

Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE BIOTECNOLOGICHE, BIOCOMPUTAZIONALI,
FARMACEUTICHE E FARMACOLOGICHE

Ciclo XXXIV

Settore Concorsuale: 03/D1 - CHIMICA E TECNOLOGIE FARMACEUTICHE,
TOSSICOLOGICHE E NUTRACEUTICO-ALIMENTARI

Settore Scientifico Disciplinare: CHIM/08 - CHIMICA FARMACEUTICA

APPLICATIONS OF NETWORK-BASED APPROACHES
IN DRUG RESEARCH

Presentata da: Chiara Cabrelle

Coordinatore Dottorato

Prof.ssa Maria Laura Bolognesi

Supervisore

Prof. Maurizio Recanatini

Esame finale anno 2022

*The knowledge could already be in the network,
but we need to connect the dots to find it.*

Kirk Borne

ACKNOWLEDGMENTS

First and foremost, I am truly grateful to my Supervisor, *Prof. Maurizio Recanatini*, for his precious guidance, encouragement and help during my PhD study. I would like to thank people at Recanatini's lab and, in particular, *Luca* for sharing honors and burdens of our PhD lives. A special thank to all the PhD academic board members and colleagues for giving me the opportunities to grow up as a scientist with our fruitful discussions. I would like to thank very warmly *Prof. Federico Giorgi* and all the people of his lab for their technical support on my study and their friendship. Then, but not in order of relevance, a very special thanks and a big hug to my family, to my parents *Gianni* and *Stefania* without whom I couldn't be who I am and to *Matteo* who has been supporting and enduring me for a long time. And since it is impossible to feel at home without friends, I can only express a huge <3 to *Erika, Luca, Daniele, Laura, Nicola, Riccardo* and *Marco*.

ABSTRACT

**APPLICATIONS OF NETWORK-BASED APPROACHES IN DRUG
RESEARCH**

CHIARA CABRELLE

Recent research trends in computer-aided drug design have shown an increasing interest towards the implementation of advanced cross-field approaches able to deal with large amount of data as network-based and machine learning methods. This demand arose from the awareness of the complexity of biological systems and from the availability of data provided by high-throughput technologies. As a consequence, drug research has embraced this paradigm shift that has occurred with systems pharmacology which exploits approaches such as that based on networks. From this concept, this thesis is built around the application of network-based approaches in drug research.

Indeed, the process of drug discovery can benefit from the implementation of network-based approaches at different steps from target identification to drug repurposing. From this broad range of opportunities, this thesis is focused on the following three main topics:

- Chemical space networks (CSNs), which are generally designed to represent and characterize bioactive compound data sets; CSNs have been used for SAR visualization and analysis;
- Drug-target interactions (DTIs) prediction through a network-based algorithm for predicting missing links; this method found application in drug repurposing as well as target identification;
- COVID-19 drug research, which was explored implementing COVIDrugNet, a network-based tool for COVID-19 related drugs, realized by our computational medicinal chemistry group, to contribute to tackle an actual health problem.

The main highlight emerged from this thesis is that network-based approaches can be considered useful methodologies to tackle different issues in drug research. In detail, CSNs are valuable coordinate-free, graphically accessible representations of the structure-activity relationships of bioactive compounds data sets especially for medium-large libraries of molecules. DTIs prediction through the *random walk with restart* algorithm on

heterogeneous networks can be a helpful method for target identification. COVIDDrugNet highlights the potential of network-based approaches for studying drugs related to a specific condition, i.e., COVID-19, and the same ‘systems-based’ approaches can be used for other diseases.

To conclude, network-based tools are proving to be suitable in many applications in drug research and provide the opportunity to model and analyze diverse drug-related data sets, even large ones, also integrating different multi-domain information.

TABLE OF CONTENTS

BACKGROUND	1
1 TOWARDS A NEW PARADIGM FOR DRUG DISCOVERY	1
2 NETWORK SCIENCE	2
2.1 NETWORK APPLICATIONS IN DRUG DISCOVERY	2
2.2 BASIC CONCEPTS OF NETWORKS	2
2.2.1 Definition of network and network types	3
2.2.2 Data structures	5
2.2.3 Network analysis	5
2.2.3.1 Density and sparsity	6
2.2.3.2 Degree and degree distribution	6
2.2.3.3 Degree assortativity	7
2.2.3.4 Subnetworks	8
2.2.3.5 Paths	9
2.2.3.6 Connectivity and components	9
2.2.3.7 Clustering coefficient	9
2.2.4 Network layouts	10
3 THE ANTIPROLIFERATIVE COMPOUND SET	10
4 TARGETING SIGNALING PATHWAYS IN LEUKEMIA	11
5 AIM OF THE STUDY	13
CHAPTER 1. CHEMICAL SPACE NETWORKS	15
6 INTRODUCTION	15
6.1 CHEMINFORMATICS	15
6.1.1 Representation of molecular structures	16
6.1.2 Molecular fingerprints	16
6.1.3 Similarity measures	17
6.2 CHEMICAL SPACE	18
6.2.1 Coordinate-based vs coordinate free representations	19
6.2.2 Chemical space networks	20
7 METHODS	22
7.1 DESIGN OF THR-CSNS	22
7.1.1 Curation of data	22
7.1.2 Similarity assessment	22
7.1.3 THR-CSN construction	23

7.1.4	Characterization of THR-CSNs	24
7.1.4.1	Network density	24
7.1.4.2	Global network properties	24
8	RESULTS AND DISCUSSION	25
8.1	DESIGN OF CSNS	25
8.2	CHARACTERIZATION OF THR-CSNS	26
8.3	COMPARISON OF THR-CSNS	29
8.4	EXPLORATION OF SARs THROUGH CSNS VISUALIZATION	30
9	CONCLUSIONS	34
<u>CHAPTER 2. DRUG-TARGET INTERACTIONS</u>		37
10	INTRODUCTION	37
10.1	DRUG-TARGET NETWORKS	37
10.2	DTIS PREDICTIONS	39
10.3	DATABASES	40
10.3.1	DrugBank	40
10.3.2	KEGG	41
11	METHODS	43
11.1	DATA COLLECTION	43
11.1.1	Data collection from DrugBank	43
11.1.2	Data collection from KEGG	44
11.2	DATA PREPARATION	44
11.2.1	Target sequence similarity matrix	44
11.2.2	Compound structure similarity matrix	45
11.2.3	Adjacency matrix	45
11.2.4	Additional similarities matrices	45
11.3	RWR ALGORITHM ON HETEROGENEOUS NETWORK	46
11.4	OVER-REPRESENTATION ANALYSIS	47
12	RESULTS AND DISCUSSION	48
12.1	EVALUATING THE PERFORMANCE OF THE ALGORITHM IN PREDICTING MISSING LINKS	49
12.2	PREDICTION FROM KEGG DATABASE	50
12.3	PREDICTION FROM DRUGBANK DATABASE	56
13	CONCLUSIONS	62
<u>CHAPTER 3. NETWORK-BASED APPROACHES FOR COVID-19 DRUG RESEARCH</u>		63
14	INTRODUCTION	63

15	COVID-19 DRUGS NETWORKER	63
16	CONCLUSIONS	67
	<u>REFERENCES</u>	<u>69</u>
	<u>SUPPLEMENTARY MATERIALS</u>	<u>79</u>

LIST OF ACRONYMS

ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
ATC	Anatomical Therapeutic Chemical code
CML	Chronic Myeloid Leukemia
CSN	Chemical Space Network
DTI	Drug-Target Interaction
DTN	Drug-Target Network
ECFP	Extended-Connectivity Fingerprint
GO	Gene Ontology
HSC	Hematopoietic Stem Cell
IC ₅₀	half maximal Inhibitory Concentration
JAK	Janus-activated kinase
KEGG	Kyoto Encyclopedia of Genes and Genomes
LSC	Leukemia Stem Cells
MAPK	Mitogen-Activated Protein Kinase
MCS-CSN	Maximum Common Substructure- Chemical Space Network
MMP-CSN	Matched Molecular Pairs-Chemical Space Network
NBI	Network-Based Inference
NRWRH	Network-based Random Walk with Restart on Heterogeneous network
PI3K	PhosphoInositide 3-Kinase
R&D	Research And Development
RWR	Random Walk with Restart
SAR	Structure-Activity Relationship
SDF	Structure Data File
SMILES	Simplified Molecular Input Line Entry Specification
STAT	Signal Transducer and Activator of Transcription
THR-CSN	THreshold-Chemical Space Network
TKI	Tyrosine Kinase Inhibitor
TV-CSN	Tversky similarity- Chemical Space Network

Background

1 TOWARDS A NEW PARADIGM FOR DRUG DISCOVERY

In the last decades, drug discovery has seen important changes in its grounds. Among the reasons that led to them, it must be included the attrition rate in the late stage of clinical trials and the consequent drop in pharmaceutical R&D productivity [1], [2]. It has been suggested that this failure can be partly due to the classic paradigm of drug discovery according to which the best drug candidate is that with a high selectivity and potency toward the desired target hypothesized to cause the disease. However, the *one gene, one drug, one disease* paradigm was challenged by the evidence gained through systems biology and by *omics* disciplines (genomics, transcriptomics, metabolomics and proteomics) [1]. Systems biology studies the interactions between biological entities at different levels (organism, cells, tissues, regulatory networks and molecular pathways) to understand the complex biology underlying physiological and pathological states. The advances in this discipline were made possible by the innovative technologies that provide the starting material for the *omics* sciences and the development of mathematical and computational models to analyze these large-scale data [3]. Such a progress set out the basis for the corresponding developments in drug discovery: systems pharmacology.

From a broad perspective, the improved understanding about the complexity of biological systems need to be reflected in drug discovery, hence the emergence of systems pharmacology. In fact, (quantitative) systems pharmacology aims at identifying and validating targets as well as at elucidating therapeutic and toxic effects of drugs on cellular networks, through computational models that integrate multiple entities (biomolecules, cells, etc.) interacting at several temporal and spatial scales [4].

Indeed, it has been demonstrated that complex diseases, i.e., cancer and central nervous system diseases, are caused by a deregulated network of proteins, not only a single target. The increasing interest in the design of drug candidates that interact with multiple targets, known as polypharmacology [5], is one of the applications driven by this new thinking on drug discovery, that complements the more traditional paradigm of target-based drug design. Hopkins defined network pharmacology as “the next paradigm in drug discovery” [1].

2 NETWORK SCIENCE

The essential intuition underlying a network is that of elements interconnected by links that can represent different relationships among the elements. When representing a real system through a network, it is crucial to find the best representation to model our data that allows addressing the goal of the study. In other words, network science provides a useful framework to represent interrelations. However, it is fundamental to know the basis of network theory to choose the network model that best fits the data. An introduction to network science is provided in the following paragraphs, since a general background helps understanding the state of the art of network-based applications in drug research.

The terminology about networks derives from graph theory, a branch of mathematics. Therefore, *network* and *graph* are often used as synonyms, even though imprecisely because with *graph* we are referring to the mathematical model under the network itself. The foundation of graph theory can be dated back to 1741 with the Leonhard Euler's problem of the seven bridges of Königsberg. He modeled a real problem, that is if it is possible to walk through the city of Königsberg crossing each bridge only once, through a network approach [6]. Since then, network science has seen a rapid expansion in different fields from mathematics, physics and computer science to biology, sociology, finance, etc.

Why are networks everywhere? Networks are a versatile and powerful tool for modeling interactions; for this reason, they are so widely employed that they pervade our everyday life. Networks have these extensive applications across disciplines because they are a useful tool to model, analyze and visualize real world systems which are intrinsically complex.

2.1 *Network applications in drug discovery*

Focusing on networks in pharmacology, a list of different applications over the entire drug design process can be drawn up. Indeed, network can be applied for target identification, drug repurposing, drug combinations and drug adverse effects predictions. Drug-target networks (DTNs) and drug-target interactions (DTIs) prediction through network-based approaches are useful for target identification and drug repurposing studies, whereas for lead search and optimization, chemical space networks (CSNs) can be explored for structure-activity relationship (SAR) visualization and analysis.

2.2 *Basic concepts of networks*

Hereafter, an overview of the most relevant concepts in network theory is reported. In particular, different types of networks, the underlying data formats as well as important

features and properties of networks are defined in order to set the ground for the drug research-related applications in the subsequent chapters of this thesis.

2.2.1 Definition of network and network types

A network (or graph) G is a set V of elements, known as *nodes* or *vertices*, together with a set E of connections between pairs of nodes, defined as *edges* or *links* representing the interactions among them. Therefore, a network is formally expressed as:

$$G = (V, E) \quad (1)$$

In graph theory, all nodes adjacent to node i are its *neighbors*, that constitute the neighbor set.

Focusing on the pair of nodes i and j , they are called *adjacent* nodes if joined by an edge:

$$e \in E, e = \{i, j\} \quad (2)$$

with $i, j \in V$.

A network is characterized by some properties as its number of nodes N that defines the *size* of the network, and the total number of edges L .

According to this description, two nodes i and j could be joined by multiple edges forming a set of edges linking i and j . Networks in which there are no *multiple edges* (or multi-edges) – more than one link between a pair of nodes – or *self-loops* – edges connecting a node to itself – are termed *simple* networks. If each node is connected to all the other nodes, the network is a *complete* graph.

Based on the type of links, it is possible to describe different classes of networks. In a network, links can be undirected or directed defining *undirected* or *directed* networks, respectively.

A network $G = (V, E)$, in which an edge $e = \{i, j\}$ is defined by an unordered pair of vertices, is termed as *undirected network*. In this case, links are bi-directional, and the order of the nodes in the pair is irrelevant.

In a *directed network* or *digraph*, a link is defined by an ordered pair of nodes and it goes from the source node i to the target node j reflecting its direction that is graphically represented by an arrow pointing to the target node. Directed graphs are commonly used to represent systems involving sequential interactions between the elements, like gene regulatory networks in systems biology.

To model the intensity of interactions, it can be useful to attribute *weights* to links. Therefore, the link joining nodes i and j is described as (i, j, w_{ij}) in which w_{ij} is a real number expressing its weight. This type of network is defined as *weighted* network. A weighted network can be

undirected or directed. When displaying a weighted network, the links are generally represented as lines of different width as in Figure 1.

Networks might have two different types of nodes in a way that the set of nodes V is partitioned into two subsets V_1 and V_2 , such that each edge connects only nodes of different type. Networks of this class are termed *bipartite* networks and are widely used in drug discovery. The analysis of bipartite networks is so complex that it is common practice to compress their information, by projecting it in the corresponding two monopartite networks, despite the possible loss of information. A projection contains only one type of node, and those nodes are connected only if they share at least one common neighbor in the bipartite network. An example of a bipartite network is a drug-target network in which drugs and targets represent the two distinct sets of nodes.

Indeed, bipartite networks can be considered a subtype of another network type: *multipartite* networks. A multipartite network is characterized by multiple types of nodes.

For the sake of completeness, *multilayer* networks must be mentioned. Multilayer networks have different types of nodes and links distributed through interconnected layers. However, if each layer contains the same nodes, the multilayer network is called a *multiplex* network. Exemplary networks of different types are shown in Figure 1.

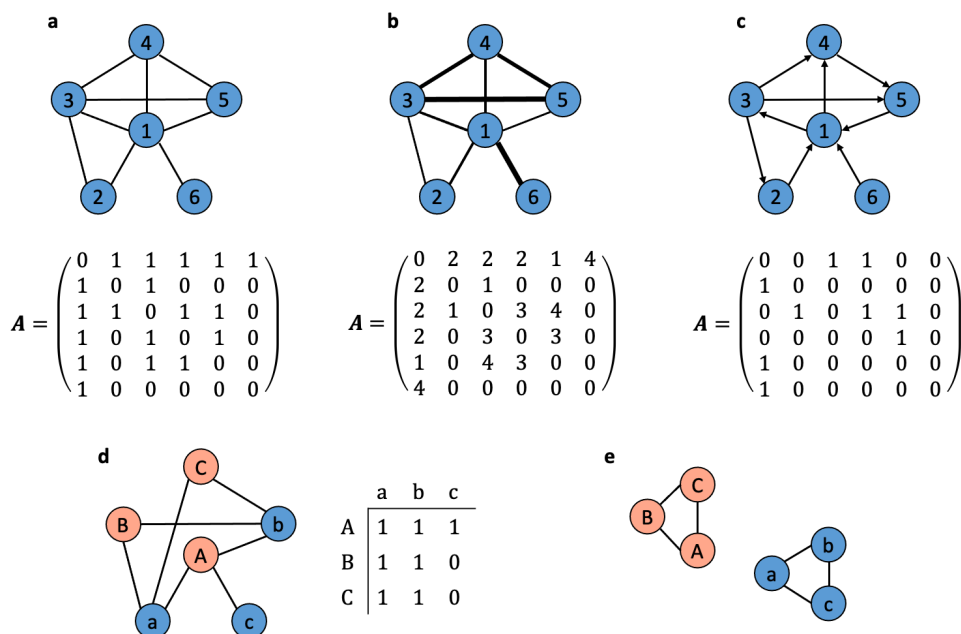


Figure 1. Network types and their adjacency matrices. **A.** an undirected network with its symmetric binary adjacency matrix; **b.** a weighted undirected network in which the width of the edges represents weights; **c.** a directed network with non-symmetric edges; **d.** a bipartite network in which nodes belonging to two different sets are illustrated with different colors and **e.** the corresponding two projections.

2.2.2 Data structures

Beneath the graphical representation of network, different types of data structure are used to store the information about nodes, connections and relative attributes in a computer readable format. Especially for large networks, the choice of data structure is strategic because it has to retain the network data preferably without needlessly consuming storage space.

Among the well-known data structures, it is impossible not to mention the *adjacency matrix*. The adjacency matrix A for a simple undirected network G is a symmetric matrix with N rows and N columns. The presence of a link connecting nodes i and j is recorded with 1, denoted as $A[i, j] = 1$, and $A[i, j] = A[j, i]$ since it is symmetric, otherwise $A[i, j] = 0$ and $A[i, i] = 0$ since no self-loops exists. By definition, in a directed network the ordered pair of nodes describes the link $A[i, j] \neq A[j, i]$, while in a weighted network the *weight* w_{ij} for the link between i and j can assume any real values, not only 1 or 0.

The adjacency matrix of a bipartite network is a $N \times M$ adjacency matrix in which N is number of nodes belonging to one group of nodes and M is the number of nodes of the others.

The adjacency matrix is not always the best choice to store a network particularly in case of very sparse and large networks. In fact, in their adjacency matrix there will be plenty of zeros, thus occupying storage space inefficiently. An alternative data structure is the *edge list*, so intuitively, a two-columns table that keeps in each row the nodes connected by a link. For weighted networks, a third column contains the weights.

Finally, the *adjacency list* is another common format in which each row records the neighbor set of a node.

The last two data structures are more compact representations for large sparse networks due to the fact that only existing links are kept, ignoring the zero of non-linked nodes that are usually stored in the adjacency matrix.

Figure 1 shows the adjacency matrices underlying the graphical representation of different types of networks.

2.2.3 Network analysis

Network theory provides a variety of network properties useful to describe networks and tools that enable to analyze their features. Properties can refer either to the whole network, e.g., the total number of nodes and links and the edge density, or to single nodes, e.g., node degree and centrality measures. Hereafter, the properties relevant for this thesis are defined, however this is not to be considered an exhaustive list.

2.2.3.1 Density and sparsity

Taking into consideration global network properties, one of the most fundamental concepts is density/sparsity. The density of a network is the fraction of links that actually exist. The *complete* graph has the maximum number of links, so it has the maximum density that equals one.

Density d is described by the following formula:

$$d = L/L_{\max} \quad (3)$$

where L is the number of links and L_{\max} is the maximum number of links.

In undirected networks, L_{\max} is given by the following formula:

$$L_{\max} = N(N - 1)/2 \quad (4)$$

where N corresponds to the number of distinct nodes, hence the density d become

$$d = 2L/N(N - 1). \quad (5)$$

In a directed network, L_{\max} corresponds to

$$L_{\max} = N(N - 1) \quad (6)$$

and its density d are calculated as

$$d = L/N(N - 1). \quad (7)$$

Overall, real world networks are usually *sparse* which means that the density is much smaller than 1 since the number of existing links can be order of magnitude smaller than L_{\max} .

2.2.3.2 Degree and degree distribution

One of the first properties to calculate when analyzing a network is the *degree*. The *degree* k_i of the node i is the number of edges connecting it to its neighbors.

In an undirected network $G = (V, E)$, the degree k_i of the node i is the sum of the values in its row or column i of the adjacency matrix A :

$$k_i = \sum_{j=1}^N A[i, j]. \quad (8)$$

Instead, in directed networks, the incoming and the outgoing links of the node i must be considered, hence its *out-degree* k_i^{out} is the number of edges leaving it and its *in-degree* k_i^{in} is the number of edges reaching it. Thus, for the node i , the out-degree is the sum of the values in the i^{th} row:

$$k_i^{out} = \sum_{j=1}^N A[i, j], \quad (9)$$

where the link is directed from i to j , while the in-degree is the sum of the values in the i^{th} column:

$$k_i^{in} = \sum_{i=1}^N A[i, j]. \quad (10)$$

Finally, in a weighted network, the degree of a node can be measured considering the weights or not.

Evaluating node degrees is essential in network analysis because it provides a way to identify *hubs*, i.e., nodes with high degree that assume a pivotal role, and the removal of which disaggregates the network.

Considering the degrees of all nodes in the network, the node *degree distribution* describes the fraction of nodes with degree k , stating the probability that a randomly chosen node has degree k .

Figure 2 illustrates degree and degree distribution for an exemplary network.

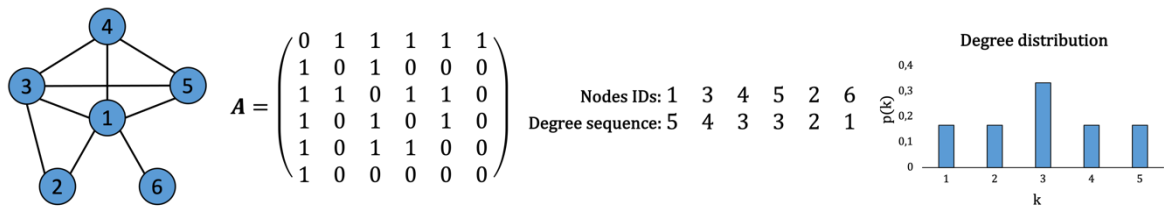


Figure 2. Degree and degree distribution. The figure shows the degree for each node – reported as degree sequence – and the degree distribution for the exemplary network with its adjacency matrix.

2.2.3.3 Degree assortativity

In a network, it is possible that nodes linked among each other tend to have similar features. This is an important property, named *assortativity*. Moreover, assortativity can be due to *homophily*, the tendency of similar nodes to be connected.

Degree assortativity is referred to the property of node degree and it is a measure of how much high-degree nodes tend to be connected to other nodes with high degree, while low-degree nodes tend to be connected to other nodes with low degree.

Networks in which this correlation is verified are called *assortative*. Assortative networks display a *core-periphery structure*. *Disassortative* networks are instead characterized by high-degree nodes connected to low-degree nodes.

The *assortativity coefficient* measures the degree assortativity as the Pearson correlation among degrees of adjacent nodes. The assortativity coefficient values are in the range $[-1, 1]$ with 0 meaning no correlation. Thus, the network is assortative if the assortativity coefficient

is positive, while it is disassortative if the assortativity coefficient is negative. Figure 3 shows exemplary assortative and disassortative networks.

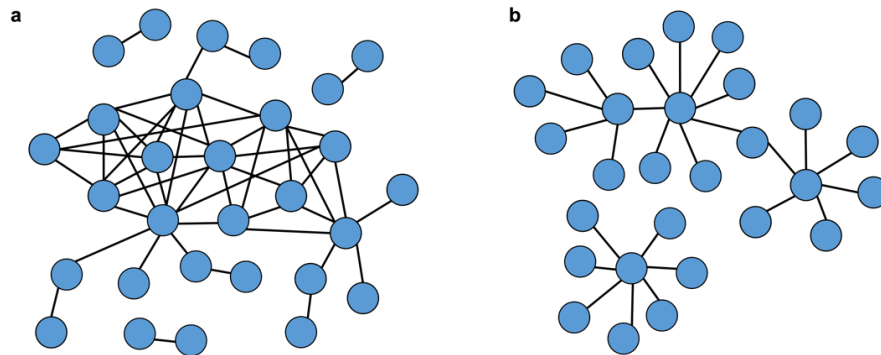


Figure 3. Degree assortativity. The figure shows exemplary **a.** assortative network and **b.** disassortative network.

2.2.3.4 Subnetworks

Sometimes, it can be convenient to focus on a subset of a network: a *subnetwork*. Basically, a subnetwork contains a subset of the nodes of the network as well as all the links connecting these nodes in the initial network. Moreover, if the subnetwork includes all possible edges, such that it is completely connected, it is named *clique*. Examples of subnetwork and clique are shown in Figure 4. The abundance of certain subnetworks is important to characterize real networks.

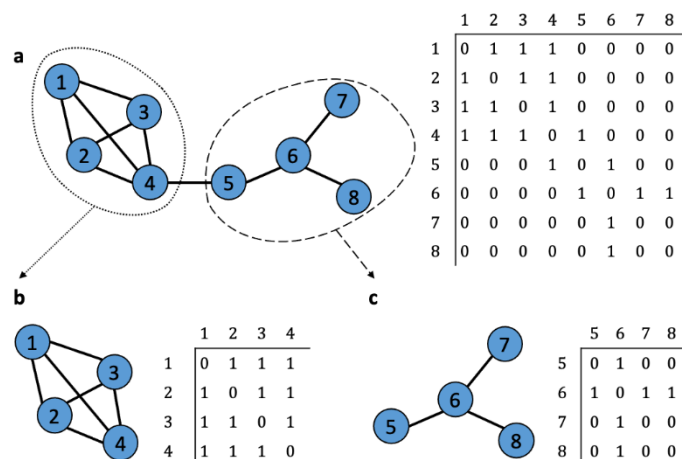


Figure 4. Subnetworks. The figure shows **a.** an exemplary network, **b.** a subnetwork example and **c.** a clique, with their respective adjacency matrix.

2.2.3.5 Paths

Considering a network as a map, it is possible to study the routes that connect a source node to a target node by traversing links. Therefore, a *path* is defined as the sequence of links to cross in order to connect two nodes in a network. The *path length* corresponds to the number of links forming the path. As a result, multiple paths with different path lengths can be traced between two nodes. To study the *distance* among nodes, it is common to refer to the *shortest path* that has the minimum length, by definition, the *shortest path length*. It is possible to find more than one shortest path between two nodes.

Other distance measures are referred to the whole network. Indeed, the *diameter* of the network is the length of the longest shortest path, whereas the *average path length* is the average of the shortest path lengths between all pairs of nodes. Figure 5 depicts the shortest path in different networks.

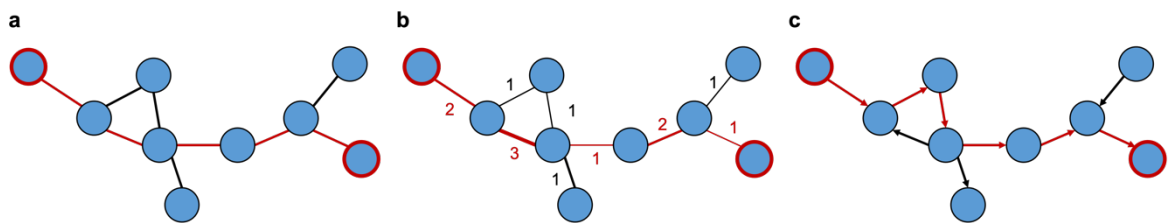


Figure 5. Shortest paths depicted as a sequence of red links among the nodes circled in red in **a.** undirected, **b.** unweighted (undirected) and **c.** directed networks.

2.2.3.6 Connectivity and components

The connectivity of a network relates its structure to its function. In a *connected* network, each node can be reached from any other nodes crossing the network through paths. Instead, a network is *disconnected* if it is composed by more than one connected component. Therefore, a *component* is a subnetwork in which there is a path connecting any pair of its nodes, but it does not exist a path that connects them to other components. Isolated nodes, named *singletons*, belong to their own components. In real networks, the largest component generally contains most nodes and is called the *giant component*.

2.2.3.7 Clustering coefficient

Concerning the local structure of a network, it is important to consider the connectivity among the neighbors of a node. The *clustering coefficient* of a node accounts for how tightly connected the nodes are and it is expressed as the fraction of pairs of neighbors of the node

i that are connected to each other. Moreover, the clustering coefficient for the nodes i is formally defined as:

$$C(i) = \frac{2\tau(i)}{k_i(k_i - 1)} \quad (11)$$

in which $\tau(i)$ is the number of triangles containing i and k_i is its degree.

The clustering coefficient ranges in the interval $0 < C(i) < 1$, and it is equal to 1 in the case of clique.

In addition, the clustering coefficient for the whole network is measured by averaging the clustering coefficient across the nodes as defined by the following formula:

$$C = \frac{\sum_{i:k_i>1} C(i)}{N_{k>1}}. \quad (12)$$

For instance, a network in which many triangles are present has a high clustering coefficient.

2.2.4 Network layouts

Network appearances are generated through layout algorithms that return coordinates for each node of the network. In general, these algorithms attempt to position nodes in order to uniform edge length and minimize edge crossing. Most visualization software and tools (e.g. Cytoscape [7], Gephi [8], Pajek [9], igraph [10], etc.) provide automatic layout algorithms including force-directed algorithms.

Fruchterman-Reingold [11] layout is a popular force-directed layout method which models the network as a physical system of steel rings and springs and applies spring forces on the rings leading the system to a minimal energy state. It calculates repulsive forces between every pair of nodes and attractive forces only between adjacent nodes [11]. Thus, the algorithm aggregates densely connected nodes in subset and separates different subsets from each other iterating the simulation until the positions are close to an equilibrium.

By the way, Fruchterman-Reingold layout has been used in this thesis by means of Python package *NetworkX* [12] for setting up node coordinates. The implementation provides an option to use edge attributes to weight the edges so that higher values mean stronger forces of attraction.

3 THE ANTIPROLIFERATIVE COMPOUND SET

In this research, 220 compounds were collected from previously published analog series that were designed, synthesized and tested *in vitro* with the aim to study potential anticancer agents in an effort involving laboratories of the Department of Pharmacy and Biotechnology, University of Bologna as well as the University of Ferrara and Palermo, mainly.

This compound collection is composed of retinoids analogs [13]–[15], stilbene derivatives [16]–[19], thiazolobenzimidazole derivatives [20], chemically modified tetracyclines [21], combretastatin analogues [22], biphenyl-based hybrid molecules (with spirocyclic ketones) [23], 2-{{(2E)-3-phenylprop-2-enoyl}amino}benzamides [24], chimeric molecules (stilbene or terphenyl derivatives with SAHA fragment) [25], iodoacetamido benzoheterocyclic derivatives [26], pimozone derivatives [27], [28]. Figure S1 shows the molecular structures of all 220 compounds.

Most of these compounds have been studied for their biological activity as modulators of cell growth and differentiation. For example, retinoids, i.e. vitamin A and its biologically active, natural or synthetic derivatives, have been demonstrated to regulate the growth and differentiation of different cell types and they have been evaluated as chemoprevention agents [29]. Thus, all-*trans*-retinoic acid has been found to be active in patients with acute promyelocytic leukemia [29]. Retinoids exert their activities by binding to two receptor subfamilies: retinoic acid receptors (RARs) α , β and γ , and retinoid X receptors (RXRs) α , β and γ . Even if the therapeutic efficacy of retinoids has been thought to be related to both the differentiation activity and the activation of apoptotic processes, the molecular mechanisms underlying apoptosis-inducing activity still remain unclear [14].

Hence, these compounds were tested *in vitro* for their antiproliferative activity into two widely used cellular models of leukemia subtypes, namely the K562 and HL60 cells, which are representative cell lineages of CML and AML, respectively.

4 TARGETING SIGNALING PATHWAYS IN LEUKEMIA

Leukemia is a broad term that encompasses multiple hematological malignancies with an abnormal proliferation of hematopoietic stem cells and a clinical picture related to cytopenias. The World Health Organization (WHO) publishes and periodically updates the classification of leukemias [30], considering clinical, prognostic, morphologic, phenotypic and genetic features. However, the four main groups of leukemia are acute lymphoblastic (ALL), acute myelogenous (AML), chronic lymphocytic (CLL) and chronic myelogenous (CML) according to a classification based on cell type and rate of growth.

According to the American Cancer Society, AML is the second most common leukemia among adults (31%) [31]. However, 5-year survival rate among adults ages 20 and older is 27% for AML [31]. A critical issue is the biological complexity of AML since it comprehends phenotypically and genetically heterogeneous disorders. Thus, the research has been focused on the investigation of AML molecular pathways involved in cell proliferation and survival to develop novel specific treatments.

