- Dreyer, D. L. (1968) Chemotaxonomy of the Rutaceae. III. Isolation of halfordinol derivatives from *Aeglopsis chevalieri*. *The Journal of Organic Chemistry*, 33, 3658-3660.
- Du, S. S., K. Yang, C. F. Wang, C. X. You, Z. F. Geng, S. S. Guo, Z. W. Deng & Z. L. Liu (2014) Chemical constituents and activities of the essential oil from *Myristica fragrans* against cigarette beetle *Lasioderma serricorne*. *Chemistry & Biodiversity*, 11, 1449-56.
- Duarte, M. C. & M. Rai. 2015. *Therapeutic Medicinal Plants: From Lab to the Market*. CRC Press.
- Einhellig, F. A. & P. C. Eckrich (1984) Interactions of temperature and ferulic acid stress on grain sorghum and soybeans. *Journal of Chemical Ecology*, 10, 161-170.
- Fakhri, S., S. Z. Moradi, M. H. Farzaei & A. Bishayee. 2020. Modulation of dysregulated cancer metabolism by plant secondary metabolites: A mechanistic review. In *Seminars in cancer biology*. Academic Press.
- Faugeras, G. & R. Paris (1968) Alkaloids and polyphenols of leguminous plants. XIV. On cinegalline,(+)(dihydroxy-3-5-methoxy-4-benzoyl)-13-oxy-lupanine, a new alkaloid of *Genista cinerea* DC. *Comptes rendus hebdomadaires des seances de l'Academie des sciences. Serie D: Sciences naturelles*, 267, 538-540.
- Fayaz, S. M., V. S. S. Kumar & K. G. Rajanikant (2014) Finding needles in a haystack: application of network analysis and target enrichment studies for the identification of potential anti-diabetic phytochemicals. *PLoS one*, 9, e112911.
- Felton, G., J. Workman & S. Duffey (1992) Avoidance of antinutritive plant defense: role of midgut pH in Colorado potato beetle. *Journal of Chemical Ecology*, 18, 571-583.
- Feuston, B. P., S. J. Chakravorty, J. F. Conway, J. C. Culberson, J. Forbes, B. Kraker, P. A. Lennon, C. Lindsley, G. B. McGaughey & R. Mosley (2005) Web enabling technology for the design, enumeration, optimization and tracking of compound libraries. *Current Topics in Medicinal Chemistry*, 5, 773-783.
- Fiehn, O. 2002. Metabolomics-the link between genotypes and phenotypes. In *Functional genomics*, 155-171. Springer.
- Freeman, B. C. & G. A. Beattie (2008) An overview of plant defenses against pathogens and herbivores. *The Plant Health Instructor*, 226.
- Fukuda, T., H. Ito, T. Mukainaka, H. Tokuda, H. Nishino & T. Yoshida (2003) Anti-tumor promoting effect of glycosides from *Prunus persica* seeds. *Biological and Pharmaceutical Bulletin*, 26, 271-273.
- Fuller, S. & J. M. Stephens (2015) Diosgenin, 4-hydroxyisoleucine, and fiber from fenugreek: mechanisms of actions and potential effects on metabolic syndrome. *Adv Nutr*, 6, 189-97.
- Gallo, M., A. Formato, D. Ianniello, A. Andolfi, E. Conte, M. Ciaravolo, V. Varchetta & D. Naviglio (2017) Supercritical fluid extraction of pyrethrins from pyrethrum flowers (*Chrysanthemum cinerariifolium*) compared to traditional maceration and cyclic pressurization extraction. *The Journal of Supercritical Fluids*, 119, 104-112.

- Garro Martinez, J. C., E. G. Vega-Hissi, M. F. Andrada & M. R. Estrada (2015) QSAR and 3D-QSAR studies applied to compounds with anticonvulsant activity. *Expert Opin Drug Discov*, 10, 37-51.
- Gatehouse, A. M. R., Y. Shi, K. S. Powell, C. Brough, V. A. Hilder, W. D. O. Hamilton, C. A. Newell, A. Merryweather, D. Boulter & J. A. Gatehouse (1993) Approaches to Insect Resistance Using Transgenic Plants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 342, 279-286.
- Gatouillat, G., A. A. Magid, E. Bertin, H. Morjani, C. Lavaud & C. Madoulet (2015) Medicarpin and millepurpan, two flavonoids isolated from *Medicago sativa*, induce apoptosis and overcome multidrug resistance in leukemia P388 cells. *Phytomedicine*, 22, 1186-1194.
- Genser, B., G. Silbernagel, G. De Backer, E. Bruckert, R. Carmena, M. J. Chapman, J. Deanfield, O. S. Descamps, E. R. Rietzschel, K. C. Dias & W. Marz (2012) Plant sterols and cardiovascular disease: a systematic review and meta-analysis. *Eur Heart J*, 33, 444-51.
- Ghelardini, C., L. Di Cesare Mannelli & E. Bianchi (2015) The pharmacological basis of opioids. *Clin Cases Miner Bone Metab*, 12, 219-21.
- Ghose, A. K., V. N. Viswanadhan & J. J. Wendoloski (1999) 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*, 1, 55-68.
- Giovannucci, E., E. B. Rimm, Y. Liu, M. J. Stampfer & W. C. Willett (2002) A prospective study of tomato products, lycopene, and prostate cancer risk. *Journal of the National Cancer Institute*, 94, 391-398.
- Glatstein, M., D. Danino, I. Wolyniez & D. Scolnik (2014) Seizures caused by ingestion of *Atropa belladonna* in a homeopathic medicine in a previously well infant: case report and review of the literature. *American Journal of Therapeutics*, 21, e196-e198.
- Glover, V., J. Liebowitz, I. Armando & M. Sandler (1982) β -Carbolines as selective monoamine oxidase inhibitors: in vivo implications. *Journal of neural transmission*, 54, 209-218.
- Gogte, V. M. 2000. *Ayurvedic pharmacology and therapeutic uses of medicinal plants (Dravyagunavigyan)*. Bharatiya Vidya Bhavan (SPARC).
- Gonzalez, M. L., D. M. A. Vera, J. Laiolo, M. B. Joray, M. Maccioni, S. M. Palacios, G. Molina, P. A. Lanza, S. Gancedo & V. Rumjanek (2017) Mechanism underlying the reversal of drug resistance in P-glycoprotein-expressing leukemia cells by pinorexinol and the study of a derivative. *Frontiers in pharmacology*, 8, 205.
- Gordon, E. M., R. W. Barrett, W. J. Dower, S. P. Fodor & M. A. Gallop (1994) Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. *Journal of Medicinal Chemistry*, 37, 1385-1401.
- Graf, E. & J. W. Eaton (1990) Antioxidant functions of phytic acid. *Free Radical Biology and Medicine*, 8, 61-69.
- Graf, E., K. L. Empson & J. W. Eaton (1987) Phytic acid. A natural antioxidant. *Journal of Biological Chemistry*, 262, 11647-11650.
- Grases, F. & A. Costa-Bauza (1999) Phytate (IP6) is a powerful agent for preventing calcifications in biological fluids: usefulness in renal lithiasis treatment. *Anticancer Res*, 19, 3717-22.

- Griffith, A. P. & M. W. Collison (2001) Improved methods for the extraction and analysis of isoflavones from soy-containing foods and nutritional supplements by reversed-phase high-performance liquid chromatography and liquid chromatography–mass spectrometry. *Journal of Chromatography A*, 913, 397-413.
- Gu, E.-J., D. W. Kim, G.-J. Jang, S. H. Song, J.-I. Lee, S. B. Lee, B.-M. Kim, Y. Cho, H.-J. Lee & H.-J. Kim (2017) Mass-based metabolomic analysis of soybean sprouts during germination. *Food chemistry*, 217, 311-319.
- Gupta, R. K., S. S. Gangoliya & N. K. Singh (2015) Reduction of phytic acid and enhancement of bioavailable micronutrients in food grains. *J Food Sci Technol*, 52, 676-84.
- Gylling, H. & P. Simonen (2015) Phytosterols, Phytostanols, and Lipoprotein Metabolism. *Nutrients*, 7, 7965-77.
- Han, Y. W. (1988) Removal of phytic acid from soybean and cottonseed meals. *Journal of Agricultural and Food Chemistry*, 36, 1181-1183.
- Hao, G., Q. Dong & G. Yang (2011) A comparative study on the constitutive properties of marketed pesticides. *Molecular informatics*, 30, 614-622.
- Harikarnpakdee, S. & C. Chuchote (2018) Oviposition Deterrent Efficacy and Characteristics of a Botanical Natural Product, *Ocimum gratissimum* (L.) Oil-Alginate Beads, against *Aedes aegypti* (L.). *The Scientific World Journal*, 2018, 9.
- Harland, B. F. & D. Oberleas (1987) Phytate in foods. *World Rev Nutr Diet*, 52, 235-59.
- Harvey, A. (2000) Strategies for discovering drugs from previously unexplored natural products. *Drug Discov Today*, 5, 294-300.
- Hemmati, S. & H. Seradj (2016) Justicidin B: A Promising Bioactive Lignan. *Molecules*, 21, 820.
- Heshmati, P., M. Nasehi & M. R. Zarrindast (2013) An overview of cognitive aspects of β -carbolines. *Archives of Advanced in Bioscience*, 5.
- Hoagland, R. E. (2009) Toxicity of tomatine and tomatidine on weeds, crops and phytopathogenic fungi. *Allelopathy J*, 23, 425-436.
- Holman, J. D., D. L. Tabb & P. Mallick (2014) Employing ProteoWizard to convert raw mass spectrometry data. *Current protocols in bioinformatics*, 13.24. 1-13.24. 9.
- Hou, S., Z. Ni, T. Ren, Z. Dong, M. Dong, Y. Gu, J. Yang & Q. Shi (2014) MALDI-MS of Pseudo-Alkaloid Taxanes from *Taxus canadensis*. *Chemistry of Natural Compounds*, 50, 1050-1055.
- Hussain, T. & B. Tan (2016) Oxidative Stress and Inflammation: What Polyphenols Can Do for Us? *Oxidative medicine and cellular longevity*, 2016, 9.
- Hussein, R. A. & A. A. El-Anssary (2018) Plants secondary metabolites: the key drivers of the pharmacological actions of medicinal plants. *Herbal Medicine*, 1, 13.
- Hutabarat, L., H. Greenfield & M. Mulholland (2000) Quantitative determination of isoflavones and coumestrol in soybean by column liquid chromatography. *Journal of chromatography A*, 886, 55-63.
- International Union for Conservation of Nature, N. R. & a. L. G. Ecosystems. 2006. *Conserving medicinal species: securing a healthy future*. IUCN.
- Ishimi, Y. (2015) Bone and Nutrition. Effect of isoflavones on bone health. *Clin Calcium*, 25, 999-1005.

