INTEGRATIVE NETWORK MODELLING OF DRUG RESPONSES IN CANCER FOR REVEALING MECHANISM OF ACTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ŞEYMA ÜNSAL BEYGE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF MEDICAL INFORMATICS

SEPTEMBER 2021

Approval of the thesis:

INTEGRATIVE NETWORK MODELLING OF DRUG RESPONSES IN CANCER FOR REVEALING MECHANISM OF ACTION

Submitted by ŞEYMA ÜNSAL BEYGE in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics** _____

Assoc. Prof. Dr. Nurcan Tunçbağ
Supervisor, **Health Informatics** _____

Examining Committee Members:

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics Dept., METU _____

Assoc. Prof. Dr. Nurcan Tunçbağ
Health Informatics Dept., METU _____

Assist. Prof. Dr. Aybar Can Acar
Health Informatics Dept., METU _____

Assoc. Prof. Dr. Ceren Sucularlı
Bioinformatics Dept., Hacettepe University _____

Assist. Prof. Dr. Tuğba Süzek
Computer Engineering Dept., Muğla Sıtkı Koçman Üni _____

Date: *06.09.2021*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Şeyma Ünsal Beyge

Signature : _____

ABSTRACT

INTEGRATIVE NETWORK MODELLING OF DRUG RESPONSES IN CANCER FOR REVEALING MECHANISM OF ACTION

ÜNSAL BEYGE, Şeyma
Ph.D., Department of Health Informatics
Supervisor: Assoc. Prof. Dr. Nurcan Tunçbağ

September 2021, 145 pages

Classification of cancer drugs is crucial for drug repurposing since the cost and innovation deficit make new drug development processes challenging. Heterogeneity of cancer causes drug classification purely based on known mechanism of action (MoA) and the list of target proteins to be insufficient. Multi-omic data integration is necessary for a systems biology perspective to understand molecular mechanisms and interactions between cellular entities underlying the disease. This study integrates drug-target interaction data with transcriptomics and phosphoproteomic data of perturbed cell lines to model drug and cell-specific subnetworks. Total 250 networks are reconstructed, including 70 small molecule drugs on six cell lines. Similarities of reconstructed networks are quantitatively calculated using a topology-based network comparison measure which scores the separation of networks using the shortest paths between network nodes. Different drugs with similar omic outcomes on variable cell lines are revealed with the aid of separation scores. Moreover, the effect of drugs on variable cell lines is discovered together with the impact of target selectivity of drugs within the same MoA group. Functional analysis of reconstructed networks for their enriched cellular pathways further indicated that drugs with different chemical structures and MoA might induce common signaling cascades. As omics data integration coupled network modeling reveals modulated pathways for specific conditions, the methodology of this study is applicable to different drug-disease research areas. Prediction of drug combinations for a given disease and inference of drug similarity based on cell line sensitivity are two applications presented in this study.

Keywords: Network reconstruction, protein-protein interactions, omics, drug repurposing, data integration

ÖZ

KANSERDE İLAÇ ETKİLERİNİN VE BENZERLİKLERİNİN BULUNMASI AMAÇLI ÇOKLU OMİK VERİ ENTEGRASYONU İLE BİYOLOJİK AĞ MODELLEME

Ünsal Beyge, Şeyma
Doktora, Sağlık Bilişimi Bölümü
Tez Yöneticisi: Doç. Dr. Nurcan Tunçbağ

Eylül 2021, 145 sayfa

Yeni ilaç geliştirme uygulamalarının pahalı ve zorlu prosedürler olması sebebiyle kanser ilaçlarının sınıflandırılması ilaç yeniden hedefleme stratejileri için büyük önem taşır. Kanser heterojenliği de sadece bilinen hedef mekanizmaları ya da hedef protein listelerinin kullanılmasının ilaçların sınıflandırılmasında yetersiz kalmasına sebep olmaktadır. Hastalığa sebep olan moleküler mekanizmaların ve hücresel etkileşimlerin anlaşılabilmesi için sistem biyolojisi perspektifinden çoklu-omik veri entegrasyonu gereklidir. Bu çalışmada, ilaç hedef etkileşim verisi ilaç uygulanmış hücre hatlarının transkriptomik ve fosfoproteomik verisi ile entegre edilerek ilaç ve hücre spesifik protein etkileşim ağları modellenmiştir. 70 küçük molekül ilaç ve altı hücre hattını kapsayan toplam 250 protein etkileşim ağı oluşturulmuştur. Oluşturulan ağların benzerliği nicel olarak ağ nodlarının arasındaki en kısa yolları ölçen topoloji temelli ağ karşılaştırma metodu ile hesaplanmıştır. Ayrışma skorları yardımıyla değişik hücre hatlarında benzer omik sonuçlara sahip ilaçlar ortaya çıkarılmıştır. Ayrıca, ilaçların farklı hücrelerdeki etkileri ve aynı mekanizma grubunda olan ilaçların hedef protein seçiciliğinin etkileri ortaya konmuştur. Protein ağlarında zenginleşen sinyal yolaklarının analizi farklı kimyasal yapıya ve mekanizmaya sahip ilaçların benzer yolakları etkileyebileceğini göstermiştir. Omik veri entegre edilmiş biyolojik ağ modelleme özel durumlarda etkilenen sinyal yolaklarının bulunmasını sağladığından, bu metodoloji farklı ilaç-hastalık ilişkisi araştırma alanlarında kullanılabilir. Belirli bir hastalık için kombine edilebilecek ilaç çiftlerinin tahmin edilmesi ve hücre hatlarının ilaca duyarlılık verisi kullanılarak ilaç benzerliğinin tahmin edilmesi bu çalışmada gösterilen iki uygulamadır.

Anahtar Sözcükler: Ağ modelleme, protein protein etkileşimleri, omik data, ilaç yeniden hedefleme, biyolojik veri entegrasyonu

To my precious daughter, Helen…

ACKNOWLEDGEMENTS

I have several people to thank and express my gratitude for their support during my time in graduate school. Without their precious help and considerations, it would not be possible to conduct this research.

First of all, I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Dr. Nurcan Tunçbağ for being an excellent mentor for me throughout my studies. She convincingly guided and encouraged me to be a professional and not give up when I hardly found the strength and time to continue my academic studies. I am very grateful for her availability as a supervisor, regardless of time and location.

Besides my supervisor, I am grateful to all my thesis progress committee members Assoc. Prof. Dr. Aybar Can Acar and Assoc. Prof. Dr. Ceren Sucularlı for their valuable discussions and insightful comments throughout my studies. Besides my thesis progress committee members, I also thank examining committee members Assoc. Prof. Dr. Yeşim Aydın Son and Assist. Prof. Dr. Tuğba Süzek for their participation and valuable comments.

The study presented in this thesis is partly supported by the Career Development Program of the Scientific and Technological Research Council of Turkey (TÜBİTAK) under the project number 117E192 and TUBA-GEBIP. I greatly acknowledge their valuable supports which helped me to accomplish these studies.

I express my deepest appreciation to my husband, Mahmut Beyge for his support both academically and psychologically. Also, I would like to thank my parents, Ayşe and Salih Ünsal, and my brother, Murat Ünsal for their endless support in every aspect of life. This thesis would not happen without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Cmap	Connectivity Map
MoA	Mechanism of Action
A375	cell line abbreviation for human malignant melanoma
A549	cell line abbreviation for human Caucasian lung carcinoma
MCF7	cell line abbreviation for human breast carcinoma
PC3	cell line abbreviation for human prostate adenocarcinoma
YAPC	cell line abbreviation for human pancreatic carcinoma
NPC	cell line abbreviation for neural progenitor cells
TCGA	The Cancer Genome Atlas
CNV	copy number variation
miRNA	microRNA
ICGC	International Cancer Genome Consortium
CCLE	Cancer Cell Line Encyclopedia
NOI	Network of Interest
PCSF	Prize-Collecting Steiner Forest
GRN	Gene Regulatory Networks
JASPAR	open-access database of curated, non-redundant transcription factor (TF) binding profiles
TRANSFAC	database of eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles
BioGRID	The Biological General Repository for Interaction Datasets
RegulonDB	database on transcriptional regulation in Escherichia coli K-12
PPIN	Protein-Protein Interaction Network
DIP	Database of Interacting Proteins
BIND	the Biomolecular Interaction Network Database
HPRD	Human Protein Reference Database
HPA	The Human Protein Atlas
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
KEGG	Kyoto Encyclopedia of Genes and Genomes
Reactome	open-source, open access, manually curated and peer-reviewed pathway database
ENZYME	Enzyme nomenclature database

BRENDA	The Comprehensive Enzyme Information System
TRANSPATH	database of mammalian signal transduction and metabolic pathways
GCN	Gene Co-expression Network
COXPRESdb	a database of comparative gene coexpression networks of eleven species for mammals
GeneFriends	a database for gene and transcript co-expression networks based on RNA-seq data
RIN	Residue Interaction Network
RING	The Residue Interaction Network Generator
RIP-MD	Residue Interactions in Protein - Molecular Dynamics
ProSNEx	Protein Structure Network Explorer
RINspector	a Cytoscape app for centrality analyses and DynaMine flexibility prediction
Ctyoscape	A Community-Based Framework for Network Modeling
REVEAL	REVerse Engineering ALgorithm
ChemChains	A platform for simulation and analysis of biochemical networks
Pint	A Static Analyzer for Transient Dynamics of Qualitative Networks
SimBoolNet	a Cytoscape plugin for dynamic simulation of signaling networks
BoolNet	an R package for generation, reconstruction and analysis of Boolean networks
CellNetAnalyzer	A Software Package for Analyzing Structure and Function of Cellular Networks
DAG	Directed Acyclic Graph
BNFinder	Bayesian Network Finder
bnlearn	an R package for learning the graphical structure of Bayesian networks
ODE	Ordinary Differential Equation
GA	Genetic Algorithm
PSO	Particle Swarm Optimization
MWSLE	Minimum Weight Solution to Linear Equation
FTSS	Fourier Transform for Stable Systems
EM	Evolutionary Modeling
ANN	Artificial Neural Network
RNN	Recurrrent Neural Network
ELM	Extreme Learning Machine
CNN	Convolutional Neural Network
FCNN	Fully Connected Neural Network
PyTorch	an open source machine learning library based on the Torch library

TensorFlow	an end-to-end open source platform for machine learning
RELEVANCE	A tool for network reconstruction based on information theory
ARACNE	Algorithm for the Reverse engineering of Accurate Cellular Network
CLR	Context Likelihood of Relatedness
MRNET	Maximum Relevance/Minimum Redundancy based method for inferring genetic networks
FyNE	Fuzzy Networks
PLS	Partial Least Square
CCA	Canonical Correlation Analysis
CIA	Co-Inertia Analysis
MCIA	Multiple Co-Inertia Analysis
MFA	Multiple Factor Analysis
PCA	Principle Component Analysis
SNF	Similarity Network Fusion
ATHENA	a tool for meta-dimensional analysis
NMF	Non-negative Matrix Factorization
JIVE	joint and individual variation explained
ssCCA	structure-constrained sparse canonical correlation analysis
sMBPLS	sparse Multi-Block Partial Least Squares
SNPLS	sparse network-regularized partial least square
MDI	Multiple Dataset Integration
Prob_GBM	Probabilistic Glioblastoma Multiforme
PSDF	Patient Specific Data Fusion
BCC	Bayesian Consensus Clustering
CONEXIC	COpy Number and EXpression In Cancer
rMKL-LPP	regularized unsupervised multiple kernel learning-Locality Preserving Projections
CNAmet	an R package for integrating copy number, methylation and expression data
iPAC	in-trans process associated and cis-correlated
SDP	Semi-Definite Programming
SVM	Support Vector Machine
FS	Feature Selection
MKL	Multiple Kernel Learning
FSMKL	MKL learning with feature selection
iBAG	integrative Bayesian analysis of genomics data
MCD	Multiple Concerted Disruption
MKGI	Metadimensional Knowledge-driven Genomic Interactions

EXP	Expression
MET	Methylation
SNP	Single Nucleotide Polymorphism
GENN	Grammatical Evolution Network Network
aCGH	Array Comparative Genomic Hybridization
LoH	Loss of Heteerozygosity
MP-ERGMs	penalized multimode Exponential Random Graph Model
LO	Linear Optimization
MINT	The Molecular INTeraction Database
GO	Gene Ontology
DAVID	Database for Annotation, Visualization and Integrated Discovery
GSEA	Gene Set Enrichment Analysis
CAR-index	Cannistrai-Alanis-Ravai index
CPA	CAR-index variant of Preferential Attachment
CRA	CAR-index variant of Resource Allocation
GCP	Global Chromatin Profiling
OLS	Ordinary Least Square
BING	Best Inferred Genes
DMSO	Dimethyl sulfoxide
SMILES	simplified molecular-input line-entry system
MACCS	Molecular ACCess System
SMARTS	SMILES arbitrary target specification
PCST	Prize-Collecting Steiner Tree
PCSF	Prize-Collecting Steiner Forest
GBM	Glioblastoma Multiforme
LP	Link Prediction
AA	Adamic/Adar
JC	Jaccard's Coefficient
PA	Preferential Attachment
RA	Resource Allocation
HCC	Hepatocellular carcinoma
NF1	Neurofibromatosis type I
GDSC	Genomics of Drug Sensitivity in Cancer
FDA	Food and Drug Administration
NMPA	National Medical Products Administration

CHAPTER 1

INTRODUCTION

1.1. Background

Characterization and quantification of a set of bio-molecules are classified according to the target biomolecule class. If the target is a genome, it is called genomics; and if proteins are targeted, it is called proteomics. The entirety of these disciplines that try to measure all the members of a type of bio-molecule (DNA, RNA, Protein, Metabolites, etc.) is collectively called "omics" [1,2]. Omics data integration can be classified into two main categories, namely horizontal meta-analysis, and vertical integrative analysis[3]. In the horizontal meta-analysis, different samples are aligned horizontally based on features (genes, proteins, etc.) to obtain their similarities and differences. These differences and similarities can be used to obtain biomarkers and for pathway detection. Multiple samples are taken from the same specimen in vertical integrative analysis to obtain multi-layers of omics data. Data integration/integrative analysis is performed to obtain intra-specimen inter-omics relations and mechanisms. Integration of different "omics" classes to understand interrelations and systems-level interpretation of a point of interest (function, subtyping, treatment, etc.) is called Multi-Omics[4,5]. The first review on integrating omics datasets[5] was published more than a decade ago and, more reviews on human[6] and plant[7] were also published recently.

Multi-Omics is a holistic approach and can be used in personalized and preventive medicine[8,9] and in the identification of new biomarkers for diseases. Biomarkers can increase the sensitivity of diagnosis and treatment by subtyping[10–15]. Multi-Omics can also be used for spatial or temporal mathematical modeling purposes[4] to gain understanding of mechanisms of diseases or pathogenesis and to identify potential candidates or targets for new therapeutics[16,17]. In the long term, the mathematical model obtained can answer questions that cannot be investigated experimentally. The need for integrated data analysis originates from the complexity of biological systems. Single-level data (genome, transcriptome, metabolome, etc.) could not fulfill necessary and sufficient conditions to explain a biological outcome such as phenotype, disease risk, response to treatment, etc. By identification and validation of essential factors and their interactions, data integration may improve the modeling of complex traits[18]. There are mainly two approaches to data integration; multi-stage analysis in which data integration is performed with hierarchical

stages[19], and meta-dimensional analysis in which multivariate model while allowing non-linear interactions from different biological levels is performed[20,21]. Analysis of multiple omics layers in parallel is recently defined as 'trans-omics' and 'integromics'[22–24].

Cancer is the generic name for a family of diseases in which rapid growth and the possible spread of abnormal cells are observed to threaten millions of lives[25]. Heterogeneity of cancer makes it difficult for effective diagnose and treatment[26,27]. Cancer caused nearly 10 million deaths in 2020 with almost 19.3 million new cases. Of cancer-caused deaths, 3.8% are due to prostate, 4.7% are due to pancreas, and 6.9% due to breast cancer[28], while the global cancer burden is expected to be 28.4 million cases in 2040[29]. Since cancer is a complex disease, a multi-stage transformation of a normal cell into malignant cells, a network of interacting proteins and genes are involved in disease progression[30,31]. Because of the high mortality and heterogeneity of cancer, personalized medicine is gaining attention for targeted treatment and management of each patient[32]. In cancer, one-dimensional omics data has given limited information regarding diagnosis, progression, treatment, and prevention. The reason for using one-dimensional data by part was the low availability of multi-dimensional data from the same bio-specimen. However, tumor specimens from over 10,000 cancer patients are analyzed with The Cancer Genome Atlas (TCGA) which include tissue exome sequencing, copy number variation (CNV), DNA methylation, gene expression, and microRNA (miRNA) expression[33]. Other than TCGA, tons of cancer multi-omics data are generated by International Cancer Genome Consortium (ICGC)[34] and Cancer Cell Line Encyclopedia (CCLE)[35].

As the size of data related to cancer increased, several challenges for data integration and analysis of multi-omics data have emerged. The different omics dimensions produce different data types; sequencing gives count data, microarrays give continuous data, and genetic variations give discrete data, all of which must be handled simultaneously. Data dimension is much higher than sample size; therefore, effective and robust computational algorithms are also needed. Moreover, data normalization and quality control preprocessing steps have to be performed before data integration [36,37].

Revealing potential disease networks and potential drugs for the Networks of Interest (NOI) becomes the primary step for designing treatment strategies for cancer. Identification of NOI involved in the type/subtype of a cancer is still a challenging task because of the spatial and temporal coverage and sampling number of modern high-throughput techniques. Especially to identify the mechanism of action of the drugs and increase the dimensionality of the high-throughput data, several small molecules have been used to treat cancer cells, and the high-throughput transcriptomic and proteomic data have been released in many databases. Treatment with perturbagens not only helps identification of potential drug targets or drug families for a given type/subtype of cancer but also gives a spatial resolution for identification of potential NOI for a given type/subtype of cancer. An adequate network may help minimize the number of drug targets and find new drug targets while studying a specific disease condition.

1.2. Motivation

Recovery of an optimal cancer treatment strategy is challenging due to the inter- and intra-heterogeneity of tumor samples. Modulations in signaling pathways and interactions between various bio-entities are critically important in the multi-stage tumor cell formation from the normal cells. Hence, multi-omic data integration is vital for understanding the molecular interactions in the cancer cell and for the development of an optimum treatment strategy. Since the development and approval of new drugs is both expensive and time-consuming, drug repurposing is a promising strategy in cancer treatment. Classification of currently available therapeutic agents is important and a challenging step necessary to index possible drugs for drug repurposing approaches. Moreover, the determination of molecular mechanisms of available drugs in different cancer types deciphers the possibilities in drug repurposing, given the heterogeneity of cancer. Conventionally, drugs are classified based on their primary targets, therapeutic actions, target specificity, nature of the interaction, molecular type, and chemical structure similarity. The effects of small drug molecules are highly dependent on cellular and physiological factors. Many drugs have multiple targets. Depending on the features selected, similar drug groups may change. Even though two drugs are present in the same group, they may modulate different signaling mechanisms within the cell.

Considering all these challenges in cancer drug repurposing, the below questions are tried to be answered in this study:

- Do chemically and functionally different drugs have similar protein-protein interaction networks or vice versa?
- Do drugs modulate different signaling pathways depending on the tumor type?
- Is it possible to emphasize signature signaling mechanisms caused by drug treatment by using a network-based approach?
- Is it possible to apply network-based analysis of omic data to different areas of drug-disease interactions?

1.3. Contributions of the Study

In this study, the transcriptomic and phosphoproteomic data of cell lines treated with small molecule drugs[38] are used to reconstruct networks by integrating the data with drug targetome and human interactome. Data integration is a challenging task since the proper integration method changes depending on the nature of the data and the network reconstruction principles. In this study, transcriptomic data is used to back trail the regulatory elements that act on the experimental hits. Phosphoproteomic hits allowed the selection of functionally active proteins that may be closely related to the drug modulatory effects. Since the drug actions start after their attack on the target proteins, the drug targetome includes the first affected cellular molecules. Also, the human interactome is referenced by the network reconstruction software, Omics Integrator[39], to map the seed proteins and find the optimum connected subnetwork. The human interactome is known for its incompleteness, and also it has bias to well-studied proteins[40,41]. Therefore, it has

many false negatives and false positives. In this study, human interactome is processed to eliminate these drawbacks. First, high-degree nodes are eliminated. For each cell line–drug condition, interactome is filtered for very low expressed genes by the aid of associated transcriptomic data, making every interactome tissue specific. After, each interactome is enriched with a link prediction approach followed by localization filters. The final processed interactome used for each cell line – drug condition has turned into both tissue and drug-specific.

Omics Integrator uses the Prize-Collecting Steiner Forest (PCSF) algorithm with reverse engineering principles. The significant benefit of using this approach is that it builds a connected subnetwork by adding extra nodes where necessary. In this way, it can find hidden elements which have roles in the cellular processes but couldn't be found via experimental methods.

Total 250 cell line-drug specific networks are reconstructed, covering 70 drugs and six cell lines. A stringent topological and pathway analysis of these reconstructed networks provided insight into the drug modulations occurring in different cell lines, including mostly cancer models. The highlights addressed by these findings are as following:

- Chemically and functionally different drugs may modulate overlapping networks.
- The target selectivity of the drug is an important factor leading to separate networks for drugs being functionally similar.
- Network-based analysis coupled with multi-omic data integration helped to reveal cell line and drug-specific hidden modulated pathways.
- Topological distance and active pathways of drug networks may guide the use of efficient drug combinations.
- The separation between networks of a drug across cell lines can help infer their resistance or sensitivity or no response to that drug.

1.4. Organization of the Dissertation

The dissertation consists of six main chapters, namely Introduction, Literature Review, Materials and Methods, Results, Discussion, and Conclusion. In Chapter 1, a brief description of the main concepts included in this study is presented. Also, the motivation and contributions of the study are included.

Chapter 2 presents the theoretical background of the concepts addressed in this study are presented. Definition of biological networks is provided, followed by the frequently used biological network types. After, the main network reconstruction and data integration methods are described. Finally, a brief information for link prediction methods is given and examples of the application of link prediction to biological networks are summarized.

Chapter 3 presents the workflow of this study. Details of the specific methods used in each step are explained.

In Chapter 4, results produced through the drug classification and network reconstruction procedures are presented. Comparison of reconstructed networks on topology and pathway level revealed that drugs have variable effects independent of their chemical structure within the cell. Moreover, selectivity on target predefines the similarity ratio of drug action to its MoA counterparts. Two different applications of network-based perturbation analysis are also illustrated in this chapter.

Chapter 5 presents the discussions on this thesis, and Chapter 6 presents conclusive remarks, the main findings of the study, limitations, and future prospects.

CHAPTER 2

LITERATURE REVIEW

This chapter briefly discusses the background and literature related to this study. The literature review is addressed in five main sections: (1) brief introduction of biological networks; (2) types of biological networks; (3) network reconstruction methods; (4) introduction to data integration and (5) link prediction.

2.1. Biological Networks

Biological systems function at several levels, i.e., genome, transcriptome, proteome, interactome, metabolome, and fluxome levels, and all of these states are in relation with each other. Study of all biological entities as a whole to understand the dynamics of living systems is defined as 'systems biology'[42]. Systems biology enables to have an understanding of how each bio-class is functioning and how these classes affect the whole system. Networks of biological entities constitute a critical aspect for a better understanding of several stages in a living organism. Networks (G(V, E)) represent relationships between entities in which nodes/vertices (V) refer to each entity of interest (genes, RNA molecules, proteins, metabolites, etc.) and edges/links (E) refer to the interactions between these. The dynamic nature of biological networks reflects the molecular traffic within the biological system[8]. Biological networks can be directed (edges include directional information) or undirected (edges do not include any directional information), weighted or unweighted (In weighted networks, edges include a quantitative value such as strength, cost, flow, or experimental reliability.). Given a biological network, the relationships of a set of biological entities leading to a phenotype can be characterized by mapping them on the network and analyzing the network properties. The global network properties include degree distribution, shortest path distances, network centralities, clustering coefficient, etc. The degree of a node is the number of other nodes it is linked to. The shortest path between two nodes is the minimal number of edges that must be traversed to travel from one node to the other. There are several network centrality measures such as degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, etc., all of which quantify each node's importance using different

aspects. The clustering coefficient can be measured locally or globally for a network where the former for a node indicates the proportion of edges between the nodes within its neighborhood divided by the number of edges that could exist between them. The latter is based on triplets of nodes[43,44].

2.2. Types of Biological Networks

There are several types of biological networks based on the type of data built upon and the behavior of interactions between biological entities. Below, the most popular biological network types are listed.

2.2.1. Gene regulatory networks (GRN)

Cells have several processes starting from transcription of a gene to a functional product. Proteins as functional units are regulated by transcription factors, cofactors, mRNAs, miRNAs, post-translational modifications, and the interplay between these regulators. Gene regulatory networks (GRN) are biological networks representing all of these interactions between regulators and genes/proteins they regulate[45,46]. GRNs can be reconstructed for specific tissues [47] and conditions such as drug treatment [48], mutations [49], gene knockouts [50], disease conditions [51], time-series gene expression levels [52], etc. [53]. Since they include all types of relationships between biological entities, GRNs can be reconstructed from all kinds of omics data such as transcriptomic, proteomic, phosphoproteomic, metabolomic data. Therefore, GRNs can be thought of as span all other types of biological networks.

Databases like JASPAR [54], TRANSFAC [55], BioGRID [56], and RegulonDB [57] provide information for gene regulatory networks. An example of GRN of *EGFR* produced with BioGRID is shown in Figure 1.

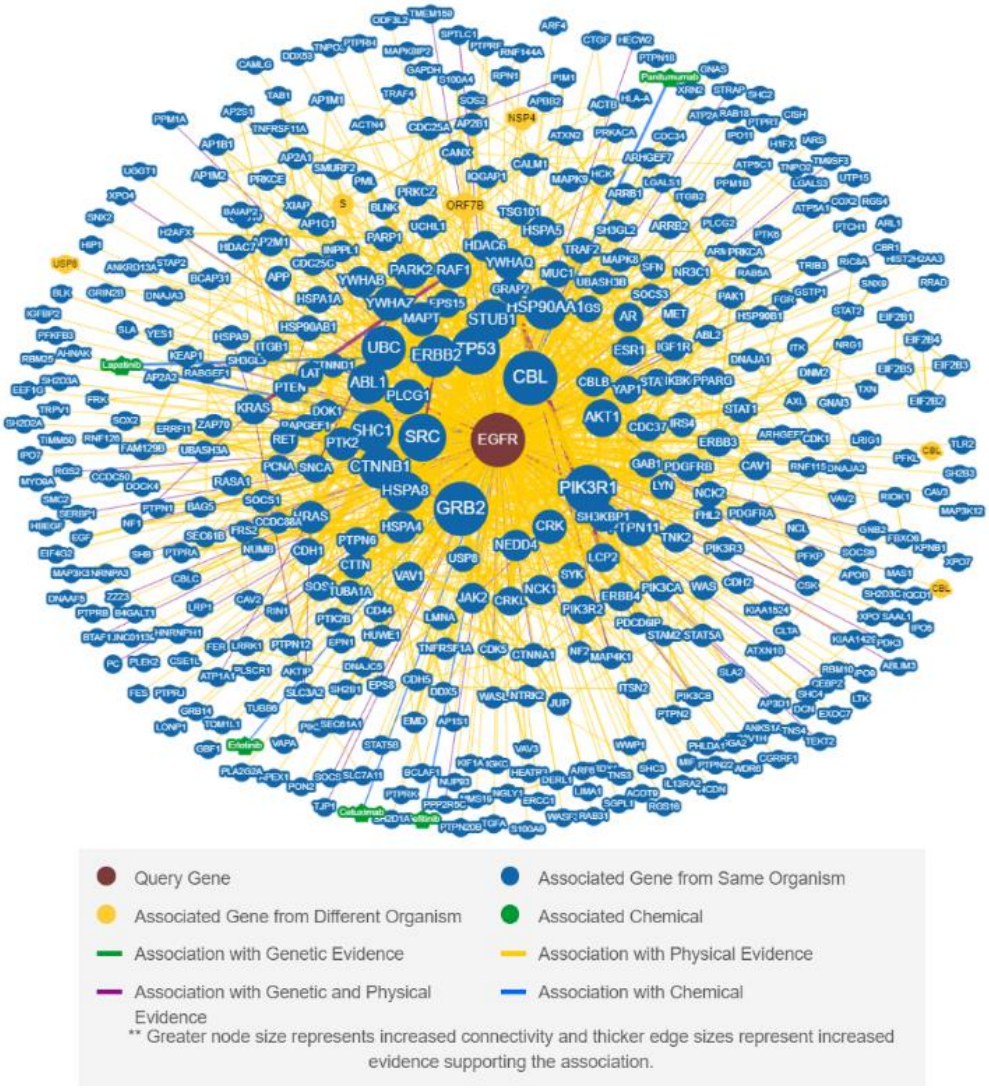


Figure 1. Example GRN for EGFR gathered from BioGRID.

2.2.2. Protein – Protein interaction networks (PPINs)

As its name implies, protein-protein interaction networks are the biological networks representing relationships between proteins. Proteins are responsible for several biological processes in the cell, such as catalysis, transportation, signal transduction, growth control, development, etc. Protein interactions can be divided into two categories based on their affinity, obligate and non-obligate, and non-obligate interactions can be transient or permanent based on the lifetime of the interaction. Permanent interactions are strong and irreversible, while transient interactions happen for a limited time and are broken easily. Transient interactions are involved in several cellular processes such as the transmission of regulatory signals and post-translational modifications [58,59].

PPINs are, by their nature, dynamic. As they reflect the cellular processes, all interactions occur in a time-based, location-based, or condition-based manner. Several methods are used to capture these binary relationships according to the nature of the interaction. For example, strong permanent interactions are detected by co-immunoprecipitation assays, functional associations in which two proteins do not necessarily interact with direct physical contact are detected by biochemical assays, condition-based interactions are detected by gene expression data, transient interactions are detected by high-throughput yeast-two-hybrid assays, and tandem affinity purification by use of chemical crosslinking followed by mass spectroscopy, etc. [59–61].

Databases like DIP [62], BIND [63], HPRD [64], HPA [65], and STRING [66] provide information for protein-protein interactions. An example of PPIN of *EGFR* produced with STRING is shown in Figure 2.

