

DEVELOPMENT OF A DATA COLLECTION AND ANALYSIS TOOL FOR
PROTEIN – LIGAND INTERACTIONS

by

Mehmet Aziz Yirik

B.S., Mathematics, Mimar Sinan Fine Arts University, 2013

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computational Science and Engineering

Boğaziçi University

2017

DEVELOPMENT OF A DATA COLLECTION AND ANALYSIS TOOL FOR
PROTEIN – LIGAND INTERACTIONS

APPROVED BY:

Assoc. Prof. Elif Özkırmılı
(Thesis Supervisor)

Prof. Kutlu Ülgen
(Thesis Co-supervisor)

Assoc. Prof. Burak Alakent

Assoc. Prof. Arzucan Özgür

Assist. Prof. Nihat Alpıgu Sayar

DATE OF APPROVAL: 17/07/2017

ACKNOWLEDGEMENTS

First, I would like to thank my thesis supervisor, Assoc. Prof. Elif Özkırmılı Ölmez and my thesis co-supervisor, Prof. Kutlu Ülgen. Their guidance, support and trust were invaluable for me to complete my study.

I am very grateful to my committee members, Assoc. Prof. Burak Alakent, Assoc. Prof. Arzucan Özgür and Assist. Prof. Nihat Alpogu Sayar for their valuable recommendations.

I would like to thank the members of KB407 for their good friendship and encouragement. I am also grateful to Hakime for her help.

Distances do not matter for such great friendships. I would like to thank my friends from different countries for their support and encouraging messages. I especially would like to thank Anwar Isied from Palestinian Neuroscience Initiative.

I would like to especially thank my family for their love and support. I dedicate my thesis to the three beloved women in my life: My grandmother, my mother and my sister.

Last, I would like to acknowledge Technological Research Council of Turkey (TUBITAK Project 115S208) for financial support.

ABSTRACT

DEVELOPMENT OF A DATA COLLECTION AND ANALYSIS TOOL FOR PROTEIN – LIGAND INTERACTIONS

Analysis of protein – ligand interactions guides the development of new drugs. For protein - ligand interaction studies, first step is the construction of an accurate dataset. This data collection process can be completed either by manual search in databases or by using computer-assisted data collection methods. Manual data collection is difficult, time consuming and prone to errors. In this work, we present a novel tool to collect protein-ligand interaction data. We first introduce a protein – ligand interaction data collection tool using UniProt, ChEMBL, PubChem, PDB and BindingDB as its source databases. In the second part, we use this tool to analyze protein – ligand interactions of sphingolipid and insulin metabolisms. First, the datasets of both metabolisms were constructed, then their ligand centric network models were built for ligand analysis. Based on these networks, first the interactions within sphingolipid metabolism proteins, then their interactions with insulin proteins were analyzed. According to the ligand analysis, specific interactions and significant drugs were highlighted. Besides promiscuous drugs interacting with too many proteins, Tamoxifen and Altretamine cancer drugs interacted with key sphingolipid proteins, namely GLCM, ARSA and AGAL. Ceritinib, used for the treatment of non-small cell lung cancer, interacted with Kit and Lyn kinases. This ligand based interaction network analysis highlighted the synergy between these two networks.

ÖZET

PROTEİN - LİGAND ETKİLEŞİMLERİ İÇİN BİR VERİ TOPLAYICI VE ANALİZ ARACI GELİŞTİRME

Protein - ligand etkileşimlerinin analizi, yeni ilaçların gelişimine dönük çalışmaları yönlendirmektedir. Protein - ligand etkileşimi ile alakalı çalışmalar için, ilk adım doğru bir veri kümesinin oluşturulmasıdır. Bu veri toplama işlemi ya veritabanlarında manuel arama ya da bilgisayar destekli veri toplama yöntemleri kullanılarak tamamlanabilir. Manuel veri toplama, zaman alıcı ve hatalara açık olması nedeniyle zordur. Bu çalışmada, protein-ligand etkileşim verileri toplamak için yeni bir araç sunulmuştur. İlk olarak UniProt, ChEMBL, PubChem, PDB ve BindingDB'yi kaynak veritabanları olarak kullanan bir protein - ligand etkileşimi veri toplama aracı tanıtılmıştır. İkinci bölümde, sfingolipid ve insülin metabolizmalarının protein - ligand etkileşimlerini analiz etmek için bu araç kullanılmıştır. İlk olarak, her iki metabolizmanın veri setleri oluşturulup, ardından ligand merkezli ağ modelleri ligand analizi için inşa edilmiştir. Bu ağlara dayanarak, önce sfingolipid metabolizma proteinleri içindeki etkileşimleri, daha sonra insülin proteinleri ile olan etkileşimleri analiz edilmiştir. Ligand analizine göre spesifik etkileşimler ve önemli ilaçlar vurgulanmıştır. Bir çok sayıda proteinle etkileşime giren ilaçların yanı sıra, Tamoksifen ve Altretamin kanseri ilaçları, önemli Sfingolipit proteinleri, yani GLCM, ARSA ve AGAL ile etkileşime girmiştir. Küçük hücreli olmayan akciğer kanserinin tedavisinde kullanılan Ceritinib, Kit ve Lyn kinazlarıyla etkileşime girmiştir. Bu ligand tabanlı etkileşim ağı analizi, bu iki ağ arasındaki sinerjiyi vurgulamıştır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	iii
ABSTRACT.....	iv
ÖZET.....	v
LIST OF TABLES.....	ix
LIST OF ACRONYMS / ABBREVIATIONS.....	xv
1. INTRODUCTION.....	1
2. THEORY.....	4
2.1. Sphingolipid and Insulin Metabolisms.....	4
2.2. Protein-Ligand Interaction (PLI) Data.....	5
2.2.1. Databases.....	5
2.2.1.1. UniProt.....	6
2.2.1.2. PDB.....	6
2.2.1.3. ChEMBL.....	6
2.2.1.4. Other Protein-Ligand Databases.....	7
2.2.2. Programmatic Access to Databases.....	7
2.2.2.1. BioServices.....	8
2.2.2.2. ChEMBL Webresource Client.....	9
2.3. Protein-Protein Interaction (PPI) Networks.....	9
2.3.1. Ligand Centric Network Model (LCNM).....	10
2.3.1.1. Identity Network Model.....	11
2.3.2. Clustering Analysis.....	12
2.3.2.1. MCL Algorithm.....	12
2.4. Ligand Analysis.....	13
2.4.1. Matplotlib.....	14
2.4.2. Protein – Ligand Docking.....	14
2.4.3. Molecular Symmetry and Chemical Features.....	16

3. METHODS.....	19
3.1. Data Collection.....	19
3.2. Database Based Protein-Ligand Interaction Data Collectors.....	21
3.2.1. ChEMBL.....	21
3.2.2. PDB.....	28
3.2.3. BindingDB.....	29
3.2.4. KEGG.....	31
3.2.5. PubChem.....	32
4. RESULTS.....	35
4.1. Sphingolipid and Insulin Data Summary.....	35
4.2. Sphingolipid and Insulin Ligand Centric Networks.....	36
4.2.1. Construction of SL-WIN Network.....	36
4.2.1.1. Interactor Analysis of Sphingolipid Network.....	38
4.2.1.2. Scaffold Analysis and ZINC Database Search.....	49
4.2.1.3. Protein – Ligand Docking of SL-WIN.....	58
4.2.2. Construction of Combined Network of Sphingolipids and Insulins.....	61
4.2.2.1. Interactor Analysis of Combined Sphingolipid and Insulin Network.....	63
4.2.2.2. Protein-Ligand Docking of Key Sphingolipid-Insulin Interactions.....	71
4.3. Comparison Between Sphingolipids and Inflammation Enzymes.....	75
4.3.1. Sphingolipid and Inflammation Data Summary.....	75
4.3.2. Construction of Inflammation Network.....	77
4.3.3. Interactor Analysis for Intersection of Sphingolipids and Inflammation Proteins.....	79
5. CONCLUSION.....	81
5.1. Conclusions.....	81
5.2. Further Studies.....	83

REFERENCES..... 84

LIST OF TABLES

Table 2.1. The list of symmetry operations.	17
Table 2.2. The list of used symmetry groups and their operations. Here, n is an ordinal number.	17
Table 3.1. The number of ligands each database includes.	20
Table 4.1. The number of protein IDs and ligands extracted for sphingolipid and insulin metabolism from UniProt and ChEMBL.	35
Table 4.2. For sphingolipid clusters, number of proteins, number of ligands and ligand pairwise similarity values are listed.	41
Table 4.3. The drugs of C1 of SL-WIN.	42
Table 4.4. The observed known drugs and the protein interactions from cluster 1 of the SL-WIN network are listed.	46
Table 4.5. The drugs of the cluster 2 are listed with their names, ChEMBL IDs and related diseases	47
Table 4.6. The specific interactions of the SL-WIN cluster 2 with their rarely seen drugs.	49

Table 4.7. The scaffolds of sphingolipid clusters are listed by their ChemSpider names and their figures.	51
Table 4.8. The number of scaffolds and the number of detected ligands are listed.	58
Table 4.9. For SL-WIN, the ligands docked into more than one protein with docking scores below -6 are listed.	59
Table 4.10. For SPHINS clusters, number of proteins, number of ligands and ligand pairwise similarity values are listed.	66
Table 4.11. 23 anticancer drugs were observed in C1-SPHINS.	67
Table 4.12. The drugs interacting less than 7 proteins pairs in C1 of SPHINS are listed with the interacted protein pairs.	68
Table 4.13. These three drugs and their protein interactions from C2 of the SPHINS network are listed.	70
Table 4.14. The list of clusters 2's proteins, ligands and their XP Gscores.	71
Table 4.15. Protein numbers of both sphingolipids and inflammations.	76
Table 4.16. The list of the proteins observed at the intersection between both protein families.	73

Table 4.17. 19 commercial drugs were observed in sphingolipid-inflammation intersection network.....	79
---	----

LIST OF FIGURES

Figure 3.1. Pseudo code of ChEMBL data collection script.....	23
Figure 3.2. The flowchart of protein-ligand interaction process.....	25
Figure 3.3. Pseudo code of ChEMBL data collection script based on threshold criteria	26
Figure 3.4. The options for threshold selection.....	27
Figure 3.5. Pseudo code of PDB data collection script.....	29
Figure 3.6. Pseudo code of BindingDB data collection script.....	30
Figure 3.7. Pseudo code of KEGG data collection script	32
Figure 3.8. Pseudo code of PubChem data collection script.....	34
Figure 4.1. The WIN of sphingolipid metabolism	37
Figure 4.2. The clusters of the SL-WIN.....	37
Figure 4.3. Number of shared ligands higher than 50.....	38

Figure 4.4. PLS distribution of SL-WIN clusters.	39
Figure 4.5. Scaffolds identified by scaffold decomposition.	50
Figure 4.6. Number of shared ligands higher than 50.....	61
Figure 4.7. (A) SPHINS-WIN, where insulins and sphingolipids are shaped as ellipse and rectangles, respectively. (B) The clusters of SPHINS- WIN.....	62
Figure 4.8. The similarity distribution of the weighted identity SPHINS clusters. X index is for the similarity values and Y is for the frequency of these similarity values.	64
Figure 4.9. The second cluster of weighted identity SPHINS network. Green hexagonal nodes represent the first sphingolipid neighbours of the insulins coloured with light blue and shaped as ellipse.	71
Figure 4.10. The third cluster of weighted identity SPHINS-WIN. Green hexagonal nodes represent the target proteins bridging both metabolisms. In the graphic, sphingolipid and insulin proteins are shaped as rectangular and ellipse, respectively.....	74
Figure 4.11. WIN of inflammation.	78

Figure 4.12. The intersection of both networks is illustrated..... 78

Figure 4.13. The frequency distribution of PLS values for Inflammation

ligand set. 79

LIST OF ACRONYMS / ABBREVIATIONS

2D	Two Dimensional
3D	Three Dimensional
AC	Activity Concentration
ACER1	Alkaline ceramidase 1
ACER2	Alkaline ceramidase 1
AGAL	Alpha-galactosidase A
AKT1	RAC-alpha serine/threonine-protein kinase
AKT2	RAC-beta serine/threonine-protein kinase
API	Application programming interface
ARSA	Arylsulfatase A
ARSB	Arylsulfatase B
ASAH1	Acid ceramidase
ASAH2	Neutral ceramidase
ASM	Sphingomyelin phosphodiesterase
BGAL	Beta-galactosidase
BindingDB	Binding Database
CEGT	Ceramide glucosyltransferase
ChEMBL	Chemical database of European Molecular Biology Laboratory
ChemSpider	Chemical Structure Database
CID	Compound Identification
CSK	Tyrosine-protein kinase CSK
DHB12	Very-long-chain 3-oxoacyl-CoA reductase
DrugBank	Drug Database

EC ₅₀	Half maximal effective concentration
ELOV	Elongation of very long chain fatty acids protein
ENPP7	Ectonucleotide pyrophosphatase/phosphodiesterase family member 7
FOXO1	Forkhead box protein O1
FYN	Tyrosine-protein kinase Fyn
GBA2	Non-lysosomal glucosylceramidase
GBA3	Cytosolic beta-glucosidase
GLCM	Glucosylceramidase
GO	Gene Ontology
GSK3A	Glycogen synthase kinase-3 alpha
HTML	HyperText Markup Language
IC ₅₀	The half maximal inhibitory concentration
ID	Identification
IDE	Insulin-degrading enzyme
IGF1R	Insulin-like growth factor 1 receptor
INSR1	Insulin receptor
Jar	JAVA Archive
Jmol	JAVA Molecule Viewer
KEGG	Kyoto Encyclopedia of Genes and Genomes
Ki	Binding affinity of the inhibitor
KIT	Mast/stem cell growth factor receptor Kit
KPCD	Protein kinase C delta type
KPCD1	Serine/threonine-protein kinase D1
LCNM	Ligand Centric Network Model

LPPI	Ligand Centric Protein-Protein Interaction
LYAM3	P-selectin
LYN	Tyrosine-protein kinase Lyn
MCL	Markov Clustering Algorithm
MS	Multiple sclerosis
NCBI	National Center for Biotechnology Information
NEUR1	Sialidase-1
NEUR2	Sialidase-2
NEUR3	Sialidase-3
NEUR4	Sialidase-4
nM	Nanometer
NPC1	Niemann-Pick C1 protein
OTAVA	OTAVA Chemical Libraries
P2RX7	P2X purinoceptor 7
PDB	Protein Data Bank
pH	potential of hydrogen
PLI	Protein-Ligand Interaction
PLS	Pairwise Ligand Similarity
PP2AA	Serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform
PPI	Protein-Protein Interaction
PTN1	Tyrosine-protein phosphatase non-receptor type 1
PTN2	Tyrosine-protein phosphatase non-receptor type 2
PTPRA	Receptor-type tyrosine-protein phosphatase alpha

PubChem	Database of chemical molecules and their activities against biological assays
SDF	Structure Data Format
SL	Sphingolipid
SL-WIN	Weighted Identity Network of Sphingolipid Metabolism
SMILES	Simplified Molecular Input Entry Specification
SPHINS	Sphingolipid and Insulin
SPHK1	Sphingosine kinase 1
SPHK2	Sphingosine kinase 2
STS	Steryl-sulfatase
TC	Tanimoto Coefficient
TY3H	Tyrosine 3-monooxygenase
UniProt	Universal Protein Knowledgebase
URL	Uniform Resource Locator
WIN	Weighted Identity Network
XML	Extendable Markup Language
ZINC	Database of commercially available compounds

1. INTRODUCTION

In the field of drug development, protein-ligand interaction based studies are the fundamental approaches to the field. Understanding the protein-ligand interactions of a metabolism highlights the important proteins and their interactors. These proteins or interactors are the good candidates for experimental studies. For a metabolic pathway or a group of proteins, working on the whole metabolism including its proteins and ligands takes too long time for experimental approaches. Thus, computational analysis is the first step to detect target molecules and proteins of metabolisms. Reliance of these computational analyses depends on how well and accurately its data set is prepared.

In this study, we proposed a data collection tool for protein-ligand interactions. Instead of manually collecting data, the programmatic way is more accurate to avoid miss of any data. By this tool, any protein family and their ligands can be extracted from different databases such as PDB [1], BindingDB [2], ChEMBL [3] and UniProt [4]. The output of this tool is also useful for ligand centric network models developed by our lab member, Hakime Ozturk. Thus, this data collection is a complementary of this network construction method. As the second step, the ligands of this protein-ligand interactions data are also analysed to specify which ligands are the key players of these interactions.

The aim of our first task is to construct a useful data collection tool for protein-ligand interactions. For this tool, web services of different databases, such as UniProt, ChEMBL and PDB, are analysed to understand how these services can be used to collect protein-ligand interaction data. For that, Python language is preferred because of its common usage in the solvation of biological problems. In addition, there are many Python packages developed to access biological databases. Based on these

services and usefulness of Python, the language is learned via trying to apply these methods. Then, Python scripts are written for different databases to collect protein-ligand interaction data. Each database has its own specific data types; therefore, new methods are learned to clearly understand the data format of each databases. For instance, PDB database stores its data in XML format which is a commonly used hierarchical tree based representation of data [5]. Therefore, first the data format is learned, then a Python package is applied to extract protein and ligand information from these XML data files. Thus, each database and their data collection scripts are considered separately. For this study, two metabolisms; which are sphingolipids and insulins, are preferred as the case studies. Protein-ligand interaction data of both metabolisms are extracted via these data collection scripts. These metabolisms are preferred due to their dense biological relations. Sphingolipids and insulins are commonly observed at the drug discovery studies because of their impacts on cancer, diabetes and neurodegenerative disorders [6]. Analyses of both metabolisms based on their protein ligand interactions are crucial to highlight key molecules and enzymes which can be crucial to analyse treatments of these diseases. A ligand binding to main proteins of these metabolisms can be the key inhibitor of many enzymes whose deficiency triggers many illnesses. Being able to detect such target enzymes of molecules is the main need of many theoretical and experimental researchers.

As the second task of this study, ligand centric networks of both metabolisms are constructed. important ligand centric networks are analysed together to detect commonly observed interactors of both metabolisms. Analyses of both metabolisms based on their interactors are crucial to highlight key molecules and enzymes of these metabolisms. First, sphingolipid ligand centric networks, then both metabolisms' data are gathered to build ligand centric networks of these metabolisms together. Analysing both metabolisms together eases the finding of shared ligands among both metabolisms as well as the key interactions between sphingolipids and insulins. Inhibitors acting on both metabolisms' enzymes are detected which can be used for development of new inhibitors to treat sphingolipids and insulins related diseases. These ligand centric networks are analysed via CytoScape [7] network visualisation

tool, and CANVAS; a cheminformatic platform, is also preferred for pairwise ligand similarity analysis [8]. According to these analyses, pairwise ligand similarity distributions of ligand sets are obtained.

The following sections of this study include detailed information about the biological and theoretical background of the research, the proposed data collection method, analysis of the results and conclusion. Background information about sphingolipids and insulin metabolisms, protein-ligand interaction networks, protein-ligand databases and their data preparation process are explained in the following sections of the thesis.

2. THEORY

2.1. Sphingolipid and Insulin Metabolisms

Sphingolipid metabolism, which includes many known proteins whose deficiency trigger various diseases such as obesity, depression, insulin resistance, cancer and more [9-12], is one of the important metabolic pathways for the discovery of drug targets. Sphingolipids are the fatty acid derivatives of sphingosine [13] and abundant in the nervous system. They play important roles in membrane reorganization and lipid-protein interaction within cellular membrane. In addition, the metabolic pathway includes many bioactive metabolites regulating the cell functions [13]. Therefore, the analysis of the sphingolipid metabolic pathway is crucial to highlight the targetable proteins for therapeutic approaches. The changes in sphingolipids such as ceramides and glycosphingolipids cause neurodegenerative diseases [14]. Brain aging is an important example to such cases. During the span of human life, structural changes in brain occur in specific regions of brain or in membrane microdomains [14]. These changes only comprise lipids such as sphingomyelin, ceramides and glycosphingolipids. For example, aberrant conversion from sphingomyelin to ceramide triggers Alzheimer's disease and Parkinson's disease [14]. Another type of disease caused by the deficiency of sphingolipids is cancer which occurs when the balance between mitosis and apoptosis is disrupted [15]. The ratio of ceramide/sphingosine-1-phosphate level affects cancer treatments since they mediate antimitogenic responses such as cell differentiation and apoptosis [6]. Thus, understanding the interactions among the sphingolipid proteins and which ligands induce these interactions might provide valuable insights for the development of new strategies for these diseases.

Sphingolipids are also one of the main factors in insulin resistance and diabetes [6,13,15,16]. For insulin signaling pathways, lipid rafts enriched by lipids and cholesterol, are required [17]. Thus, sphingolipids are the key players of the insulin

signaling pathways and affect insulin resistance, since they are located at cell membranes. Thus, the decline or incline in the amount of ceramide has the parallel impacts on insulin resistance [13,16,17]. It was also found that obese patients have increased concentration of ceramide [17]. We can conclude that the number of sphingolipids such as ceramide and sphingomyelin influences the insulin resistance, diabetes and obesities.

2.2. Protein-Ligand Interaction (PLI) Data

Proteins generally have many ligands binding to them. For example, ligands of an enzyme might be its inhibitors, putative drugs or known drugs. For a group of proteins, their protein-ligand interactions are defined based on the ligands binding to the proteins. There can be a variety of protein-ligand interaction types. For instance, PLI data can be prepared depending on the docking site of ligand to the proteins or the types of ligands such as inhibitors, known drugs and so on. In this study, we have extracted proteins and their ligands without having any specification. Thus, this PLI data can be the basis for broad range of studies. By using our data collection tool, proteins and their ligands are collected.