- Iwashina, T. (2000) The structure and distribution of the flavonoids in plants. *Journal of Plant Research*, 113, 287.
- Jhanwar, B., V. Sharma, R. Singla & B. Shrivastava (2011) QSAR-Hansch analysis and related approaches in drug design. *Pharmacology online*, 1, 306-44.
- Jiang, H., G. Zheng, J. Lv, H. Chen, J. Lin, Y. Li, G. Fan & X. Ding (2016) Identification of *Centella asiatica*'s Effective Ingredients for Inducing the Neuronal Differentiation. *Evidence-Based Complementary and Alternative Medicine*, 2016, 9.
- Jorgensen, W. L. (2009) Efficient drug lead discovery and optimization. *Accounts of Chemical Research*, 42, 724-733.
- Joshi, T., M. R. Fitzpatrick, S. Chen, Y. Liu, H. Zhang, R. Z. Endacott, E. C. Gaudiello, G. Stacey, H. T. Nguyen & D. Xu (2014) Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding. *Nucleic Acids Research*, 42, D1245-D1252.
- Joshi, T., Q. Yao, D. F. Levi, L. Brechenmacher, B. Valliyodan, G. Stacey, H. Nguyen & D. Xu. 2010. SoyMetDB: the soybean metabolome database. In *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 203-208. IEEE.
- Julsing, M. K., A. Koulman, H. J. Woerdenbag, W. J. Quax & O. Kayser (2006) Combinatorial biosynthesis of medicinal plant secondary metabolites. *Biomolecular engineering*, 23, 265-279.
- Kadowaki, E., Y. Yoshida, T. Nitoda, N. Baba & S. Nakajima (2003) (-)-Olivil and (+)-1-Acetoxy-pinoreosinol from the Olive Tree (*Olea europaea* LINNE; Oleaceae) as Feeding Stimulants of the Olive Weevil (*Dyscerus perforatus*). *Bioscience, biotechnology, and biochemistry*, 67, 415-419.
- Kala, C. P., P. P. Dhyani & B. S. Sajwan (2006) Developing the medicinal plants sector in northern India: challenges and opportunities. *Journal of Ethnobiology and Ethnomedicine*, 2, 32.
- Kalaiselvan, V., M. Kalaivani, A. Vijayakumar, K. Sureshkumar & K. Venkateskumar (2010) Current knowledge and future direction of research on soy isoflavones as a therapeutic agents. *Pharmacogn Rev*, 4, 111-7.
- Kang, J., T. M. Badger, M. J. Ronis & X. Wu (2010) Non-isoflavone phytochemicals in soy and their health effects. *Journal of Agricultural and Food Chemistry*, 58, 8119-8133.
- Kapral, M., J. Wawrzczyk, M. Jurzak, A. Hollek & L. Weglarz (2012) The effect of inositol hexaphosphate on the expression of selected metalloproteinases and their tissue inhibitors in IL-1 β -stimulated colon cancer cells. *International journal of colorectal disease*, 27, 1419-1428.
- Karade, D., D. Vijayasarathi, N. Kadoo, R. Vyas, P. Ingle & M. Karthikeyan (2020) Design of Novel Drug-like Molecules using Informatics Rich Secondary Metabolites Analysis of Indian Medicinal and Aromatic Plants. *Combinatorial Chemistry & High Throughput Screening*, 23, 1113 - 1131.
- Karthikeyan, M., D. Nimje, R. Pahujani, K. Tyagi, S. Bapat, R. Vyas & K. Pillai Padmakumar (2015a) Chemoinformatics Approach for Building Molecular Networks from Marine Organisms. *Comb Chem High Throughput Screen*, 18, 673-684.
- Karthikeyan, M., D. Pandit & R. Vyas (2015b) ChemScreener: A distributed computing tool for scaffold based virtual screening. *Combinatorial chemistry & high throughput screening*, 18, 544-561.

- Karthikeyan, M. & R. Vyas. 2014. Chemoinformatics approach for the design and screening of focused virtual libraries. In *Practical Chemoinformatics*, 93-131. Springer.
- Karthikeyan, M. & R. Vyas (2015) Role of open source tools and resources in virtual screening for drug discovery. *Combinatorial chemistry & high throughput screening*, 18, 528-543.
- Katajamaa, M. & M. Oresic (2007) Data processing for mass spectrometry-based metabolomics. *Journal of chromatography A*, 1158, 318-328.
- Katayama, T. (1999) Hypolipidemic action of phytic acid (IP6): prevention of fatty liver. *Anticancer Res*, 19, 3695-8.
- Kaur, S., I. Grover & S. Kumar (1997) Antimutagenic potential of ellagic acid isolated from Terminalia arjuna. *Indian journal of experimental biology*, 35, 478-482.
- Khan, A. V., Q. U. Ahmed, M. W. Khan & A. A. Khan (2014) Herbal cure for poisons and poisonous bites from Western Uttar Pradesh, India. *Asian Pacific Journal of Tropical Disease*, 4, S116-S120.
- Khan, M. S. & I. Ahmad (2011) In vitro antifungal, anti-elastase and anti-keratinase activity of essential oils of Cinnamomum-, Syzygium- and Cymbopogon-species against Aspergillus fumigatus and Trichophyton rubrum. *Phytomedicine*, 19, 48-55.
- Kim, S. M., C. Kim, H. Bae, J. H. Lee, S. H. Baek, D. Nam, W. S. Chung, B. S. Shim, S. G. Lee & S. H. Kim (2015) 6-Shogaol exerts anti-proliferative and pro-apoptotic effects through the modulation of STAT3 and MAPKs signaling pathways. *Molecular carcinogenesis*, 54, 1132-1146.
- Klingberg, S., H. Andersson, A. Mulligan, A. Bhaniani, A. Welch, S. Bingham, K. Khaw, S. Andersson & L. Ellegård (2008) Food sources of plant sterols in the EPIC Norfolk population. *European journal of clinical nutrition*, 62, 695-703.
- Kochev, N. A., S.; Jeliaskova, N. (2017) Combinatorial generation of molecules by virtual software reactor. *Sci Work Union Sci Bulg Plovdiv*, 11, 214-219.
- Kostoff, R. N. 2005. Method for data and text mining and literature-based discovery. Google Patents.
- Koul, O. & S. Walia (2009) Comparing impacts of plant extracts and pure allelochemicals and implications for pest control. *CAB Reviews: Perspectives in agriculture, veterinary science, nutrition and natural resources*, 4, 1-30.
- Kris-Etherton, P. M., K. D. Hecker, A. Bonanome, S. M. Coval, A. E. Binkoski, K. F. Hilpert, A. E. Griel & T. D. Etherton (2002) Bioactive compounds in foods: their role in the prevention of cardiovascular disease and cancer. *The American journal of medicine*, 113, 71-88.
- Kubinyi, H. 1993. *3D QSAR in drug design: volume 1: theory methods and applications*. Springer Science & Business Media.
- Kugler, F., F. C. Stintzing & R. Carle (2004) Identification of betalains from petioles of differently colored Swiss chard (*Beta vulgaris* L. ssp. *cicla* [L.] Alef. cv. Bright Lights) by high-performance liquid chromatography– Electro spray ionization mass spectrometry. *Journal of Agricultural and Food Chemistry*, 52, 2975-2981.
- Kumar, D. (2015) Nuclear magnetic resonance (NMR) spectroscopy: Metabolic profiling of medicinal plants and their products. *Crit Rev Anal Chem*, 46, 400-412.
- Kumar, H., D. Saini, S. Jain & N. Jain (2013) Pyrazole scaffold: a remarkable tool in the development of anticancer agents. *Eur J Med Chem*, 70, 248-58.