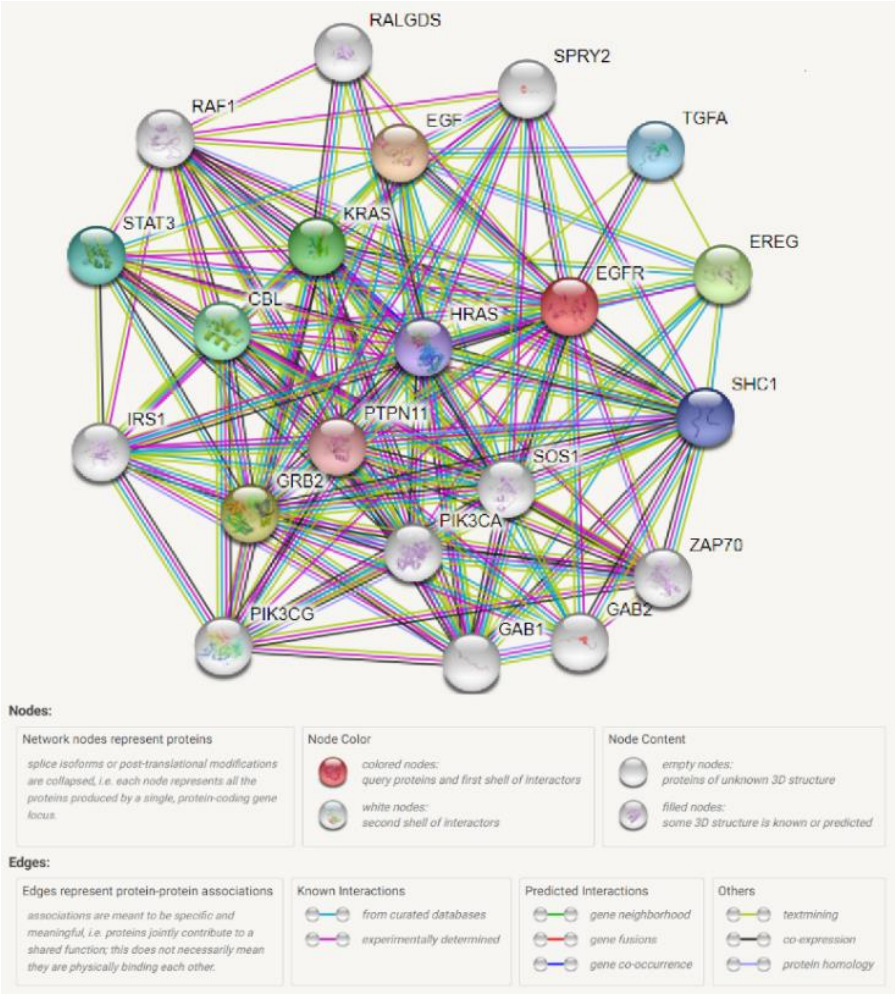


Figure 2. Example PPIN for EGFR gathered from STRING.

2.2.3. Metabolic networks

Metabolic networks include all chemical reactions of metabolism in the organism of interest that represents biochemical and physiological processes within the cell. Nodes usually represent metabolites, while edges represent catalytic reactions converting one into the other. Metabolic networks generally include enzymes, substrates and products of reactions, information about the reaction and enzymes, substrate specificity, and cellular localization information [67].

Databases like KEGG [68], Reactome [69], ENZYME [70], and BRENDA [71] provide information for metabolic networks. Metabolic network of *EGFR* produced with Reactome is shown in Figure 3.

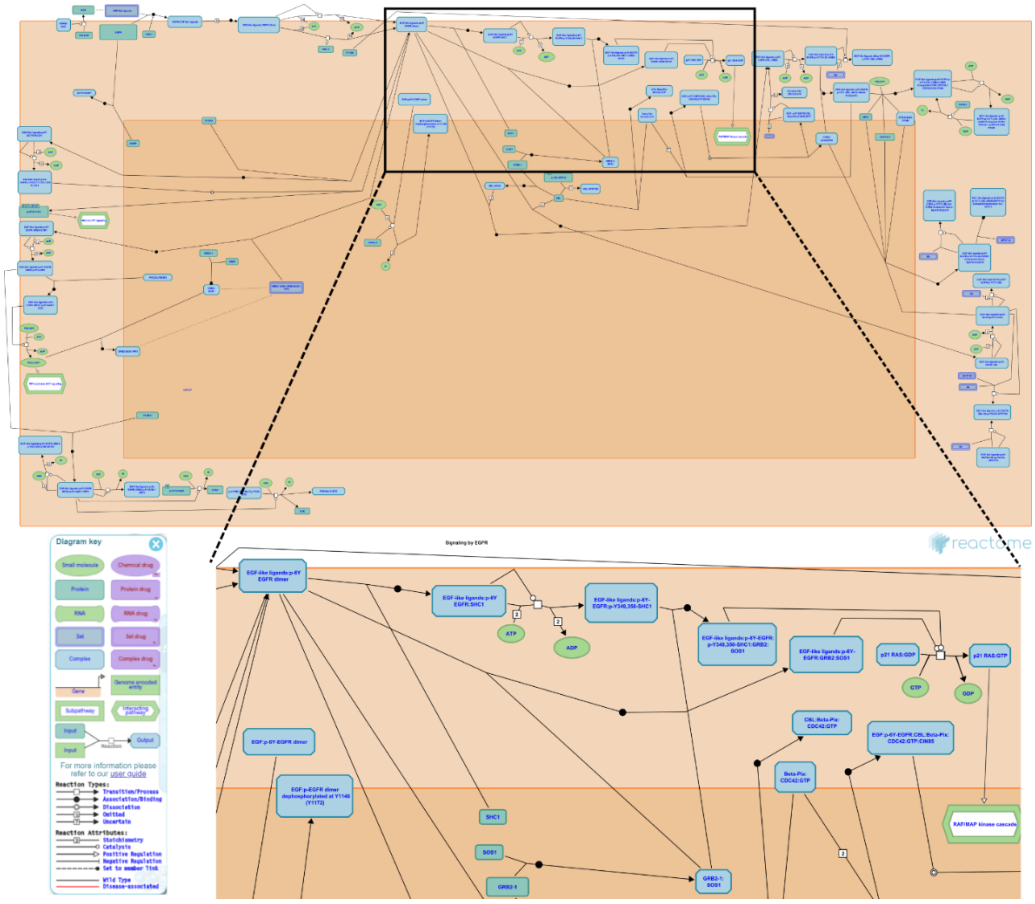


Figure 3. Example metabolic network for EGFR signaling gathered from Reactome.

2.2.4. Cell signaling networks

Cells acquire integrated processes to make decisions for homeostatic balance, development, and survival. In order to maintain cellular mechanisms necessary for the current situation of the cell, it responds to different internal and/or external signals; either chemical, mechanic, or electrical and related signaling pathways are activated. Upon primary signal response, information flows through several signaling pathways with molecular interactions as well as biochemical reactions. As cellular processes are interconnected within the cell, signaling pathways produce complex networks[72,73]. To better understand the response mechanisms of the cells against different conditions such as perturbation, mutation, and disease[74–78], cell signaling networks are studied using various network analysis methods.

Reactome [69], KEGG [68], and TRANSPATH [79] are examples of databases for cell signaling information. Cell signaling network for *EGFR* Tyrosine Kinase inhibitor resistance pathway (hsa01521) from KEGG database is shown in Figure 4.

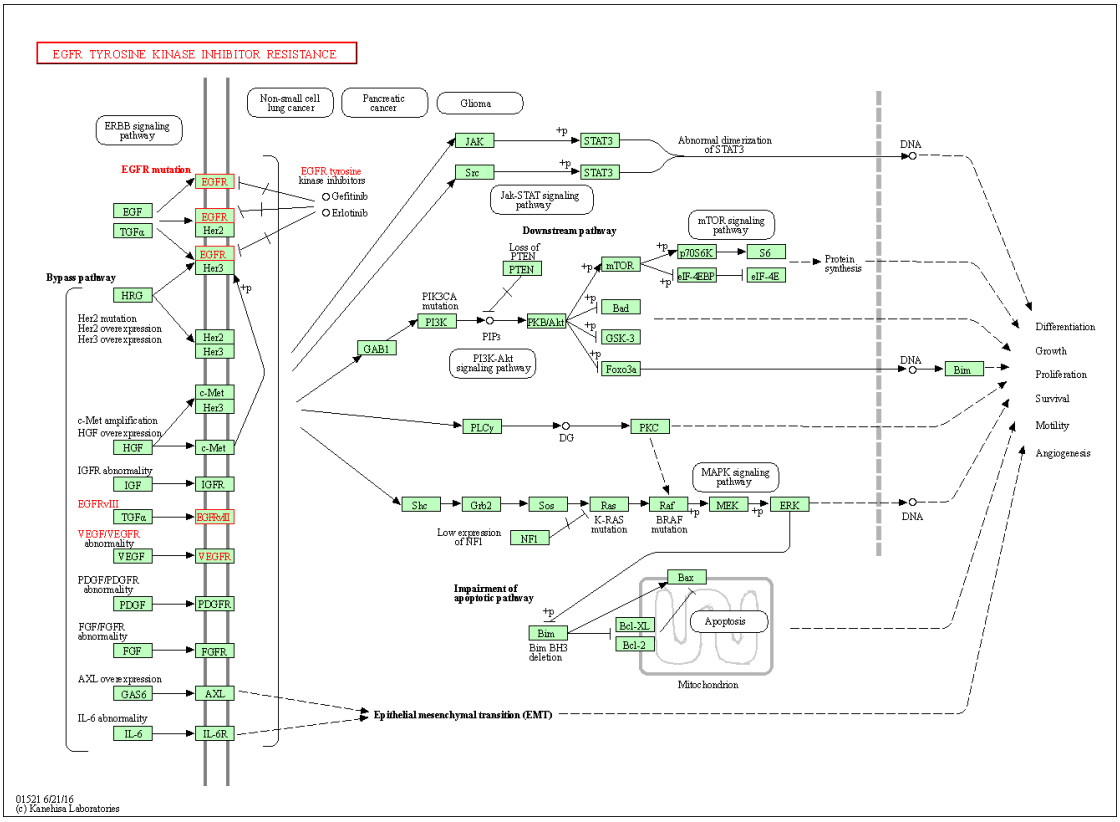


Figure 4. Example cell signaling network for EGFR Tyrosine Kinase inhibitor resistance pathway (hsa01521) gathered from KEGG database.

2.2.5. Gene co-expression networks

Gene co-expression networks (GCNs) are built from gene expression values in which genes represent nodes, and those having correlated expression patterns have an edge between them. These kinds of biological networks are generally undirected networks, and they can be either weighted or unweighted. To measure the correlation between the transcripts across experimental conditions or samples, the frequently used methods are Pearson's correlation coefficient, Spearman's rank correlation coefficient, Mutual information, and Euclidean distance[80]. All genes for each condition/sample constitute an expression vector, and these vectors are scored to construct the GCNs. GCNs can be used to detect highly connected hub-genes which are candidates as biomarkers for the condition of interest [81–84]. On the other hand, genes having similar expression patterns may not necessarily mean that they are also functionally related. Transcription of a gene may not directly reflect its corresponding protein levels, and also, two genes may share the same cis-regulatory DNA motifs leading to co-expression by chance [85].

COXPRESdb [86] and GeneFriends [87] are some of the databases that can be used to get information for gene co-expression.

2.2.6. Residue interaction networks (RINs)

Residue Interaction Networks (RINs) are biological networks to study protein structures in which nodes represent aminoacids and edges represent non-covalent interactions such as hydrogen bonds, Van der Waals interactions, and secondary structures. Since the 3D structure of a protein determines its function, topological analysis of RINs provides valuable information about folding, functionally important and central residues, interface regions, allosteric and active sites, effects of mutations on the structure etc.[88–95]. While constructing RINs, different thresholds can be specified for atomic distances such as C-C and C-S, distances of hydrogen bonds, ionic interactions, π–π stacking, π–cation interactions, etc. Also, the decision on whether to use additional features while reconstructing RINs, such as the presence of water molecules may provide more detail on the properties of the protein [96].

RING [90], RIP-MD [97], and ProSNEx [98] are three examples of tools, and RINspector [99] is an example of the Cytoscape [100] app to study residue interaction networks. RIN of *EGFR* produced with RING is shown in Figure 5.

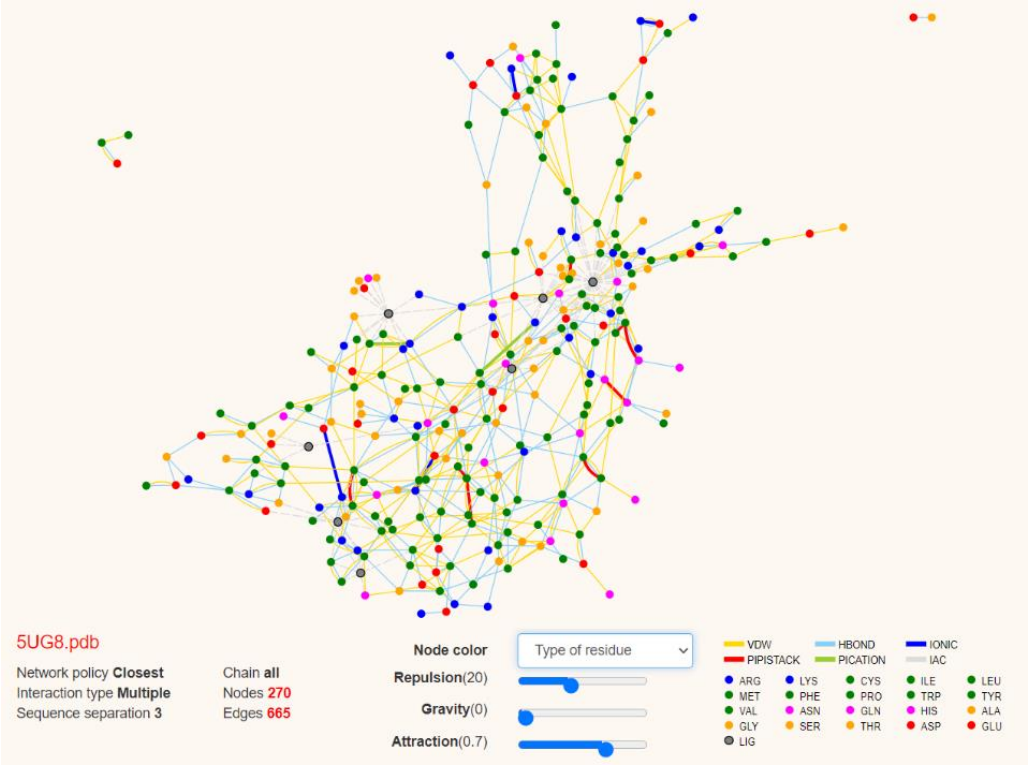


Figure 5. Residue interaction network of EGFR with pdb id 5UG8 generated with RING web tool

2.3. Network Reconstruction Methods

Recovering interactions within systems using prior bibliomic and experimental data and building a graph-based infrastructure is known as network reconstruction/modeling. Network reconstruction is based on the idea of reverse engineering. It is sometimes called as 'network inference' [101].

Network reconstruction plays an essential role in understanding the behavior of several cellular components at the systems level. It is used to mimic and take a snapshot of cellular activities at a given point in time. Sometimes networks are modeled in a time-series manner to reveal a real-time situation such as drug treatment, disease condition, etc. At this point, selecting an appropriate model for network reconstruction according to the data in hand is both crucial and challenging. The reconstruction of an accurate network according to the real data may help investigate network motifs and conservation of cellular compartments and study evolutionary relationships between them [102–104]. It can also increase the reliability of cellular component interactions found via experiments and may lead to new interaction findings that would provide time and cost gaining in wet-lab

experiments. Also, an adequate network may help minimize the number of drug targets and find new drug targets while studying a specific disease condition [105,106].

There are several methods used for network reconstruction purposes. These methods can be divided into five main categories: (i) Boolean network models, (ii) Bayesian network models, (iii) Differential equation models, (iv) Information theory models, and (v) Neural network models. Below each of these methods is basically explained.

2.3.1. Boolean Network Models

Boolean network models are dated back to 1969, firstly proposed by Kauffman [107]. They are dynamic networks and rely on discrete data. Boolean networks are built onto a variable set ($X= \{x_1, x_2, x_3, \ldots, x_n\}$, $x_i \in B$), and each variable represents the actual state of its corresponding bio entity. For state designation, binary variables {0, 1} are used, e.g., on/off switches of genes and active/inactive state of proteins.

Boolean networks are known with their ease to implement [46]. They are widely used for network inference from time-series gene expression data. However, they have limitations on the discretization of the data. For example, if gene expression data is used to reconstruct a network, it should be transformed into binary data. Since genes do not always have strict on/off states and expression levels may lie within a range leading to different regulatory conditions, using Boolean network strategy may cause loss of information [45,46]. To overcome this issue, fuzzy operators have been incorporated into the Boolean network models to reconstruct biological networks with continuous variables[108,109]. Also, noise in data arises as another problem. Akutsu et al.[110] proposed algorithms to reconstruct Boolean networks with noises. Shmulevich et al.[111] studied probabilistic Boolean networks for gene regulatory network reconstruction that can manage uncertainty. Leifeld et al.[112] showed the application of different types of prior knowledge for biological network inference using Boolean models and considered noisy measurements and missing data points. There are several reviews explaining the concepts for Boolean network modeling and presenting examples of application to biological data [113–115].

REVEAL (REVerse Engineering ALgorithm)[116] is one of the algorithms implementing Boolean network models. Several Boolean network analysis tools are developed in different programming languages such as ChemChains[117], Pint[118], SimBoolNet[119], BoolNet[120], CellNetAnalyzer[121], etc.[113].

2.3.2. Bayesian Network Models

Bayesian networks use Bayes theorem of probability, then combine graph theory to qualitatively model network properties [122]. Bayesian networks are built onto directed acyclic graphs (DAGs). Each interaction in the graph has a probabilistic dependency. The algorithm needs to process three essential parts for learning; model selection (a DAG as a candidate), parameter fitting (using the candidate DAG and data, recover the best fitting conditional probabilities), and fitness rating (scored network models) [45,46]. All models

produced with the algorithm are scored, and the highest score refers to the best fitting network (DAG) selected.

Bayesian learning can be performed from both discrete and continuous data. Therefore, Bayesian networks are advantageous over their flexibility. Different types of omics data can be combined as well as prior knowledge to reconstruct Bayesian networks. Also, Bayesian networks are advantageous with their ability to handle missing data[123]. However, computational complexity increases with increasing data size. Constraint-based methods can be used to reduce the search space by using restrictions to deal with the computing load of score-based Bayesian network modeling. However, constraints come with their disadvantage to increase the potential for compounding error [46,124,125].

Bayesian Network Webserver[126] is an example of tools for biological network modeling using the Bayesian algorithm. BNFinder[127] is a python library, and bnlearn[128] is an R package designed as a tool for learning Bayesian networks.

2.3.3. Differential Equation Models

Ordinary Differential Equation (ODE) approaches are widely used in biological network reconstruction, especially for modeling continuous dynamic networks from gene expression data. The current ODE-based methods include mainly three categories: (i) the *law of mass action*[129,130], (ii) *Hill function*[131], and (iii) *Michaelis-Menten Kinetics*[132]. The choice for the correct ODE function depends on the biological question and/or nature of experimental data[133]. According to the *law of mass action*, the reaction rate depends on the probability of the collision of the reactants. *Hill function* relates binding affinities of ligands to receptors with the concentration of the ligands. Hill functions are mostly used to model signaling networks since a protein's activation or inhibition is induced by their upstream parental nodes which Hill function[133] can characterize. *Michelis-Menten kinetics* describes reaction rates with enzyme kinetics.

When the network scale gets larger, ODE-based models become disadvantageous because they need to determine many parameters. The high number of parameters leads to a high computational burden, which in turn causes lower prediction accuracy[46,133]. Also, ODE-based models generally use linear equations and rarely specific types of non-linear equations[45]. However, regulatory interactions usually include complex non-linear dynamics, which need more complex versions of models such as Genetic Algorithm (GA)[134], Particle Swarm Optimization (PSO)[135], and scatter search[136]. There are several variations that apply ODEs. Some examples are Minimum Weight Solution to Linear Equation (MWSLE), Fourier Transform for Stable Systems (FTSS)[137], and Evolutionary Modeling (EM)[138].

2.3.4. Neural Network Models

Neural network models are inspired by real neural structures observed in the brain. Neural network models are advantageous over their ability to use several inputs/data structures

and give one output. They are flexible as they can recognize input patterns[122,139]. Other advantages of neural network models are that they can model any functional relationships and data structure, acquire nonlinear and dynamic interactions, and resist to noise. However, the learning rate must be defined for different data situations, making it challenging to perform efficient training. Also, it consumes high computing power when data gets large. As a result, neural networks are most useful for small networks[122].

Neural network models include two main strategies: Artificial Neural Network (ANN) and Recurrent Neural Network (RNN). ANNs constitute a pure neural approach, while RNNs use fuzzy logic[140]. RNNs can handle feedback mechanisms which are natural regulatory processes within the cell, allowing the modeling of non-linear and dynamic interactions[139,141,142].

New approaches to neural network models have been proposed recently. The Extreme Learning Machine (ELM) method, a supervised neural model, was shown to have a higher learning rate and better performance in terms of predictive power[143]. Also, deep-learning methods mainly include ANN models. Convolutional neural networks (CNNs) and fully connected neural networks (FCNNs) are other classes of neural networks. FCNNs have problems such as overfitting or trapping in the local optimum when the network depth increases, resulting in a higher number of parameters. CNNs take input data in the form of multidimensional arrays. PyTorch and TensorFlow are two widely used frameworks for deep learning which allow building neural networks [43].

2.3.5. Information Theory Models

Information theory-based network modeling approaches generally depend on the dependencies between bio entities. Determination of dependencies is mainly measured with correlation coefficients such as Pearson, Spearman, or Kendall. Moreover, mutual information and euclidean distances are the other methods used for network reconstruction. Information theory models are the most common methods because the application is straightforward and does not need much computing power. They can be used for very large networks. However, they are static models and cannot infer dynamic interactions, directionality, causality, or the multiple bio entities in one regulatory interaction[46,144].

Examples of tools using information theory for network reconstruction include RELEVANCE[145], ARACNE (Algorithm for the Reverse engineering of Accurate Cellular Network)[146], CLR (Context Likelihood of Relatedness)[147], MRNET (based on the maximum relevance/minimum redundancy)[148], and FyNE (Fuzzy NEtworks)[149].

2.4. Data Integration

Meta-dimensional analysis can be classified into four main categories according to data integration and statistical approaches; Multivariate methods, Concatenation-based methods, Transformation-based methods, Model-based methods. Some multivariate methods[150] are; Partial Least Squares (PLS)[151,152] and Canonical Correlation Analysis (CCA)[153,154]. In the literature, several approaches to PLS-based methods[152,155–160] and CCA-based methods[154,161,162] can be found. Some of the concatenation-based integration[6] methods are low rank-based approximation[163], latent factor analysis[164], Co-Inertia Analysis (CIA)[165], Multiple Co-Inertia Analysis (MCIA)[166], and also Multiple Factor Analysis (MFA)[167,168], which is based on Principle Component Analysis (PCA). Transformation-based methods basically use graph or kernel matrix[169–174]. An example of the transformation-based approach is Similarity Network Fusion (SNF)[170]. Model-based approaches are based on learning strategies (generally Bayesian approaches are used[175–177]), in which a model is obtained by a training phase, and the model is used for data integration. An example of a model-based approach tool is ATHENA[21].

Tablo 1. Current summary of data integration tools
(adapted from Huang et al.,2017 [178], detailed version can be found in Appendix A)

Name	Data Type	Reference
Joint NMF	Multi-data	172,179
iCluster	EXP, CNV	180
iCluster+	Multi-data	181
JIVE	Multi-data	182
Joint Bayes Factor	EXP, MET, CNV	183
ssCCA	Sequence data	169
CCA sparse Group	Two types of data	162
sMBPLS	Multi-data	157
SNPLS	EXP, drug response, gene network info	184
MDI	Multi-data	185
Prob_GBM	EXP, CNV, miRNA, SNP	186
PSDF	EXP, CNV	187
BCC	EXP, MET, miRNA, proteomics	188
CONEXIC	EXP, CNV	175
PARADIGM	Multi-data	189
SNF	EXP, MET, miRNA	170
Lemon-Tree	EXP, CNV/miRNA/methyl	190
rMKL-LPP	Multi-data	171
CNAmet	EXP, MET, CNV	191
iPAC	EXP, CNV	192
ATHENA	EXP, CNV, MET, miRNA	21

Table 1. Current summary of data integration tools (adapted from Huang et al.,2017 [178], detailed version can be found in Appendix A)-(cont.)

jActiveModules	EXP, PPI, protein-DNA interactions	[193]
Network Propagation	Gene expression, mutation, PPI	[194]
SDP/SVM	EXP, protein sequence, protein interactions, hydropathy profile	[195]
FSMKL	EXP, CNV, Clinic feature (ER status)	[196]
iBAG	Multi-data	[197]
MCD	MET, CNV, LoH	[198]
Anduril	EXP,MET, miRNA,exon, aCGH,SNP	[199]
GeneticInterPred	EXP, PPI, protein complex data	[200]
Graph-based Learning	EXP, CNV, MET, miRNA	[201]
CoxPath	EXP, CNV, MET, miRNA	[202]
MKGI	EXP, CNV, MET, miRNA	[203]

2.5. Link Prediction

Link prediction is the process of scoring the likelihood of a future/missing edge between two nodes, knowing that there is no edge between those nodes in the current state of the graph[204]. Link prediction strategies are used in several areas, such as "friend-finder" features in social networks[205], recommendation systems features in e-commerce[206], and identifying hidden groups of criminals in the security domain[207]. In bioinformatics, link prediction approaches have become popular because different types of biological networks are used in several bioinformatics areas. It is known that molecular interactions are currently incomplete, which is a limiting factor to understand molecular mechanisms underlying diseases, perturbations, internal and external signaling, etc.[208–213]. Experimental procedures that are used to extract molecular interactions have many challenges. For example, interactions between bio-entities may be missed due to localization differences while collecting them during experimental setup or due to the nature of the interaction (transient interactions). Also, since biological systems are very complex, interactions between bio-entities may have difficulties while capturing by experiments, and bias for highly studied bio-entities may prevent finding out interactions of other low-studied ones. Besides, biological/experimental noise can result in a loss of interactions. Link prediction in bioinformatics provides an opportunity to find interactions between bio-entities that are invisible due to missing data, experimental errors, etc.[214]. These techniques take advantage of many network properties such as degree, clustering, and path lengths. Link prediction algorithms generally score the likelihood of interaction of node pairs, and these scores are listed in decreasing order.

Link prediction methods are classified into two main categories; supervised and unsupervised. In figure 6, the main methods of each category are illustrated. There are several studies on supervised link prediction methods. For example, Kashima et al.[215,216] proposed (i) parameterized probabilistic model for supervised link prediction and (ii) "Link Propagation" as an adaptation of label propagation which is a semi-supervised learning method. They applied their method to two different biological data, i.e., the metabolic pathways of the yeast *S. Cerevisiae* in the KEGG database and a protein-protein interaction network dataset constructed by another group. Shojaie[217] performed link prediction in biological networks using a penalized multimode exponential random graph model (MP-ERGMs). In [218], they investigated the effect of four neural network models on the performance of a neural link predictor in biomedical graphs.

Unsupervised methods have node-based and path-based topological similarity metrics. From node-based topological similarity methods (local metrics), Common Neighbors illustrates the situation such that two stranger people can be introduced to each other with a common friend, which has the effect of "closing a triangle" in the graph. Jaccard's Coefficient measures the probability both nodes *u* and *v* have a neighbor *y* for a randomly selected neighbor *y* that either *u* or *v* has. The Adamic-Adar method is a frequency-weighted version of the Common Neighbors method. Adamic/Adar method weights rarer properties more heavily. For example, if two non-interacting nodes, *u* and *v*, have a common neighbor *y* which is a node with a low degree, it's more probable that *u* and *v* are interacting. The Preferential Attachment method relies on the assumption that nodes with higher degrees tend to interact with more nodes which increases the possibility of hub formation.

General formulas for node-based methods are:

- Common Neighbors:

(1) $$Score(x, y) = |Neighbors(x) \cap Neighbors(y)|$$

- Adamic/Adar:

(2) $$Score(x, y) = \sum_{w \in Neighbors(x) \cap Neighbors(y)} \frac{1}{\log|Neighbors(w)|}$$

- Jaccard's Coefficient:

(3) $$Score(x, y) = \frac{|Neighbors(x) \cap Neighbors(y)|}{|Neighbors(x) \cup Neighbors(y)|}$$

- Preferential Attachment:

(4) $$Score(x, y) = |Neighbors(x)| \cdot |Neighbors(y)|$$

Path-based topological similarity methods use global network properties. There are three widely known methods in this category: Katz, Hitting time, and Rooted PageRank. Katz method is proposed by Leo Katz[219]. As a measure of score between nodes u and v, the Katz method sums all possible path lengths between *u* and *v* and penalizes longer paths by a factor β. Hitting time is built upon the concept of random walks on the graph. To measure the score of a predicted edge between the nodes *u* and *v*, it counts the expected time for a random walk starting from *u* to reach *v*. Smaller the score of hitting time; it is more probable that *u* and *v* are connected. Hitting time is not a symmetric metric. That's why the commute time is used in undirected graphs, which is $C_{x,y} = -(H_{x,y} + H_{y,x})$. Rooted PageRank is an alternative to hitting time. It is designed to escape from the problem of walking to a node *z* that has high stationary probability and is far away from *u* and *v*.

General formulas for path-based methods are:

- Katz:

(5) $$Score(x,y) = \sum_{l=1}^{\infty} \beta^l \cdot |paths_{x,y}^{<l>}|$$

- Hitting Time:

(6) $$Score(x,y) = -H_{x,y}$$

- Rooted PageRank:

(7) $$Score(x,y) = (1-\beta)(1-\beta N)^{-1}$$

(where D is a diagonal degree matrix with $D[i,j] = \sum_j A[i,j]$ and let $N = D^{-1}A$ be the adjacency matrix with row sums normalized to 1, β is the probability of walking to a random neighbor.)