2.2.1. Databases

There are many databases such as PDB, UniProt, BindingDB and ChEMBL for proteins and ligands. Some of these databases are constructed for general purposes such as UniProt which is a universal protein knowledgebase. However, there are also particularly developed databases such as PDB which is the only source comprising structural data of macromolecules [1]. For PLI data collection, UniProt, PDB, ChEMBL, BindingDB, KEGG and PubChem are preferable. Each of these databases and their data types are explained in the coming subsection to clarify readers mind

about specificities of these databases, why they are required and what types of data are needed to be parsed to extract specific protein-ligand information from them.

2.2.1.1. UniProt. UniProt is a universal protein knowledgebase which is the main data source for protein based studies [4]. The database consists of almost all the worldwide protein data with detailed information. For each protein, the database provides biological process, molecular functionality, enzyme & pathway databases, gene information, organism, access to protein-protein interaction databases, sequential information.

2.2.1.2. PDB. UniProt comprises worldwide protein knowledge without having any specificity; however, PDB only includes structurally validated data of macromolecules [1]. Therefore, protein structure based studies usually prefer to work on PDB data. Proteins and their active sites on their structures are also given in PDB. These active sites are the docking sites for ligands. With the active site information, the structural ligand data are also provided with the residues they interact with. However, PDB only includes ligands whose structure is well-defined in 3D form. Therefore, these ligand information is limited to structural ligand data which is not enough for comprehensive protein-ligand interaction studies. For our study, we have worked on general protein information instead of structure based approaches. Moreover, ChEMBL database is preferred as the source database for ligands rather than PDB.

2.2.1.3. ChEMBL. ChEMBL is the globally preferred bioactivity database developed by European Bioinformatics Institute same as UniProt [3]. Therefore, this database is on the ligand side of EBI's bioinformatics studies. Bioactivity information of ligands with their experimentally validated data is accessible from ChEMBL database. In this database, proteins are defined as targets and ligands are named as compounds. ChEMBL lists compounds with their activity concentration (AC) which shows the binding affinity of these ligands onto the proteins such as enzymes. Calculated potency measurements of these activity concentrations are IC₅₀, EC₅₀ and K_i and so on.

Without being stacked into variety of these measurements, all ligands with different binding affinity measurements are extracted to work on. For further specific analysis of small number of ligands, these measurements can be analysed deeply to decide which measurement is preferable to inhibit an enzyme. Another globally preferred bioactivity database is PubChem; however, ChEMBL comprises also all the PubChem data of active ligands with their activity concentration information. Thus, ChEMBL database is the most adequate database to work on protein-ligand interactions and ligand activities.

2.2.1.4. Other Protein-Ligand Databases. KEGG, BindingDB and PubChem are also well-known protein and ligand databases; however, these databases are constructed for specific purposes such as gene database like KEGG [18] or universal compound database like PubChem including bioactivity data [19]. Binding DB also provides binding affinity data for protein-ligand complexes. These data are collected from articles and experimental studies[2]; however, the size of the datasets Binding DB and KEGG have are smaller than the number of ligands ChEMBL database consists. Moreover, ChEMBL database comprises all the PubChem data having activity concentration information. Thus, considering only ChEMBL database is enough for ligand activity based studies.

2.2.2. Programmatic Access to Databases

Rather than manually searching in databases, programmatic access to the databases is more accurate. For that, web services are developed to access a database via web [20]. These services are developed to use via any artificial languages such as Python, Java or Perl. Web services based searching requires specific URLs, web addresses, to directly access the webpage of these databases. URL pathways are generally divided into two or more locations in which user can add input, operation type, data format specification, data domain and output type. After filling the subdivisions of URL paths, data can be parsed from the specified URL pathway. For

this URL based accession to databases, there are many packages like requests, written in Python language. Thus, first users need to import requests package into their scripts to be able to call functions of these package. However, packages should be installed into your computers. All the Python packages can be install via command prompt by just typing pip install 'package name'. After installing packages, their functions can be imported and applied into the working environments such as scripts. The simple function of requests package is called "get" by which data can be called via the provided URLs. The response of the function is returned in XML format which is a commonly preferred file format for data storage and HTML based studies [5]. XML is an extensible markup language in which all data are listed in a tree format starting by a root and having many sub children. As an example, BindingDB and PDB databases return data in XML format for each of their entries. Protein information is the root of the XML file, then its sub children include protein's ligand name, ligand ID, molecule's binding affinity, its full name and its SMILES which is a string representation of molecules. In brief, this XML file consists of all the information of the protein and its ligand information. However, only required protein-ligand information should be parsed from these files. Python has a package called xml.etree to parse a XML file. Thus, by using the functions of the package, protein ID, ligand ID, and ligand SMILES are parsed by iterating over all the children of the XML file. This methodology is applied for data collection process from many databases such as PDB and BindingDB.

2.2.2.1. BioServices. BioServices is a Python package developed for programmatic accession to biological databases via their web services. Each database has its own web services; however, their data type or the storing way of data might be different. For that, functions of this package allow users to search directly in biological databases such as UniProt and PDB without worrying about the data types. For each database, there are sub classes of this python package. Therefore, users can import any provided databases' function class into their scripts; then the functions of BioServices can be performed to search in the databases. BioServices is developed by a research group from European Bioinformatics Institute (EBI) and the package does not include all of

bioinformatics databases. The list of protein and ligand based databases, for which the EBI group developed BioServices, are UniProt, PDB, ChEMBL, KEGG. Except KEGG, all the databases are developed by EBI research groups. That is why BioServices is also specifically a product of EBI. However, some errors have received while trying to apply ChEMBL and KEGG functions of the package. The functions did not extract any data from these databases. Then, these errors were reported to the research group as the issues of their package. As our B plan, these databases were accessed via their own clients and web services rather than BioServices.

2.2.2.2. ChEMBL Webresource Client. Client is a software accessing services made available by servers. For ChEMBL web services, a python package called ChEMBL web resource client is designed by ChEMBL group to provide a programmatic as well as an easy way to users. While working on a set of compounds in a script, functions of ChEMBL client can be called to search for these compounds in ChEMBL. For this study, SMILES, activity concentration, and molecule ChEMBL IDs are collected from ChEMBL via functions of the client. There are many sub classes of these client for each group of specific data set of ChEMBL. The database stores data in dictionaries. Unlike strings indexing its entries by numbers, dictionaries label each entry by keywords which are called when users need to get any specific data from the dictionary. To illustrate, a variable called resource object of the database should be first defined such as target, compound or assay, then via this object in the script, the data can be accessed. For instance, if compound data is searched, then compound variable called compound() should be indicated in the Python script.

2.3. Protein-Protein Interaction (PPI) Networks

Network models provide a practical environment for the analysis of protein interactions, thus leading to emergence of protein-protein interaction (PPI) networks. Among the many benefits of network representation, identification of the central nodes (e.g., proteins) and clusters constitute important steps into the better understanding of

the interactions. For instance, insulin like peptides are the regulators of insulin signaling pathway[21], which is known by its regulatory function in diverse physiological processes in humans. An insulin-based PPI network might provide an information about the significance of peptides as regulators in the insulin metabolism. In a simple way, protein-protein interaction networks can be built with respect to the data worked on. For instance, Durek and Walther constructed [22] metabolic interaction networks (MIN) which can be analyzed by constructing the PPI network of metabolisms in order to understand the changes in the information flow among all the proteins. According to their study, analogy between these PPI networks and metabolic pathways are observed.

PPI networks are not only useful to understand metabolic interactions but also for analyses of different biological problems. For instance, PPI networks can be modelled based on gene neighborhood, protein structure or domain-domain interactions of the proteins and more [23]. The important genes for diseases can be also detected via the analysis of these gene based networks (GBN). For instance, Fenger and Jeppesen [24] used gene-based networks to analyze the importance of sphingolipid metabolism on hypertension. GBNs were constructed to clarify which gene interactions are significant for the regulation of blood pressure.

2.3.1. Ligand Centric Network Model (LCNM)

In this study, Ligand-centric PPI (LPPI) networks are constructed to investigate the interactions between the proteins in terms of the ligands binding to the proteins. In this model, proteins are connected via their shared ligands. Moreover, the edge weights of these networks are set based on the number of shared ligands. There are two main and three sub network types provided by LCNM. These networks are called identity and similarity network models; and their sub types are explained step by step.

2.3.1.1. Identity Network Model. Construction of this network depends on the number of shared identical or chemically ligands between proteins. For that, there are two main network types called identity and similarity. For this study, we preferred to apply identity networks since construction of similarity networks takes too long time for big protein-ligand data sets. Edge weights of identity networks are set in three different types; namely, unweighted, weighted and normalized weighted. Based on to which type of network is required, user can decide for edge weight sets of networks. The definition of these network types is listed below.

Unweighted Identity Network Model: In this model, two proteins are connected by an edge if they share at least an identical ligand. Whether the number of shared ligands is 1 or more than 1, all the edge weights are set to 1. This way of setting briefly indicates which proteins share ligands or not.

Weighted Identity Network Model: Different than unweighted network model, edges of weighted networks are set by the number of shared identical ligands. The edge weight of the network is the number of shared ligands between proteins. In addition to the number, the IDs of these ligands also provided on the network. This information allows us to search for specific protein interactions and their interactors.

Normalized Weighted Identity Model: For normalized weighted identity network, edge weights of weighted identity network are normalized by the total number of ligands binding to proteins. As an example, it is assumed that there are two proteins, A and B, in a weighted identity network; moreover, number of ligands binding to A and B are 14 and 16 respectively. In addition, the number of shared ligands between these proteins is assumed to be 4. Then, the normalization of the edge weight is $4 / (14 + 16 - 4)$ which equals 0.15. This is how the edge weights of the normalized weighted identity network are set.

For this study, we focus on weighted identity networks since the precise number of shared ligands and their IDs are required. Finding the most dominant interactors of a metabolism is the first step to build new molecules for the disease caused by the metabolic disorders. The interactors of the main enzymes in a metabolism are the target molecules to estimate the inhibition level of enzymes. Inhibition level of the main enzymes are the triggers of metabolic disorders and related illnesses. Therefore, each network and their clusters are considered to detect these interactors for further studies such as molecule generation.

2.3.2. Clustering Analysis

For both sphingolipid and insulin metabolisms, their protein interaction networks are constructed as ligand centric network. First, sphingolipid network, then both metabolisms' data are gathered to construct sphingolipid & insulin PPI network. These networks are visualized and analysed in CytoScape which is a well-known network visualization and analysis platform[7]. This software includes many plugins via which network parameters and their clusters can be calculated.

2.3.2.1. MCL Algorithm. For our protein interaction networks, MCL algorithm; provided via ClusterMaker plugin, is applied to find clusters of these networks. There are not many options for the analysis of weighted networks, therefore, well-known MCL algorithm is preferred. MCL is a network clustering algorithm which works on flows of the network [25]. In other words, edge weights of a network are considered for clustering process. In this algorithm, the flows also called edge weights of the input network are considered to build flow matrix of the network. The algorithm is implemented for given number of iterations. For each iteration, first, the matrix is expanded by algebraic matrix multiplication to itself. Then, each non-zero elements of the new matrix are raised to a power which is an input of the algorithm called granularity inflation. Increasing the inflation value causes the emerging of new clusters of the network. Therefore, user can try different inflation values to understand which

approach is more appropriate for its dataset. For the metabolic data sets, 2.5 is chosen as the granularity inflation value. Then the clusters of these metabolic networks are calculated.

2.4. Ligand Analysis

For each clusters of these PPI networks, their ligands are collected from CytoScape. First, each clusters of the networks are extracted to create subnetworks of these networks. For each cluster, then the edge attributes are collected and saved as sdf files. For each cluster, its ligand sets are prepared to analyse. These analyses are performed in CANVAS which is a computational cheminformatics platform [8]. CANVAS requires sdf or mol file formats to import data into the software. Each ligand sets are imported into CANVAS, then their similarity matrices are calculated. For this calculation, the binary fingerprints of these SMILES are required since these matrices are constructed based on the fingerprints[8]. Binary fingerprints are binary string representation of substructures belonging to the analysed molecule. All possible chemical substructures are defined in this CANVAS fingerprint algorithm; moreover, presence and absence of these substructures in the given molecule are indicated with 1s and 0s in its fingerprint. For a list of molecules, first their fingerprints are computed, then pairwise similarity comparisons are performed to construct fingerprints based similarity matrices. The similarity measurement is chosen as Tanimoto coefficient which is calculated by dividing intersection of two vectors to their union or summation [26].

For each cluster and their ligand sets, this process is implemented and their similarity matrices are obtained via CANVAS. These similarity matrices are exported as csv files then these matrices are plotted as similarity distribution via a Python script. This script easily takes the upper diagonal side of any matrix by reading each row of the matrices, then plot the distribution of the entries taken from the upper part of the matrix. These similarity matrices are symmetric matrices since the entries include

pairwise Tanimoto similarity values of each ligands belonging to a ligand set. That is why only the upper part of the matrices are considered. In addition, these matrices are generally too big to analyse directly by looking at them. For instance, the size of the similarity matrix calculated for 5000 ligands is 5000*5000. Therefore, plotting these similarity distributions is the simple but effective way to analyse big data matrices.

2.4.1. Matplotlib

Commonly used Python library for plotting variables; such as arrays, matrices, is Matplotlib. This library is a 2D graphics package and used for image generation for Python scripts [27]. A function of this package called “plt” is used for visualization of variables in a script. For this study, upper triangular similarity matrices are considered; moreover, entries of these upper triangular matrices are stored in arrays. These arrays are plotted by using the plt function of Matplotlib. This function also has many features by which user can select what type of plot model is needed or even the titles and labels of these plots can be added by using the features. For instance, histogram model is preferred for the similarity matrices and this preference is denoted by “plt.hist” in the script. For naming the x and y axes of these plots, “xlabel” and “ylabel” features of plt function are added to the plt function as “plt.xlabel” and “plt.ylabel”.

2.4.2. Protein – Ligand Docking

After analysing protein-ligand interaction networks, some of the specific interactions and interactors are detected. These proteins and ligands are evaluated to find the molecules having impacts on more than one protein. This sort of molecules might inhibit many enzymes in the metabolic network. After detecting the molecules, new drug candidates can be designed.

For a group of detected proteins and ligands, these ligands should be docked into proteins if they can. These docking scores help us to demonstrate which ligands are the efficient candidates to work on. After having such computational cheminformatics approaches, these candidate drugs or well-docked ligands might be considered for further experimental approaches instead of trying to dock the whole list of ligands.

Schrödinger Maestro [28] is commonly preferred cheminformatics platform for protein-ligand interactions. First, proteins are prepared, then their ligands are also prepared to dock them into the proteins. For a group of proteins, the docking process is separately completed. Maestro provides Protein Preparation Wizard for protein preparation, and Lig-Prep for ligand set preparation [29]. First, the protein's crystal structure is imported into the environment by entering its PDB ID in Protein Preparation Wizard. PDB consists multiple structures of proteins, thus first the well-designed form of a protein should be decided. The mostly preferred structures are ligand bonds ones since the proteins' active sides are already provided. The docked sides of the ligands are the possible docking sides to which new ligands are tried to dock. Missing side chains and missing loops are also added to the structures, then they are optimized as well as minimized for pH 7 for natural pH of cytoplasm and pH 5 for natural pH of lysosome. Furthermore, water molecules in the crystal structure were also removed since it eases the complicatedness of docking process. Proteins usually comprise more than one chains in the structures; however, sometimes these chains might be identical. For that, working on only one of them is preferred for preventing waste of time. If the protein working on has identical chains, one of the chains can be erased via Protein Preparation Wizard.

For ligands, their sdf file is imported into Maestro, then these ligands are prepared to dock via Lig-Prep. First, imported ligand set is selected from the entry list of the Maestro project. Then, the possible states of ligands are generated in pH 7 for cytoplasmic and pH 5 for lysosomal proteins. The subcellular locations of the proteins are searched in UniProt. During this process, specified chiral centres of the molecules

are retained. After preparing the ligands, receptor grid generation of GLIDE is performed to detect active sites to which ligands are docked [30]. For that, docking sites of the ligands, provided by protein structures in PDB, are selected to which the list of prepared ligands is docked. From the docking types, XP (Extra Precision) docking is preferred since it offers more extensive ligand-receptor shape based docking process. According to the outputs of docking process, top 5 ligands are considered based on their XP Gscores. Lowest XP Gscores are the best docking scores since well-docked ligands require lower energies.

2.4.3. Molecular Symmetry and Chemical Features

Geometric properties of molecules determine the molecules' chemical features such as chirality and polarity. Non-superimposable molecules; whose mirror image cannot be superimposed to itself, are defined as chiral molecules. Mirror images of the molecules, also called stereoisomers, may react differently in chemical processes [31]. Thus, chiral molecules are generally analysed by their two isomers to evaluate which isomer of the ligand plays role in the chemical reaction. Moreover, natural products are commonly chiral. That is why drug developer and synthesisers prefer chiral molecules since synthesizing isomers of chiral molecules is less expensive [31]. For the detection of geometric features of ligands, group theory is commonly preferred by computational chemist to detect molecular symmetries using symmetry groups [32].

In mathematics, group is an algebraic structure with a non-empty set. The elements of this set are equipped with an operation which combines any two elements to form a third element. In addition, the set and the operation satisfies the four axioms of group theory: closure, associativity, identity and invertibility. For symmetry groups, we have symmetry operations; such as identity, reflection and rotation, as the elements of the set. Moreover, any two symmetry operations are combined to obtain a third symmetry operation. For instance, rotation-reflection symmetry is obtained by

combining both rotation and reflection symmetries. These symmetries are comprised by symmetry group. For objects such as molecules, these symmetries map the objects to themselves by rearranging their vertices. In addition, all possible symmetry groups are well-defined in group theory. The list of symmetry operations is given in Table 2.1 [33]. The list of used symmetry groups and their operations are also given in Table 2.2.

Table 2.1. The list of symmetry operations.

Symbol	Operation
E	Identity operation (do nothing)
Σ	reflection through a mirror plane
C_n	rotation around n-fold axis
I	inversion through a centre of symmetry
S_n	rotation around n-fold axis + reflection in a plane perpendicular to axis of rotation

Table 2.2. The list of used symmetry groups and their operations. Here, n is an ordinal number.

Symbol	Symmetry Operation
C_1	E
C_n	E, C_n
C_s	E, σ
D_n	E, C_n , $n \cdot C_2$

Symmetry groups are also classified based on chemical features. For example, C_1 , C_n , and D_n symmetry groups are called chiral groups. Thus, if symmetry group of a molecule is one of these groups, then the molecule is considered as chiral [33]. Similar to chiral groups, polar groups are also determined: C_1 , C_n and C_s . Polar molecules are water soluble and non-polar molecules are fat soluble molecules. This information is important at the synthesis of the molecules.

In molecular symmetry, symmetries of molecules are calculated to define the molecular symmetry with respect to symmetry groups. The Java tool called Jmol allows users to calculate the molecular symmetry based on SMILES [34]. First, Jmol jar file is installed from its webpage. For symmetry group calculation, java scripts are written in which the list of SMILES are read and their symmetry groups are calculated via functions of Jmol. Then, these java scripts and Jmol jar file are called via command prompt. This command prompt based usage of Jmol and the scripts are given as the supplementary files.

3. METHODS

3.1. Data Collection

The binding sites of ligands and their pharmacological effects are important for protein activity [35]. According to the impacts, the possible pharmaceutical drug compounds can be found. However, the first and the fundamental step is to build or collect accurate data to work on. For that, there are web services to collect data from databases such as APIs and clients written in languages such as Python, Java and R. These web services are developed for data collection from these databases; however, the important need is to collect specific data belonging to a protein family from a variety of databases at a single time. For that, there are some tools such as PLI [36] and LigDig [37] which are both designed for different aims. PLI detects the binding pockets of a protein for a ligand extracted from PDB. Different than PLI, LigDig analyze the protein-ligand interaction from different points such as searching for inhibitors of a protein, finding 3D structures of protein-ligand interactions. The former works only on PDB data and the latter analyze a protein or a list of proteins. In this study, the main proposed method is the protein-ligand interaction data collector which collect protein-ligand interaction data from public databases ChEMBL [3], UniProt [4], PDB [1] and BindingDB [2]. Manual collection of protein-ligand interaction data from several different databases is an inefficient and unpractical task due to the possible redundancy/miss cases and continuous updates. Therefore, with this data collector, we provide obtaining protein-ligand interaction (PLI) data which is accelerated and simplified with the promise of an up-to-date data. The data collector is written in Python.

For proteins, UniProt database is preferred since it comprises all the universal protein knowledge as can be understood from its name “Universal Protein

Knowledgebase”. PDB is also another option but it only includes the proteins having available 3D shape and structural knowledge. Thus, UniProt is preferred as the source database for protein information.

ChEMBL, PDB, PubChem and BindingDB are commonly used source databases for ligand data. First, all the databases are compared to detect the most dominant database for ligands. The numbers of ligands these databases comprise are listed below in the Table 3.1.

Table 3.1. The number of ligands each database includes.

Database	Number of Ligands
PDB	22.876
ChEMBL	2.036.512
BindingDB	520.000

As a case study, different protein families, such as sphingolipids, and their ligands are searched in these databases. Then, the results are compared depending on the intersection of these databases. For instance, all the sphingolipids’ ligands extracted from PDB and BindingDB are also consisted in ChEMBL database.