Different factors are implicated in AML like genetic and environmental conditions or patients' clinical history of hematologic disorder. As regards the pathogenesis of AML, it has been hypothesized that it depends on two cooperating processes: the uncontrolled cell proliferation due to class I mutations (FMS-like tyrosine kinase 3 (*FLT3*), *NRAS*, *c-KIT*) and the blocked myeloid differentiation related to class II mutations (*RUNXI-RUNXIT1*, *CEBPA*, *TP53*). Moreover, genetic mutations reveal distinct AML subgroups (mutations such as *NPM1*, *FLT3*, isocitrate dehydrogenase 1 (*IDH1*), *IDH2* and *TP53*) [32]. Furthermore, chemotherapy resistance and disease relapse have been related to self-renewing leukemia stem cells (LSCs) that are characterized by a CD34+, CD38– and CD123+ immunophenotype [33]. Apoptosis, receptor tyrosine kinase (RTK) signaling, Hedgehog (HH) pathway, mitochondrial function, DNA repair, and c-Myc signaling have been reported as targetable signaling pathways for novel therapies [34]. In the past, the AML standard drug therapy consisted of the combination of daunorubicin, a DNA intercalating agent and topoisomerase II inhibitor, and cytarabine, an antimetabolite. Hypomethylating agents as azacitidine and decitabine have been used in combination with venetoclax, an inhibitor of the antiapoptotic protein Bcl-2. A sticking point is the myelosuppression, caused by traditional antineoplastic agents, which hinders the development of effective drugs. Fortunately, targeted therapies for AML were introduced with the approval of midostaurin and gilteritinib (FLT3 inhibitors), gemtuzumab and ozogamicin (antibody–drug conjugates), enasidenib and ivosidenib (IDH1 and IDH2 inhibitors), glasdegib (Hedgehog receptor inhibitors) and CPX-351 (liposomal formulation of cytarabine and daunorubicin) [34].

CML accounts for about 15% of leukemia in adults with an incidence of 1-2 cases per 100000. Compared to AML, CML has a higher 5-year survival rate (70%) in adults ages 20 and older [31]. CML is characterized by the t(9;22)(q34;q11) reciprocal chromosomal translocation which results in the Philadelphia (Ph) chromosome. The *BCR-ABL* fusion gene encodes for Bcr-Abl1 protein with a constitutively tyrosine kinase activity. Abl1, a non-receptor tyrosine-protein kinase, forms dimers or tetramers, then autophosphorylates, resulting in an uncontrolled signaling to multiple downstream proteins: for instance, the perturbation of the Ras–mitogen-activated protein kinase (MAPK) leading to increased proliferation, of the Janus-activated kinase (JAK)–Signal Transducer and Activator of Transcription (STAT) pathway leading to impaired transcriptional activity, and of the phosphoinositide 3-kinase (PI3K)–Akt pathway resulting in increased apoptosis. Given the importance of aberrant Abl1 signaling in causing CML, research was focused on potential drugs targeting this pathway.

Tyrosine Kinase Inhibitors (TKIs) impairs Bcr-Abl1, blocking the ATP binding pocket, thus inhibiting phosphorylation and leading to cell death. Imatinib (Gleevec) is the first-in-class treatments of TKIs for CML treatment, but unfortunately disorder relapses are caused by resistance to imatinib by point mutations. Other TKIs have been therefore developed to overcome this issue, i.e., dasatinib, nilotinib and bosutinib with different efficacy to mutants.

5 AIM OF THE STUDY

The leitmotiv of this thesis concerns the application of network-based approaches in drug research. Moreover, the focus has been to explore the opportunities and challenges of network science in a real case scenario consisting of a set of antiproliferative compounds.

The first chapter of this thesis is devoted to the investigation of the bioactive set through chemical space networks (CSNs). Therefore, a brief introduction to cheminformatics, as the framework in which chemical space exploration is placed in, is given. Then, the concept of chemical space is presented, and CSNs, in contrast to coordinate-based representations, are explained. After that, the methods to design, analyze and graphically explore CSNs are presented with a focus on SAR analysis. Finally, the networks of the antiproliferative compounds under investigation are discussed. To conclude, pros and cons of this network-based approach in this cheminformatic application are outlined.

The second chapter is focused on the prediction of putative targets of these molecules by means of network-based approaches that integrate chemical and biological data. So, drug-target networks (DTNs) are presented as tools to explore the complexity of drug actions in the framework of systems pharmacology. Consequently, the concept of drug-target interactions (DTIs) prediction is introduced and the methods, from the collection of drug and target information to the link prediction algorithm, are reported. Then, the predicted targets are examined in the attempt to shed light on the potential pathways in which they are involved that might clarify the experimental activities of these compounds and help understanding their mechanisms of action. Finally, the highlights and drawbacks of the considered network-based prediction method are summarized.

The third chapter concerns the application of network-based approaches on the study of drugs currently in clinical trials for COVID-19. The COVID-19 Drugs Networker (COVIDrugNet) web tool is briefly explained. Finally, an application of COVIDrugNet for studying pharmacological options to treat COVID-19, in light of the biological evidence elucidating virus infection mechanisms, is proposed to prove that network-based tools can provide effective strategies in drug research context.

Chemical Space Networks

6 INTRODUCTION

At the foundation of the emergence of computer science in life science disciplines there has been the urge to design, manipulate and store data. Precisely, the exponential growth of the amount of data being generated thanks to high-throughput screening and combinatorial chemistry together with the increase of computer power formed the basis for the development of computational methodologies in drug design and discovery.

This premise leads us into the realm of cheminformatics. In particular, to deal with the large volume of chemical information, network-based approaches were also implemented in cheminformatics. Rather, network theory is at the ground of cheminformatics since molecules are usually computationally represented as molecular graphs as described below.

6.1 Cheminformatics

Cheminformatics denotes the computational approaches employed to address chemical problems. Although it is difficult to trace exactly when cheminformatics was founded, it was defined as following:

“The mixing of information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug lead identification and optimisation.” [35];

“Chem(o)informatics is a generic term that encompasses the design, creation, organisation, management, retrieval, analysis, dissemination, visualisation and use of chemical information.” [36], [37].

In broad terms, cheminformatics deals with datasets of small molecules. It embraces several aspects like the computational representation of chemical structures, similarity searching as well as the calculation of molecular descriptors and molecular similarity. These issues are at the basis of different applications in drug discovery as Quantitative Structure-Activity Relationships (QSAR) study, virtual screening and chemical space description which is the focus of this part of the thesis.

6.1.1 Representation of molecular structures

One of the most important aspects of cheminformatics is the digitalization of chemical structures. They are commonly represented as molecular graphs, where nodes correspond to the atoms and edges to bonds, and to which properties, i.e., atomic numbers and bond orders can be associated as attributes. Two exemplary data structures to store molecular graphs in digital format are connection tables and line notations. The essential blocks of a connection table are the list of atoms – one line for each atom describing x , y , z coordinates – and the list of bonds – one line for each bond describing the first and the second atoms and the bond type. This is the core structure of chemical table files as, e.g., *molfiles* and *SDF* (Structure Data File). As regards line notation, the *Simplified Molecular Input Line Entry Specification* (SMILES) is likely to be the most used chemical line notation language so that a SMILES is a string written in ASCII code. The SMILES system has its encoding rules [38]; basically, atoms are represented by their atomic symbols (upper case if aliphatic atoms, lower if aromatic) and single, double, and triple bonds are $-$, $=$, $\#$ respectively, but single bonds are usually omitted as well as hydrogen atoms. Branches are specified between round parenthesis and cyclic structures are represented by breaking a bond and specifying the same digit after the atomic symbols of the two atoms involved in the broken bond. By following these rules, different valid SMILES can be written for the same molecule. Therefore, to generate *canonical SMILES*, i.e. unique SMILES for a specific chemical structure, a canonicalization algorithm exists. Moreover, additional SMILES rules are optionally used to specify chirality, configuration of double bounds and isotopism. The SMILES strings generated by these rules are referred as *isomeric SMILES*.

SMILES files and SDF are the starting material of this research - containing the molecular information of the dataset - from which molecular fingerprints have been computed.

6.1.2 Molecular fingerprints

Molecular descriptors are used in cheminformatics for several applications from database handling with full or substructure searching to similarity searching. Different types of 2D fingerprints exist and they are classified in two families: *structural keys* and *hashed fingerprints*. The former employs boolean array or, more precisely, bitmap that encodes the presence (1) or absence (0) of specific structural features, i.e., fragments, in a binary string. In general, the computation of structural keys is time consuming, and further, the search speed is strongly affected by the choice of structural features. Structural keys lack of generality since the optimum fragment dictionary is often dataset dependent. However, structural features allow a straightforward interpretation since each bit corresponds to a

fragment. By contrast, hashed fingerprints are still bitmaps, but each bit is not assigned to a specific chemical feature. Precisely, the hashed fingerprints do not depend on a predefined fragment dictionary, thus, in principle, being appropriate for any type of molecular structure so that any of its fragment will be encoded. Nevertheless, for this reason, fingerprints suffer from the lack of interpretability since it is not possible to recall the substructural fragment from the bit position of the fingerprint. Despite this, fingerprints can be rapidly calculated and provide fast screening.

The *Extended-Connectivity Fingerprints* (ECFPs) [39] are circular topological fingerprints that offer various advantages over other types of fingerprints, like simplicity and flexibility, since the method can be customized and readily computed. The ECFPs are based on the Morgan algorithm which implements two essential steps:

1. Assigning atom numbering to all atoms iterating this procedure:
 - a. A random atom is selected and enumerated as 1;
 - b. Its neighbors are numerated in random order;
 - c. The neighbors of the atoms, labeled in sequential order on the previous step, are, in turn, numbered.

This step is repeated for every atom, yielding a set of atom assignments;

2. The algorithm iteratively eliminates assignments until one remains.

In the procedure, each atom is assigned an initial connectivity value, i.e., the number of heavy atoms attached to it. Subsequent iterations update the connectivity values with the sum of neighbor connectivity values until the number of unique connectivity values decreases; the connectivity values in the last iteration are used to allocate the unique atom assignments for that molecule.

In ECFPs, each atom in a molecule is seen as the center of a circle with a predefined radius. The ECFPs algorithm differs from the Morgan for two aspects: the intermediate atom identifiers are retained, and it does not require a unique set of atom identifiers.

ECFPs with bond diameter 4 (ECFP4) were computed for our bioactive compound dataset to assess the similarity among the molecules.

6.1.3 Similarity measures

One important task in a wide range of cheminformatics applications is to determine how similar a molecule is to another, regardless to the presence of a specific substructure. In order to quantify the similarity, methods require numerical descriptors for comparing molecules and a similarity coefficient that measures the extent of similarity based on the descriptors. It is important to note that similarity measures are independent of the molecular descriptor

used. Therefore, similarity methods are characterized by different molecular descriptors and similarity coefficients.

In particular, similarity coefficients can be calculated over 2D fingerprints since they are useful descriptors encoding the presence/absence of patterns in the compound structure. Among similarity metrics, *Tanimoto coefficient* is one of the early and most frequently used measures of similarity. The Tanimoto coefficient (T_c) measures the number of fragments in common between two molecules over the number of total fragments, hence the ratio of intersection over union:

$$T_c = \frac{c}{a + b - c} \quad (13)$$
$$0 \leq T_c \leq 1$$

where a is the count of bits set to 1 in the fingerprint of the molecule A, b the count of bits set to 1 in the fingerprint of the molecule B and c is the number of bits 1 in common to both A and B. The T_c values range from 0 to 1, where 0 indicates that the two molecules have no bits in common and 1 indicates the maximum similarity denoting that the two molecules have an identical fingerprint. However, identical fingerprints do not imply that the two molecules are identical.

In this thesis, the T_c was considered to assess the pairwise similarities within our antiproliferative molecule set with the aim of designing the chemical space networks.

6.2 Chemical space

Chemical space is one of the pillars of cheminformatics. The theoretical concept of chemical space relies on the idea of a *chemical universe* populated by all possible compounds. It was estimated that the chemical space of small organic molecules – i.e. compounds with a maximum number of 30 C, N, O and S atoms – is so huge to encompass about 10^{60} compounds [40], however very different estimates were proposed about how big is the chemical space [41]. Beyond the concept of chemical space for which different definitions have been proposed, it has served various purposes in cheminformatics and computational medicinal chemistry. Several applications have been reported in [41] and, to briefly mention the most relevant in drug discovery, chemical spaces are considered in the context of chemical library design, ligand-based virtual screening, diversity assessment, compound selection and, primarily, in the visualization and analysis of structure-property and structure-activity relationships (SPR, SAR).

In the last decade, indeed, there has been growing interest in the design, characterization and representation of chemical spaces populated by compounds with biological activities that have been referred to as *biologically relevant chemical spaces* [42].

6.2.1 Coordinate-based vs coordinate free representations

Conventionally, chemical space is envisioned as multi-dimensional space defined by the choice of chemical descriptors [42]. In this reference space, each dimension represents a descriptor, whereas compounds are represented as feature vectors and assigned coordinates (Figure 6a). However, this representation suffers from some drawbacks in describing chemical space, e.g. (i) the multi-dimensional space is continuous, whereas the number of compounds, however large it is, is discrete, (ii) chemical space strongly depends on the choice of molecular representations used to encode compounds, (iii) computed chemical space is generally a high dimensional space, hence dimensionality reduction is required for its interpretation leading to a loss of information, (iv) vector components, being continuous, have different units requiring to be scaled, and (v) if vector components have categorical values (molecular fingerprints) must be converted to coordinate systems of lower dimension with a consequent loss of information [43]. To overcome these limitations, coordinate-free representation (Figure 6b) of chemical spaces was then proposed by contrast to coordinate-based representation (Figure 6a) described above.

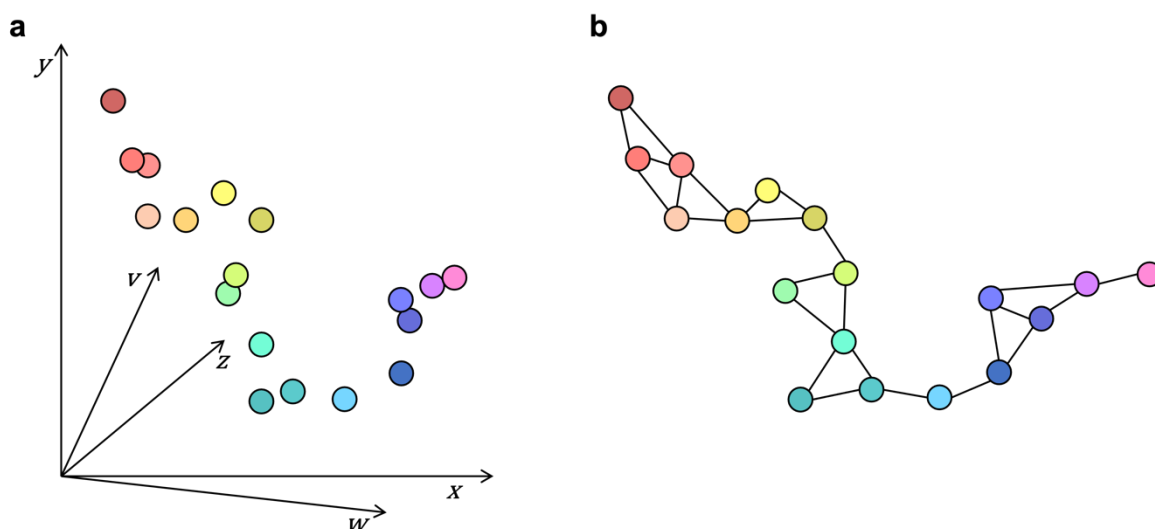


Figure 6. Chemical spaces. **a.** Exemplary coordinate-based representation: a schematic illustration of a multi-dimensional chemical space in which the axis are molecular descriptors. **b.** Prototypical coordinate-free similarity network in which edges are similarity relationships. In both panels the compounds are represented as circles.

The coordinate-free chemical space is based on pairwise compound similarities between molecules which replace feature vectors [44]. To graphically represent coordinate-free chemical space, chemical space networks (CSNs) were introduced. The representation of chemical space through CSN provides many advantages compared to coordinate-based

representations: (i) it captures the inherently discrete nature of chemical space, (ii) it is not affected by the issues related to high-dimensional spaces, and, in general, (iii) it allows an easy to interpret graphical representation. However, CSNs are influenced by the choice of the molecular representation, like multi-dimensional spaces. This remains a drawback, as well as the choice of the similarity coefficient that affects the design and analysis of CSNs. This dependence has been described as the lack of invariance of the chemical space representation, in general. Moreover, designing chemical spaces as networks provides the possibility to characterize chemical spaces through statistical properties offered by network science [43].

6.2.2 *Chemical space networks*

In CSNs compounds are represented as nodes and edges are pairwise similarity relationships. As mentioned before, the main advantage of CSNs is that they can be visualized and, consequently, easily interpreted compared to multi-dimensional space. However, this is true for compound sets of hundreds to thousands compounds since much larger sets threaten their interpretability. Even though large sets make the visual interpretation less effective, network analysis provides efficient tools to characterize large networks.

Concerning characterization of CSNs, network measures and properties can help revealing latent characteristics of chemical spaces. Latent characteristics are defined as those properties associated with nodes that are not considered for network design. Originating from social networks, the principle of *homophily* describes the tendency of nodes with similar latent characteristics to be connected to each other rather than to other nodes. Homophily principle finds a translation into chemical space when considering biological activities as an emergent property of CSN [44]. However, nodes with similar biological activities to be highly connected need to be similar according to the definition of similarity-based molecular networks (i.e., CSNs). This is exactly what states the *similarity properties principle* in cheminformatics establishing that structurally similar compounds tend to have similar properties, in this case, biological activities [44].

In the last decades, the increased interest in network as attractive tool for the exploration of chemical spaces paved the way to different studies focused on the design of CSNs. To trace the efforts in this topic from the beginning, it has to be said that the main variable in CSN design consists in the estimation of molecular similarity relationships [44].

At first, CSNs were designed as threshold chemical space networks (THR-CSNs) [45]. In this case, molecular fingerprints are primarily calculated as molecular descriptors and then the Tanimoto coefficient (T_c) is used as similarity function. To generate a CSN based on numerical continuous T_c values, a threshold value must be set as a criterion to choose how

many edges to display. Briefly, changing the threshold led to a series of THR-CSNs with different topologies, hence network analysis is strongly affected by this choice.

To introduce an alternative to THR-CSN, CSNs based upon matched molecular pairs (MMPs) were designed [46]. The MMP is a substructure-based similarity measure and it encodes a pair of compounds that differ only by a structural change at a single site. MMPs differ from continuous T_c values since MMP is a binary value, hence this results in a single CSN in which similarity relationships cannot be adjusted. It has been shown that MMP-CSN and the THR-CSN at same edge density share similar topologies, but local similarity relationships often differ.

Another example of CSN is MCS-CSNs [47] that were implemented by calculating the similarity measure as a variant of T_c upon maximum common substructure (MCS) between pairs of compounds. MSC-CSNs were found to improve cluster structures for data sets rich in structurally similar compounds.

To complete the list of CSN types reported in literature, TV-CSNs [48] are based on normalized Tversky similarity values; Tversky coefficient (T_v) is an asymmetric index calculated through a formula in which the parameters α and β determine the relative weights on two compounds being compared. In TV-CSNs, edges represent only asymmetric similarity relationships and with respect to a threshold value. A summary of CSN types and their similarity measure is proposed in Table 1.

CSN types	Similarity methods (incl. molecular representation - similarity function)
THR-CSN [45]	2D fingerprints - Tanimoto similarity upon threshold values
MMP-CSN [46]	Matched Molecular pairs
MCS-CSN [47]	Maximum Common Substructure - Tanimoto similarity variant
TV-CSN [48]	2D fingerprints - Tversky similarity

Table 1. Types of chemical space network and their relative similarity methods.

The design of CSNs relative to biological relevant compound data sets are often aimed at SAR studies [49]. In fact, one of the primary strengths of CSNs is that they can be annotated with compound biological activities. More importantly, from CSNs, specific compound clusters can be selected to study SAR information. Indeed, at low edge density, CSNs show well-defined communities that encode different local SARs [44], [50].

7 METHODS

7.1 Design of THR-CSNs

7.1.1 Curation of data

The core data set under study here was compiled from previously published studies whose common goal was to design, synthesize and biologically assay molecules endowed with antiproliferative potential (described in paragraph *The Antiproliferative Compound Set*).

To prepare the data, the compound structures were checked to identify and correct any structural errors. Then, duplicate analysis was performed to find potential duplicates and to remove them. The curated data set consists of 220 compounds; their structures and reference papers are reported in Figure S1.

Moreover, the small library was assembled taking care of collecting compounds whose differentiating, cytotoxicity and apoptotic activities showed low experimental variability. Indeed, these small molecules were tested *in vitro* for their antiproliferative activity by the same laboratory. The biological assays were conducted into two widely used cellular models of leukemia subtypes, namely the K562 and HL60 cells. In particular, for 140 out of 220 compounds the antiproliferative activities have been recorded as IC₅₀ (concentration able to inhibit cell growth by 50%) in K562 cells, whereas for 139 out of 220 IC₅₀ were measured in HL60 cells. The IC₅₀ values were converted in logarithmic scale: the obtained pIC₅₀ K562 values ranges from 4.000 to 7.699, while the pIC₅₀ HL60 values ranges from 4.000 to 7.523, as listed in Figure S1. This procedure of data curation was mainly performed using the Maestro [51] software.

7.1.2 Similarity assessment

As introduced above, THR-CSNs were based on similarity values that were generally calculated from 2D fingerprints. For this purpose, the extended connectivity fingerprint with bond diameter 4 (ECFP4) is commonly used as molecular descriptor to represent bioactive compound datasets, as reported in [44], [45], then it can be considered a standard for THR-CSN generation. In particular, ECFP4 is a topological atom environment feature set fingerprint of compound-specific size with higher structural resolution than, i.e., MACCS thus ECFP4 is more appropriate to represent bioactive compounds. In this case, the Morgan fingerprint with radius 2 – similar to ECFP4 – was calculated for each compound by the open-source cheminformatics software *RDKit* [52] in a Conda environment with Python 3.8. Then, the T_c was considered to measure the *structural overlap* for each pair of compounds according to conventional protocol for THR-CSNs.

7.1.3 THR-CSN construction

Herein, the mathematical formalism under network construction is explained in more details. Algorithmically, it is similar to the pipeline previously reported in [45].

According to graph theory, threshold networks can be defined as $G = (N, E)$ in which N is the set of nodes representing the compound set of n distinct molecules and E is the set of edges representing similarity relationships, so that m is the number of edges. Since each edge is described as unordered pair of nodes $e_{ij} = \{i, j\}$, the threshold network is undirected. The network is also simple as loops are not allowed.

Mathematically, T_c similarity values for each pair are usually assembled into a square $n \times n$ symmetric similarity matrix:

$$\mathbf{S} = \begin{pmatrix} s_{11} & \cdots & s_{1j} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{i1} & \cdots & s_{ij} & \cdots & s_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nj} & \cdots & s_{nn} \end{pmatrix} \quad (14)$$

in which s_{ij} is the similarity value between i and j and since in threshold networks s_{ij} is a T_c value, it ranges in the interval $0 \leq s_{ij} \leq 1$. In the leading diagonal, s_{ii} is the self-similarity thus it has the maximum value. If two non-identical molecules have the same molecular fingerprints, $s_{ij} = 1$ even if $i \neq j$ so the presence of off-diagonal ones. Being a symmetric matrix, the upper or lower triangular matrix completely defines all pairs of similarity values. From the similarity matrix \mathbf{S} , varying the threshold value t results in a sequence of threshold networks H :

$$\begin{aligned} G_t, 0 \leq t \leq 1 \\ H = \{G_0, \dots, G_1\}. \end{aligned} \quad (15)$$

Each threshold network has the same set of nodes N , but different set of edges E_t depending on the threshold value t so that a threshold network can be represented formally by the following symmetric adjacency matrix:

$$\mathbf{A}_t = \begin{pmatrix} a_{11}(t) & \cdots & a_{1j}(t) & \cdots & a_{1n}(t) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i1}(t) & \cdots & a_{ij}(t) & \cdots & a_{in}(t) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n1}(t) & \cdots & a_{nj}(t) & \cdots & a_{nn}(t) \end{pmatrix} \quad (16)$$

in which

$$a_{ij} = \begin{cases} s_{ij} & \text{if } i \neq j \text{ and } s_{ij} \geq t \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

and given that $a_{ij} = s_{ij}$ for $s_{ij} \geq t$, the edge linking i and j is represented by

$$e_{ij} = \{i, j, s_{ij}\} \quad (18)$$

representing how similar two compounds are according to the T_c value, resulting in a weighted network.

Within the sequence of threshold networks, G_0 is the complete network - maximum number of edges, $m_0 = n(n - 1)/2$ - when $t = 0$, whereas sparser networks are obtained by increasing t . Indeed, as the threshold value t increases, the number of edges m_t decreases as well as the network density since the number of nodes n is fixed. This is due to the fact that $G_{t'}$ is a subnetwork of G_t for $t' > t$ with a set of edges $E_{t'} \subseteq E_t$. Finally, when threshold value t assumes the maximum value, $t = 1$, the resulting network G_1 contains only disconnected cliques derived from molecules with identical fingerprints.

7.1.4 Characterization of THR-CSNs

7.1.4.1 Network density

The definition of network density was given in the paragraph 2.2.3.1 above. However, for each threshold network G_t belonging to the same sequence as described below edge density can be written as:

$$\rho(G_t) = 2m_t/n(n - 1) \quad (19)$$

in which n is the number of nodes common to every network of the series and m_t is the number of edges for G_t . As mentioned before, increasing t leads to sparser networks, i.e., the edge density decreases reaching its minimum $\rho_{min}(G_1)$ for $t = 1$, while $\rho(G_0) = 1$ for $t = 0$. Since $\rho(G_t)$ is a monotonic function in t , it is possible to calculate the threshold value from a given edge density $\bar{\rho}$ through the inverse function:

$$t(\bar{\rho}) = \max\{t|\rho(G_t) \geq \bar{\rho}\} \quad (20)$$

Pragmatically, the similarity threshold can be adjusted to obtain a given network density (Equation 20), and vice versa the edge density can be measured for the threshold network at a chosen t (Equation 19).

Network topology is affected by edge density: to compare networks of diverse data sets through their topological properties, it is required to set the same network density, not the threshold value.

7.1.4.2 Global network properties

With the aim to perform a network-based analysis on threshold network exploiting the potential of network-based approaches derived from having represented the chemical space as THR-CSN, different network properties depicting global features were calculated by means of package *NetworkX* [12] in Python 3.8. The network properties measured for each

threshold value are the following: (i) number of edges, (ii) edge density, (iii) number of connected components (CC), (iv) maximum degree, (v) degree assortativity and degree distribution, (vi) clustering coefficient, (vii) average path length.

8 RESULTS AND DISCUSSION

8.1 Design of CSNs

Using the whole dataset, the 220 antiproliferative compounds under study, a sequence of threshold networks was built from the similarity matrix by varying the similarity threshold value. Figure 7 illustrates four threshold networks of this sequence in which the nodes represent the molecules, and the edges are displayed if their pairwise similarity values exceed or equal the threshold value (t). Intuitively, high threshold values lead to a network in which almost all compounds are singletons as depicted in Figure 7d when $t = 1$ and the network density reaches the minimum.

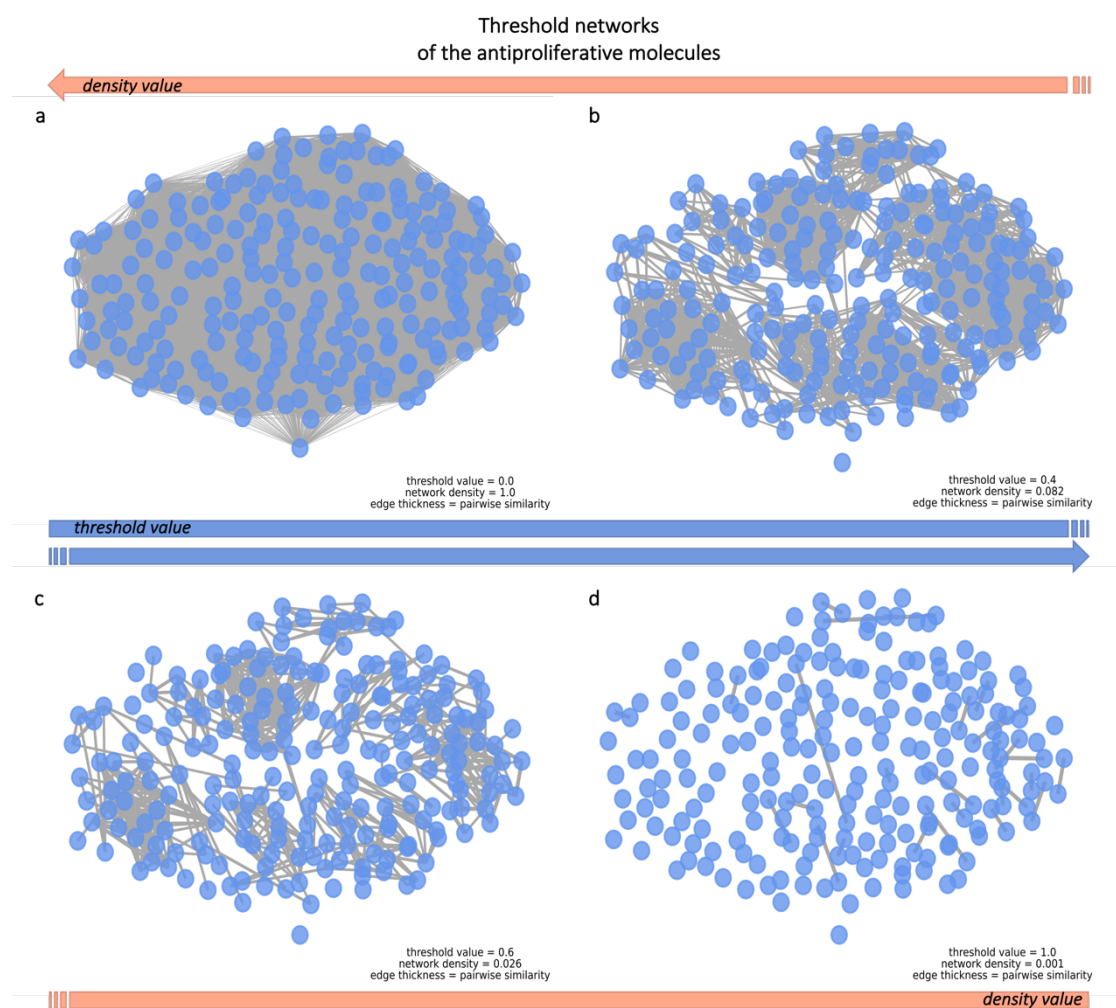


Figure 7. Threshold networks for antiproliferative molecules. Node positions were set for the complete graph only and then kept fixed for all the other threshold networks to help understanding the disaggregation of nodes. Node positions were

assigned through Fruchterman-Reingold force-directed layout considering the pairwise similarity values, i.e., edge attributes, in a way that larger values mean stronger attractive forces. The networks were produced by means of the Python package NetworkX [12].

At $t = 1$, the network contains only edges linking compounds with identical fingerprints which are not necessarily identical molecules since 2D fingerprints algorithm is not able to encode completely structural information. Actually, our dataset contains *cis*- and *trans*-isomers for which the similarity values correspond to unit. Furthermore, Figure 7a at $t = 0$ displays a single component in which all nodes are connected to each other, and the links are 24090 reaching the maximum. This shows how too high or too low threshold values lead to threshold networks that are not really informative. By contrast, intermediate threshold values (Figure 7b and 7c) result in networks in which clusters appear. Therefore, increasing the threshold value leads to decreasing edge density, thus allowing the network topology to emerge. Network properties are strongly influenced by the distribution of the similarity values, so that tweaking the threshold value affects the analysis of the network and, consequently, of the underlying molecules set.

8.2 Characterization of THR-CSNs

The properties of the THR-CSNs at different threshold values varying from 0 to 1 by 0.1 were computed and analyzed. Table 2 reports the properties including among others, edge density, degree assortativity, clustering coefficient and average path length for each CSN at a given threshold.

As explained before, network density ranges from 0 to 1; therefore, it reaches the maximum for the complete network that displays a fully connected component, to decrease at increasing values of t providing networks that become more and more sparse. In fact, the fewer the edges on a network, the sparser it is, resulting in a network density much smaller than unit. This behavior for edge density in relation to threshold value is reported in literature for different set of target-specific compound activity classes retrieved from ChEMBL and for randomly selected compounds from ZINC [45]. To sum up, changing threshold values results in variation of the edge density making possible to modulate network topology as shown by Figure 7 and Figure 8a.

From the data reported in Table 2, we also see how threshold variation, and, in turn, edge density affects other features of CSNs, like number of connected components (CC) and nodes in the largest connected component. At increasing threshold values, networks become

more and more disconnected, so that the number of CCs grows up as long as the number of nodes in the largest component drops.

Furthermore, it is worth to observe the degree assortativity (Figure 8b) that, except for extreme similarity threshold values, results in assortative networks, in which hubs are connected to each other and low-degree nodes to other low-degree nodes, resulting in a core-periphery structure. Then, high assortativity values might be indicative of the presence of series of analogs in the bioactive compound set that form tightly connected clusters.

In addition, clustering coefficient (see Table 2) rapidly increases at low edge densities then small changes in edge density result in large variation of the tendency of neighbors' nodes to form clusters.

In our case, clustering coefficient (Figure 8a) and degree assortativity (Figure 8b) shows different curves compared with what reported in literature for randomly selected ZINC compounds (Figure 4 in [45]) and this might be due to the nature of the dataset. In fact, as the authors described by comparing random selected and bioactive compounds from ChEMBL database at constant edge density, bioactive compound datasets differ from random compound samples for the presence of structural scaffolds that are responsible of the biological activities resulting in higher clustering coefficient and assortativity. This is confirmed by the authors in [44].

Our analysis suggests that the THR-CSN at an edge density of 0.05 is characterized by high assortativity and clustering coefficient as well as a number of connected components and nodes in the largest components that indicate the network is not disaggregated. In our opinion, this edge density results in an interpretable threshold network for our antiproliferative compound dataset.