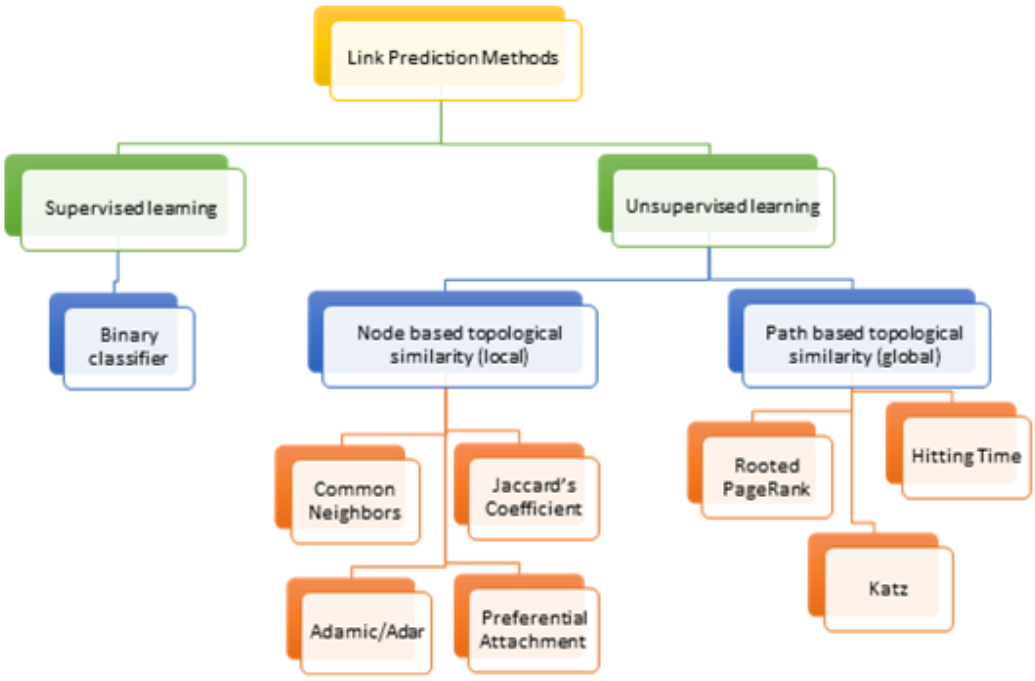


Figure 6. Classification of link prediction methods

There are several studies involving node and path-based link prediction applications on biological networks. Cannistraci et al.,2013 introduced the CAR index and applied it to adapt widely known node similarity-based link prediction methods, namely Common Neighbors, Adamic-Adar, Jaccard index, and Preferential Attachment. CAR index assembles node-based and link-based prospects. It is defined as the possibility of being linked for two nodes is higher when their common-first-neighbors are included in a 'local-community'. The links within local communities are named as 'local-community-links'. Local-community is a collection of robustly linked nodes. They demonstrated the power of CAR index applied node-based link prediction methods in brain connectomes. They randomly destroy a certain percentage of synapses and measured the correct prediction rates of these methods. They observed that up to 50% of original connectome synapses could be predicted back, and the CAR-index variant of Preferential Attachment (CPA) performed better than others. They also tested CAR index variants on protein interactomes and concluded that CAR-index increases the overall prediction performance[220]. For example, Lu et al., 2017 used bipartite networks in which drugs and targets constitute two disjoint sets of nodes. They applied modified versions of four unsupervised link prediction methods (Common Neighbors, Jaccard index, Preferential Attachment, and Katz index) to predict drug-target interactions[221]. Kovacs et al., 2019 proposed the L3 method for link prediction in protein interactomes based on the idea that two proteins interact when one is similar to the other's partners. According to the L3 method, which relies on path lengths of three, if two proteins have interacting partners, then it is probable for them to interact,

too. They compared L3 prediction against other node similarity-based methods and CAR-index variant of Resource Allocation (CRA) and showed that the L3 method performs better than other methods[222]. Pech et al.,2019 proposed an optimization to L3 prediction, considering that the probability of two nodes being linked is related to a linear combination of similarities between the neighbors of each of them. They performed a linear optimization (LO) method to produce a link likelihood matrix [223]. Chen et al.,2020 used the information gathered from proteins' complementary interfaces and gene duplication and proposed Sim as a link prediction method for PPI networks. They used Jaccard Similarity in Sim for sequence alignment. Seven link predictors – L3, LO, Sim, SimCN, SimPA, SimRA, and L3 + Sim (integration of Sim with L3) – are compared in the study, and Sim+L3 performed better on prediction and robustness for only PPI networks[224]. Yuen & Jannson, 2020 proposed the ExactL3 link prediction method for PPI networks and tested its performance on four PPIs of STRING, BioGRID, IntAct/HuRI, and MINT. They suggest that ExactL3 infers undiscovered PPIs better and characterizes protein interactions biologically by using only network topology. They measured the biological relevance of edge predictions found by ExactL3 in terms of Gene Ontology (GO) Semantic Similarity (GOSemSim), STRING Confidence Scores, and Gene Essentiality. They showed its superiority over other methods such as L3, CRA, Sim, and CN[225].

Overall, all link prediction methods have their properties to predict edges accurately, depending on the mechanism for how the biological network grows[226]. Molecular interactions may evolve based on different properties according to the interaction types and data types[227]. Therefore, it is very important to understand the growing nature underlying the biological network and select the link prediction method accordingly[228]. Also, the density of the reference graph is significant for the necessary computing power. When the graph is dense, there are not many edges left to predict, so the probability of randomly choosing the correct edge from this remaining possible edge list is high. However, when the graph is sparse such that $E \approx V$, there are almost V^2 edges non-existed in the current graph, and the possibility of selecting the correct edge out of V^2 edges is lower.

Linkpred[229] and LPmade[230] are two of the computational tools for link prediction on biological networks. Moreover, the networkx[231] module in python has its own functions for some of the link prediction methods.

CHAPTER 3

MATERIALS AND METHODS

The main workflow of this study for the integrative network modeling of drug responses to reveal drug modulation based on drug and tumor type is illustrated in Figure 7. Firstly, the transcriptomic and phosphoproteomic dataset for small molecule perturbation on six different cell lines (CMap) is retrieved and statistically preprocessed for significant treatment conditions across associated control treatment. Secondly, drugs are classified according to two traditional methods, i.e., chemical similarity-based classification and literature curated mechanism of action (MoA)-based classification. After, network reconstruction studies by use of Omics Integrator software are performed. Two input files are needed for network reconstruction: seed proteins and a reference interactome. These two inputs are prepared by the application of several approaches. All cell line and drug-specific networks are modeled and used for further classification procedures. Finally, drug groups produced by different methodologies are compared to understand the advantages of network-based studies. All methods used for these steps are detailed within the below sections.

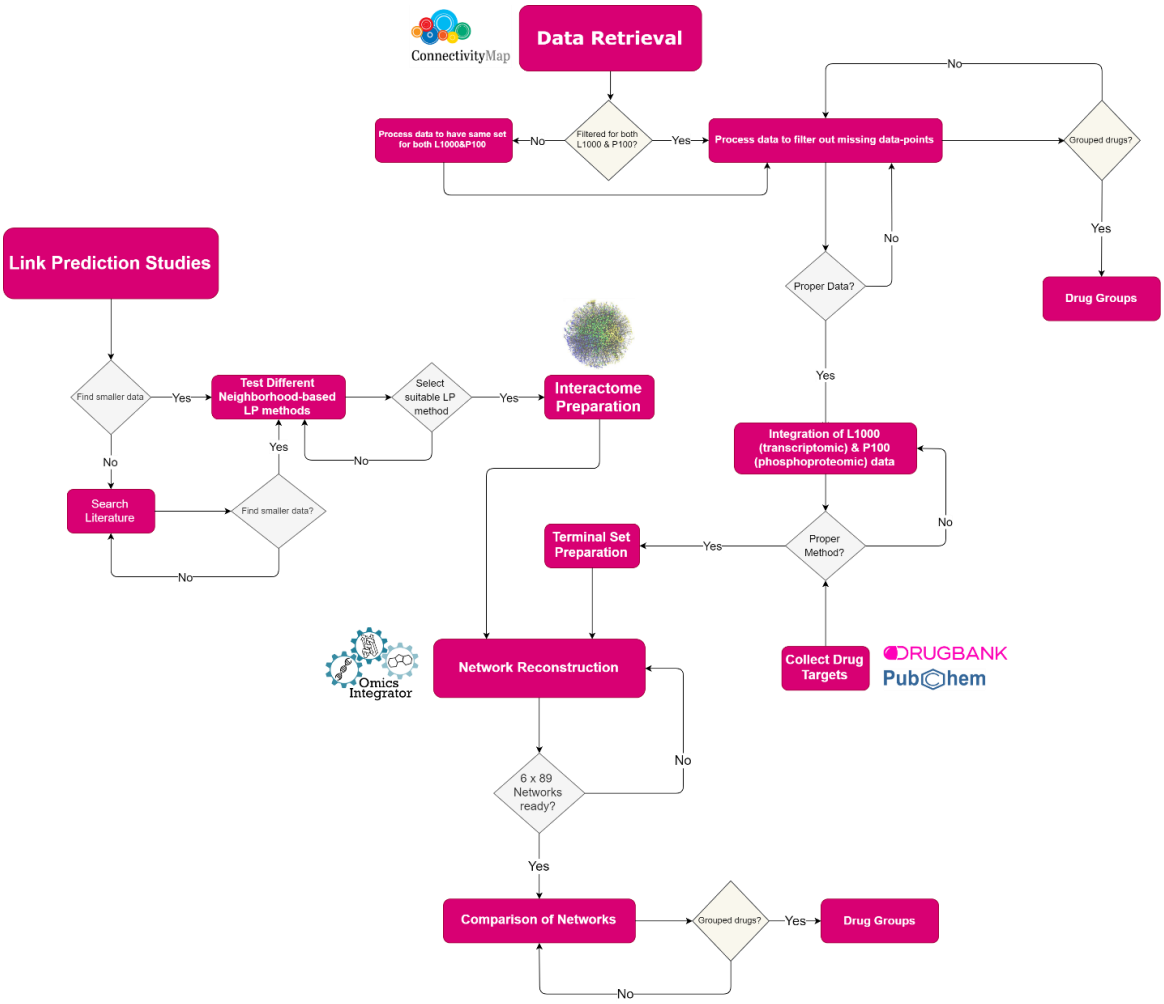


Figure 7. Workflow of thesis study. After data retrieval and preprocessing for collecting significant datasets and excluding missing information, drugs are classified using traditional classification methods to later compare with network-based classification. Meanwhile, two input files necessary for network reconstruction studies are prepared by integrating different data types to collect terminal set and application link prediction strategies on the human interactome and other filtering protocols.

3.1. Data

The ‘omic’ data is obtained from Connectivity Map (CMap), which contains 1.3 M L1000 data for nine different core cell types, treated with 27927 perturbagen. Within CMap, the Touchstone V1.1 dataset contains 8388 perturbagens (well-annotated genetic and small-molecular perturbagens) for nine cell lines for three time points and three replicates; and P100 dataset contains proteomic data for six different core cell types, treated with 90

perturbagens[38,232,233]. In addition to transcriptional data (L1000), proteomic and phosphoproteomic data for a subset of these perturbagens (P100) were released on 4/13/18 with the accession number GSE101406.

P100 dataset was used in this study which contains already filtered transcriptional and phosphoproteomic data for the same set. The P100 dataset represents three different read-outs (phosphosignaling (P100), chromatin modifications (GCP), and transcriptional changes(L1000)) of the same 90 small-molecule perturbations in six cell lines. A list of these small molecules and cell lines is provided below in Table 1 and Table 2. The duration of treatment was 3 hours for P100, 24 hours for GCP, and 6 hours for L1000. These data are available at multiple levels of processing: level 1 is fluorescence intensity (for L1000) or mass spectrometry extracted ion chromatogram traces (for P100, GCP); level 2 is gene expression or proteomic values without normalization; level 3 is normalized; and level 4 is differential (i.e., each sample is compared to all other samples on a plate). Normalized (level3) data of transcriptomics(L1000) and phosphoproteomics(P100) are used in this study.

L1000 assay measures transcription levels of 978 genes (referred to as Landmark genes) and 80 control transcripts directly. It then infers the expression levels of the remaining 11350 genes via Ordinary Least Square (OLS) regression. From 11350 inferred genes, 9196 genes are considered well inferred and called as Best Inferred Genes (BING). Landmark genes are used in the terminal set preparation procedure (section 3.4.1). Phosphoproteomic data generated by the P100 assay consists of 96 phosphosites and is a reduced representation of phosphoproteomics, targeting common signaling pathways.

The drug-target interactions are retrieved mainly from CLUE Drug Repurposing tool[234], which curates information from several databases such as DrugBank[235] & PubChem[236] (https://clue.io/repurposing-app). This tool provides an opportunity to browse drugs and gives various information about drugs, including the target molecules. The reconstructed networks have been oriented through these targets for each drug.

The primary reference human interactome used is the edge-weighted protein-protein interaction network retrieved from iRefWeb, v13.0, which has 15,404 nodes (proteins) and 175,820 weighted edges (protein interactions) without self-loops. The weights of edges representing how real an interaction is based on the MIScore function. It is downloaded from Omics Integrator GitHub repository (https://github.com/fraenkel-lab/OmicsIntegrator/tree/master/data).

Tablo 2. List of Small-Molecule Perturbagens

1271738-62-5	vu0155056	etoposide	methylstat	semagacestat
bafilomycin a1	4,5,6,7-tetrabromobenzotriazole	everolimus	niclosamide	sirolimus
bms-906024	alpelisib	EX-527	nilotinib	sotrastaurin
BRD-K68548958	alvocidib	exifone	olaparib	staurosporine
BRD-K73261812	AR-A014418	geldanamycin	OSI-027	tacedinaline
calpain inhibitor ii	belinostat	ginkgetin	palbociclib	tacrolimus
compound e	BIX-01294	GSK-J4	pazopanib	TG-101348
cpi-169	BIX-01338	I-BET-151	PD-0325901	tofacitinib
CYT387	BMS-345541	I-BET-762	pravastatin	tretinoin
epz-5676	CC-401	IKK-inhibitor-X	pyrazolanthrone	trichostatin-a
Gossypetin	CHIR-99021	IPI-145	resveratrol	UNC-0321
gsk126	curcumin	JQ1-(+)	RG-4733	UNC-1215
GSK-2110183	decitabine	KN-62	RGFP-966	VE-822
NVP-BEZ235	dexamethasone	KN-93	rolipram	vemurafenib
okadaic acid	dinaciclib	KU-55933	roscovitine	verteporfin
sch 900776	entinostat	lenalidomide	ruxolitinib	vorinostat
smer-3	epz004777	losmapimod	salermide	zebularine
unc-0646	EPZ-005687	LY-294002	selumetinib	DMSO

Tablo 3. List of Cell Lines

Cell Line	Description
A375	human malignant melanoma, a solid tumor from a 54-year-old female
A549	human Caucasian lung carcinoma, isolated in 1972 from a 58-year-old Caucasian male
MCF7	human breast carcinoma, isolated in 1970 from a 69-year-old Caucasian woman
PC3	human prostate adenocarcinoma, isolated in 1979 from a 62-year-old Caucasian male
YAPC	human pancreatic carcinoma, isolated in 1993 from a 43-year-old Japanese man
NPC	neural progenitor cells

3.2. Statistical Analyses and Data Processing

All statistical analyses were performed with python SciPy module[237].

In the network reconstruction protocol, processing of L1000 and P100 data is performed. L1000 data includes fluorescence values of several replicas for transcriptional

measurements of both drug-treated samples, and DMSO treated samples to be used as control. These two conditions for each landmark (genes that are directly measured by L1000 assay) gene were compared with one way-Anova method after verifying the assumptions (*scipy.stats.f_oneway* function). The data for p-values for each gene as drug-treated condition compared against control condition is generated. Genes with a lower p-value than the selected threshold were collected and labeled as 'significantly transcribed genes'. These 'significantly transcribed genes' are listed for each cell line-drug pair. Also, log2 fold changes of average transcription values of each gene relative to the control condition are calculated and stored to be later used in the prize designation of terminals. The same procedure is performed for P100 data to be able the collect 'significantly phosphorylated proteins' for $p<.05$ and calculate their log2 fold changes against the control treatment.

Regression analysis performed in this study is also performed by use of *scipy.stats.pearsonr* function of python SciPy module.

Clustering of matrices is performed by use of scipy.cluster.hierarchy function of python SciPy module where necessary.

3.3. Chemical Structure-Based Drug Classification

Two different metrics were used for the investigation of chemical structure similarity. 70 drugs out of 90 are listed after the Anova significance test. SMILES signatures of 70 drugs are curated from chemical databases such as DrugBank[235] and PubChem[236].

3.3.1. Tanimoto Similarity Calculation

Tanimoto similarity[238] of two compounds are calculated with their SMILES signatures. It calculates the Jaccard's coefficient between fingerprints of two compounds.

Rdkit[239], an open-source python library, is used for the construction of the Tanimoto similarity distance matrix.

Hierarchical clustering on the matrix is performed using python scipy.cluster.hierarchy function of SciPy module[237].

3.3.2. MACCS key distances

MACCS (Molecular ACCess System) keys (166-bit 2D structural fingerprints) distance is a SMARTS (SMILES arbitrary target specification)-based implementation of the 166 public MACCS keys, which are sometimes referred to as the MDL keys[240].

Rdkit[239], an open-source python library, is used for the construction of the MACCS key distance matrix.

Hierarchical clustering on the matrix is performed using python scipy.cluster.hierarchy function of SciPy module[237].

3.4. Network Reconstruction

Network reconstruction is performed by use of Omics Integrator software[39]. Omics Integrator searches for a solution to the prize-collecting Steiner tree problem to find an optimum tree. Nodes obtained from the experiments (terminal nodes) and nodes not detected in experiments and obtained by the algorithm (Steiner nodes) are determined by this process.

Let G = (V, E, c, p) be an undirected graph, with the vertices/nodes V associated with non-negative prizes p(v), and with the edges E associated with non-negative costs c(e). The Prize-Collecting Steiner Tree problem (PCST) consists of finding a connected subgraph T = (V' , E') of G, that maximizes profit(T) which is defined as the sum of all node-prizes taken into the solution minus the costs of the edges needed to establish the network (Figure 8). It can also be defined as the minimization of the weight of T, which is the sum of its edge costs plus the sum of the penalties of the vertices of G outside of the solution. There are both the *rooted* and *unrooted* variants for each problem where T must contain v_0 in rooted variant, and T can be any subtree in unrooted variant[241].

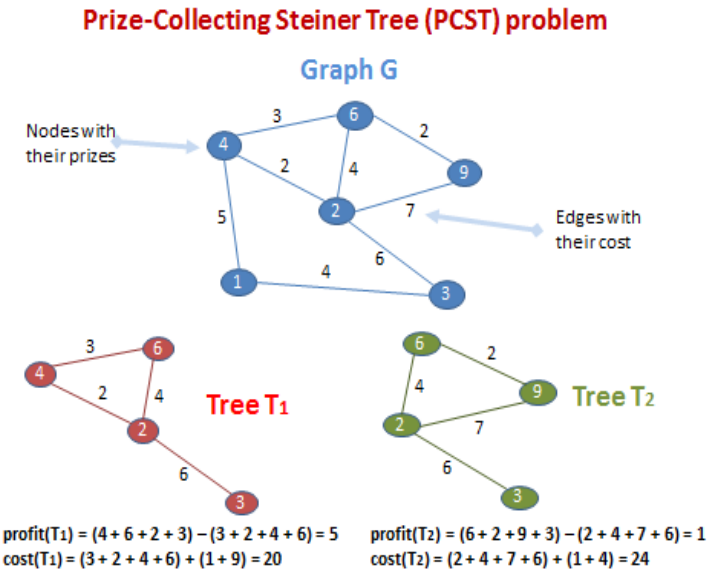


Figure 8. Illustration of PCST problem

Omics Integrator is a software package that applies the Prize-Collecting Steiner Forest (PCSF) approach to construct the most biologically relevant protein-protein interaction network. This tool is efficient in the integration of different omics data using a reference interactome.

There are two distinct tools within the Omics Integrator package: Garnet and Forest. Garnet and Forest complement each other to integrate different experimental results derived from measurements of mRNA, proteins, metabolites, etc. They use omics data either gathered from experiments or derived from several databases (Figure 9).

Garnet tool allows users to estimate transcription factors that possibly affect gene expression levels. It can use epigenetic data such as phosphoproteomic data. Garnet is important while integrating different omics data. This tool includes two steps. First, it computationally predicts transcription factor-DNA interactions from epigenetic data and a set of DNA binding motifs. Then it estimates regulator activities by correlating these predictions with mRNA expression changes in genes neighboring the predicted binding sites.

The forest tool constructs the interaction network using user-defined omic data, the seed protein list containing their importance (prizes), and interactome data including edges' significance levels. Forest combines user-defined hits with the output of Garnet. Each protein given as input to the Forest tool is defined as 'terminal'. If necessary, Forest can add extra nodes from the interactome data, which are called 'Steiner nodes'. When constructing the network, the algorithm optimizes the score by calculating the sum of the costs of edges included and the prizes of nodes not included. The algorithm seeks to minimize the score to find the most optimum and biologically relevant protein-protein interaction network. There is always a possibility to include 'hub' nodes, highly connected nodes, but this can be avoided by applying negative-valued prizes to them. Forest uses a generalized prize function to decide whether these 'hub' nodes are in fact essential and should be included in the network or not. This generalized prize function assigns negative weights to nodes according to the number of connections they have in the interactome. The prize function is:

(8) $$p'(v) = \beta \cdot p(v) - \mu \cdot degree(v)$$

where β and μ are scaling parameters and degree(v) is the number of connections of node v in the interactome. β is used to calibrate the effect of terminal nodes, and μ is used to calibrate the impact of hub nodes. μ is 0 in the default parameters where hub correction is disabled; if it increases, the algorithm excludes these hub nodes. β enables the algorithm to include terminal nodes, and increasing β facilitates more terminal nodes to be included in the final network.

Given a directed, partially directed, or undirected network G(V, E, c(e), p'(v)), it is aimed to find a forest F(V_F, E_F) that minimizes the objective function:

(9) $$f'(F) = \sum_{v \notin V_F} p'(v) + \sum_{e \in E_F} c(e) + \omega \cdot \kappa$$

where $p'(v)$ is defined in Eq.1, $c(e)$ is costs of each edge, κ is the number of trees in the forest, and ω is another scaling parameter which is a uniform edge cost of each node connected to a dummy node (explained below).

The Forest tool has six PCSF parameters; however, ω, β and D are the ones that at least need to be defined by the user. D is the depth parameter which is the maximum path length from v_0 to terminal nodes. The other three optional parameters are μ, g (reinforcement parameter, default is 1e-3), and *garnetBeta* (scales the Garnet output prizes relative to the provided protein prizes, default is 0.01).

Dummy node is an artificial node v_0, a starter node connected to a subset of nodes N. Forest tool builds F to be a tree rooted at v_0. After the network optimization is performed, the root node (v_0) and all its edges are removed. Dummy node by default is set to connect to all of the input terminals. However, one can also define a subset of terminals to link the dummy node by using the --dummyMode option.

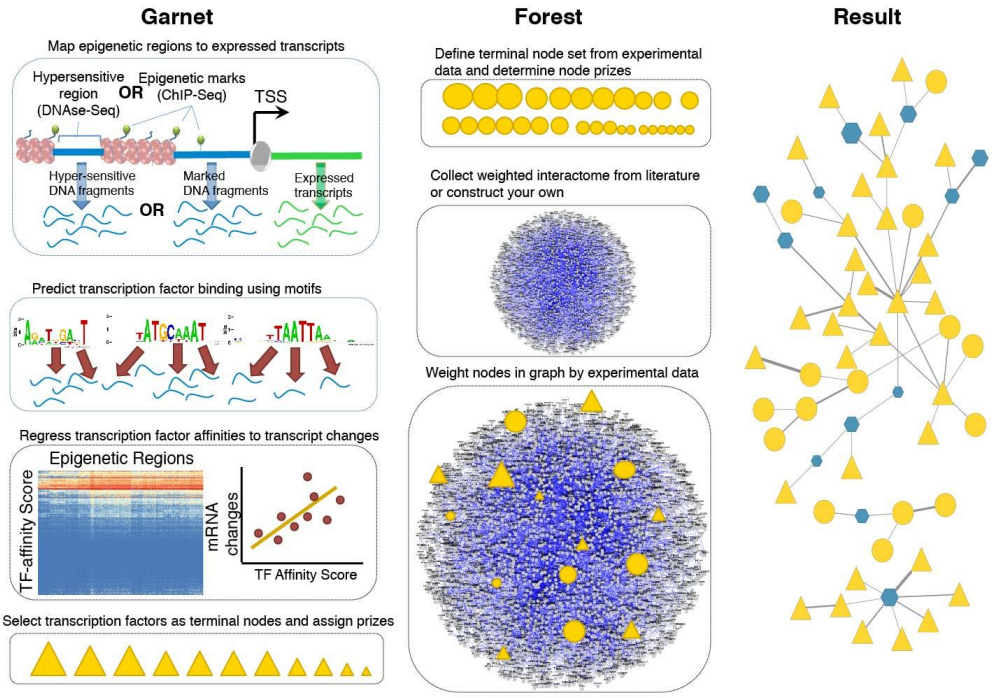


Figure 9. Representation of Forest and Garnet tools found in Omics Integrator [39]

For each cell line and drug condition, the Forest tool is used with the seed proteins and interactome produced by the associated date. The algorithm is run for all combinations of D=10, w=[1,2,3], β=[2,3,4,5,6,7,8,9,10], and μ=[0.0, 0.005, 0.010, 0.015, …., 0.095, 0.1]. After all networks of each parameter combination are collected, the ones with the highest number of terminal nodes and lowest number of hub proteins(degree>100) are selected,

and their union network is defined as the network of the corresponding cell line – drug pair. The network reconstruction protocol is illustrated in Figure 10.

3.4.1. Collection of Seed Proteins

Using the 'significantly transcribed genes' (explained in section 3.2) and transcription factor regulatory network[242], transcription factors regulating any of these 'significantly transcribed genes' are collected. Then, if a transcription factor regulates at least three 'significantly transcribed genes', it is used as a terminal, and mean log2 fold changes of its interactors are designated as its prize.

Also, 'significantly phosphorylated proteins' (explained in section 3.2) are collected, and their associated log2 fold changes are used as their prizes.

Finally, the target(s) of the drug of interest is appended to the seed protein list with a uniform prize inferred from the overall prize distribution of all proteins.

The seed protein list preparation procedure for each cell line–drug condition is performed with its associated data.

3.4.2. Preparation of Condition-Specific Reference Human Interactome

As a primary reference interactome, iRefWeb (version 13.0) is used. It is a human protein-protein interactome with edge weights representing their confidence values based on the MIScore function. After removing self-loops, it consists of 15,404 nodes (proteins) and 175,820 weighted edges.

First, interactome is finetuned for highly connected proteins, hubs. As defined in Hristov & Singh, 2017, proteins that have degrees higher than 900 and more than ten standard deviations away from the mean ('*UBC*', '*APP*', '*ELAVL1*', '*SUMO2*', '*CUL3*')[243] are excluded. These hub proteins constitute 13,738 interactions in the reference interactome. All edges that include at least one of these hub proteins were excluded.

After, the interactome is processed for low expressed genes to have hub-free, tissue and drug-specific versions. For this purpose, transcriptomic data (L1000) is used. L1000 includes fluorescence measurements ranging between 0.0 and 15.0 inferred from direct measurements of landmark genes. iRefWeb interactome and L1000 genes share 11,002 genes in common. Among those common genes, those with expression level below 2.0 and related edges are excluded from the interactome. At this step, hub-free, tissue, and drug-specific human interactomes included ~160,800 interactions.

Next, the link prediction approach is applied. Three case studies are performed for the selection of link prediction algorithm from the node-based topological similarity (local) methods, i.e., Jaccard's index, Adamic/Adar, Preferential Attachment, and Resource Allocation (section 3.4.2.1). For the application of these methods, built-in methods of python networkx module[231] are used.

The selected link prediction algorithm is applied on the hub-free, tissue, and drug-specific interactome. Predictions are sorted descending, and the same number of best scoring predicted edges as the number of edges in the original interactome are selected. These predictions are further filtered based on their subcellular localization information. Localization information was gathered from the Human Protein Atlas[65]. Using this information, predictions in which two proteins that do not have any common location were filtered out. The predictions in which at least one of the proteins do not have any localization information were collected since it is currently unknown and it has a possibility to be found in the same location with other protein and the predictions in which two proteins having at least one common location were also included in the predicted edge list. The final list of predicted edges is appended to the original interactome by scaling their scores between 0.0 and 0.5 as edge weight.

For each cell line–drug condition, the interactome preparation procedure is repeated with its associated data.

3.4.2.1. Case Studies for the Selection of Link Prediction Algorithm

Two different case studies for node-based topological similarity methods (four built-in methods found in networkx module[231] in python: Adamic/Adar, Jaccard's index, Preferential Attachment, Resource Allocation) are performed to select the most suitable link prediction approach for this study. Since the reference iRefWeb (v13.0) interactome is very large, small-sized networks are chosen for each. These two case studies are performed as explained below.

GBM Data Trial. Omics Integrator has been released with several example network data. One of them is the GBM case (one can download from the link: https://github.com/fraenkel-lab/OmicsIntegrator/tree/master/example/GBM). This example data comes with its script to run the Forest tool with a pre-defined parameter set. Using the script and prize file provided with the original iRefWeb (v.13.0) human interactome, its network is reconstructed to be used as a reference network. This reference network has 81 nodes. For the preparation of a new interactome with link prediction application, three protocols are followed for four link prediction methods such that:

- Protocol 1: Four link prediction algorithms are applied, and exactly 100,000 predicted edges from the descending score list of predicted edges for each. Two versions are used for these top 100,000 edges, as '*UBC*' included and '*UBC*' not included.
- Protocol 2: Same procedure as Protocol 1 is performed; however, 175,853 (the number of edges in the original interactome) predicted edges are appended (two versions of '*UBC*' included and '*UBC*' not included).
- Protocol 3: The original interactome is revised such that edges with weights lower than 0.5 are excluded resulting in 39,803 edges remaining. The same number of edges are appended after predicted edges are produced with descending scores (two versions of '*UBC*' included and '*UBC*' not included).

For '*UBC*' not included cases, the number of appended edges decreases by the number of edges including '*UBC*'.

For each trial, predicted edges are appended to the original interactome with constant weights, i.e., 0.1, 0.2, 0.3, …,0.9, 1.0. All interactomes are used separately for new network reconstruction, and they are compared against the reference network.

PI3K/AKT/mTOR Data Trial. In this trial, the performances of link prediction methods are measured by dividing edges randomly into two parts as 90% training and 10% test sets. PI3K/AKT/mTOR pathway network has 1578 edges in total, making 1420 edges in the training set and 158 edges in the test set. For this purpose, the below steps are performed for each link prediction method:

- All edges are assigned to random probability values differing in each trial.
- According to these probabilities, edges are sorted in decreasing order to separate the training and test sets by 90%.
- The first 90% of the edges are selected to be the training set, while the remaining are the test set.
- All nodes are extracted from both training and test sets.
- LP algorithm is used to score the possible edges, and predictions are sorted in decreasing order.
- From the sorted list of predictions, the first 158 artificial edges (since 158 is the length of the test set) are selected.
- The intersection of the highest scoring 158 predictions and the test set found, and the length of intersecting list is recorded.
- The protocol is repeated ten times.