Another disadvantage of BindingDB is the ID mapping from BindingDB ID to other databases. BindingDB IDs and their equivalent PubChem IDs are listed as a supplementary file of this database which can be obtained by following the way in its webpage: Download – All Compounds and Data – Lists and Identifier Mappings – BindingDB_CID. CID is the ID type used for compounds in PubChem. However, this file; we call monomer file, is not frequently updated. The updating the BindingDB ID editing for new molecules takes 2 weeks; then adding its equivalent PubChem ID into

the monomer file takes also 2 weeks more; since the updating process is completed each month for total number of new entries. Thus, BindingDB is not a recommendable source database for ligand studies to build up to date ligand datasets. The main importance of such data collection tools is to easily obtain updated data sets from databases.

For this ligand centric study, protein data are collected from UniProt, then ligands are collected from ChEMBL for each UniProt IDs as illustrated in Figure 3.2. However, the scripts prepared for the other databases BindingDB and PDB are also proposed. Studies may require ligand data which are specifically collected from BindingDB and PDB databases. The data collection scripts of these databases are demonstrated in the following section.

3.2. Database Based Protein-Ligand Interaction Data Collectors

In this section, the scripts, which are developed to extract ligand data from different databases, are clearly demonstrated. The main difference for these scripts is the ligand database. For these scripts, UniProt database is preferred as the source protein database. The searched ligand databases are ChEMBL, PDB and BindingDB. For each database, separate scripts are written; therefore, the flowchart of each script is clearly described in the following subsections.

3.2.1. ChEMBL

First, some python packages; namely BioServices and chembl-webresource-client, are imported into the script. Then the resource objects are described via which the data are searched in both UniProt and ChEMBL. On other words, resource objects are just variables of a script. The resource objects of ChEMBL script are TargetResource() and UniProt(). TargetResource() is implemented to search for target

data; called also protein data, in ChEMBL. On the other hand, UniProt () is used for searching the keywords defined for the searched protein data. These keywords are searched UniProt via the UniProt () resource object to collect list of related proteins and their IDs.

First, the keywords are searched in UniProt to collect related protein IDs, then these IDs are input to search for their ligands in ChEMBL. For the UniProt search, data type and some information should be specified first. More than just searching the keywords, organism id, data format and its columns should be selected. Each entry of UniProt includes many information; however, relevant data should be only parsed from the database. For this study, human organism is searched by its taxonomy ID 9606, the data format is chosen as “tab” format, and the selected data columns are “id, entry name, database(chembl)”. Thus, UniProt ID, short name, and their equivalent ChEMBL IDs are parsed. ChEMBL target IDs are required for checking whether proteins have compound data in ChEMBL or not. All target data do not have to have compound information. For some of target data, there is no available compound data since the compound data depend on experimental studies.

After collecting UniProt ID, protein short name and their ChEMBL target ID, these data are appended into separated arrays from which the required information can be used. During these classification process, ChEMBL target IDs are searched in ChEMBL to check whether there are compound data or not. Via the TargetResource() object, each target ID is checked; then if a target ID has any compound data, its equivalent UniProt ID is stored in an array called “input”. This array is the input of ligand data collection from ChEMBL.

For each entry of the array, first their target data are obtained from ChEMBL via TargetResource() object and the filter functions of the ChEMBL client. There are two functions called “target.filter” and “activity.filter”. The former filters search UniProt ID; and the latter filters activity data of these targets. Both filters are performed in

ChEMBL database. The returns of this activity filtering are in dictionary format. Thus, instead of considering number as the index of the array type, each activity information such as standard unit type, activity concentration and their SMILES, are stored by their short names in this array. Thus, each specific data can be called via their short names by indicating the name into the index brackets of the array. To illustrate, it is assumed that there are activity arrays from which ligand ChEMBL IDs and ligands' SMILES are required to be parsed. Thus, printing `activity['smiles']` and `activity['molecule_chembl_id']` elements of these arrays returns us the ChEMBL ID of the molecule and its SMILES. For our study, we parsed ChEMBL IDs and their SMILES from the activity dictionaries. The data collection process is summarized by the flowchart given in Figure 3.2. In addition, its pseudo code is also given in Figure 3.1.

```

INPUT: Protein's keywords and organism ID

OUTPUT: Proteins, their ligands with ChEMBL IDs, and ligands' SMILES

/* First UniProt IDs and ChEMBL IDs are extracted via BioServices by searching the given
keywords. Then these UniProt IDs and ChEMBL IDs are stored in arrays called uni and chem.
Another array is also created in which the UniProt IDs having compound data in ChEMBL are
stored. */

for list of proteins in chem array do:

    if ChEMBL ID has compound data in ChEMBL:

        append its UniProt ID into the input array

for UniProt IDs in input array do:

    search the UniProt ID's activity in ChEMBL

    parse the ChEMBL ID and SMILES of the protein's ligands

    print UniProt ID, ChEMBL ID belonging to ligands, SMILES

```

Figure 3.1. Pseudo code of ChEMBL data collection script

There is also a tricky approach to extract significant ligands from ChEMBL. The database provides activity concentration of ligands since it is a bioactivity dataset. Thus, defining a threshold for these ligand collection process reduces the size of ligand data set. This reduction allows us to work on more effective ligand data set to detect best ligands of the proteins. Regarding to the leader of ChEMBL research group who is Anne Hersey, 1000 nM can be selected as the cut-off value for the activity concentrations; however, considering the concentration distribution for each ligand sets is more reliable since the activity information can be varied for different protein families and organisms. For that, two options are offered to the users. First, concentration values of search ligands can be stored to calculate their median, then this median can be used as the threshold for the ligand collection process. However, if users would like to see the frequency distribution of these values, it is offered a Python script collecting the concentration data of the ligands to plot their distribution. Thus, after this analysis of the distribution, user can decide for the threshold value by plotting the frequency distribution of the standard values. These three options for threshold defining are summarized in Figure 3.4. If users would like to collect ligand data with respect to a threshold of the data, they can apply threshold based modified ChEMBL script whose pseudo code is given in Figure 3.3. This script is more preferable since threshold data are more reliable. Users can also modify the script based on their needs. The threshold can be chosen as 1000 nM or the median of activity concentration values. For calculating median of a set, numpy package of Python and its functions are used. User can choose threshold from the script. If they would like to plot the activity concentration values as the frequency distribution of these values, it is also proposed another script for that purpose. For all the threshold based client, the output of the script is the same. The output includes protein ID (UniProt ID), ligand ID (ChEMBL ID) and its SMILES. This format is also the input format of ligand centric network model. Thus, this output can be also used for ligand centric network construction of these interactions. After selecting the script, they prefer, the ligand centric networks can be built.

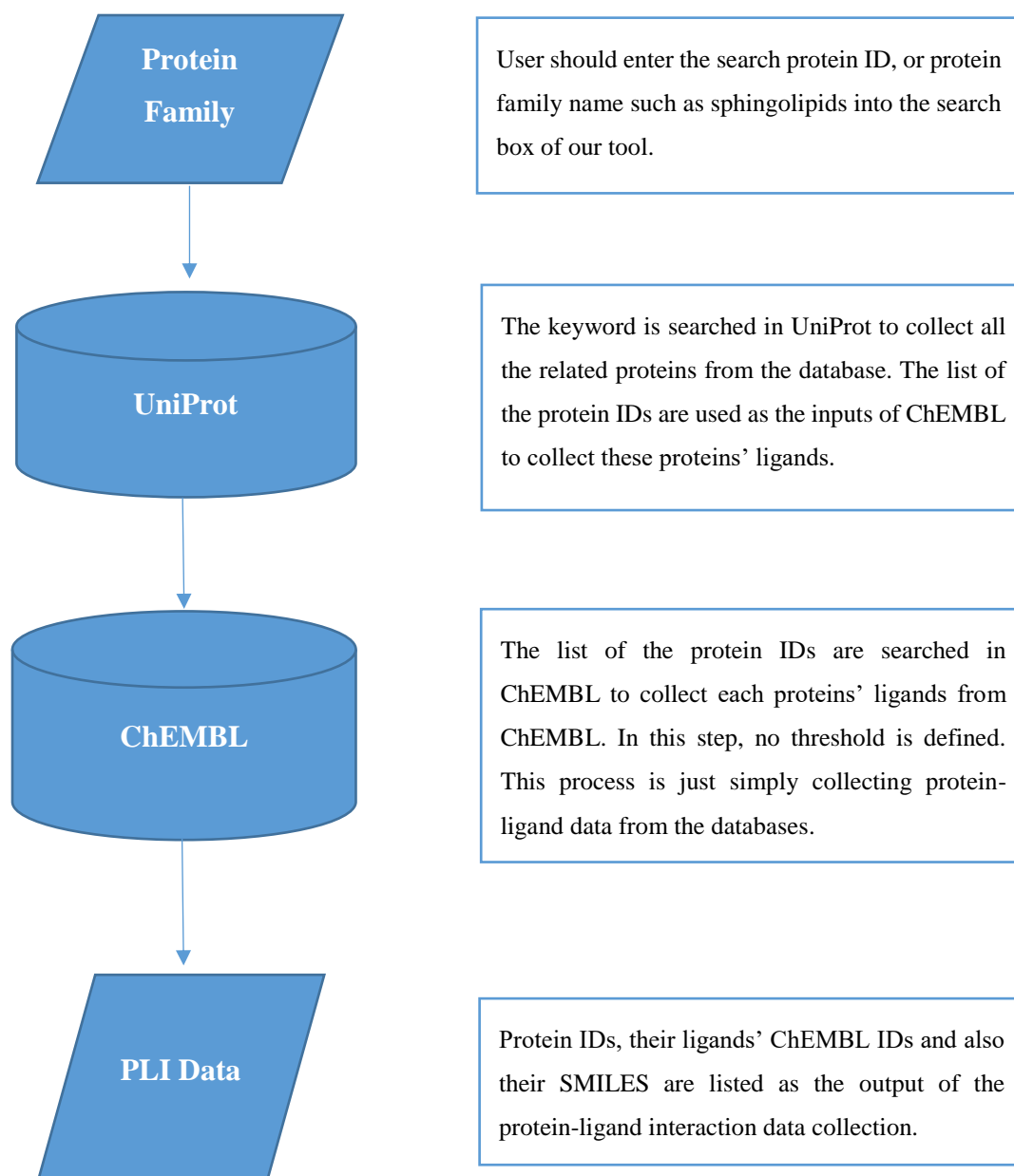


Figure 3.2. The flowchart of protein-ligand interaction process

```

INPUT: Protein's keywords and organism ID

OUTPUT: Proteins, their ligands with ChEMBL IDs, and ligands' SMILES

/* First UniProt IDs and ChEMBL IDs are extracted via BioServices by searching the
given keywords. Then these UniProt IDs and ChEMBL IDs are stored in arrays called
uni and chem. Another array is also created in which the UniProt IDs having compound
data in ChEMBL are stored. */

for list of proteins in chem array do:

    if ChEMBL ID has compound data in ChEMBL:

        append its UniProt ID into the input array

/* Array called active is defined to store activity concentration values of ligands. Then,
according to the preferred threshold*/

for UniProt IDs in input array do:

    search the UniProt ID's activity in ChEMBL

    parse & append activity concentration of ligands into active array

/* The median of the active array values is calculated, then the median is signed as the
threshold. Users may use median or default value (1000nM) as the threshold of the
ligands. */

for UniProt IDs in input array do:

    search the UniProt ID's activity in ChEMBL

    parse activity concentration of ligands

    if the activity value <= threshold:

        parse the ChEMBL ID and SMILES of the protein's ligands

        print UniProt ID, ChEMBL ID belonging to ligands, SMILES

```

Figure 3.3. Pseudo code of ChEMBL data collection script based on threshold criteria

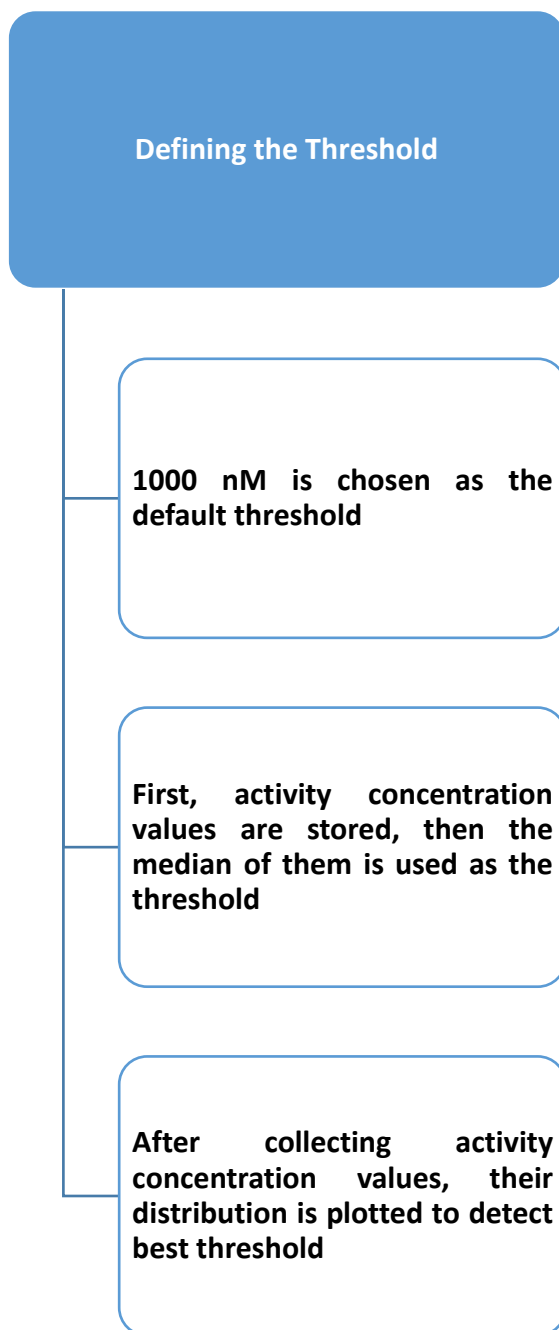


Figure 3.4. The options for threshold selection.

3.2.2. PDB

Same as the ChEMBL script, first of all the Python packages; which are bioservices and xml.etree, are imported to the script. Then, PDB based data collector first searches protein keywords in UniProt like ChEMBL script. The difference between these scripts is the data columns of the protein data. For these script, only PDB IDs of the proteins are collected from UniProt as indicated by the column “database(pdb)”. In addition to the UniProt () resource object, PDB() object is also added. The latter is required for the data search in PDB database. In this script, the keywords are searched in UniProt, then if there is any protein data for them, the proteins’ PDB IDs are only collected from the database. These PDB IDs are stored in an array called input to use them as the input of ligand data search in PDB. For each entry of the input array, their ligands are searched in PDB. For that, xml.etree package is imported to the environment since PDB returns data in XML format. For each entry of the input data, their ligands are searched by get_ligands function of BioServices. This function returns XML file of the protein data. Each XML file includes all the related information about the search proteins. This information also consists proteins’ ligand information including ligands’ chemical names, their formulas and SMILES. For this study, ligand IDs and their SMILES are parsed from the XML files of proteins; then protein IDs, ligand IDs and SMILES are returned. Different than ChEMBL, PDB data does not provide activity concentration value of the ligands. Thus, thresholds are not considered for the data collection process. This is also another reason why PDB is not preferable for bioactivity data analysis of ligands. The script for this database is easier than ChEMBL script since this database does not include any bioactivity data. The scripts’ pseudocode is given in Figure 3.5 and each step of the data collection is explained step by step. This script is the simplest one for the protein-ligand data collection and also return less amount of data than the other databases since the amount of data PDB has is lower than the others.

```

INPUT: Protein's keywords and organism ID

OUTPUT: Proteins, their ligands with PDB IDs, and ligands' SMILES

/* First PDB IDs are extracted from UniProt via BioServices by searching the given keywords.
Then these PDB IDs are stored in the array called input. */

for list of proteins in input array do:

    search the PDB ID in PDB to collect their protein data

        parse the ligands' PDB IDs and their SMILES

            print UniProt ID, ChEMBL ID belonging to ligands, SMILES

```

Figure 3.5. Pseudo code of PDB data collection script.

3.2.3. BindingDB

BindingDB provides binding affinity data of protein-ligand complexes; which are used as the bases of drug development studies.[2] . As mentioned in the earlier sections, this database has smaller sized data sets when compared with ChEMBL and PubChem. Moreover, it is not frequently updated. Thus, this updating problem cause mapping error when the molecules are collected from Binding DB. Each molecule should have a commonly used ID type to which all the data can be mapped to as the unique ID type of the data set. The Binding DB script is also provided; which may be required by some researchers specifically. For this script, first identifier mapping list should be obtained from Binding DB webpage. Then this list is read line by line to append BindingDB IDs and their equivalent PubChem IDs into two arrays called monomer and CID. First, Python packages which are xml.etree and requests are imported into the script, then the BindingDB IDs called monomer IDs and their equivalent are stored in two separate arrays. These arrays are used for the ID mapping in the script. Like ChEMBL and PDB scripts, first the related protein IDs are collected from UniProt by searching the protein families' keywords. Then, ligand data are

collected from BindingDB for these protein IDs. For that, URL based web services are preferred. Via using the requests package, data are collected from the defined URLs. For BindingDB, its URL format and the information are given at its webpage under the Web Services section of its webpage. The web service called getLigandsByUniprot is used to collect ligand data from BindingDB by searching the UniProt IDs in the database. For each protein ID, the URL format is deformed by changing the protein ID entry of the URL. Then, the ligand ID of the protein is obtained via the protein's URL and the data are returned as XML files. Via the functions of xml.etree package, the specific ligand information, ligand's monomer ID and its SMILES, are parsed. For each monomer ID, its index in monomer array is found, then its equivalent CID is called by using the index number for CID array. Each monomer ID and their equivalent CIDs are appended into separate arrays but their index numbers are equal. That is how the monomer IDs of ligands are mapped to CIDs. Then, the UniProt ID, CID, and SMILES are printed. The script's pseudocode is given in Figure 3.6.

```

INPUT: Protein's keywords and organism ID

OUTPUT: Proteins, their ligands with PubChem IDs (CID), and ligands' SMILES

/* First UniProt IDs are extracted via BioServices by searching the given keywords. These
UniProt IDs are stored in input array. Then, each of these IDs is searched in BindingDB via its
URL based web services. */

for list of proteins in input array do:

    search the UniProt ID in BindingDB to collect their ligand data

        parse the ligands' monomer ID and their SMILES

            map the monomer ID to CID

                print UniProt ID, ligands' CID, SMILES

```

Figure 3.6. Pseudo code of BindingDB data collection script

3.2.4. KEGG

KEGG is the knowledge base for gene functions, genomic interactions and linking genomic information with higher order functional information like pathways, compounds and ligands information [18]. KEGG database is preferable especially for metabolic pathway analysis. One can obtain the specific enzymes and compounds of a pathway via KEGG database. However, the number of compounds KEGG provides is lesser than the amount of ChEMBL data. For specific purposes, KEGG database can be required as the source database. Thus, KEGG data collector script is also coded. Similar to BindingDB script, first the python packages called xml.etree and requests are imported. Then, the URL based web service of KEGG is added to the script. KEGG provides a variety of URL formats to search specific data types. KEGG has many sub datasets such as GENES, LIGAND, and PATHWAY. In this script, first, keywords are searched via the pathway based URL to collect related pathway information. The data are collected via requests package's functions. As the return of this pathway search, the related pathways' names and their IDs are listed. These pathway IDs are parsed, then these pathway IDs are searched in KEGG COMPOUND dataset by using the COMPOUND based URL format. For each pathway ID, its compounds are collected by requesting the data via the defined URL. These compound data include compound name, its pathway, its formula, its molecular weight, related diseases, and its equivalent IDs in databases such as ChEBI and PubChem. For this study, PubChem IDs of the compounds are parsed, then these PubChem IDs called CIDs are searched in PubChem database to obtain their SMILES. Similar to KEGG URL formats, PubChem also has URL based web services. For that, PubChem's URL format is defined, then the PubChem compound IDs are searched in compound dataset of PubChem to collect the compounds' SMILES information. The return of this script is the pathway IDs, its compounds' IDs and SMILES. Thus, different than the previous scripts, this script returns pathway IDs rather than protein IDs. The script is summarized in Figure 3.7.

```

INPUT: Pathway keywords

OUTPUT: Pathway IDs, their ligands with PubChem IDs (CID), and ligands'
SMILES

/* First Pathway IDs are extracted via KEGG's URL web service by searching the
given keywords. Then, these Pathway IDs are searched in KEGG COMPOUND data
set via its URL based web services */

for list of pathway IDs collected from KEGG PATHWAY do:

    search and collect the PATHWAY ID in KEGG COMPOUND to collect their
ligand IDs

        for each ligand ID do:

            get compound data to parse its PubChem IDs

                for each PubChem ID do:

                    get ligand's SMILES from PubChem

        print Pathway ID, ligands' CID, SMILES

```

Figure 3.7. Pseudo code of KEGG data collection script

3.2.5. PubChem

ChEMBL consist only activity data of an assay which is a data type for experimental studies different then PubChem's assay data types. PubChem includes inactive, active, unspecified or inconclusive data of an assay. ChEMBL also comprises PubChem's active data rather than the inactive and inconclusive data of PubChem. The mentioned activity information dedicates the binding activity of ligands. Thus, ChEMBL is appropriate data source for protein-ligand interactions; however, users may require all the data of an experimental study without having any restrictions based on ligand activities. For that, PubChem data collector script is coded. This script also includes requests package to access the database via its URLs. Like KEGG and