- Lagunin, A. A., R. K. Goel, D. Y. Gawande, P. Pahwa, T. A. Glorizova, A. V. Dmitriev, S. M. Ivanov, A. V. Rudik, V. I. Konova & P. V. Pogodin (2014) Chemo-and bioinformatics resources for in silico drug discovery from medicinal plants beyond their traditional use: a critical review. *Natural product reports*, 31, 1585-1611.
- Langenbach, C., H. Schultheiss, M. Rosendahl, N. Tresch, U. Conrath & K. Goellner (2016) Interspecies gene transfer provides soybean resistance to a fungal pathogen. *Plant Biotechnol J*, 14, 699-708.
- Latif, S., G. Chiapusio & L. Weston (2017) Allelopathy and the role of allelochemicals in plant defence. *Advances in botanical research*, 82, 19-54.
- Lavecchia, A. & C. Di Giovanni (2013) Virtual screening strategies in drug discovery: a critical review. *Current medicinal chemistry*, 20, 2839-2860.
- Law, V., C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson & V. Neveu (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, 42, D1091-D1097.
- Lawless, M. S., M. Waldman, R. Fraczkiwicz & R. D. Clark (2016) Using Cheminformatics in Drug Discovery. *Handb Exp Pharmacol*, 232, 139-68.
- Leach, A. R., J. Bradshaw, D. V. Green, M. M. Hann & J. J. Delany (1999) Implementation of a system for reagent selection and library enumeration, profiling, and design. *Journal of Chemical Information and Computer Sciences*, 39, 1161-1172.
- Lee, G. A., K. A. Hwang & K. C. Choi (2016a) Roles of Dietary Phytoestrogens on the Regulation of Epithelial-Mesenchymal Transition in Diverse Cancer Metastasis. *Toxins (Basel)*, 8, 162.
- Lee, M.-L. & G. Schneider (2001a) Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries. *Journal of Combinatorial Chemistry*, 3, 284-289.
- Lee, M. L. & G. Schneider (2001b) Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries. *Journal of combinatorial chemistry*, 3, 284-289.
- Lee, S. K., D.-H. Kim & H. H. Yoo (2011) Comparative metabolism of sildenafil in liver microsomes of different species by using LC/MS-based multivariate analysis. *Journal of Chromatography B*, 879, 3005-3011.
- Lee, Y.-M., Y. Yoon, H. Yoon, H.-M. Park, S. Song & K.-J. Yeum (2017) Dietary anthocyanins against obesity and inflammation. *Nutrients*, 9, 1089.
- Lee, Y. H., K. J. Lee, Y. H. Min, H. C. Ahn, Y. D. Sohn, W. W. Lee, Y. T. Oh, G. C. Cho, J. Y. Seo & D. H. Shin (2016b) Refractory ventricular fibrillation treated with esmolol. *Resuscitation*, 107, 150-155.
- Leicach, S. R. & H. D. Chludil. 2014. Plant secondary metabolites: Structure–activity relationships in human health prevention and treatment of common diseases. In *Studies in natural products chemistry*, 267-304. Elsevier.
- Letcher, R., D. Widdowson, B. Deverall & J. Mansfield (1970) Identification and activity of wyerone acid as a phytoalexin in broad bean (*Vicia faba*) after infection by *Botrytis*. *Phytochemistry*, 9, 249-252.
- Lewis, K., J. Tzilivakis, A. Green & D. Warner (2006) Pesticide Properties DataBase (PPDB).
- Li, H.-H., M. Inoue, H. Nishimura, J. Mizutani & E. Tsuzuki (1993) Interactions of trans-cinnamic acid, its related phenolic allelochemicals, and abscisic acid in

- seedling growth and seed germination of lettuce. *Journal of Chemical Ecology*, 19, 1775-1787.
- Liao, C., B. Liu, L. Shi, J. Zhou & X.-P. Lu (2005) Construction of a virtual combinatorial library using SMILES strings to discover potential structure-diverse PPAR modulators. *European Journal of Medicinal Chemistry*, 40, 632-640.
- Lin, L.-Z., J. M. Harnly, M. S. Pastor-Corrales & D. L. Luthria (2008) The polyphenolic profiles of common bean (*Phaseolus vulgaris* L.). *Food chemistry*, 107, 399-410.
- Lionta, E., G. Spyrou, D. K. Vassilatis & Z. Cournia (2014) Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr Top Med Chem*, 14, 1923-38.
- Lipinski, C. A., F. Lombardo, B. W. Dominy & P. J. Feeney (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23, 3-25.
- Luo, S.-Q., H.-F. Jin, H. Kawai, H. Sino & N. Ōtake (1987) Isolation of new saponins from the aerial part of *Bupleurum kunningense* Y. Li et SL Pan. *Agricultural and biological chemistry*, 51, 1515-1519.
- Lusher, S. J., R. McGuire, R. Azevedo, J. W. Boiten, R. C. van Schaik & J. de Vlieg (2011) A molecular informatics view on best practice in multi-parameter compound optimization. *Drug Discov Today*, 16, 555-68.
- Mahendran, G., G. Thamocharan, S. Sengottuvelu & V. N. Bai (2014) Anti-diabetic activity of *Swertia corymbosa* (Griseb.) Wight ex CB Clarke aerial parts extract in streptozotocin induced diabetic rats. *J Ethnopharmacol*, 151, 1175-1183.
- Mahmoud, A. M., W. Yang & M. C. Bosland (2014) Soy isoflavones and prostate cancer: a review of molecular mechanisms. *J Steroid Biochem Mol Biol*, 140, 116-32.
- Maia, M. F. & S. J. Moore (2011) Plant-based insect repellents: a review of their efficacy, development and testing. *Malaria Journal*, 10, 1.
- Makarevich, I. F., Y. I. Khadzhai, V. V. Pavlova & A. V. Nikolaeva. 1979. Cardenolide and bufadienolide derivatives of ajmaline and process for producing same. Google Patents.
- Makishima, M., N. Takahashi & T. Kawada (2010) Diosgenin, the Main Aglycon of Fenugreek, Inhibits LXRA Activity in HepG2 Cells and Decreases Plasma and Hepatic Triglycerides in Obese Diabetic Mice. *The Journal of nutrition*, 141, 17-23.
- Manayi, A., M. Vazirian & S. Saeidnia (2015) *Echinacea purpurea*: Pharmacology, phytochemistry and analysis methods. *Pharmacogn Rev*, 9, 63-72.
- Mannina, L., A. P. Sobolev & D. Capitani (2012) Applications of NMR metabolomics to the study of foodstuffs: truffle, kiwifruit, lettuce, and sea bass. *Electrophoresis*, 33, 2290-313.
- Märki, H., A. Binggeli, B. Bittner, V. Bohner-Lang, V. Brey, D. Bur, P. Coassolo, J. Clozel, A. D'Arcy & H. Doebeli (2001) Piperidine renin inhibitors: from leads to drug candidates. *Il Farmaco*, 56, 21-27.
- Marney, L. C., J. C. Hoggard, K. J. Skogerboe & R. E. Synovec (2014) Methods of discovery-based and targeted metabolite analysis by comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry detection. *Methods Mol Biol*, 1198, 83-97.

- Martin, Y. C. & S. Muchmore (2009) Beyond QSAR: lead hopping to different structures. *QSAR & Combinatorial Science*, 28, 797-801.
- Martinez-Mayorga, K., A. Madariaga-Mazon, J. L. Medina-Franco & G. Maggiora (2020) The impact of chemoinformatics on drug discovery in the pharmaceutical industry. *Expert Opinion on Drug Discovery*, 15, 293-306.
- Martinez Dominguez, B., M. V. Ibanez Gomez & F. Rincon Leon (2002) Phytic acid: nutritional aspects and analytical implications. *Arch Latinoam Nutr*, 52, 219-31.
- Martins, N., S. Petropoulos & I. C. Ferreira (2016) Chemical composition and bioactive compounds of garlic (*Allium sativum* L.) as affected by pre-and post-harvest conditions: A review. *Food Chemistry*, 211, 41-50.
- Matano, Y., T. Okuyama, S. Shibata, M. Hoson, T. Kawada, H. Osada & T. Noguchi (1986) Studies on coumarins of a Chinese Drug "Qian-Hu"; VIII. Structures of new coumarin-glycosides of Zi-Huan Quian-Hu and effect of coumarin-glycosides on human platelet aggregation. *Planta medica*, 52, 135-138.
- Matsumura, A., A. Ghosh, G. Pope & P. Darbre (2005) Comparative study of oestrogenic properties of eight phytoestrogens in MCF7 human breast cancer cells. *The Journal of steroid biochemistry and molecular biology*, 94, 431-443.
- Mayo-Bean, K., K. Moran-Bruce, W. Meylan, P. Ranslow, M. Lock, V. Nabholz, J. Von Runnen, L. Cassidy & J. Tunkel (2017) Methodology document for the ecological structure–activity relationship model (ECOSAR) class program, Ver 2. *Office of Pollution Prevention and Toxics, US Environmental Protection Agency, Washington DC*.
- McLafferty, F. W. (1981) Tandem mass spectrometry. *Science*, 214, 280-287.
- Meetei, P. A., R. S. Rathore, N. P. Prabhu & V. Vindal (2016) Modeling of babesipain-1 and identification of natural and synthetic leads for bovine babesiosis drug development. *J Mol Model*, 22, 71.
- Menendez, J. A., A. Vazquez-Martin, C. Oliveras-Ferraros, R. Garcia-Villalba, A. Carrasco-Pancorbo, A. Fernandez-Gutierrez & A. Segura-Carretero (2008) Analyzing effects of extra-virgin olive oil polyphenols on breast cancer-associated fatty acid synthase protein expression using reverse-phase protein microarrays. *International journal of molecular medicine*, 22, 433-439.
- Mishra, B. B. & V. K. Tiwari (2011) Natural products: an evolving role in future drug discovery. *Eur J Med Chem*, 46, 4769-4807.
- Mishra, R. K., A. Kumar & A. Kumar (2012) Pharmacological activity of Zingiber officinale. *International Journal of Pharmaceutical and Chemical Sciences*, 1, 1073-1078.
- Moga, M. A., O. G. Dimienescu, C. A. Arvatescu, A. Mironescu, L. Dracea & L. Ples (2016) The Role of Natural Polyphenols in the Prevention and Treatment of Cervical Cancer-An Overview. *Molecules*, 21, 1055.
- Mohanraj, K., B. S. Karthikeyan, R. Vivek-Ananth, R. B. Chand, S. Aparna, P. Mangalapandi & A. Samal (2018) IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry and Therapeutics. *Scientific reports*, 8, 4329.
- Monge, P., R. Scheline & E. Solheim (1976) The metabolism of zingerone, a pungent principle of ginger. *Xenobiotica*, 6, 411-423.
- Moreno, M. & V. Rodriguez (1981) Yiamoloside B, a fungistatic saponin of *Phytolacca octandra*. *Phytochemistry*, 20.
- Morimoto, S., K. Suemori, J. Moriwaki, F. Taura, H. Tanaka, M. Aso, M. Tanaka, H. Suemune, Y. Shimohigashi & Y. Shoyama (2001) Morphine metabolism in