3.5. Robustness Test of Reconstructed Networks

In order to understand the robustness of reconstructed networks, a kind of noisy edge application to the interactome is performed, and each network is tested with the same parameter set chosen in the initial network reconstruction step.

For a given cell line – drug condition, its own interactome is processed such that the weights of randomly chosen 10% of the edges are increased by 10^{-3}, and a network is reconstructed by use of the noisy interactome and original terminal set with the parameters used in initial network reconstruction step. This random noise addition is performed 100 times for each parameter set. For example, given one cell line–drug network is reconstructed with merging six parameter sets, noise addition to the weights of random 10% of the edges will be performed for 6 x 100 = 600 times, resulting in 600 networks. By using edges found in these 600 networks, the frequency of each edge is calculated and used to assess the reliabilities of the edges in the original network. If the algorithm mostly

selects an interaction despite the noisy edges; that is, the edge frequency is closer to 1, then that interaction can be considered reliable.

3.6. Comparison of Reconstructed Networks

For comparison purposes, two approaches are used: topology-based and pathway-based.

3.6.1. Topology-based Comparison

A pairwise network similarity method called separation score proposed by Menche et al. 2015 is used in this study. The separation score is based on mapping the network on the reference interactome and scoring their overlap by measuring the mean shortest distances[244]. The formula is:

(10) $$s_{AB} \equiv \langle d_{AB} \rangle - \frac{\langle d_{AA} \rangle + \langle d_{BB} \rangle}{2}$$

where s_{AB} is the separation score of the networks A and B, d_{AB} is the shortest distances between A-B proteins, and d_{AA} and d_{BB} are the shortest distances between proteins within A and B, respectively. For the calculation of d_{AB}, if a protein is shared between both networks (A and B), the distance d_{AB} of that protein is 0. The scoring method is also illustrated in Figure 10.

For comparison purposes, all pairwise network separation scores are calculated. Several separation score matrices are prepared for all tumor types, drugs across different cell lines, and cell line-drug networks. All matrices are subjected to hierarchical clustering using python scipy.cluster.hierarchy function of SciPy module[237].

3.6.2. Pathway-based Comparison

Functional analysis for network nodes based on KEGG pathways is performed to compare reconstructed networks on pathway-level for each network using DAVID source code[245,246]. After collecting union of all significant pathways ($p<.05$) enriched in networks, cell line–drug pair versus KEGG term matrix was produced such that it contains values of either -log10(p-value) of the KEGG term if that network is enriched in or 0.0 if that network is not enriched in. NMF Consensus clustering was then applied on the KEGG term matrix using NMFConsensus module on GenePattern of Broad Institute[247]. The clustering was computed for $k = 2$ to $k = 10$, and the cophenetic coefficient was considered to decide the best number of k in the clustering process.

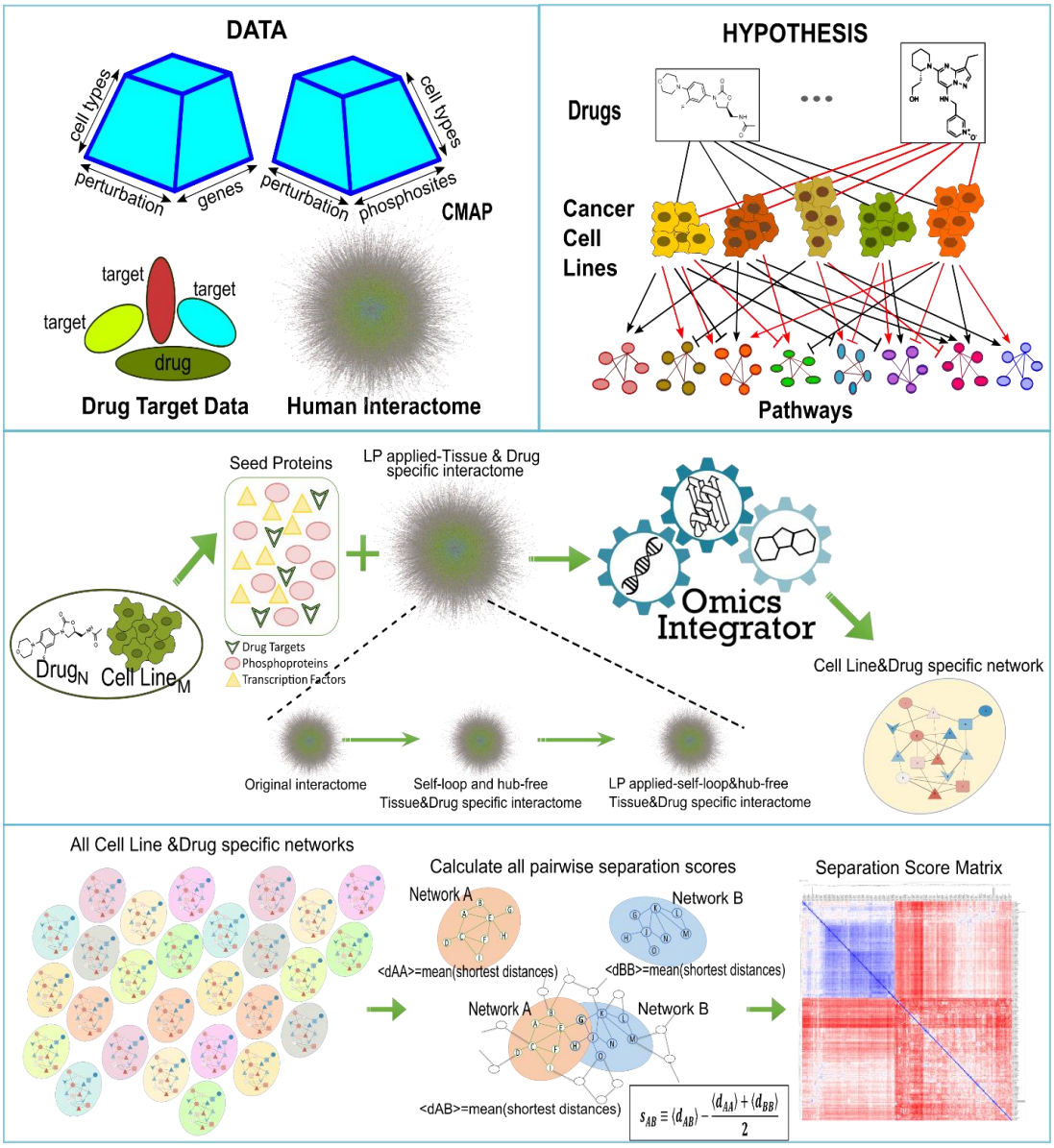


Figure 10. General overview of the methods used in the network reconstruction and network comparison parts of this study

3.7. Data and Code Availability

All datasets used in this work are publicly available from the following sources: The CMap data was downloaded using accession code GSE101406. Installation details and documentation for Omics Integrator software can be found on this link: https://github.com/fraenkel-lab/OmicsIntegrator. The PPI network used in network

reconstruction is downloaded from https://github.com/fraenkel-lab/OmicsIntegrator/tree/master/data (IRefIndex Version 13.0) and drug–target data is curated from Cmap Drug Repurposing Tool (https://clue.io/repurposing-app). The script for parameter tuning performed for network reconstruction can be downloaded from https://github.com/gungorbudak/forest-tuner. Separation score method is adapted from the source code provided in the supplementary material of the study of Menche et al,2015[244] (http://science.sciencemag.org/content/suppl/2015/02/18/347.6224.1257601.DC1). All data and code used to conduct this study can be found from the GitHub repository on this link: https://github.com/metunetlab/IntegrativeNetworkModeling

CHAPTER 4

RESULTS

This chapter briefly presents the results acquired through this study. These results are divided into three main categories: (i) classification of drugs based on prior knowledge, i.e., chemical structures and mechanism of action, (ii) selection of link prediction method that is later used in network reconstruction procedure, and (iii) analysis of networks to reveal the drug modulations.

4.1. Classification of Drugs based on Prior Knowledge

There are 70 drugs that are statistically significant. They are all small molecules targeting proteins. The list of these drugs is given in Table 4 below. These 70 drugs are clustered based on their chemical structures using three metrics: (i) Tanimoto similarities, (ii) MACCS key distances, and (iii) Mechanism of Action (MoA).

4.1.1. Tanimoto Similarity-based Classification

Pairwise Tanimoto similarity scores calculated with SMILES signatures of drugs are transformed to a matrix and clustered to observe the chemically similar drug groups. This analysis revealed that 70 drugs could not be efficiently classified by using the Tanimoto similarity metric. There are small groups of drugs containing 2-3 members. For example, decitabine and zebularine constitute one drug group, while everolimus, sirolimus, and tacrolimus constitute another. The chemical structures of BIX-01294 and UNC-0321 are also similar. The heatmap of pairwise Tanimoto similarities of 70 drugs is given in Figure 11a.

Tablo 4. Drugs that are statistically significant

4,5,6,7-tetrabromobenzotriazole	everolimus	niclosamide	semagacestat
alpelisib	EX-527	nilotinib	sirolimus
alvocidib	exifone	olaparib	sotrastaurin
AR-A014418	geldanamycin	OSI-027	staurosporine
belinostat	ginkgetin	palbociclib	tacedinaline
BIX-01294	GSK-J4	pazopanib	tacrolimus
BIX-01338	I-BET-151	PD-0325901	TG-101348
BMS-345541	I-BET-762	pravastatin	tofacitinib
CC-401	IKK-inhibitor-X	pyrazolanthrone	tretinoin
CHIR-99021	IPI-145	resveratrol	trichostatin-a
curcumin	JQ1-(+)	RG-4733	UNC-0321
decitabine	KN-62	RGFP-966	UNC-1215
dexamethasone	KN-93	rolipram	VE-822
dinaciclib	KU-55933	roscovitine	vemurafenib
entinostat	lenalidomide	ruxolitinib	verteporfin
epz004777	losmapimod	salermide	vorinostat
EPZ-005687	LY-294002	selumetinib	zebularine
etoposide	methylstat		

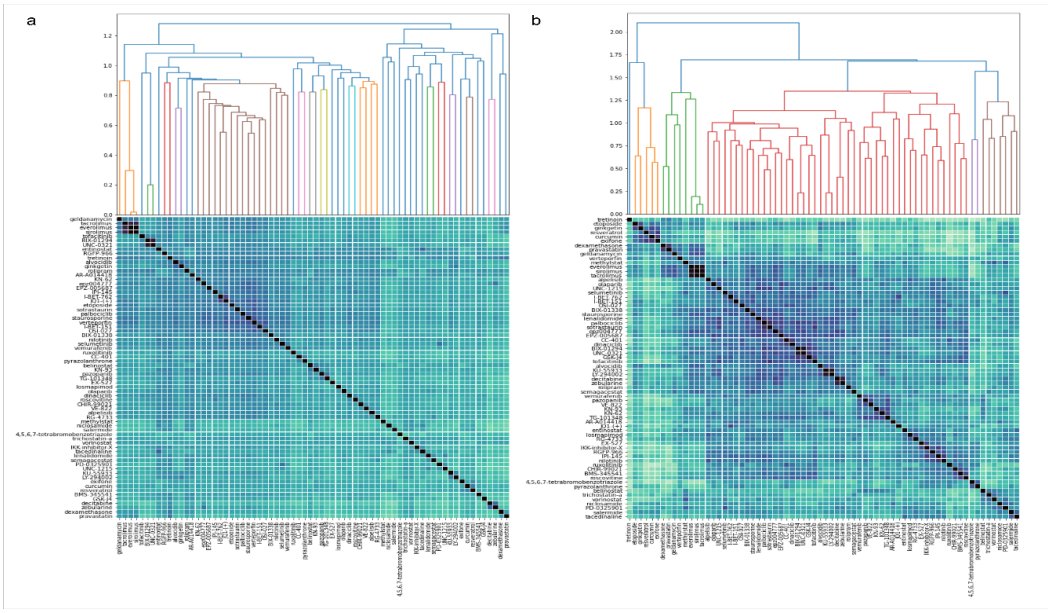


Figure 11. Chemical similarity matrices of 70 drugs **a)** Heatmap of pairwise Tanimoto similarities **b)** Heatmap of pairwise MACCS key distances (Larger figures can be found in Appendix E-Figure 1&2)

4.1.2. MACCS key distance-based Classification

MACCS key distances calculated by 166-bit 2D structural fingerprints are transformed to a matrix and clustered as in Tanimoto similarities. MACCS keys provided a more clear clustering pattern such that 70 drugs are classified into five main groups, one of which only contains Tretinoin. The other drug group includes 47 out of 70 drugs, while the numbers of drugs included in the remaining three groups are 5, 8, and 9. The compound groups in size two to three from Tanimoto similarities are also present together in the MACCS key distance groups. The heatmap of pairwise MACCS key distances of 70 drugs is given in Figure 11b, and the drugs in each group are listed in Table 5.

MACCS key-based classification performed better in the classification of the drug analogs. In this study, open-source MACCS key fingerprints are used, noting that the larger dataset may have more capacity to classify drugs.

However, it should be noted that many drug analogs are developed based on the biological activity of the drug of interest. That's why inactive forms of drugs may present chemically dissimilar structures while they change conformation in their active forms and show similarity. The other important aspect of drug development is water or lipid solubility. Some drugs are developed to preserve desired solubility patterns despite keeping the same pharmacophore, and modifications made for solubility concerns may affect the chemical similarities between drugs. Therefore, considering only chemical structures for the classification of drugs would be insufficient and not reflect the biological relevance or similarity in their mechanisms of action.

Tablo 5. Drug groups produced by MACCS keys distances

Group1	Group2	Group3			Group4	Group5
Etoposide	Tretinoin	CC-401	Lenalidomide	Losmapimod	4,5,6,7-tetrabromobenzotriazole	Dexamethasone
Ginkgetin		Dinaciclib	Palbociclib	RG-4733	Pyrazolanthrone	Pravastatin
Curcumin		BIX-01294	Decitabine	EX-527	Belinostat	Geldanamycin
Exifone		BIX-01338	Zebularine	IKK-inhibitor-X	Trichostatin-a	Verteporfin
Resveratrol		I-BET-151	Alpelisib	IPI-145	Vorinostat	Methylstat
		I-BET-762	Olaparib	Nilotinib	Niclosamide	Everolimus
		UNC-0321	Alvocidib	RGFP-966	PD-0325901	Sirolimus
		UNC-1215	KU-55993	AR-A014418	Salermide	Tacrolimus
		GSK-J4	LY-294002	JQ1	Tacedinaline	
		Tofacitinib	Rolipram	KN-62		
		OSI-027	Semagacestat	KN-93		
		Selumetinib	CHIR-99021	TG-101348		
		Staurosporine	BMS-345541	Pazopanib		
		Sotrastaurin	Roscovitine	VE-822		
		epz004777	Ruxolitinib	Vemurafenib		
		EPZ-005687	Entinostat			

4.1.3. MoA-based Classification

MoAs of 70 drugs are curated to understand whether knowledge of MoA is sufficient to classify drugs efficiently and whether the MoA clusters are consistent with the omic data. Seventy drugs are classified into three main groups: kinase inhibitors, epigenetic modulators, and variable MoA categories named others. These three groups are named as main category classes, and their specific sub-groups are designated as sub-categories. For example, the main category of kinase inhibitors has three sub-categories: serine/threonine kinases, tyrosine kinases, dual-specificity kinases, and other type kinases. An abstract view on 70 compounds in our analysis is illustrated in Figure 12 based on their main- and sub-categories and MoA.

Drugs are distributed in a balanced manner to all MoAs such that each MoA has 1-3 drugs. However, HDAC inhibitors, histone lysine methylase inhibitors, CDK inhibitors have more drugs, while MoAs in the other category usually have only one drug. When compared to MACCS clusters, the MoA of drugs is variable. For example, everolimus and tacrolimus have different MoAs, mTOR inhibitor and calcineurin inhibitor, respectively. However, they are included in the same MACCS cluster with dexamethasone, verteporfin, pravastatin, and geldanamycin. Moreover, HDAC inhibitors are separated into two different clusters based on their MACCS key distances.

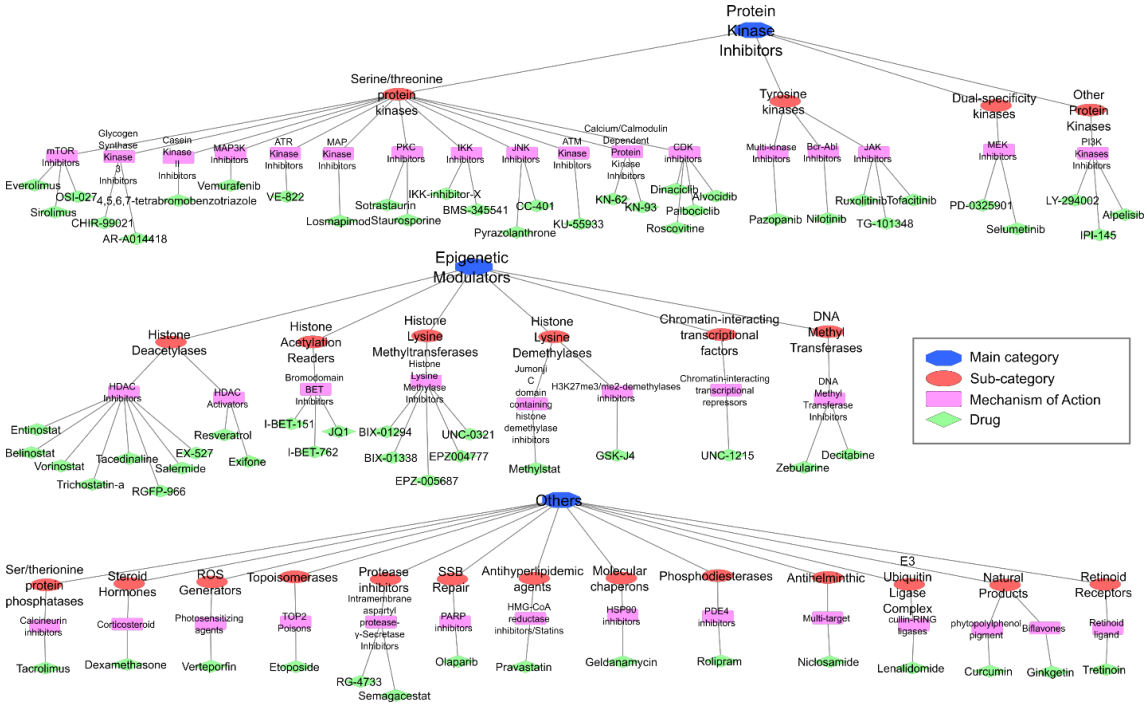


Figure 12. Drug groups classified based on mechanisms of action. (larger figure can be found in Appendix E-Figure 3)

4.2. Selection of Link Prediction Method

Three different tests with different network data are performed to select the most suitable node-based link prediction algorithm.

GBM data trial. Three interactome preparation procedure was used such that different numbers of predicted (artificial) edges were appended to the original iRefWeb interactome with two options ('*UBC'* included and '*UBC'* not included). Eighty networks are reconstructed (4 link prediction (LP) methods, ten different weights for predicted edges, and two w parameter values provided in the original example file) for each procedure. Each network is compared against its associated reference network reconstructed with the same parameters and the original interactome. The nodes and the number of predicted and original edges found in the networks are used for comparison purposes. Also, commonalities of predicted edges found by four LP algorithms are compared using Venn diagrams.

Table 6 summarizes the number of predicted/artificial edges and real edges found in the reconstructed networks. Weights column refers to the edge weights used for predicted edges; parameter column refers to the w parameter used in the network reconstruction step (other parameters are defined constant in the given GBM example); AA is Adamic/Adar method; JC is Jaccard's Coefficient; PA is Preferential Attachment, and RA is Resource Allocation. According to their equations, AA and RA are very similar as the equation of RA is a version of AA such that logarithm is not used in the calculation. The number of predicted edges found in the final network is higher in RA when compared to AA. PA method adds more predicted edges, however by its definition, it may encourage hub formation. The JC method surprisingly does not prefer to use predicted edges. According to the results shown in Table 6, AA and PA methods were more applicable.

Afterward, the commonalities of predicted/artificial edges found in each protocol ('*UBC'* not included) for three LP methods are investigated. In Figure 13, the number of edges appended with each method represents the number in the protocol minus the number of edges, including '*UBC'*. From the difference between AA edges in protocol 1 and protocol 2, it can be concluded that the top-scored 100,000 edges do not include '*UBC*' while the remaining 75,853 edges constituting the additional edges protocol 2 includes 70,335 edges with '*UBC*'. Based on the results shown in the Venn diagrams, the number of common edges is proportionally similar to each other, which means that the user can define the number of predicted edges to append to the original interactome.

Tablo 6. The number of artificial and original edges found in the reconstructed networks for each LP algorithm using protocol 2 in which the same number of artificial edges are appended to the original interactome.

Weights	Parameter	AA		JC		PA		RA	
		Artificial	Real	Artificial	Real	Artificial	Real	Artificial	Real
0.1	w=2	0	67	0	76	1	62	0	65
	w=3	0	67	0	76	1	62	0	66
0.2	w=2	0	68	0	76	3	60	2	63
	w=3	0	67	0	76	3	60	3	61
0.3	w=2	0	65	0	76	0	62	0	61
	w=3	0	69	0	76	0	61	0	61
0.4	w=2	3	61	0	76	6	56	6	57
	w=3	3	63	0	76	9	53	5	57
0.5	w=2	9	57	0	76	16	48	26	45
	w=3	10	55	0	78	16	48	20	41
0.6	w=2	17	52	0	76	21	40	29	33
	w=3	21	45	0	76	20	41	29	33
0.7	w=2	0	69	0	77	0	62	0	64
	w=3	0	71	0	76	0	62	0	64
0.8	w=2	34	38	0	76	27	36	48	23
	w=3	33	37	0	76	27	36	48	22
0.9	w=2	41	31	0	76	30	37	62	15
	w=3	40	32	0	75	30	37	63	15
1	w=2	53	22	2	74	37	31	88	6
	w=3	59	22	1	75	37	31	84	5

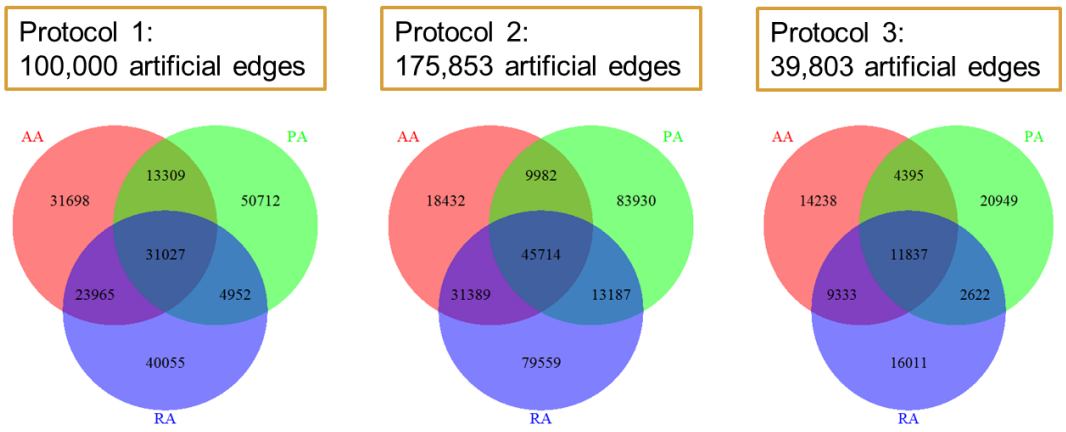


Figure 13. Commonalities of appended edges found with three LP methods as 'UBC' excluded.

PI3K/AKT/mTOR Data Trial. In this trial, recovery rates of LP methods were tested using 90% of the network edges as the training set and the remaining 10% of the network

edges as the test set. Out of the 158 top-scoring predicted edges, the numbers of the true positives are given in Table 7. Average values remain around seven even though the trial number is increased. From these results, none of the three methods, AA, PA, and RA, are superior, while JC could not predict correctly.

Tablo 7. Number of correct predictions in 10 trials with 4 LP methods

(Total number of edges in the network: 1578, Number of edges in the training set: 1420, Number of edges in the test set: 158, Number of predictions with highest scores: 158.)

#TestSet: 158	**LP Methods**			
Trials	**Adamic/Adar**	**Jaccard's Coefficient**	**Preferential Attachment**	**Resource Allocation**
1	7	0	3	7
2	10	0	10	9
3	10	0	12	10
4	8	0	4	7
5	8	0	7	9
6	7	0	6	7
7	4	0	8	3
8	2	0	5	3
9	7	0	7	8
10	8	0	8	7
Average	7,1	0	7	7

Based on the results of these two case studies Adamic/Adar link prediction method is selected. While tests could not reveal a clear distinction between three methods, AA, PA, and RA, by its definition, AA is more suitable for this study. AA is known for its ability to weigh rarer features more heavily and exclude hubs. The scoring formula of Adamic/Adar is given in Eq.2. Since the prediction capacity is no worse than others, the study is carried out using AA in the interactome preparation protocol.

4.3. Analysis of Reconstructed Network

The aim of this study is to reveal network-level commonalities and differences of drugs across several tumor types and demonstrate the advantages of network-based studies over the listed traditional chemical and functional comparisons. Data, network reconstruction, and topological comparison protocols used in this study are represented in Figure 10 (see in section 3.6).

Although data includes 89 drugs and five cell lines, statistical analyses resulted in 70 significantly affecting drugs. Also, the number of proteins in the seed protein lists (section 3.4.1) is very limited for some cell line–drug conditions leading to insufficient input data for network modeling. In order to increase experimental hits, less stringent thresholds are tested; however, still networks of some cell line–drug conditions could not be modeled. A549-decitabine, YAPC-ginkgetin and MCF7-vemurafenib are examples that do not have a reconstructed network. Experimental hits (seed proteins) ranged from 6 to 214 for each cell line–drug condition, and within these seed proteins, the number of phosphoproteins ($p<.05$) was limited with a maximum of 20 hits. All drugs were forced to reconstruct their associated networks while modeling networks for the A375 cell line. However, it is observed that reconstructed networks are very small-sized and do not reveal sufficient knowledge about drug modulation when the input terminal number is below ~15. Thresholds used to collect transcriptomic hits and the total number of seed proteins are provided in Appendix A. Figure 14 summarizes the conditions that do or do not have reconstructed networks. Overall, we constructed 250 subnetworks for each pair of 70 drugs and six cell line combinations. However, not all cell lines have the same number of subnetworks. Out of 250 networks, cell line A375 has 70, A549 has 46, MCF7 has 43, PC3 has 59, YAPC has 18, and NPC has 14 drug-specific subnetworks. A summary of the topological properties of these 250 networks is given in Appendix B.

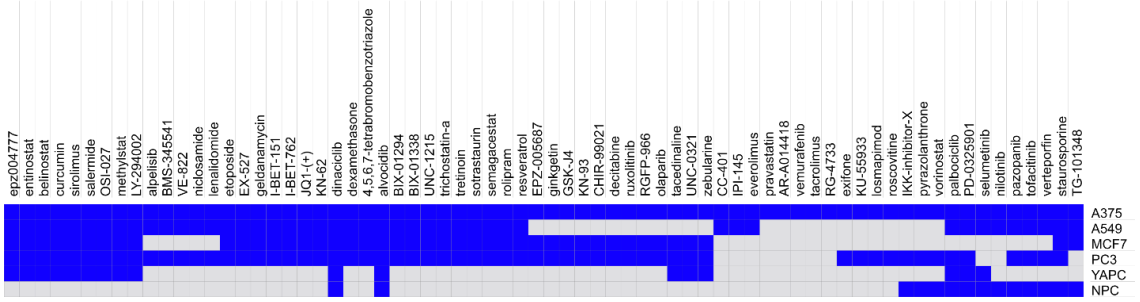


Figure 14. Heatmap for drugs if its PPI network could be modeled for each cell line. Blue means that the drug has network information for that cell line.

4.3.1. Analysis of most frequently found 100 proteins observed in all cell line-drug pair networks

The initial analysis performed for networks is a general node-based commonality investigation. 98.4% of network pairs share at least one protein, and some of the proteins are found in a high number of networks. Out of 1908 proteins across all 250 networks, the most frequent 100 proteins (Figure 15) are used for functional annotation. The most frequently present 100 proteins are enriched in MAPK signaling pathway, AMPK signaling pathway, mTOR signaling pathway, PPAR signaling pathway, insulin signaling pathway, cell cycle, and some other cancer-related pathways. These results are consistent with the known active signaling pathways in cancer cells. For example, the PPAR signaling pathway is related to fatty acid metabolism. Cancer cells need to produce more energy, so the fatty acid degradation process is activated. Insulin signaling is also related to the energy needs of the cells. Additionally, we observed that transcription regulation, DNA-binding, and protein heterodimerization activity are mostly active functions. Molecular function, biological process, cellular compartment, and KEGG pathway enrichments of these most frequent 100 proteins can be seen in Figure 16.

4.3.2. Analysis of networks based on separation score method that uses node similarity

Since the analysis of networks solely based on total node frequencies provides only a weak explanation of cancer-related activities, pairwise network similarities are searched via the separation score method proposed by Menche et al.[244] (section 3.6.2). If the separation score is negative, then two networks share a common set of nodes, meaning they overlap. The magnitude of the negative score reflects the overlap ratio, which means the closer to -1, the larger the overlap between two networks. Topological similarity also leads to pathway-level similarity, which lets to understand the similarities of drug modulations.

Transforming the networks into a matrix of separation scores provides a better understanding of both pairwise similarities between drugs across different tumor types. After calculating all pairwise separation scores, matrices for all 250 networks and cell line-specific networks are produced, then hierarchically clustered to easily detect similarly acting cell line–drug conditions (Figures 17 and 18). Four types, T1, T2, T3 and T4, are defined to illustrate the tumor types and MoA types. T1 refers to the same cell line and same MoA, T2 refers to the same cell line and different MoA, T3 refers to different cell lines and the same MoA, T4 refers to both different cell lines and MoA.