BindingDB, PubChem also has URL based web services [38]. By using the web services, gene information, assay data, and compounds' SMILES are collected from PubChem database. First, Python packages called `entrez`, `requests`, and `xml.etree` are imported into the environment. `Entrez` package provides codes to access National Center for Biotechnology Information (NCBI) which consists many databases [39]. For this study, PubChem Bioassay, called `pcassay`, database is preferred to collect assay IDs by searching protein keywords in the database. This database short name, protein keyword and organism ID are indicated in the search function of `entrez`. After obtaining the list of assay IDs via `entrez`'s search function, these IDs are used as the inputs to obtain each assays' detailed information such as target protein name, compound IDs, gene IDs, activity concentration values and so on. For that, PubChem's URL based we service is used. Its URL path is modified specifically for assay ID searching in PubChem. From these data, target protein names, and compound IDs are parsed. However, PubChem sometimes does not provide the target protein of molecules. Thus, target protein name is parsed if it is provided. After that, the SMILES of these compounds are collected again via URL based web service of PubChem. Moreover, URL path of PubChem is particularly modified for SMILES data extraction. By collecting only active data for ligands, ChEMBL data are collected from PubChem as explained earlier. However, without specifying the activity data type, all the compounds are extracted from PubChem. The output of this script includes target protein name, compounds' CID, and SMILES. These target protein names are general names of proteins without given any further detailed information about the protein types. For instance, Arylsulfatase types like arylsulfatase A and B are only named as Arylsulfatase under the target protein name. Thus, PubChem provides general information about the proteins. Instead of this general protein information, specific enzyme names and their detailed information are required for the protein interactions. Thus, PubChem data are not preferable for protein interactions and their network construction. The brief explanation of PubChem script is given in Figure 3.8.

```
INPUT: Protein keywords and organism ID  
OUTPUT: Target protein ID, their ligands with PubChem IDs (CID), and  
ligands' SMILES  
/* First keywords and organism ID are searched in PubChem Bioassay via  
entrez package. Then, the related assay IDs are obtained. */  
for list of assay IDs do:  
    search the detailed information about the assays  
    if target protein ID is given do:  
        parse target protein IDs and compound IDs  
        for each compound ID do:  
            get SMILES of the compound  
            print target protein ID, compound ID, and SMILES
```

Figure 3.8. Pseudo code of PubChem data collection script

4. RESULTS

4.1. Sphingolipid and Insulin Data Summary

As the case studies of our data collection tool, sphingolipid and insulin metabolisms' data were collected (December/ 2016). For sphingolipid, “sphingolipid”, “sphingomyelin”, “glycosphingolipid” and sphingolipid metabolic process (GO ID :0006665) were searched for only humans with organism number 9606. For insulin, only two GO IDs were searched which were insulin metabolic process (GO: 1901142) and insulin receptor signalling pathway (GO: 0008286) were searched for only humans with organism number 9606. The number of proteins and ligands extracted for both metabolisms are given in the Table 4.1. However, not all the proteins have ligand data in ChEMBL. As explained in methods section, first, the proteins were collected from UniProt via functions of BioServices. However, the UniProt IDs having compound data in ChEMBL were detected to collect their ligand information from ChEMBL. The numbers of the proteins having compound data in ChEMBL are also given in the Table 4.1.

Table 4.1. The number of protein IDs and ligands extracted for sphingolipid and insulin metabolism from UniProt and ChEMBL.

Protein Family	Number of Proteins	Number of Proteins Having ChEMBL Compound Data	Number of Ligands
Sphingolipid	294	51	84397
Insulin	101	20	44419
Sphingolipid & Insulin	395	71	128816

4.2. Sphingolipid and Insulin Ligand Centric Networks

4.2.1. Construction of SL-WIN Network

The ligand centric networks of sphingolipids were constructed by using the data obtained from ChEMBL data collection script. For that, jar file of ligand centric network models was used via command prompt. For this, the input of the network jar file was the sphingolipid data file. The network model provides three types of networks; namely unweighted, weighted and normalized weighted. Weighted identity network (WIN) was chosen to construct sphingolipids networks to be able to get ligand information of protein pairs. The same process was applied to combined sphingolipid & insulin data. This SPHINS network was prepared by gathering sphingolipid and insulin data collected via ChEMBL data collector. Same as sphingolipids, both metabolisms' weighted identity networks were constructed. The aim of gathering both metabolisms was to understand their interactions as well as their interactors. The networks' visualization and analysis were completed in CytoScape version 3.4.0 [7].

First, the sphingolipids' weighted identity network was built as illustrated in Figure 4.1. The number of shared ligand between protein pairs was set as edge weight. The clusters of the network were calculated by MCL algorithm and biologically evaluated to highlight the crucial interactors (Figure 4.2). For MCL algorithm, the inflation value was chosen as 2.5; after an optimization process. Different inflation values were tried ranged from 1.5 to 4. The results of each inflation values were considered based on the biological meaning of these clusters. After this optimization process, the output of 2.5 inflation value returned biologically more relevant sphingolipid clusters.

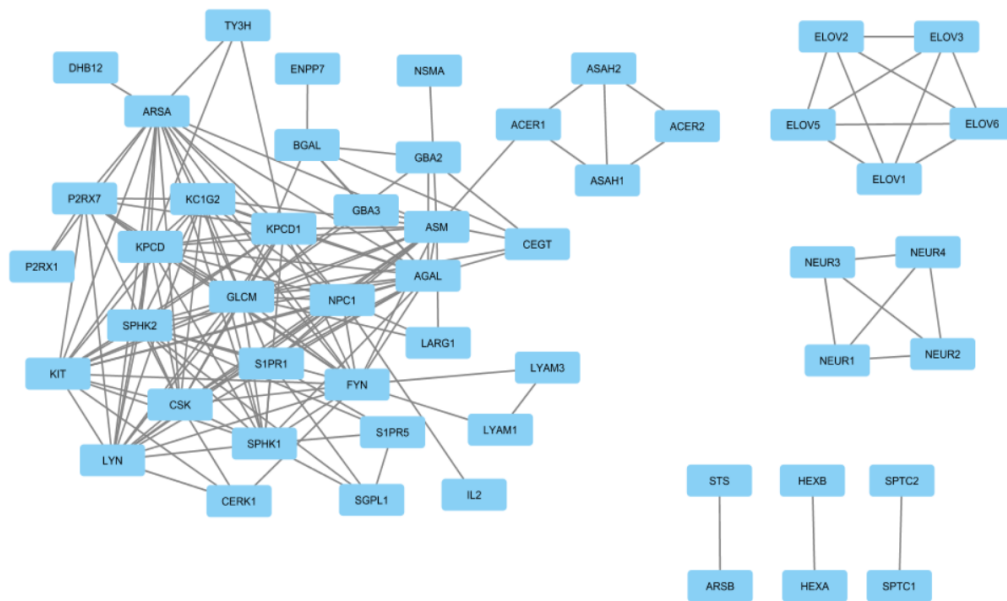


Figure 4.1. The WIN of sphingolipid metabolism

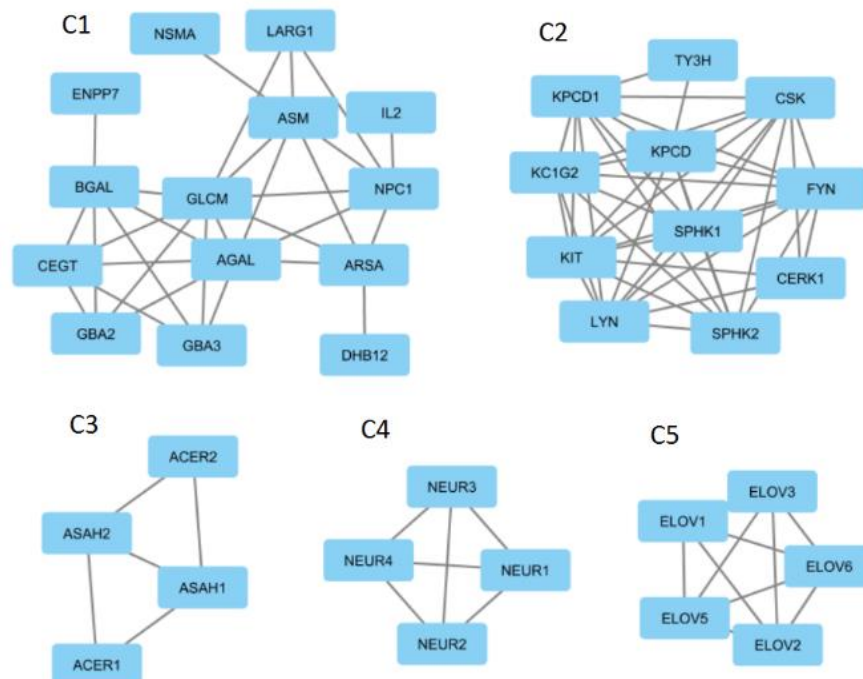


Figure 4.2. The clusters of the SL-WIN

4.2.1.1. Interactor Analysis of Sphingolipid Network. For each cluster of sphingolipids, the shared ligands of proteins given as the edge attributes were collected from CytoScape. For these interactors, their ChEMBL IDs were only given. For the list of ChEMBL IDs, their SMILES were collected from ChEMBL by using its webresource client. This process was iterated for each cluster to obtain their ligands' ChEMBL IDs and SMILES. These ligand sets of clusters were saved as sdf files which is a suitable file format to import sets into CANVAS. For each ligand set, fingerprint calculations of molecules and their similarities were analysed in CANVAS.

First, the binary fingerprints of these ligands were calculated, then their Tanimoto coefficient based similarities were also calculated. Their similarities were output as similarity matrices. These similarity matrices were plotted as the similarity distributions. These similarity plots of the clusters are given in Figure 4.4. The frequency distribution of proteins having number of shared ligands equal or higher than 50 is also given in Figure 4.3.

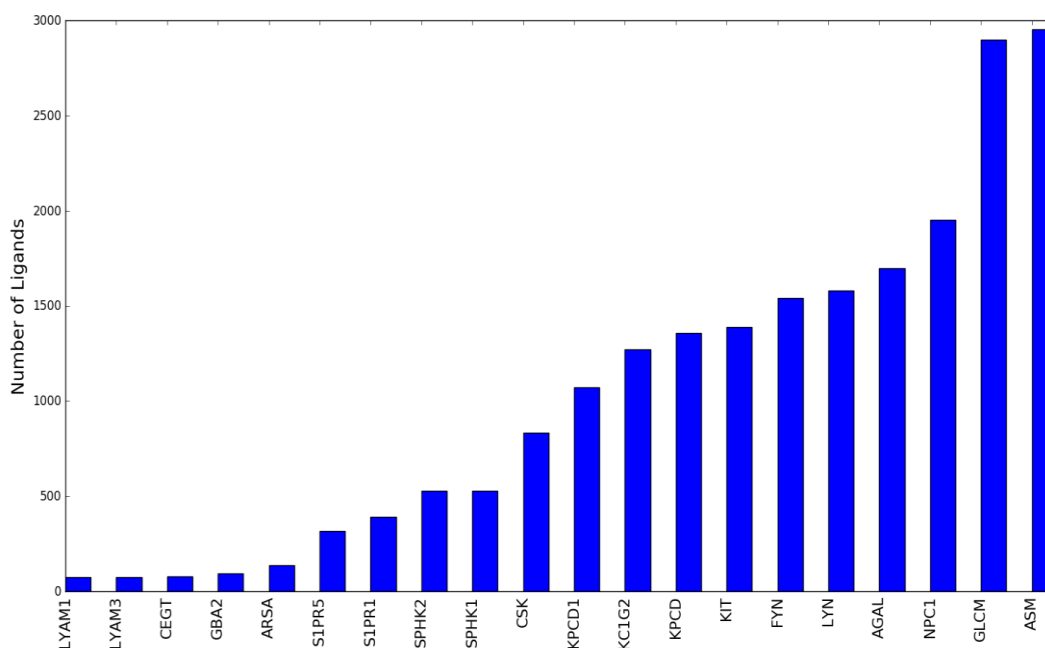


Figure 4.3. Number of shared ligands higher than 50.

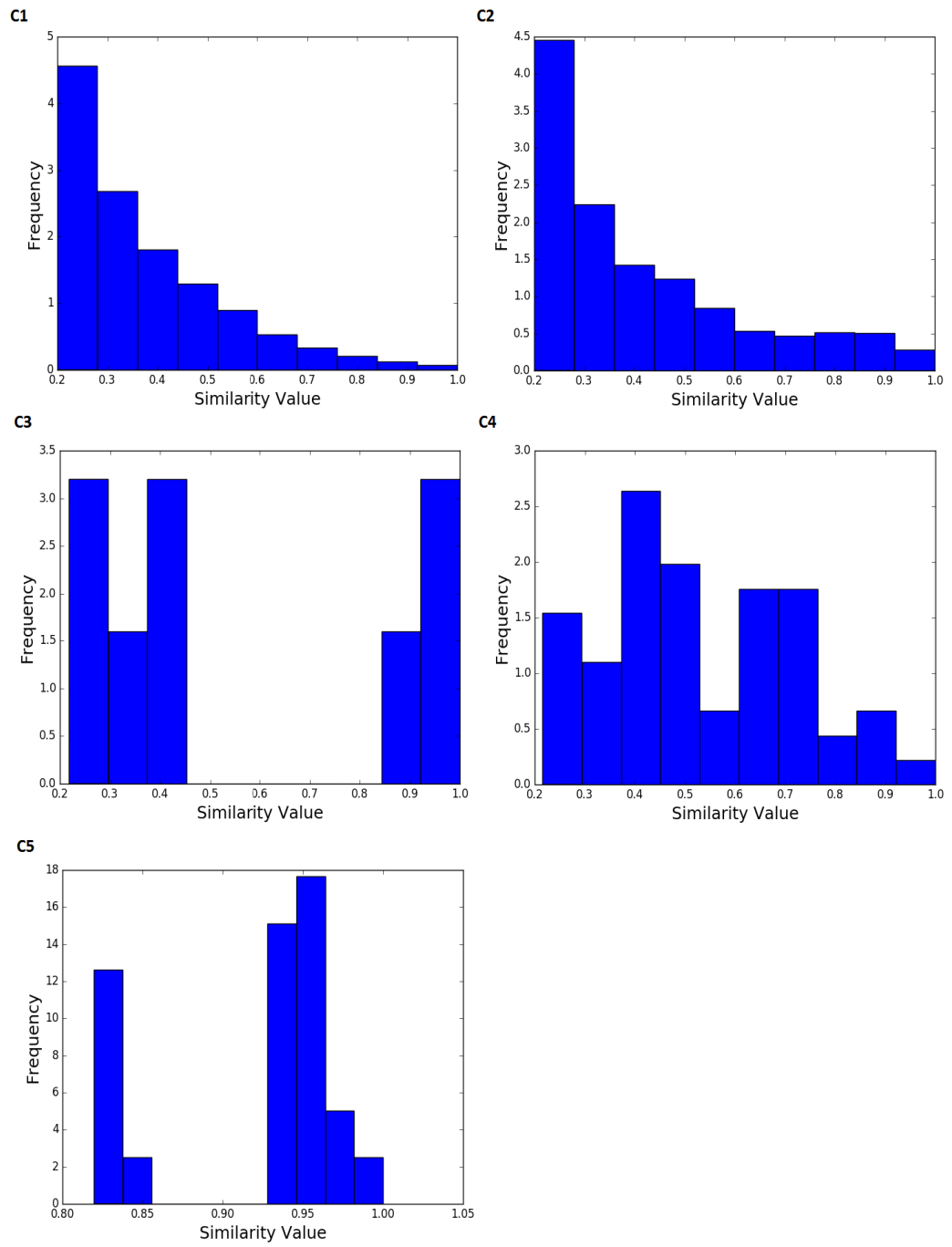


Figure 4.4. PLS distribution of SL-WIN clusters.

Sphingolipid weighted identity network (SL-WIN) has five clusters, that are dominated by some protein groups namely glycosidases, kinases, ELOV proteins, ceramidases and sialidases (Figure 4.2). The first cluster of SL-WIN (14 proteins) consists of glycosidases, namely alpha galactosidase (AGAL), beta galactosidase (BGAL), non-lysosomal glucosylceramidase (GBA2) and sphingomyelin phosphodiesterase (ASM) [40]. The second cluster of SL-WIN (11 proteins) includes all the kinases of the metabolic pathway: sphingosine kinases, protein kinases and tyrosine kinases as well as tyrosine 3-hydroxylase (TY3H). These clusters include 4526 and 2037 ligands; respectively with average pairwise ligand similarity (PLS) score around 0.035. In the first cluster, the ligand pairs with similarity above 0.7 includes drugs namely Miglustat, Prazosin, Amiloride and Terazosin. Each drug was paired with their derivatives and targeted the same proteins in the cluster. Miglustat and its derivatives (CHEMBL408500, CHEMBL1086997, CHEMBL1076754) targets CEGT, GBA2 and GLCM. The other drugs and their derivatives (CHEMBL1558 for Prazosin, CHEMBL540851 and CHEMBL1398126 for Amiloride, and CHEMBL1256665, CHEMBL1554413, and CHEMBL1201091 for Terazosin) interacted with GLCM, ARSA, ASM and AGAL. In the second cluster, the ligand pairs with similarity above 0.7 includes Sunitinib and Lapatinib cancer drugs. These drugs were paired with their derivatives (CHEMBL13485 and CHEMBL1721885 for Sunitinib, CHEMBL212250 and CHEMBL215814 for Lapatinib) and targeted almost all the proteins in the cluster, except TY3H and CERK1. The remaining clusters (C3- 4 proteins, C4- 4 proteins, and C5- 5 proteins) include specific groups of proteins namely elongation of very long chain fatty acids proteins (ELOV), ceramidases (ACER1, ACER2, ASAH1, ASAH2) and sialidases (NEUR1, NEUR2, NEUR3, NEUR4) respectively. These clusters comprise 11, 9, and 18 ligands and their average pairwise ligand similarities (APLSs) are 0.31389, 0.22181 and 0.38108. Only sialidase cluster includes a drug called Zanamivir used for treatment of influenza [41]. The ligand pairs with similarity above 0.7 includes Zanamivir and its derivative CHEMBL96712. Zanamivir target only NEUR2 and NEUR3; however, its derivative targets also NEUR4. The pairwise ligand similarity distribution of the SL-WIN clusters are also illustrated in Figure 4.4.

Based on the results, the ligand sets of the first two clusters are less similar than the other clusters' sets. Since the last three clusters belonged to specific group of proteins from the same families such as ceramidases, sialidases, and ELOV proteins, it was expected to obtain higher similarity score for these three clusters. The number of ligands and their APLSs are listed for each cluster in Table 4.2. The percentage of the ligand pairs having PLS above 0.7 is also given in the Table 4.2.

Table 4.2. For SL clusters, protein and ligand numbers, and PLSs.

Cluster No	Number of Proteins	Number of Ligands	The Average Pairwise Ligand Similarity	Number of Unique PLS Values	Percentage (%) and the Number of PLS Above 0.7
1	14	4526	0.0350	10.262.715	3.105 / 0.03
2	11	2037	0.0379	2.778.903	2.530 / 0.09
3	4	9	0.3139	36	6 / 17
4	4	18	0.2218	153	14 / 9
5	5	11	0.3811	55	22 / 40

As the next step, the drugs connecting these proteins were considered. For each cluster, their ligands were searched in ChEMBL using its web services to find known drugs. For each cluster, drugs and the diseases the drugs cure were searched in DrugBank database [42]. For cluster 1 of SL-WIN, the full name of the drugs, their ChEMBL IDs, and diseases are also listed in Table 4.3. In addition, the list of the proteins and their interactors, drugs, is also given in Table 4.4.

Table 4.3. The drugs of C1of SL-WIN.

Drug Name	ChEMBL ID	Diseases
Estrone	CHEMBL1405	Estrogen Metabolic Disorders
Niclosamide	CHEMBL1448	Tapeworm Infection
Riluzole	CHEMBL744	Amyotrophic Lateral Sclerosis
Leflunomide	CHEMBL960	Immune Disease
Hydralazine	CHEMBL276832	Cardiovascular Disease
Methotrexate	CHEMBL34259	Crohn' Disease
Folic Acid	CHEMBL1622	(It is just a vitamin B)
Prazosin Hydrochloride	CHEMBL1558	Cardiovascular Disease
Terazosin	CHEMBL611	Cardiovascular Disease
Amiloride	CHEMBL945	Chronic Kidney Disease
Doxazosin Mesylate	CHEMBL1200561	Hypertension
Terfenadine	CHEMBL17157	Allergic Skin Disorders
Carbetapentane	CHEMBL173234	Cough Suppressant
Bepridil	CHEMBL1008	Hypertension

Table 4.3. The drugs of C1of SL-WIN (cont.).