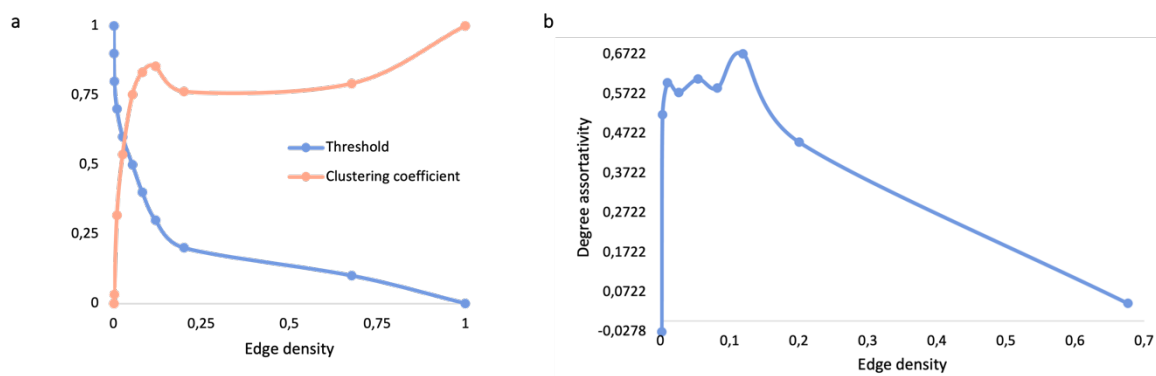


Figure 8. Global properties of threshold networks for our bioactive compound data set of interest. **a** shows the curves of similarity threshold and clustering coefficient versus edge density; **b** illustrates degree assortativity values at varying edge density.

Threshold	N. edges	Edge density	Max degree	N. of CC	N. nodes in the largest CC	Degree assortativity	Clustering coefficient	Average path length (SPL _{average})	Average of SPL _{average}
0	24090	1	219	1	220	N.D.	1	1.0	1
0.1	16296	0.6765	210	1	220	0.0439	0.7917	1.3235	1.3235
0.2	4825	0.2003	95	1	220	0.4511	0.7628	2.2944	2.2944
0.3	2874	0.1193	61	8	130	0.6737	0.8547	2.3024	0.9199
0.4	1977	0.0821	45	10	116	0.5873	0.833	3.0229	0.9628
0.5	1298	0.0539	32	11	114	0.6102	0.7529	4.5513	1.2344
0.6	626	0.026	20	29	36	0.5763	0.5359	2.4952	1.0534
0.7	240	0.01	12	88	23	0.601	0.3177	1.9091	0.6774
0.8	65	0.0027	4	159	11	0.5206	0.0318	2.8545	0.3219
0.9	37	0.0015	2	183	3	-0.0278	0	1.3333	0.1985
1	30	0.0012	1	190	2	N.D.	0	1.0	0.1579

Table 2. Network properties for THR-CSNs at different threshold value for our set of 220 molecules. The number of edges, edge density, maximum degree, number of connected components, number of nodes in the largest connected component, degree assortativity, clustering coefficient, average path length in the largest connected component and average of the average path length over all components are reported for each CSN at a given threshold.

8.3 Comparison of THR-CSNs

As explained above, variation of the similarity threshold value leads to different network topology and by consequence it strongly affects edge density and network properties. However, similarity value distribution depends on compound class, thus the CSNs of two different compound data sets at the same similarity threshold might result in non-interpretable networks. This is a critical aspect to consider when comparing biologically relevant chemical spaces of different sets by means of CSNs. However, since edge density can be interpreted as the probability of the presence of a link between two compounds, networks with the same edge density can be compared.

For this reason, given that it is possible to choose the threshold parameter after setting the density at a desired value, CSNs have been compared at constant edge density. As a result, topological properties of CSNs can be compared even if they are generated from compound collections characterized by different size and similarity distribution.

This concept turns out to be useful to study our collection of 220 compounds, since for 139 compounds out of 220 the antiproliferative activity was measured in HL60 cell lines, whereas for 140 out of 220 the antiproliferative activity was measured in K562 cell lineage; thus, for 59 compounds the activity was tested in both cell lines. This means that the chemical space networks of K562- and HL60- tested compounds can be compared by means of network analysis at a constant edge density. Therefore, sequence of threshold networks for both HL60 and K562 cell-tested molecules were designed separately and then their CSNs were compared at a predefined edge density value of 0.05.

Table 3 reports network properties for both HL60- and K562- cell-tested molecules networks. Both networks have high degree assortativity and clustering coefficient as well as similar maximum node degree. Moreover, similar distributions have been found for node degrees as shown in Figure 9. The threshold values corresponding to the predefined edge density of 0.05 differ only slightly, being 0.559 and 0.577 for HL60- and K562- cell-tested compounds, respectively.

Network features	HL60 cell-tested molecules network	K562 cell-tested molecules network
Threshold	0.559	0.577
Number of nodes	139	140
Number of edges	488	492
Edge density	0.05	0.05
Max degree	23	22
Number of CC	14	18

Number of nodes in the largest CC	93	26
Degree assortativity	0.5522	0.635477059
Clustering coefficient	0.6583	0.562402912

Table 3. Network properties of the HL60- and K562- cell-tested compounds network at constant edge density 0.05.

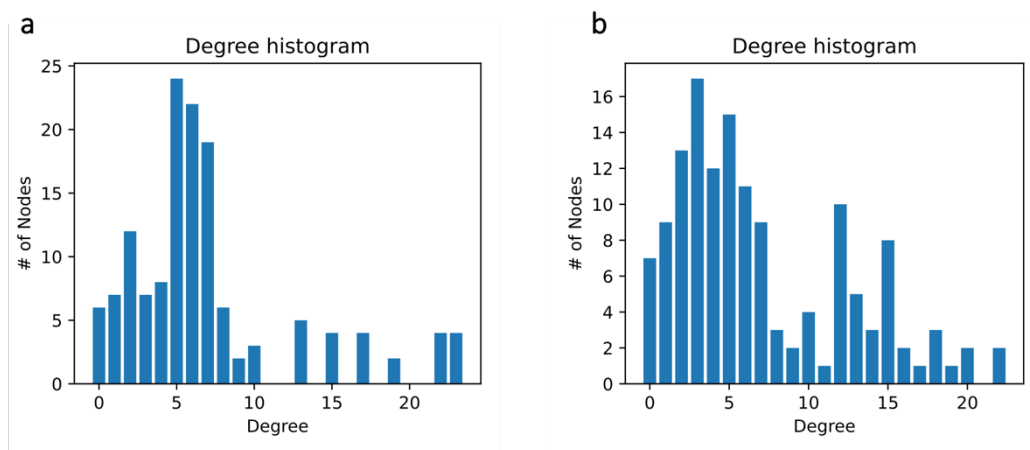


Figure 9. Degree distribution plots for THR-CSNs at edge density 0.05 of **a.** HL60 cell-tested molecules and **b.** K562 cell-tested molecules.

8.4 Exploration of SARs through CSNs visualization

The primary goal under the investigation of biologically relevant chemical spaces by means of network-based strategies is probably the SAR analysis. In detail, the main strength of chemical space networks is the opportunity to study the SAR through an interactive graphical exploration by taking advantage of the network topology which, in turn, relies on the structure similarity relationships. In addition, CSN visualization allows to identify clusters and to investigate their associated SAR information. Interestingly, clusters might represent local regions of continuous or discontinuous SARs that coexist in essentially all compound activity classes [50].

Therefore, for the purpose of studying the potential of CSNs as network-based approach for SAR analysis, THR-CSNs have been generated for both HL60 and K562 cell-tested molecules separately, as mentioned above.

Figure 10 shows the threshold network for the HL60 cell-tested molecules in our antiproliferative compound collection. The network is generated by setting the edge density to 0.05 that corresponds to a 0.559 similarity threshold value. The network consists of 139 nodes, that represent the compounds with a measured antiproliferative activity in HL60 cell

line, and of 488 edges representing pairwise compound similarities equal or higher than the threshold. To display the compound activities, nodes are colored from red to green according to the range of pIC_{50} of 4.0-7.52.

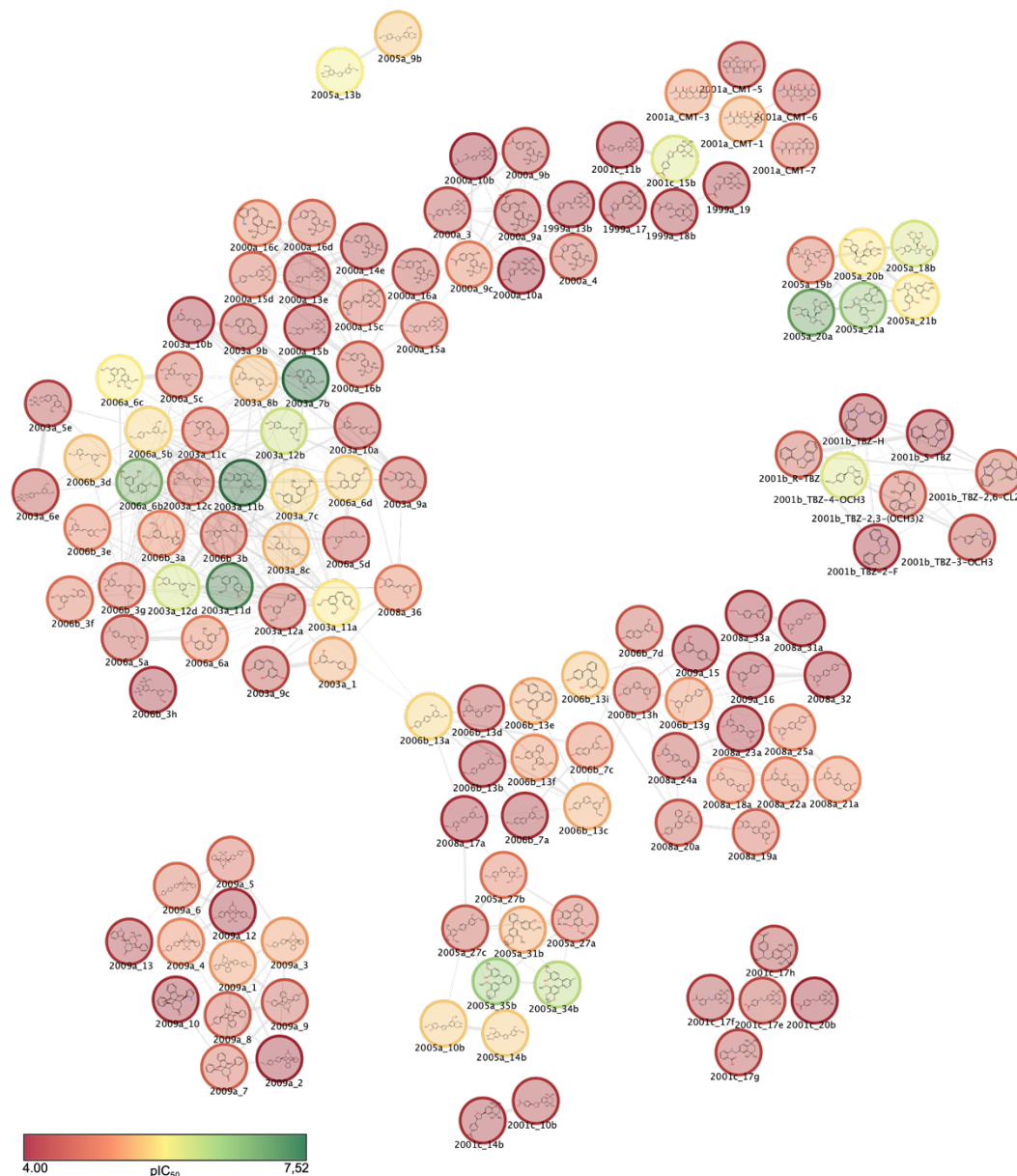


Figure 10. THR-CSN of HL60 cell-tested molecules at edge density 0.05 resulting on similarity threshold value 0.559. Nodes represent compounds and are colored according to antiproliferative activities from red (lowest potency) over yellow to green (highest potency). Isolated nodes are not shown. Edges represent pairwise similarity value over the threshold value. The network was rendered through Cytoscape [7] and the molecule structures depicted inside nodes by means of chemViz2 plugin [53] for Cytoscape app.

Similarly, Figure 11 illustrates the THR-CSN for the K562 cell-tested molecules included in the antiproliferative set. Also in this case, edge density is set at 0.05 that corresponds instead to a 0.577 similarity threshold value.

It is interesting to notice that molecules with different scaffolds are pretty confined in different clusters, thus the network structure reflects the different series of analogs in an efficient way.

Interestingly, Figure 10 and Figure 11 show that CSNs allow to identify continuous and discontinuous SAR regions that correspond to clusters in which the biological activities have homogeneous or heterogeneous values, respectively.

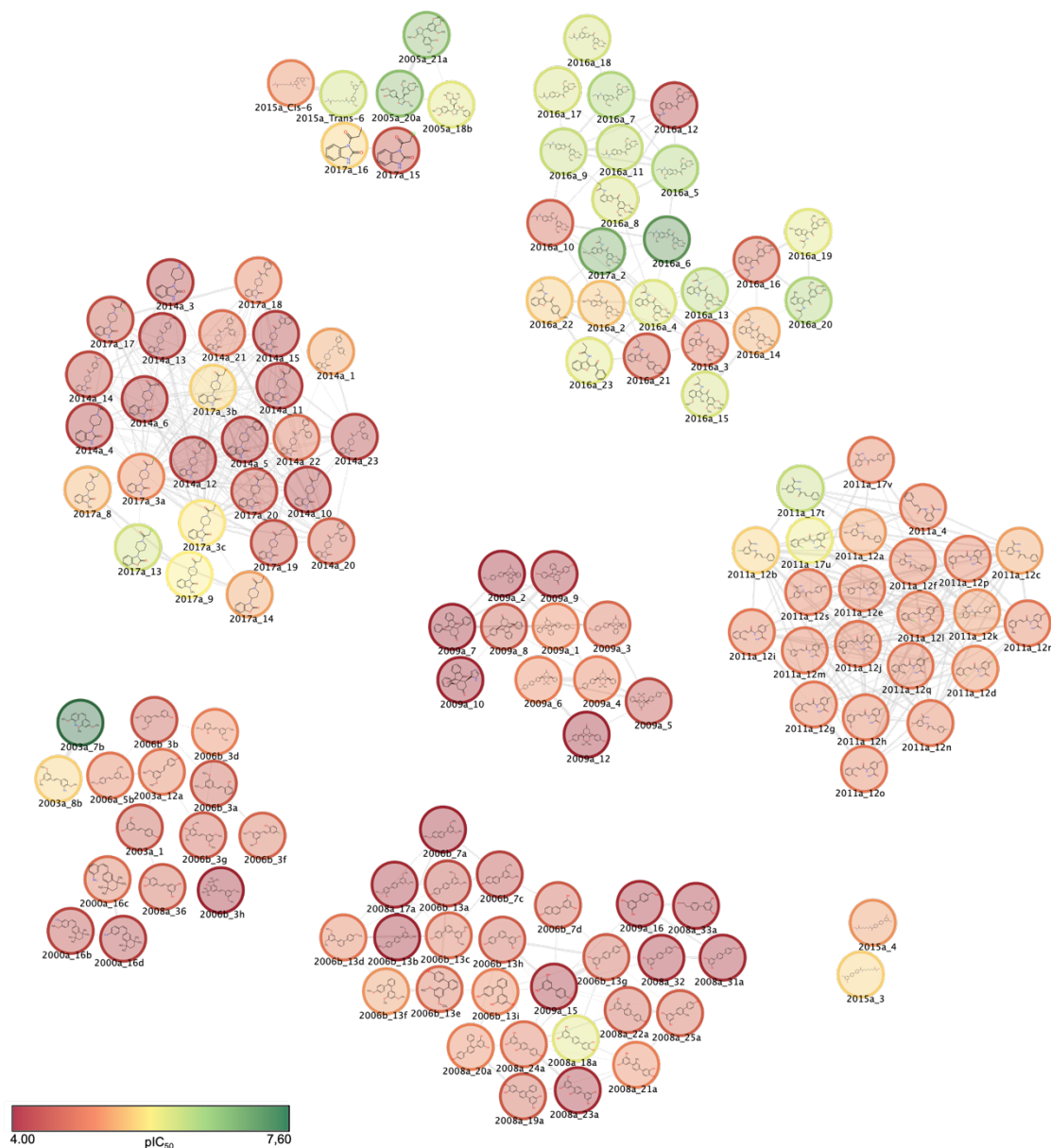


Figure 11. THR-CSN of K562 cell-tested molecules at edge density 0.05 resulting on similarity threshold value 0.577. Nodes represent compounds and are colored according to antiproliferative activities from red (lowest potency) over yellow to green (highest potency). Isolated nodes are not shown. Edge thickness represents pairwise similarity values over the threshold value. The network was rendered through Cytoscape [7] and the molecule structures depicted inside nodes by means of chemViz2 plugin [53] for Cytoscape app.

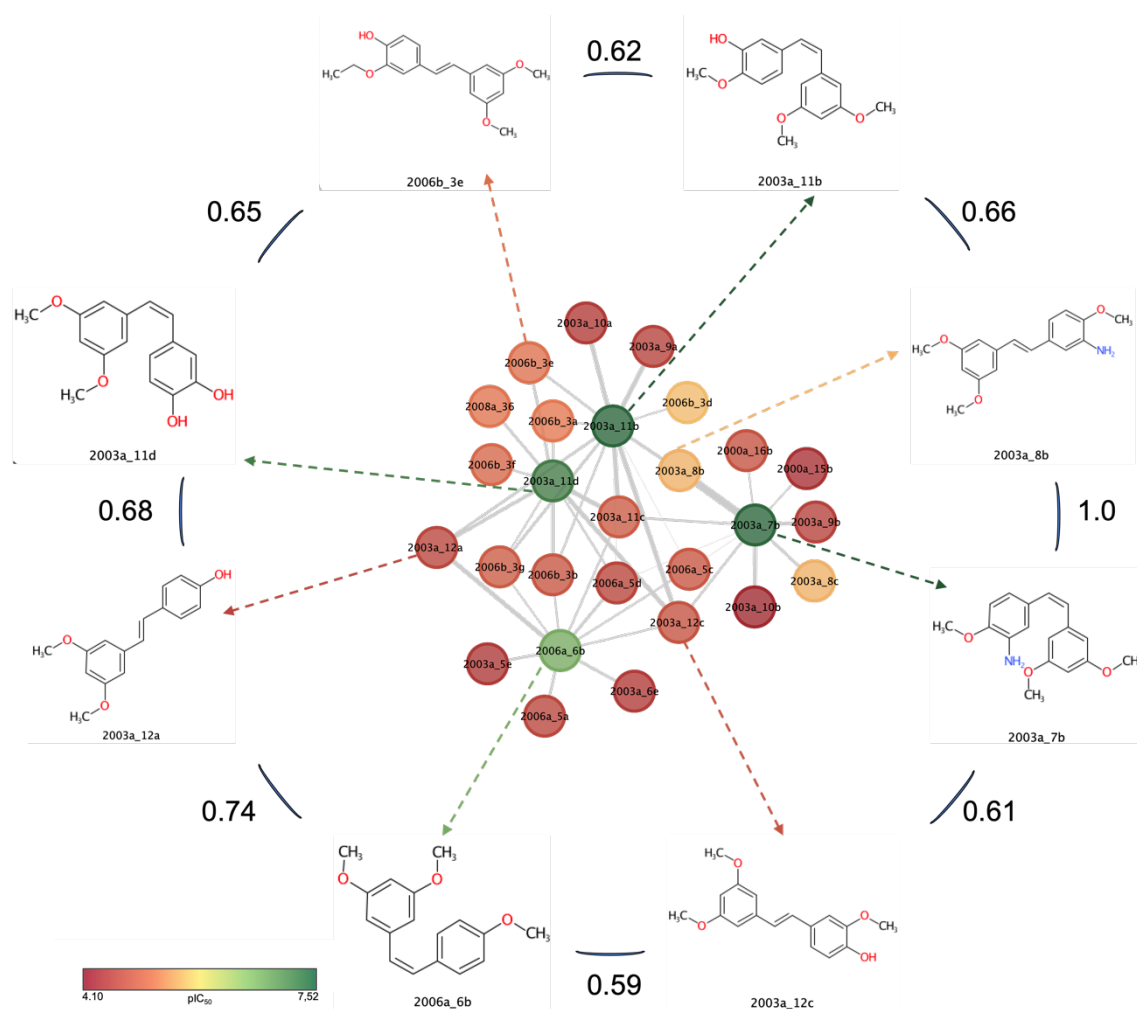


Figure 12. The largest connected component containing activity cliffs from HL60 cell-tested threshold network at edge density 0.05 and similarity threshold 0.559. Nodes represent compounds and are colored according to antiproliferative activities from red (lowest potency) over yellow to green (highest potency). Edge thickness represents pairwise similarity values over the threshold value. The network was rendered through Cytoscape [7] and the molecule structures by means of chemViz2 plugin [53] for Cytoscape app.

Furthermore, even if it is well accepted that similar compounds should share similar biological activities, it is also possible that pairs of structurally similar compounds display large differences in potency, i.e., activity cliffs [54]. Activity cliffs are considered SAR-informative indicators. Thus, CSNs can be transformed in activity cliff networks in which only the edges connecting compound pairs having at least a 100-fold difference in potency are shown.

The threshold network of HL60 cell-tested molecules was then analyzed looking for activity cliffs. Figure 12 shows the largest connected component containing activity cliffs. Thus, it includes the pairs of compounds characterized by a 100-fold difference in antiproliferative activities measured as the concentration able to inhibit cell growth by 50%, so IC_{50} values. In detail, similar compounds display large differences in potency (pIC_{50} values are reported

in Supplementary Figure S1), for, e.g., compound 2006a_6b with $pIC_{50}=6.82$ and compound 2003a_12a with $pIC_{50}=4.456$ show a pairwise similarity value of 0.74. Structurally, 2003a_12a is the *trans* isomer and contains the hydroxyl group in para position of the phenyl ring, while 2006a_6b presents the methoxy group in para and a *cis* configuration of the double bond. However, they show experimentally a more than 100-fold difference in potency. In turn, 2003a_12a is connected to the node 2003a_11d, but the latter has high potency ($pIC_{50}=7.301$) compared to the first.

To conclude, this is an example of how CSNs give the opportunity to highlight the relationship between compound chemical structures and biological activities for biologically relevant compound data sets.

9 CONCLUSIONS

With the aim to exploit the potential of network-based approaches in the framework of chemical spaces focusing on a real case scenario, CSNs of our antiproliferative compound set were designed and built as threshold networks. Briefly, the decision to design THR-CSNs was driven by the following considerations:

- THR-CSNs are based on continuous similarity values. In fact, the T_c is a numerical value that ranges from 0 to 1, in contrast to, e.g., MMPs which are binary values and are used to build MMP-CSNs. Moreover, the Tanimoto similarity in our THR-CSNs is calculated from 2D fingerprints that do not rely on predefined fragment dictionary, since Morgan fingerprints employed in this thesis are not based merely on the presence/absence of structural keys.
- THR-CSNs give the chance to modify the T_c threshold value. This constitutes, at the same time, an advantage and a drawback of the method, since on one hand the choice of the threshold value must be carefully evaluated to make the network interpretable, but on the other hand the possibility to adjust the threshold confers flexibility to the methods.
- Threshold networks enable the comparison of chemical spaces originating from different biologically relevant compound data sets. For this purpose, their respective CSNs are evaluated by setting a predefined edge density and not a similarity threshold value.

Starting from these considerations, the influence of similarity threshold on the chemical space of our small library of antiproliferative compounds was studied taking into account the network density and other network properties. This step is essential considering that the choice of the threshold value is dataset-dependent since molecular representation and

similarity measure affect network structure. Then, the network analysis helps to identify the similarity threshold and, by consequence, the edge density that results into an informative CSN for the bioactive compound data set of interest. This CSN is characterized by an interpretable network topology with defined communities.

Finally, with the dual purposes of proving the relevance of CSNs in drug research, more specifically, in SAR analysis and to propose a simple SAR study for our antiproliferative molecules, K562 and HL60 cell-tested compounds, CSNs at constant edge density were examined investigating *in vitro* activities and structure similarity relationships by exploiting the network formalism. In conclusion, THR-CSNs have been proven to be a powerful tool among SAR visualization and analysis methodologies.

Drug-target Interactions

10 INTRODUCTION

As seen before, network-based approaches help to understand the chemical information and to relate it to the biological activities for SAR analysis. However, network methods also enable the integration of different levels of information and pave the way for different opportunities in drug research in the context of systems pharmacology. For instance, network-based approaches allow to integrate the chemical and the biological spaces into a wide system. Thus, network science provides a framework to explore the system as a whole, revealing its peculiar features, that is the *emergent properties* derived from its inherent complexity, and not from its single entities.

10.1 Drug-target networks

Unveiling the molecular mechanisms through which drugs exert their therapeutic effects is a key issue in drug design and discovery. In this context, the cornerstone of the drug discovery paradigm is the process of DTI.

In the past decade, DTNs have been proposed as bipartite graphs that model drug-target associations retrieved from databases of experimental information on drugs, like, e.g., DrugBank [55]. By means of network visualization and analysis, inductive studies on the drug landscape can be conducted gaining different insights on the pharmacological space. To facilitate network analysis, the DTN can be projected into the target and the drug networks, in which only target nodes or drug nodes are represented and connected to each other if they share at least one drug or one target in the DTN, respectively. DTN analysis can reveal, e.g., trends in drug discovery and how they change, since DTN is like a screenshot of the molecular drug-target space, which necessarily reflects the advances of drug knowledge. Furthermore, the set of protein targets in DTN, representing the druggable genome, can be compared with the protein-protein network representing the interactome as reported in [56]. DTNs can be employed in the fields of drug combinations, drug repurposing, or adverse effects evaluation.

As an example, Figure 13 shows a DTN of approved drugs (small molecules) and human protein targets. The network contains 3627 nodes, 1636 of which represents drugs, and the remaining being targets so that links represents 7521 DTIs illustrating a global picture of the DTN extracted from DrugBank v. 5.1.5 [55]. The DTN is composed of a giant component, the largest connected component of 3368 nodes, 1510 of which are drugs. It is easy to notice that most drugs have at least one target in common. Moreover, coloring the drugs according to the first level of the anatomical therapeutic chemical (ATC) code highlights the presence of targets linking drugs of different therapeutic groups, and the tight clustering of neurological (light pink), cardiovascular (brown), and respiratory system (aquamarine) drugs.

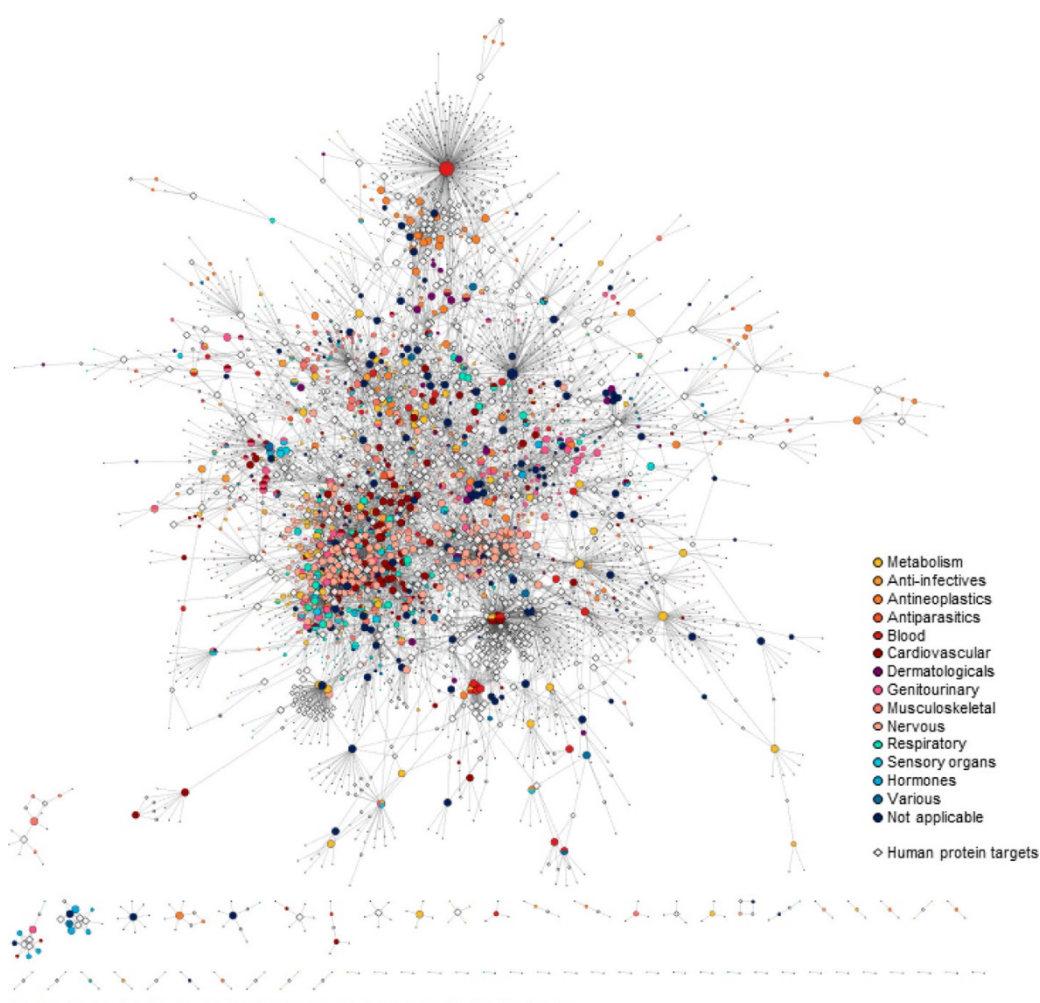


Figure 13. Drug-target network from DrugBank [55]. Approved small molecules drugs are represented as circle and human protein targets as white diamonds. The legend shows the color code for drug nodes accounting for the first level of the ATC code as reported in DrugBank. The nodes size displays increasing node degree. Network rendering was realized through Cytoscape version 3.7.2 [7]. Figure extracted from [57].

10.2 DTIs predictions

Identification of DTIs is a crucial step in drug discovery to detect novel targets for existing drugs or to discover candidate drugs for targets associated with diseases. Unfortunately, even though experimental methods for inferring DTIs exist, they are particularly costly and time-consuming. Computational approaches, in contrast, have been proved to efficiently predict DTIs, thus they can reduce the efforts for DTIs identification, providing a guide for wet-lab experimental validation.

Currently, *in silico* methods that deal with the study of DTIs are structure-based and ligand-based approaches as well as chemogenomic approaches [58].

Structure-based strategies such as molecular docking and molecular dynamics simulations are widely used to investigate DTIs when the 3D structure of the target protein is available. Ligand-based methods rely on the principle that similar molecules tend to have similar properties and thus bind similar proteins. 3D structure-activity relationships (3D QSAR) and pharmacophore modeling are useful to model the interactions between ligand molecules and the protein target of interest, for which the 3D structures is not known. However, ligand-based approaches lead to inaccurate predictions when the number of known binding ligands is low.

Instead, chemogenomic methods are based on large datasets of chemical structures and sequences for drugs and targets, respectively. The idea behind machine learning and network-based approaches which belong to this class of DTIs prediction strategies is that if drug d interacts with target t , then drugs similar to d are likely to interact with t ; proteins similar to t are likely to interact with drug d and drug similar to d are likely to interact with targets similar to t [59]. Chemogenomic methods can be classified into five types of models: neighborhood, bipartite local, matrix factorization, feature-based classification and network diffusion models [58].

Referring to network-based models, DTIs prediction can be seen as a link prediction problem. Among the most popular network-based prediction methods, we may mention the similarity-based algorithms based on either recommendation [60] or network propagation [61] approaches.

The first ones, i.e., network recommendation algorithms, exploit similarity scores to predict a node's preferences for unconnected nodes. To this class of methods belongs the network-based inference (NBI) algorithm implemented by Cheng et al. [62] to predict DTIs from a bipartite DTN built from the adjacency matrix containing known DTIs. NBI was further improved by Alaimo et al. [63] integrating into the model chemical and target similarity measures, thus incorporating domain-dependent biological knowledge.

The latter ones, i.e., network propagation algorithms, simulate the spread of information across the network. As an example, the well-known Google's PageRank algorithm [64] is based on random walk in the network of web pages that ranks web pages simulating a web surfer who randomly clicks hyperlinks. Most notably, Random Walk with Restart (RWR) algorithm has been successfully applied to DTNs with the aim of predicting DTIs. As a result, it estimates the relevance of each node with respect to the source node (or set of nodes), computing probability scores representing proximity measures between drugs and targets, thus being relevant for drug-target association prediction. Chen et al. [65] described a Network-based RWR method on Heterogeneous network (NRWRH) that was further optimized by Seal et al. [66]. It is important to note that this network-based approach is not based on the 3D structures of protein targets. This represents an advantage of NRWRH compared to, e.g., structure-based approaches since NRWRH can be applied when the 3D protein structures are not available. Another important aspect is that it overcomes the issue of selecting negative samples that affects the prediction accuracy of supervised machine learning methods in which unknown DTIs are considered as negative samples.

Leveraging these strengths, NRWRH was chosen as network-based link prediction method with the goal of predicting compound-target interactions for our collection of antiproliferative compounds.

10.3 Databases

Since chemogenomic approaches are based on large datasets, drug-related databases are the starting materials for any DTIs prediction task. Databases can be classified according to the type of knowledge they collect, e.g., DTIs (ChEMBL, DrugBank, KEGG, STITCH, SuperTarget), drug data (DrugCentral, PubChem) or target data (BRENDA, Pharos), drug-target binding affinity data (BindingDB) and so on [59]. However, in this thesis, two comprehensive databases, i.e., DrugBank [55] and KEGG [67], [68], are used and, therefore, introduced below in greater detail.