- the opium poppy and its possible physiological function biochemical characterization of the morphine metabolite, bismorphine. *Journal of Biological Chemistry*, 276, 38179-38184.
- Munro, I. C., M. Harwood, J. J. Hlywka, A. M. Stephen, J. Doull, W. G. Flamm & H. Adlercreutz (2003) Soy isoflavones: a safety review. *Nutrition Reviews*, 61, 1-33.
- Murkies, A. L., G. Wilcox & S. R. Davis (1998) Phytoestrogens 1. *The Journal of Clinical Endocrinology & Metabolism*, 83, 297-303.
- Mutha, R. E., A. U. Tatiya & S. J. Surana (2021) Flavonoids as natural phenolic compounds and their role in therapeutics: an overview. *Future Journal of Pharmaceutical Sciences*, 7, 1-13.
- Nadeem, M. & A. Zeb (2018) Impact of maturity on phenolic composition and antioxidant activity of medicinally important leaves of *Ficus carica* L. *Physiology and molecular biology of plants*, 24, 881-887.
- Najar, Z. A. & S. Agnihotri (2012) Need and importance of conservation of endangered tree *Oroxylum indicum* (Linn.) Vent. *Asian Journal of Plant Science and Research*, 2, 220-223.
- Nawrocka-Musial, D. & M. Latocha (2012) Phytic acid--anticancer nutraceutical. *Pol Merkur Lekarski*, 33, 43-7.
- Neves, B. J., R. C. Braga, C. C. Melo-Filho, J. T. Moreira-Filho, E. N. Muratov & C. H. Andrade (2018) QSAR-based virtual screening: advances and applications in drug discovery. *Frontiers in pharmacology*, 9, 1275.
- Niedzwiecki, A., M. W. Roomi, T. Kalinovsky & M. Rath (2016) Anticancer Efficacy of Polyphenols and Their Combinations. *Nutrients*, 8, 552.
- Niedźwiedz-Siegień, I. (1998) Cyanogenic glucosides in *Linum usitatissimum*. *Phytochemistry*, 49, 59-63.
- Niemeyer, H. M. (1988) Hydroxamic acids (4-hydroxy-1, 4-benzoxazin-3-ones), defence chemicals in the Gramineae. *Phytochemistry*, 27, 3349-3358.
- Nisbet, A. J. (2000) Azadirachtin from the neem tree *Azadirachta indica*: its action against insects. *Anais da Sociedade Entomológica do Brasil*, 29, 615-632.
- Nuraini, F. D., S. Rahayu & M. Rifai (2019) Anti-inflammatory activity of elicited soybean (*Glycine max*) extract on Balb/C mice (*Mus musculus*) with high-fat and -fructose diet. *Cent Eur J Immunol*, 44, 7-14.
- Oak, M. H., J. Bedoui, S. F. Madeira, K. Chalupsky & V. Schini-Kerth (2006) Delphinidin and cyanidin inhibit PDGFAB-induced VEGF release in vascular smooth muscle cells by preventing activation of p38 MAPK and JNK. *British journal of pharmacology*, 149, 283-290.
- Ohnishi, M. & Y. Fujino (1982) Sphingolipids in immature and mature soybeans. *Lipids*, 17, 803-810.
- Okada, K., H. Kawaide, K. Miyamoto, S. Miyazaki, R. Kainuma, H. Kimura, K. Fujiwara, M. Natsume, H. Nojiri, M. Nakajima, H. Yamane, Y. Hatano, H. Nozaki & K. Hayashi (2016) HpDTC1, a Stress-Inducible Bifunctional Diterpene Cyclase Involved in Momilactone Biosynthesis, Functions in Chemical Defence in the Moss *Hypnum plumaeforme*. *Sci Rep*, 6, 25316.
- Oliveros, J. (2015) Venny 2.1. 0. An interactive tool for comparing lists with Venn's diagrams. *BioinfoGP of CNB-CSIC*.
- Oomah, B. D., G. Mazza & E. O. Kenaschuk (1992) Cyanogenic compounds in flaxseed. *Journal of Agricultural and Food chemistry*, 40, 1346-1348.
- Opletal, L., M. Locarek, A. Frankova, J. Chlebek, J. Smid, A. Host'alkova, M. Safratova, D. Hulcova, P. Kloucek, M. Rozkot & L. Cahlikova (2014)

- Antimicrobial activity of extracts and isoquinoline alkaloids of selected papaveraceae plants. *Nat Prod Commun*, 9, 1709-12.
- Oprea, T. I. (2000) Property distribution of drug-related chemical databases. *Journal of computer-aided molecular design*, 14, 251-264.
- Osbourn, A. E. (1996) Preformed antimicrobial compounds and plant defense against fungal attack. *The plant cell*, 8, 1821.
- Osman, S. & R. Moreau (1985) Potato phytoalexin elicitors in *Phytophthora infestans* spore germination fluids. *Plant science*, 41, 205-209.
- Ostlund Jr, R. E. (2004) Phytosterols and cholesterol metabolism. *Current Opinion in Lipidology*, 15, 37-41.
- Pacirc, M., A. E. Pacirc, O. Rosca-Casian, L. Vlase & G. Groza (2010) Antifungal activity of *Allium obliquum*. *Journal of Medicinal Plants Research*, 4, 138-141.
- Padalia, R. C., R. S. Verma, A. Chauhan, P. Goswami, C. S. Chanotiya, A. Saroj, A. Samad & A. Khaliq (2014) Compositional variability and antifungal potentials of *ocimum basilicum*, *O. tenuiflorum*, *O. gratissimum* and *O. kilimandscharicum* essential oils against *Rhizoctonia solani* and *Choanephora cucurbitarum*. *Natural Product Communications*, 9, 1507-10.
- Panda, A. K. & S. K. Debnath (2010) Overdose effect of aconite containing Ayurvedic medicine ('Mahashankha Vati'). *International Journal of Ayurveda Research*, 1, 183.
- Parasuraman, S., G. S. Thing & S. A. Dhanaraj (2014) Polyherbal formulation: Concept of ayurveda. *Pharmacognosy Reviews*, 8, 73.
- Paré, P. W. & J. H. Tumlinson (1999) Plant volatiles as a defense against insect herbivores. *Plant physiology*, 121, 325-332.
- Patel, K., V. Kumar, M. Rahman, A. Verma & D. K. Patel (2018) New insights into the medicinal importance, physiological functions and bioanalytical aspects of an important bioactive compound of foods 'Hyperin': Health benefits of the past, the present, the future. *Beni-Suef University Journal of Basic and Applied Sciences*, 7, 31-42.
- Pathania, S., S. M. Ramakrishnan & G. Bagler (2015) Phytochemica: a platform to explore phytochemicals of medicinal plants. *Database*, 2015.
- Patton, C. A., T. G. Ranney, J. D. Burton & J. F. Walgenbach (1997) Natural pest resistance of *Prunus* taxa to feeding by adult Japanese beetles: role of endogenous allelochemicals in host plant resistance. *Journal of the American Society for Horticultural Science*, 122, 668-672.
- Patwardhan, B., A. D. Vaidya & M. Chorghade (2004) Ayurveda and natural products drug discovery. *CURRENT SCIENCE-BANGALORE*, 86, 789-799.
- Pigatto, A. G., C. C. Blanco, L. A. Mentz & G. L. Soares (2015) Tropane alkaloids and calystegines as chemotaxonomic markers in the Solanaceae. *An Acad Bras Cienc*, 87, 2139-49.
- Piironen, V. & A.-M. Lampi. 2004. *Occurrence and levels of phytosterols in foods*. Marcel Dekker: New York, NY, USA.
- Polur, H., T. Joshi, C. T. Workman, G. Lavekar & I. Kouskoumvekaki (2011) Back to the roots: prediction of biologically active natural products from ayurveda traditional medicine. *Molecular informatics*, 30, 181-187.
- Potterat, O. & M. Hamburger. 2008. Drug discovery and development with plant-derived compounds. In *Natural Compounds as Drugs Volume I*, 45-118. Springer.