Drug Name	ChEMBL ID	Diseases
Haloperidol	CHEMBL54	Schizophrenia
Oxymetazoline	CHEMBL762	Nasal Congestion
Astemizole	CHEMBL296419	Allergy Symptoms
Thonzonium Bromide	CHEMBL1200883	Otic Disease
Miglustat	CHEMBL1029	Gaucher Disease
Amitriptyline	CHEMBL629	Emotional Expression Disorder
Tetracaine	CHEMBL698	Haemorrhoids
Loratadine	CHEMBL998	Allergic Rhinitis
Mibefradil	CHEMBL45816	High Blood Pressure
Pimozide	CHEMBL1423	Tourette's Disorder
Promazine	CHEMBL564	Disturbed Behaviours
Ketotifen	CHEMBL534	Asthma, Skin Allergies
Thioridazine	CHEMBL479	Schizophrenia
Amlexanox	CHEMBL1096	Ulcers
Digoxin	CHEMBL1751	Cardiac Insufficiency
Lansoprazole	CHEMBL480	Acid Reflux Disorders

Table 4.3. The drugs of C1 of SL-WIN (cont.).

Drug Name	ChEMBL ID	Diseases
Dithiazanine Iodide	CHEMBL421701	Heartworms and Threadworms
Triamterene	CHEMBL585	Congestive Heart Failure
Nordihydroguaiaretic Acid	CHEMBL52	Actinic Keratoses
Tolazamide	CHEMBL817	Type 2 Diabetes
Clotrimazole	CHEMBL104	Vaginal Yeast Infections
Cimetidine	CHEMBL30	Acid-Reflux Disorders
Altretamine	CHEMBL1455	Ovarian Cancer
Nabumetone	CHEMBL1070	Osteoarthritis and Rheumatoid Arthritis
Prazosin	CHEMBL2	Hypertension
Methysergide Maleate	CHEMBL1200938	Vascular Headache
Phenoxybenzamine Hydrochloride	CHEMBL1200787	Hypertension
Amiloride Hydrochloride	CHEMBL1398126	Congestive Heart Failure, Hypertension

Table 4.3. The drugs of C1 of SL-WIN (cont.).

Drug Name	ChEMBL ID	Diseases
Terazosin Hydrochloride Hydrate	CHEMBL1201091	Benign Prostatic Hyperplasia
Acrisorcin	CHEMBL1201038	Fungal Infection
Alfuzosin	CHEMBL709	Benign Prostatic Hyperplasia
Fluphenazine	CHEMBL726	Psychoses
Dantrolene	CHEMBL1201288	Malignant Hyperthermia
Tamoxifen	CHEMBL83	Breast Cancer

The proteins listed in Table 4.4 play role in crucial diseases such as immune system disease, cardiovascular disease, and cancer. Leflunomide drug used for the treatment of immune disease is an interactor of NPC1-ARSA protein pairs [43]. The same enzymes are also connected by Hydralazine drug used for the treatment of cardiovascular disease [44]. Thus, more than just immune system therapies, cardiovascular diseases are also effected by the deficiencies of the protein pair, NPC1-ARSA. Terazosin drug is also used for cardiovascular disease; moreover, observed in many protein interactions of C1 of SL-WIN [45]. The protein pairs interacted by Terazosin are ASM-AGAL, ASM-ARSA, ASM-GLCM, AGAL-GLCM, AGAL-ARSA and GLCM-ARSA. The same protein pairs are also important for chronic kidney disease since Amiloride drug is one of the interactors of the pairs [46]. Haloperidol, interacting with ASM, GLCM and ARSA enzymes, is used for the treatment of schizophrenia [47]. For ASM-ARSA pair, Amitriptyline and Thioridazine are observed as the interactor of the pair. The former is used for emotional expression disorder and the latter cures schizophrenia [48,49].

Table 4.4. Known drugs and their interactions in C1.

Interactions	ChEMBL IDs
ARSA- DHB12	CHEMBL1405
NPC1-ARSA	CHEMBL1448, CHEMBL744, CHEMBL 960, CHEMBL276832
ASM-AGAL	CHEMBL34259, CHEMBL611, CHEMBL945
ASM-GLCM	CHEMBL611, CHEMBL945, CHEMBL17157, CHEMBL73234, CHEMBL1008, CHEMBL54, CHEMBL762, CHEMBL296419
CEGT-GBA2	CHEMBL1029
ASM-ARSA	CHEMBL611, CHEMBL945, CHEMBL17157, CHEMBL1008, CHEMBL54, CHEMBL629, CHEMBL698, CHEMBL998, CHEMBL45816, CHEMBL1423, CHEMBL564, CHEMBL534, CHEMBL479
AGAL-NPC1	CHEMBL1096, CHEMBL1751, CHEMBL480
AGAL-ARSA	CHEMBL611, CHEMBL945, CHEMBL585, CHEMBL817, CHEMBL104, CHEMBL30, CHEMBL1455
GLCM-CEGT	CHEMBL1029
GLCM-GBA2	CHEMBL1029
GLCM-NPC1	CHEMBL1096, CHEMBL1070
GLCM-AGAL	CHEMBL611, CHEMBL945, CHEMBL1096, CHEMBL585, CHEMBL2, CHEMBL1201038, CHEMBL709
GLCM-ARSA	CHEMBL611, CHEMBL945, CHEMBL17157, CHEMBL1008, CHEMBL54, CHEMBL585, CHEMBL726, CHEMBL1201288, CHEMBL83

Miglustat drug used for treatment of Gaucher disease is rarely seen as the interactors of the protein pairs [50]. CEGT- GBA2 and GLCM proteins are interacted by this drug in the cluster 1. Another important disease is ulcers cured by Amlexanox drug [51]. This drug interacts AGAL, NPC1, and GLCM enzymes in the cluster 1.

Among the drugs belonging to the cluster 1, there are only two drugs, Tamoxifen and Altretamine, used for the treatment of cancers such as ovarian and breast cancers [52,53]. These drugs are the interactors of the proteins: AGAL, GLCM, and ARSA. Same as the first cluster of SL-WIN, the second cluster's drugs and their diseases are also listed in Table 4.5.

Table 4.5. For C2, drug names, ChEMBL IDs and diseases.

Drug name	ChEMBL ID	Diseases
Sunitinib	CHEMBL535	Cell cancer, neuroendocrine cancer
Imatinib	CHEMBL941	Leukaemia and stomach cancer
Sorafenib	CHEMBL1336	Thyroid cancer, cell cancer
Nilotinib	CHEMBL255863	Leukaemia
Vandetanib	CHEMBL24828	Thyroid cancer
Pazopanib	CHEMBL477772	Kidney cancer
Dasatinib	CHEMBL1421	Leukemia cancer
Tofacitinib	CHEMBL221959	Infection
Erlotinib	CHEMBL553	Pancreatic cancer and lung cancer

Table 4.5. For C2, drug names, ChEMBL IDs and diseases (cont.).

Drug name	ChEMBL ID	Diseases
Gefitinib	CHEMBL939	Lung cancer
Lapatinib	CHEMBL554	Breast cancer
Bosutinib	CHEMBL288441	Leukaemia
Afatinib	CHEMBL1173655	Blocking cancer cell growth
Ruxolitinib	CHEMBL1789941	Infection
Axitinib	CHEMBL1289926	Renal cell carcinoma
Crizotinib	CHEMBL601719	Small cell lung cancer
Nintedanib	CHEMBL502835	Idiopathic Pulmonary Fibrosis
Sirolimus	CHEMBL413	Lymphoma and skin cancer
Niclosamide	CHEMBL1448	Worm infection (KIT-LYN)
Ceritinib	CHEMBL2403108	Non-small cell lung cancer (KIT-LYN)
Fingolimod	CHEMBL314854	MS (SPHK1-SPHK2)

Different than the first cluster, the drugs of the second cluster are frequently observed in the cluster 2 of SL-WIN as the interactors. Except the drugs, Niclosamide, Ceritinib, and Fingolimod, all the drugs are interacted at least 21 enzyme pairs in the cluster 2. However, these three drugs; namely, Niclosamide, Ceritinib, and Fingolimod, are only connectors of specific protein pairs. Niclosamide is used for worm infection [54], Ceritinib is used for non-small lung cancer [55] and Fingolimod

is used for the treatment of MS [56]. These protein pairs and their drugs are listed in Table 4.6.

Table 4.6. For C2, the drugs interacting with less than two protein pair.

Interactions	Drug Names
KIT-LYN	Niclosamide, Ceritinib
SPHK1-SPHK2	Fingolimod

In brief, the second cluster of SL-WIN comprises almost all the sphingolipids which are crucial for cancer treatments. Comparison between the first and the second clusters of SL-WIN also proves that the cancer related sphingolipids belong to the second cluster. The second cluster comprises all the kinases of sphingolipid metabolism. Sphingolipid metabolism and its kinases are the target enzymes for cancer treatment strategies [57]. Sphingolipids particularly kinases cell cycle progression, oncogenesis, and drug resistance in cancer biology.

The other clusters C3 and C5 of SL-WIN does not include any drug as an interactor. Only C4 includes a drug as its interactor. In this cluster, Zanamivir (CHEMBL222813) interacts NEUR2 and NEUR3. This drug is used for treatment of influenza A and B [41].

4.2.1.2. Scaffold Analysis and ZINC Database Search. After the similarity analysis, the scaffold decompositions of each ligand sets were also obtained in CANVAS. First, the binary fingerprints of these ligand sets were calculated, then the scaffold decompositions were constructed. The decomposition lists all the scaffolds in order to the number of their rings. Among all the scaffolds, the most complex as well as commonly observed scaffolds were selected. A scaffold can be directly a ligand;

therefore, the scaffolds frequently seen in more than one ligand were preferred. The list of scaffolds is provided as the supplementary files of the thesis. In the scaffold decomposition, the observation frequency of each scaffold is illustrated by the number of vertical bar, below the scaffold figure. Therefore, frequently shared scaffolds are collected as the basis for ZINC database search. ZINC is a commonly used drug like compound database in which selected substructures are search for the ligands consisting them [58]. From the decomposition, the more complex and frequently observed scaffolds with low binding affinities are selected. For each scaffold, the vertical bars are variously distanced with respect to the average binding affinity shown by the triangles. The binding affinities of the vertical bars belong to the ligands consisting the scaffolds as their substructures. The lower binding affinities are preferred since the ligands with low binding affinities can easily interact with proteins. For the vertical bars, blue and orange areas mean the lowest and the highest binding affinities; respectively. To illustrate, scaffolds identified by scaffold decomposition belonging to the fourth cluster is given in Figure 4.5. The selected scaffold of the cluster is signed by * in the figure.

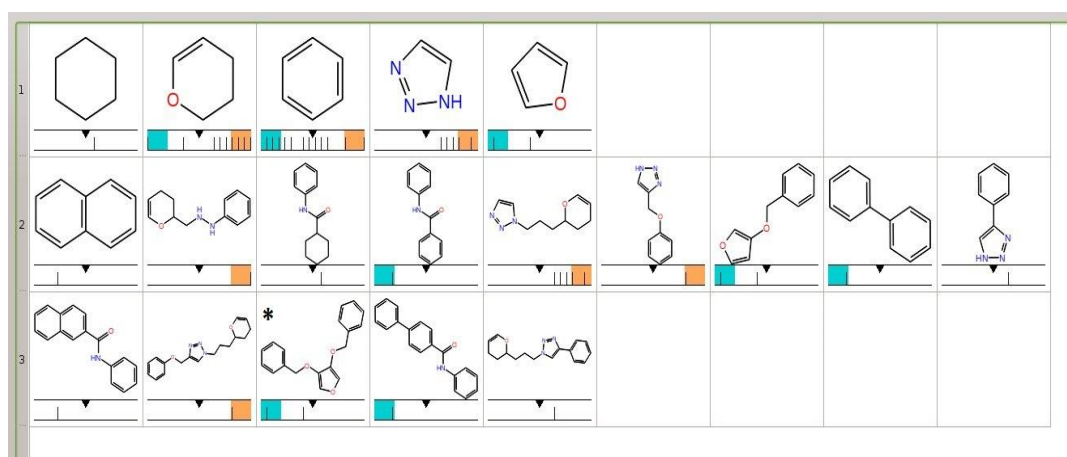


Figure 4.5. Scaffolds identified by scaffold decomposition.

For each sphingolipid cluster, the scaffolds are searched in ChemSpider [59]. For each sphingolipid cluster, the scaffolds found in ChemSpider are listed in the Table 4.7.

Table 4.7. The scaffolds of sphingolipid clusters are listed by their ChemSpider names and their figures.

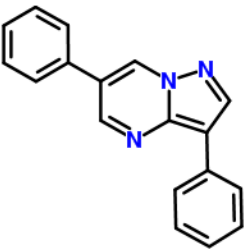
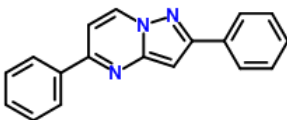
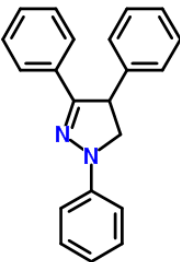
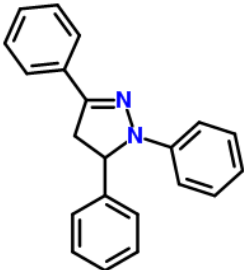
Cluster No	ChemSpider Name	Figure
1	3,6-Diphenylpyrazolo[1,5-a]pyrimidine	 The structure shows a fused bicyclic system consisting of a pyrazole ring fused to a pyrimidine ring. Two phenyl rings are attached to the pyrazole ring at the 3 and 6 positions.
1	2,5-Diphenylpyrazolo[1,5-a]pyrimidine	 The structure shows a fused bicyclic system consisting of a pyrazole ring fused to a pyrimidine ring. Two phenyl rings are attached to the pyrazole ring at the 2 and 5 positions.
1	1,3,4-Triphenyl-4,5-dihydro-1H-pyrazole	 The structure shows a five-membered pyrazole ring with a nitrogen atom at the bottom position. Three phenyl rings are attached to the ring at the 1, 3, and 4 positions.
1	1,3,5-triphenyl-2-pyrazoline	 The structure shows a five-membered pyrazoline ring with a nitrogen atom at the top position. Three phenyl rings are attached to the ring at the 1, 3, and 5 positions.

Table 4.7. The scaffolds of sphingolipid clusters are listed by their ChemSpider names and their figures (cont.).

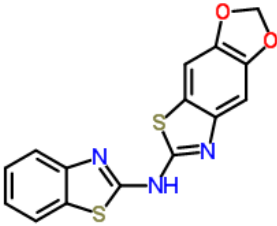
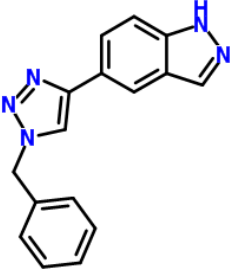

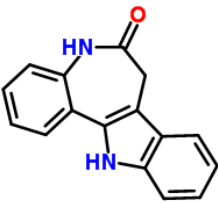
Cluster No	ChemSpider Name	Figure
1	N-(1,3-Benzothiazol-2-yl)[1,3]dioxolo[4,5-f][1,3]benzothiazol-6-amine	 <p>The structure shows a 1,3-benzothiazole ring system connected at its 2-position to the 6-position of a [1,3]dioxolo[4,5-f][1,3]benzothiazole ring system. The nitrogen at the 6-position is substituted with an amino group (-NH-).</p>
2	5-(1-Benzyl-1H-1,2,3-triazol-4-yl)-1H-indazole	 <p>The structure features a 1H-indazole ring system with a 1,2,3-triazole ring attached at its 5-position. The 1-position of the triazole ring is substituted with a benzyl group (-CH2-Ph).</p>
2	N-Phenyl-4-(pyrazolo[1,5-b]pyridazin-3-yl)-2-pyrimidinamine	 <p>The structure consists of a pyrazolo[1,5-b]pyridazine ring system connected at its 4-position to the 2-position of a pyrimidinamine ring system. The nitrogen at the 2-position of the pyrimidinamine is substituted with a phenyl group (-Ph).</p>
2	7,12-Dihydroindolo[3,2-d][1]benzazepin-6(5H)-one	 <p>The structure shows a complex polycyclic system with a benzazepine core fused to an indole ring system. A carbonyl group (=O) is attached to the 6-position of the benzazepine ring, and the nitrogen at the 5-position is hydrogenated (5H).</p>

Table 4.7. The scaffolds of sphingolipid clusters are listed by their ChemSpider names and their figures (cont.).

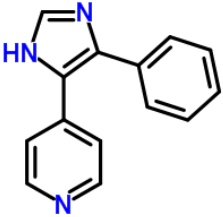
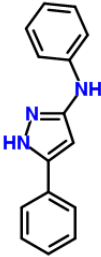
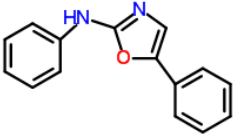
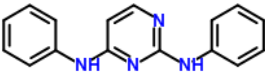
Cluster No	ChemSpider Name	Figure
2	4-(4-Phenyl-1H-imidazol-5-yl)pyridine	 <p>The structure shows a pyridine ring substituted at the 4-position with a 4-phenyl-1H-imidazol-5-yl group. The imidazole ring is connected to the pyridine ring at its 5-position, and a phenyl ring is attached to the imidazole ring at its 4-position.</p>
2	N,5-Diphenyl-1H-pyrazol-3-amine	 <p>The structure shows a 1H-pyrazole ring substituted at the 3-position with an amino group (-NH-) and at the 5-position with a phenyl ring. The amino group is further substituted with another phenyl ring.</p>
2	N,5-Diphenyl-1,3-oxazol-2-amine	 <p>The structure shows a 1,3-oxazole ring substituted at the 2-position with an amino group (-NH-) and at the 5-position with a phenyl ring. The amino group is further substituted with another phenyl ring.</p>
2	N,N'-Diphenyl-2,4-pyrimidinediamine	 <p>The structure shows a 2,4-pyrimidinediamine ring substituted at the 2 and 4 positions with amino groups (-NH-). Each amino group is further substituted with a phenyl ring.</p>

Table 4.7. The scaffolds of sphingolipid clusters are listed by their ChemSpider names and their figures (cont.).

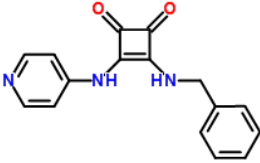
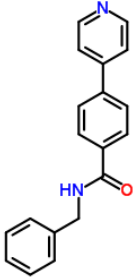
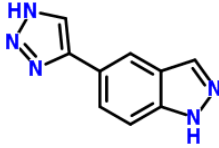
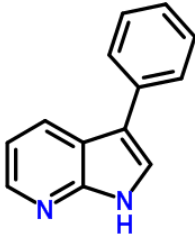
Cluster No	ChemSpider Name	Figure
2	3-(Benzylamino)-4-(4-pyridinylamino)-3-cyclobutene-1,2-dione	
2	N-Benzyl-4-(4-pyridinyl)benzamide	
2	5-(1H-1,2,3-Triazol-4-yl)-1H-indazole	
2	3-Phenyl-1H-pyrrolo[2,3-b]pyridine	

Table 4.7. The scaffolds of sphingolipid clusters are listed by their ChemSpider names and their figures (cont.).

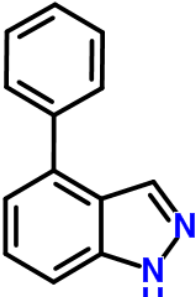
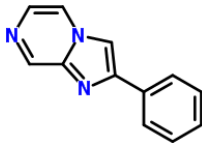
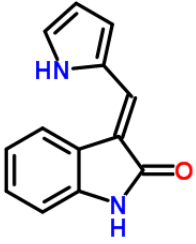
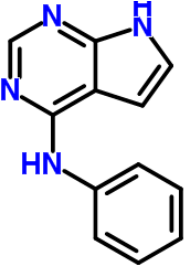
Cluster No	ChemSpider Name	Figure
2	4-Phenyl-1H-indazole	
2	2-Phenylimidazo[1,2-a]pyrazine	
2	(3E)-3-(1H-Pyrrol-2-ylmethylene)-1,3-dihydro-2H-indol-2-one	
2	4-(phenylamino)-7H-pyrrolo(2,3-d)pyrimidine	

Table 4.7. The scaffolds of sphingolipid clusters are listed by their ChemSpider names and their figures (cont.).

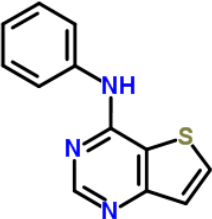
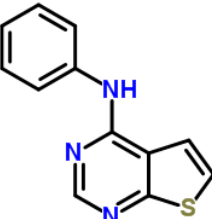
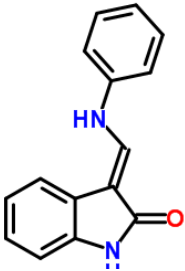
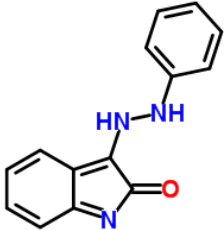
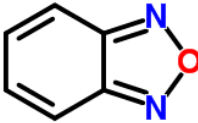
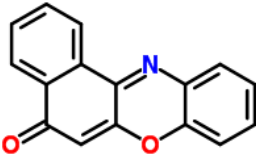
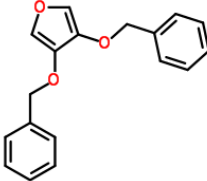
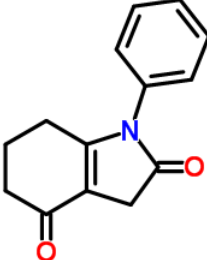
Cluster No	ChemSpider Name	Figure
2	N-Phenylthieno[3,2-d]pyrimidin-4-amine	 <p>The structure shows a thieno[3,2-d]pyrimidine core. The nitrogen at position 4 is substituted with a phenyl group. The nitrogen at position 2 is also highlighted in blue.</p>
2	N-Phenylthieno[2,3-d]pyrimidin-4-amine	 <p>The structure shows a thieno[2,3-d]pyrimidine core. The nitrogen at position 4 is substituted with a phenyl group. The nitrogen at position 2 is also highlighted in blue.</p>
2	(3E)-3-(Anilinomethylene)-1,3-dihydro-2H-indol-2-one	 <p>The structure shows a 1,3-dihydro-2H-indol-2-one core. The nitrogen at position 3 is substituted with a phenyl group via a methylene bridge. The nitrogen at position 1 is highlighted in blue.</p>
2	3-(2-Phenylhydrazino)-2H-indol-2-one	 <p>The structure shows a 2H-indol-2-one core. The nitrogen at position 3 is substituted with a phenyl group via a hydrazino bridge. The nitrogen at position 1 is highlighted in blue.</p>

Table 4.7. The scaffolds of sphingolipid clusters are listed by their ChemSpider names and their figures (cont.).