10.3.1 DrugBank

DrugBank [55] is a highly reliable and freely available drug database that includes primarily drugs and drug targets information. Precisely, DrugBank was originally conceived as an extensive and curated resource of cheminformatics and bioinformatics data [69]. With this dual-purpose, it was considered as an integrative collection of data about drugs - i.e., chemical structures and physical properties, mechanisms of action and pharmacological profiles - and drug targets - i.e. sequences, structures and mechanistic data - as reported for its first release in 2006 [69]. From then on, it was significantly improved to follow the great

advances in drug development and to meet the needs of its user base, since it has been accessed by pharmacists, pharmacologists, chemists and physicians other than pharmaceutical researchers. The later versions have been enhanced both in terms of expansion of data size and coverage and in terms of the information quality and reliability. As regards the data coverage, the database counts 215 data fields on the last version (DrugBank 5.0) [55]. To briefly sum up the growth in data content among versions, pharmacological information, ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) and QSAR (Quantitative Structure-Activity Relationship) parameters, drug and drug-food interactions as well as pharmacogenomic, pharmacometabolomic, pharmacotranscriptomic and pharmacoproteomic data were integrated into the database. Concerning data size, DrugBank 5.0 contains 2358 approved drugs including both small molecules and biotech drugs. Moreover, DrugBank collects different categories other than approved drugs, i.e., investigational, experimental, nutraceutical, withdrawn and illicit drugs.

Regarding drug targets, DrugBank 5.0 counts 4563 unique drug targets (including proteins, RNA, DNA, etc.) and 2242 compounds with drug-target binding constant data. According to DrugBank, by definition, a target is the biological entity to which a drug binds or interacts with, altering its function and producing intended therapeutic effects or unwanted adverse effects. Therefore, drug targets account for on- and off- targets, recording both the positive and negative effects of drug actions. However, since ver. 3.0 [70] on, drug targets have been classified into targets that have been found to exert the desired pharmacological effects and targets with unknown/unwanted effects. Furthermore, proteins that are involved in the delivery, transport and metabolism of drugs are separated from targets implicated in drug action, and collected into the carrier, transporter, and metabolizing enzyme classes, respectively. Nowadays, thanks to its breadth and reliability, it is widely used for *in silico* drug discovery, drug screening, drug target prediction, drug metabolism prediction, drug interaction prediction and *-omics* applications.

In view of the high-quality of data and its nature as integrated drug and drug target database, DrugBank was selected as source database for this research aimed at target identification.

10.3.2 KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) [67], [68] is an integrative resource originally designed to support the interpretation of biological systems from genome sequence data by assigning higher-level functions. KEGG consists of sixteen databases (as listed in Table 4) classified in four categories:

- *systems information* including PATHWAY, BRITE, MODULE databases,

- *genomic information* including KEGG ORTHOLOGY (KO), GENES and GENOMES databases,
- *chemical information* including COMPOUND, GLYCAN, REACTION/RCLASS and ENZYME databases,
- *health information* covering KEGG NETWORK, VARIANT, DISEASE, DRUG /DGROUP databases,

modeling the biological systems into an informatic representation.

Each database includes entries called KEGG objects, whose identifiers are composed of a database-dependent prefix and a five-digit number. For instance, D08389, in which “D” stands for KEGG DRUG, identifies pirenzepine.

KEGG Database	Content	KEGG prefix
System information		
PATHWAY	Pathway maps	map
BRITE	Functional hierarchies and tables	br
MODULE	KEGG modules and reaction modules	M
Genomic information		
ORTHOLOGY (KO)	Functional orthologs	K
GENES	Genes and proteins	locus_tag/GeneID
GENOME	Organisms and viruses	organism code/ T number
Chemical information		
COMPOUND	Metabolites and small molecules	C
ENZYME	Enzyme nomenclature	EC
GLYCAN	Glycans	G
REACTION	Biochemical reactions	R
RCLASS	Reaction classes	RC
Health information		
NETWORK	Disease-related network variations	N / nt
VARIANT	Human gene variants	GeneID+variant number
DISEASE	Human diseases	H
DRUG	Drugs	D
DGROUP	Drug groups	DG

Table 4. Architecture of the KEGG database.

The most peculiar feature of KEGG pertains to system information; it consists of manually processed experimental knowledge about the systemic functions of biological systems - intended as cell or organism - built upon networks of molecular interactions, reactions and relations in the shape of KEGG pathway maps, BRITE hierarchies and KEGG modules.

In contrast to other categories, health information was made up of human-specific data with the aim of providing a translational bioinformatics resource both for practical applications and to shed light on the molecular mechanisms of diseases and on drug actions. From the point of view of KEGG architecture, diseases are considered as perturbed states of molecular networks caused by genetic and environmental factors and drugs are view as perturbagens.

The most used databases in this research are KEGG DRUG, KEGG GENES and KEGG BRITE. Herein, a more detailed description of these reference resources is given.

KEGG DRUG is devoted to approved drugs in Japan, USA and Europe. As reported in Table 4, each drug entry is recorded with a D identifier and is annotated for therapeutic targets, drug metabolism and molecular interaction network data. At the time of writing (29/12/2021), KEGG DRUG includes 11802 entries, 6300 of which with reported targets (5095 with human gene targets) for a total of 1106 unique human genes (hsa IDs).

KEGG also provides additional drug classifications, for instance well-established ATC classification for prescription drugs as well as those based on drug targets, drug group and other features. This introduces the usefulness of KEGG BRITE, the collection of functional hierarchies of biological entities. It contains hierarchical text/table files for the classification of KEGG objects bridging genes, proteins, drug, diseases, etc.

KEGG GENES comprehends KEGG organism category that contains sequenced genomes derived from RefSeq, Genbank. The organism is codified by a three/four letter, e.g. *hsa* for *Homo sapiens* and each identifier is 'org:gene' where org is the organism and gene represents the gene identifier, e.g. the GeneID for genes in complete eukaryotic genomes.

11 METHODS

11.1 Data collection

11.1.1 Data collection from DrugBank

DrugBank [55] provides free access to drug-related information through the website, but also it allows its users to download the entire database in *xml* format. Thus, DrugBank (version 5.1.9 released on January 04, 2022) was parsed in R through *dbparser* package [71]. First of all, each class of proteins involved in drug actions (targets, enzyme, transporters and carriers), was filtered by organism to retain only human proteins. Then, only drugs classified

as approved, i.e., authorized for commercialization in at least one jurisdiction at a given time, were considered. Moreover, drugs were filtered to retain only small molecule drugs, discarding biotech ones. So, protein classes were further filtered to keep only proteins interacting with small molecule approved drugs. In this way, a drug-target edge list was acquired for each protein class. Henceforth, UniProt IDs are used as protein identifiers, and DrugBank IDs for drugs.

Afterward, to build the similarities matrices for amino acid sequences and drugs, protein sequence *FASTA* file as well as SDF for approved drugs were retrieved from DrugBank too.

11.1.2 Data collection from KEGG

The KEGG BRITE hierarchical table *br08310* (last updated: April 24, 2021) containing the target-based classification of drug was used as DTIs resource. The file contains drugs catalogued according to different levels of target information. The first and most general level is represented by the following classes kept separated for DTIs prediction: G protein-coupled receptors, ion channels, nuclear receptors, protein kinases, cytokines and receptors, cell surface molecules and ligands, transporters, enzymes, nucleic acids and not elsewhere classified. Nucleic acids class was not considered since it contains a microRNAs target with two drugs only. The second and third levels correspond to families and subfamilies, respectively. The latter contains grouped KEGG gene IDs for which their drugs are listed. This reference was used to generate a drug-target edge list for each target class. Then, the amino acid sequence was retrieved for each target of the drug-target edge lists by means of *KEGGREST* package [72] in order to obtain the *FASTA* file for each target class.

In the same way, KEGG DRUG was filtered keeping only the drugs in the drug-target list for which the database records a MOL file. In this case, for each KEGG DRUG ID, the corresponding MOL file was imported via *URL* in R and concatenated into an *SDFset* container. The ChemmineR package [73] was used to perform this task. As a result, the SDF was generated for each target class, to compute pairwise structure similarities later.

11.2 Data preparation

11.2.1 Target sequence similarity matrix

The sequence similarity scores for all pairs of amino acid sequences were calculated using the Smith-Waterman algorithm [74] for local sequence alignment. To compute the pairwise sequence alignments the BLOSUM62 was used as substitution matrix with default values for gap opening and gap extension costs. This task was performed with the R Biostrings

package [75] as reported in [66]. Thus, the normalized Smith-Waterman score between proteins i and j was computed as following:

$$S_t^s[i, j] = \frac{SW(i, j)}{\sqrt{SW(i, i)}\sqrt{SW(j, j)}} \quad (21)$$

in which $SW(-; -)$ is the original Smith-Waterman score according to the pipeline in [76]. In this way, the sequence similarity matrix denoted as S_t^s was generated for each protein target class.

11.2.2 Compound structure similarity matrix

The SDF of our antiproliferative compound set was merged to the ones containing drugs of each protein target classes. Hence, the compound similarity matrix denoted as S_d^c was built for the drugs of each target class by computing the T_c for all pairs of compounds upon the Morgan fingerprints with radius 2. Thus, $S_d^c[i, j]$ contains the similarity value between compounds i and j . This task was realized using the package *RDKit* [52] in Python 3.8 within a Conda environment.

11.2.3 Adjacency matrix

The adjacency matrices of the different target classes were built by considering the compounds in S_d^c and the targets in S_t^s . In the adjacency matrix, $A[i, j]$ will be 1 if drug j is reported to target the protein i , otherwise it will be 0, according to the presence/absence of the drug-target interaction in the edge list created before.

11.2.4 Additional similarities matrices

The main advantage of NRWRH with respect to classic RWR is that it integrates the similarity metrics derived from the biological and chemical knowledge domains with the similarity based on network information which derives from having shared targets for drugs and shared drugs for targets.

Thus, network-based similarity matrices for drugs and targets were generated from the adjacency matrix to exploit the information about known drug-target interactions.

Therefore, S_d^n is the drug-drug similarity matrix in which $S_d^n[i, j]$ is the number of shared targets between drugs i and j and S_t^n is the target-target similarity matrix in which $S_t^n[i, j]$ is the number of shared drugs between targets i and j . Hence, the similarity between i and j is measured by the Jaccard coefficient [66].

Then, the similarities based on chemical structure and based on network are integrated by the linear combination into S_d as following:

$$S_d = \omega_d S_d^c + (1 - \omega_d) S_d^n \quad (22)$$

and the same for target similarities according to which the integrated target-target similarities S_t is formulated as:

$$S_t = \omega_d S_t^s + (1 - \omega_t) S_t^n \quad (23)$$

where ω_d and ω_t represent the weights of the chemical or the sequence similarities, respectively, in the integrated similarity matrices.

This step is directly computed by the *Netpredictor* RWR function on bipartite networks.

11.3 RWR algorithm on heterogeneous network

Network-based random walk with restart on heterogeneous network (NRWRH) proposed by Chen et al. [65] to predict potential drug-target associations is reported below. First, the heterogeneous network is built by connecting the above-described drug and target networks through the DTN: these networks are defined by their respective integrated similarity matrices S_d and S_t and the adjacency matrix.

The method simulates a walker that, after starting from given source nodes, randomly walks from its current node to neighbors. However, with a probability r , the walk restarts from the source node. Thus, the initial probability matrix p_0 is defined as follows:

$$p_0 = \begin{bmatrix} (1 - \eta)u_0 \\ \eta v_0 \end{bmatrix} \quad (24)$$

where v_0 is the initial probability of drug network in which 1 is assigned to source nodes and 0 to the others; u_0 is the initial probability of target network in which the target nodes connected to drug source nodes are considered, in turn, source nodes in the target network with an equal probability value, the sum of which is 1; η sets the importance of drug and target network and its value is between 0 and 1.

Furthermore, the resource diffuses to neighbors according to a probability matrix, namely the transition matrix; if the current node is connected to a neighbor of the other type, the transition to the latter is allowed according to probability λ , while with probability $1 - \lambda$ it walks to a neighbor of the same type.

The transition matrix of the heterogeneous network denoted as W is defined as:

$$W = \begin{bmatrix} W_{TT} & W_{TD} \\ W_{DT} & W_{DD} \end{bmatrix} \quad (25)$$

where W_{TT} and W_{DD} denote the transition matrices which specify the probability the random walker jumps within the target network and drug network respectively, while W_{DT} is the transition matrix from drug to target and W_{TD} the vice versa.

The transition probability within the target network from target j to target i is

$$W_{TT}(i, j) = p(t_j | t_i) \quad (26)$$

$$= \begin{cases} S_t(i,j)/\sum_j S_t(i,j) & \text{if } \sum_j A(i,j) = 0 \\ (1-\lambda) S_t(i,j)/\sum_j S_t(i,j) & \text{otherwise} \end{cases}$$

The transition probability within the drug network from drug j to drug i is

$$W_{DD}(i,j) = p(d_j|d_i) = \begin{cases} S_d(i,j)/\sum_j S_d(i,j) & \text{if } \sum_j A(j,i) = 0 \\ (1-\lambda) d(i,j)/\sum_j S_d(i,j) & \text{otherwise} \end{cases} \quad (27)$$

The transition probability from drug j to target i is

$$W_{TD}(i,j) = p(d_j|t_i) = \begin{cases} \lambda A(i,j)/\sum_j A(i,j) & \text{if } \sum_j A(i,j) \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (28)$$

The transition probability from drug i to target j is

$$W_{DT}(i,j) = p(t_j|d_i) = \begin{cases} \lambda A(j,i)/\sum_j A(j,i) & \text{if } \sum_j A(j,i) \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (29)$$

Thus, the probability of finding the walker at node i at step t is defined by the i -th element of probability vector p_t . Therefore, to compute the probability, the NRWRH equation is implemented as follows:

$$p_{t+1} = (1-r)W^T p_t + r p_0. \quad (30)$$

The probability p_∞ indicates the steady probability defined as:

$$p_\infty = \begin{bmatrix} (1-\eta)u_\infty \\ \eta v_\infty \end{bmatrix} \quad (31)$$

achieved iterating the NRWRH algorithm until the change between p_t and p_{t+1} , measured by the Frobenius norm, is less than 10^{-10} .

Finally, targets are ranked based on u_∞ or, more pragmatically, the probability for the i -th element in p_∞ represents the *proximity score* between node i and the source node.

In this work, the implementation of NRWRH in *Netpredictor* R package [77] as proposed by Chen et al. [65] was applied to each of the different target classes of both KEGG and DrugBank databases, separately. The parameters are set to default values.

11.4 Over-representation analysis

The pool of predicted protein targets for the antiproliferative compounds only was converted to the corresponding coding genes and then evaluated to find a significant enrichment in

annotated biological pathways and curated gene sets through Over-Representation Analysis (ORA). Given that a gene set is an unordered collection of functionally related genes, a pathway can be interpreted as a gene set without considering functional relationship between them. The over-representation analysis was carried out by means of *enrichr* package [78]. *Enrichr* contains a huge collection of gene sets (about 400000) and gene-set libraries (about 300). The latter are classified into six categories: transcriptional, pathways, ontologies, diseases/drugs, cell types and miscellaneous.

The functional enrichment of pathways was conducted using the following gene-set libraries: *KEGG_2021_Human*, *WikiPathway_2021_Human* and *Reactome_2016*. For cell type enrichment *ARCHS4_Cell-lines* and *Azimuth_Cell_Types_2021* were used, whereas for diseases/drugs enrichment *MSigDB_Hallmark_2020*, *OMIM_Disease*.

Pathways were considered statistically significant if the p-values adjusted by the false discovery rates (FDR) was lower than 0.05. Pathways were prioritized according to the combined score which is defined by *enrichr* [79] as:

$$c = \log(p) * z \quad (32)$$

where c is the combined score, p the p-value calculated through the Fisher exact test and z is the z-score computed by evaluating the deviation from the expected rank.

Through *ClusterProfiler* R package [80], GO analysis for biological processes, cellular components and molecular functions were performed.

12 RESULTS AND DISCUSSION

Known DTIs for different pharmaceutically relevant classes of drug targets were retrieved both from KEGG according to its target-based classification of drugs, and from DrugBank. RWR is based on the assumption that similar drugs target similar proteins [65]. The prediction algorithm is implemented on a heterogeneous network, which is composed of the drug and target networks connected by the DTN.

Actually, Chen et al. [65] used four pharmaceutically useful drug target classes as gold standard, namely, enzymes, ion channels, GPCRs (G protein-coupled receptors) and nuclear receptors that had been compiled before by Yamanishi et al. [76]. Chen et al. also provides examples of the similarity nature of DTIs (similar drugs that target similar target proteins into heterogeneous networks) within the four classes of data.

Similarly, the predictions of DTIs for our compounds of interest were performed for each target classes separately, as described in Methods. Hereafter, the prediction performance of the NRWRH is assessed and the results for each database are illustrated and discussed separately.

12.1 Evaluating the performance of the algorithm in predicting missing links

The prediction performance of the NRWRH was evaluated on the KEGG and DrugBank datasets by using the statistical analysis proposed in *Netpredictor* R package [77]. The performance test randomly removes known DTIs from the dataset and assesses the ability of the algorithm to predict them. The DTIs removal takes into account the frequency of the DTIs so that the test datasets are built by removing only drug-target relationships for drugs with more than one target.

For evaluating prediction performances, AUAC (Area Under the Accumulation Curve), AUC (Area Under the Curve), AUC Top 10% (AUC up to the first 10% of the false positives), BEDROC (Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic curve) [81] and Enrichment Factor (EF) metrics were calculated. AUC ranges from 0 to 1 with a value of 0.5 corresponding to random performance. Although, AUC metric was assessed for the performance evaluation of RWR algorithms [65], other metrics are more appropriate when the aim is to assess the ability in ranking target candidates early in an ordered list (namely, early recognition problem) [81]. Indeed, EF measures the enrichment of annotated associations in the top-ranking list: the higher the EF is, the better the prediction performance of the algorithm in detecting true positives at the top of the ordered list compared to random selection. Similarly, BEDROC is useful since the metric can be applied for early recognition problem [81].

The performance test was executed for each dataset randomly removing 20% of the DTIs and computing the means of the performance metrics over 50 repeats.

The results of the performance evaluation are reported in Table 5, which shows good performances for each dataset both for KEGG and DrugBank databases. Overall, the resulting performances of the RWR algorithm implemented in *Netpredictor* [77] in the datasets generated before, suggest it is a reliable network-based DTI prediction method.

Datasets	AUAC	AUC	AUCTOP	BEDROC	EF
KEGG					
Cell surface molecules and ligands	0.838	0.609	0.551	0.479	7.073
Cytokines and receptors	0.690	0.943	0.369	0.348	4.809
Enzymes	0.944	0.907	0.845	0.747	9.223
GPCRs	0.940	0.903	0.707	0.585	9.260
Ion channels	0.928	0.968	0.402	0.364	6.784
Not elsewhere classified	0.927	0.855	0.398	0.334	9.748
Nuclear receptors	0.779	0.747	0.361	0.215	5.052
Protein kinases	0.923	0.926	0.630	0.527	8.628
Transporters	0.886	0.917	0.578	0.436	8.267
DRUGBANK					
Carriers	0.798	0.835	0.526	0.490	6.236
Enzymes	0.915	0.962	0.645	0.554	8.608
Targets	0.854	0.919	0.601	0.562	7.212
Transporters	0.921	0.958	0.610	0.528	8.698

Table 5. Performance test. AUAC; AUC, AUC Top 10%, BEDROC and EF are reported for each KEGG and DrugBank databases.

12.2 Prediction from KEGG database

To first characterize the DTNs, the number of DTIs, targets and drugs for each target classes of KEGG were considered and reported in Table 6. To make a comparison, the DTNs proposed in [76] account for a number of known DTIs of 2926, 1476, 635 and 90 for enzymes, ion channels, GPCRs and nuclear receptors, respectively, while in our case the number of DTIs are 2317, 3615, 3292, 604, respectively. Although in our case only KEGG database was considered as source, while KEGG BRITE, BRENDA, SuperTarget and DrugBank were used by Yamanishi et al. [76], the number of DTIs in our survey is higher, except for enzyme class. It is very likely that the continuous updating of drug-related databases on the basis of the evolving knowledge in drug research is the reason of the reported differences.

Nevertheless, a small number of DTIs for cell surface molecules and ligands, and cytokines and receptors classes were retrieved.

Therefore, a critical issue for DTIs prediction in some of the datasets is the small number of drugs with known targets. Indeed, considering the total number of drugs which includes the 220 antiproliferative compounds with unknown recognized targets, the percentages of drugs having at least one target are 12% (Cell surface molecules and ligands), 21.4% (Cytokines and receptors), 81.7% (Enzymes), 87.7% (GPCRs), 67.2% (Ion channels), 32.9% (Not elsewhere classified), 67.7% (Nuclear receptors), 58% (Protein kinases), 56.9% (Transporters) for the target classes.

Target classes	Number of targets	Number of drugs	Number of DTIs
Cell surface molecules and ligands	101	30	53
Cytokines and receptors	123	60	76
Enzymes	315	985	2317
GPCRs	122	1562	3292
Ion channels	129	450	3615
Not elsewhere classified	103	108	499
Nuclear receptors	20	462	604
Protein kinases	111	306	739
Transporters	42	291	464

Table 6. Collected data from KEGG database.

Despite these considerations, the RWR was applied through *Netpredictor* R package [77] on each dataset using the following set of predefined parameters: the restart probability $r = 0.8$, the jumping probability $\lambda = 0.2$, $\omega_d = 0.5$, $\omega_t = 0.5$ and $\eta = 0.01$. Actually, this set of parameters was previously tested in [66] to achieve the best prediction rate.

As reported in the Methods section, the probability vector u_∞ for each compound resulting from RWR makes up its target profile such that the relevance of a target i for the given drug is the i -th element of the vector. Unfortunately, using the KEGG database to build the networks lead to poor prediction results for some target classes. Indeed, NRWRH did not predict any targets for our antiproliferative compounds in enzymes, GPCRs, ion channels, nuclear receptors and transporters classes. It is interesting to note that these classes are those with the greatest number of known DTIs. On the other hand, the predicted compound-target interactions for cell surface molecules and ligands, cytokines and receptors, not elsewhere classified and protein kinases are 1056, 659, 780, 134, respectively. Figure 14 illustrates the

network of predicted compound-target interactions for cytokines and receptors, while Figure S2, S3, S4 shows the networks for the other classes.

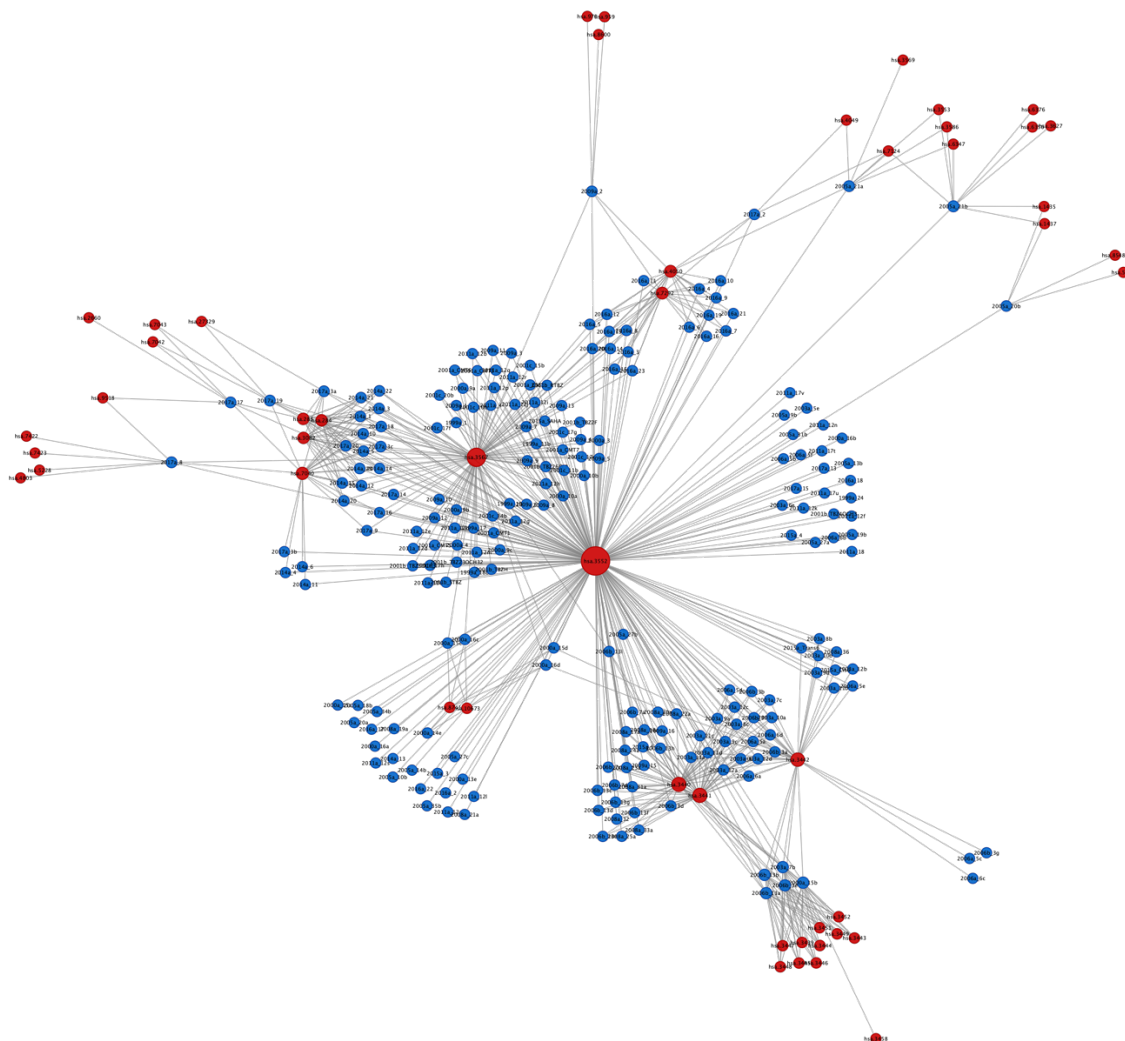


Figure 14. Predicted compound-target interaction network for cytokines and receptors class. Nodes representing the antiproliferative compounds are colored in blue, whereas predicted protein targets in red. Node size represents the node degree. Links represent the predicted compound-target associations. The image was rendered through Cytoscape 3.8.2 [7].

The compound-target network for cytokines and receptors in Figure 14 has a central hub, namely hsa:3552 (IL1A) with a degree of 216, and other predicted target nodes as hsa:3567 (IL5) whose node degree is 97, hsa:3440 (IFNA2) and hsa:3441 (IFNA4) both with a node degree equal to 48. According to KEGG, IL1A (interleukin 1 alpha) is involved in MAPK signaling pathway as well as hematopoietic cell lineage; IL5 (interleukin 5) has been annotated for JAK-STAT signaling pathway and pathways in cancer as well as for network of CML (nt06276). Among the pathways annotated in KEGG for IFNA2 and IFNA4

(interferon alfa 2 and 4, respectively), it is worth to mention PI3K-Akt and JAK-STAT signaling pathways.

Additionally, among the predicted targets for cell surface molecules and ligands depicted in Figure S2, hsa:3383 (ICAM1), hsa:3123 (HLA-DRB1), hsa:933 (CD22 antigen) and hsa:3134 (HLA-F) are targeted by most compounds of the antiproliferative set. Indeed, they are characterized by high node degree of 220, 210, 189, 179, respectively. Notably, intercellular adhesion molecule 1 (ICAM1) is involved in TNF and NF-kappa B signaling pathways and natural killer cell mediated cytotoxicity, according to KEGG pathways. Moreover, HLA-DRB1 (Major Histocompatibility Complex, class II, DR Beta 1) is involved in hematopoietic cell lineage pathway, Th1 and Th2 cell differentiation and Th17 cell differentiation. Furthermore, CD22 antigen is also annotated for hematopoietic cell lineage and B cell receptor signaling pathways, while HLA-F (an MHC class I antigen) is involved in cellular senescence.

As regards protein kinases, the network shown in Figure S3 is a star network in which the central hub is the only predicted target, namely hsa:93 (ACVR2B). Activin receptor type-2B is a serine/threonine kinase annotated for TGF-beta signaling pathway and signaling pathways regulating pluripotency of stem cells.

At a first glance, it appears clear that predicted targets mainly belong to cytokines and receptors, cell surface molecules and ligands and protein kinases, which are consistently altered in leukemia. In fact, mutations and chromosomal translocations in leukemic cells result in elevated expression or constitutive activation of several growth factor receptors and downstream kinases [82], and also dysregulation of the complex interactions between pro-inflammatory cytokines, such as IL-1 β , TNF- α and IL-6 and anti-inflammatory mediators like TGF- β and IL-10, which collectively govern the AML aggressiveness and progression [83].

These mutated genes affect diverse signaling cascades. The PI3K-Akt pathway is an intracellular pathway involved in cell growth and survival both in physiological as well as in pathological conditions. Indeed, it is a key regulator of survival in cellular stress which is, in turn, a frequent factor in malignant cells. In hematological disorders, PI3K-Akt pathway was reported to be abnormally upregulated and constitutive PI3K activation is detectable in 50% of AML samples.

Thus, targeting this pathway with effective inhibitors is a strategic pharmacological option to suppress leukemic cell proliferation [82], [84], [85]. However, the heterogeneity of intracellular signaling regulating cell growth frequently impairs the efficacy of treatments targeting just a single signaling pathway. In fact, leukemia cells might show alterations in

other survival pathways, as it happens, e.g., for the JAK-STAT pathway, that is also constitutively activated in hematological malignancies, often as a result of mutations in upstream genes [82], [86]. Moreover, the MAPK pathway is at the crossroads of signal transduction cascades integrating multiple extracellular stimuli to proliferation, differentiation, and survival [87]. The MAPK/ERK pathway (Ras-Raf-MEK-ERK pathway) is involved in sensitivity and resistance to leukemia treatments [88]. The pharmacological modulation of the components of these pathways has been explored to improve leukemia therapy effectiveness and achieve a better clinical course of the disease.

After that, an over-representation analysis was conducted to evaluate the pool of target as a whole. By joining all target classes, the predicted targets account for 98 genes. The goal of ORA is to determine whether the inferred targets show a significant enrichment of functional terms correlated to antiproliferative or other antitumoral pathways, that might be also relevant to target signaling pathways of leukemia. This should help to identify putative molecular functions or biological processes that could provide hints for the interpretation of the mechanism of action of our molecules of interest. Figure 15a shows the functional enrichment in predicted targets for the gene-set library *MSigDB_Hallmark_2020*. It is interesting to notice that different biological states and pathways involved in leukemia were included among the significant (defined by the * as $p\text{-value} \leq 0.05$) enriched terms, namely IL-6/JAK/STAT3 and IL-2/STAT5 signaling, hence the JAK-STAT pathway mentioned above, as well as apoptosis and TNF-alpha signaling via NF-kB which are involved in many tumoral pathways. NF-kB (nuclear factor kappa B) regulates the expression of different genes implicated in cell proliferation and antiapoptotic signaling and its constitutive activation in AML leads to resistance to apoptosis hence the need of drugs that inhibit NF-kB activity [89]. Moreover, Figure 15b shows top ten GO biological processes enriched in our predicted targets; they recall the previously reported signaling involving JAK-STAT pathways as well as multiple pathways relative to the regulation of cell proliferation or differentiation consistent with the antiproliferative activities observed through biological assays. To conclude, it is worth to mention the enrichment in proteins interacting with the hub protein ERBB2 found in the Hub Proteins in Protein-Protein Interaction database, as shown in Figure 15c. As described in literature, this receptor tyrosine kinases, when activated, leads to an increased RAS/MAPK, PI3K/AKT and JAK/STAT downstream signaling [90]. Additionally, SYK (spleen tyrosine kinase) was reported to be a critical regulator of FLT3 in AML since it is transactivated by SYK by direct binding [91].