Cell line-specific separation score matrices (Figure 17) revealed that some drugs have a high overlap for the cell lines A375, MCF7, and PC3, although the chemical similarities of those drugs are limited. Drugs in A549 generally modulate different mechanisms within the cell as the networks are separated with positive separation scores. That means 46 drugs have more specific actions on the A549 cell line even though they have similar MoAs or chemical structures. The number of drug networks in YAPC and NPC is very limited; that's why the investigation of drug effects on these cell lines is not feasible. The list of

drugs with topologically similar networks according to the separation score matrix of 250 networks is given in Table 8. As A375 has the most coverage in terms of the number of drug networks, analysis for drug modulation within the same tumor type is mainly performed for A375. For example, Tacedinaline and geldanamycin are two drugs that have different targets. However, they have very similar omic output (separation score=-0.61), leading to several common signaling pathways modulated.

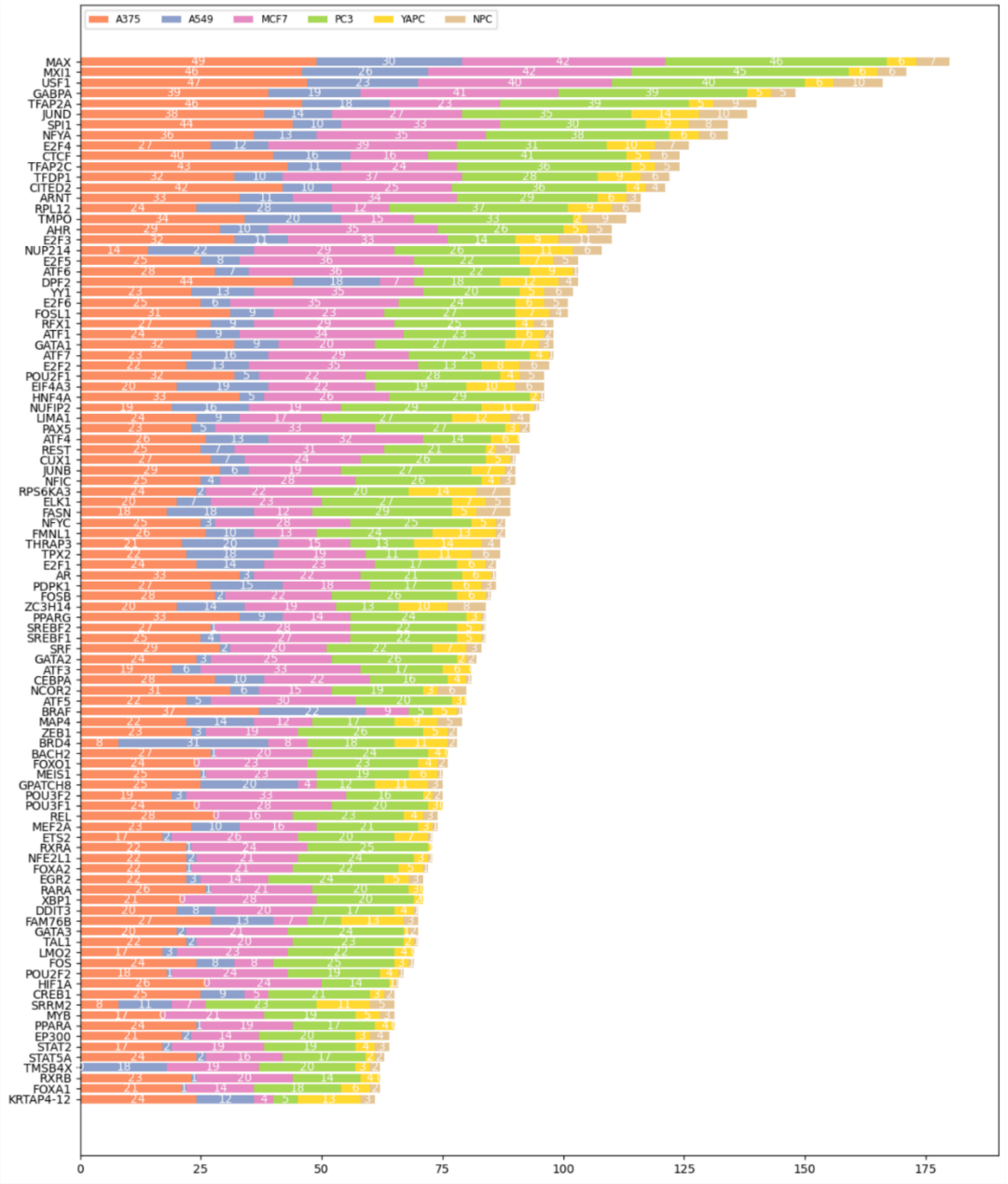


Figure 15. Mostly found 100 proteins observed in all cell line–drug networks.

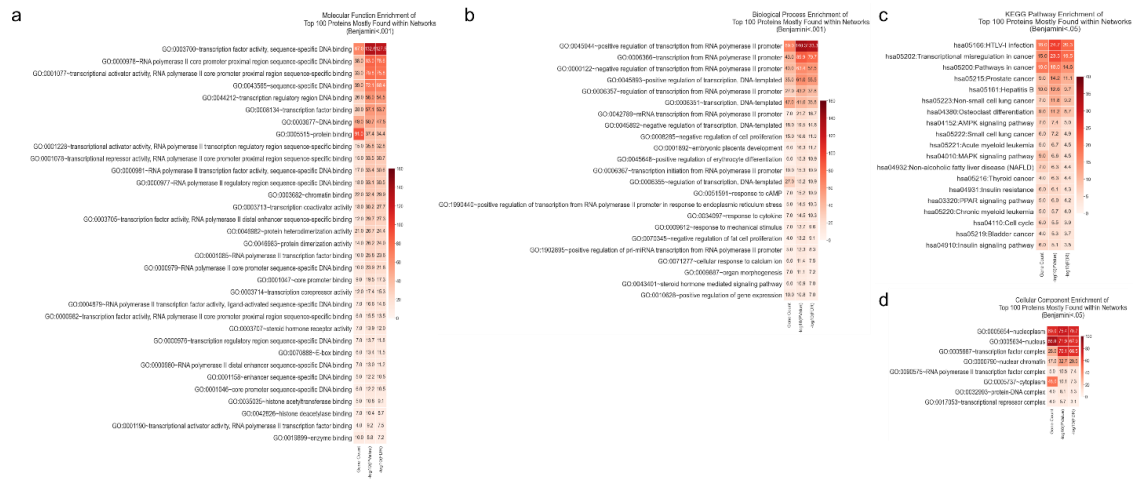


Figure 16. Functional analysis of mostly found 100 proteins **a)** Molecular function enrichments **b)** Biological process enrichments **c)** KEGG pathway enrichments **d)** Cellular compartment enrichments (Larger figures can be found in Appendix E-Figure 4-7)

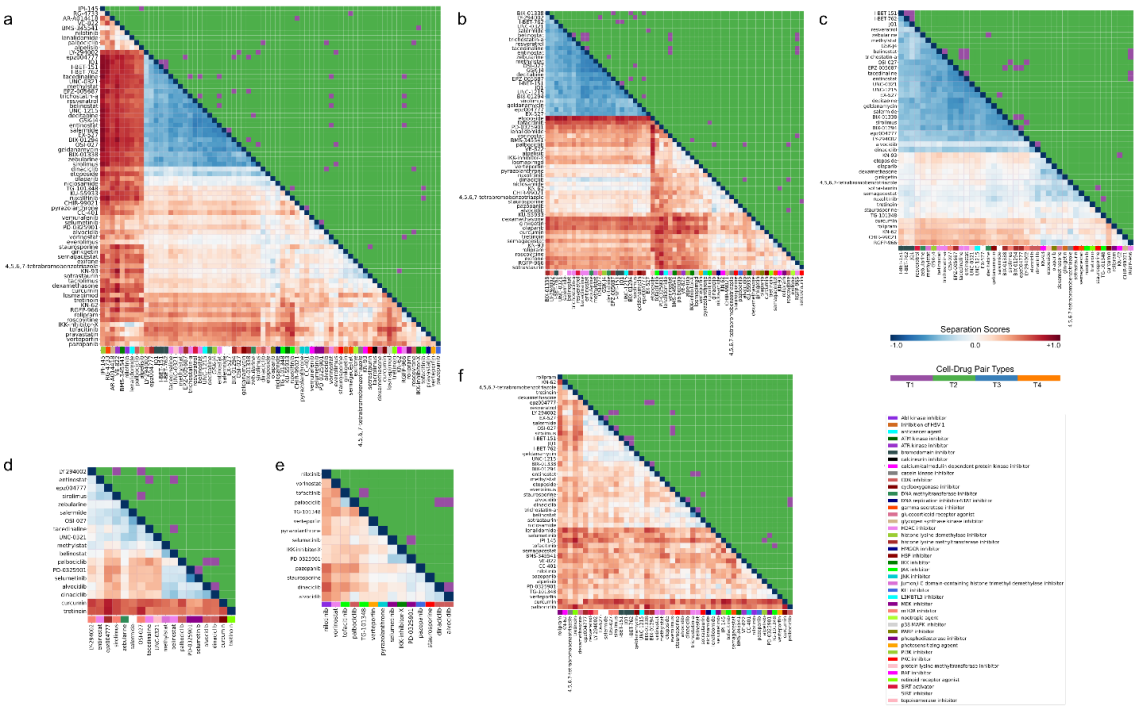


Figure 17. Separation score matrices of 6 cell lines **a)** A375 **b)** PC3 **c)** MCF7 **d)** YAPC **e)** NPC **f)** A549. (Larger figures can be found in Appendix E-Figure 8-14)

Figure 18. Separation score matrix of 250 cell line – drug network

Tacedinaline is an HDAC inhibitor that selectively targets HDAC1, while geldanamycin is a heat-shock protein (HSP) inhibitor that targets HSP90AA1 and HSP90AA1. HDACs and HSPs are shown to have an interplay within cellular processes to be affected by both HDAC and HSP90 inhibitor drugs[248–254]. With overlapping networks of tacedinaline and geldanamycin, common proteins are analyzed for the cellular phenotypes with the aid of CancerGeneNet[255]. CancerGeneNet measures the shortest paths between genes and phenotypes. Proteins shared by tacedinaline and geldanamycin networks on A375 are downstream transcription factors of both HDAC1 and HSP90s, which are on the path toward cell death and cell differentiation pathways (Figure 19A). Pathways commonly enriched in these networks are MAPK, AMPK, TG-beta signaling pathways and their transcriptomic data shows a highly positive correlation (Pearsons R=.94, p-value=0.0). Thus, these two drugs with different MoA cause several common modulations within the A375 cell line. At the same time, Tacedinaline differs from geldanamycin with chemokine, neurotrophin, ErbB, TNF, FoxO, PPAR, prolactin, and thyroid hormone signaling pathways (Figure 19B).

Tablo 8. Drugs that have topologically similar networks

A375	**MCF7**	**PC3**	**Other cells**
	MCF7_alvocidib		YAPC_alvocidib
A375_belinostat	MCF7_belinostat	PC3_belinostat	YAPC_belinostat
A375_BIX-01294	MCF7_BIX-01294	PC3_BIX-01294	YAPC_dinaciclib
A375_BIX-01338	MCF7_BIX-01338	PC3_BIX-01338	NPC_dinaciclib
A375_decitabine	MCF7_decitabine	PC3_decitabine	YAPC_methylstat
A375_etoposide	MCF7_dinaciclib		YAPC_PD-0325901
A375_entinostat	MCF7_entinostat	PC3_entinostat	YAPC_selumetinib
A375_epz004777	MCF7_epz004777	PC3_epz004777	A549_staurosporine
A375_EPZ-005687	MCF7_EPZ-005687	PC3_EPZ-005687	
A375_EX-527	MCF7_EX-527	PC3_EX-527	
A375_geldanamycin	MCF7_geldanamycin	PC3_geldanamycin	
A375_GSK-J4	MCF7_GSK-J4	PC3_GSK-J4	
A375_I-BET-151	MCF7_I-BET-151	PC3_I-BET-151	
A375_I-BET-762	MCF7_I-BET-762	PC3_I-BET-762	
A375_JQ1	MCF7_JQ1	PC3_JQ1	
A375_LY-294002	MCF7_LY-294002	PC3_LY-294002	
A375_methylstat	MCF7_methylstat	PC3_methylstat	
A375_OSI-027	MCF7_OSI-027	PC3_OSI-027	
A375_resveratrol	MCF7_resveratrol	PC3_resveratrol	
A375_salermide	MCF7_salermide	PC3_salermide	
A375_sirolimus	MCF7_sirolimus	PC3_sirolimus	
A375_tacedinaline	MCF7_tacedinaline	PC3_tacedinaline	
A375_trichostatin-a	MCF7_trichostatin-a	PC3_trichostatin-a	
A375_UNC-0321	MCF7_UNC-0321	PC3_UNC-0321	
A375_UNC-1215	MCF7_UNC-1215	PC3_UNC-1215	
A375_zebularine	MCF7_zebularine	PC3_zebularine	

The network-level overlap between drug pairs in the same cell line still exists across different cell lines, but their overlap ratio is relatively lower. Resveratrol is another example of drugs having overlapping networks with geldanamycin (in A375, separation score=-0.60; in MCF7, separation score=-0.47; in PC3, separation score=-0.54). However, the separation between resveratrol and geldanamycin across different cell lines gets higher such that the separation score between A375-resveratrol and MCF7-geldanamycin is -0.40 separation score between A375-resveratrol and PC3-geldanamycin is -0.43.

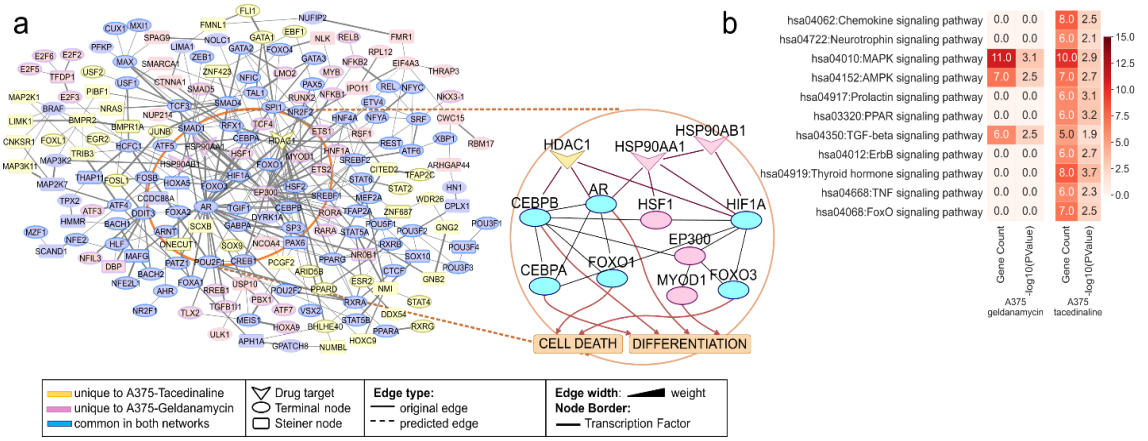


Figure 19. Illustration of the comparison of A375-Tacedinaline and A375-Geldanamycin **a)** Representation of the union of two networks with a focus on common transcription factors leading to cellular phenotypes **b)** Pathways enriched in two networks.

Moreover, drugs with the same MoAs may have different outcomes. Target selectivity is an important aspect leading to separate networks within the same MoA group. That's why the comparison strategy for the remaining conditions turned out to the analysis of MoA groups both within the same cell lines and across all cell lines. The selected MoAs are below explored for comparison of drugs both within the individual cell lines and across all cell lines.

4.3.2.1. Analysis of Mechanism of Action: HDAC inhibitors

There are six HDAC inhibitor (HDACi) drugs: entinostat, belinostat, vorinostat, trichostatin-a, tacedinaline, and RGFP-966. Within these, tacedinaline is a selective HDAC1 inhibitor, and RGFP-966 is a selective HDAC3 inhibitor, while others are broad HDACi drugs. Network separation scores between these drugs are predominantly determined by target selectivity. Networks of broad HDACi drugs have high overlap both within the same cell line and between different cell lines. Tacedinaline also has similar networks with others in A375, MCF7, and PC3 cell lines.

Tacedinaline treatment resulted in similar modulated mechanisms on A375, MCF7, and PC3, but the YAPC-Tacedinaline network is more distant to others with a separation score of 0.38 across A375. Tacedinaline treatment on YAPC cells affected a small subnetwork that is only enriched for the Osteoclast differentiation pathway. Recently, osteoclast differentiation has been shown to be induced in pancreatic cancer-derived exosomes[256]. On the contrary, the A375-Tacedinaline network is enriched in several pathways such as MAPK, ErbB, FoxO, TGF-beta, and TNF signaling pathways (Figure20).

RGFP-966 is a slow-on/slow-off, competitive tight-binding HDAC3 inhibitor with an IC50 of 0.08 μM and shows no effective inhibition of any other HDAC at concentrations

up to 15μM[257]. It is reported to inhibit the EGFR signaling pathway and suppress proliferation and migration of HCC cells[258]. It has networks in A375, MCF7, and PC3 cell lines. All three networks are separated with a very small number of shared proteins (separation score between A375-MCF7 is 0.45, A375-PC3 is 0.28, and MCF7-PC3 is 0.42). RGFP-966 treatment also results in different omic outputs than the other HDACi drugs. When it is pairwisely compared to broad HDACi drugs, separation scores vary around 0.4. The most distant pairs usually belong to A375 and PC3 cell lines. As an example, PC3-RGFP-966 and A375-Trichostatin-a networks have a separation score of 0.46. They share no common pathway enrichments. PC3-RGFP-966 is only enriched in the cell cycle pathway, while A375-Trichostatin-a is enriched in several pathways such as estrogen, neurotrophin, and TNF signaling pathways (Figure 21A-B). When Tacedinaline and RGFP-966 networks are compared as two selective drugs, more distant networks are obtained with separation scores up to 0.62. PC3-RGFP-966 and A375-Tacedinaline networks only share 3 proteins and there are no commonly enriched pathways (Figure 21C-D).

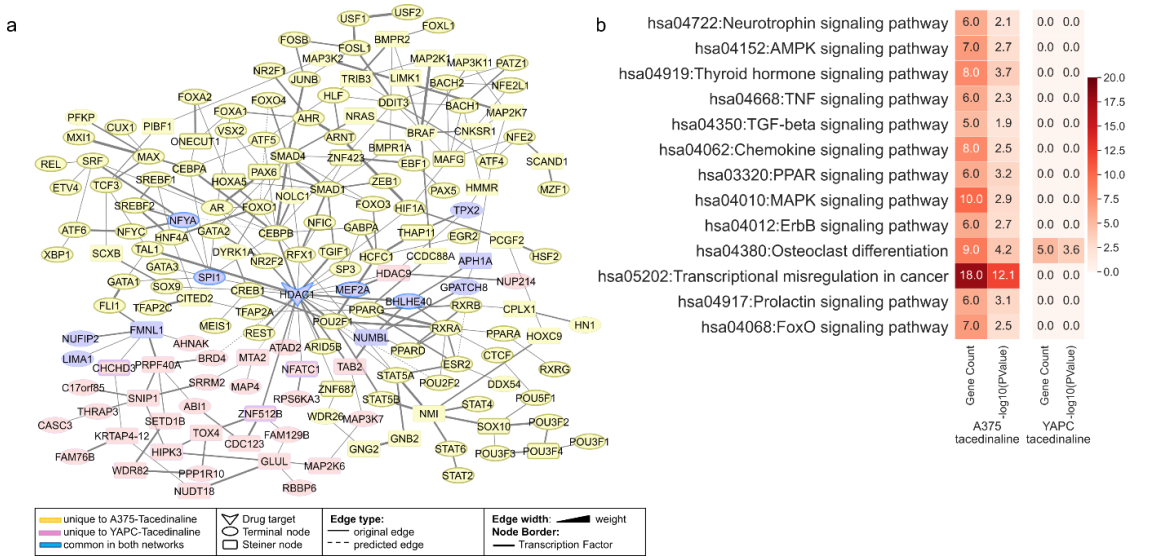


Figure 20. Merged network of Tacedinaline networks in A375 and YAPC and their associated pathway enrichments

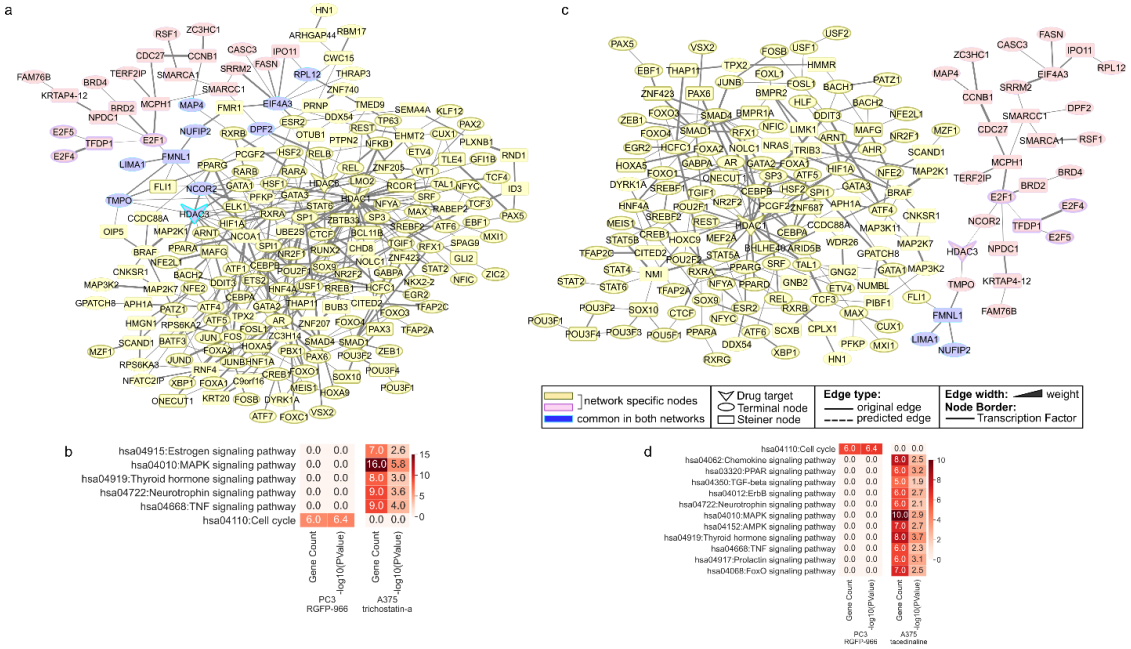


Figure 21. Merged network of two cell line – drug conditions and their associated pathway enrichments **a-b)** PC3-RGFP-966 and A275-Trichostatin-a **c-d)** PC3-RGFP-966 and A375-Tacedinaline (Larger figures can be found in Appendix E-Figure 15-16)

4.3.2.2. Analysis of Mechanism of Action: JAK inhibitors

There are three JAK inhibitor drugs; ruxolitinib, tofacitinib, and TG-101348 (fedratinib). Tofacitinib is a marketed drug used to treat rheumatoid arthritis. Ruxolitinib is marketed for the treatment of intermediate or high-risk myelofibrosis. TG-101348 is also a marketed drug used to treat myeloproliferative diseases, including myelofibrosis. They all target *JAK1*, *JAK2*, *JAK3*, and *TYK2* proteins with varying affinities; however, TG-101348 is known for its selectivity on JAK2 protein[259–264]. These three JAK inhibitors generally have distant networks both within the same cell line and across cell lines. The network pair which has the lowest separation score (-0.15) is A549-TG-101348 and A549-tofacitinib. TG-101348 and tofacitinib networks in A549 have 23 proteins in common besides the drug targets, and they are commonly enriched in Jak-STAT and Notch signaling pathways. Those in A375 are more separated (separation score is 0.40), and the only commonly enriched pathway is the Jak-STAT signaling pathway, while the A375-tofacitinib network differs with ErbB, mTOR, PI3K-Akt signaling pathways.

The most distant networks are A375-tofacitinib and MCF7-ruxolitinib, with a separation score of 0.70. Ruxolitinib and tofacitinib modulate different cellular mechanisms, as it can be understood from that pairwise separation scores are always higher than 0.35. They are separated with a score of 0.68 in A375 and only share target proteins, JAK1 and JAK2. A375-Tofaticinib is differently enriched in Jak-STAT, ErbB, PI3K-Akt signaling pathways.

Network visualizations and associated pathway enrichments of JAK inhibitor drugs can be found in Figure 22.

4.3.2.3. Analysis of Mechanism of Action: MEK inhibitors

There are two MEK (mitogen-activated protein kinase kinase enzyme-MAPKK) inhibitor drugs, selumetinib and PD-0321901(Mirdametinib), both target MAP2K1. Selumetinib is an FDA approved drug. It was very recently approved in April 2020 for neurofibromatosis type I (NF1). NF1 is a genetic disorder of the nervous system in which tumors grow on nerves. Networks of four cell lines (A375, A549, YAPC, NPC) treated with selumetinib could be reconstructed. The most distant networks for selumetinib are those of A549 and YAPC, with a separation score of 0.40 (Figure 23A-B). A549-Selumetinib network is a small network and is not significantly enriched for any networks, while YAPC-Selumetinib has several pathways enrichments covering critical signaling pathways. The lowest separation score (-0.05) came from A375 and YAPC (Figure 23C-D). A375-Selumetinib network is differently enriched in mTOR signaling, B cell and T cell receptor signaling pathways, and choline metabolism in cancer pathway. In contrast, the YAPC-Selumetinib network differs with Wnt, TGF-Beta, PI3K/Akt, cGMP-PKG, Jak-STAT, estrogen, prolactin, chemokine signaling pathways. They also share other signaling pathway enrichments such as TNF, MAPK, neurotrophin signaling pathways, transcriptional misregulation in cancer, etc.

PD-0325901 completed Phase2 trials in 2019 for the same conditions as selumetinib. There are also studies for different cancer types such as non-small cell lung cancer and head and neck cancer[265,266]. Five cell line networks could be reconstructed for PD-0325901: A375, A549, PC3, YAPC, NPC. Within these five networks, the most separated networks are of A549 and PC3 (separation score=0.44, Figure 24A), while the most similar networks are of YAPC and NPC (separation score=-0.27, Figure 24C). PC3 and YAPC networks are also distant from each other (separation score=0.43). PC3 and A549 networks share neurotrophin and insulin signaling pathway enrichments. Although the PC3-PD0325901 network is small-sized, it is significantly enriched for several pathways such as cGMP-PKG, sphingolipid, FoxO, GnRH, MAPK, Fc epsilon RI, Ras, ErbB, cAMP, Rap1, VEGF, oxytocin, prolactin, chemokine, thyroid hormone signaling pathways. A549-PD-0325901 network is differently enriched than PC3 with AMPK and mTOR signaling pathways (Figure24 B).

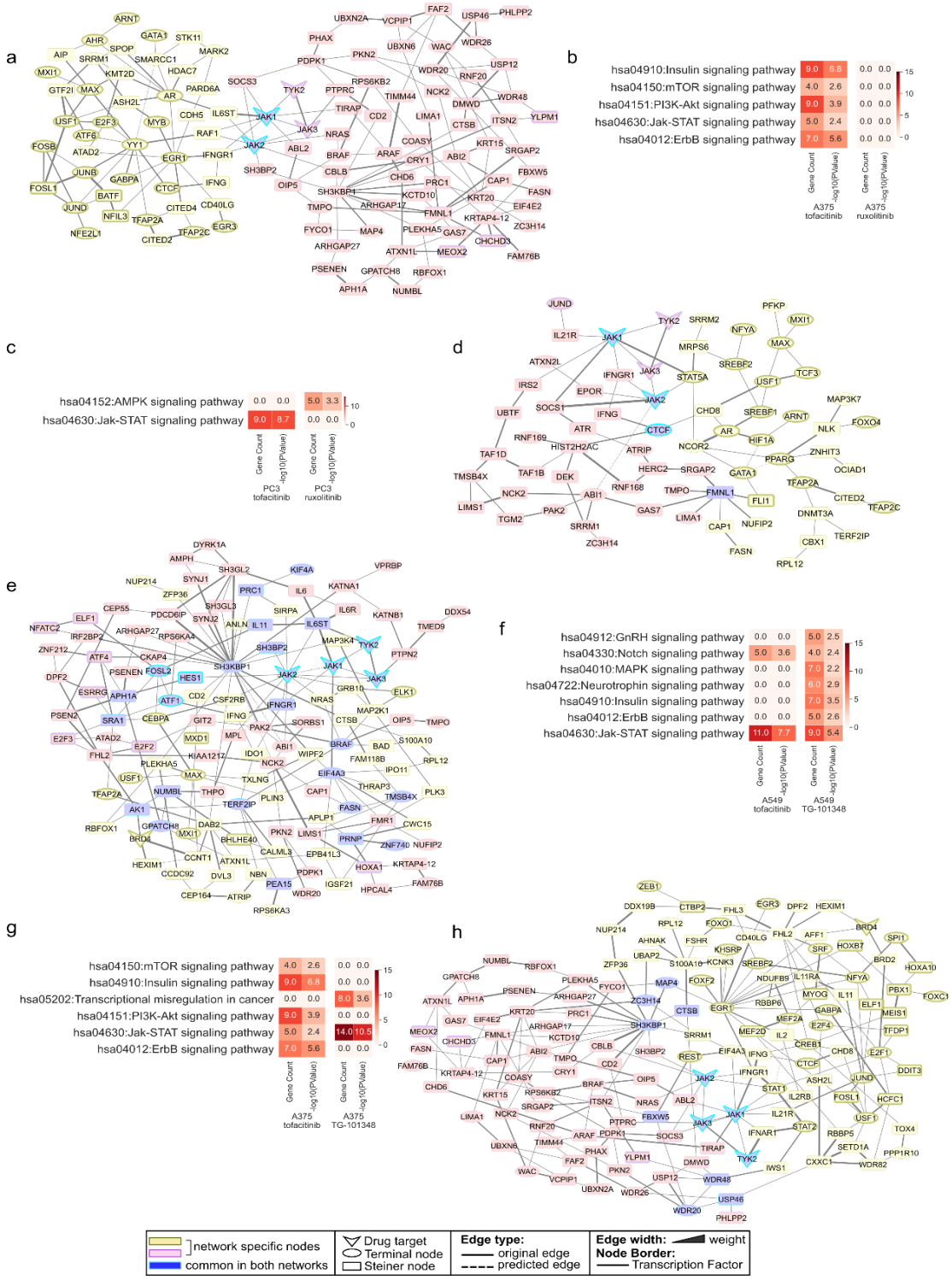


Figure 22. Network pairs and their pathway enrichments **a-b)** A375-Tofacitinib and A375-Ruxolitinib **c-d)** PC3-Tofacitinib and PC3-Ruxolitinib **e-f)** A549-Tofacitinib and A549-TG-1010348 **g-h)** A375-Tofacitinib and A375-TG-1010348

When selumetinib and PD-0325901 networks are compared against each other, a high similarity is observed within the same cell lines. Still, networks of two drugs with different cell lines have separate networks. The separation scores between two drugs on the same cells reach up to -0.64 for YAPC. Comparison of selumetinib and PD-0325901 networks between the same and different cell lines suggest that selumetinib/PD-0325901 treatment modulates different mechanisms in different tumor types. However, there are consistent similarities between the two drugs for the same cancer types, which supports clinical studies.