Cluster No	ChemSpider Name	Figure
3	Benzofurazan	
3	5H-Benzo[a]phenoxazin-5-one	
4	3,4-Bis(benzyloxy)furan	
5	1-Phenyl-3,5,6,7-tetrahydro-1H-indole-2,4-dione	

The dominant substructures of these scaffolds are pyrazole, purine, and pyrimidine with electronegative atoms, N, O and S, having the tendency to attract a bonding pair of electrons. The inhibition of cellular replication is the main tactic for anticancer agents killing cancer cells by inhibiting their DNA synthesis [60]. For the synthesis of deoxyribonucleotides used for DNA synthesis, pyrimidines and purines can be salvaged in the human cells. These nucleotides and their analogues are important class of anticancer agents. Pyrazoles are also crucial for cancer treatment since it is an anti-angiogenesis [61]. Angiogenesis is the production of new blood vessels whose increase in the tumor cells encourage the tumor growth. Thus, pyrazoles are commonly used for cancer treatment to block tumor growth.

For each sphingolipid ligand set, these selected scaffolds were searched in ZINC database [58]. The scaffolds were searched in the database to detect the ligands including these scaffolds as their substructures. The number of scaffolds and the number of related ligands including the scaffolds as substructures are listed in the Table 4.8. For the second cluster, there is no matches in ZINC database.

Table 4.4. The number of scaffolds and the number of detected ligands.

Cluster	Number of Selected Scaffolds	Number of Ligands Extracted from ZINC
Cluster 1	9	2965
Cluster 3	2	4055
Cluster 4	3	9
Cluster 5	5	215

4.2.1.3. Protein – Ligand Docking of SL-WIN. These ZINC ligand sets and OTAVA library were used for sphingolipid protein-ligand docking studies. For each

cluster, their ligand sets were docked to their proteins. The docking process was performed in Maestro. The proteins having PDB structures were used. For each cluster, their ligand sets collected from ZINC were docked to the clusters' proteins. For the ligand set of C2, OTAVA dataset was used since there was no compounds in ZINC database, comprising the scaffolds of C2 [62]. The ligands, well-docked into more than 1 protein with docking scores below -6 kcal/mole are listed in the Table 4.9. These results can guide experimental approaches to sphingolipid protein-ligand interactions.

Table 4.5. For SL-WIN, the ligands docked into more than one protein with docking scores below -6 are listed.

LIGAND ID	TARGET PROTEINS	DOCKING SCORES
ZINC72267284	ASM, BGAL, AGAL	-6.231, -8.496, -7.807
ZINC72267285	ASM, BGAL, AGAL	-6.230, -8.496, -7.807
ZINC22058728 (NPC)	ASM, BGAL, AGAL	-6.215, -8.925, -6.583
ZINC01625746	ASM, AGAL	-6.884, -6.564
ZINC13000556	ASM, AGAL	-6.884, -6.564
ZINC03826691 (Belotecan)	AGAL, BGAL	-6.128, -7.949
ZINC05924106	BGAL, GLCM	-6.828, -6.395
ZINC13832891	BGAL, GLCM	-6.828, -6.395

Table 4.9. For SL-WIN, the ligands docked into more than one protein with docking scores below -6 are listed (cont.).

LIGAND ID	TARGET PROTEINS	DOCKING SCORES
OTAVA_Drug-Like_Green_Collection.118826	FYN, GBA3	-6.007, -8.586
OTAVA_PrimScreen1.394	FYN, GBA3	-6.157, -7.471
OTAVA_Drug-Like_Green_Collection.22943	FYN, GBA3	-6.628, -7.14
ZINC01611274 (Topotecan)	ASM, AGAL	-6.150, -6.204
ZINC09427866	GLCM, AGAL	-6.773, -6.029
ZINC04833656	GLCM, BGAL	-6.961, -7.005
ZINC04980573	GLCM, BGAL	-7.080, -6.244
ZINC05105116	GLCM, BGAL	-6.347, -6.178
ZINC05580265	AGAL, BGAL	-6.792, -6.142
ZINC15772765	ASM, AGAL	-6.631, -8.038

The listed ligands are simply the inhibitors of the enzymes listed as short protein names; however, there are also three drugs: ZINC01611274 (Topotecan), ZINC03826691 (Belotecan), and ZINC22058728 (NPC). Topotecan and Belotecan are used for the treatment of small cell lung cancer [63,64]. NPC is used for the treatment of nasopharynx cancer [65]. These three drugs were docked into AGAL.

Molecular Symmetry Based Chirality and Polarity Detection: For the listed ligands (Table 4.9), their symmetry groups were calculated using Jmol. The symmetry group of almost all the ligands is C_1 . Thus, almost all the ligands are chiral and polar molecules except OTAVA_PrimaryScreen1.394 having symmetry group C_s . Therefore, this ligand is polar meaning also water soluble.

4.2.2. Construction of Combined Network of Sphingolipids and Insulins

Same as sphingolipid networks, combined sphingolipid and insulin weighted identity network was also constructed. This network is illustrated in Figure 4.7 (A). First, the clustering algorithm MCL was performed in CytoScape. Inflation value was selected as 2.5 for MCL algorithm. Similar to SL-WIN network, number of inflation values ranged from 1.5 to 4 were tried, and their output clusters were biologically considered to decide more relevant inflation value for the network. The output of MCL algorithm, i.e. cluster, is given in Figure 4.7 (B). The frequency distribution of proteins having number of shared ligands equal or higher than 50 is also given in Figure 4.6.

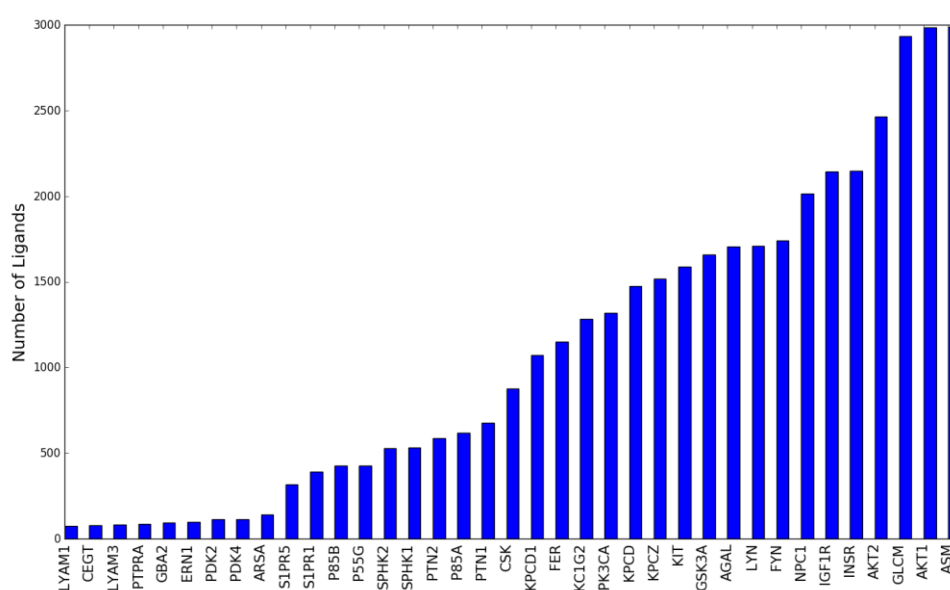


Figure 4.6. Number of shared ligands higher than 50

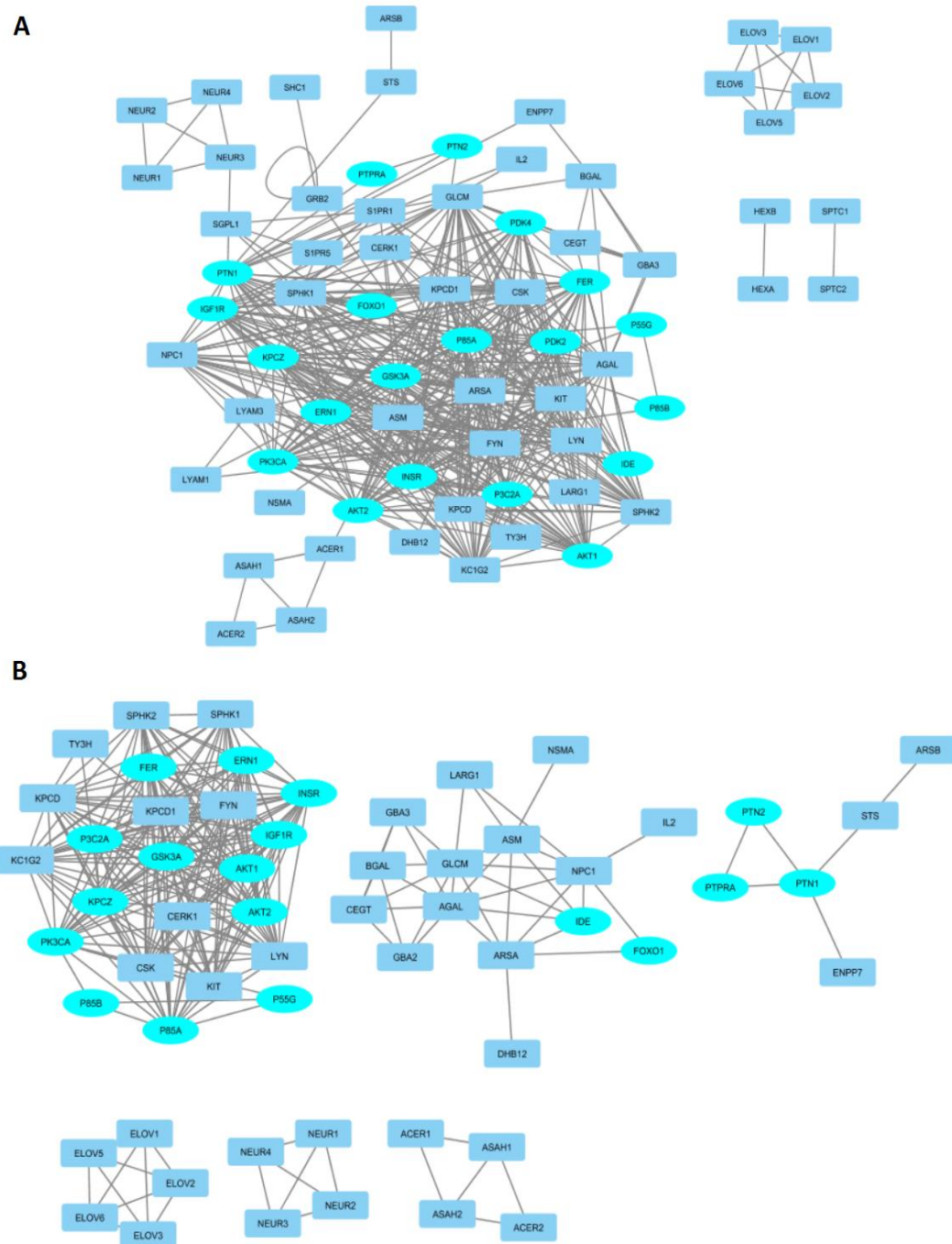


Figure 4.7. (A) SPHINS-WIN, where insulins and sphingolipids are shaped as ellipse and rectangles, respectively. (B) The clusters of SPHINS-WIN

4.2.2.1. Interactor Analysis of Combined Sphingolipid & Insulin Network. In these clusters, insulins are distributed into the sphingolipids different than sphingolipid clusters. First three clusters of sphingolipid & insulin (SPHINS) network include sphingolipids and insulins together. The other clusters of SPHINS network are identical to the sphingolipid clusters. Thus, first three clusters are analysed to demonstrate the interactions between both metabolisms. For these clusters, their shared ligands are obtained from CytoScape file, then SMILES of ligands are collected from ChEMBL via its webresource client. These ligand IDs and their SMILES are listed and saved as sdf file. These ligand sets of these clusters are analysed in CANVAS, and their similarity distributions are obtained similar to the process completed for sphingolipid clusters. The similarity distribution of SPHINS ligands are given in Figure 4.8.

The first cluster (C1) of the weighted identity SPHINS network includes all the kinases of sphingolipid and insulin metabolisms together. Additional to the kinases, there are also two insulin receptor proteins, namely IGF1R and INSR. The number of shared ligands the cluster has is 5010 with the average pairwise ligand similarity, 0.0365. The ligand pairs having similarity above 0.7 includes two drugs Palbociclib and Sunitinib. These drugs were paired with similar ligands (CHEMBL365847 for Palbociclib and CHEMBL1721885 for Sunitinib). Sunitinib and CHEMBL1721885 interacted with almost all the proteins except P85A, P85B and TY3H. Palbociclib targeted GSK3A, FYN and IGF1R; however, CHEMBL365847 interacted with AKT1, AKT2, KPCD, KPCD1, and KPCZ. The similarity distribution of the first cluster is given in Figure 4.8.

C1 of the SL-WIN appears as the second cluster (C2) of the weighted identity SPHINS network. Additional to the sphingolipid cluster, the second cluster of SPHINS network consists two insulin proteins; FOXO1 and IDE. This cluster comprises glycosidases and these insulin enzymes. One of molecular functions of IDE is glycoprotein binding [66]. Thus, it is biologically relevant to have this enzyme with

GLCM and GBA2 in the same cluster. FOXO1 is a transcription factor which is a significant target of insulin signaling as well as a regulator of metabolic homeostasis [67].

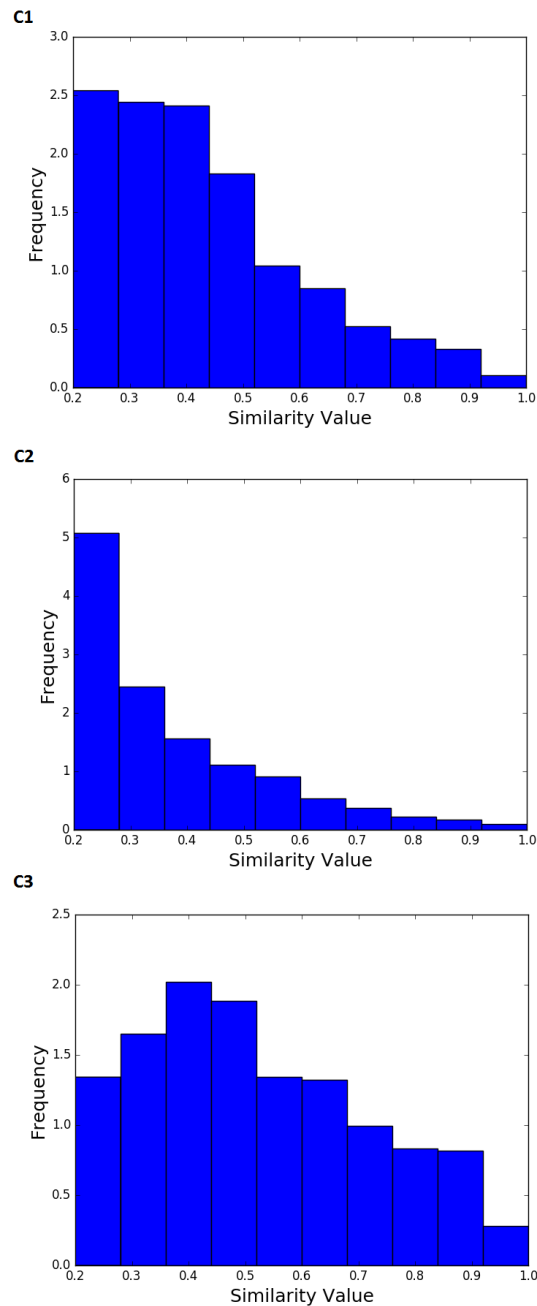


Figure 4.8. The similarity distribution of the weighted identity SPHINS clusters. X index is for the similarity values and Y is for the frequency of these similarity values.

For SPHINS network, the pairwise ligand similarity values of the second cluster are lower than the first cluster. The cluster has 2257 ligands. Furthermore, it does not include many insulin proteins. The sphingolipid members of this cluster are GLCM related proteins. To compare the pairwise ligand similarity of both sphingolipid networks and SPHINS, the average PLS of the first cluster belonging to the weighted identity sphingolipid network was 0.0350; on the other hand, the average similarity of the second cluster belonging to the weighted identity SPHINS was 0.0349. In the cluster, the ligand pairs with similarity above 0.7 includes drugs namely Miglustat, Prazosin, Amiloride and Terazosin. Each drug was paired with their derivatives and targeted the same proteins in the cluster. Miglustat and its derivatives (CHEMBL408500, CHEMBL1086997, CHEMBL1076754) targets CEGT, GBA2 and GLCM. The other drugs and their derivatives (CHEMBL1558 for Prazosin, CHEMBL540851 and CHEMBL1398126 for Amiloride, and CHEMBL1256665, CHEMBL1554413, and CHEMBL1201091 for Terazosin) interacted with GLCM, ARSA, ASM and AGAL. The similarity distributions of both clusters are almost same since the only difference was these two insulins. To illustrate the similarity results obtained from the similarity matrix, similarity distribution of this second cluster belonging to the weighted identity SPHINS is given in Figure 4.8.

The third cluster (C3) of the SPHINS network has phosphatases namely ENPP7, PTN1, PTPRA, PTN2 and PP2AA as well as two sulfatases, STS and ARSB. Having phosphatases and sulfatases together in a cluster is an expected outcome since the inhibition of phosphatases causes the decrease in the sulfatase activity [68]. The cluster does not have a dominant metabolism since there are three insulins and four sphingolipids. The ligand set of this cluster has 661 compounds; moreover, the average PLS of them is 0.500 which is the highest average PLS value obtained for both sphingolipid and insulin data. It was expected to obtain the result since the compounds playing role at the inhibition of phosphatases affect also the sulfatase activities [68]. For this cluster, the plot of the pairwise ligand similarity matrix is given in Figure 4.8. The remaining clusters of SPHINS network, C4, C5 and C6 are the same as the clusters of sphingolipid metabolism; namely, ELOV proteins, sialidases and ceramidases.

In brief, the first and the second clusters of the weighted identity SPHINS have almost the same similarity distributions when they are compared with the sphingolipid clusters. The only difference is specifically obtained for the third cluster. The results of these pairwise ligand similarity matrices are summarized in Table 4.10. The percentage of the ligands having PLS above 0.7 is also given in the Table 4.10.

Table 4.10. For SPHINS clusters, number of proteins, number of ligands and ligand pairwise similarity values are listed.

Cluster No	Number of Proteins	Number of Ligands	The Average Pairwise Ligand Similarity	Number of Unique PLS Values	Percentage (%) and the Number of PLS Above 0.7
1	24	5010	0.0365	12.547.545	16.871 / 0.135
2	15	2257	0.0350	10.605.315	3.072 / 0.038
3	6	661	0.5000	218.130	1.614 / 0.746

Similar to SL clusters, these clusters were analyzed. First cluster of the SPHINS network comprises kinases. The cluster includes 23 drugs; 18 out of which are used for the treatment of cancer. There are also 3 drugs curing infections, one for MS (multiple sclerosis) and one for idiopathic pulmonary fibrosis. These drugs, listed in Table 4.11 were frequently observed for these sphingolipid-insulin interactions. Almost all the drugs have at least 20 interactions with protein pairs. Besides these promiscuous drugs, 4 of them interacted with less than 7 proteins pairs. These 4 drugs and the protein interactions in which they were observed are listed in Table 4.12.

Table 4.11. 23 anticancer drugs were observed in C1-SPHINS.

Drug Name	ChEMBL ID	Diseases
Tofacitinib	CHEMBL221959	Infection
Lapatinib	CHEMBL554	Breast cancer
Sorafenib	CHEMBL1336	Thyroid cancer, cell cancer
Sunitinib	CHEMBL535	Cell cancer, neuroendocrine cancer
Vandetanib	CHEMBL24828	Thyroid cancer
Dasatinib	CHEMBL1421	Leukemia cancer
Erlotinib	CHEMBL553	Pancreatic cancer and lung cancer
Gefitinib	CHEMBL939	Lung cancer
Pazopanib	CHEMBL477772	Kidney cancer
Imatinib	CHEMBL941	Leukaemia and stomach cancer
Bosutinib	CHEMBL288441	Leukaemia
Ruxolitinib	CHEMBL1789941	Infection
Nilotinib	CHEMBL255863	Leukaemia
Afatinib	CHEMBL1173655	Blocking cancer cell growth
Crizotinib	CHEMBL601719	Small cell lung cancer
Axitinib	CHEMBL1289926	Renal cell carcinoma
Sirolimus	CHEMBL413	Lymphoma and skin cancer

Table 4.11. 23 anticancer drugs were observed in C1-SPHINS (cont.).