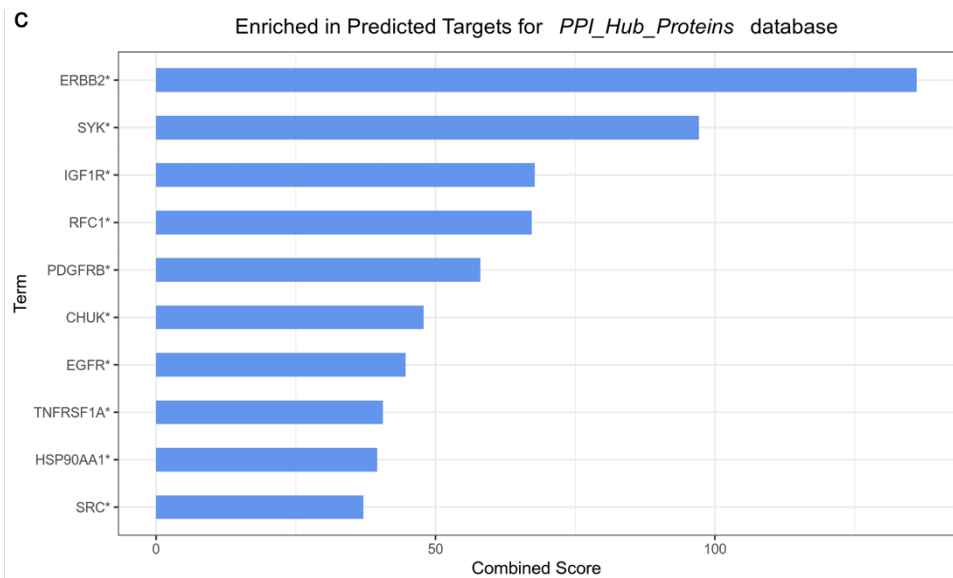
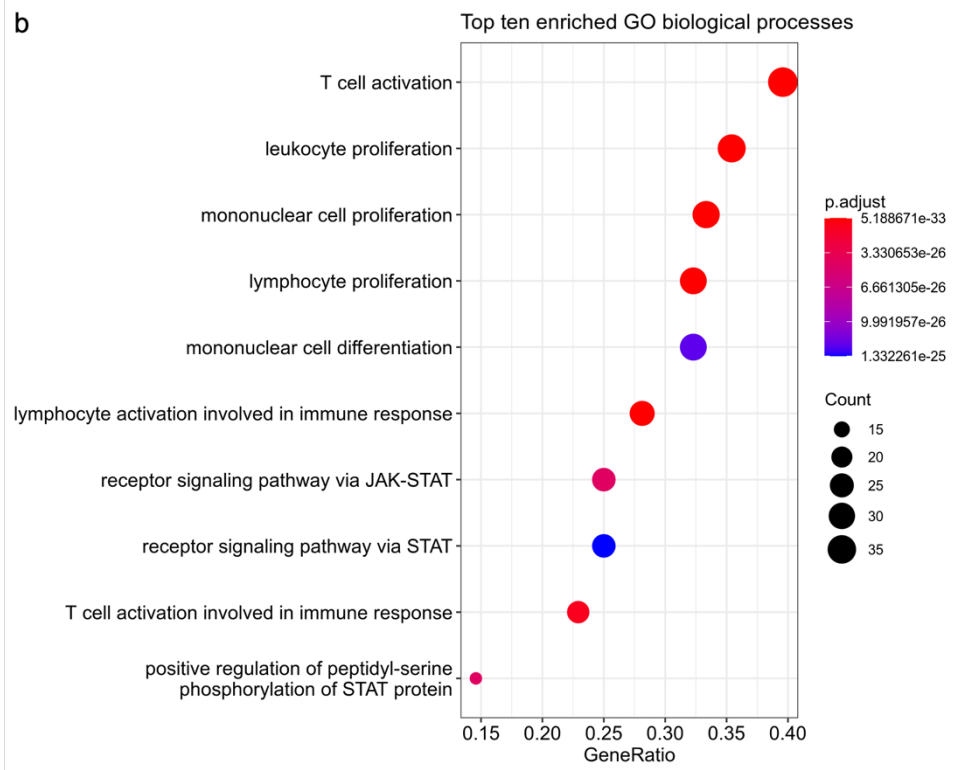
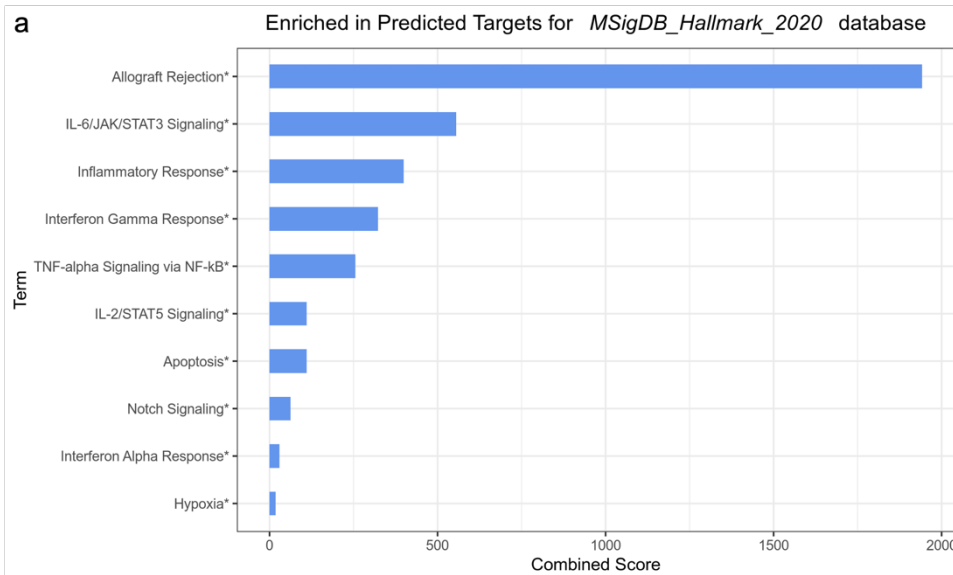


Figure 15. Functional enrichment plots. The ten pathways with highest combined score are shown for **a.** Molecular Signature Database Hallmark gene-set library and **c.** Hub Proteins in Protein-Protein Interaction Data. These bar plots were realized by means of *enrichr* R package [78]. The dot plot for Gene Ontology (GO) biological processes is shown in **b.** in which enriched terms are ranked in function of the number of genes entering a given pathway (Gene Ratio). The size of the dot is proportional to the size of the enrichment and are color coded on the basis of the adjusted p-value. The enrichment for GO biological processes was computed through *ClusterProfiler* R package [80] and plotted by means of *enrichplot* [92].

12.3 Prediction from DrugBank database

The analysis of the predicted compound-target interactions through RWR in DrugBank was focused on the target class since, by definition, it contains targets that have been found to exert the desired pharmacological effects and targets with unknown/unwanted effects, whereas the other classes include proteins involved in the delivery, transport and metabolism of drugs and not implicated in drug molecular mechanism of action. However, the numbers of DTIs, targets and drugs for each class of DrugBank were considered and reported in Table 7.

Target classes	Number of targets	Number of drugs	Number of DTIs
Carriers	75	482	709
Enzymes	388	1281	4506
Targets	2794	1648	7133
Transporters	257	867	2807

Table 7. Collected data from DrugBank database.

After incorporating the antiproliferative compounds, the percentages of drugs having at least one target are 68.7% (Carriers), 85.2% (Enzymes), 88.2% (Targets), 79.7% (Transporters) for the target classes.

Also in this case, the RWR on bipartite network was applied through *Netpredictor* R package on each dataset using the predefined parameters.

Therefore, the target profile for each compound resulting from RWR was computed by ranking the predicted targets according to their steady-state probabilities, i.e., proximity scores. Indeed, NRWRH predicts a total of compound-target interactions for enzymes, targets and transporters are 220, 426372, 36, respectively, whereas no target predictions are obtained for carriers. Hence, the top ten predicted targets per antiproliferative compounds

were considered representative of their target profile. Figure 16 illustrates the network containing the top ten predicted targets per compound for DrugBank targets.

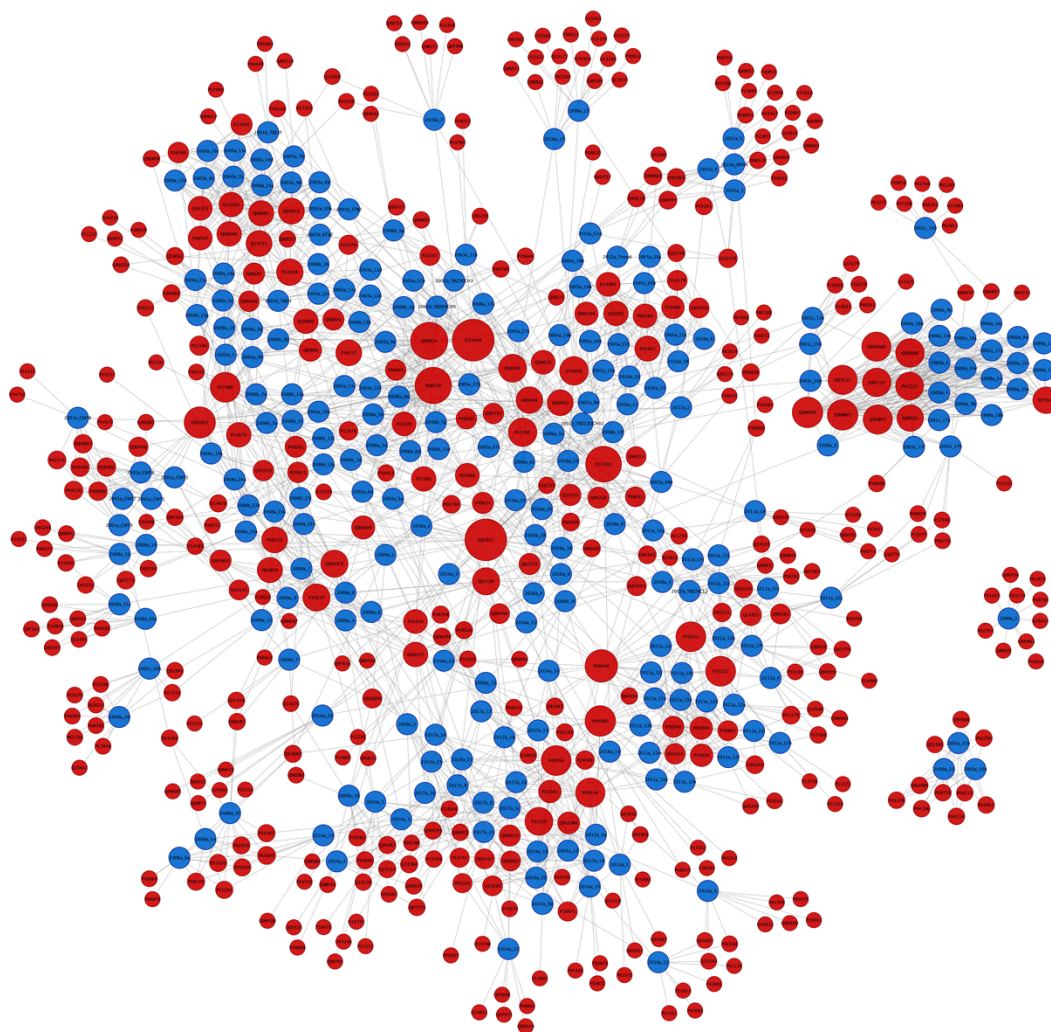


Figure 16. Top ten predicted targets per compound network for DrugBank targets. Nodes representing the antiproliferative compounds are colored in blue, whereas predicted protein targets in red. Node size represents the node degree. Links represent the predicted compound-target associations. The image was rendered through Cytoscape 3.8.2 [7].

What emerges from the network topology (Figure 16) is that some targets have been predicted for many compounds: SLC5A5 or NIS (Q92911), ABCC5 or MRD5 (O15440), EPAS1 (Q99814) proteins have highest node degree, i.e. 41, 41, 34, 33, 32. Interestingly, the ATP-binding cassette (ABC) transporters are transmembrane proteins implicated in multidrug resistance (MDR) against anticancer drugs acting as efflux pumps for xenobiotics, including chemotherapeutic agents, whereas solute carrier (SLC) transporters are considered gatekeepers of the cellular milieu implicated in changes in cell homeostasis thus in MDR too. It was previously reported that multidrug resistance proteins (MRPs) MRP1, MRP2, MRP3, MRP5 and P-glycoprotein (P-gp) contribute to a net resultant pump function in AML

and P-gp and MRPs activities were correlated with the maturation stage as defined by immune phenotype [93].

Even if ABC and SLC transporters are involved in cancer drug resistance, it has been challenging to find direct evidence of NIS and ABCC5 involvement on leukemogenesis, thus these compound-target interaction predictions are likely to be relevant not for the desired pharmacological effects, but for plausible unwanted effects. Moreover, polymorphism of ABCC5 was found to influence doxorubicin metabolism and functional pathways involved in anthracycline-induced cardiotoxicity in pediatric patient with ALL [94].

Furthermore, hypoxia has been shown to impact on cell proliferation, differentiation and resistance to chemotherapeutic agents not only in solid tumors, but also in leukemia, regulating the hematopoietic stem cell (HSC) niche and thus affecting the growth of LSCs that arise from HSCs and are involved in the relapse of AML. There has been a great interest in clarifying the role of hypoxia-inducible factors (HIFs) implicated in hypoxic signaling to advance therapeutic strategies targeting this process. HIFs are indeed the master regulators of cell response to hypoxia. Actually, HIF1A, HIF2A (alias EPAS1, endothelial PAS domain-containing protein 1) and HIF3A are three oxygen-regulated HIF-alpha subunits forming the heterodimer complexes and HIF-beta subunit. HIFs are regulated either by oxygen-dependent and oxygen-independent mechanisms. Tumor suppressor genes, i.e. p53, or GSK3 might cause HIFs downregulation, whereas PI3K/AKT or mTORC1 might cause its upregulation [95].

Then, the pool of targets was studied by means of the over-representation analysis. The predicted targets account for 1949 target genes.

Figure 17a displays the functional enrichment in predicted targets for the gene-set library *MSigDB_Hallmark_2020* highlighting significant biological pathways that could be related to leukemia according to our literature search. For instance, oxidative phosphorylation, fatty acids metabolism, glycolysis and peroxisome as well as mTOR signaling and apoptosis are enriched terms that can be easily reconducted to the metabolic impairments that were previously described for different hematological malignancies, comprising myeloid disorders [96]. Similarly, Figure 17b displays significant biological processes related to metabolic pathways as annotated by Gene Ontology.

Importantly, metabolic adaptation is a hallmark of cancer cells. The metabolic rewiring of leukemic cells not only affects the initiation, but also the disease progression involving cellular factors, i.e. oncogenic mutations, and microenvironmental factors, i.e. hypoxia, as mentioned before; these factors affect the suppression of immune responses against

leukemia. However, targeting bioenergetic pathways that differs between normal hematopoietic and leukemic cells, offers a therapeutic window to treat especially non responsive patients affected by leukemia disorders [96].

Considering the hub proteins related with our gene set, (Figure 17c) multiple enriched terms deal with protein kinases, namely PRKACA (protein kinase cAMP-activated catalytic subunit A), PRKCA and PRKCE (protein kinase C alfa and epsilon respectively), SRC (tyrosine protein kinase SRC), RP66KA3 (ribosomal protein S6 kinase A3) of the RSK family of serine/threonine kinases that phosphorylates members of the MAPK pathway, EGFR (epidermal growth factor receptor) and GRB2 (growth factor receptor bound protein 2) that binds EGFR. Among the other genes related to leukemia according to literature, STAT3 is included in the JAK/STAT signaling cascade.

Additionally, Figure S5 shows the functional enrichment in predicted targets for the gene-set library *NCI-Nature_2016*. Multiple pathways involved in leukemia were included among the significant (defined by the * as $p\text{-value} \leq 0.05$) enriched terms such as pathways of RXR and RAR, VEGFR1 and VEGFR2, ERBB1 (EGFR) and PDGFR-beta and osteopontin.

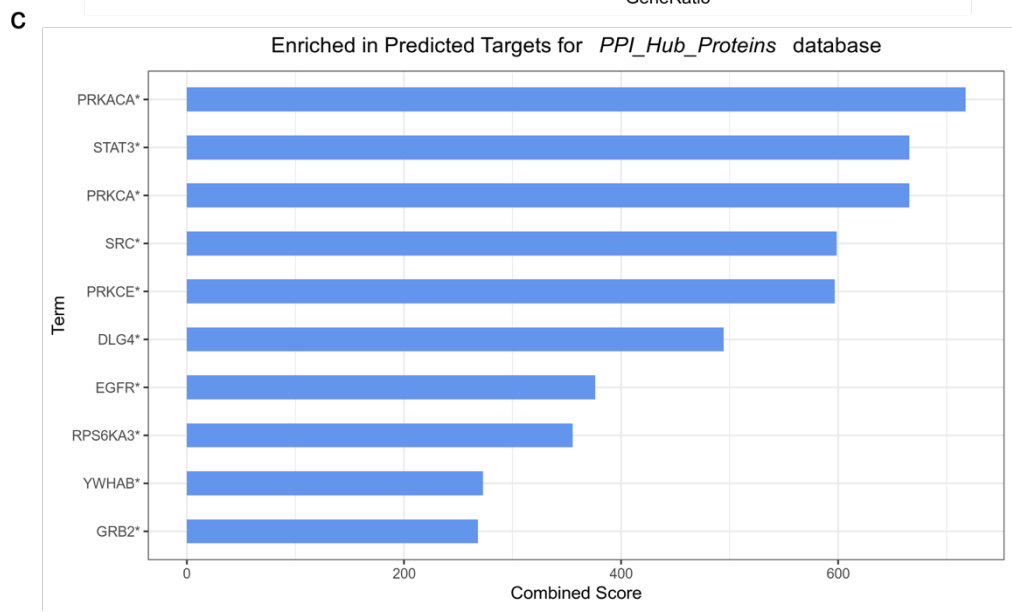
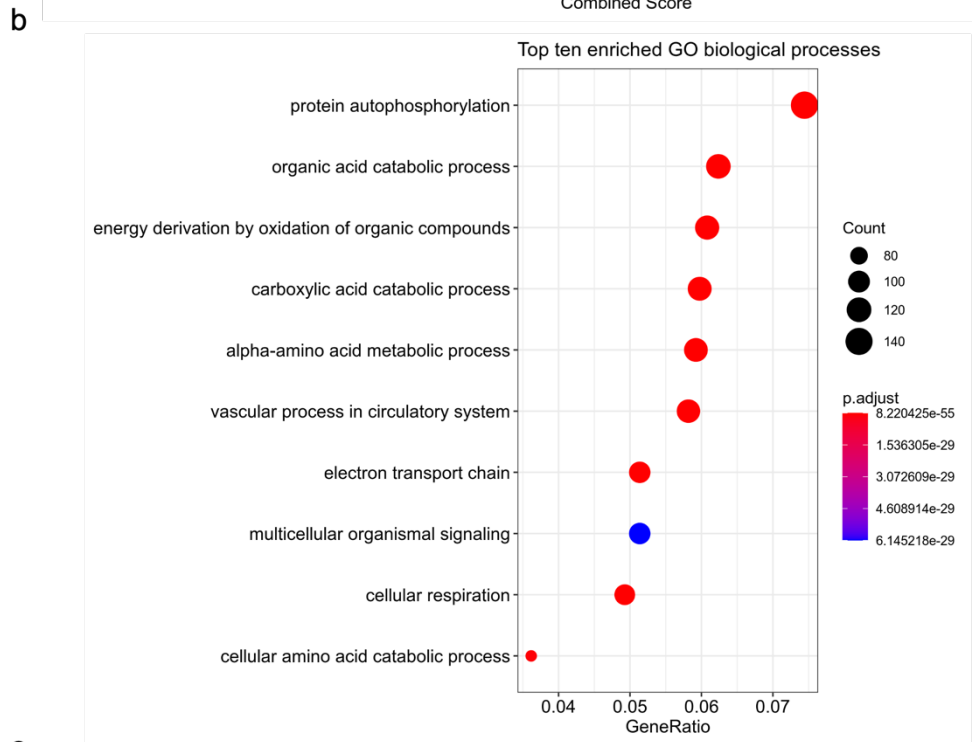
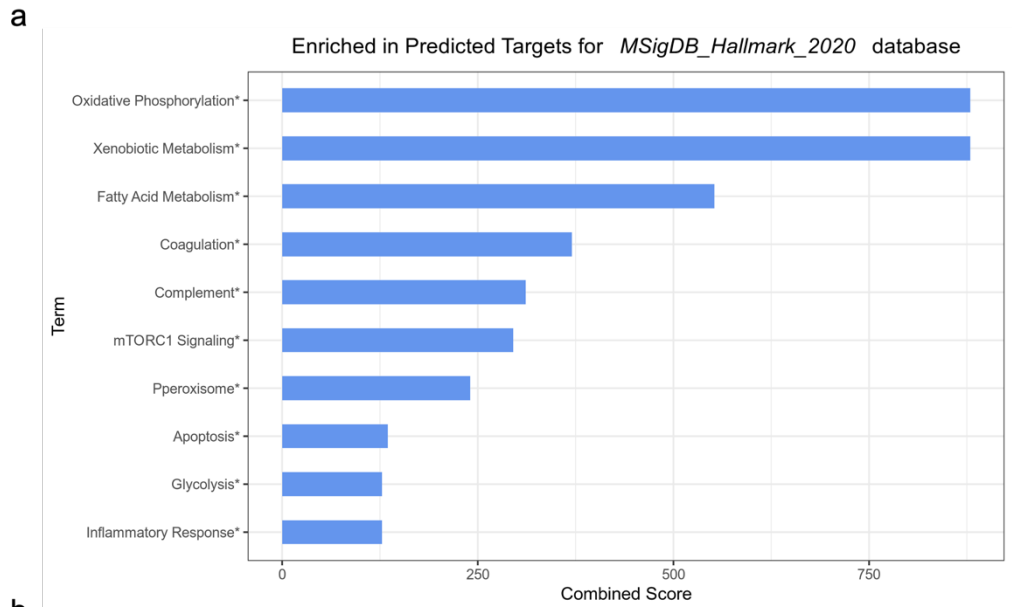


Figure 17. Functional enrichment plots. The ten pathways with highest combined score are shown for **a.** Molecular Signature Database Hallmark gene-set library and **c.** Hub Proteins in Protein-Protein Interaction Data. These bar plots were realized by means of *enrichr* R package [78]. The dot plot for Gene Ontology (GO) biological processes is shown in **b.** in which enriched terms are ranked in function of the number of genes entering a given pathway (Gene Ratio). The size of the dot is proportional to the size of the enrichment and are color coded on the basis of the adjusted p-value. The enrichment for GO biological processes was computed through *ClusterProfiler* R package [80] and plotted by means of *enrichplot* [92].

All the above considered, it appears that the analysis of compound-target interaction predictions through KEGG and DrugBank databases leads to coherent and complementary results as suggested also by over-representation analyses. These show enriched pathways that can be reconducted to leukemias as expected for our compounds that were tested on K562 and/or HL60 cell lines for their antiproliferative activities.

Another aspect that further supports our compound-target predictions is that these compounds were specifically designed and synthesized with the purpose of inhibiting targets involved in many types of cancer, including CML. For instance, the pimozone derivatives in [27], [28] and the iodoacetamido benzoheterocyclic derivatives in [26] were originally designed as STAT5 inhibitors, since STAT5 is a transcription factor of the STAT family constitutively activated in BCR-ABL positive CML and the inhibition of STAT5 phosphorylation induces apoptosis in CML cells. Additionally, retinoid analogues with apoptotic activity were proposed in [15] since retinoids are a class of vitamin A analogues structurally related to all-trans-retinoic acid involved in the growth and differentiation in both normal and malignant cell types. Indeed, retinoids were found to inhibit cell proliferation and to induce differentiation and apoptosis at the cellular level; moreover, they act by binding to specific nuclear retinoic acid receptors (RARs) and retinoid X receptors (RXRs). Therefore, it is important to note how previous literature about these antiproliferative compounds is coherent with what emerges from our analyses. Indeed, according to the over-representation analyses, predicted targets are involved in biological pathways like apoptosis, cell proliferation and differentiation as well as oxidative phosphorylation and pathways related with STAT family, RXR and RAR receptors.

Finally, on one hand, this work confirms that RWR is a valuable network-based approach for compound-target interactions prediction, and, on the other hand, it enables to hypothesize possible molecular mechanisms that might enlighten the antiproliferative activities of our compound collection in view of further experimental validation.

13 CONCLUSIONS

Here, a network-based link prediction method that has been previously validated and exploited to predict DTIs with the aim of identifying potential targets for existing drugs was applied to infer potential targets for antiproliferative compounds, whose mechanisms of action are not yet clarified. With the aim of investigating the molecular basis of their biological activities without reducing the complexity of compounds action, a network-based approach was applied.

In summary, this work is an attempt to predict potential targets also for compounds having no known targets using the RWR approach. The great strength of RWR is that it exploits network structure information to infer missing links and indeed DTIs prediction can be seen as a problem of link prediction in complex networks. In effect, network-based prediction for a drug/compound by means of the RWR approach allows to take into account not only the closeness of molecules to the potential targets, but also the multiple paths connecting the former to the latter. From this point of view, RWR is capable of predicting potential targets even if the drug has no known target.

Thus, the target profile of each antiproliferative compound was investigated considering compound-target interaction networks. In broad terms, some targets were found to be predicted for many compounds. This behavior is coherent with the fact that a certain degree of similarity is present within the compound dataset.

Additionally, the predicted targets were further investigated with the aim of advancing some hypotheses that might shed light on the mechanisms of action underlying their antiproliferative activities.

In conclusion, this network-based approach is a valuable method to make a hypothesis about the mechanisms of action of these antiproliferative compounds, setting the stage for further experimental studies aimed at validating them.

Network-based approaches for COVID-19 drug research

14 INTRODUCTION

The outbreak of the COVID-19 pandemic prompted the scientific community to join the efforts for fighting the public health emergency.

In this context, to contribute to facilitate the exploration of the knowledge about the continuously evolving scenario of the ongoing drug clinical trials, as a research group, we developed a web tool, namely COVID-19 Drugs Networker, for COVID-19 related drugs now available online at <http://compmedchem.unibo.it/covidrugnet> and published in [97].

15 COVID-19 DRUGS NETWORKER

COVID-19 Drugs Networker (COVIDrugNet) was designed to keep the user up to date on the progresses of drug development to contain the SARS-CoV-2 infection.

The web application provides a network-based approach to study drugs repurposed for COVID-19 listed on DrugBank database. COVIDrugNet supports the graphical exploration of the interactive networks and their analysis.

The COVIDrugNet core is the interactive bipartite drug-target network (DTN shown in Figure 18a) which contains drug and target nodes whose additional features are reported in the *Node Info* box providing drug and target data. The DTN (Figure 18a) is indeed built connecting drugs currently in clinical trials present in the COVID-19 Dashboard of DrugBank [55] and their retrieved targets. Currently, the DTN is composed by 292 drugs and 1193 targets with a giant component including 211 drugs 1040 targets. Since bipartite networks are usually analyzed by means of the two monopartite networks called projections, COVIDrugNet offers the interactive Drug and Target Networks (DN and TN, as depicted in Figure 18b and c, respectively).

To further expand the potential of network exploration, different information related to therapeutic, biological, and network features has been annotated for nodes, so that the user has multiple node coloring options to render the networks being assisted in network analysis. Among coloring options, network properties as degree, centrality measures or node grouping are available. In addition, Figure 18b illustrates how therapeutic information could be

examined at a glance by coloring the DN according to ATC code, whereas Figure 18c displays the TN in which nodes are color coded based on the protein class. In addition, COVIDrugNet offers other coloring option according to protein family or cellular location. Furthermore, the main section of COVIDrugNet contains (i) the *Charts and Plots* in which pie or bar charts are updated on the basis of the displayed network properties allowing the user to capture the relative proportions among variables of that property as well as the degree distribution plot, and (ii) the *Graph Properties* reporting centrality measures accounting for network topology.

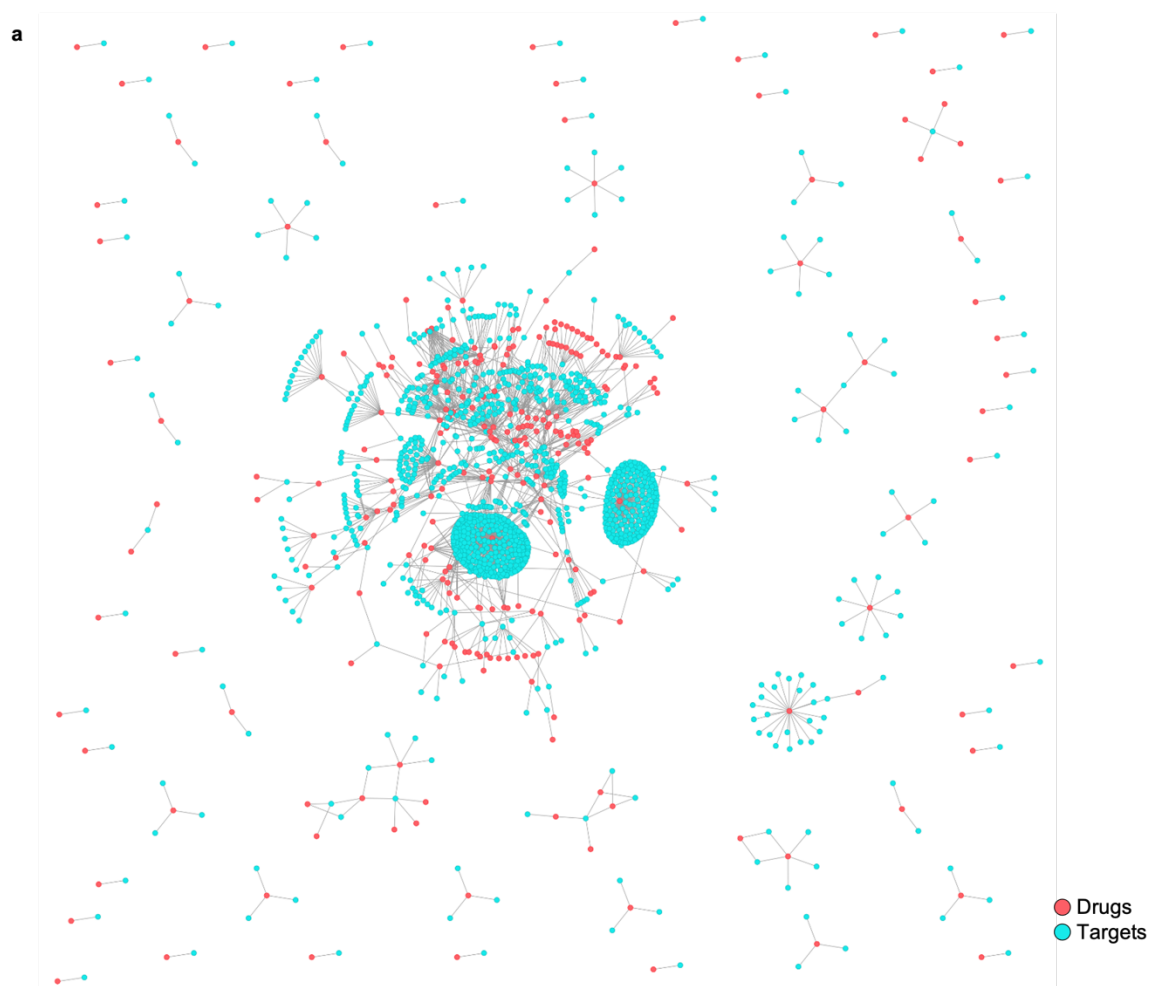




Figure 18. COVIDDrugNet networks provided by the web tool updated on 31st January 2022. The figure shows the drug-target bipartite network **(a)** in which red and cyan nodes represent drugs and targets, respectively; the drug network **(b)** which contains only drug nodes colored according to the first level ATC codes reported in DrugBank; the target network **(c)** that includes only targets. In the latter, the nodes are color coded according to their protein class collected from ChEMBL [98].

Moreover, the web tool gives the opportunity to perform network analysis through advanced tools. The drop-down *Advanced Tools* section contains *Clustering*, *Advanced Degree Distribution* and *Current Virus–Host–Drug Interactome*. *Clustering* is designed for grouping analysis that might reveal possible trends in repurposed drugs. The networks partitioning algorithms implemented are spectral analysis combined with K-means clustering [99], Girvan–Newman [100] and greedy modularity community detection [101] methods. *Advanced Degree Distribution* contains the degree distribution interactive chart with different distribution fittings compared to those of an Erdős–Rényi equivalent graph.

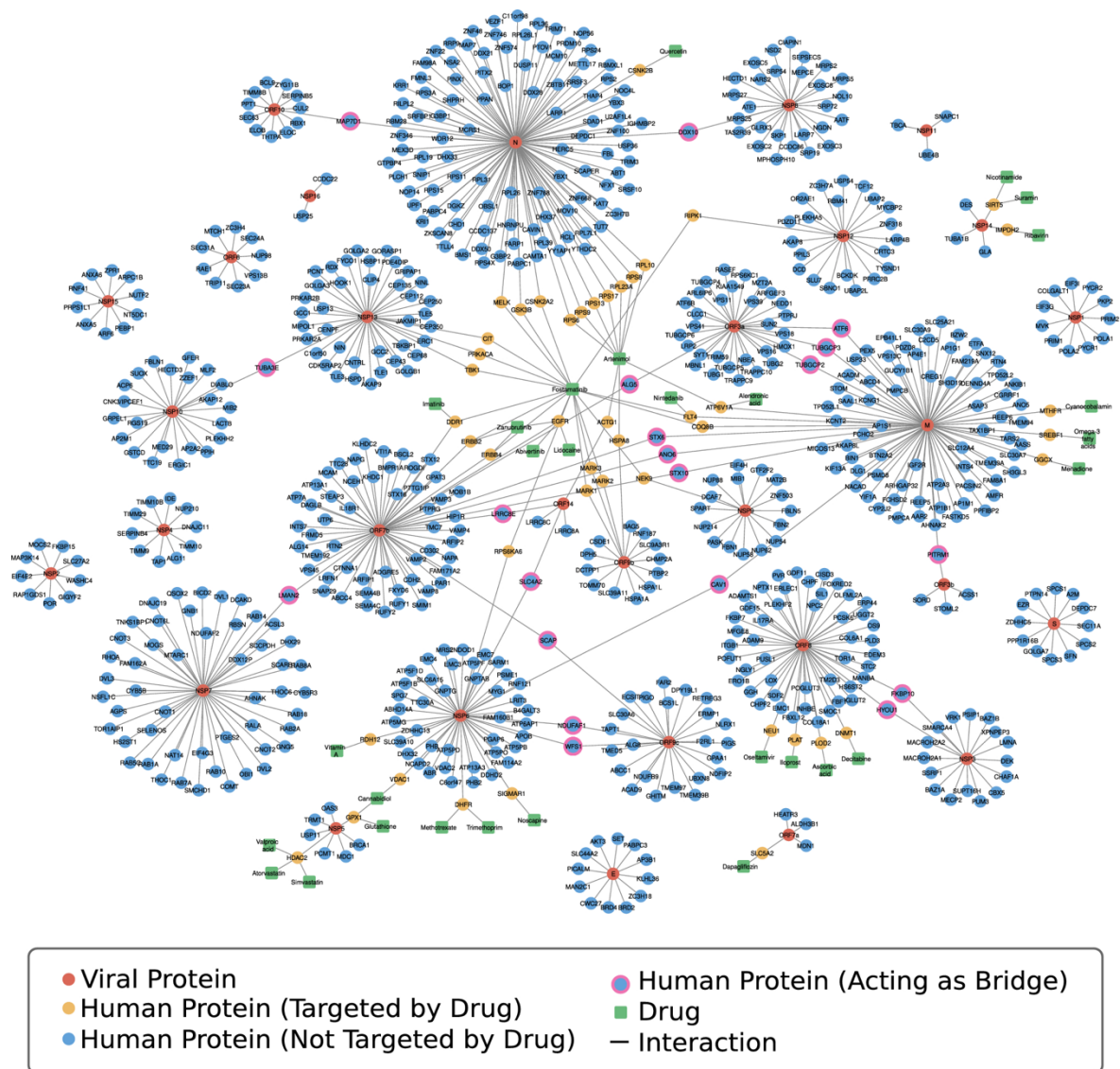


Figure 19. Virus–Host–Drug network adapted from [97]. The network is based on the virus–host interactome of Gordon et al. [102] and Chen et al. [103] from which the virus proteins shown as red nodes and the human proteins were retrieved from. The human proteins are represented as blue nodes if not targeted by any repurposed drugs and as yellow nodes if targeted by drug according to the TN; the human proteins linking two viral proteins, thus acting as bridge, are distinguished by a pink border. Drugs are depicted as square green nodes. The network was rendered by means of Cytoscape 3.8.2 [7].