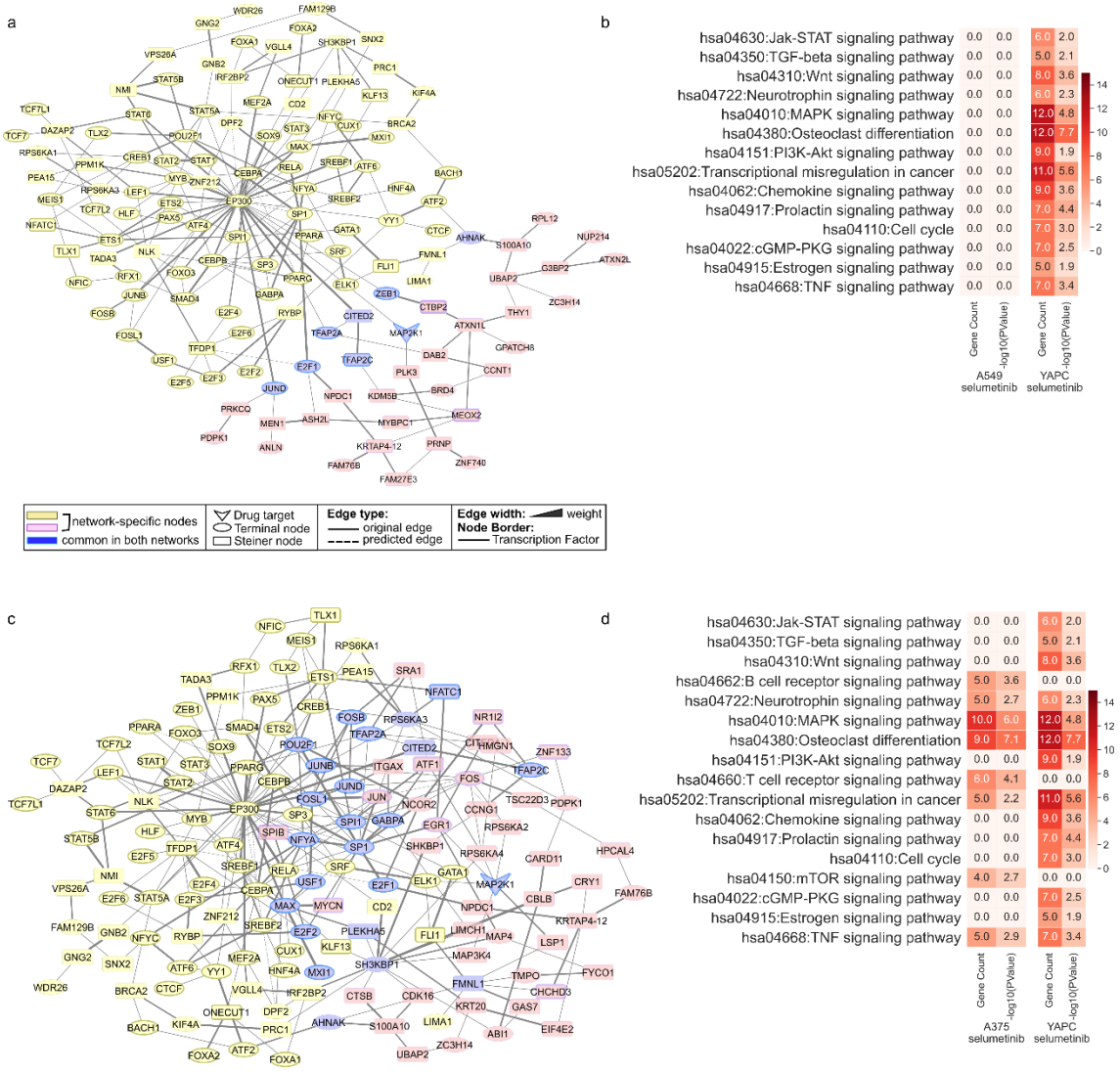


Figure 23. Network pairs and their pathway enrichments **a-b)** A549-Selumetinib and YAPC- Selumetinib **c-d)** A375- Selumetinib and YAPC- Selumetinib

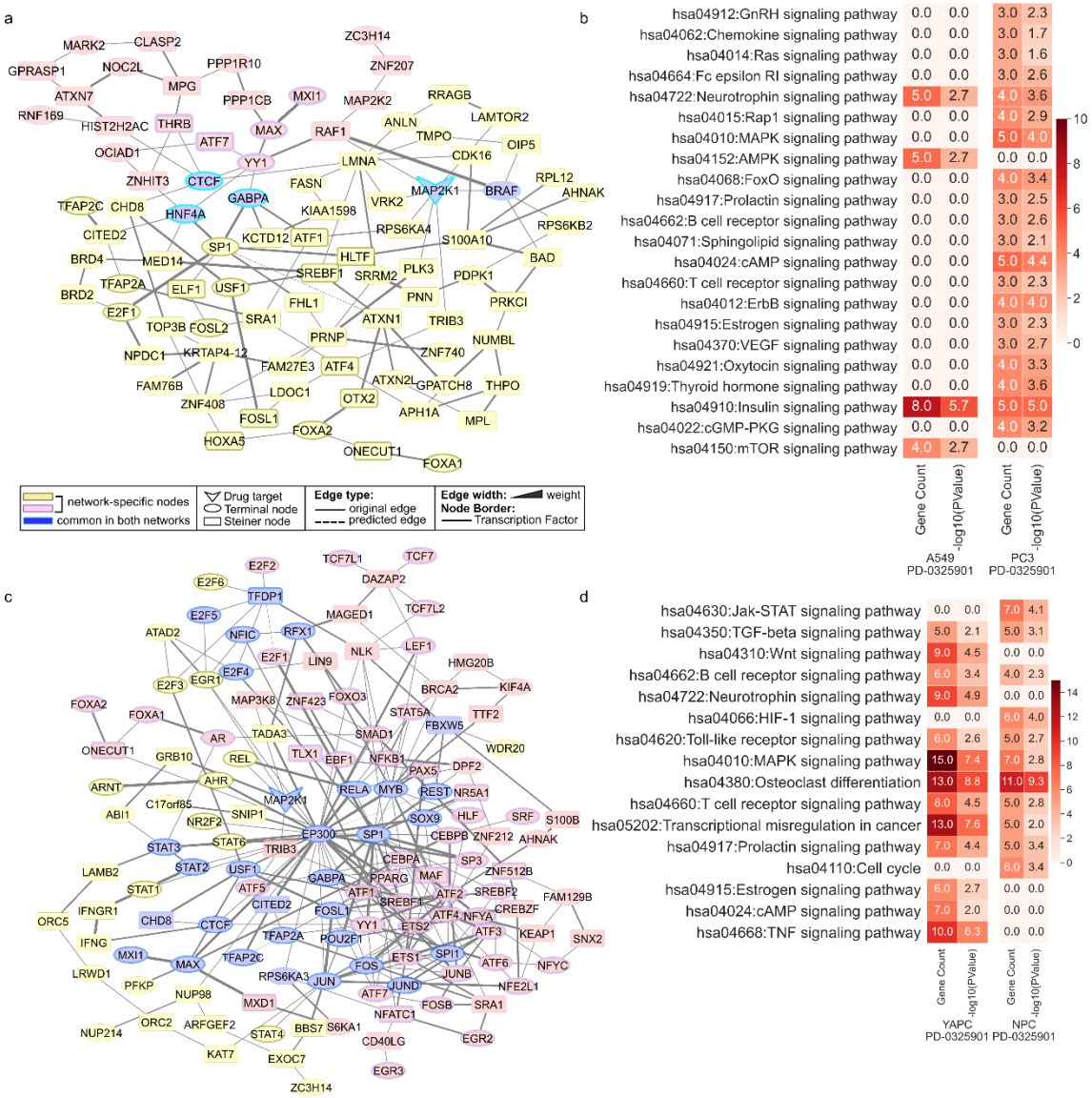


Figure 24. Network pairs and their pathway enrichments **a-b)** A549-PD-0325901 and PC3-PD-0325901 **c-d)** YAPC-PD-0325901 and NPC-PD-0325901

4.3.2.4. Analysis of Mechanism of Action: mTOR inhibitors

Everolimus, sirolimus and OSI-027 are three mTOR drugs. Everolimus is marketed with different trade names for variable purposes. The primary usage is an immunosuppressant to prevent rejection of organ transplants and as an antitumor medication to treat renal cell cancer and other tumors. For example, it is used for breast cancer in combination with another drug, exemestane, in postmenopausal women[267,268], and to treat certain advanced neuroendocrine tumors of pancreatic, gastrointestinal, or lung origin[269–271]. Networks for

A375 and A549 cell lines could be modeled for everolimus, and the separation score between them is -0.14.

Sirolimus is known as rapamycin. It is a natural macrocyclic lactone produced by the bacterium *Streptomyces hygroscopicus* and used to prevent rejection (anti-rejection medicine) in people 13 years of age and older who have received a kidney transplant. Networks for five cancer cell lines (A375, A549, MCF7, PC3, YAPC) could be modeled for sirolimus. Separation scores between cell lines range between -0.45 to 0.45. A375 network is very similar to the MCF7 network (separation score=-0.45) while the most distant network to A375 is the YAPC-sirolimus network (separation score=0.45).

Phase I clinical trials of OSI-027 were completed in 2013 for the investigation on patients with advanced solid tumors or lymphoma. Its activity on pancreatic ductal adenocarcinoma as an inhibitor of cell proliferation has been shown in [272]. OSI-027 has five cell line networks: A375, A549, MCF7, PC3, and YAPC. Separation scores of OSI-027 networks between cell lines range from -0.50 (A375-PC3) to 0.35 (A375-YAPC).

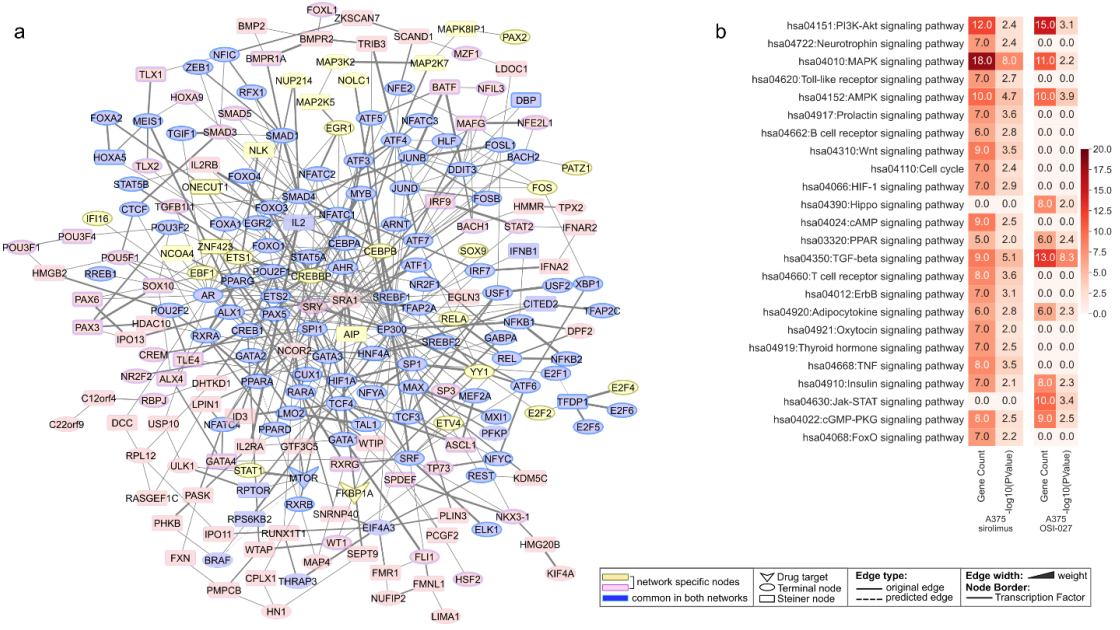


Figure 25. Network pairs (a) and their pathway enrichments (b) of A375-sirolimus and A375-OSI-027. (Larger figures can be found in Appendix E-Figure 17)

Binary relationships between cell lines for sirolimus and OSI-027 have a high similarity as they are very similar drugs. In A375, sirolimus and OSI-027 networks have a high overlap (separation score=-0.56) with many common proteins (Figure 25A). Two networks share several signaling pathways: MAPK signaling, PI3K-Akt signaling, and cGMP-PKG signaling pathways. In contrast, the OSI-027 network differs with Hippo

signaling, Jak-STAT signaling pathways, and sirolimus is additionally enriched in Wnt signaling, ErbB signaling pathways, and some immunity-related pathways such as Toll-like receptor signaling pathways (Figure 25B).

4.3.3. Analysis of separation score calculation in terms of the difference between the network-based approach used in this thesis study against the list of seed proteins

The method proposed by Menche et al.2015 only needs two sets of biomolecules and a reference network to calculate the separation scores[244] when these two sets are mapped on the reference network. For this reason, two different applications of the method, network-based, and lists of seed proteins, are used to construct separation score matrices and compare the efficiency of the approaches. The same reference interactome, iRefWeb v.13.0, is used in both.

Seed protein lists prepared for the network reconstruction step of each cell line-drug condition are used in the direct application of separation score calculation. These seed proteins include transcription factors regulating significantly transcribed genes, significantly phosphorylated proteins, and drug targets. If a network-based approach were not applied in the study, the seed proteins would demonstrate the biological entities reflecting the outcome of drug perturbation on the given cell line.

In the network-based application of separation score calculation, the proteins found in the reconstructed networks are used. These proteins may or may not include all the seed proteins as well as there may be additional Steiner proteins found via the PCSF algorithm.

Even though the network reconstruction procedure starts with the seed proteins, the two sets used to calculate separation scores would differ due to the missing or additional proteins in the resulting networks. These differences may lead to different pairwise scores and different clustering patterns between conditions. That's why separation scores between 236 conditions of cancer cell models are calculated to understand if the network-based approach is advantageous. The networks reconstructed with NPC cell line, a non-cancer cell model, are excluded from this analysis since it may create discrepancies in the clustering and mislead this analysis.

Figure 26 includes the hierarchically clustered heatmaps of pairwise separation scores calculated by both applications. The network-based approach resulted in a small subset of cell line-drug conditions with overlapping networks (Figure 26A). On the other hand, using the seed proteins resulted in very similar scores such that most of the separation scores are in the negative range leading to the higher similarity between conditions (Figure 26B). It is not expected for these 236 cell line-drug conditions to similarly modulate the same pathways due to the molecular heterogeneity. Therefore, the results collected with the direct application of the separation score calculation include higher false positives across cancer types and drugs.

The difference between the two approaches stems mainly from the Steiner nodes found in the networks. In the direct application, spurious and low-confidence interactions within the reference interactome may be used to find the shortest distances between proteins. In contrast, in the network-based application, PCSF finds an optimal network with high confidence interactions using Steiner nodes. The Steiner nodes have the potential to reflect the off-target effects of the drugs.

To evaluate the performance of the method, ground truth is defined such that drugs with the same MoA would have similar outputs. The drug pairs having the same MoA and similar resulting networks are defined as true positives (TP); those having the same MoA and separated networks are defined as false negatives (FN). The drug pairs having different MoA and overlapping networks are defined as false positives (FP), and those having different MoA and separated networks are defined as true negatives (TN). The performance evaluation metrics are calculated for different separation score thresholds ranging from -1.0 to +1.0. Plots for performance evaluation can be found in Appendix B. Because the data is imbalanced (200 positive cases - pairs with same MoA, 6017 negative cases - pairs with different MoA), the best performance is found based on Matthew's correlation coefficient (MCC).

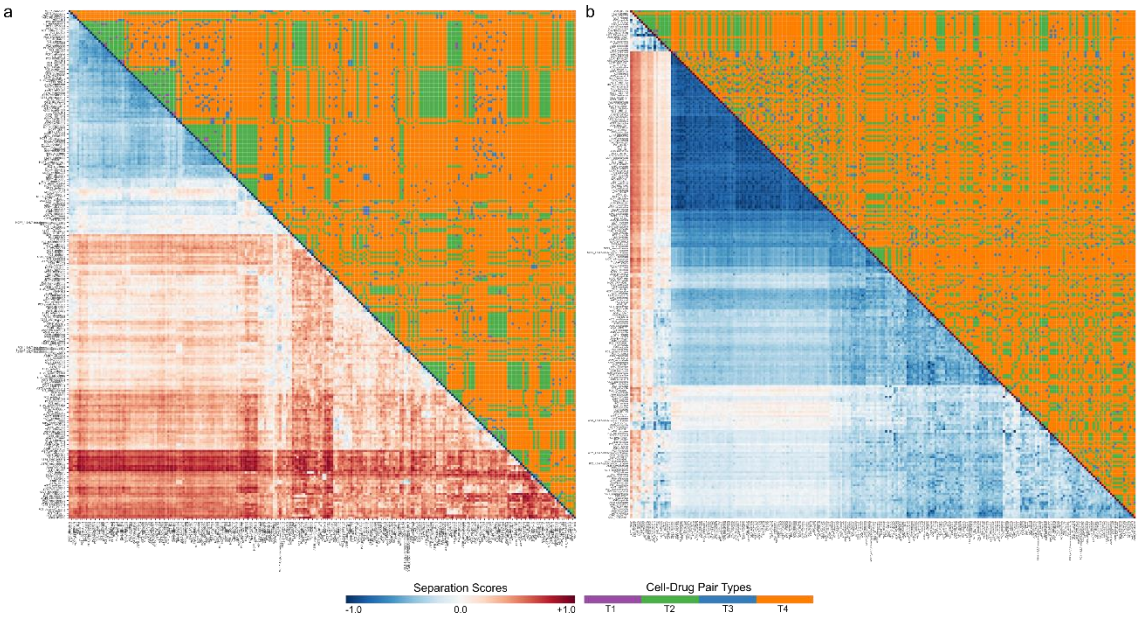


Figure 26. Separation score matrices depicting comparison between 236 cell line – drug conditions. a) separation score matrix of reconstructed networks and b) pairwise separation score matrix of seed protein lists

The best performance of the network-based classification was achieved with the separation score threshold of -0.45 (precision=0.138, recall=0.400, accuracy=0.900,

sensitivity=0.400, specificity=0.917, TPR= 0.400, FPR=0.083, F1-score=0.205, and MCC=0.192). When the same threshold is used for the direct application to the seed proteins without network reconstruction, the performance is significantly low with many false positives (MCC=0.116 and FPR=0.358). As a result, integrative network modeling provides more stringent and biologically relevant results.

4.3.4. Analysis of networks based on pathway enrichments

NMFConsensus clustering was performed on the KEGG term matrix for k=2 to k=10. The cophenetic coefficients of the results are considered to decide which k gave the best result. Best cophenetic coefficients are observed for k=5 and k=7; 0.9789 and 0.9791, respectively. To select the most suitable clustering and understand the difference between five and seven clusters, two cases are individually compared against the networks with topologically overlapping structures (Table 8). The results of these five and seven clusters are given in Figure 27A-B. Given in Figure 27C that is a comparison of the k=5 case, cluster 5 mostly consists of overlapping networks such that 71 out of 92 networks have pairwisely negative separation score and the remaining 12 networks are distributed to the first four clusters. In the case of k=7, the results are more or less the same in terms of the coverage of overlapping networks. Again cluster 5 includes most of the topologically similar networks, while clusters 1, 6, and 7 do not include any. The similar networks coverages and cophenetic coefficients of two conditions (k=5 and k=7) are almost the same, and the results shown in Figure 27A-B show k=5 is more precise. Therefore, it is decided to continue the pathway-level comparison analysis with five clusters.

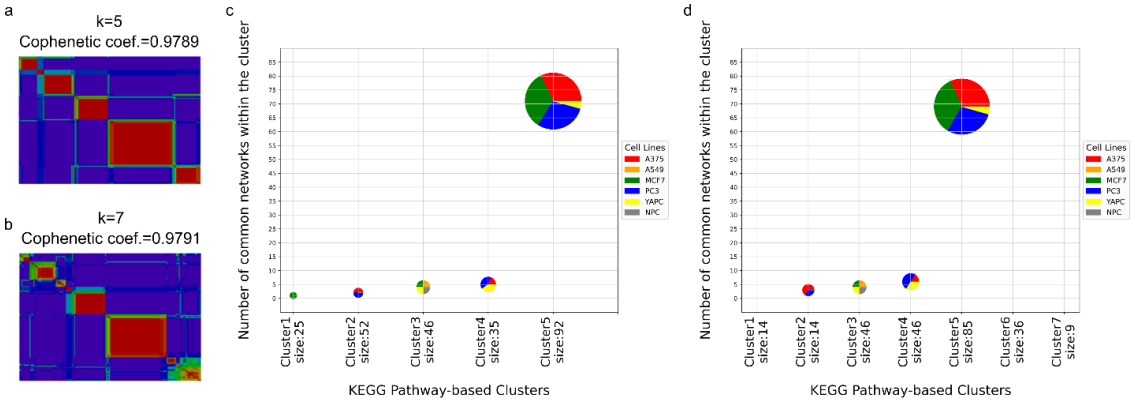


Figure 27. NMFConsensus Clustering Results for k=5 and k=7. **a)** NMFConsensus output of five clusters **b)** NMFConsensus output of seven clusters **c)** Comparison of five clusters to the list of networks with negative pairwise separation scores (overlapping networks) **d)** Comparison of seven clusters to the list of networks with negative pairwise separation scores (overlapping networks) (Larger figures can be found in Appendix E-Figure 18-19)

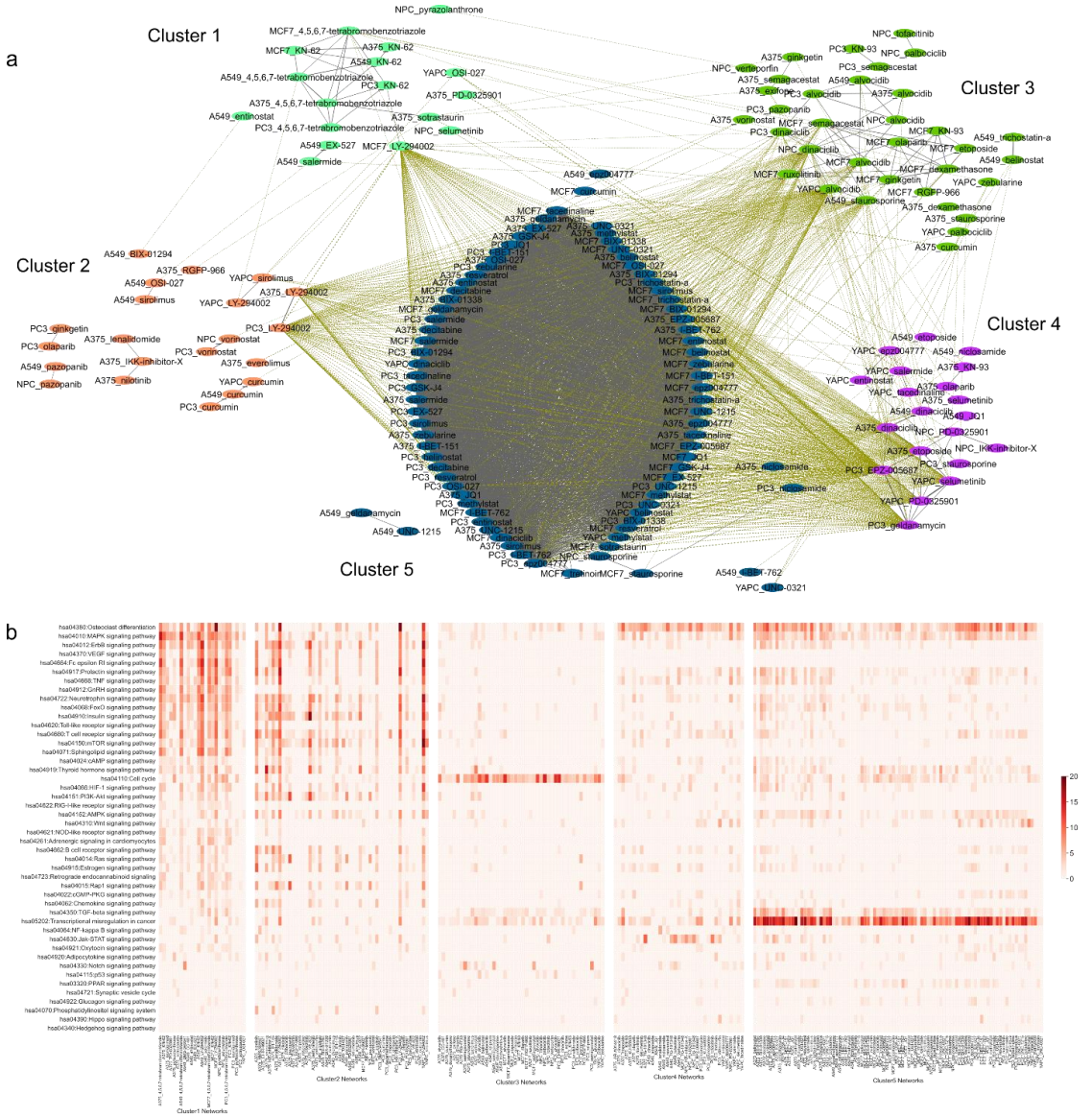


Figure 28. Relationships between drug-cell line pair networks and their clusters a) Network representation of KEGG clusters with edges representing overlapping cell line-drug pair networks. b) Heatmap of KEGG terms enriched in each cell line-drug pair network (KEGG terms on y-axis and cell line-drug pairs on x-axis). Networks are ordered by cluster number and by alphabetical order. (Larger figures can be found in Appendix E-Figure 20-21)

Pathway-level analysis of 250 networks revealed similar networks consistently with the topology-based analysis. 85% of topologically similar networks are found within the same

pathway-based cluster. A network illustration is produced in which networks are denoted as nodes. Pathway-based clusters are shown using different colors in nodes, the layout is specified to visualize pathway-based clusters separately, and topological similarity is reflected with edges. An edge is added between two networks if their separation score is lower than -0.20. In-cluster edges are reflected with solid gray lines, while between-cluster edges are reflected with dashed olive lines. Edge width refers to the separation scores between networks. (Figure 28A). Some cell line-drug pair networks do not have an edge to their corresponding cluster members while having a topological overlap between other cell line-drug pair networks included in another cluster. For example, A549-BIX-01294 is included in Cluster 2, and it does not have any within-cluster edge. However, it is connected to A549-entinostat found in Cluster 1, which means they have a separation score lower than -0.20.

Moreover, the KEGG terms enriched in each network are back-investigated by their associated cluster. The heatmap in Figure 28B shows the pathway enrichments. Clusters are grouped. The networks within each cluster are ordered alphabetically on the x-axis, and the KEGG terms are listed on the y-axis. By this analysis, it is possible to observe the signature pathways of the clusters. Cluster 1 and 2 are enriched in several pathways, so it is not possible to assign specific pathways as the signatures of clusters 1 and 2. On the other hand, cluster 3 is highly enriched in the pathway cell cycle. Cluster 4 shows a consistent enrichment for osteoclast differentiation. Finally, the signature pathways of cluster 5 that represent most of the overlapping networks are transcriptional misregulation in cancer, MAPK signaling, and osteoclast differentiation. Within cluster 5, A375 networks show additional pathway enrichments such as prolactin, TNF, neurotrophin, and FoxO signaling pathways, and MCF7 networks show additional enrichments in cell cycle, thyroid hormone signaling pathways, and PC3 networks are additionally enriched in Wnt, AMPK, and cGMP-PKG signaling pathways.

4.4. Application of Network-based Perturbed Cell Analysis to Combinatorial Drug Treatment

Using two or more drugs as a combination is usually advantageous for disease treatment when the application of one drug continuously and/or with high dosages causes resistance, recurrence, and when the drug has adverse side-effects[273–275]. Cancer drugs usually cause resistance[276,277] upon continuous administration, and they cause variable side-effects dependent on the drug mechanism of action, dosage, and some other factors such as genetic factors, administration route, etc[278]. Network-based investigation of omics data of perturbed cancer cells reveals the cellular mechanisms affected by the perturbation. Combinatorial drug therapy is an effective cancer treatment strategy to decrease the resistance and side effects of drugs. Two drugs may be efficient to deal with cancer when they modulate different pathways, which both have a relation with the disease. In terms of the separation score method, drugs to be used as a combination should affect a different

set of proteins leading to different drug modules, while these proteins should be close enough on the interactome to overlap with the disease module. This situation can be called as 'attacking the disease from different sides'[279,280]. By the analysis of reconstructed networks of perturbed cells, drugs affecting disease-related different pathways could be inferred.

Cheng et al., 2019 investigated the efficacy of drug combinations with the application separation score method on the drug and disease modules[281]. They characterized the drug module by the drug target proteins and the disease module by the disease proteins. Then they calculated the drug-drug and drug-disease separation to understand drug-drug-disease interactions for experimentally validated efficient drug combinations. Out of six classes of drug-drug-disease interactions, they reported that one class shows correlation with therapeutic effects, which they called as 'complementary exposure'. According to 'complementary exposure', two drug modules are separated with a positive separation score, and each drug module has an overlap with the disease module (drug-disease modules have negative separation scores.).

Using the experimentally validated cancer drug pair data they provided, the intersection with CMap data is extracted to further analyze in terms of perturbed networks in this study. There are four drug pairs both data commonly have, and these drug pairs could be analyzed for a total of seven cases such that some of the drug pairs have network information in different cell lines (Table 9).

Tablo 9. FDA approved or experimentally validated drug combination cases

Drug1	**Drug2**	**Cell Line**
Alvocidib	Vorinostat	A375
Alvocidib	Vorinostat	PC3
Alvocidib	Vemurafenib	A375
Vorinostat	Vemurafenib	A375
Lenalidomide	Dexamethasone	A375
Lenalidomide	Dexamethasone	A549
Lenalidomide	Dexamethasone	PC3

To be able to perform a similar analysis, drug and disease modules should be specified. As a general cancer drug combination analysis, cancer driver genes could be used as disease genes constituting the disease module. Cancer Genome Interpreter[282] data is used to collect the cancer driver gene list. The reconstructed networks constitute a subnetwork of cell line treated with the drug. Each cell line–drug network includes many hidden proteins and omic hits together with the drug targets and is a merged network of the drug module and a subnetwork of disease module. Therefore, the coverage of the cell line–drug networks is higher than the drug modules described in the above study[281]. This analysis is carried out with the idea that to be an effective drug combination, two cell line–drug

networks should have at least one cancer driver protein in common, and they should be topologically separated enough, including different cancer driver proteins (Figure 29A).