Drug Name	ChEMBL ID	Diseases
Niclosamide	CHEMBL1448	Worm infection
Mitoxantrone	CHEMBL58	Prostate cancer and leukaemia
Ceritinib	CHEMBL2403108	Non-small cell lung cancer
Palbociclib	CHEMBL189963	Breast cancer
Fingolimod	CHEMBL314854	MS
Nintedanib	CHEMBL502835	Idiopathic Pulmonary Fibrosis

Table 4.12. The drugs interacting less than 7 proteins pairs in C1 of SPHINS.

ChEMBL ID	Interactions
CHEMBL1448 (Niclosamide)	KIT-LYN, AKT1-LYN, AKT1-KIT
CHEMBL2403108 (Ceritinib)	KIT-LYN, IGF1R-LYN, IGF1R-KIT, INSR-LYN, INSR-KIT, INSR-IG1FR
CHEMBL189963 (Palbociclib)	IGF1R-GSK3A, IGF1R-FYN, GSK3A-FYN
CHEMBL314854 (Fingolimod)	SPHK1-SPHK2

These interactions and drugs highlight important enzymes for the diseases cured by the drugs. KIT (stem cell growth factor receptor Kit), LYN (Tyrosine-protein kinase Lyn), and AKT1 (alpha serine/threonine-protein kinase) enzymes are interacted with Niclosamide used for worm infection [54]. However, KIT and LYN are interacted also by Ceritinib curing non-small lung cancer [55]. It was assumed to have a tyrosine protein kinase as the targets of the drug since the drug is a tyrosine protein kinase

inhibitor. Tyrosine kinases catalyse the cell growth and proliferation which also increases cancer cell growth [69]. Moreover, tyrosine kinases inhibitors block oncogenic activations in cancer cells. In addition to the proteins, IGF1R (insulin-like growth factor 1 receptor) and INSR (insulin receptor) are also interacted via this cancer drug. Activation and cell growth of cancer cells are also stemmed from the activity of insulin receptors IGF1R and INSR; moreover controlling these receptors does not only prevent cancer cell growth, cancer progression and resistance to cancer treatment but also controlling insulin resistance [70]. Cancer treatments target also insulin resistance and insulin signaling pathway disorders. Another drug called Palbociclib is used for breast cancer treatment [71]. This drug interacts with IGF1R, GSK3A (Glycogen synthase kinase-3 alpha) and FYN (Tyrosine-protein kinase Fyn). Similar to Ceritinib drug, Palbociclib interacts with a tyrosine-protein kinase and insulin proteins, IGF1R and GSK3A. Fingolimod used for treatment of MS illness [56]; moreover, this drug interacts with SPHK1 (Sphingosine kinase 1) and SPHK2 (Sphingosine kinase 2). Thus, these kinases are obviously preferable for enzyme therapies of MS patients. Fingolimod is phosphorylated by sphingosine kinases and then interacts with sphingosine 1-phosphatase, sphingosine 3-phosphatase and sphingosine 5-phosphatase receptors [72]. Fingolimod is a sphingosine 1-phosphatase receptor modulator, moreover, it reduces the relapse rates of MS patients [56].

On the other hand, in the second and the third clusters of this network, there are a few insulins interacting with sphingolipids. These insulins and their shared ligands were chosen as the target molecules to highlight the significance of these specific interactions. In the second cluster given in Figure 4.9, there are only two insulins interacting directly with five sphingolipids which are glycosidases. For these seven proteins and their interactions, there are 85 unique ligands interacting them. These ligands were docked into the proteins to range the ligands based on their docking scores. According to the docking scores, the well-docked ligands were detected. Before docking them, these ligands were searched in ChEMBL via its client to classify putative and known drugs. 3 of these 85 ligands were known drugs. These ligands are CHEMBL1563 (Daunoxome), CHEMBL1200883 (Tonzonium Bromide), and

CHEMBL1607 (Hycamptin). These drugs were observed in 4 protein-protein interactions in the cluster. These interactions and the observed drugs are listed in Table 4.13.

Table 4.13. Three drugs and their protein interactions from C2 of SPHINS.

Interaction	Drug(s)
ASM -IDE	CHEMBL1200883, CHEMBL1607
NPC1 – IDE	CHEMBL1607, CHEMBL1200883
GLCM – IDE	CHEMBL1200883, CHEMBL1607
FOXO1 – NPC1	CHEMBL1563

CHEMBL1607 (Hycamptin) is an anticancer drug used for the treatment of ovarian cancer and lung cancer [73]. The drug diminishes cancer cells. CHEMBL1563 (Daunoxome) is also an antitumor antibiotic having cytotoxic impacts on cancer cells to prevent cancer cell growth [74].

Different than the previous drugs, CHEMBL1200883 (Tonzonium Bromide) is an antibiotic curing infections [75]. From Table 4.13, it is proved that these cancer drugs are targeting both insulin and sphingolipid protein in the treatment of cancers. The frequently interacted insulin enzyme is called insulin degrading enzyme (IDE). This enzyme interacts with sphingolipid phosphodiesterase (ASM), glucosylceramidase (GLCM), and Niemann Pick protein 1 (NPC1). NPC1 is a cholesterol transporter which acts also on insulin signaling pathway because of the interaction between insulin and cholesterol metabolisms [76]. Thus, NPC1 protein also affects both insulin based disorders and cancers as a sphingolipid protein. These listed anticancer drugs inhibit both insulin degrading enzyme and also sphingolipid proteins to prevent cancer cell growth in human body.

4.2.2.2. Protein-Ligand Docking of Key Sphingolipid-Insulin Interactions. For docking process, first, these 7 proteins were prepared in Maestro from Protein Preparation Wizard, then the ligand set comprising these 85 ligands were imported into the environment. According to that, these ligands were prepared for docking via Lig-Prep. Then, these ligands were docked into the generated grid receptor area. The top 5 XP Gscores are listed for each protein in Table 4.14.

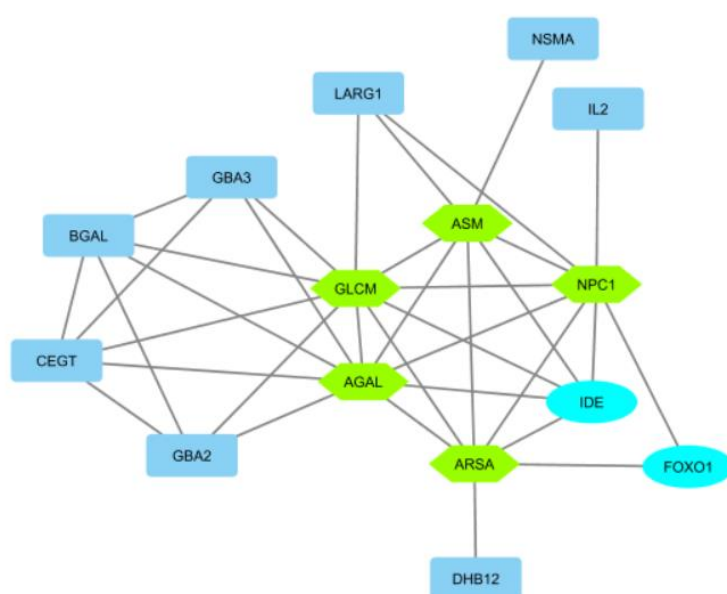


Figure 4.9. The second cluster of weighted identity SPHINS network. Green hexagonal nodes represent the first sphingolipid neighbours of the insulins coloured with light blue and shaped as ellipse.

Table 4.14. The docking results of the C2-SPHINS proteins.

Protein Name	Protein- PDB ID	Ligand- ChEMBL ID	XP Gscore
IDE	4PES	CHEMBL1418096	-8.165
IDE	4PES	CHEMBL250711	-7.836

Table 4.14. The docking results of the C2-SPHINS proteins (cont.).

Protein Name	Protein- PDB ID	Ligand- ChEMBL ID	XP Gscore
IDE	4PES	CHEMBL1504972	-7.695
IDE	4PES	CHEMBL1367989	-7.61
IDE	4PES	CHEMBL1563	-7.581
AGAL	4NXS	CHEMBL1491847	-8.942
AGAL	4NXS	CHEMBL1514790	-8.852
AGAL	4NXS	CHEMBL1491847	-8.485
AGAL	4NXS	CHEMBL1563	-8.045
AGAL	4NXS	CHEMBL86464	-7.6539
GLCM	2XWD	CHEMBL1553406	-8.6715
GLCM	2XWD	CHEMBL1514790	-7.8364
GLCM	2XWD	CHEMBL1591898	-7.3637
GLCM	2XWD	CHEMBL2000525	-7.2902
GLCM	2XWD	CHEMBL1389865	-7.1456
ASM	5I81	CHEMBL1563	-8.2484
ASM	5I81	CHEMBL1966241	-7.7208

Table 4.14. The docking results of the C2-SPHINS proteins (cont.).

Protein Name	Protein- PDB ID	Ligand- ChEMBL ID	XP Gscore
ASM	5I81	CHEMBL1514790	-7.330
ASM	5I81	CHEMBL1572827	-7.2156
ASM	5I81	CHEMBL1991601	-7.1011
ARSA	2AIK	CHEMBL2000525	-5.632
ARSA	2AIK	CHEMBL1563	-3.678
ARSA	2AIK	CHEMBL3189447	-3.47
ARSA	2AIK	CHEMBL1553406	-3.454
ARSA	2AIK	CHEMBL1966241	-3.285

Based on the results, not all the ligands were docked to the proteins. For instance, none of the ligands were docked to NPC1. In Table 4.14, some ligands are well-docked more than one protein. These ligands are CHEMBL589694, CHEMBL1367989, CHEMBL250711, CHEMBL1563, and CHEMBL1360013. Among these ligands, CHEMBL250711 (N-Oleoyldopamine) and CHEMBL1563 (Daunoxome) are known drugs. N-Oleoyldopamine is a lipid producing hyperalgesia meaning the increased sensitivity to pain [77]. It is a lipid found in bovine brain. Daunoxome is used for the treatment of leukemias [74]. Daunoxome is well-docked into two proteins; namely AGAL and IDE. N-Oleoyldopamine is well-docked into IDE and ARSA.

In the third cluster given in Figure 4.10, there are three sphingolipids and three insulins respectively. Form these nodes, the key interactions, bridging both

metabolisms, belonged to PTN1, STS and ENPP7 proteins. PTN1 is called tyrosine-protein phosphatase non-receptor type 1 belonging to insulin metabolism. On the other hand, STS is a sulfatase and ENPP7 is a phosphodiesterase belonging to sphingolipid metabolism. The shared ligand connecting SST and ENPP7 with PTN1 is a dye called evans blue (CHEMBL1200712).

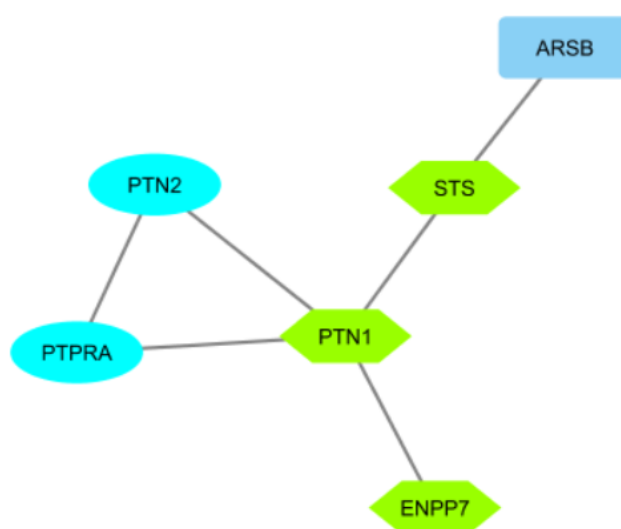


Figure 4.10. The third cluster of weighted identity SPHINS-WIN. Green hexagonal nodes represent the target proteins bridging both metabolisms. In the graphic, sphingolipid and insulin proteins are shaped as rectangular and ellipse, respectively.

Molecular Symmetry Based Chirality and Polarity Detection: The chirality and polarity of the molecules were also calculated via Jmol tool. Among these 85 ligands, the symmetry group of CHEMBL1369972, CHEMBL1349451 and CHEMBL1730051 ligands is C_s having identity symmetry operation and mirror plan reflection. Thus, these molecules are polar meaning also water soluble according to the chemistry based classification of symmetry groups. The other molecules' symmetry groups are identical; which is C_1 comprising only identity symmetry

operation. C_1 is a polar and chiral symmetry group, therefore, these molecules are also water soluble and chiral.

4.3. Comparison Between Sphingolipids and Inflammation Enzymes

For the study of our lab member, Begum Yagci, the data of both sphingolipids and inflammation enzymes were collected; then the ligand centric network models of both metabolic pathways were constructed same as previous metabolic networks. Sphingolipids, insulins and inflammation are all related to each other since obesity increases secretion of inflammatory cytokines whose role is in the communication between cells via cell signaling [78]. This increase of the cytokines causes also chronic inflammation in patients' bodies. My task was to intersect both metabolic networks, sphingolipids and inflammations, to highlight commonly shared proteins between both networks.

4.3.1. Sphingolipid and Inflammation Data Summary

First, proteins and ligands were extracted for both sphingolipids and inflammation keywords. The data extraction processes were completed separately. The keywords used for sphingolipids were sphingolipid, glycosphingolipid and sphingomyelin; on the other hand, inflammation was just preferred for inflammation proteins. In addition to the keywords, GO IDs were also search with these keywords. For sphingolipids, GO:0006665-sphingolipid metabolic process, and for inflammation, GO:0006954-inflammatory response, were preferred since these IDs are the basic metabolic processes of both protein families. The total number of proteins and the number of proteins having ChEMBL compound information are given in the Table 4.15. Data of both protein families were intersected. First, there were 9 proteins 3 of which did not have any ChEMBL compound information. These proteins are ASM3B, D3DWC4 (unreviewed), A0A024RDA0 (unreviewed). The other proteins

were 6 proteins; namely, LYAM3, P2RX7, LYN, SPHK1, KPCD1, and KIT. These proteins are well known enzymes in their enzyme families. LYN and KIT are from the second cluster of SL-WIN network. In addition, KPCD1 is a protein kinase D1 and SPHK1 is a sphingosine kinase 1. These kinases were commonly observed in both sphingolipid and insulin data sets. Different than the kinases, P2RX7 is a purinoceptor protein from sphingolipid metabolism. The full description of these proteins is also listed in Table 4.16.

Table 4.15. Protein numbers of both sphingolipids and inflammations.

Family	Number of total proteins	Number of proteins having ChEMBL IDs	Number of Ligands
Sphingolipids	327	75	51383
Inflammation	1059	270	250576

Table 4.16. The list of the proteins observed at the intersection between both protein families.

ENTRY NAMES	PROTEIN NAMES
LYAM3	P-selectin (CD62 antigen-like family member P)
P2RX7	P2X purinoceptor 7
LYN	Tyrosine-protein kinase Lyn (EC 2.7.10.2)
SPHK1	Sphingosine kinase 1 (SK 1) (SPK 1) (EC 2.7.1.91)
KPCD1	Serine/threonine-protein kinase D1 (EC 2.7.11.13) (Protein kinase C mu type)

Table 4.16. The list of the proteins observed at the intersection between both protein families (cont.).

ENTRY NAMES	PROTEIN NAMES
ASM3B	Acid sphingomyelinase-like phosphodiesterase 3b (ASM-like phosphodiesterase 3b) (EC 3.1.4.-)
D3DWC4	Phosphatidylcholine:ceramide cholinephosphotransferase 1 (Transmembrane protein 23, isoform CRA_a)
A0A024RDA0	V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog, isoform CRA_a
KIT	Mast/stem cell growth factor receptor Kit (SCFR) (EC 2.7.10.1)

4.3.2. Construction of Inflammation Network

First, the ligand centric network models of inflammation proteins were constructed. Then, the weighted identity network was considered for the ligand based analysis. The weighted identity inflammation network is given in Figure 4.11. The name of the proteins cannot be seen since inflammation network has intense protein interactions and high number of interactors. After constructing the inflammation network, its intersection with sphingolipid weighted identity network was also obtained from CytoScape providing a tool to merge or intersect networks. The intersection of both metabolisms is also illustrated in Figure 4.12.

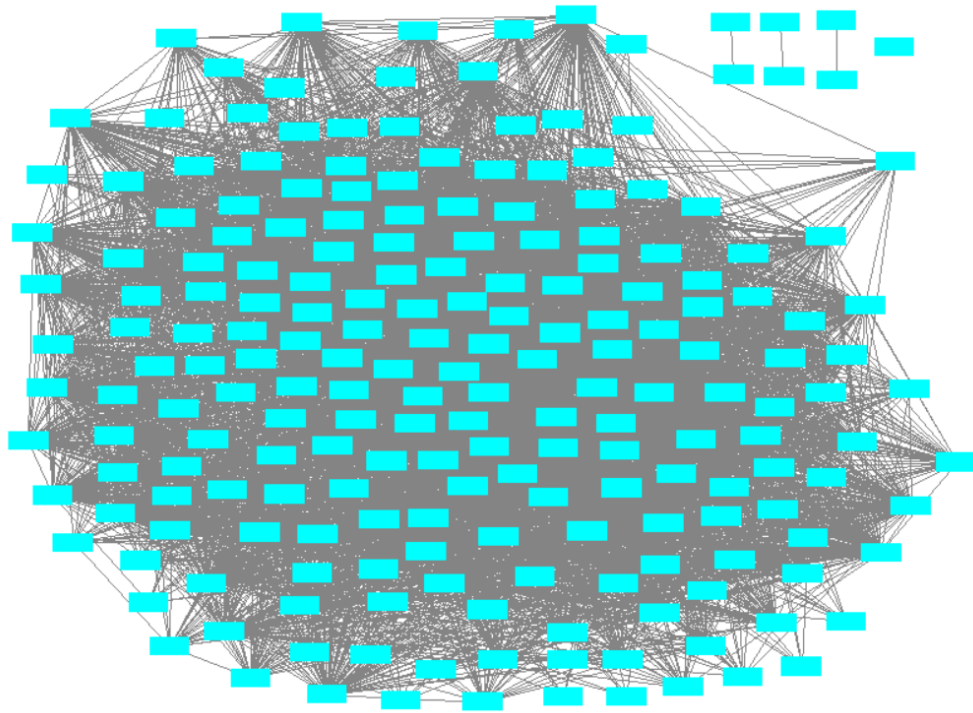


Figure 4.11. WIN of inflammation.

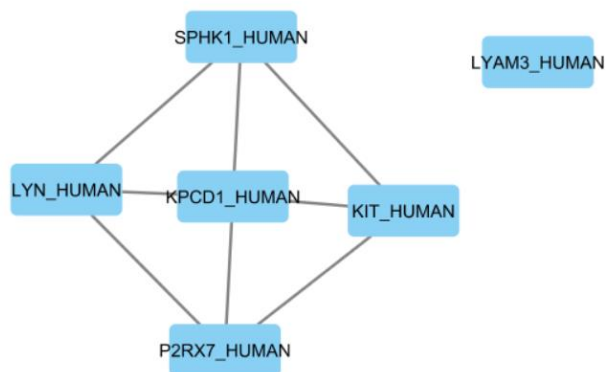


Figure 4.12. The intersection of both networks is illustrated.

4.3.3. Interactor Analysis for Intersection of Sphingolipids and Inflammation Proteins

The ligand set of the intersection network was analysed to understand pairwise ligand similarity (PLS). The total number of unique ligands are 685, 18 of which are commercial drugs searched in DrugBank to detect their diseases (Table 4.17). All the drugs were interacted with only KIT, LYN and KPCD1 proteins. Almost all of them are used for the treatment of cancers; however, there are only 2 drugs curing infection, namely Ruxolitinib and Tafocitinib. KIT, LYN and KPCD1 interacted with these infectious drugs. First, these ligands were imported into CANVAS to analyze their pairwise ligand similarity (PLS). Based on the similarity matrix, %0.18 of these ligand pairs have PLS above 0.7. Among these ligand pairs, there are only two drugs; namely Lapatinib and Pazopanib. These drugs are used for the treatment of breast cancer and kidney cancer, respectively [79,80]. These drugs were interacted with KIT, LYN and KPCD1 proteins. The PLS distribution of them is also illustrated and given in Figure 4.13.

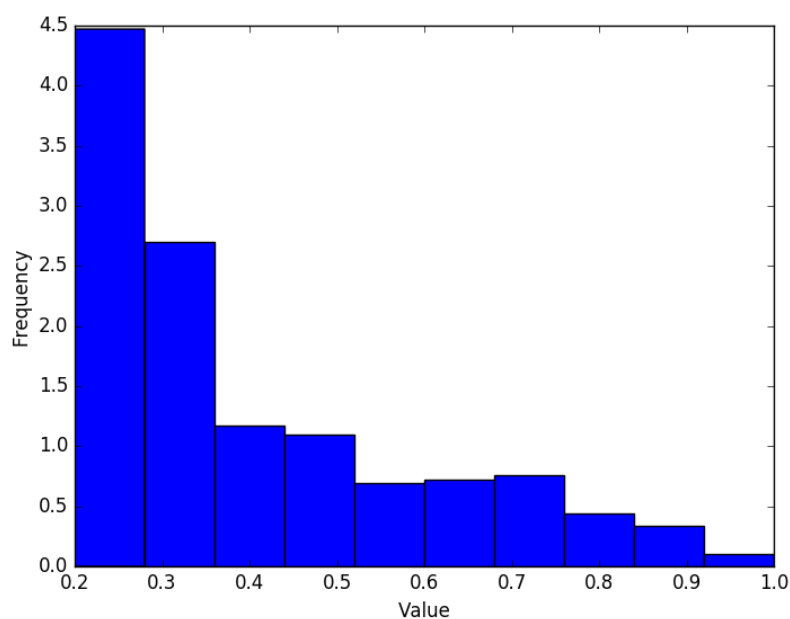


Figure 4.13. The frequency distribution of PLS values for Inflammation ligand set.