Thanks to COVIDrugNet, we carried out a dual perspective analysis joining the data on repurposed drugs from COVIDrugNet itself and the molecular experimental data on SARS-CoV-2. The latter were retrieved from the works of Gordon et al. [102] and Chen et al. [104] which published virus-host interactomes. Merging their data, 732 human proteins that bound directly to the viral proteome were extracted. This list was compared with that of the human targets of the COVIDrugNet TN and, as a result, only 45 out of 732 proteins matched. Figure 19 shows the *virus-host-drug* network highlighting the 45 proteins in common which are depicted as yellow nodes. From the drug-target network emerges that these human proteins are targeted by 29 repurposed drugs (represented by square green nodes in Figure 19). The *virus-host-drug* network also shows 20 human proteins that act as bridges between two viral proteins appearing to be potential targets to disrupt the network of host-virus PPIs. The *Current Virus-Host-Drug Interactome* is displayed in the last section of COVIDrugNet. In an effort to provide current information, we are updating COVIDrugNet every two weeks. We observe that there have not been changes since last 31st January 2022.

16 CONCLUSIONS

Novel coronavirus is still hampering the healthcare system and the worldwide socio-economic condition. Since COVID-19 emerged in December 2019, huge steps forward have been made on the understanding of SARS-CoV-2 viral infection and replication mechanisms, epidemiology and clinical manifestation. Concurrently, COVID-19 vaccines and pharmacological treatments have been developed and received approval for emergency use.

In this context, however, we noticed an unmet need of depicting the continuous evolving scenario of the ongoing drug clinical trials. Hence, we developed COVIDrugNet, as a freely accessible online tool that helps to monitor drug research progresses about COVID-19 repurposed drugs.

Here, a brief description of how COVIDrugNet features could be employed to have a global insight into the molecular networks of drugs in clinical trials and their targets was proposed. Then, it was suggested that COVIDrugNet might be useful to probe the consistency of actual treatments with the biological evidence on virus infection. The target network has been *overlapped* with the host-virus interactome to offer new perspectives on drugs to be proposed for clinical investigation. For instance, human proteins linking two viral proteins might be promising targets for anti-COVID-19 drugs.

Finally, network-based approach to drug-related data as COVIDrugNet might help to understand the molecular implications of several COVID-19 pharmacological interventions that are still lacking a solid pharmacological rationale [105].

In conclusion, COVIDrugNet was designed to keep the user up to date on the advances of drug repurposing to contrast SARS-CoV-2 infection.

REFERENCES

- [1] A. L. Hopkins, “Network pharmacology: the next paradigm in drug discovery,” *Nat. Chem. Biol.* 2008 411, vol. 4, no. 11, pp. 682–690, Oct. 2008, doi: 10.1038/nchembio.118.
- [2] J. van der Greef and R. N. McBurney, “Rescuing drug discovery: In vivo systems pathology and systems pharmacology,” *Nat. Rev. Drug Discov.*, vol. 4, no. 12, pp. 961–967, Dec. 2005, doi: 10.1038/nrd1904.
- [3] E. C. Butcher, E. L. Berg, and E. J. Kunkel, “Systems biology in drug discovery,” *Nat. Biotechnol.* 2004 2210, vol. 22, no. 10, pp. 1253–1259, Oct. 2004, doi: 10.1038/nbt1017.
- [4] P. K. Sorger *et al.*, “Quantitative and Systems Pharmacology in the Post-genomic Era: New Approaches to Discovering Drugs and Understanding Therapeutic Mechanisms,” *An NIH white Pap. by QSP Work. Gr.*, vol. 48, pp. 1–47, 2011.
- [5] A. Anighoro, J. Bajorath, and G. Rastelli, “Polypharmacology: Challenges and Opportunities in Drug Discovery,” *J. Med. Chem.*, vol. 57, no. 19, pp. 7874–7887, Oct. 2014, doi: 10.1021/JM5006463.
- [6] L. Euler, “Solutio problematis ad geometriam situs pertinentis,” in *Commentarii academiae scientiarum Petropolitanae*, 1741, pp. 128–140.
- [7] P. Shannon *et al.*, “Cytoscape: A software Environment for integrated models of biomolecular interaction networks,” *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.
- [8] M. Bastian, S. Heymann, and M. Jacomy, “Gephi: An Open Source Software for Exploring and Manipulating Networks,” *Third Int. AAAI Conf. Weblogs Soc. Media*, 2009, doi: 10.1136/qshc.2004.010033.
- [9] A. Mrvar and V. Batagelj, “Analysis and visualization of large networks with program package Pajek,” *Complex Adapt. Syst. Model.*, vol. 4, no. 1, pp. 1–8, Dec. 2016, doi: 10.1186/S40294-016-0017-8.
- [10] G. Csardi, T. Nepusz, and others, “The igraph software package for complex network research,” *InterJournal, complex Syst.*, vol. 1695, no. 5, pp. 1–9, 2006.
- [11] T. M. J. Fruchterman and E. M. Reingold, “Graph drawing by force-directed placement,” *Softw. Pract. Exp.*, vol. 21, no. 11, pp. 1129–1164, Nov. 1991, doi: 10.1002/SPE.4380211102.
- [12] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring network structure, dynamics, and function using NetworkX,” in *Proceedings of the 7th Python in Science*

- Conference (SciPy 2008)*, 2008, pp. 11–15.
- [13] D. Simoni *et al.*, “Structure-activity relationship studies of novel heteroretinoids: Induction of apoptosis in the HL-60 cell line by a novel isoxazole-containing heteroretinoid,” *J. Med. Chem.*, vol. 42, no. 24, pp. 4961–4969, Dec. 1999, doi: 10.1021/jm991059n.
- [14] D. Simoni *et al.*, “Programmed cell death (PCD) associated with the stilbene motif of arotinoids: Discovery of novel apoptosis inducer agents possessing activity on multidrug resistant tumor cells,” *Bioorganic Med. Chem. Lett.*, vol. 10, no. 23, pp. 2669–2673, Dec. 2000, doi: 10.1016/S0960-894X(00)00547-3.
- [15] D. Simoni *et al.*, “Heterocycle-containing retinoids. Discovery of a novel isoxazole arotinoid possessing potent apoptotic activity in multidrug and drug-induced apoptosis-resistant cells,” *J. Med. Chem.*, vol. 44, no. 14, pp. 2308–2318, Jul. 2001, doi: 10.1021/jm0010320.
- [16] M. Roberti *et al.*, “Synthesis and biological evaluation of resveratrol and analogues as apoptosis-inducing agents,” *J. Med. Chem.*, vol. 46, no. 16, pp. 3546–3554, Jul. 2003, doi: 10.1021/jm030785u.
- [17] M. Roberti *et al.*, “Identification of a terphenyl derivative that blocks the cell cycle in the G0-G1 phase and induces differentiation in leukemia cells,” *J. Med. Chem.*, vol. 49, no. 10, pp. 3012–3018, May 2006, doi: 10.1021/jm060253o.
- [18] D. Simoni *et al.*, “Stilbene-based anticancer agents: Resveratrol analogues active toward HL60 leukemic cells with a non-specific phase mechanism,” *Bioorganic Med. Chem. Lett.*, vol. 16, no. 12, pp. 3245–3248, Jun. 2006, doi: 10.1016/j.bmcl.2006.03.028.
- [19] D. Pizzirani *et al.*, “Antiproliferative Agents That Interfere with the Cell Cycle at the G1→S Transition: Further Development and Characterization of a Small Library of Stilbene-Derived Compounds,” *ChemMedChem*, vol. 3, no. 2, pp. 345–355, Feb. 2008, doi: 10.1002/cmde.200700258.
- [20] S. Grimaudo *et al.*, “Apoptotic effects of thiazolobenzimidazole derivatives on sensitive and multidrug resistant leukaemic cells,” *Eur. J. Cancer*, vol. 37, no. 1, pp. 122–130, Jan. 2001, doi: 10.1016/S0959-8049(00)00372-5.
- [21] M. Tolomeo *et al.*, “Effects of chemically modified tetracyclines (CMTs) in sensitive, multidrug resistant and apoptosis resistant leukaemia cell lines,” *Br. J. Pharmacol.*, vol. 133, no. 2, pp. 306–314, 2001, doi: 10.1038/sj.bjp.0704068.
- [22] D. Simoni *et al.*, “Heterocyclic and phenyl double-bond-locked combretastatin analogues possessing potent apoptosis-inducing activity in HL60 and in MDR cell

- lines,” *J. Med. Chem.*, vol. 48, no. 3, pp. 723–736, Feb. 2005, doi: 10.1021/jm049622b.
- [23] D. Pizzirani *et al.*, “Identification of biphenyl-based hybrid molecules able to decrease the intracellular level of Bcl-2 protein in Bcl-2 overexpressing leukemia cells,” *J. Med. Chem.*, vol. 52, no. 21, pp. 6936–6940, Nov. 2009, doi: 10.1021/jm900907s.
- [24] D. Raffa *et al.*, “Synthesis, antiproliferative activity, and mechanism of action of a series of 2-{[(2E)-3-phenylprop-2-enoyl]amino}benzamides,” *Eur. J. Med. Chem.*, vol. 46, no. 7, pp. 2786–2796, Jul. 2011, doi: 10.1016/j.ejmech.2011.03.067.
- [25] E. Giacomini *et al.*, “Novel antiproliferative chimeric compounds with marked histone deacetylase inhibitory activity,” *ACS Med. Chem. Lett.*, vol. 5, no. 9, pp. 973–978, Sep. 2014, doi: 10.1021/ml5000959.
- [26] R. Romagnoli *et al.*, “Novel iodoacetamido benzoheterocyclic derivatives with potent antileukemic activity are inhibitors of STAT5 phosphorylation,” *Eur. J. Med. Chem.*, vol. 108, pp. 39–52, Jan. 2016, doi: 10.1016/j.ejmech.2015.11.022.
- [27] R. Rondanin *et al.*, “Inhibition of activated STAT5 in Bcr/Abl expressing leukemia cells with new pimozone derivatives,” *Bioorg. Med. Chem. Lett.*, vol. 24, pp. 4568–4574, 2014, doi: 10.1016/j.bmcl.2014.07.069.
- [28] R. Rondanin *et al.*, “Effects of Pimozone Derivatives on pSTAT5 in K562 Cells,” *ChemMedChem*, vol. 12, no. 15, pp. 1183–1190, Aug. 2017, doi: 10.1002/cmdc.201700234.
- [29] T. R. J. Evans and S. B. Kaye, “Retinoids: present role and future potential,” *Br. J. Cancer*, vol. 80, no. 1–2, p. 1, 1999, doi: 10.1038/SJ.BJC.6690312.
- [30] D. A. Arber *et al.*, “The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia,” *Blood*, vol. 127, no. 20, pp. 2391–2405, May 2016, doi: 10.1182/BLOOD-2016-03-643544.
- [31] American Cancer Society, “Cancer Facts & Figures 2022,” *Atlanta, Ga: American Cancer Society*, 2022. <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2022/2022-cancer-facts-and-figures.pdf> (accessed Jan. 24, 2022).
- [32] E. Papaemmanuil *et al.*, “Genomic Classification and Prognosis in Acute Myeloid Leukemia,” *N. Engl. J. Med.*, vol. 374, no. 23, pp. 2209–2221, Jun. 2016, doi: 10.1056/NEJMOA1516192.
- [33] D. Thomas and R. Majeti, “Biology and relevance of human acute myeloid leukemia stem cells,” *Blood*, vol. 129, no. 12, pp. 1577–1585, Mar. 2017, doi: 10.1182/BLOOD-2016-10-696054.

- [34] J. L. Carter *et al.*, “Targeting multiple signaling pathways: the new approach to acute myeloid leukemia therapy,” *Signal Transduct. Target. Ther.* 2020 51, vol. 5, no. 1, pp. 1–29, Dec. 2020, doi: 10.1038/s41392-020-00361-x.
- [35] F. K. Brown and others, “Chemoinformatics: what is it and how does it impact drug discovery,” *Annu. Rep. Med. Chem.*, vol. 33, pp. 375–384, 1998.
- [36] Bf. Begam and Js. Kumar, “A Study on Cheminformatics and its Applications on Modern Drug Discovery,” *Procedia Eng.*, vol. 38, pp. 1264–1275, 2012, doi: 10.1016/j.proeng.2012.06.156.
- [37] V. Arulmozhi and R. Rajesh, “Chemoinformatics - A quick review,” *ICECT 2011 - 2011 3rd Int. Conf. Electron. Comput. Technol.*, vol. 6, pp. 416–419, 2011, doi: 10.1109/ICECTECH.2011.5942128.
- [38] W. F. Lunnion, J. Brunvoll, S. J. Cyvin, B. N. Cyvin, and A. T. Balaban, “SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules,” *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: 10.1021/CI00057A005.
- [39] D. Rogers and M. Hahn, “Extended-Connectivity Fingerprints,” *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, May 2010, doi: 10.1021/CI100050T.
- [40] R. S. Bohacek, C. Mcmartin, and W. C. Guida, “The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective,” *Med. Res. Rev.*, vol. 16, no. 1, pp. 3–50, 1996, doi: 10.1002/(SICI)1098-1128(199601)16:1.
- [41] J. Medina-Franco, K. Martinez-Mayorga, M. Giulianotti, R. Houghten, and C. Pinilla, “Visualization of the Chemical Space in Drug Discovery,” *Curr. Comput. Aided-Drug Des.*, vol. 4, no. 4, pp. 322–333, Dec. 2008, doi: 10.2174/157340908786786010.
- [42] C. M. Dobson, “Chemical space and biology,” *Nature*, vol. 432, no. 7019, pp. 824–828, Dec. 2004, doi: 10.1038/NATURE03192.
- [43] G. M. Maggiora and J. Bajorath, “Chemical space networks: A powerful new paradigm for the description of chemical space,” *J. Comput. Aided. Mol. Des.*, vol. 28, no. 8, pp. 795–802, Jun. 2014, doi: 10.1007/S10822-014-9760-0.
- [44] M. Vogt, D. Stumpfe, G. M. Maggiora, and J. Bajorath, “Lessons learned from the design of chemical space networks and opportunities for new applications,” *J. Comput. Aided. Mol. Des.*, vol. 30, no. 3, pp. 191–208, Mar. 2016, doi: 10.1007/S10822-016-9906-3.
- [45] M. Zwierzyna, M. Vogt, G. M. Maggiora, and J. Bajorath, “Design and characterization of chemical space networks for different compound data sets,” *J.*

- Comput. Aided. Mol. Des.*, vol. 29, no. 2, pp. 113–125, 2015, doi: 10.1007/S10822-014-9821-4.
- [46] B. Zhang, M. Vogt, G. M. Maggiora, and J. Bajorath, “Comparison of bioactive chemical space networks generated using substructure- and fingerprint-based measures of molecular similarity,” *J. Comput. Aided. Mol. Des.*, vol. 29, no. 7, pp. 595–608, Jul. 2015, doi: 10.1007/s10822-015-9852-5.
- [47] B. Zhang, M. Vogt, G. M. Maggiora, and J. Bajorath, “Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures,” *J. Comput. Aided. Mol. Des.*, vol. 29, no. 10, pp. 937–950, Oct. 2015, doi: 10.1007/S10822-015-9872-1.
- [48] M. Wu, M. Vogt, G. M. Maggiora, and J. Bajorath, “Design of chemical space networks on the basis of Tversky similarity,” *J. Comput. Aided. Mol. Des.*, vol. 30, no. 1, pp. 1–12, Jan. 2016, doi: 10.1007/S10822-015-9891-Y.
- [49] D. Stumpfe and J. Bajorath, “Recent developments in SAR visualization,” *Medchemcomm*, vol. 7, no. 6, pp. 1045–1055, Jun. 2016, doi: 10.1039/C6MD00108D.
- [50] M. Wawer, L. Peltason, N. Weskamp, A. Teckentrup, and J. Bajorath, “Structure–Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure–Activity Relationship Indices,” *J. Med. Chem.*, vol. 51, no. 19, pp. 6075–6084, Oct. 2008, doi: 10.1021/JM800867G.
- [51] “Schrödinger Release 2021-1: Maestro.” Schrödinger, LLC, New York, NY, 2021.
- [52] “RDKit: Open-source cheminformatics.” <https://www.rdkit.org>.
- [53] “ChemViz2: Cheminformatics App for Cytoscape.” <http://www.rbvi.ucsf.edu/cytoscape/chemViz2/>.
- [54] D. Stumpfe, Y. Hu, D. Dimova, and J. Bajorath, “Recent progress in understanding activity cliffs and their utility in medicinal chemistry,” *J. Med. Chem.*, vol. 57, no. 1, pp. 18–28, Jan. 2014, doi: 10.1021/jm401120g.
- [55] D. S. Wishart *et al.*, “DrugBank 5.0: a major update to the DrugBank database for 2018,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018, doi: 10.1093/NAR/GKX1037.
- [56] M. A. Yildirim, K. Il Goh, M. E. Cusick, A. L. Barabási, and M. Vidal, “Drug-target network,” *Nat. Biotechnol.*, vol. 25, no. 10, pp. 1119–1126, Oct. 2007, doi: 10.1038/NBT1338.
- [57] M. Recanatini and C. Cabrelle, “Drug research meets network science: Where are we?,” *J. Med. Chem.*, vol. 63, no. 16, pp. 8653–8666, Aug. 2020, doi: 10.1021/acs.jmedchem.9b01989.

- [58] A. Ezzat, M. Wu, X.-L. Li, and C.-K. Kwoh, “Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey,” *Brief. Bioinform.*, vol. 20, no. 4, pp. 1337–1357, Jul. 2019, doi: 10.1093/bib/bby002.
- [59] M. Bagherian, E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska, and K. Najarian, “Machine learning approaches and databases for prediction of drug–target interaction: a survey paper,” *Brief. Bioinform.*, vol. 22, no. 1, pp. 247–269, Jan. 2021, doi: 10.1093/bib/bbz157.
- [60] S. Alaimo, R. Giugno, and A. Pulvirenti, “Recommendation Techniques for Drug–Target Interaction Prediction and Drug Repositioning,” *Methods Mol. Biol.*, vol. 1415, pp. 441–462, Aug. 2016, doi: 10.1007/978-1-4939-3572-7_23.
- [61] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, “Network propagation: a universal amplifier of genetic associations,” *Nat. Publ. Gr.*, vol. 18, 2017, doi: 10.1038/nrg.2017.38.
- [62] F. Cheng *et al.*, “Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference,” *PLOS Comput. Biol.*, vol. 8, no. 5, p. e1002503, May 2012, doi: 10.1371/JOURNAL.PCBI.1002503.
- [63] S. Alaimo, A. Pulvirenti, R. Giugno, and A. Ferro, “Drug–target interaction prediction through domain-tuned network-based inference,” *Bioinformatics*, vol. 29, no. 16, pp. 2004–2008, Aug. 2013, doi: 10.1093/BIOINFORMATICS/BTT307.
- [64] L. Page and S. Brin, “The anatomy of a large-scale hypertextual Web search engine,” *Comput. Networks ISDN Syst.*, vol. 30, no. 1–7, pp. 107–117, Apr. 1998, doi: 10.1016/S0169-7552(98)00110-X.
- [65] X. Chen, M. X. Liu, and G. Y. Yan, “Drug–target interaction prediction by random walk on the heterogeneous network,” *Mol. Biosyst.*, vol. 8, no. 7, pp. 1970–1978, Jun. 2012, doi: 10.1039/C2MB00002D.
- [66] A. Seal, Y. Y. Ahn, and D. J. Wild, “Optimizing drug–target interaction prediction based on random walk on heterogeneous networks,” *J. Cheminform.*, vol. 7, no. 1, pp. 1–12, Aug. 2015, doi: 10.1186/S13321-015-0089-Z.
- [67] M. Kanehisa, “Toward understanding the origin and evolution of cellular organisms,” *Protein Sci.*, vol. 28, no. 11, pp. 1947–1951, Nov. 2019, doi: 10.1002/PRO.3715.
- [68] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/NAR/28.1.27.
- [69] D. S. Wishart *et al.*, “DrugBank: a comprehensive resource for in silico drug discovery and exploration,” *Nucleic Acids Res.*, vol. 34, no. Database issue, 2006, doi: 10.1093/nar/gkj067.

- [70] C. Knox *et al.*, “DrugBank 3.0: a comprehensive resource for ‘Omics’ research on drugs,” *Nucleic Acids Res.*, vol. 39, no. suppl_1, pp. D1035–D1041, Jan. 2011, doi: 10.1093/NAR/GKQ1126.
- [71] M. Ali and A. Ezzat, “DrugBank Database XML Parser,” *Dainanahan*. 2020, [Online]. Available: <https://cran.r-project.org/package=dbparser>.
- [72] Dan Tenenbaum and Bioconductor Package Maintainer, “KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG).” 2021.
- [73] Y. Cao, A. Charisi, L. C. Cheng, T. Jiang, and T. Girke, “ChemmineR: a compound mining framework for R,” *Bioinformatics*, vol. 24, no. 15, pp. 1733–1734, Aug. 2008, doi: 10.1093/BIOINFORMATICS/BTN307.
- [74] T. F. Smith and M. S. Waterman, “Identification of common molecular subsequences,” *J. Mol. Biol.*, vol. 147, no. 1, pp. 195–197, Mar. 1981, doi: 10.1016/0022-2836(81)90087-5.
- [75] H. Pagès, P. Aboyou, R. Gentleman, and S. DebRoy, “Biostrings: Efficient manipulation of biological strings.” 2021, [Online]. Available: <https://bioconductor.org/packages/Biostrings>.
- [76] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, “Prediction of drug-target interaction networks from the integration of chemical and genomic spaces,” *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, Jul. 2008, doi: 10.1093/bioinformatics/btn162.
- [77] A. Seal and D. J. Wild, “Netpredictor: R and Shiny package to perform drug-target network analysis and prediction of missing links,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–10, Jul. 2018, doi: 10.1186/S12859-018-2254-7.
- [78] Z. Xie *et al.*, “Gene Set Knowledge Discovery with Enrichr,” *Curr. Protoc.*, vol. 1, no. 3, p. e90, Mar. 2021, doi: 10.1002/CPZ1.90.
- [79] E. Y. Chen *et al.*, “Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool,” *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–14, Apr. 2013, doi: 10.1186/1471-2105-14-128.
- [80] T. Wu *et al.*, “clusterProfiler 4.0: A universal enrichment tool for interpreting omics data,” *Innov.*, vol. 2, no. 3, p. 100141, Aug. 2021, doi: 10.1016/J.XINN.2021.100141.
- [81] J. F. Truchon and C. I. Bayly, “Evaluating virtual screening methods: Good and bad metrics for the ‘early recognition’ problem,” *J. Chem. Inf. Model.*, vol. 47, no. 2, pp. 488–508, Mar. 2007, doi: 10.1021/CI600426E.
- [82] L. S. Steelman *et al.*, “Contributions of the Raf/MEK/ERK, PI3K/PTEN/Akt/mTOR and Jak/STAT pathways to leukemia,” *Leuk. 2008 224*, vol. 22, no. 4, pp. 686–707,

- Mar. 2008, doi: 10.1038/leu.2008.26.
- [83] S. Binder, M. Luciano, and J. Horejs-Hoeck, “The cytokine network in acute myeloid leukemia (AML): A focus on pro- and anti-inflammatory mediators,” *Cytokine Growth Factor Rev.*, vol. 43, pp. 8–15, Oct. 2018, doi: 10.1016/J.CYTOGFR.2018.08.004.
- [84] S. Park *et al.*, “Role of the PI3K/AKT and mTOR signaling pathways in acute myeloid leukemia,” *Haematologica*, vol. 95, no. 5, p. 819, May 2010, doi: 10.3324/HAEMATOL.2009.013797.
- [85] I. Nepstad, K. J. Hatfield, I. S. Grønningseter, and H. Reikvam, “The PI3K-Akt-mTOR Signaling Pathway in Human Acute Myeloid Leukemia (AML) Cells,” *Int. J. Mol. Sci.*, vol. 21, no. 8, Apr. 2020, doi: 10.3390/IJMS21082907.
- [86] W. Vainchenker and S. N. Constantinescu, “JAK/STAT signaling in hematological malignancies,” *Oncogene*, vol. 32, no. 21, pp. 2601–2613, May 2013, doi: 10.1038/ONC.2012.347.
- [87] M. Milella *et al.*, “Therapeutic targeting of the MEK/MAPK signal transduction module in acute myeloid leukemia,” *J. Clin. Invest.*, vol. 108, no. 6, p. 851, 2001, doi: 10.1172/JCI12807.
- [88] L. S. Steelman *et al.*, “Roles of the Ras/Raf/MEK/ERK pathway in leukemia therapy,” *Leuk. 2011 257*, vol. 25, no. 7, pp. 1080–1094, Apr. 2011, doi: 10.1038/leu.2011.66.
- [89] M. C. J. Bosman, J. J. Schuringa, and E. Vellenga, “Constitutive NF- κ B activation in AML: Causes and treatment strategies,” *Crit. Rev. Oncol. Hematol.*, vol. 98, pp. 35–44, Feb. 2016, doi: 10.1016/J.CRITREVONC.2015.10.001.
- [90] S. K. Joshi *et al.*, “ERBB2/HER2 mutations are transforming and therapeutically targetable in leukemia,” *Leuk. 2020 3410*, vol. 34, no. 10, pp. 2798–2804, May 2020, doi: 10.1038/s41375-020-0844-7.
- [91] A. Puissant *et al.*, “SYK Is a Critical Regulator of FLT3 in Acute Myeloid Leukemia,” *Cancer Cell*, vol. 25, no. 2, pp. 226–242, Feb. 2014, doi: 10.1016/J.CCR.2014.01.022.
- [92] G. Yu, “enrichplot: Visualization of Functional Enrichment Result.” 2022, [Online]. Available: <https://yulab-smu.top/biomedical-knowledge-mining-book/>.
- [93] D. M. Van Der Kolk *et al.*, “Activity and expression of the multidrug resistance proteins P-glycoprotein, MRP1, MRP2, MRP3 and MRP5 in de novo and relapsed acute myeloid leukemia,” *Leuk. 2001 1510*, vol. 15, no. 10, pp. 1544–1553, Oct. 2001, doi: 10.1038/sj.leu.2402236.
- [94] M. Krajcinovic *et al.*, “Polymorphisms of ABCC5 and NOS3 genes influence doxorubicin cardiotoxicity in survivors of childhood acute lymphoblastic leukemia,”

- Pharmacogenomics J.* 2015 166, vol. 16, no. 6, pp. 530–535, Sep. 2015, doi: 10.1038/tpj.2015.63.
- [95] M. Deynoux, N. Sunter, O. Hérault, and F. Mazurier, “Hypoxia and hypoxia-inducible factors in leukemias,” *Front. Oncol.*, vol. 6, no. FEB, p. 41, 2016, doi: 10.3389/FONC.2016.00041.
- [96] M. Soltani, Y. Zhao, Z. Xia, M. Ganjalikhani Hakemi, and A. V. Bazhin, “The Importance of Cellular Metabolic Pathways in Pathogenesis and Selective Treatments of Hematological Malignancies,” *Front. Oncol.*, vol. 11, p. 4665, Nov. 2021, doi: 10.3389/FONC.2021.767026.
- [97] L. Menestrina, C. Cabrelle, and M. Recanatini, “COVIDrugNet: a network-based web tool to investigate the drugs currently in clinical trial to contrast COVID-19,” *Sci. Reports 2021 111*, vol. 11, no. 1, pp. 1–15, Sep. 2021, doi: 10.1038/s41598-021-98812-0.
- [98] A. Gaulton *et al.*, “The ChEMBL database in 2017,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, Jan. 2017, doi: 10.1093/nar/gkw1074.
- [99] U. Von Luxburg, “A tutorial on spectral clustering,” *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007, doi: 10.1007/s11222-007-9033-z.
- [100] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002, doi: 10.1073/pnas.122653799.
- [101] M. E. J. Newman, “Fast algorithm for detecting community structure in networks,” *Phys. Rev. E*, vol. 69, no. 6, p. 066133, Jun. 2004, doi: 10.1103/PhysRevE.69.066133.
- [102] D. E. Gordon *et al.*, “A SARS-CoV-2 protein interaction map reveals targets for drug repurposing,” *Nature*, vol. 583, no. 7816, pp. 459–468, Jul. 2020, doi: 10.1038/s41586-020-2286-9.
- [103] Z. Chen *et al.*, “Interactomes of SARS-CoV-2 and human coronaviruses reveal host factors potentially affecting pathogenesis,” *EMBO J.*, vol. 40, no. 17, Sep. 2021, doi: 10.15252/EMBJ.2021107776.
- [104] Z. Chen *et al.*, “Comprehensive analysis of the host-virus interactome of SARS-CoV-2,” *bioRxiv*, Jan. 2021, doi: 10.1101/2020.12.31.424961.
- [105] S. P. H. Alexander *et al.*, “A rational roadmap for SARS-CoV-2/COVID-19 pharmacotherapeutic research and development: IUPHAR Review 29,” *Br. J. Pharmacol.*, vol. 177, no. 21, pp. 4942–4966, Nov. 2020, doi: 10.1111/BPH.15094.

SUPPLEMENTARY MATERIALS

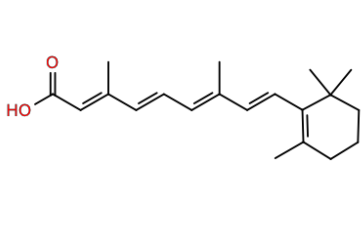
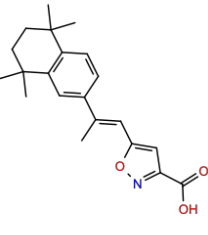
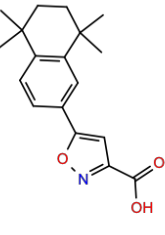
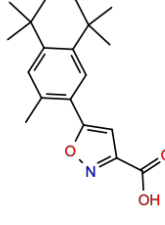
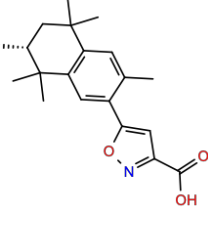
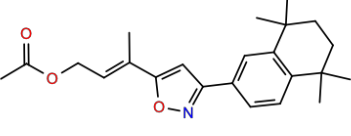
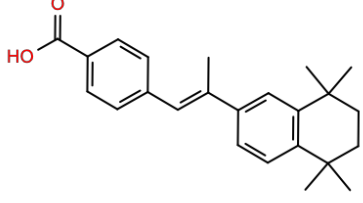
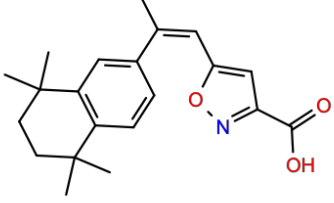
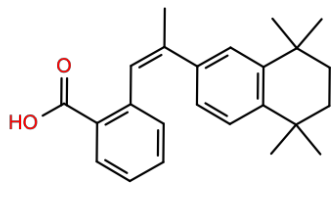
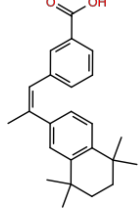
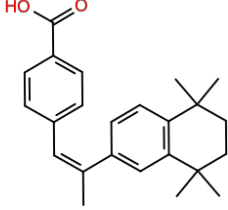
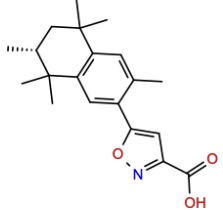
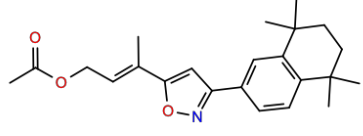
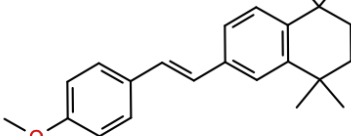
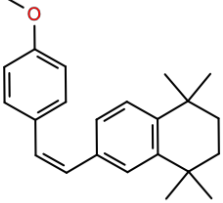
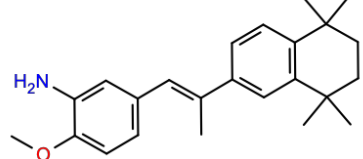
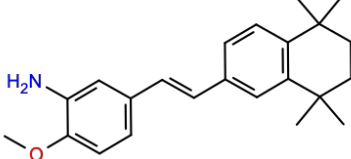
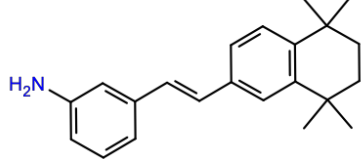
Supplementary Figure S1. The collection of 220 antiproliferative compounds. Their chemical structures are shown, and their antiproliferative activities are reported as pIC₅₀ values.