The first case is the Alvocidib-Vorinostat pair in A375. Alvocidib is a CDK inhibitor, and vorinostat is an HDAC inhibitor drug. The networks of two drugs in A375 have a separation score of -0.09, the lowest score among the seven cases. Modulated pathways by each drug show a complementarity, with only cell cycle pathway being in common in both. A375-Alvocidib network is enriched in p53, FoxO, ErbB, PI3K-Akt, insulin, and T cell receptor signaling pathways, while the A375-Vorinostat network is enriched in MAPK, Wnt, Toll-like receptor, estrogen, notch, prolactin, and thyroid hormone signaling pathways (Figure 29B). These two networks have common cancer driver proteins (RB1, MAX, and BRD4) and other cancer driver proteins unique to each network (Figure 29C). The same drug pair shows a similar situation in PC3 cells such that their networks are separated with a score of 0.37, and they share one common cancer driver protein.

Among all seven cases, the pairwise separation scores lie between -0.09 and 0.68. While the separation score gets closer to +1, the commonality decreases as expected. The number of common nodes between networks of drug pairs ranges from three to 16 proteins, and only one case does not involve a shared cancer driver protein (A375: Lenalidomide-Dexamethasone, separation score:0.527). All seven cases consistently have topologically disjoint regions, including cancer driver proteins, and their pathway enrichments reveal different signaling pathways modulated.

The complementary exposure concept and the above seven cases' criteria can be used as a guideline to predict new drug pairs. From 6,308 possible drug pairs for the same cell line, 1,441 drug pairs are rated with the following filtering: (i) the separation score between drug networks in the same cell line should be between 0.0 and 0.7; (ii) two networks shouldn’t share more than two modulated networks, and (iii) size of both networks should be larger than 40, and also the size of at least one network should be larger than 100. I-BET-762 and lenalidomide drug pair in the A375 cell line is rated first when all predictions are ordered with decreasing separation scores. Two networks (A375-lenalidomide and A375-I-BET-762) have a separation score of 0.69, and there is no commonly modulated pathway. Lenalidomide network is enriched in several important pathways such as PI3K-Akt, ErbB, MAPK, Ras, Rap1, VEGF, and FoxO signaling pathways. In contrast, the I-BET-762 network is enriched in Jak-STAT signaling and transcriptional misregulation in cancer pathways (Figure 30A). Moreover, two networks have three proteins in common, and one out of these is a cancer driver protein (*BRAF*). The cancer driver proteins that are unique to each network are also shown in Figure 30B. Synergistic activities of BET inhibitors in combination with lenalidomide are shown in different studies of multiple myeloma[283,284]. Moreover, the combination of lenalidomide with CPI-203, a primary amide analog of (+)-JQ1 and having the same mechanism of action as I-BET-762, is shown effective to synergistically induce cell death in bortezomib-resistant mantle cell lymphoma[285].

More examples within the highest ranked 100 drug pairs have evidence for effective drug combinations based on their MoAs. Three cases of those are geldanamycin and tofacitinib

(A375, separation score=0.66), trichostatin-a and tofacitinib (A375, separation score=0.56), and resveratrol and ginkgetin (PC3, separation score=0.53). Geldanamycin is an HSP inhibitor, and tofacitinib is a JAK inhibitor drug. HSP90 and JAK2 inhibition is shown to synergistically overcome resistance to JAK2-TKI in human myeloproliferative neoplasm cells[286]. Although tofacitinib is not selective for JAK2, its selective MoA counterpart, TG-101348, and geldanamycin pair is also found within the predicted drug pairs. The second example, trichostatin-a and tofacitinib, includes a combination of HDAC and JAK inhibition. HDAC and JAK dual inhibition has also been studied in several studies to improve treatment strategies[287–290]. Furthermore, resveratrol and ginkgetin combination against colorectal cancer is shown to synergistically suppress VEGF-induced angiogenesis[291]. A list of highest ranked first 100 predicted drug pairs can be found in Appendix C.

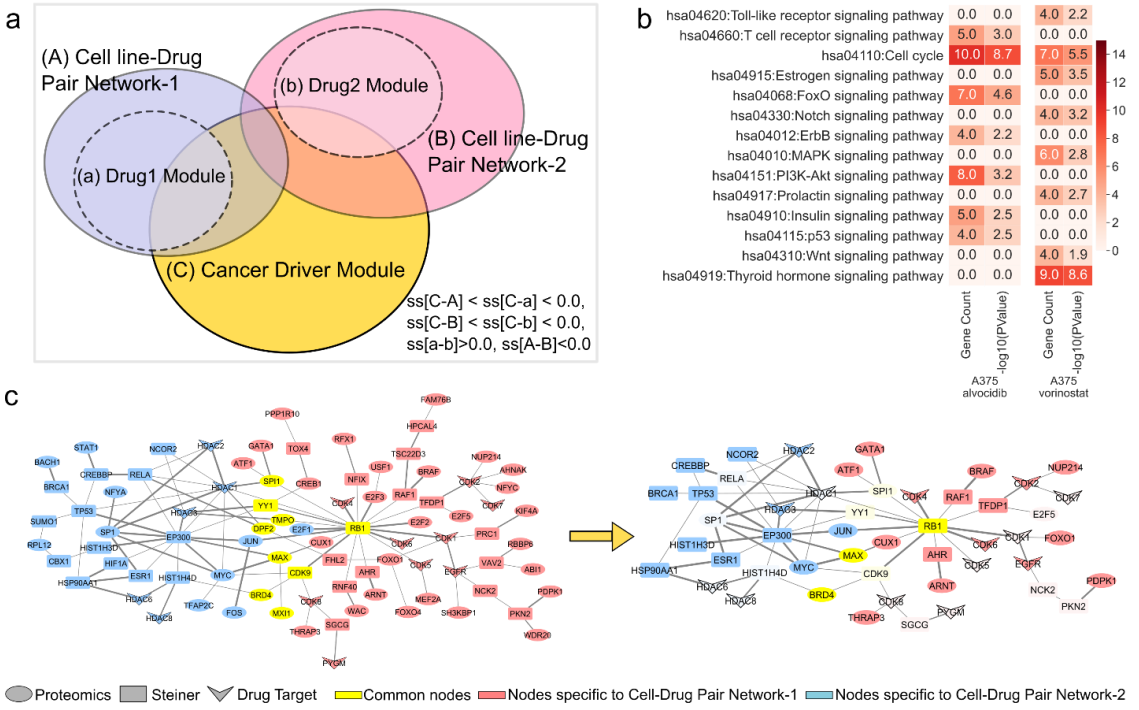


Figure 29. Illustration of the network-based analysis of efficient drug combination **a)** Illustration of the hypothesized overlaps of drug and disease modules together with cell line-drug pair networks **b)** Signaling pathway enrichments of Alvocidib and Vorinostat networks on A375 **c)** Merged network maps of Alvocidib and Vorinostat treated A375 cell lines and the subnetwork that includes cancer driver genes. Yellow nodes represent commonly found proteins in two cell line-drug pair networks, blue nodes represent proteins unique to the A375-Vorinostat network, and pink nodes represent proteins unique to the A375-Alvocidib network.

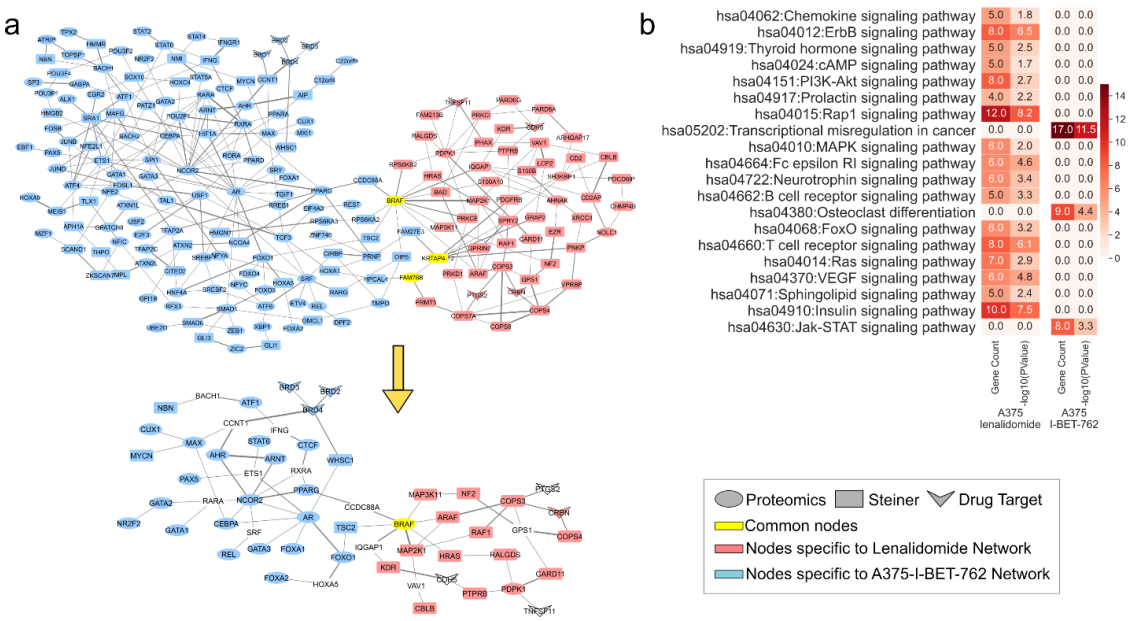


Figure 30. Illustration of the network-based analysis of predicted drug combination **a)** Merged network maps of Alvocidib and Vorinostat treated A375 cell lines and the subnetwork that includes cancer driver genes. Yellow nodes represent commonly found proteins in two cell line-drug pair networks, blue nodes represent proteins unique to the A375-I-BET-762 network, and pink nodes represent proteins unique to the A375-Lenalidomide network. **b)** Signaling pathway enrichments of Lenalidomide and I-BET-762 networks on A375

4.5. Application of Network-based Perturbed Cell Analysis to Drug Sensitivity Analysis of Cell Lines

Network-based analysis of perturbed cells can also be used for the investigation of drug sensitivity mechanisms. Drug sensitivity scores of five cancer cell lines are downloaded from the Genomics of Drug Sensitivity in Cancer (GDSC) Database[292], which provides anti-cancer drug sensitivity information for more than 1000 genetically characterized human cancer cell lines. There are 30 common drugs in CMap P100 and GDSC datasets, and among those 30 drugs, four cell lines are significantly sensitive or resistant to only five drugs. The available cell line – drug conditions are as follows:

- A375 is significantly sensitive to two drugs: PD-0325901 and CHIR-99021.
  The reconstructed networks include four cell line networks of PD-0325901 (A375, A549, PC3, and YAPC) and three cell line networks of CHIR-99021 (A375, MCF7 and PC3).
- MCF7 is significantly sensitive to Trichostatin-a.

The reconstructed networks include three cell line networks of Trichostatin-a (MCF7, A375 and PC3).

- PC3 is significantly resistant to Staurosporine.
  The reconstructed networks include four cell line networks of Staurosporine (A375, A549, MCF7 and PC3).
- YAPC is significantly resistant to Dinaciclib.
  The reconstructed networks include five cell line networks of Dinaciclib (A375, A549, MCF7, PC3 and YAPC).

For PD-0325901 (Mirdametinib), to which A375 cells are sensitive, several signaling pathways are observed as activated or inactivated depending on cancer cell type. PD-0325901 is a selective MAP2K1 (MEK1) inhibitor that is directly related to cell proliferation. As MEK is in the middle of the RAS/RAF/MEK/ERK pathway, which is on the upstream of several cellular mechanisms for cell proliferation and cell survival, it is expected to have several pathways affected. The neurotrophin signaling pathway is enriched in all cells. However, cells differentiate from each other with some of the other signaling pathways. For example, PC3 cells differ with several pathways such as cGMP-PKG, sphingolipid, Rap1, Ras, VEGF, Oxytocin, FoxO, and Fc epsilon RI signaling. Sphingolipid signaling is shown to be involved in the resistance of prostate cancer cell lines to antineoplastic drug treatment (z-score of PC3 is 1.95, very close to the resistance threshold)[293]. A549 cells are only enriched in mTOR signaling, insulin signaling, neurotrophin signaling, and AMPK signaling pathways. YAPC cells are also enriched in several pathways, but YAPC differs from other cells with Wnt signaling, TGF-beta signaling, Toll-like receptor, and TNF signaling pathways. MAPK signaling, mTOR signaling, GnRH signaling, ErbB signaling, insulin signaling, and cAMP signaling pathways are active in sensitive A375 cells (Figure 31A).

CHIR-99021 is a glycogen synthase kinase 3 inhibitor targeting GSK3B. GSK3 is on the downstream of several signaling pathways such as Wnt, PI3K/Akt, growth factors, hedgehog, etc. It is known as a Wnt antagonist and is also involved in cell cycle regulation. Several studies show the relationship of GSK3B with neurodegeneration[294], neurotransmission[295], several neurological disorders[296–298], prolactin signaling[299], and TGF-beta induced differentiation[300]. These pathways are enriched in different cell lines. Sensitive A375 cells are enriched in the synaptic vesicle cycle, while MCF7 cells are enriched in Wnt signaling, and PC3 cells are enriched in prolactin and TGF-beta signaling pathways. Inhibition of GSK3B by CHIR-99021 may cause several pathways to be more active depending on the state of the cell. GSK3 inhibition also has a role in cell proliferation together with several growth factors, and this may be the reason for cell cycle and prolactin signaling pathways to be active in PC3 cells as it is shown that GSK3 expression is increased in prostate cancer[301]. (Figure 31B)

There are three networks to use for the comparison of Trichostatin-a (TSA): MCF7, A375, and PC3. Three cell lines differ from each other with active TGF-beta signaling pathway in MCF7, neurotrophin, estrogen, MAPK, and TNF signaling pathways in A375, and AMPK and JAK-STAT signaling pathways in PC3. TSA targets class I and class II

HDACs. Treatment with HDAC inhibitors is reported to restore TGF-beta signaling in breast cancer, so it is meaningful to observe the TGF-beta signaling pathway enriched in MCF7 cells[302]. Also, the transcriptomic profile of TSA treated SK-MEL-3 cells carrying BRAF mutation is examined in Mazzio et al., 2018[303], and TSA is shown to down-regulate proteins in MAPK/MEK/BRAF pathway. (Figure 31C)

Staurosporine is a broad spectrum, ATP competitive protein kinase inhibitor and a relatively nonspecific inhibitor of protein kinase C, targeting CHEK1, MARK3, PDPK1. PC3 is resistant to staurosporine and shows no significantly active signaling pathways, while A375 and A549 are enriched for the cell cycle, and MCF7 is enriched for the MAPK signaling pathway.

Dinaciclib is a CDK inhibitor. It is shown to suppress tumor growth for several cancer types such as pancreatic cancer[304], breast cancer[305], ovarian cancer[306], and thyroid cancer[307]. When we look into signaling pathways enriched in dinaciclib networks, YAPC is not enriched for the cell cycle compared to other cells. On the other hand, YAPC is highly enriched in the Wnt signaling pathway together with PC3. A375 differs from others with MAPK signaling pathway, and MCF7 differs with Jak-STAT and TGF-beta signaling pathways. Wnt signaling stimulates cell proliferation which in turn it is also related to cell cycle[308]. Some of the Wnt signaling pathway components are shown to be effective in apoptotic regulation[309]. As YAPC cells are resistant to dinaciclib, it may be expected for cell cycle regulation not to be active but for Wnt signaling to be active due to a side effect or another mechanism as discussed in Bryja et al.,2017[308].

The drug resistance of the cells may be affected by several cellular mechanisms, and it is not possible to directly infer resistance mechanisms from the reconstructed networks. On the other hand, the modulated pathways in drug-sensitive cancer cells may hint at the reasons leading to sensitivity. For this reason, analysis of the drugs that at least one cancer cell line is sensitive to (CHIR-99021, PD-0325901, and Trichostatin-a) is more feasible. However, there is no direct relationship between enriched pathways observed in sensitive cells across the enriched pathways observed in non-sensitive cells. Among the networks of three sensitive drug cases, Wnt signaling is not enriched in any networks of sensitive cell lines (MCF7-Trichostatin-a, A375-CHIR-99021, and A375-PD-0325901).

The z-scores of cell lines are compared pairwisely against the separation score between the corresponding networks. An increasing separation score is observed in cases of CHIR-99021 and PD-0325901 with the increasing delta z-score. The same condition is not valid for TSA since it is one of the drugs with highly overlapping networks across all cell lines (Figure 31D). Inspiring from the relationship of z-scores and separation scores in PD-0325901 and CHIR-99021 cases, all possible data, including pairwise z-score differences and separation scores, is analyzed. Within 30 drugs with sensitivity data of cancer cell lines, those with reconstructed networks are collected, and those with one positive z-score and one negative z-score are filtered. Then the pairs with separation scores are lower than -0.45 are excluded to avoid the networks with highly overlapping topology. The regression analysis is performed with the remaining data, and a moderately positive correlation is

observed such that separation score increases with increasing delta z-scores (R=.29, p-value=0.011) (Figure 31E).

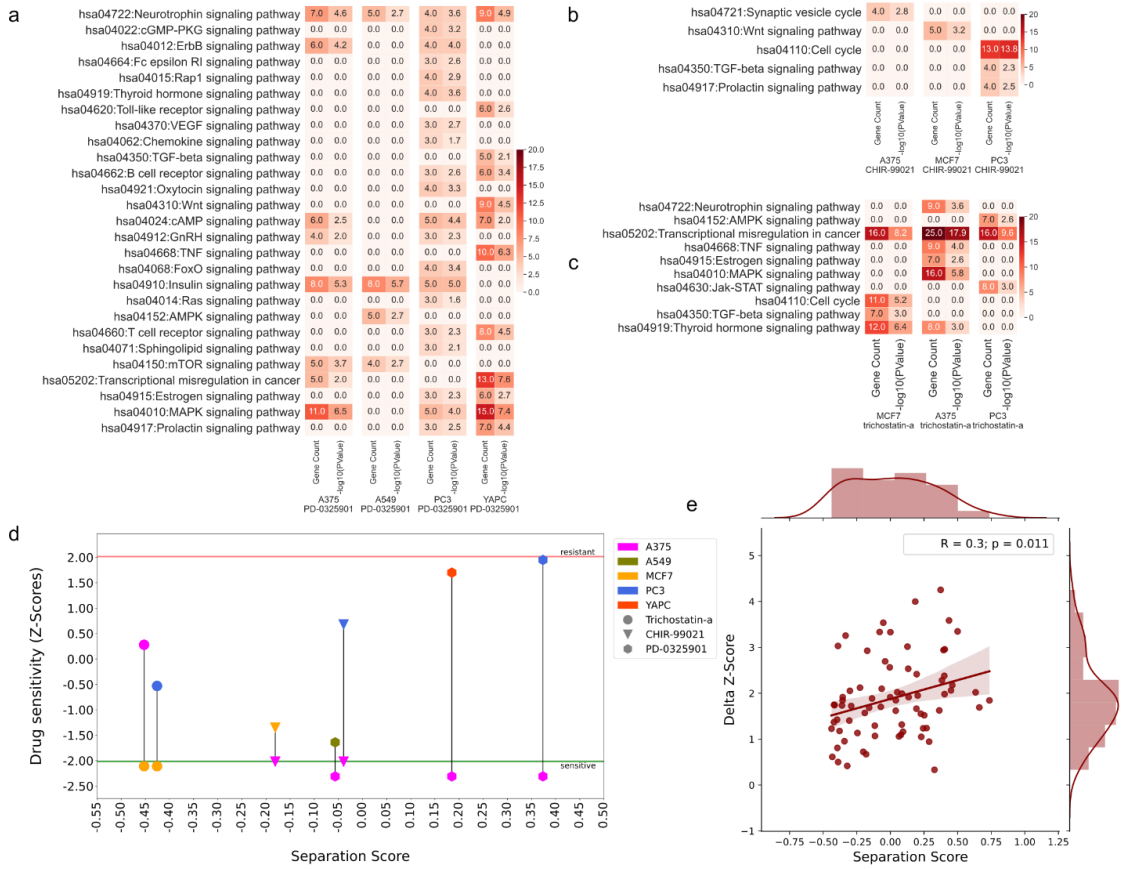


Figure 31. Analysis of drugs that cell lines are sensitive to. **a-c)** Signaling pathways enriched in drug networks per cell type: PD-0325901, Trichostatin-a, and CHIR-99021, respectively. **d)** Separation scores for network pairs of each drug in sensitive cell lines and non-sensitive cell lines versus drug sensitivity plot. Drugs are mapped with different shapes, and cell lines are mapped with different colors. **e)** Regression plot of pairwise separation scores and drug sensitivity score differences for conditions in which one drug of two cell line-drug pair networks has a negative z-score on its corresponding cell line and the other drug has a positive z-score on its corresponding cell line and in which the separation score is higher than -0.45. (Larger figures can be found in Appendix E-Figure 22-23)

CHAPTER 5

DISCUSSION

Drug classification studies hold an essential key for drug repurposing. There are a vast number of drugs used for cancer treatment strategies. When these drugs are classified according to known target molecules and/or mechanisms of action, they either constitute a small number of very crowded groups under main categories or a large number of groups with a minimal number of drugs due to highly detailed classification methods. If drugs are classified based on their chemical structures, they cannot be classified appropriately since most of them are derivatives of each other. Also, drug groups may change when the drug features used for the classification are changed[310–314]. The CMap P100 dataset is used to analyze the drug modulations on six cell lines in this study. P100 dataset includes 89 small molecule drug perturbations and one associated control treatment. After the statistical analysis of transcriptomics and phosphoproteomics data included in the P100 dataset, there were 70 significant drugs to study further. These 70 drugs were firstly classified according to two categories; their chemical structures and mechanisms of action (MoAs). For chemical structure-based classification, two metrics were used: (i) Tanimoto similarities and (ii) MACCS key distances. Tanimoto similarities didn't discriminate between drugs, while MACCS key distances provided five drug clusters with variable sizes. MoA-based classification discriminated drugs into specific categories with clusters included one to five drugs. However, this kind of clustering does not still allow understanding the changes in transcriptomics or phosphoproteomics. Moreover, drugs may induce different mechanisms according to the tumor type and usually affect variable pathways leading to different phenotypes[315–317]. Therefore, it is not possible to predict the outcome of a drug by only using its chemical structure and MoA.

Modulations lead by the perturbation on drug targets are not local since cells include complex interactions between molecules. The first inhibition/activation starting from the attack to drug target(s) diffuse to different locations within the cell and cause changes in the steady-state activities of several signaling pathways. The interplay between signaling pathways is also important in the total cellular effects of drugs. When the perturbation effects of drugs are investigated based on the activated/inactivated signal mechanisms, better resolution can be acquired for the classification of drugs, which helps find out similarly acting drugs on different cancer types. Protein-protein interaction subnetworks

of perturbed cell lines are reconstructed by integrating transcriptomic and phosphoproteomic data with drug targets using the Omics Integrator software to infer the similarly acting drug groups. Omics Integrator software takes a reference interactome to build networks. iRefWeb human interactome is used after some condition-specific filters are applied to decrease the false positives. Adamic/Adar link prediction method is performed to enrich the interactome for false negatives. The reconstructed cell line–drug networks include the hidden proteins besides the experimental hits. According to the topological similarity method used in this study, it is hypothesized that similarity between two networks on node level can reveal the shared cellular mechanisms. Reconstructed networks also allow understanding modulated pathways caused by drug perturbations.

Topology-based and pathway-based comparison of the networks revealed that a subgroup of drugs modulates overlapped networks even though their chemical structure and MoA are different. A comparison of tacedinaline and geldanamycin networks on A375 is illustrated to show that drugs with different drug targets/MoAs may lead to overlapped networks and may affect common transcription factors responsible for similar phenotypes. Moreover, drugs with the same MoAs and targets modulate different pathways besides the common pathways, which differentiates these drugs from each other (i.e., sirolimus and OSI-027). It is shown that in A375, sirolimus modulates Wnt, cAMP, ErbB, and FoxO signaling pathways while the OSI-027 network is differently enriched in Jak-STAT and Hippo signaling. As well as effects of different drugs could be compared against each other in the same cell type, the effect of a drug in different cell types could also be compared. In this sense, tacedinaline networks in A375 and YAPC cell lines are examined, and pathway-level differences are listed.

Four types of drug mechanisms of action categories are separately examined to understand both within cell lines and across cell lines differences of drugs with the same MoAs. In this context, networks of HDAC inhibitors, JAK inhibitors, MEK inhibitors, and mTOR inhibitors are analyzed. Both overlapping and distant network examples within the same MoA groups are visualized. Target selectivity is the primary determinant factor of network similarity. If two drugs are selective for different target proteins even though they are included in the same MoA group, their networks modulate different signaling pathways, leading to different cellular phenotypes. For example, RGFP-966 is a selective HDAC3 inhibitor, while Belinostat, vorinostat, Trichostatin-a are broad-spectrum HDACi drugs. Networks of broad-spectrum HDACi drugs have overlapping regions in most of the cell lines. However, RGFP-966 networks are distant from other HDACi drugs in the same cell lines and to each other across different cell lines.

KEGG pathway enrichments of 250 networks are also used to compare cell line–drug conditions. Significant signaling pathways enriched in all networks are transformed into a matrix including zero if the network is not enriched in the KEGG term and $-\log_{10}$(p-value) if the network is enriched in. This matrix is exposed to NMFConsensus clustering, and the signature pathways of each cluster are found out. One of the pathway-based clusters is highly consistent with the topologically overlapping network cluster. The

signature pathway of this cell line–drug cluster is transcriptional misregulation in cancer and TGF-beta and MAPK signaling pathways.

Both topology- and pathway-based network analysis reveal several important pieces of information about the action mechanisms of drugs on different cell lines. Menche et al.,2015 reported molecular similarity of the diseases could be evaluated by the overlap of disease modules[244]. This approach is also applied to the assessment of drug pair efficacies for variable diseases[281]. In this study, the same topological similarity calculation method is used to check the consistency of drug-drug-disease relationships of the experimentally validated drug pairs with the networks reconstructed in this study. It is shown that reconstructed networks of perturbed cell lines are applicable to combinatorial drug prediction. Experimentally validated drug pairs constitute topologically separate networks with a little overlap consisting of at least one disease gene. There was a limited amount of drug pairs to evaluate the consistency; however, it is still clear that for two drugs to be used for the treatment of a disease, the perturbed networks should be distant enough to be able to target the disease with a complementary effect. Also, their enriched pathways show a complementarity in which each network is uniquely enriched in signaling pathways affecting the disease. Treatment strategies combining drugs are used to decrease adverse effects and prevent resistance to drugs[318–321]. Recently, a large combinatorial cell line screening dataset has been provided for the DREAM AstraZeneca-Sanger drug combination prediction challenge, which supports the studies to predict synergistic drug combinations[322]. Several machine learning or network-based approaches are applied within this challenge[323–326]. In this study, combinatorial drug pairs are predicted by the application of several filtering on the separation scores, enriched pathways, and the number of nodes. The approach of this study can be used to decrease the number of combinatorial drug predictions. It can be coupled with the combinatorial cell line screening dataset and learning-based methods to validate the treatment outcome of drug pairs.

The second application of network-based analysis of drugs performed in this study is the analysis of cell lines in terms of drug sensitivities. Networks of three drugs that cell lines are significantly sensitive to are compared against the networks of the same drugs in different cell lines, and differences in pathway enrichments are illustrated. Moreover, sensitivity scores (z-score) of drugs are collected from GDSC[292], and differences between z-scores are compared against separation scores of network pairs. Regression analysis revealed that separation score increases with the increase in the differences in z-scores.

Overall, network reconstruction of cell line perturbation with the multi-omics integration holds a significant potential to find pharmacodynamic similarity of drugs, predict drug sensitivity and infer combinatorial drugs for cancer. This approach can be applied to both patient-specific data and variable disease conditions.

CHAPTER 6

CONCLUSION

Small molecule-targeted anti-cancer drugs constitute a large group of therapeutics in cancer since the first small molecule drug, imatinib, was approved by the US Food and Drug Administration (FDA) in 2001. Eighty-nine small-molecule targeted antitumor drugs have been approved by the US FDA and the National Medical Products Administration (NMPA) of China by the end of 2020, and many more are already in clinical trials. However, these drugs have some problems, such as drug resistance and low response rate[327]. To overcome these challenges in cancer therapy, understanding the molecular modulations resulting from small molecule perturbations of particular cell types is very important. With a proper classification of drugs based on their omic outputs, the desired activity within the cell in disease conditions can be achieved by the selection of the correct drug or drug combinations. There are studies aiming to prioritize small molecules for drug repurposing strategies[328–330]. For example, Lv et al. 2012[328] used a mathematical model to incorporate breast cancer-related pathways collected using gene set enrichment analysis (GSEA) and small molecule-related pathways collected using transcriptomics data of small molecules applied on breast cancer cell lines. However, these kinds of studies do not involve omics relationships and are only applicable to a generalized understanding of small molecule actions in a specific cancer type. As biological systems have several molecular functions, including different bio-entity interactions, the use of variable omic data such as transcriptome, proteome, metabolome, and drug targetome becomes crucial in modeling the dynamics of perturbations. Therefore, this study aimed to reconstruct the most biologically relevant networks of drug-treated cell lines and use these networks for stratification of the effectiveness of the drugs at the network level beyond the list of altered molecules.

250 protein-protein interaction networks are reconstructed in this thesis study from six cell lines and 70 drugs. First, all proteins observed in at least one network are collected, and the most frequent 100 proteins are analyzed for their molecular function, biological process, cellular compartment, and KEGG pathway enrichment. These functional annotations revealed that the proteins observed in networks are consistently related to cancer-related functions, and pathways, as expected since 236 out of 250 networks, are reconstructed from cancer cell lines. After, the separation score method is applied to build

network distance matrices of all pair comparisons and cell line-specific comparisons. Also, KEGG pathway enrichments of all networks are analyzed using DAVID source code[246].

Modulated subnetworks resulting from the small molecule perturbations of cell lines are reconstructed by integrating available transcriptomic and phosphoproteomic data with drug targetome and human interactome. The human interactome is modified such that a node similarity-based link prediction method enriched the interactome for false negatives, and some filtering was performed to decrease possible false positives. Rigorous topology and pathway level comparison of networks revealed several significant findings. First, chemically and functionally different drugs may modulate overlapping networks. Second, drugs targeting the same group of target proteins with different selectivity ratios may modulate distant networks. Third, network reconstruction with multi-omic integration helped reveal cell line and drug-specific hidden modulated pathways and the connection between them. Finally, the methodology used in this study can be applied to several drug-disease interaction studies, such as combinatorial drug discovery and cell line sensitivity.