- Putnam, A. R. (1988) Allelochemicals from plants as herbicides. *Weed technology*, 510-518.
- Qiang, Z., S. O. Lee, Z. Ye, X. Wu & S. Hendrich (2012) Artichoke extract lowered plasma cholesterol and increased fecal bile acids in Golden Syrian hamsters. *Phytotherapy Research*, 26, 1048-1052.
- Racette, S. B., X. Lin, L. Ma & R. E. Ostlund, Jr. (2015) Natural Dietary Phytosterols. *JAOAC Int*, 98, 679-84.
- Raffa, K. F. (1987) Influence of host plant on deterrence by azadirachtin of feeding by fall armyworm larvae (Lepidoptera: Noctuidae). *Journal of economic entomology*, 80, 384-387.
- Rajemiarimiraho, M., J.-T. Banzouzi, M.-L. Nicolau-Travers, S. Ramos, Z. Cheikh-Ali, C. Bories, O. L. Rakotonandrasana, S. Rakotonandrasana, P. A. Andrianary & F. Benoit-Vical (2014) Antiprotozoal activities of *Millettia richardiana* (Fabaceae) from Madagascar. *Molecules*, 19, 4200-4211.
- Rajemiarimiraho, M., J. T. Banzouzi, S. R. Rakotonandrasana, P. Chalard, F. Benoit-Vical, L. H. Rasoanaivo, A. Raharisololalao & R. Randrianja (2013) Pyranocoumarin and triterpene from *Millettia richardiana*. *Natural product communications*, 8, 1934578X1300800817.
- Ramos-López, M., S. Pérez, G. Rodríguez-Hernández, P. Guevara-Fefer & M. A. Zavala-Sanchez (2010) Activity of *Ricinus communis* (Euphorbiaceae) against *Spodoptera frugiperda* (Lepidoptera: Noctuidae). *African Journal of Biotechnology*, 9.
- Ramprasath, V. R. & A. B. Awad (2015) Role of Phytosterols in Cancer Prevention and Treatment. *JAOAC Int*, 98, 735-738.
- Ramteke, V., V. Kurrey & S. Kar (2015) Jamun: A Traditional Fruit and Medicine. *Popular Kheti*, 3, 188-190.
- Rao, J. & J. Cooper (1995) Soybean nodulating rhizobia modify nod gene inducers daidzein and genistein to yield aromatic products that can influence gene-inducing activity. *Molecular plant-microbe interactions: MPMI (USA)*, 8, 855-862.
- Raut, J. S. & S. M. Karuppayil (2014) A status review on the medicinal properties of essential oils. *Industrial Crops and Products*, 62, 250-264.
- Rawat, G. (2008) Special habitats and threatened plants of India. *ENVIS Bulletin: Wildlife and Protected Areas*, 11, 239.
- Reddy, A. S., S. P. Pati, P. P. Kumar, H. Pradeep & G. N. Sastry (2007) Virtual screening in drug discovery-a computational perspective. *Current Protein and Peptide Science*, 8, 329-351.
- Reddy, N. R., S. K. Sathe & D. K. Salunkhe (1982) Phytates in legumes and cereals. *Adv Food Res*, 28, 1-92.
- Ren, J.-L., A.-H. Zhang, L. Kong & X.-J. Wang (2018) Advances in mass spectrometry-based metabolomics for investigation of metabolites. *RSC advances*, 8, 22335-22350.
- Rietjens, I. M., J. Lousse & K. Beekmann (2016) The potential health effects of dietary phytoestrogens. *Br J Pharmacol*, 174, 1263-1280.
- Rishi, R. (2002) Phytoestrogens in health and illness. *Indian journal of pharmacology*, 34, 311-320.
- Robert, A. & P. Wither (2007) StatistiXL, version 1.8, a powerful statistics and statistical analysis add-in for Microsoft Excel. *Microsoft, Washington DC*.

- Rollinger, J. M., H. Stuppner & T. Langer. 2008. Virtual screening for the discovery of bioactive natural products. In *Natural compounds as drugs Volume I*, 211-249. Springer.
- Ross, K., and Giuseppe Mazza (2010) Characteristics of lignin from flax shives as affected by extraction conditions. *International Journal of Molecular Sciences*, 11, 4035-4050.
- Rossi, D. T. & M. Sinz. 2001. *Mass spectrometry in drug discovery*. CRC Press.
- Roy, K. & R. N. Das (2014) A review on principles, theory and practices of 2D-QSAR. *Curr Drug Metab*, 15, 346-79.
- Roy, K. & I. Mitra (2011) On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb Chem High Throughput Screen*, 14, 450-74.
- Rozanski, A. (1966) Gas Chromatographic Determination of Campesterol, β -Sitosterol, and Stigmasterol. *Analytical chemistry*, 38, 36-40.
- Saade, D. & C. Joshi (2015) Pure cannabidiol in the treatment of malignant migrating partial seizures in infancy: a case report. *Pediatric neurology*, 52, 544-547.
- Sahoo, M. R., S. Dhanabal, A. N. Jadhav, V. Reddy, G. Muguli, U. Babu & P. Rangesh (2014) Hydnocarpus: An ethnopharmacological, phytochemical and pharmacological review. *J Ethnopharmacol*, 154, 17-25.
- Sakai, T. & M. Kogiso (2008) Soy isoflavones and immunity. *The Journal of Medical Investigation*, 55, 167-173.
- Salim, A. A., Y.-W. Chin & A. D. Kinghorn. 2008. Drug discovery from plants. In *Bioactive molecules and medicinal plants*, 1-24. Springer.
- Samal, J. (2015) Role of AYUSH workforce, therapeutics, and principles in health care delivery with special reference to National Rural Health Mission. *Ayu*, 36, 5-8.
- Sánchez-Campillo, M., J. Gabaldon, J. Castillo, O. Benavente-García, M. Del Baño, M. Alcaraz, V. Vicente, N. Alvarez & J. Lozano (2009) Rosmarinic acid, a photo-protective agent against UV and other ionizing radiations. *Food and chemical toxicology*, 47, 386-392.
- Saric, S. & R. K. Sivamani (2016) Polyphenols and Sunburn. *Int J Mol Sci*, 17, 1521.
- Scarpa, A. & A. Guerci (1982) Various uses of the castor oil plant (*Ricinus communis* L.) a review. *Journal of Ethnopharmacol*, 5, 117-137.
- Schauer, N. & A. R. Fernie (2006) Plant metabolomics: towards biological function and mechanism. *Trends in plant science*, 11, 508-516.
- Schuller, A., V. Hahnke & G. Schneider (2007) SmiLib v2. 0: A Java-Based Tool for Rapid Combinatorial Library Enumeration. *QSAR & Combinatorial Science*, 26, 407-410.
- Sengottayan, S.-N. (2013) Physiological and biochemical effect of neem and other Meliaceae plants secondary metabolites against Lepidopteran insects. *Frontiers in physiology*, 4, 359.
- Senthil-Nathan, S. (2013) Physiological and biochemical effect of neem and other Meliaceae plants secondary metabolites against Lepidopteran insects. *Frontiers in physiology*, 4.
- Shamsuddin, A. M. & I. Vucenik (1999) Mammary tumor inhibition by IP6: a review. *Anticancer Res*, 19, 3671-4.
- Shamsuddin, A. M., I. Vucenik & K. E. Cole (1997) IP6: a novel anti-cancer agent. *Life Sci*, 61, 343-54.
- Shang, S. Z., H. Chen, C. Q. Liang, Z. H. Gao, X. Du, R. R. Wang, Y. M. Shi, Y. T. Zheng, W. L. Xiao & H. D. Sun (2013) Phenolic constituents from *Parakmeria*

- yunnanensis and their anti-HIV-1 activity. *Archives of pharmacal research*, 36, 1223-1230.
- Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski & T. Ideker (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13, 2498-2504.
- Sheppard, D. W. & J. A. MacRitchie (2013) Building in molecular diversity for targeted libraries. *Drug Discov Today Technol*, 10, e461-6.
- Shin, S.-A., S. Y. Moon, W.-Y. Kim, S.-M. Paek, H. H. Park & C. S. Lee (2018) Structure-based classification and anti-cancer effects of plant metabolites. *International journal of molecular sciences*, 19, 2651.
- Shiva, M. 1998. *Inventory of Forest Resources for Sustainable Management & Biodiversity Conservation with Lists of Multipurpose Tree Species Yielding Both Timber & Non-timber Forest Products (NTFPs), and Shrub & Herb Species of NTFP Importance*. Indus Publishing.
- Shlyankevich, M. 1995. Pharmaceutical compositions and dietary soybean food products for the prevention of osteoporosis. Google Patents.
- Silva, E., J. P. da Graça, C. Porto, R. M. do Prado, C. B. Hoffmann-Campo, M. C. Meyer, E. de Oliveira Nunes & E. J. Pilau (2020) Unraveling Asian Soybean Rust metabolomics using mass spectrometry and Molecular Networking approach. *Scientific reports*, 10, 1-11.
- Silva, R. R., F. Jourdan, D. M. Salvanha, F. Letisse, E. L. Jamin, S. Guidetti-Gonzalez, C. A. Labate & R. Z. Vencio (2014) ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics*, 30, 1336-1337.
- Singh, B. & R. Rastogi (1970) Cardenolides—glycosides and genins. *Phytochemistry*, 9, 315-331.
- Singh, N., R. Guha, M. A. Giulianotti, C. Pinilla, R. A. Houghten & J. L. Medina-Franco (2009) Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *Journal of Chemical Information and Modeling*, 49, 1010-1024.
- Sledz, W., E. Los, A. Paczek, J. Rischka, A. Motyka, S. Zoledowska, J. Piosik & E. Lojkowska (2015) Antibacterial activity of caffeine against plant pathogenic bacteria. *Acta Biochimica Polonica*, 62, 605-612.
- Sobenin, I. A., V. A. Myasoedova & A. N. Orekhov (2016) Phytoestrogen-Rich Dietary Supplements in Anti-Atherosclerotic Therapy in Postmenopausal Women. *Curr Pharm Des*, 22, 152-63.
- Sood, A. & A. Ghosh (2006) Literature search using PubMed: an essential tool for practicing evidence-based medicine. *Journal Association of Physicians of India*, 54, 303.
- Sparvoli, F. & E. Cominelli (2015) Seed Biofortification and Phytic Acid Reduction: A Conflict of Interest for the Plant? *Plants (Basel)*, 4, 728-55.
- Speijers, G., B. Bottex, B. Dusemund, A. Lugasi, J. Toth, J. Amberg-Müller, C. Galli, V. Silano & I. M. Rietjens. 2010. Risk Assessment of Phytochemicals in Food: Novel Approaches. In *Wiley-Blackwell*.
- Sravanthi, T. V. & S. L. Manju (2016) Indoles - A promising scaffold for drug development. *Eur J Pharm Sci*, 91, 1-10.
- Stafford, H. A. (1997) Roles of flavonoids in symbiotic and defense functions in legume roots. *The Botanical Review*, 63, 27-39.