Table 4.17. 19 commercial drugs were observed in sphingolipid-inflammation intersection network.

Drug Name	ChEMBL ID	Diseases
Afatinib	CHEMBL1173655	Blocking cancer cell growth
Axitinib	CHEMBL1289926	Renal cell carcinoma
Bosutinib	CHEMBL288441	Leukemia
Crizotinib	CHEMBL601719	Small cell lung cancer
Dasatinib	CHEMBL1421	Leukemia
Erlotinib	CHEMBL553	Pancreatic cancer and lung cancer
Gefitinib	CHEMBL939	Lung cancer
Imatinib	CHEMBL941	Leukemia
Lapatinib	CHEMBL554	Breast cancer
Nilotinib	CHEMBL255863	Leukaemia
Nintedanib	CHEMBL502835	Idiopathic Pulmonary Fibrosis
Pazopanib	CHEMBL477772	Kidney cancer
Ruxolitinib	CHEMBL1789941	Infection
Sirolimus	CHEMBL413	Lymphoma and skin cancer
Sorafenib	CHEMBL1336	Thyroid cancer, cell cancer
Sunitinib	CHEMBL535	Cell cancer
Tofacitinib	CHEMBL221959	Infection
Vandetanib	CHEMBL24828	Thyroid cancer

5. CONCLUSION

5.1. Conclusions

With this study, protein-ligand interaction data collection tool has been developed. For five major ligand databases, namely BindingDB, PDB, PubChem, KEGG, and ChEMBL scripts were coded. For all the scripts, UniProt database was used as protein database. First, the keywords or identifiers of a selected protein family are searched; then their ligands are collected. The output includes protein ID, ligand ID and SMILES which are used as the basis for cheminformatic studies such as chemical similarity analysis. The output can be used with the ligand centric network model, developed previously in our group. By only searching the selected keywords or identifiers of a protein family, dataset and its network models can be constructed.

Among these five ligand databases; BindingDB, PDB, PubChem, KEGG, and ChEMBL, ChEMBL database is the most preferable database for protein – ligand interactions. PDB only consists of ligands with crystal structures and all entries of PDB are also stored in ChEMBL and PubChem databases. ChEMBL collects data from medicinal chemistry journals and BindingDB collects binding data from chemical biology journals. BindingDB shares its data with ChEMBL and has less amount of data than ChEMBL has. KEGG is a genomics knowledgebase. It provides gene, protein and ligand information of biological pathways with less amount of data than ChEMBL and PubChem consists. PubChem stores four types of ligands based on their binding affinity kinds such as unspecified, insufficient, inactive and active. For protein – ligand interactions are evaluated via their biologically active ligands and their binding affinities. ChEMBL provides all the active ligand data PubChem has. ChEMBL also consists of ligand data from both commercial and academic studies.

Thus, selecting ChEMBL as the source database for the protein-ligand interaction studies is reasonable.

As a case study, sphingolipid and insulin metabolism proteins were collected and examined with this tool. Sphingolipids are crucial for membrane structure and they are implicated in the treatment of cancer and neurodegenerative disorders. The deficiencies in sphingolipid proteins cause diseases such as Alzheimer's, Parkinson's and MS diseases. Sphingolipid proteins also interact with insulin proteins. Insulin signaling pathway plays a role in insulin resistance, diabetes, and cancer. Lipid rafts enriched by lipids and cholesterol, are required for the cell signaling of insulin proteins. Thus, the deficiencies in sphingolipid proteins cause neurodegenerative disorders and these disorders affect the insulin signaling pathway. In this study, first the sphingolipid protein interactions and then the sphingolipid – insulin protein interactions were analyzed to highlight bridging interactions between both metabolisms.

First, the protein – ligand interactions of sphingolipid and insulin metabolisms were collected from ChEMBL. Then, the ligand centric weighted identity networks were constructed for the sphingolipid network, then the combined sphingolipid and insulin network. Based on these networks, key interactions and interactor drugs of sphingolipid protein pairs and sphingolipid – insulin protein pairs were detected. For all the interactions, proteins, drugs and related diseases were listed. Besides the promiscuous drugs in the sphingolipid network, Tamoxifen and Alzetamine cancer drugs interacted with GLCM, ARSA and AGAL. Fingolimod, used for the treatment of MS, interacted with SPHK1 and SPHK2. Ceritinib, used for the treatment of non-small cell lung cancer, and Niclosamide, curing worm infection were interacted with KIT and LYN. In sphingolipid – insulin interactions, Ceritinib interacted also with IGF1R and INSR proteins.

5.2. Further Studies

We collected and analyzed sphingolipid and insulin metabolisms to highlight important interactions. Based on the detected specific protein pairs and their interactors, more than just listing interactions and their drugs, the role of specific protein pairs in diseases aimed to be biologically analyzed. Molecular dynamics analysis and experimental confirmations.

REFERENCES

1. H. M. Berman *et al.*, “The Protein Data Bank.,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.
2. T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, “BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities,” *Nucleic Acids Research*, vol. 35, no. SUPPL. 1, pp. 198–201, 2007.
3. A. P. Bento *et al.*, “The ChEMBL bioactivity database: An update,” *Nucleic Acids Research*, vol. 42, no. D1, pp. 1083–1090, 2014.
4. R. Apweiler *et al.*, “UniProt: the Universal Protein knowledgebase.,” *Nucleic Acids Research*, vol. 32, no. Database issue, pp. D115-9, 2004.
5. A. Brown, “XML in serial publishing: Past, present and future,” *OCLC Systems and Services*, vol. 19, no. 4, pp. 149–154, 2003.
6. T. Kolter and K. Sandhoff, “Sphingolipid metabolism diseases,” *Biochimica et Biophysica Acta - Biomembranes*, vol. 1758, no. 12, pp. 2057–2079, 2006.
7. C. Christmas, Rowan; Avila-Campillo, Iliana; Bolouri, Hamid; Schwikowski, Benno; Anderson, Mark; Kelley, Ryan; Landys, Nerius; Workman, Chris; Ideker, Trey; Cerami, Ethan; Sheridan, Rob; Bader, Gary D.; Sander, “Cytoscape: a software environment for integrated models of biomolecular interaction networks,” *American Association for Cancer Research Education Book*, no. Karp 2001, pp. 12–16, 2005.

8. J. Duan, S. L. Dixon, J. F. Lowrie, and W. Sherman, "Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods," *Journal of Molecular Graphics and Modelling*, vol. 29, no. 2, pp. 157–170, 2010.
9. P. L. Jernigan *et al.*, "Sphingolipids in Major Depression," *NeuroSignals*, vol. 23, no. 1, pp. 49–58, 2015.
10. D. I. Kuzmenko and T. K. Klimentyeva, "Role of ceramide in apoptosis and development of insulin resistance," *Biochemistry*, vol. 81, no. 9, pp. 913–927, 2016.
11. S. Kim *et al.*, "Targeting cancer metabolism by simultaneously disrupting parallel nutrient access pathways," *Journal of Clinical Investigation*, vol. 126, no. In Press, 2016.
12. M. Park *et al.*, "A Role for Ceramides, but NOT Sphingomyelins, as antagonists of insulin signaling and mitochondrial metabolism in C2C12 myotubes," *Journal of Biological Chemistry*, p. jbc.M116.737684, 2016.
13. C. R. Gault, L. M. Obeid, and Y. A. Hannun, "An overview of sphingolipid metabolism: From synthesis to breakdown," *Advances in Experimental Medicine and Biology*, vol. 688, pp. 1–23, 2010.
14. E. Posse de Chaves and S. Sipione, "Sphingolipids and gangliosides of the nervous system in membrane function and dysfunction," *FEBS Letters*, vol. 584, no. 9, pp. 1748–1759, 2010.

15. R. Gerl and D. L. Vaux, "Apoptosis in the development and treatment of cancer," *Carcinogenesis*, vol. 26, no. 2, pp. 263–270, 2005.
16. P. E. P. E. Bickel, "Lipid rafts and insulin signaling," *American Journal of Physiology- Endocrinology And Metabolism*, vol. 282, no. 1, pp. E1–E10, 2002.
17. M. Fuller, "Sphingolipids: the nexus between Gaucher disease and insulin resistance.," *Lipids in Health and Disease*, vol. 9, no. 1, p. 113, 2010.
18. H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.
19. S. Kim *et al.*, "PubChem substance and compound databases," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1202–D1213, 2016.
20. E. Christensen, F. Curbera, G. Meredith, and S. Weerawarana, "Web services description language (WSDL) 1.1." 2001.
21. D. A. Fernandes de Abreu *et al.*, "An Insulin-to-Insulin Regulatory Network Orchestrates Phenotypic Specificity in Development and Physiology," *PLoS Genetics*, vol. 10, no. 3, pp. 17–19, 2014.
22. P. Durek and D. Walther, "The integrated analysis of metabolic and protein interaction networks reveals novel molecular organizing principles.," *BMC Systems Biology*, vol. 2, p. 100, 2008.

23. V. S. Rao, K. Srinivas, G. N. Sujini, and G. N. S. Kumar, "Protein-Protein Interaction Detection: Methods and Analysis," *International Journal of Proteomics*, vol. 2014, no. ii, pp. 1–12, 2014.
24. M. Fenger, A. Linneberg, and J. Jeppesen, "Network-based analysis of the sphingolipid metabolism in hypertension," *Frontiers in Genetics*, vol. 5, no. FEB, pp. 1–14, 2015.
25. J. H. Morris *et al.*, "clusterMaker: a multi-algorithm clustering plugin for Cytoscape.," *BMC Bioinformatics*, vol. 12, no. 1, p. 436, 2011.
26. D. J. Rogers and T. T. Tanimoto, "A Computer Program for Classifying Plants.," *Science*, vol. 132, no. 3434, pp. 1115–8, 1960.
27. J. D. Hunter, "Matplotlib: A 2D graphics environment," *Computing in Science and Engineering*, vol. 9, no. 3, pp. 99–104, 2007.
28. D. M. D. System and D. E. S. Research, "Schrödinger Maestro." Schrödinger, New York.
29. T. *et al.* Madhavi Sastry, G., Adzhigirey, M., Day, "Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments," *Journal of Computer Aided Molecular Design*, 2013.
30. R. A. Friesner *et al.*, "Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes," *Journal of Medicinal Chemistry*, vol. 49, no. 21, pp. 6177–6196, 2006.

31. N. M. Davies and X. Wei Teng, "Importance of Chirality in Drug Therapy and Pharmacy Practice: Implications for Psychiatry," *Advances in Pharmacy*, vol. 1, no. 3, pp. 242–252, 2003.
32. C. Robert, *Molecular Symmetry and Group Theory*. New York, 1998.
33. D. H. Johnston, "Symmetry Resources at Otterbein College," 2008.
34. A. Herraes, "Biomolecules in the computer: Jmol to the rescue," *Biochemistry and Molecular Biology Education*, vol. 34, no. 4, pp. 255–261, 2006.
35. S. F. Sousa *et al.*, "Protein-ligand docking in the new millennium--a retrospective of 10 years in the field.," *Current Medicinal Chemistry*, vol. 20, no. 18, pp. 2296–314, 2013.
36. A. M. Gallina, P. Bisignano, M. Bergamino, and D. Bordo, "PLI: A web-based tool for the comparison of protein-ligand interactions observed on PDB structures," *Bioinformatics*, vol. 29, no. 3, pp. 395–397, 2013.
37. J. C. Fuller, M. Martinez, S. Henrich, A. Stank, S. Richter, and R. C. Wade, "LigDig: A web server for querying ligand-protein interactions," *Bioinformatics*, vol. 31, no. 7, pp. 1147–1149, 2015.
38. S. Kim, P. A. Thiessen, E. E. Bolton, and S. H. Bryant, "PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem," *Nucleic Acids Research*, p. gkv396, 2015.

39. M. Murphy, G. Brown, and C. Wallin, "Entrez Gene Help : Integrated Access to Genes of Genomes in the Reference Sequence Collection," *Gene Help*, no. Md, pp. 1–49, 2006.
40. S. Divakar, "Enzymatic transformation," *Enzymatic Transformation*, vol. 9788132208, pp. 1–284, 2013.
41. K. Hata *et al.*, "Limited inhibitory effects of oseltamivir and zanamivir on human sialidases," *Antimicrobial Agents and Chemotherapy*, vol. 52, no. 10, pp. 3484–3491, 2008.
42. D. S. Wishart *et al.*, "DrugBank: A knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, no. SUPPL. 1, pp. 901–906, 2008.
43. K. J. Smith and M. Germain, "Leflunomide: an immune modulating drug that may have a role in controlling secondary infections with review of its mechanisms of action.," *Journal of Drugs in Dermatology*, vol. 14, no. 3, pp. 230–236, 2015.
44. L. M. Parker, H. A. Damanhuri, S. P. S. Fletcher, and A. K. Goodchild, "Hydralazine administration activates sympathetic preganglionic neurons whose activity mobilizes glucose and increases cardiovascular function," *Brain Research*, vol. 1604, pp. 25–34, 2015.
45. L. M. Al-Harbi, E. H. El-Mossalamy, A. Y. Obaid, and M. A. El-Ries, "Thermal Decomposition of Some Cardiovascular Drugs (Telmisartane, Cilazapril and Terazosin HCL)," *American Journal of Analytical Chemistry*, vol. 4, no. 7, p. 337, 2013.

46. M. C. Chappell, "Amiloride and the diabetic kidney," *Journal of Hypertension*, vol. 34, no. 8, pp. 1500–1501, 2016.
47. M. Dold, M. T. Samara, C. Li, M. Tardy, and S. Leucht, "Haloperidol versus first generation antipsychotics for the treatment of schizophrenia and other psychotic disorders," *The Cochrane Library*, 2015.
48. N. J. Talley *et al.*, "Effect of amitriptyline and escitalopram on functional dyspepsia: a multicenter, randomized controlled study," *Gastroenterology*, vol. 149, no. 2, pp. 340–349, 2015.
49. S. Leucht *et al.*, "Sixty Years of Placebo-Controlled Antipsychotic Drug Trials in Acute Schizophrenia: Systematic Review, Bayesian Meta-Analysis, and Meta-Regression of Efficacy Predictors," *American Journal of Psychiatry*, p. appi-ajp, 2017.
50. D. J. Kuter *et al.*, "Miglustat therapy in type 1 Gaucher disease: clinical and safety outcomes in a multicenter retrospective cohort study," *Blood Cells, Molecules, and Diseases*, vol. 51, no. 2, pp. 116–124, 2013.
51. J. Bell, "Amlexanox for the treatment of recurrent aphthous ulcers," *Clinical Drug Investigation*, vol. 25, no. 9, pp. 555–566, 2005.
52. E. B. C. T. C. Group, "Aromatase inhibitors versus tamoxifen in early breast cancer: patient-level meta-analysis of the randomised trials," *The Lancet*, vol. 386, no. 10001, pp. 1341–1352, 2015.

53. V. Talwar, V. Goel, S. Raina, N. Patnaik, and D. C. Doval, "Altretamine in advanced pretreated epithelial ovarian carcinoma patients: Experience from a center in north India," *Current Medicine Research and Practice*, vol. 6, no. 3, pp. 109–112, 2016.
54. T. Gwisai *et al.*, "Repurposing niclosamide as a versatile antimicrobial surface coating against device-associated, hospital-acquired bacterial infections," *Biomedical Materials*, 2017.
55. A. T. Shaw *et al.*, "Ceritinib in ALK-rearranged non-small-cell lung cancer," *New England Journal of Medicine*, vol. 370, no. 13, pp. 1189–1197, 2014.
56. L. Kappos, "Oral fingolimod (FTY720) for relapsing multiple sclerosis," *New England Journal of Medicine*, vol. 355, pp. 1124–1140, 2006.
57. B. Oskouian and J. D. Saba, "Cancer treatment strategies targeting sphingolipid metabolism," *Advances in Experimental Medicine and Biology*, vol. 688, pp. 185–205, 2010.
58. J. J. Irwin and B. K. Shoichet, "ZINC - A free database of commercially available compounds for virtual screening," *Journal of Chemical Information and Modeling*, vol. 45, no. 1, pp. 177–182, 2005.
59. H. E. Pence and A. Williams, "ChemSpider: an online chemical information resource." *ACS Publications*, 2010.
60. W. B. Bowne, J. Michl, M. H. Bluth, M. E. Zenilman, and M. R. Pincus, "Treatment of Cancer," *Cancer Therapy*, vol. 5B, no. 718, pp. 331–344, 2007.

61. K. M. Kasiotis, E. N. Tzanetou, and S. A. Haroutounian, "Pyrazoles as potential anti-angiogenesis agents: a contemporary overview," *Frontiers in Chemistry*, vol. 2, no. September, pp. 1–7, 2014.
62. J. Hert, J. J. Irwin, C. Laggner, M. J. Keiser, and B. K. Shoichet, "Quantifying biogenic bias in screening libraries," *Nature Chemical Biology*, vol. 5, no. 7, pp. 479–483, 2009.
63. J. von Pawel *et al.*, "Randomized phase III trial of amrubicin versus topotecan as second-line treatment for patients with small-cell lung cancer," *Journal of Clinical Oncology*, vol. 32, no. 35, pp. 4012–4019, 2014.
64. Y. H. Park *et al.*, "Lesser Toxicities of Belotecan in Patients with Small Cell Lung Cancer: A Retrospective Single-Center Study of Camptothecin Analogs," *Canadian Respiratory Journal*, vol. 2016, 2016.
65. P. Blanchard *et al.*, "Chemotherapy and radiotherapy in nasopharyngeal carcinoma: an update of the MAC-NPC meta-analysis," *The Lancet Oncology*, vol. 16, no. 6, pp. 645–655, 2015.
66. Q. Li, M. A. Ali, and J. I. Cohen, "Insulin Degrading Enzyme Is a Cellular Receptor Mediating Varicella-Zoster Virus Infection and Cell-to-Cell Spread," *Cell*, vol. 127, no. 2, pp. 305–316, 2006.
67. K. Valis *et al.*, "Hippo/Mst1 stimulates transcription of the proapoptotic mediator NOXA in a FoxO1-dependent manner," *Cancer Research*, vol. 71, no. 3, pp. 946–954, 2011.

68. S. Bhattacharyya, L. Feferman, and J. K. Tobacman, "Inhibition of phosphatase activity follows decline in sulfatase activity and leads to transcriptional effects through sustained phosphorylation of transcription factor MITF," *PLoS One*, vol. 11, no. 4, pp. 1–22, 2016.
69. M. K. Paul and A. K. Mukhopadhyay, "Tyrosine kinase - Role and significance in Cancer.," *International Journal of Medical Sciences*, vol. 1, no. 2, pp. 101–115, 2004.
70. A. Belfiore and R. Malaguarnera, "Insulin receptor and cancer," *Endocrine-Related Cancer*, vol. 18, no. 4, pp. 125–147, 2011.
71. G. Qin *et al.*, "Palbociclib inhibits epithelial-mesenchymal transition and metastasis in breast cancer via c-Jun/COX-2 signaling pathway," *Oncotarget*, vol. 6, no. 39, pp. 41794–41808, 2015.
72. Y. P. Wu, K. Mizugishi, M. Bektas, R. Sandhoff, and R. L. Proia, "Sphingosine kinase 1/S1P receptor signaling axis controls glial proliferation in mice with Sandhoff disease," *Human Molecular Genetics*, vol. 17, no. 15, pp. 2257–2264, 2008.
73. M. Beran and H. M. Kantarjian, "Topotecan (Hycamptin) and Topotecan Containing Regimens in the Treatment of Hematologic Malignancies," *Annals of the New York Academy of Sciences*, vol. 922, no. 1, pp. 247–259, 2000.
74. A. Fassas and A. Anagnostopoulos, "The use of liposomal daunorubicin (DaunoXome) in acute myeloid leukemia," *Leukemia & Lymphoma*, vol. 46, no. 6, pp. 795–802, 2005.

75. D. Plouffe *et al.*, “In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 26, pp. 9059–9064, 2008.
76. H. Gylling *et al.*, “Insulin sensitivity regulates cholesterol metabolism to a greater extent than obesity: lessons from the METSIM Study.,” *Journal of Lipid Research*, vol. 51, no. 8, pp. 2422–7, 2010.
77. C. J. Chu *et al.*, “N-oleoyldopamine, a novel endogenous capsaicin-like lipid that produces hyperalgesia,” *Journal of Biological Chemistry*, vol. 278, no. 16, pp. 13633–13639, 2003.
78. S. C. Kang, B. R. Kim, S. Y. Lee, and T. S. Park, “Sphingolipid metabolism and obesity-induced inflammation,” *Frontiers in Endocrinology*, vol. 4, no. JUN, pp. 1–11, 2013.
79. W. M. Linehan, R. Srinivasan, and L. S. Schmidt, “The genetic basis of kidney cancer: a metabolic disease,” *Nature Reviews Urology*, vol. 7, no. 5, pp. 277–285, 2010.
80. C. E. Geyer *et al.*, “Lapatinib plus capecitabine for HER2-positive advanced breast cancer,” *New England Journal of Medicine*, vol. 355, no. 26, pp. 2733–2743, 2006.