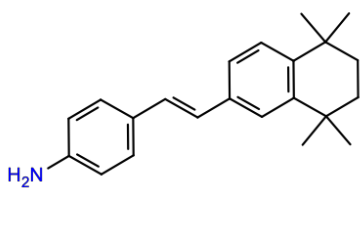
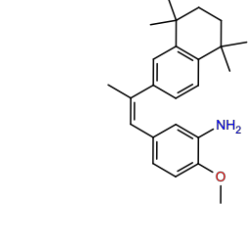
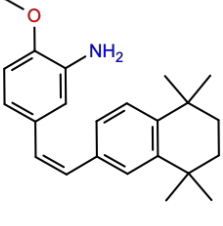
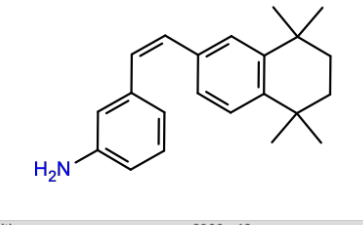
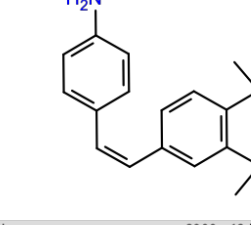
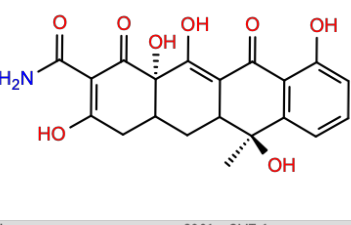
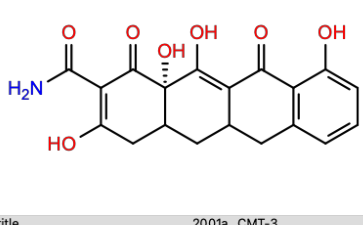
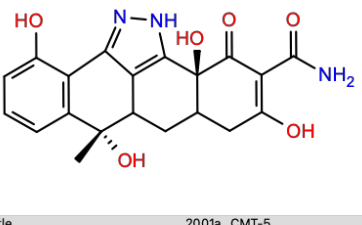
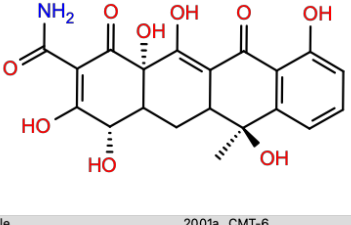
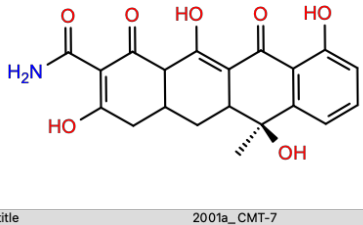
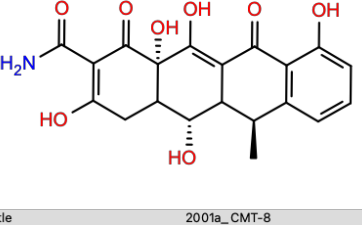
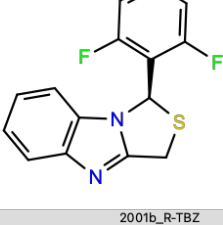
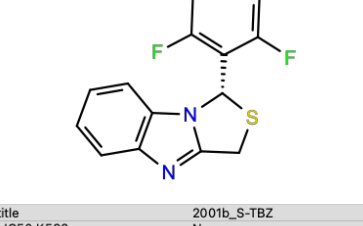
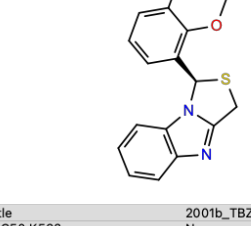
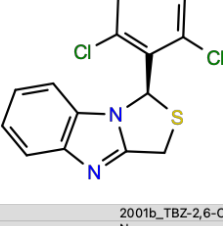
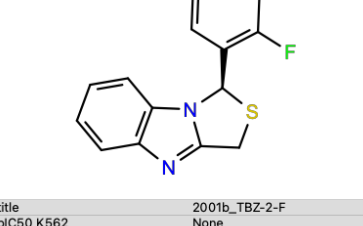
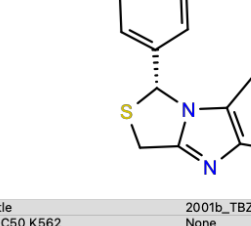
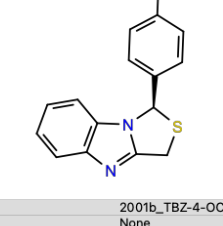
Supplementary Figure S2. Predicted compound-target interaction network for cell surface molecules and ligands class.

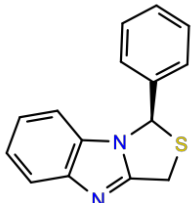
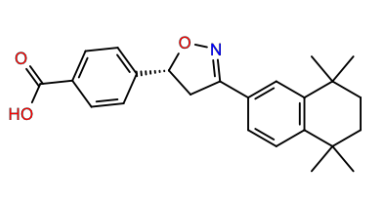
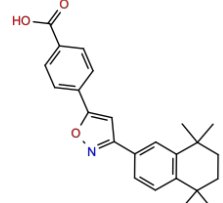
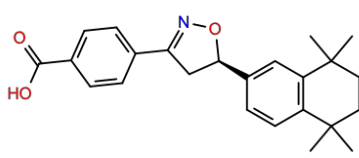
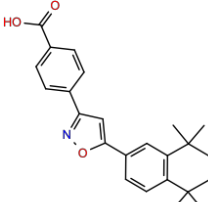
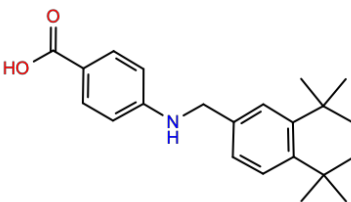
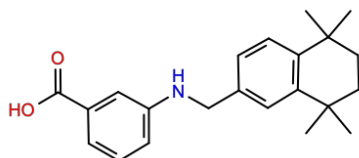
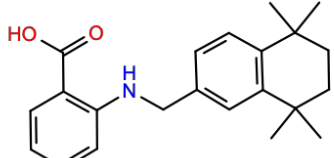
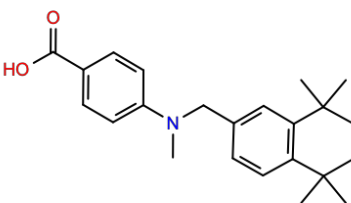
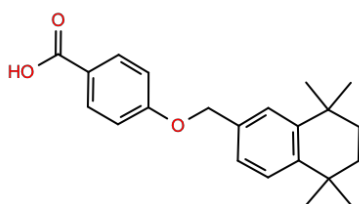
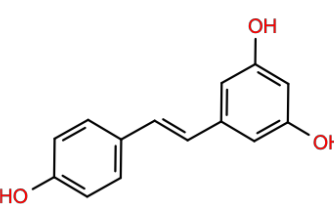
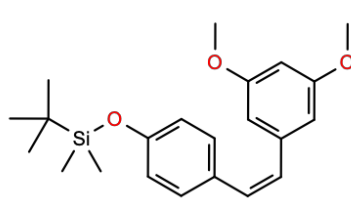
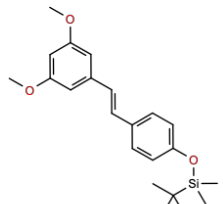
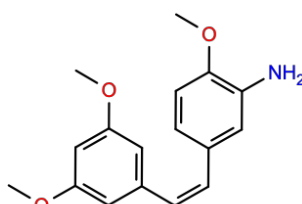
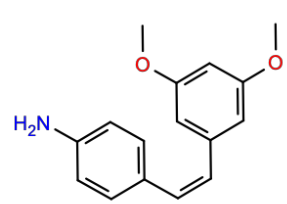
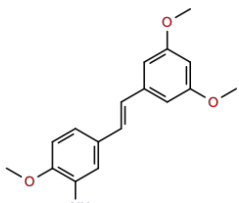
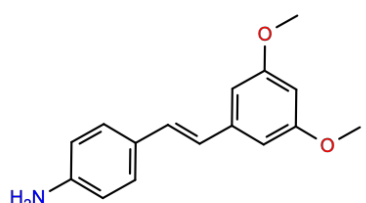
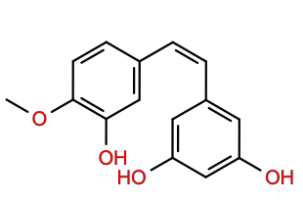
Supplementary Figure S3. Predicted compound-target interaction network for protein kinases class.

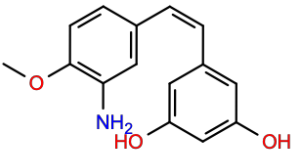
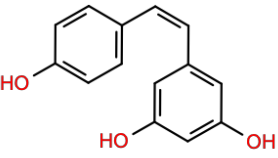
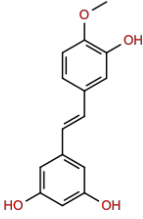
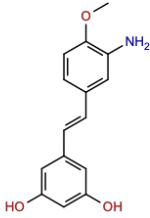
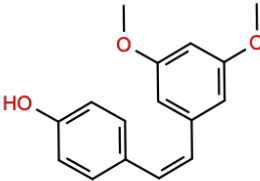
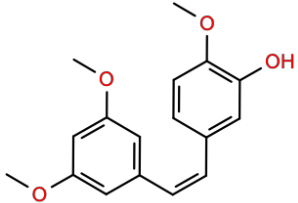
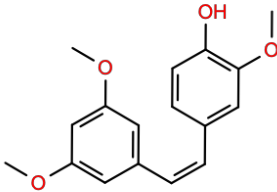
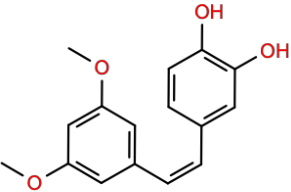
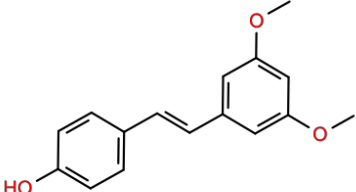
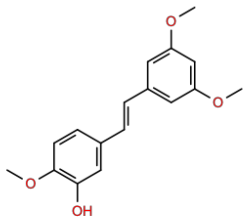
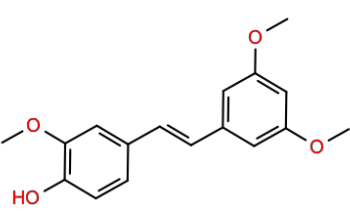
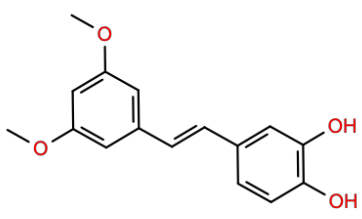
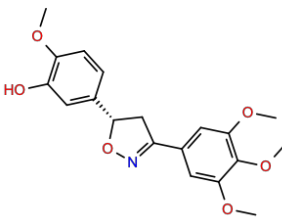
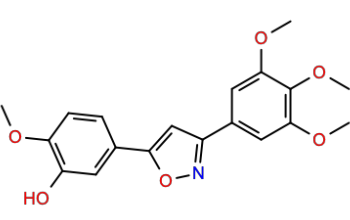
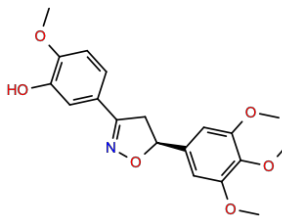
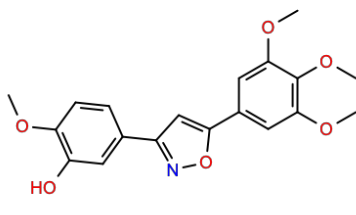
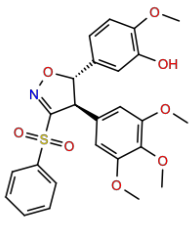
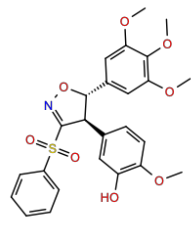
Supplementary Figure S4. Predicted compound-target interaction network for not elsewhere classified targets class.

Supplementary Figure S5. Functional enrichment bar plot for DrugBank target class for NCI-Nature pathways gene-set library.

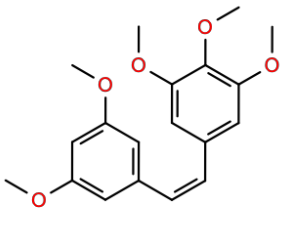
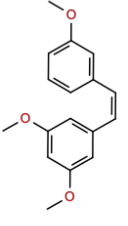
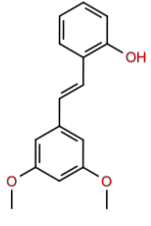
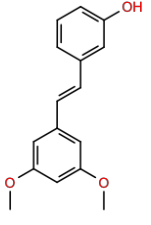
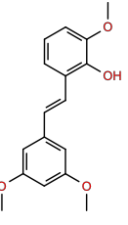
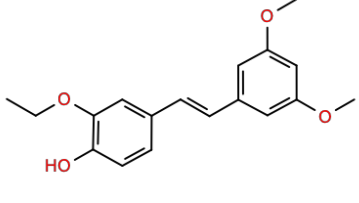
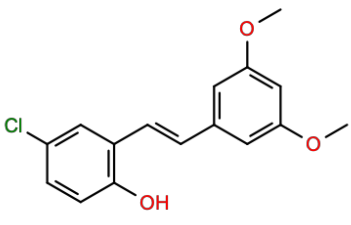
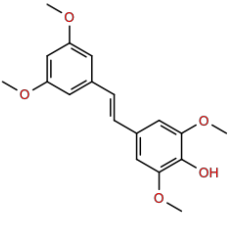
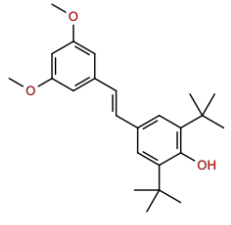
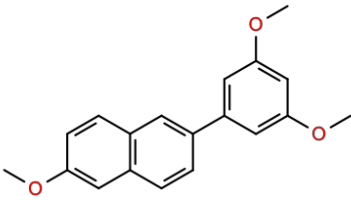
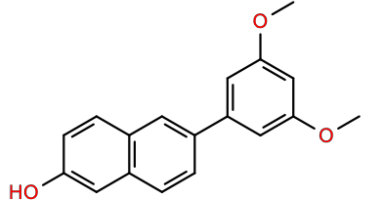
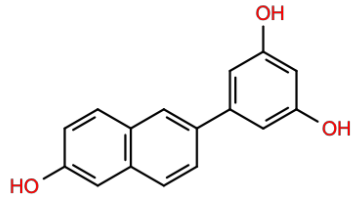
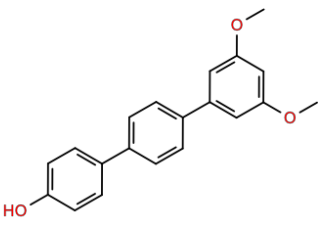
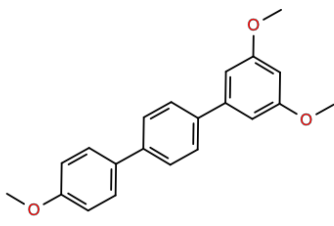
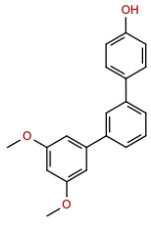
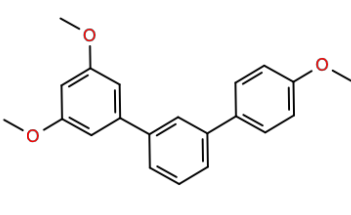
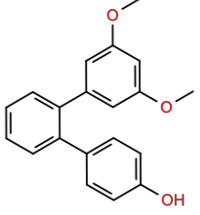
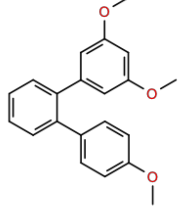
		
title 1999a_1 pIC50 K562 None pIC50 HL60 4.959	title 1999a_13b pIC50 K562 None pIC50 HL60 4.131	title 1999a_17 pIC50 K562 None pIC50 HL60 4.125
		
title 1999a_18b pIC50 K562 None pIC50 HL60 4.143	title 1999a_19 pIC50 K562 None pIC50 HL60 4.137	title 1999a_24 pIC50 K562 None pIC50 HL60 4.26
		
title 2000a_3 pIC50 K562 None pIC50 HL60 4.337	title 2000a_4 pIC50 K562 None pIC50 HL60 4.585	title 2000a_9a pIC50 K562 None pIC50 HL60 4.284
		
title 2000a_9b pIC50 K562 None pIC50 HL60 4.301	title 2000a_9c pIC50 K562 None pIC50 HL60 5.0	title 2000a_10a pIC50 K562 None pIC50 HL60 4.0
		
title 2000a_10b pIC50 K562 None pIC50 HL60 4.0	title 2000a_13e pIC50 K562 None pIC50 HL60 4.301	title 2000a_14e pIC50 K562 None pIC50 HL60 4.301
		
title 2000a_15a pIC50 K562 None pIC50 HL60 4.699	title 2000a_15b pIC50 K562 None pIC50 HL60 4.215	title 2000a_15c pIC50 K562 None pIC50 HL60 4.699

		
title 2000a_15d pIC50 K562 None pIC50 HL60 4.745	title 2000a_16a pIC50 K562 None pIC50 HL60 4.377	title 2000a_16b pIC50 K562 4.377 pIC50 HL60 4.602
		
title 2000a_16c pIC50 K562 4.886 pIC50 HL60 5.0	title 2000a_16d pIC50 K562 4.301 pIC50 HL60 4.745	title 2001a_CMT-1 pIC50 K562 None pIC50 HL60 5.301
		
title 2001a_CMT-3 pIC50 K562 None pIC50 HL60 5.155	title 2001a_CMT-5 pIC50 K562 None pIC50 HL60 4.398	title 2001a_CMT-6 pIC50 K562 None pIC50 HL60 4.42
		
title 2001a_CMT-7 pIC50 K562 None pIC50 HL60 4.658	title 2001a_CMT-8 pIC50 K562 None pIC50 HL60 5.456	title 2001b_TBZ pIC50 K562 None pIC50 HL60 4.602
		
title 2001b_TBZ-2-F pIC50 K562 None pIC50 HL60 4.125	title 2001b_TBZ-2,3-(OCH3)2 pIC50 K562 None pIC50 HL60 4.638	title 2001b_TBZ-2,6-Cl2 pIC50 K562 None pIC50 HL60 4.602
		
title 2001b_TBZ-2-F pIC50 K562 None pIC50 HL60 4.097	title 2001b_TBZ-3-OCH3 pIC50 K562 None pIC50 HL60 4.398	title 2001b_TBZ-4-OCH3 pIC50 K562 None pIC50 HL60 6.0

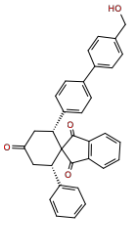
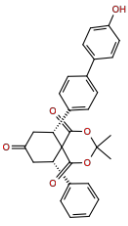
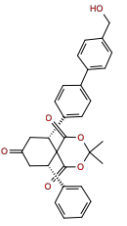
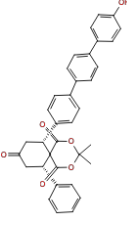
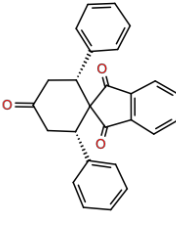
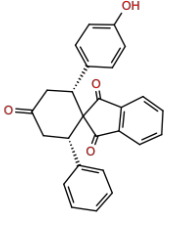
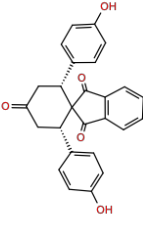
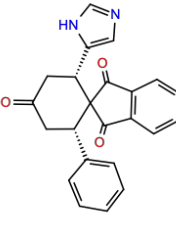
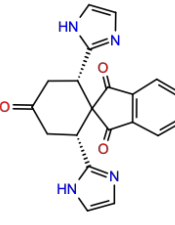
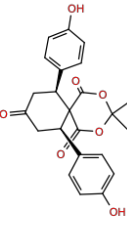
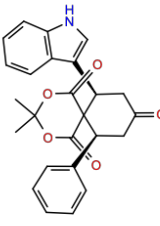
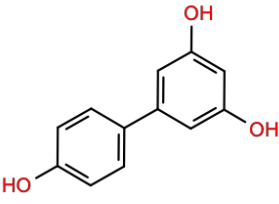
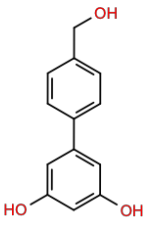
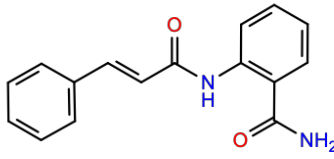
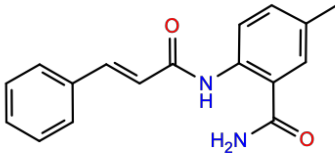
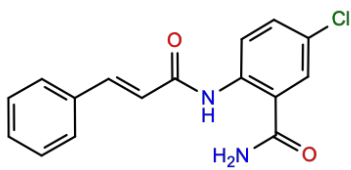
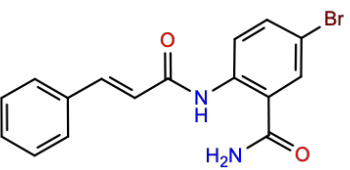
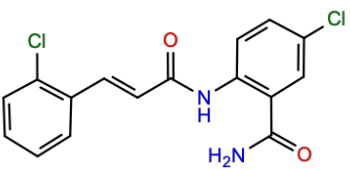
		
title 2001b_TBZ-H pIC50 K562 None pIC50 HL60 4.125	title 2001c_10b pIC50 K562 None pIC50 HL60 4.167	title 2001c_11b pIC50 K562 None pIC50 HL60 4.097
		
title 2001c_14b pIC50 K562 None pIC50 HL60 4.125	title 2001c_15b pIC50 K562 5.638 pIC50 HL60 6.0	title 2001c_17e pIC50 K562 None pIC50 HL60 4.398
		
title 2001c_17f pIC50 K562 None pIC50 HL60 4.222	title 2001c_17g pIC50 K562 None pIC50 HL60 4.31	title 2001c_17h pIC50 K562 None pIC50 HL60 4.237
		
title 2001c_20b pIC50 K562 None pIC50 HL60 4.0	title 2003a_1 pIC50 K562 4.553 pIC50 HL60 5.301	title 2003a_5e pIC50 K562 None pIC50 HL60 4.301
		
title 2003a_6e pIC50 K562 None pIC50 HL60 4.319	title 2003a_7b pIC50 K562 7.602 pIC50 HL60 7.523	title 2003a_7c pIC50 K562 None pIC50 HL60 5.602
		
title 2003a_8b pIC50 K562 5.602 pIC50 HL60 5.398	title 2003a_8c pIC50 K562 None pIC50 HL60 5.398	title 2003a_9a pIC50 K562 None pIC50 HL60 4.398

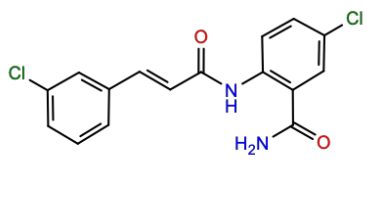
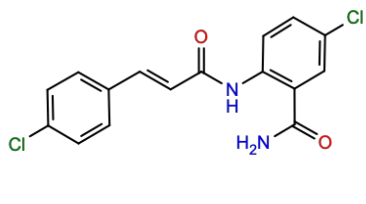
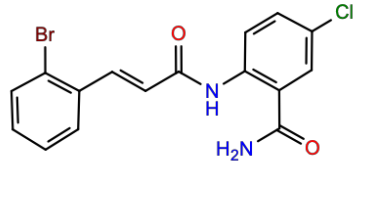
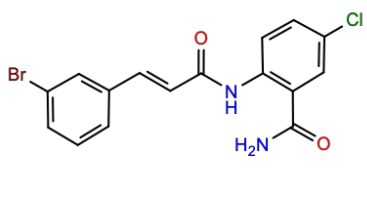
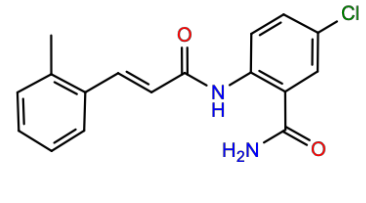
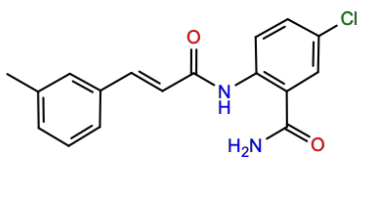
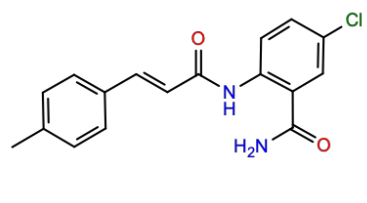
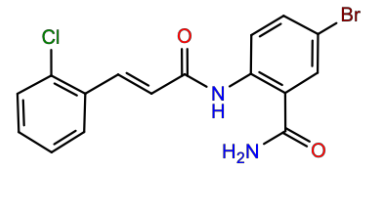
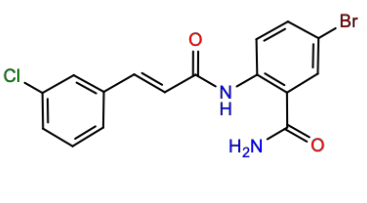
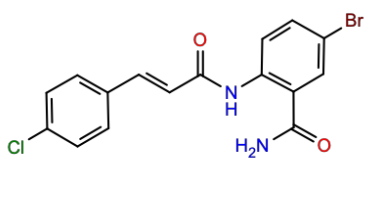
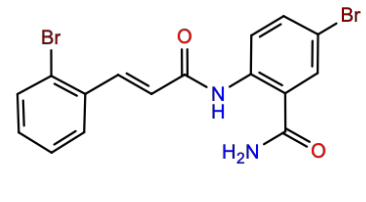
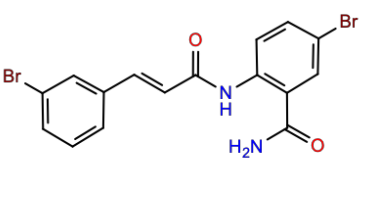
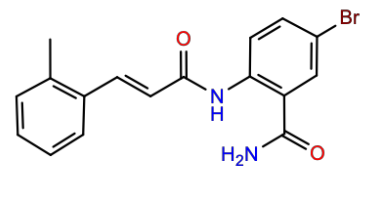
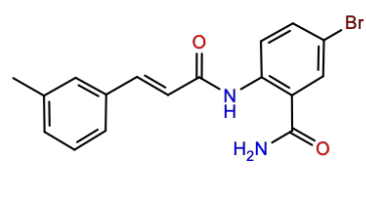
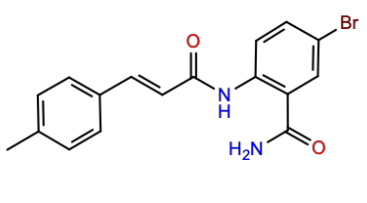
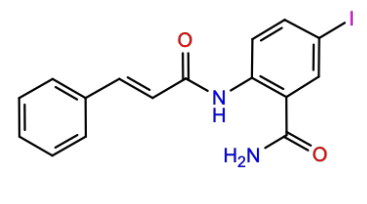
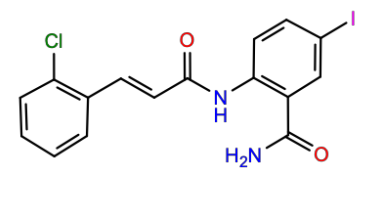
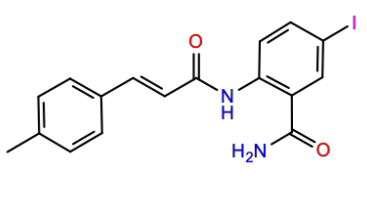
		
title 2003a_9b pIC50 K562 None pIC50 HL60 4.398	title 2003a_9c pIC50 K562 None pIC50 HL60 4.377	title 2003a_10a pIC50 K562 None pIC50 HL60 4.319
		
title 2003a_10b pIC50 K562 None pIC50 HL60 4.097	title 2003a_11a pIC50 K562 None pIC50 HL60 5.699	title 2003a_11b pIC50 K562 None pIC50 HL60 7.523
		
title 2003a_11c pIC50 K562 None pIC50 HL60 4.699	title 2003a_11d pIC50 K562 None pIC50 HL60 7.301	title 2003a_12a pIC50 K562 5.0 pIC50 HL60 4.456
		
title 2003a_12b pIC50 K562 None pIC50 HL60 6.155	title 2003a_12c pIC50 K562 None pIC50 HL60 4.602	title 2003a_12d pIC50 K562 None pIC50 HL60 6.097
		
title 2005a_9b pIC50 K562 None pIC50 HL60 5.495	title 2005a_10b pIC50 K562 None pIC50 HL60 5.523	title 2005a_13b pIC50 K562 None pIC50 HL60 5.824
		
title 2005a_14b pIC50 K562 None pIC50 HL60 5.523	title 2005a_18b pIC50 K562 6.0 pIC50 HL60 6.046	title 2005a_19b pIC50 K562 None pIC50 HL60 4.824

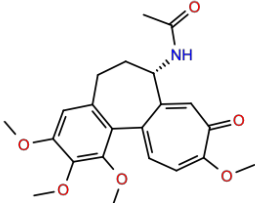
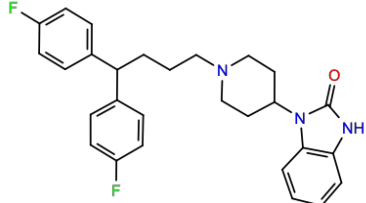
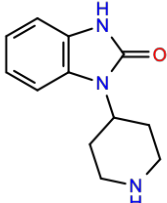
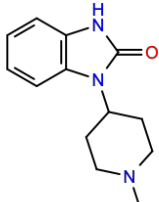
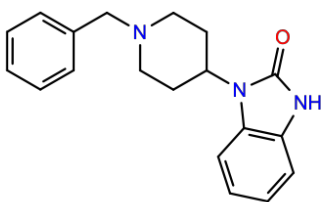
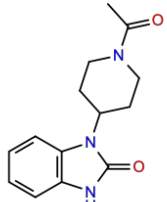
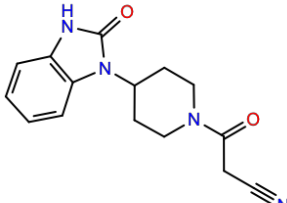
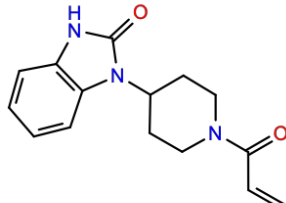
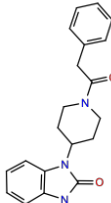
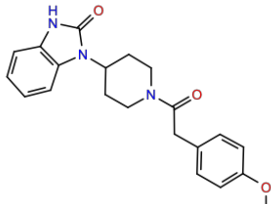
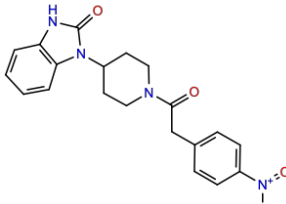
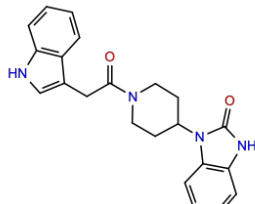
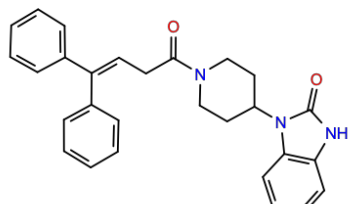
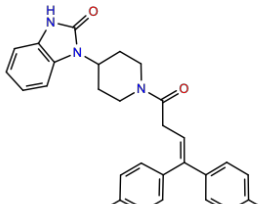
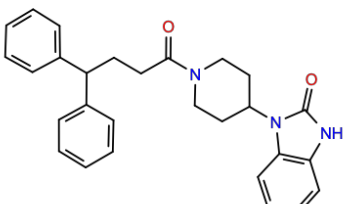
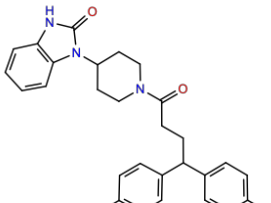
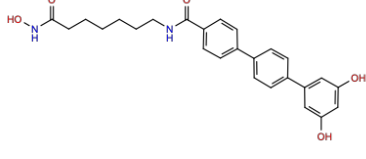
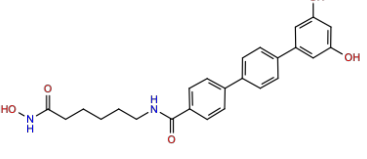
	<p>title 2005a_20a</p> <p>pIC50 K562 6.699</p> <p>pIC50 HL60 7.0</p>		<p>title 2005a_20b</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 5.699</p>		<p>title 2005a_21a</p> <p>pIC50 K562 6.699</p> <p>pIC50 HL60 6.602</p>
	<p>title 2005a_21b</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 5.699</p>		<p>title 2005a_27a</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 4.699</p>		<p>title 2005a_27b</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 4.921</p>
	<p>title 2005a_27c</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 4.523</p>		<p>title 2005a_31b</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 5.301</p>		<p>title 2005a_34b</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 6.301</p>
	<p>title 2005a_35b</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 6.523</p>		<p>title 2006a_5a</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 4.456</p>		<p>title 2006a_5b</p> <p>pIC50 K562 5.0</p> <p>pIC50 HL60 5.602</p>
	<p>title 2006a_5c</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 4.658</p>		<p>title 2006a_5d</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 4.432</p>		<p>title 2006a_5e</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 5.456</p>
	<p>title 2006a_5f</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 5.602</p>		<p>title 2006a_6a</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 5.0</p>		<p>title 2006a_6b</p> <p>pIC50 K562 None</p> <p>pIC50 HL60 6.824</p>