- Stegelmeier, B. L. (2011) Pyrrolizidine alkaloid-containing toxic plants (Senecio, Crotalaria, Cynoglossum, Amsinckia, Heliotropium, and Echium spp.). *Vet Clin North Am Food Anim Pract*, 27, 419-28, ix.
- Stevenson, J. M. & P. D. Mulready. 2003. Pipeline Pilot 2.1 By Scitegic, 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123-1365. www.scitegic.com. See Web Site for Pricing Information. 1437-1438.
- Stijve, T. & A. De Meijer (1999) Hydrocyanic acid in mushrooms, with special reference to wild-growing and cultivated edible species. *Deutsche Lebensmittel-Rundschau (Germany)*, 95, 366-373.
- Su, G., J. H. Morris, B. Demchak & G. D. Bader (2014) Biological network exploration with Cytoscape 3. *Current protocols in bioinformatics*, 47, 8.13. 1-8.13. 24.
- Sud, M., E. Fahy & S. Subramaniam (2012) Template-based combinatorial enumeration of virtual compound libraries for lipids. *Journal of Cheminformatics*, 4, 23.
- Sugimoto, N., S. Miwa, Y. Hitomi, H. Nakamura, H. Tsuchiya & A. Yachie (2014) Theobromine, the primary methylxanthine found in Theobroma cacao, prevents malignant glioblastoma proliferation by negatively regulating phosphodiesterase-4, extracellular signal-regulated kinase, Akt/mammalian target of rapamycin kinase, and nuclear factor-kappa B. *Nutr Cancer*, 66, 419-23.
- Sun, H., G. Tawa & A. Wallqvist (2012) Classification of scaffold-hopping approaches. *Drug Discovery Today*, 17, 310-324.
- Swamy, M. K. & U. R. Sinniah (2015) A Comprehensive Review on the Phytochemical Constituents and Pharmacological Activities of Pogostemon cablin Benth.: An Aromatic Medicinal Plant of Industrial Importance. *Molecules*, 20, 8521-47.
- Tao, W., X. Xu, X. Wang, B. Li, Y. Wang, Y. Li & L. Yang (2013) Network pharmacology-based prediction of the active ingredients and potential targets of Chinese herbal Radix Curcumae formula for application to cardiovascular disease. *J Ethnopharmacol*, 145, 1-10.
- Tautenhahn, R., G. J. Patti, D. Rinehart & G. Siuzdak (2012) XCMS Online: a web-based platform to process untargeted metabolomic data. *Analytical chemistry*, 84, 5035-5039.
- Teo, C. C., S. N. Tan, J. W. Hong Yong, T. Ra, P. Liew & L. Ge (2011) Metabolomics analysis of major metabolites in medicinal herbs. *Analytical Methods*, 3, 2898-2908.
- Thulesius, O., J. Gjöres & E. Berlin (1979) Vasoconstrictor effect of midodrine, ST 1059, noradrenaline, etilefrine and dihydroergotamine on isolated human veins. *European Journal of Clinical Pharmacology*, 16, 423-424.
- Tian, H., S. M. Lam & G. Shui (2016) Metabolomics, a Powerful Tool for Agricultural Research. *Int J Mol Sci*, 17, 1871.
- Tisler, T. & J. Zagorc-Koncan (1997) Comparative assessment of toxicity of phenol, formaldehyde, and industrial wastewater to aquatic organisms. *Water, Air, and Soil Pollution*, 97, 315-322.
- Torjesen, I. (2015) Drug development: the journey of a medicine from lab to shelf. *Pharmaceutical Journal*.
- Truchon, J. & C. Bayly (2006) GLARE-A free open source software for combinatorial library design. *Journal of Chem Inf Model*, 46, 1536-1548.

- Truszkowski, A., K. V. Jayaseelan, S. Neumann, E. L. Willighagen, A. Zielesny & C. Steinbeck (2011) New developments on the cheminformatics open workflow environment CDK-Taverna. *Journal of cheminformatics*, 3, 54.
- Uddin, Q., L. Samiulla, V. Singh & S. Jamil (2012) Phytochemical and pharmacological profile of *Withania somnifera* Dunal: a review. *Journal of Applied Pharmaceutical Science*, 2, 170-175.
- Umbelliferae, A. *Bacopa monnieri* (Linn.) Penn.
- Vacca, R. A., D. Valenti, S. Caccamese, M. Daglia, N. Braidy & S. M. Nabavi (2016) Plant polyphenols as natural drugs for the management of Down syndrome and related disorders. *Neurosci Biobehav Rev*, 71, 865-877.
- Valitova, J., A. Sulkarnayeva & F. Minibayeva (2016) Plant sterols: diversity, biosynthesis, and physiological functions. *Biochemistry (Moscow)*, 81, 819-834.
- Van Drie, J. H. & M. S. Lajiness (1998) Approaches to virtual library design. *Drug Discovery Today*, 3, 274-283.
- Verma, P., A. K. Mathur, S. P. Jain & A. Mathur (2012) In vitro conservation of twenty-three overexploited medicinal plants belonging to the Indian sub continent. *The Scientific World Journal*, 2012, 10.
- Verpoorte, R. (1998) Exploration of nature's chemodiversity: the role of secondary metabolites as leads in drug development. *Drug Discovery Today*, 3, 232-238.
- Vlase, L., M. Parvu, E. Parvu & A. Toiu (2013) Chemical constituents of three *Allium* species from Romania. *Molecules*, 18, 114-127.
- Wachira, S. W., S. Omar, J. W. Jacob, M. Wahome, H. T. Alborn, D. R. Spring, D. K. Masiga & B. Torto (2014) Toxicity of six plant extracts and two pyridone alkaloids from *Ricinus communis* against the malaria vector *Anopheles gambiae*. *Parasites & Vectors*, 7, 312.
- Wadood, A., N. Ahmed, L. Shah, A. Ahmad, H. Hassan & S. Shams (2013) In-silico drug design: An approach which revolutionarised the drug discovery process. *OA drug design & delivery*, 1, 3-7.
- Wang, X. B., G. H. Li, L. J. Zheng, K. Y. Ji, H. Lü, F. F. Liu, L. Z. Dang, M. H. Mo & K. Q. Zhang (2009) Nematicidal Cardenolides from *Nerium indicum* Mill. *Chemistry & biodiversity*, 6, 431-436.
- Weber, L. 2008. JChem Base-ChemAxon. 65-66. ROYAL SOC CHEMISTRY THOMAS GRAHAM HOUSE, SCIENCE PARK, MILTON RD, CAMBRIDGE
- Wei, C.-H., H.-Y. Kao & Z. Lu (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41, W518-W522.
- Wen, Z., M. R. Berenbaum & M. A. Schuler (2006) Inhibition of CYP6B1-mediated detoxification of xanthotoxin by plant allelochemicals in the black swallowtail (*Papilio polyxenes*). *Journal of chemical ecology*, 32, 507.
- Wickham, H. (2011) ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3, 180-185.
- Wink, M. (1988) Plant breeding: importance of plant secondary metabolites for protection against pathogens and herbivores. *Theoretical and applied genetics*, 75, 225-233.
- Wishart, D. S. (2007) Introduction to cheminformatics. *Current protocols in bioinformatics*, 18, 14-1.
- Wishart, D. S., Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li & Z. Sayeeda (2017) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46, D1074-D1082.

- Wojdylo, A., J. Oszmianski & R. Czemerys (2007) Antioxidant activity and phenolic compounds in 32 selected herbs. *Food Chemistry*, 105, 940-949.
- Woodard, R., M. Tsai, H. Floss, P. Crooks & J. Coward (1980) Stereochemical course of the transmethylation catalyzed by catechol O-methyltransferase. *Journal of Biological Chemistry*, 255, 9124-9127.
- Wu, J. & I. T. Baldwin (2010) New insights into plant responses to the attack from insect herbivores. *Annual review of genetics*, 44, 1-24.
- Wu, W., Q. Zhang, Y. Zhu, H.-M. Lam, Z. Cai & D. Guo (2008) Comparative metabolic profiling reveals secondary metabolites correlated with soybean salt tolerance. *Journal of agricultural and food chemistry*, 56, 11132-11138.
- Xu, J. & A. Hagler (2002) Chemoinformatics and drug discovery. *Molecules*, 7, 566-600.
- Yamaya, A., Y. Endo, K. Fujimoto & K. Kitamura (2007) Effects of genetic variability and planting location on the phytosterol content and composition in soybean seeds. *Food chemistry*, 102, 1071-1075.
- Yang, F., Z.-G. Yue, X. Wang, X.-P. Zhang, J. Chai, J.-C. Cui, X.-M. Song & Q.-B. Mei (2014) Chemical constituents of leaf of *Eucommia ulmoides*. *Zhongguo Zhong yao za zhi= Zhongguo zhongyao zazhi= China journal of Chinese materia medica*, 39, 1445-1449.
- Yang, M., C. Cheng, J. Yang & D. A. Guo (2012) Metabolite profiling and characterization for medicinal herbal remedies. *Curr Drug Metab*, 13, 535-57.
- Yasri, A., D. Berthelot, H. Gijzen, T. Thielemans, P. Marichal, M. Engels & J. Hoflack (2004) REALISIS: a medicinal chemistry-oriented reagent selection, library design, and profiling platform. *Journal of Chemical Information and Computer Sciences*, 44, 2199-2206.
- Yoo, I., P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang & L. Hua (2012) Data mining in healthcare and biomedicine: a survey of the literature. *Journal of Medical Systems*, 36, 2431-2448.
- Yoshikawa, M., N. Nishida, H. Shimoda, M. Takada, Y. Kawahara & H. Matsuda (2001) Polyphenol constituents from *Salacia* species: quantitative analysis of mangiferin with alpha-glucosidase and aldose reductase inhibitory activities. *Yakugaku zasshi: Journal of the Pharmaceutical Society of Japan*, 121, 371-378.
- Yu, H. & A. Adedoyin (2003) ADME-Tox in drug discovery: integration of experimental and computational technologies. *Drug Discovery Today*, 8, 852-861.
- Yuan, T. T., N. D. Zhang, Y. J. He, M. Li, H. T. Xu & Q. Y. Zhang (2014) [Research progress of phytoestrogens-like chemical constituents in natural medicines]. *Zhongguo Zhong Yao Za Zhi*, 39, 4526-31.
- Zagrobelny, M. & B. L. Møller (2011) Cyanogenic glucosides in the biological warfare between plants and insects: The Burnet moth-Birdsfoot trefoil model system. *Phytochemistry*, 72, 1585-1592.
- Zainal, B., M. Abdah, Y. Taufiq-Yap, A. Roslida & K. Rosmin (2014) Anticancer agents from non-edible parts of *Theobroma cacao*. *Natural Products Chemistry & Research*, 2, 2-8.
- Zare, K., A. Movafeghi, S. A. Mohammadi, S. Asnaashari & H. Nazemiyeh (2014) New Phenolics from *Linum mucronatum* subsp. *orientale*. *Bioimpacts*, 4, 117-22.

- Zavala, J. A., P. D. Nability & E. H. DeLucia (2013) An emerging understanding of mechanisms governing insect herbivory under elevated CO₂. *Annual Review of Entomology*, 58, 79-97.
- Zemer, D., A. Livneh, Y. L. Danon, M. Pras & E. Sohar (1991) Long-term colchicine treatment in children with familial mediterranean fever. *Arthritis & Rheumatism*, 34, 973-977.
- Zenk, M. H. & M. Juenger (2007) Evolution and current status of the phytochemistry of nitrogenous compounds. *Phytochemistry*, 68, 2757-72.
- Zhao, J., P. Jiang & W. Zhang (2010) Molecular networks for the study of TCM pharmacology. *Brief Bioinform*, 11, 417-30.
- Zhong, H. A., V. Mashinson, T. A. Woolman & M. Zha (2013) Understanding the molecular properties and metabolism of top prescribed drugs. *Curr Top Med Chem*, 13, 1290-307.
- Zhou, J. R. & J. W. Erdman, Jr. (1995) Phytic acid in health and disease. *Crit Rev Food Sci Nutr*, 35, 495-508.
- Zhou, Y., J. Zheng, Y. Li, D. P. Xu, S. Li, Y. M. Chen & H. B. Li (2016) Natural Polyphenols for Prevention and Treatment of Cancer. *Nutrients*, 8, 515.
- Zloh, M., E. G. Samaras, J. Calvo-Castro, A. Guirguis, J. L. Stair & S. B. Kirton (2017) Drowning in diversity? A systematic way of clustering and selecting a representative set of new psychoactive substances. *RSC Advances*, 7, 53181-53191.
- Zou, C., Y. Wang, H. Zou, N. Ding, N. Geng, C. Cao & G. Zhang (2019) Sanguinarine in *Chelidonium majus* induced antifeeding and larval lethality by suppressing food intake and digestive enzymes in *Lymantria dispar*. *Pesticide biochemistry and physiology*, 153, 9-16.

Supplementary Data

The supplementary data are available on GitHub and Zenodo as a citable resource, as follows:

1. GitHub link: Ph.D. Thesis Data – <https://github.com/DivyaKarade/PhD-Thesis-Data>
2. Chapter-1: Introduction and Review of Literature - <https://doi.org/10.5281/zenodo.5637086>
3. Chapter-2: Design of Novel Drug-like Molecules using Informatics Rich Secondary Metabolites Analysis of Indian Medicinal and Aromatic Plants - <https://doi.org/10.5281/zenodo.5637174>
4. Chapter-3: Bridging In-Silico and Experimental: Chemoinformatics Investigation for Mass Spectrometry-Based Metabolomics Study of Soybean - <https://doi.org/10.5281/zenodo.5637185>
5. Chapter-4: Chemoinformatics Investigation on Chemical Defense in Plants - <https://doi.org/10.5281/zenodo.5637187>

Mass Spectrometry Data Availability: Chapter-3

The metabolomics data (MS-1) reported in this chapter are available at METASPACE with Datasets ID: soy_LS_pos and soy_LS_neg (<https://metaspace2020.eu/datasets?q=soy>) and MS-2 data in the GNPS-MassIVE repository with MassIVE ID: MSV000087957 (<https://massive.ucsd.edu/ProteoSAFe/dataset.jsp?task=b593b656b3e24e4fab682c13c4acef42>).

Supplementary Table S1.1: Bioactive compounds identified from food crops and medicinal plants with their protein targets involved in respective diseases: Bioactive compounds identified from food crops and medicinal plants with their references.

Supplementary Table S1.2: Bioactive compounds identified from food crops and medicinal plants with their protein targets involved in respective diseases- Protein targets for their respective bioactive compounds with their references.

Supplementary Table S1.3: Bioactive compounds identified from food crops and medicinal plants with their protein targets involved in respective diseases- Protein targets involved in various diseases through their respective pathways with their references.

Supplementary file S2.1: Computational Protocol for Identification of ‘specific’ metabolites and building ‘virtual library’ with predicted bioactivity.

Supplementary Table S2.1.1: Bioactivity studies done on Indian medicinal plants (n=10).

Supplementary Table S2.1.2: Comparison of Indian medicinal plants of the present study by FRLHT (<http://www.medicinalplants.in/>) with IMPATT database of Indian medicinal plants.

Supplementary Table S2.2: Plant molecules (n= 3459) text mined by PubTator from their respective medicinal plants (n=104) PubMed abstracts.

Supplementary Table S2.3.1: SMILES strings for plant metabolites (n= 1665) mined from PubMed literature for 104 Indian medicinal plants with their extracted scaffolds (209) and functional groups (97).

Supplementary Table S2.3.2: SMILES strings for approved drug molecules (n=2354) downloaded from DrugBank with their extracted scaffolds (306) and functional groups (291).

Supplementary Table S2.4.1: Indian medicinal plants families with their PubMed abstracts.

Supplementary Table S2.4.2: Classification of Indian medicinal plants according to their therapeutic properties.

Supplementary Table S2.5: A 2D principal component analysis (PCA) of 1665 ring containing metabolites by generating their descriptors (n=16).

Supplementary Table S2.6: SMILES of common plant molecules and scaffolds identified in Indian medicinal and aromatic plants molecules, drug molecules and scaffolds merged network as depicted in organic and edge-weighted spring embedded layout (for selected nodes only) in Cytoscape.

Supplementary Table S2.7.1: Plant-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Plant-based clustering (6 clusters).

Supplementary Table S2.7.2: Plant-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 1 (525).

Supplementary Table S2.7.3: Plant-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 2 (400).

Supplementary Table S2.7.4: Plant-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 3 (575).

Supplementary Table S2.7.5: Plant-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 4 (625).

Supplementary Table S2.7.6: Plant-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 5 (575).

Supplementary Table S2.7.7: Plant-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 6 (415).

Supplementary Table S 2.7.8: Plant-based clusters of virtual libraries with their total sum of toxicophoric, chemophoric and pharmacophoric fingerprints.

Supplementary Table S2.7.9: Dendrograms for Toxicophoric (a), Pharmacophoric (b) and Chemophoric (c) fingerprints of virtual library molecules based on plant-based clustering

Supplementary Table S2.8.1: Scaffold or fragment-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Scaffold or fragment-based clustering (6 clusters).

Supplementary Table S2.8.2: Scaffold or fragment-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 1 (285) in scaffold or fragment-based clustering pattern.

Supplementary Table S2.8.3: Scaffold or fragment-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 2 (165) in scaffold or fragment-based clustering pattern.

Supplementary Table S2.8.4: Scaffold or fragment-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 3 (425) in scaffold or fragment-based clustering pattern.

Supplementary Table S2.8.5: Scaffold or fragment-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 4 (63) in scaffold or fragment-based clustering pattern.

Supplementary Table S2.8.6: Scaffold or fragment-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 5 (6) in scaffold or fragment-based clustering pattern.

Supplementary Table S2.8.7: Scaffold or fragment-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Virtual library for Cluster 6 (88) in scaffold or fragment-based clustering pattern.

Supplementary Table S2.8.8: Scaffold or fragment-based clusters of virtual libraries with their Toxicophoric, chemophoric and pharmacophoric fingerprints: Toxicophoric, chemophoric and pharmacophoric fingerprints difference classified according to their six scaffold or fragment-based clustering of molecules.

Supplementary Table S3.1.1: Reported Soybean small molecules [SoyKb (http://soykb.org/search/fuzzy_search.php?metabolite=)+SoyCyc (<https://pmn.plantcyc.org/SOY/class-tree?object=Compounds#>)] + Text mined soy molecules] (n=1622).