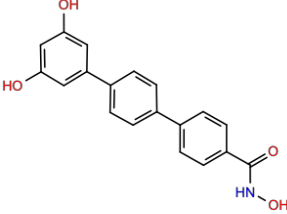
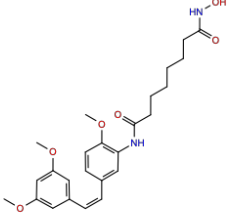
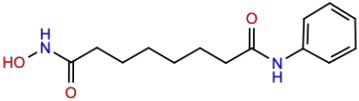
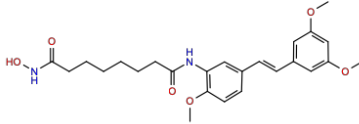
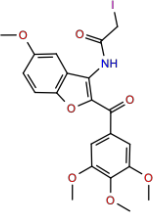
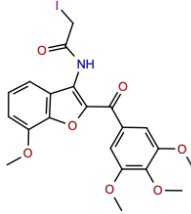
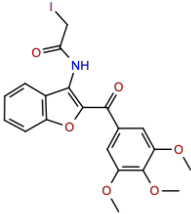
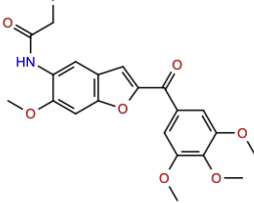
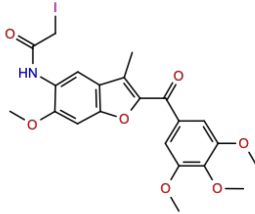
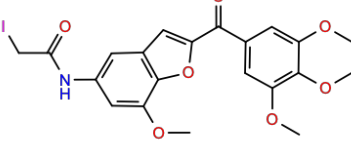
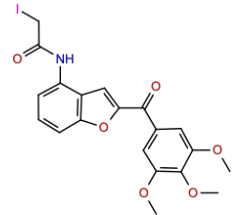
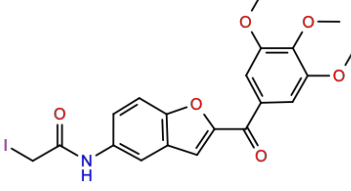
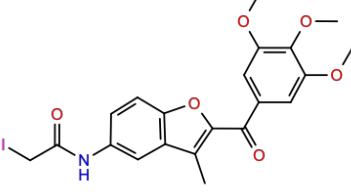
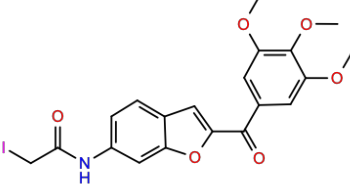
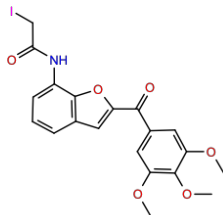
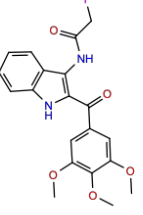
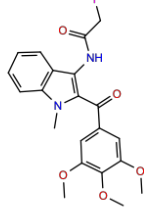
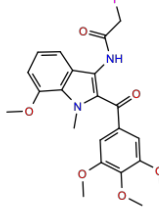
		
title 2006a_6c pIC50 K562 None pIC50 HL60 5.745	title 2006a_6d pIC50 K562 None pIC50 HL60 5.553	title 2006b_3a pIC50 K562 4.62 pIC50 HL60 5.046
		
title 2006b_3b pIC50 K562 4.585 pIC50 HL60 4.638	title 2006b_3d pIC50 K562 5.097 pIC50 HL60 5.456	title 2006b_3e pIC50 K562 4.745 pIC50 HL60 4.959
		
title 2006b_3f pIC50 K562 4.796 pIC50 HL60 4.854	title 2006b_3g pIC50 K562 4.638 pIC50 HL60 4.699	title 2006b_3h pIC50 K562 4.0 pIC50 HL60 4.076
		
title 2006b_7a pIC50 K562 4.0 pIC50 HL60 4.119	title 2006b_7c pIC50 K562 4.377 pIC50 HL60 5.0	title 2006b_7d pIC50 K562 4.699 pIC50 HL60 4.553
		
title 2006b_13a pIC50 K562 4.456 pIC50 HL60 5.602	title 2006b_13b pIC50 K562 4.0 pIC50 HL60 4.125	title 2006b_13c pIC50 K562 4.854 pIC50 HL60 5.319
		
title 2006b_13d pIC50 K562 5.0 pIC50 HL60 4.301	title 2006b_13e pIC50 K562 4.854 pIC50 HL60 5.301	title 2006b_13f pIC50 K562 5.222 pIC50 HL60 5.155

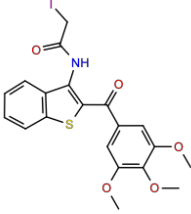
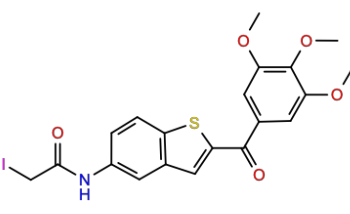
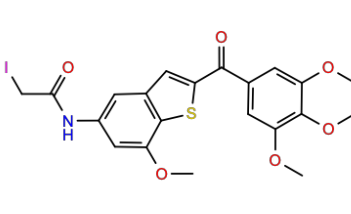
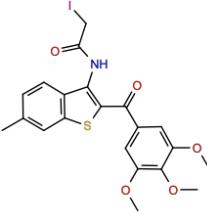
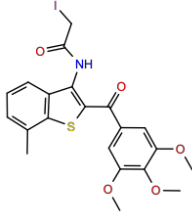
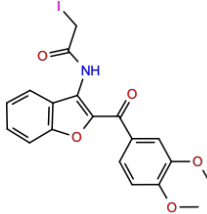
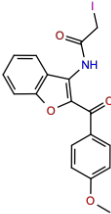
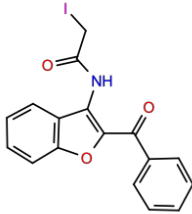
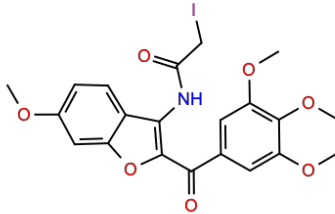
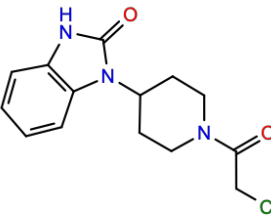
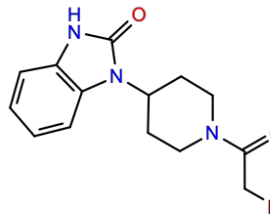
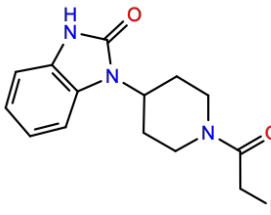
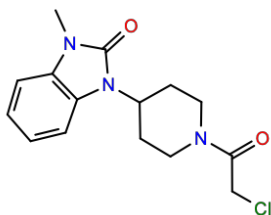
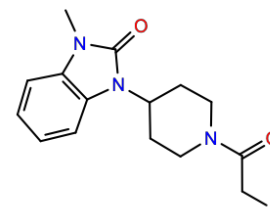
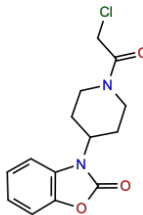
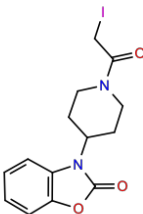
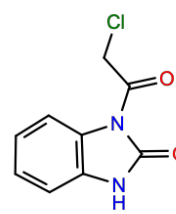
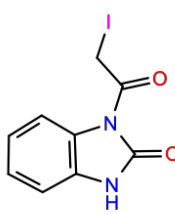
	<p>title 2006b_13g</p> <p>pIC50 K562 4.699</p> <p>pIC50 HL60 5.155</p>		<p>title 2006b_13h</p> <p>pIC50 K562 4.699</p> <p>pIC50 HL60 4.569</p>		<p>title 2006b_13i</p> <p>pIC50 K562 5.097</p> <p>pIC50 HL60 5.456</p>
	<p>title 2008a_17a</p> <p>pIC50 K562 4.097</p> <p>pIC50 HL60 4.097</p>		<p>title 2008a_18a</p> <p>pIC50 K562 6.0</p> <p>pIC50 HL60 5.0</p>		<p>title 2008a_19a</p> <p>pIC50 K562 4.658</p> <p>pIC50 HL60 4.699</p>
	<p>title 2008a_20a</p> <p>pIC50 K562 5.097</p> <p>pIC50 HL60 4.523</p>		<p>title 2008a_21a</p> <p>pIC50 K562 5.097</p> <p>pIC50 HL60 5.097</p>		<p>title 2008a_22a</p> <p>pIC50 K562 4.699</p> <p>pIC50 HL60 5.0</p>
	<p>title 2008a_23a</p> <p>pIC50 K562 4.097</p> <p>pIC50 HL60 4.097</p>		<p>title 2008a_24a</p> <p>pIC50 K562 4.921</p> <p>pIC50 HL60 4.347</p>		<p>title 2008a_25a</p> <p>pIC50 K562 4.745</p> <p>pIC50 HL60 5.0</p>
	<p>title 2008a_31a</p> <p>pIC50 K562 4.097</p> <p>pIC50 HL60 4.097</p>		<p>title 2008a_32</p> <p>pIC50 K562 4.097</p> <p>pIC50 HL60 4.097</p>		<p>title 2008a_33a</p> <p>pIC50 K562 4.097</p> <p>pIC50 HL60 4.097</p>
	<p>title 2008a_36</p> <p>pIC50 K562 4.658</p> <p>pIC50 HL60 5.0</p>		<p>title 2009a_1</p> <p>pIC50 K562 5.097</p> <p>pIC50 HL60 5.222</p>		<p>title 2009a_2</p> <p>pIC50 K562 4.0</p> <p>pIC50 HL60 4.0</p>

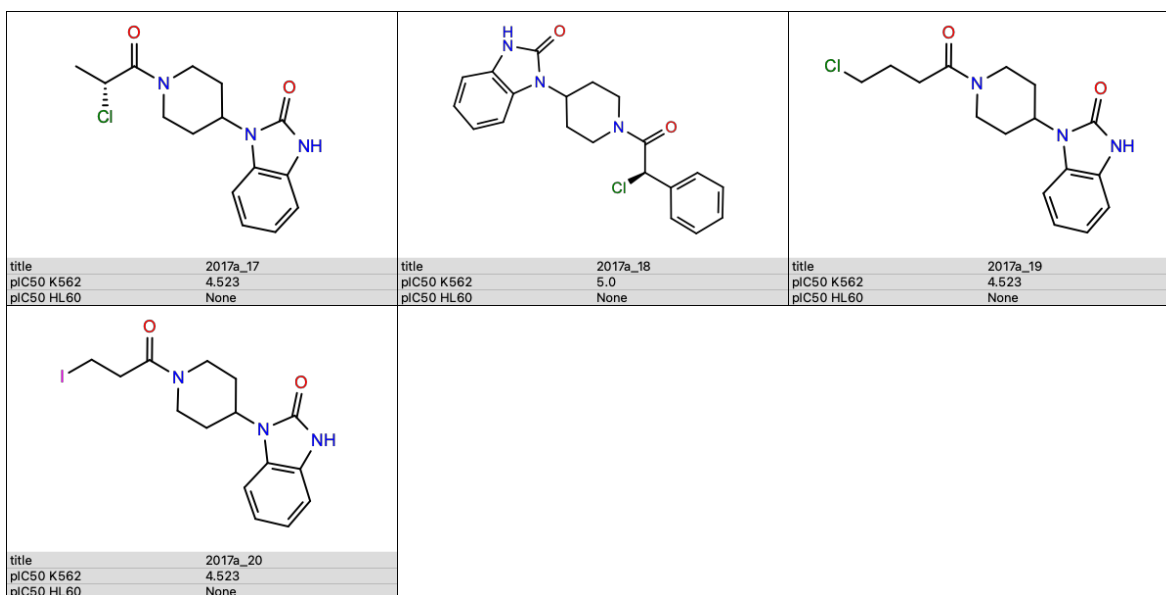
		
title 2009a_3 pIC50 K562 4.854 pIC50 HL60 5.222	title 2009a_4 pIC50 K562 5.046 pIC50 HL60 5.097	title 2009a_5 pIC50 K562 4.42 pIC50 HL60 4.699
		
title 2009a_6 pIC50 K562 5.097 pIC50 HL60 4.796	title 2009a_7 pIC50 K562 4.009 pIC50 HL60 4.699	title 2009a_8 pIC50 K562 4.602 pIC50 HL60 4.638
		
title 2009a_9 pIC50 K562 4.137 pIC50 HL60 4.553	title 2009a_10 pIC50 K562 4.022 pIC50 HL60 4.066	title 2009a_11 pIC50 K562 4.0 pIC50 HL60 4.0
		
title 2009a_12 pIC50 K562 4.0 pIC50 HL60 4.06	title 2009a_13 pIC50 K562 4.0 pIC50 HL60 4.208	title 2009a_15 pIC50 K562 4.097 pIC50 HL60 4.097
		
title 2009a_16 pIC50 K562 4.097 pIC50 HL60 4.097	title 2011a_4 pIC50 K562 5.0 pIC50 HL60 None	title 2011a_12a pIC50 K562 5.26 pIC50 HL60 None
		
title 2011a_12b pIC50 K562 5.602 pIC50 HL60 None	title 2011a_12c pIC50 K562 5.301 pIC50 HL60 None	title 2011a_12d pIC50 K562 5.131 pIC50 HL60 None

		
title 2011a_12e pIC50 K562 5.0 pIC50 HL60 None	title 2011a_12f pIC50 K562 5.0 pIC50 HL60 None	title 2011a_12g pIC50 K562 5.0 pIC50 HL60 None
		
title 2011a_12h pIC50 K562 5.0 pIC50 HL60 None	title 2011a_12i pIC50 K562 5.022 pIC50 HL60 None	title 2011a_12j pIC50 K562 5.0 pIC50 HL60 None
		
title 2011a_12k pIC50 K562 5.201 pIC50 HL60 None	title 2011a_12l pIC50 K562 5.092 pIC50 HL60 None	title 2011a_12m pIC50 K562 5.0 pIC50 HL60 None
		
title 2011a_12n pIC50 K562 5.0 pIC50 HL60 None	title 2011a_12o pIC50 K562 5.0 pIC50 HL60 None	title 2011a_12p pIC50 K562 5.0 pIC50 HL60 None
		
title 2011a_12q pIC50 K562 5.0 pIC50 HL60 None	title 2011a_12r pIC50 K562 5.0 pIC50 HL60 None	title 2011a_12s pIC50 K562 5.0 pIC50 HL60 None
		
title 2011a_17t pIC50 K562 6.244 pIC50 HL60 None	title 2011a_17u pIC50 K562 5.921 pIC50 HL60 None	title 2011a_17v pIC50 K562 5.0 pIC50 HL60 None

		
title 2011a_18 pIC50 K562 7.699 pIC50 HL60 None	title 2014a_1 pIC50 K562 5.301 pIC50 HL60 None	title 2014a_3 pIC50 K562 4.301 pIC50 HL60 None
		
title 2014a_4 pIC50 K562 4.301 pIC50 HL60 None	title 2014a_5 pIC50 K562 4.301 pIC50 HL60 None	title 2014a_6 pIC50 K562 4.301 pIC50 HL60 None
		
title 2014a_10 pIC50 K562 4.301 pIC50 HL60 None	title 2014a_11 pIC50 K562 4.301 pIC50 HL60 None	title 2014a_12 pIC50 K562 4.301 pIC50 HL60 None
		
title 2014a_13 pIC50 K562 4.301 pIC50 HL60 None	title 2014a_14 pIC50 K562 4.456 pIC50 HL60 None	title 2014a_15 pIC50 K562 4.301 pIC50 HL60 None
		
title 2014a_20 pIC50 K562 4.745 pIC50 HL60 None	title 2014a_21 pIC50 K562 4.886 pIC50 HL60 None	title 2014a_22 pIC50 K562 4.854 pIC50 HL60 None
		
title 2014a_23 pIC50 K562 4.301 pIC50 HL60 None	title 2015a_3 pIC50 K562 5.602 pIC50 HL60 None	title 2015a_4 pIC50 K562 5.222 pIC50 HL60 None

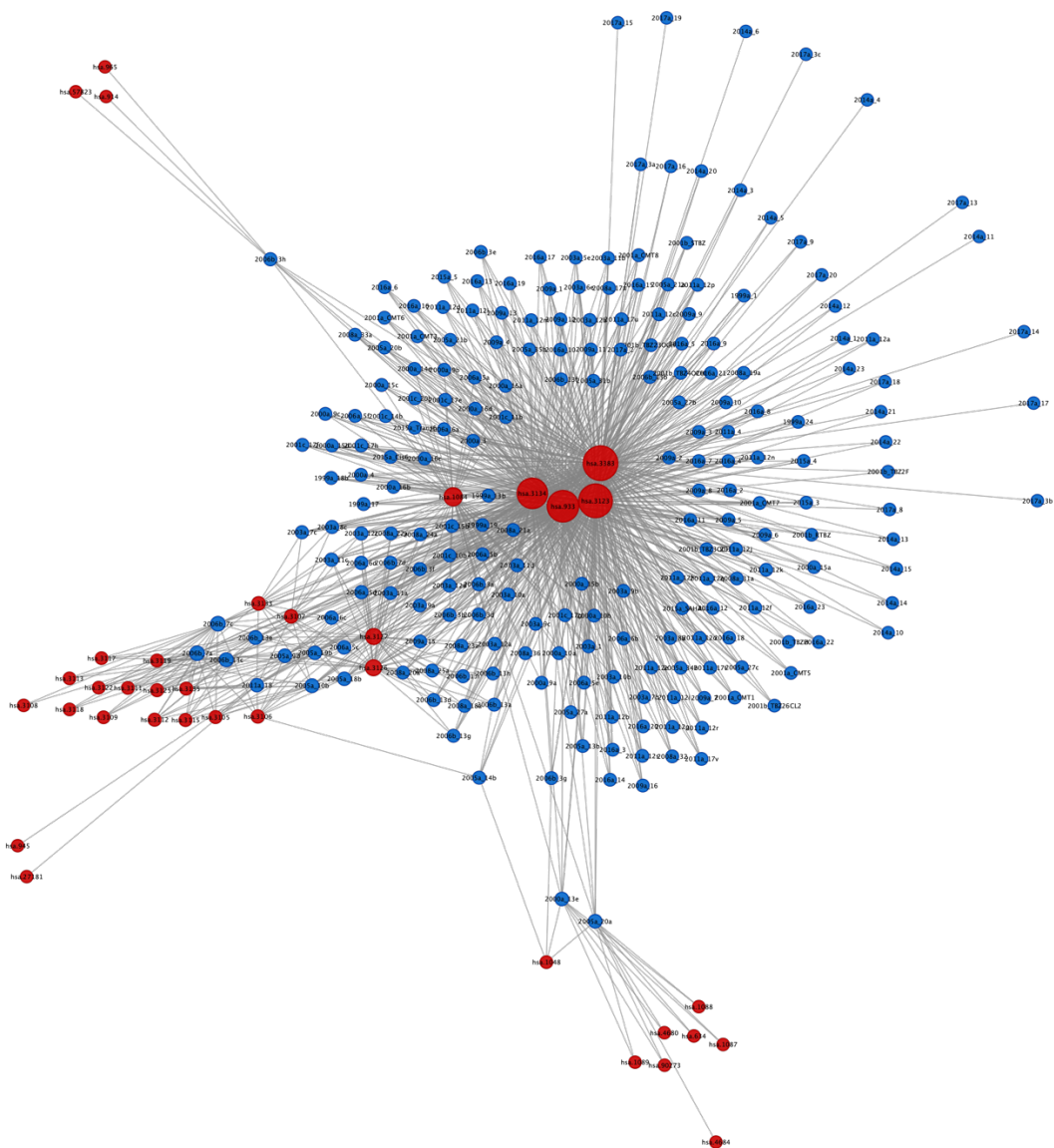
		
title 2015a_5 pIC50 K562 4.699 pIC50 HL60 None	title 2015a_Cis-6 pIC50 K562 5.125 pIC50 HL60 None	title 2015a_SAHA pIC50 K562 6.0 pIC50 HL60 None
		
title 2015a_Trans-6 pIC50 K562 6.155 pIC50 HL60 None	title 2016a_2 pIC50 K562 5.495 pIC50 HL60 None	title 2016a_3 pIC50 K562 4.854 pIC50 HL60 None
		
title 2016a_4 pIC50 K562 6.046 pIC50 HL60 None	title 2016a_5 pIC50 K562 6.398 pIC50 HL60 None	title 2016a_6 pIC50 K562 7.155 pIC50 HL60 None
		
title 2016a_7 pIC50 K562 6.398 pIC50 HL60 None	title 2016a_8 pIC50 K562 6.097 pIC50 HL60 None	title 2016a_9 pIC50 K562 6.222 pIC50 HL60 None
		
title 2016a_10 pIC50 K562 4.745 pIC50 HL60 None	title 2016a_11 pIC50 K562 6.222 pIC50 HL60 None	title 2016a_12 pIC50 K562 4.301 pIC50 HL60 None
		
title 2016a_13 pIC50 K562 6.301 pIC50 HL60 None	title 2016a_14 pIC50 K562 5.367 pIC50 HL60 None	title 2016a_15 pIC50 K562 6.097 pIC50 HL60 None

		
title 2016a_16 pIC50 K562 4.745 pIC50 HL60 None	title 2016a_17 pIC50 K562 6.097 pIC50 HL60 None	title 2016a_18 pIC50 K562 6.108 pIC50 HL60 None
		
title 2016a_19 pIC50 K562 6.0 pIC50 HL60 None	title 2016a_20 pIC50 K562 6.398 pIC50 HL60 None	title 2016a_21 pIC50 K562 4.658 pIC50 HL60 None
		
title 2016a_22 pIC50 K562 5.538 pIC50 HL60 None	title 2016a_23 pIC50 K562 6.0 pIC50 HL60 None	title 2017a_2 pIC50 K562 6.921 pIC50 HL60 None
		
title 2017a_3a pIC50 K562 5.137 pIC50 HL60 None	title 2017a_3b pIC50 K562 5.62 pIC50 HL60 None	title 2017a_3c pIC50 K562 5.745 pIC50 HL60 None
		
title 2017a_8 pIC50 K562 5.398 pIC50 HL60 None	title 2017a_9 pIC50 K562 5.796 pIC50 HL60 None	title 2017a_13 pIC50 K562 6.125 pIC50 HL60 None
		
title 2017a_14 pIC50 K562 5.301 pIC50 HL60 None	title 2017a_15 pIC50 K562 4.523 pIC50 HL60 None	title 2017a_16 pIC50 K562 5.602 pIC50 HL60 None

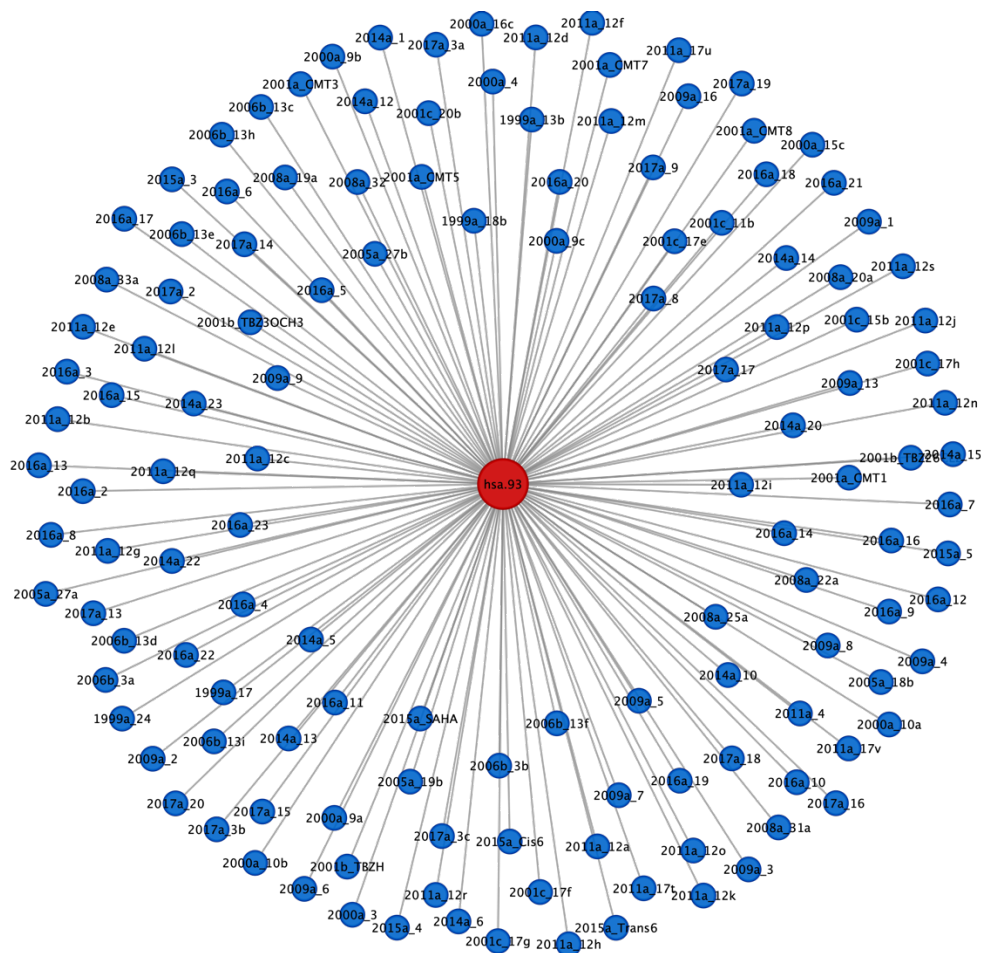


Supplementary Figure 1. The collection of 220 antiproliferative compounds. Their chemical structures are shown, and their antiproliferative activities are reported as pIC50 values.

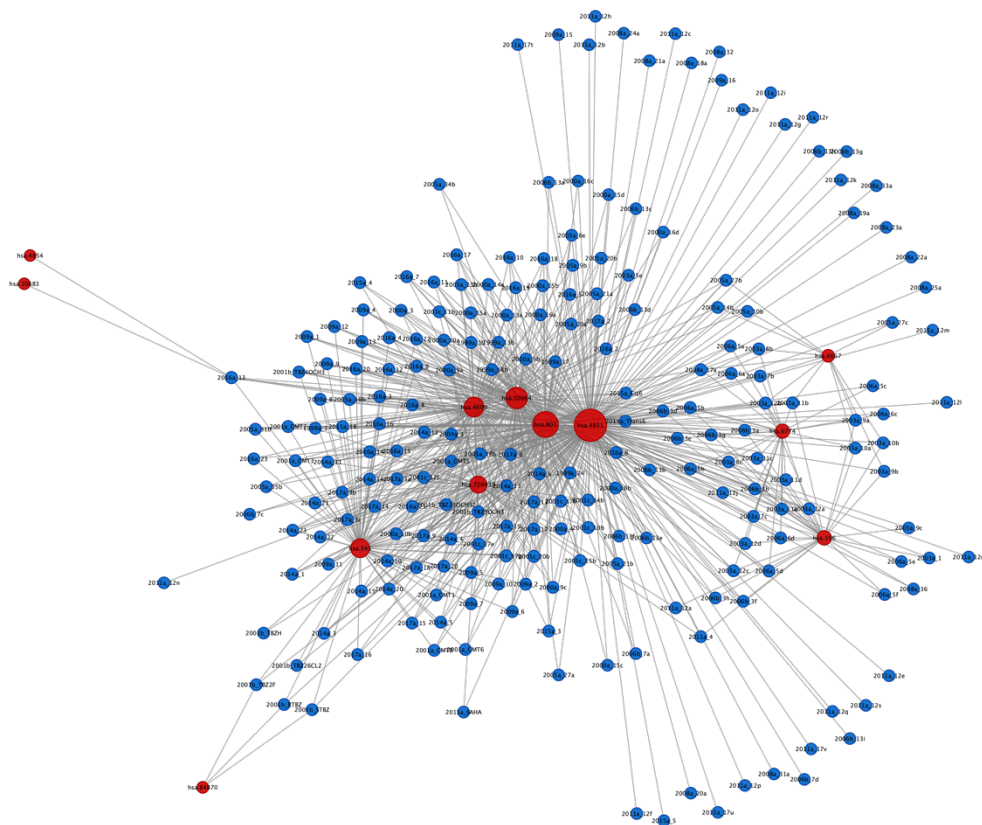
Reference articles according to molecules' titles: 1999a [13], 2000a [14], 2001a [21], 2001b [20], 2001c [15], 2003a [16], 2005a [22], 2006a [18], 2006b [17], 2008a [19], 2009a [23], 2011a [24], 2014a [27], 2015a [25], 2016a [26], 2017a [28].



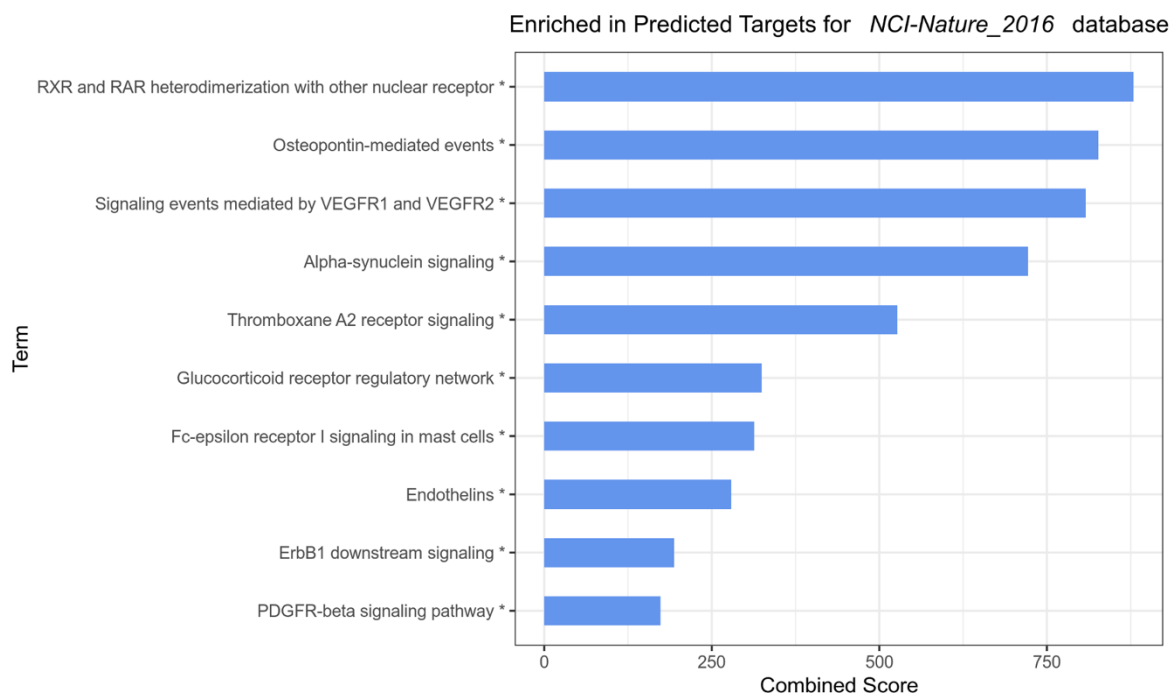
Supplementary Figure S2. Predicted compound-target interaction network for cell surface molecules and ligands class. Nodes representing the antiproliferative compounds are colored in blue, whereas predicted protein targets in red. Node size represents the node degree. Links represent the predicted compound-target associations. The image was rendered through Cytoscape 3.8.2 [7].



Supplementary Figure S3. Predicted compound-target interaction network for protein kinases class. Nodes representing the antiproliferative compounds are colored in blue, whereas predicted protein targets in red. Node size represents the node degree. Links represent the predicted compound-target associations. The image was rendered through Cytoscape 3.8.2 [7].



Supplementary Figure S4. Predicted compound-target interaction network for not elsewhere classified class. Nodes representing the antiproliferative compounds are colored in blue, whereas predicted protein targets in red. Node size represents the node degree. Links represent the predicted compound-target associations. The image was rendered through Cytoscape 3.8.2 [7].



Supplementary Figure S5. Functional enrichment bar plot for DrugBank target class. The ten pathways with highest combined score are shown for NCI-Nature pathway gene-set library. The analysis was performed by means of *enrichr* R package [78].