Supplementary Table S3.1.2: SMILES strings for approved drug molecules (n= 2354) downloaded from DrugBank.

Supplementary Table S3.1.3: Filtered soybean molecules on the basis of molecular weight ≤ 1000 and ring content ≤ 6 (n= 660) with other 2D descriptors.

Supplementary Table S3.1.4: Approved Drug molecules with their 2D descriptors.

Supplementary Table S3.1.5: SMILES strings for scaffolds (n= 58) and functional groups (n= 59) extracted from soybean molecules (660).

Supplementary Table S3.1.6: SMILES strings for scaffolds (n= 306) and functional groups (n= 291) extracted from drug molecules.

Supplementary Table S3.2.1: XCMS report for positive ion mode.

Supplementary Table S3.2.2: XCMS report for negative ion mode.

Supplementary Table S3.3.1: List of identified metabolites by ProbMetab and PUTMEDID-LCMS methods in seed and leaf tissue samples of four soybean varieties.

Supplementary Table S3.3.2: Small organic molecules identified by combining two methods [ProbMetab- R package (for probability score:1) and PUTMEDID LC-MS in Taverna workflow (up to 2 ppm error)] in soybean leaf (n=18,020) and seed (n=14,847) tissue samples.

Supplementary Table S3.3.3: List of identified metabolites (KEGG IDs) by ProbMetab method in four varieties of soybean.

Supplementary Table S3.4.1: Small organic molecules (7185) containing reported (557) and non-reported (6628) molecules identified and annotated putatively by

combining two methods (ProbMetab R package and PUTMEDID-LCMS) with highest probability score =1 and up to 2 ppm error.

Supplementary Table S3.4.2: Small organic molecules [(n=20):-Reported (6) and Novel small organic molecules (14)] validated by tandem mass spectrometry experiments.

Supplementary Table S3.4.3: Separation of compounds (EIC - embedded as picture) by UHPLC-MS for validated molecules by MS/MS (n=20) with ppm error=2.

Supplementary Table S3.5: SMILES of common plant molecules and scaffolds identified in soybean small molecules, drug molecules and scaffolds merged network as depicted in an organic layout in Cytoscape.

Supplementary Table S3.6: Virtual library of novel molecules (n= 1225) built from soybean metabolites(n=5) with their drug-like and lead-like properties (PDL: Progressive Drug like, PLL: Progressive Lead-like, T: Toxicophoric, P: Pharmacophoric, C: Chemophoric fingerprints etc.) and other descriptors.

Supplementary Table S4.1.1: List of plants (n=162) related to chemical defense and allelochemicals (n=280) identified from them.

Supplementary Table S4.1.2: List of unique pesticide molecules (n=1985, collected from PPDB database) screened with molecular rings- 5 to 6, and mol. weight <=1000.

Supplementary Table S4.1.3: Phytochemical classification of allelochemicals.

Supplementary Table S4.1.4: List of scaffolds (n=74) and functional groups (n=33) extracted from allelochemicals (n=280).

Supplementary Table S4.1.5: List of scaffolds (n=56) and functional groups (n=62) extracted from pesticide molecules (n=1985).

Supplementary Table S4.1.6: Common scaffolds and molecules identified through scaffold molecule network.

Supplementary Table S4.1.7: List of allelochemicals (n=39/280) marketed as biopesticides.

Supplementary Table S4.2.1: List of novel molecules of virtual library (VL) (n=380) built from selected allelochemicals scaffolds (n=5) and their functional groups (n=5).

Supplementary Table S4.2.2: TPC model graphs of VL and pesticides.

Supplementary Table S4.2.3: Comparison model of TPC proportions in allelochemicals specific VL molecules and pesticides.

Supplementary Table S4.2.4: List of VL molecules (n=208/380) screened with pesticide-likeness according to Hao et al. rules.

Supplementary Table S4.2.5: Pesticidal toxic activities (LC50, EC50) for lower aquatic organisms of VL pesticide-like molecules (n=208) calculated from ECOSAR.

Supplementary Table S4.2.6: Virtually screened molecules (n=169) from VL pesticide-like molecules (n=208) with pesticidal toxic activities (LC50, EC50) ≥ 1 for lower aquatic organisms calculated from ECOSAR.

Abstract

Name of the Student: Divya Karade

Registration No. : 10BB15J26034

Faculty of Study: Biological science

Year of Submission: 2021

AcSIR academic center/CSIR Lab: CSIR NCL, Pune

Name of the Supervisor(s): Dr. M. Karthikeyan (Supervisor), Dr. Narendra Y. Kadoo (Co-supervisor)

Title of the thesis: Chemoinformatics based investigation of plant metabolites for their medicinal and crop protection values

Bioactive compound scaffolds are the key to optimize chemical diversity in natural products. Furthermore, large data of published findings are available, which needs to be analyzed and managed for the best use and efficacy of the available products. We used chemoinformatics methods to design novel molecules (drug-like, lead-like, and pesticide-like) based on the metabolomics of Indian medicinal and aromatic plants, food crops i.e., soybean, and plants involved in chemical defense. To validate the results, we also performed mass spectrometry experiments of four Indian varieties of soybean to confirm the presence of the molecules in soybean seed and leaf tissues grown in a plot at the Biochemical Sciences Division in CSIR-NCL, Pune, India. LC-MS data were also used for multivariate analysis. This study suggests a combinatorial approach employing high-throughput metabolomics and chemoinformatics methods to efficiently identify new drug-like plant metabolites for targeted drug development. Similarly, we have also designed allelochemicals-specific environment-friendly novel molecules inspired by pesticidal activities through the chemoinformatics approach. Thus, this study has predicted several potential novel molecules concerning drugs and pesticides, which can be subsequently purified and experimentally evaluated in detail in the future.

Details of the Publications

List of publications

1. **Divya Karade**, Durairaj Vijayasarithi, Narendra Kadoo, Renu Vyas, P.K. Ingle and Muthukumarasamy Karthikeyan*, “Design of Novel Drug-like Molecules Using Informatics Rich Secondary Metabolites Analysis of Indian Medicinal and Aromatic Plants”, *Combinatorial Chemistry & High Throughput Screening* (2020) 23: 1113. <https://doi.org/10.2174/1386207323666200606211342>
2. **Divya Karade**, Chemoinformatics based investigation of plant metabolites for their medicinal and crop protection values - Introduction and Review of Literature. Zenodo. (2021). (<https://doi.org/10.5281/zenodo.5637086>)
3. **Divya Karade**, Design of Novel Drug-like Molecules using Informatics Rich Secondary Metabolites Analysis of Indian Medicinal and Aromatic Plants. Zenodo. (2021). (<https://doi.org/10.5281/zenodo.5637174>)
4. **Divya Karade**, Bridging In-Silico and Experimental: Chemoinformatics Investigation for Mass Spectrometry-Based Metabolomics Study of Soybean. Zenodo. (2021). (<https://doi.org/10.5281/zenodo.5637185>)
5. **Divya Karade**. Chemoinformatics Investigation on Chemical Defense in Plants. Zenodo. (2021). (<https://doi.org/10.5281/zenodo.5637187>)
6. **Divya Karade**, Swapnil Mundhe, Narendra Kadoo, Milind Ratnaparkhe, Renu Vyas and Muthukumarasamy Karthikeyan*, “Bridging In-Silico and Experimental: Chemoinformatics Analysis for Mass Spectrometry-Based Metabolomics study of Soybean” (2021). (*Under review*)

List of posters presented

1. **Divya Karade**, N. Kadoo, M. Karthikeyan. Chemoinformatics Investigation of Organic Metabolites from Soybean Reveals Common Drug Scaffolds presented at CSIR-National Chemical Laboratory, Pune on the “National Science Day” 2017.

Abstract

Soybean (*Glycine max* L. Merr.) is a globally important edible crop due to its high seed protein and oil contents. Soybean is also known for its medicinal importance and contains various organic metabolites, which are valuable sources for drug development. In the present study, we performed rigorous chemoinformatics analyses to identify drug-like and lead-like compounds from soybean. For this purpose, a set of molecules from ChEMBL, DrugBank, soybean phytochemicals mined from literature, and other soybean metabolites from PMN were used to extract corresponding molecular scaffolds. The scaffolds were supplied with linkers and functional groups to enumerate diverse virtual libraries and prioritize the molecules by annotation with drug-like and lead-like scores. A scaffold-drug-molecule network was constructed and analyzed to reveal common scaffolds among drug molecules, organic soybean metabolites, and other bioactive molecules. This comprehensive study will be useful to design novel bioactive molecules inspired by organic metabolites of soybean.

2. **Divya Karade**, N. Kadoo, M. Karthikeyan. Drug Design based on Metabolomics of Indian Medicinal and Aromatic Plants: A Chemoinformatics Approach presented at CSIR-National Chemical Laboratory, Pune on the “National Science Day” 2017.

Abstract

Indian medicinal plant species are used in medicinal practices from the most ancient traditional times. Drug discovery from natural origin is one of the important factors for human development and evolution. In the present study, we successfully performed and advanced the step of drug discovery. Herein, we reviewed the recent development of chemoinformatics tools and methods. We also performed the chemoinformatics analysis and designed novel molecules inspired by organic metabolites of Indian medicinal and aromatic plants. For this purpose, 887 ring-containing molecules identified from text mining of literature related to 104 plant species were used to extract molecular scaffolds. With the help of chemoinformatics tools, diverse scaffolds were used for building a focused virtual library. Virtual screening was done with cluster analysis to predict drug-like and lead-like molecules from these metabolites in the context of drug discovery.