Although the network-based analysis of small molecule perturbations of different cell lines revealed several important findings for discrimination of drugs with the same MoA and inference of cellular activity similarity of drugs with different chemical structures and functional properties, there were some limitations. Since cancer is a heterogeneous disease across the genetic background, tumor and tissue types, network reconstruction from cancer cell models does not entirely capture the expected outcomes of the small molecule on the given cancer type for all possible conditions. These findings only provide an insight into the similarities in drug modulations. The strength and direction of the modulation may change depending on the dosage and duration of the drug treatment and other factors related to the cellular states.

For a more detailed condition-specific analysis, mutation profiles of cell lines can be incorporated. Moreover, a more stringent prize determination procedure can be performed. For example, target selectivity of drugs can be reflected in the prizes of proteins in the drug targetome for a more fine-tuned network reconstruction protocol. Furthermore, learning-based methods can be used to perform predictive analysis for the various application areas of network-based perturbation analyses, such as combinatorial drug treatment and cell line specificity. Combinatorial drug pairs predicted in this study (Appendix D) are proposed as a hypothesis such that given the modulation of each drug on the specific cancer types, they may have a synergistic effect to treat the disease; however, their actual combinatorial application to the same cancer cell line is not currently analyzed. The corresponding transcriptomic and phosphoproteomic data of combinatorial application of drug pairs should be further analyzed to have a more detailed investigation and get more accurate predictions.

REFERENCES

1. Zhang W, Li F, Nie L. Integrating multiple “omics” analysis for microbial biology: Application and methodologies. *Microbiology*. 2010;156(2):287-301. doi:10.1099/mic.0.034793-0
2. Koek MM, Jellema RH, van der Greef J, Tas AC, Hankemeier T. Quantitative metabolomics based on gas chromatography mass spectrometry: Status and perspectives. *Metabolomics*. 2011;7(3):307-328. doi:10.1007/s11306-010-0254-3
3. Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res*. 2012;40(9):3785-3799. doi:10.1093/nar/gkr1265
4. Waters KM, Pounds JG, Thrall BD. Data merging for integrated microarray and proteomic analysis. *Briefings Funct Genomics Proteomics*. 2006;5(4):261-272. doi:10.1093/bfgp/ell019
5. Joyce AR, Palsson B. The model organism as a system: Integrating “omics” data sets. *Nat Rev Mol Cell Biol*. 2006;7(3):198-210. doi:10.1038/nrm1857
6. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*. 2015;16(2):85-97. doi:10.1038/nrg3868
7. Rajasundaram D, Selbig J. More effort - more results: Recent advances in integrative “omics” data analysis. *Curr Opin Plant Biol*. 2016;30:57-61. doi:10.1016/j.pbi.2015.12.010
8. Hood L, Balling R, Auffray C. Revolutionizing medicine in the 21st century through systems approaches. *Biotechnol J*. 2012;7(8):992-1001. doi:10.1002/biot.201100306
9. Westont AD, Hood L. Systems Biology, Proteomics, and the Future of Health Care: Toward Predictive, Preventative, and Personalized Medicine. *J Proteome Res*. 2004;3(2):179-196. doi:10.1021/pr0499693
10. Spielmann N, Wong D. Saliva: Diagnostics and therapeutic perspectives. *Oral Dis*. 2011;17(4):345-354. doi:10.1111/j.1601-0825.2010.01773.x

11. Katoh M, Katoh M. Bioinformatics for cancer management in the post-genome era. *Technol Cancer Res Treat*. 2006;5(2):169-175. doi:10.1177/153303460600500208

12. Hayhoe FGJ. Classification of acute leukaemias. *Blood Rev*. 1988;2(3):186-193. doi:https://doi.org/10.1016/0268-960X(88)90024-0

13. H Yan, ZG Peng, YL Wu, et al. Hypoxia-simulating agents and selective stimulation of arsenic trioxide-induced growth arrest and cell differentiation in acute promyelocytic leukemic cells. *Haematologica*. 2005;90(12 SE-Articles):1607-1616. doi:10.3324/%x

14. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J Clin Oncol*. 2014;5(3):412-424. doi:10.5306/wjco.v5.i3.412

15. Hoadley KA, Yau C, Wolf DM, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. 2015;158(4):929-944. doi:10.1016/j.cell.2014.06.049.Multi-platform

16. Sefcik LS, Wilson JL, Papin JA, Botchwey EA. Harnessing systems biology approaches to engineer functional microvascular networks. *Tissue Eng - Part B Rev*. 2010;16(3):361-370. doi:10.1089/ten.teb.2009.0611

17. Gabhann F Mac, Annex BH, Popel AS. Gene therapy from the perspective of systems biology. *Curr Opin Mol Ther*. 2010;12(5):570-577.

18. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*. 2015;16(2):85-97. doi:10.1038/nrg3868

19. Neuman RJ, Sung YJ. Multistage analysis strategies for genome-wide association studies: Summary of group 3 contributions to genetic analysis workshop 16. *Genet Epidemiol*. 2009;33(SUPPL. 1):1-8. doi:10.1002/gepi.20467

20. Holzinger ER, Dudek SM, Frase AT, Fridley B, Chalise P, Ritchie MD. Comparison of methods for meta-dimensional data analysis using in silico and biological data sets. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2012;7246 LNCS:134-143. doi:10.1007/978-3-642-29066-4_12

21. Holzinger ER, Dudek SM, Frase AT, Krauss RM, Medina MW, Ritchie MD. ATHENA: A tool for meta-dimensional analysis applied to genotypes and gene expression data to predict HDL cholesterol levels. *Pacific Symp Biocomput*. 2013;065962:385-396. doi:10.1142/9789814447973_0038

22. Sciacovelli M, Gonçalves E, Johnson TI, et al. Fumarate is an epigenetic modifier that elicits epithelial-to-mesenchymal transition. *Nature*. 2016;537(7621):544-547. doi:10.1038/nature19353

23. Kawata K, Hatano A, Yugi K, et al. Trans-omic Analysis Reveals Selective

Responses to Induced and Basal Insulin across Signaling, Transcriptional, and Metabolic Networks. *iScience*. 2018;7:212-229. doi:10.1016/j.isci.2018.07.022

24. Vitrinel B, Koh HWL, Kar FM, et al. Exploiting interdata relationships in next-generation proteomics analysis. *Mol Cell Proteomics*. 2019;18(8):S5-S14. doi:10.1074/mcp.MR118.001246

25. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136(5):E359-E386. doi:10.1002/ijc.29210

26. Bedard PL, Hansen AR, Ratain MJ, Siu LL. Tumour heterogeneity in the clinic. *Nature*. 2013;501(7467):355-364. doi:10.1038/nature12627

27. Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*. 2013;501(7467):338-345. doi:10.1038/nature12625

28. GLOBOCAN. The Global Cancer Observatory - All cancers. *Int Agent Res Cancer - WHO*. 2020;419:199-200. https://gco.iarc.fr/today/home.

29. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;0(0):1-41. doi:10.3322/caac.21660

30. Ideker T, Sharan R. Protein networks in diseases. *Genome Res*. 2008;1(2):644-652. doi:10.1101/gr.071852.107.Freely

31. Sarkar S, Horn G, Moulton K, et al. Cancer development, progression, and therapy: An epigenetic overview. *Int J Mol Sci*. 2013;14(10):21087-21113. doi:10.3390/ijms141021087

32. Chin L, Andersen JN, Futreal PA. Cancer genomics: From discovery science to personalized medicine. *Nat Med*. 2011;17(3):297-303. doi:10.1038/nm.2323

33. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*. 2013;45(10):1113-1120. doi:10.1038/ng.2764

34. Hudson TJ, Anderson W, Artez A, et al. International network of cancer genome projects. *Nature*. 2010;464(7291):993-998. doi:10.1038/nature08987

35. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603-607. doi:10.1038/nature11003

36. Kohl M, Megger DA, Trippler M, et al. A practical data processing workflow for multi-OMICS projects. *Biochim Biophys Acta - Proteins Proteomics*. 2014;1844(1 PART A):52-62. doi:10.1016/j.bbapap.2013.02.029

37. Chawade A, Alexandersson E, Levander F. Normalyzer: A tool for rapid evaluation of normalization methods for omics data sets. *J Proteome Res*. 2014;13(6):3114-

3120. doi:10.1021/pr401264n

38. Litichevskiy L, Peckner R, Abelin JG, et al. A Library of Phosphoproteomic and Chromatin Signatures for Characterizing Cellular Responses to Drug Perturbations. *Cell Syst*. 2018;6(4):424-443.e7. doi:10.1016/j.cels.2018.03.012

39. Tuncbag N, Gosline SJC, Kedaigle A, Soltis AR, Gitter A, Fraenkel E. Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLoS Comput Biol*. 2016;12(4). doi:10.1371/journal.pcbi.1004879

40. Madeddu L, Stilo G, Velardi P. Network-based methods for disease-gene prediction. 2019. http://arxiv.org/abs/1902.10117.

41. Wang X, Gulbahce N, Yu H. Network-based methods for human disease gene prediction. *Brief Funct Genomics*. 2011;10(5):280-293. doi:10.1093/bfgp/elr024

42. Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*. 2001;2:343-372. doi:10.1146/annurev.genom.2.1.343

43. Liu C, Ma Y, Zhao J, et al. Computational network biology: Data, models, and applications. *Phys Rep*. 2020;846:1-66. doi:10.1016/j.physrep.2019.12.004

44. Junker BH, Schreiber F. *Analysis of Biological Networks*. Wiley; 2008. https://books.google.com.tr/books?id=2DloLXaXSNgC.

45. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: Data inteHecker, M., Lambeck, S., Toepfer, S., van Someren, E., & Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models-A review. BioSystems, 96(1), 86–103. https://doi.org/10.1016/j.bi. *BioSystems*. 2009;96(1):86-103. doi:10.1016/j.biosystems.2008.12.004

46. Delgado FM, Gómez-Vela F. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artif Intell Med*. 2019;95(June 2018):133-145. doi:10.1016/j.artmed.2018.10.006

47. Guan Y, Gorenshteyn D, Burmeister M, et al. Tissue-Specific Functional Networks for Prioritizing Phenotype and Disease Genes. *PLoS Comput Biol*. 2012;8(9). doi:10.1371/journal.pcbi.1002694

48. Lee JH, Park YR, Jung M, Lim SG. Gene regulatory network analysis with drug sensitivity reveals synergistic effects of combinatory chemotherapy in gastric cancer. *Sci Rep*. 2020;10(1):1-10. doi:10.1038/s41598-020-61016-z

49. Hollander M, Hamed M, Helms V, Neininger K. MutaNET: A tool for automated analysis of genomic mutations in gene regulatory networks. *Bioinformatics*. 2018;34(5):864-866. doi:10.1093/bioinformatics/btx687

50. Chen X, Li M, Zheng R, et al. A novel method of gene regulatory network structure

inference from gene knock-out expression data. *Tsinghua Sci Technol*. 2019;24(4):446-455. doi:10.26599/TST.2018.9010097

51. Xie Y, Wang R, Zhu J. Construction of breast cancer gene regulatory networks and drug target optimization. *Arch Gynecol Obstet*. 2014;290(4):749-755. doi:10.1007/s00404-014-3264-y

52. Bansal M, Gatta G Della, di Bernardo D. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*. 2006;22(7):815-822. doi:10.1093/bioinformatics/btl003

53. Emmert-Streib F, Dehmer M, Haibe-Kains B. Gene regulatory networks and their applications: Understanding biological and medical problems in terms of networks. *Front Cell Dev Biol*. 2014;2(AUG):1-7. doi:10.3389/fcell.2014.00038

54. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*. 2004;32(Database issue):D91-D94. doi:10.1093/nar/gkh012

55. Wingender E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief Bioinform*. 2008;9(4):326-332. doi:10.1093/bib/bbn016

56. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*. 2006;34(Database issue):D535-9. doi:10.1093/nar/gkj109

57. Gama-Castro S, Salgado H, Peralta-Gil M, et al. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res*. 2011;39(Database issue):D98-105. doi:10.1093/nar/gkq1110

58. Nooren IMA, Thornton JM. Diversity of protein - protein interactions. 2003;22(14).

59. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C. Transient Protein-Protein Interactions : Structural , Functional , and Network Properties. *Struct Des*. 2010;18(10):1233-1243. doi:10.1016/j.str.2010.08.007

60. Brückner A, Polge C, Lentze N, Auerbach D, Schlattner U. Yeast Two-Hybrid , a Powerful Tool for Systems Biology. 2009:2763-2788. doi:10.3390/ijms10062763

61. Worthington AS, Rivera H, Torpey JW, Alexander MD, Burkart MD. Mechanism-based protein cross-linking probes to investigate carrier protein-mediated biosynthesis. *ACS Chem Biol*. 2006;1(11):687-691. doi:10.1021/cb6003965

62. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. *Nucleic Acids Res*. 2000;28(1):289-291. doi:10.1093/nar/28.1.289

63. Bader GD, Donaldson I, Wolting C, Ouellette BFF, Pawson T, Hogue CW V.

BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res*. 2001;29(1):242-245. doi:10.1093/nar/29.1.242

64. Peri S, Navarro JD, Amanchy R, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*. 2003;13(10):2363-2371. doi:10.1101/gr.1680803

65. Thul PJ, Åkesson L, Wiking M, et al. A subcellular map of the human proteome. *Science*. 2017;356(6340). doi:10.1126/science.aal3321

66. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607-D613. doi:10.1093/nar/gky1131

67. Hatzimanikatis V, Li C, Ionita JA, Broadbelt LJ. Metabolic networks: Enzyme function and metabolite structure. *Curr Opin Struct Biol*. 2004;14(3):300-306. doi:10.1016/j.sbi.2004.04.004

68. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27-30. doi:10.1093/nar/28.1.27

69. Fabregat A, Jupe S, Matthews L, et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res*. 2018;46(D1):D649-D655. doi:10.1093/nar/gkx1132

70. Bairoch A. The ENZYME data bank. *Nucleic Acids Res*. 1994;22(17):3626-3627.

71. Scheer M, Grote A, Chang A, et al. BRENDA, the enzyme information system in 2011. *Nucleic Acids Res*. 2011;39(Database issue):D670-6. doi:10.1093/nar/gkq1089

72. Weng G, Bhalla US, Iyengar R. Complexity in biological signaling systems. *Science*. 1999;284(5411):92-96. doi:10.1126/science.284.5411.92

73. Azeloglu EU, Iyengar R. Signaling networks: Information flow, computation, and decision making. *Cold Spring Harb Perspect Biol*. 2015;7(4):1-14. doi:10.1101/cshperspect.a005934

74. Young HL, Rowling EJ, Bugatti M, et al. An adaptive signaling network in melanoma inflammatory niches confers tolerance to MAPK signaling inhibition. 2017:1691-1710.

75. Zaman N, Li L, Jaramillo ML, et al. Signaling Network Assessment of Mutations and Copy Number Variations Predict Breast Cancer Subtype-Specific Drug Targets. *Cell Rep*. 2013;5(1):216-223. doi:https://doi.org/10.1016/j.celrep.2013.08.028

76. Mansoori F, Rahgozar M, Kavousi K. FoPA: identifying perturbed signaling pathways in clinical conditions using formal methods. *BMC Bioinformatics*. 2019;20(1):92. doi:10.1186/s12859-019-2635-6

77. Eshleman N, Luo X, Capaldi A, Buchan JR. Alterations of signaling pathways in response to chemical perturbations used to measure mRNA decay rates in yeast. *RNA*. 2020;26(1):10-18. doi:10.1261/rna.072892.119

78. Garrett JT, Arteaga CL. Resistance to HER2-directed antibodies and tyrosine kinase inhibitors: mechanisms and clinical implications. *Cancer Biol Ther*. 2011;11(9):793-800. doi:10.4161/cbt.11.9.15045

79. Krull M, Pistor S, Voss N, et al. TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res*. 2006;34(Database issue):D546-D551. doi:10.1093/nar/gkj107

80. Weirauch MT. Gene Coexpression Networks for the Analysis of DNA Microarray Data. *Appl Stat Netw Biol*. April 2011:215-250. doi:https://doi.org/10.1002/9783527638079.ch11

81. Chen J, Yu L, Zhang S, Chen X. Network Analysis-Based Approach for Exploring the Potential Diagnostic Biomarkers of Acute Myocardial Infarction . *Front Physiol* . 2016;7:615. https://www.frontiersin.org/article/10.3389/fphys.2016.00615.

82. Niu X, Zhang J, Zhang L, et al. Weighted Gene Co-Expression Network Analysis Identifies Critical Genes in the Development of Heart Failure After Acute Myocardial Infarction . *Front Genet* . 2019;10:1214. https://www.frontiersin.org/article/10.3389/fgene.2019.01214.

83. Tang J, Kong D, Cui Q, et al. Prognostic Genes of Breast Cancer Identified by Gene Co-expression Network Analysis. *Front Oncol*. 2018;8:374. doi:10.3389/fonc.2018.00374

84. Liu X, Hu A-X, Zhao J-L, Chen F-L. Identification of Key Gene Modules in Human Osteosarcoma by Co-Expression Analysis Weighted Gene Co-Expression Network Analysis (WGCNA). *J Cell Biochem*. 2017;118(11):3953-3959. doi:10.1002/jcb.26050

85. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science*. 2003;302(5643):249-255. doi:10.1126/science.1087447

86. Obayashi T, Kagaya Y, Aoki Y, Tadaka S, Kinoshita K. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res*. 2019;47(D1):D55-D62. doi:10.1093/nar/gky1155

87. van Dam S, Craig T, de Magalhães JP. GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Res*. 2015;43(Database issue):D1124-32. doi:10.1093/nar/gku1042

88. Shcherbinin D, Veselovsky A. Analysis of Protein Structures Using Residue Interaction Networks BT - Structural Bioinformatics: Applications in Preclinical

Drug Discovery Process. In: Mohan CG, ed. Cham: Springer International Publishing; 2019:55-69. doi:10.1007/978-3-030-05282-9_3

89. del Sol A, Fujihashi H, Amoros D, Nussinov R. Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families. *Protein Sci*. 2006;15(9):2120-2128. doi:10.1110/ps.062249106

90. Piovesan D, Minervini G, Tosatto SCE. The RING 2.0 web server for high quality residue interaction networks. *Nucleic Acids Res*. 2016;44(W1):W367-W374. doi:10.1093/nar/gkw315

91. Soundararajan V, Raman R, Raguram S, Sasisekharan V, Sasisekharan R. Atomic Interaction Networks in the Core of Protein Domains and Their Native Folds. *PLoS One*. 2010;5(2):e9391. https://doi.org/10.1371/journal.pone.0009391.

92. Swint-Kruse L. Using Networks To Identify Fine Structural Differences between Functionally Distinct Protein States. *Biochemistry*. 2004;43(34):10886-10895. doi:10.1021/bi049450k

93. Hu G, Yan W, Zhou J, Shen B. Residue interaction network analysis of Dronpa and a DNA clamp. *J Theor Biol*. 2014;348:55-64. doi:10.1016/j.jtbi.2014.01.023

94. Xue W, Jiao P, Liu H, Yao X. Molecular modeling and residue interaction network studies on the mechanism of binding and resistance of the HCV NS5B polymerase mutants to VX-222 and ANA598. *Antiviral Res*. 2014;104:40-51. doi:https://doi.org/10.1016/j.antiviral.2014.01.006

95. Stetz G, Verkhivker GM. Computational Analysis of Residue Interaction Networks and Coevolutionary Relationships in the Hsp70 Chaperones: A Community-Hopping Model of Allosteric Regulation and Communication. *PLOS Comput Biol*. 2017;13(1):e1005299. https://doi.org/10.1371/journal.pcbi.1005299.

96. Brysbaert G, Blossey R, Lensink MF. The Inclusion of Water Molecules in Residue Interaction Networks Identifies Additional Central Residues . *Front Mol Biosci* . 2018;5:88. https://www.frontiersin.org/article/10.3389/fmolb.2018.00088.

97. Contreras-Riquelme S, Garate J-A, Perez-Acle T, Martin AJM. RIP-MD: a tool to study residue interaction networks in protein molecular dynamics. *PeerJ*. 2018;6:e5998-e5998. doi:10.7717/peerj.5998

98. Aydınkal RM, Serçinoğlu O, Ozbek P. ProSNEx: a web-based application for exploration and analysis of protein structures using network formalism. *Nucleic Acids Res*. 2019;47(W1):W471-W476. doi:10.1093/nar/gkz390

99. Brysbaert G, Lorgouilloux K, Vranken WF, Lensink MF. RINspector: a Cytoscape app for centrality analyses and DynaMine flexibility prediction. *Bioinformatics*. 2018;34(2):294-296. doi:10.1093/bioinformatics/btx586

100. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*.

2003;13(11):2498-2504. doi:10.1101/gr.1239303

101. Tieri P, Farina L, Petti M, Astolfi L, Paci P, Castiglione F. Network Inference and Reconstruction in Bioinformatics. In: *Encyclopedia of Bioinformatics and Computational Biology*. ; 2019.

102. Patra S, Mohapatra A. Review of tools and algorithms for network motif discovery in biological networks. *IET Syst Biol*. 2020;14(4):171-189. doi:https://doi.org/10.1049/iet-syb.2020.0004

103. Thompson D, Regev A, Roy S. *Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution*. Vol 31.; 2015. doi:10.1146/annurev-cellbio-100913-012908

104. Gao S, Wu Z, Feng X, Kajigaya S, Wang X, Young NS. Comprehensive network modeling from single cell RNA sequencing of human and mouse reveals well conserved transcription regulation of hematopoiesis. *BMC Genomics*. 2020;21(11):849. doi:10.1186/s12864-020-07241-2

105. Zickenrott S, Angarica VE, Upadhyaya BB, del Sol A. Prediction of disease–gene–drug relationships following a differential network analysis. *Cell Death Dis*. 2016;7(1):e2040-e2040. doi:10.1038/cddis.2015.393

106. Emig D, Ivliev A, Pustovalova O, et al. Drug Target Prediction and Repositioning Using an Integrated Network-Based Approach. *PLoS One*. 2013;8(4):e60618. https://doi.org/10.1371/journal.pone.0060618.

107. Kauffman S. Homeostasis and differentiation in random genetic control networks. *Nature*. 1969;224(5215):177-178. doi:10.1038/224177a0

108. Poret A, Sousa CM, Boissel J. Enhancing Boolean networks with fuzzy operators and edge tuning. *arXiv Mol Networks*. 2014.

109. Xu H, Wang S. Fuzzy Logical on Boolean Networks as Model of Gene Regulatory Networks. In: *2009 International Joint Conference on Artificial Intelligence*. ; 2009:501-505. doi:10.1109/JCAI.2009.78

110. Akutsu T, Miyano S, Kuhara S. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*. 2000;16(8):727-734. doi:10.1093/bioinformatics/16.8.727

111. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002;18(2):261-274. doi:10.1093/bioinformatics/18.2.261

112. Leifeld T, Zhang Z, Zhang P. Identification of Boolean Network Models From Time Series Data Incorporating Prior Knowledge . *Front Physiol* . 2018;9:695. https://www.frontiersin.org/article/10.3389/fphys.2018.00695.

113. Schwab JD, Kühlwein SD, Ikonomi N, Kühl M, Kestler HA. Concepts in Boolean network modeling: What do they all mean? *Comput Struct Biotechnol J*.

2020;18:571-582. doi:https://doi.org/10.1016/j.csbj.2020.03.001

114. Krawitz P, Shmulevich I. Boolean Modeling of Biological Networks BT - Encyclopedia of Complexity and Systems Science. In: Meyers RA, ed. New York, NY: Springer New York; 2009:599-608. doi:10.1007/978-0-387-30440-3_40

115. Bornholdt S. Boolean network models of cellular regulation: Prospects and limitations. *J R Soc Interface*. 2008;5(SUPPL. 1). doi:10.1098/rsif.2008.0132.focus

116. Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*. 1998:18-29.

117. Helikar T, Rogers JA. ChemChains: a platform for simulation and analysis of biochemical networks aimed to laboratory scientists. *BMC Syst Biol*. 2009;3(1):58. doi:10.1186/1752-0509-3-58

118. Paulevé L. Pint: A Static Analyzer for Transient Dynamics of Qualitative Networks with IPython Interface BT - Computational Methods in Systems Biology. In: Feret J, Koeppl H, eds. Cham: Springer International Publishing; 2017:309-316.

119. Zheng J, Zhang D, Przytycki PF, Zielinski R, Capala J, Przytycka TM. SimBoolNet—a Cytoscape plugin for dynamic simulation of signaling networks. *Bioinformatics*. 2010;26(1):141-142. doi:10.1093/bioinformatics/btp617

120. Müssel C, Hopfensitz M, Kestler HA. BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics*. 2010;26(10):1378-1380. doi:10.1093/bioinformatics/btq124

121. Klamt S, Saez-Rodriguez J, Gilles ED. Structural and functional analysis of cellular networks with CellNetAnalyzer. *BMC Syst Biol*. 2007;1:2. doi:10.1186/1752-0509-1-2

122. Chai LE, Loh SK, Low ST, Mohamad MS, Deris S, Zakaria Z. A review on the computational approaches for gene regulatory network construction. *Comput Biol Med*. 2014;48(1):55-65. doi:10.1016/j.compbiomed.2014.02.011

123. Alakwaa FM. Modeling of Gene Regulatory Networks: A Literature Review. *J Comput Syst Biol*. 2015;1(1). doi:10.15744/2455-7625.1.102

124. Sriram A. Predicting Gene Relations Using Bayesian Networks. 2011. http://rave.ohiolink.edu/etdc/view?acc_num=akron1302619630.

125. Deeter A, Dalman M, Haddad J, Duan Z-H. Inferring gene and protein interactions using PubMed citations and consensus Bayesian networks. *PLoS One*. 2017;12(10):e0186004. https://doi.org/10.1371/journal.pone.0186004.

126. Ziebarth JD, Bhattacharya A, Cui Y. Bayesian Network Webserver: a comprehensive tool for biological network modeling. *Bioinformatics*. 2013;29(21):2801-2803. doi:10.1093/bioinformatics/btt472

127. Frolova A, Wilczyński B. Distributed Bayesian networks reconstruction on the whole genome scale. *PeerJ*. 2018;6:e5692. doi:10.7717/peerj.5692

128. Nagarajan R, Scutari M. *Bayesian Networks in {R} with Applications in Systems Biology*. New York: Springer; 2013. doi:10.1007/978-1-4614-6446-4

129. Koh G, Lee D-Y. Mathematical modeling and sensitivity analysis of the integrated TNFα-mediated apoptotic pathway for identifying key regulators. *Comput Biol Med*. 2011;41(7):512-528. doi:10.1016/j.compbiomed.2011.04.017

130. Peng H, Peng T, Wen J, et al. Characterization of p38 MAPK isoforms for drug resistance study using systems biology approach. *Bioinformatics*. 2014;30(13):1899-1907. doi:10.1093/bioinformatics/btu133

131. Shao H, Peng T, Ji Z, Su J, Zhou X. Systematically studying kinase inhibitor induced signaling network signatures by integrating both therapeutic and side effects. *PLoS One*. 2013;8(12):e80832. doi:10.1371/journal.pone.0080832

132. Sun X, Bao J, Nelson KC, Li KC, Kulik G, Zhou X. Systems modeling of anti-apoptotic pathways in prostate cancer: psychological stress triggers a synergism pattern switch in drug combination therapy. *PLoS Comput Biol*. 2013;9(12):e1003358. doi:10.1371/journal.pcbi.1003358

133. Ji Z, Yan K, Li W, Hu H, Zhu X. Mathematical and Computational Modeling in Complex Biological Systems. *Biomed Res Int*. 2017;2017:5958321. doi:10.1155/2017/5958321

134. Zinner RG, Barrett BL, Popova E, et al. Algorithmic guided screening of drug combinations of arbitrary size for activity against cancer cells. *Mol Cancer Ther*. 2009;8(3):521-532. doi:10.1158/1535-7163.MCT-08-0937

135. Peng T, Peng H, Choi DS, Su J, Chang C-CJ, Zhou X. Modeling cell-cell interactions in regulating multiple myeloma initiating cell fate. *IEEE J Biomed Heal informatics*. 2014;18(2):484-491. doi:10.1109/JBHI.2013.2281774

136. Laguna M, Martí R. Experimental Testing of Advanced Scatter Search Designs for Global Optimization of Multimodal Functions. *J Glob Optim*. 2005;33(2):235-255. doi:10.1007/s10898-004-1936-z

137. Chen T, He HL, Church GM. Modeling gene expression with differential equations. *Pac Symp Biocomput*. 1999:29-40.

138. Ando S, Sakamoto E, Iba H. Evolutionary modeling and inference of gene network. *Inf Sci (Ny)*. 2002;145(3):237-259. doi:https://doi.org/10.1016/S0020-0255(02)00235-9

139. Vohradsky J. Neural model of the genetic network. *J Biol Chem*. 2001;276(39):36168-36173. doi:10.1074/jbc.M104391200

140. Biswas S, Acharyya S. Neural model of gene regulatory network: a survey on supportive meta-heuristics. *Theory Biosci*. 2016;135(1-2):1-19.

doi:10.1007/s12064-016-0224-z

141. Ressom HW, Zhang Y, Xuan J, Wang Y, Clarke R. Inference of Gene Regulatory Networks from Time Course Gene Expression Data Using Neural Networks and Swarm Intelligence. In: *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*. ; 2006:1-8. doi:10.1109/CIBCB.2006.330969

142. Kordmahalleh MM, Sefidmazgi MG, Harrison SH, Homaifar A. Identifying time-delayed gene regulatory networks via an evolvable hierarchical recurrent neural network. *BioData Min*. 2017;10(1):29. doi:10.1186/s13040-017-0146-4

143. Rubiolo M, Milone DH, Stegmayer G. Extreme learning machines for reverse engineering of gene regulatory networks from expression time series. *Bioinformatics*. 2018;34(7):1253-1260. doi:10.1093/bioinformatics/btx730

144. Budden DM, Crampin EJ. Information theoretic approaches for inference of biological networks from continuous-valued data. *BMC Syst Biol*. 2016;10(1):89. doi:10.1186/s12918-016-0331-y

145. BUTTE AJ, KOHANE IS. MUTUAL INFORMATION RELEVANCE NETWORKS: FUNCTIONAL GENOMIC CLUSTERING USING PAIRWISE ENTROPY MEASUREMENTS. In: *Biocomputing 2000*. WORLD SCIENTIFIC; 1999:418-429. doi:doi:10.1142/9789814447331_0040

146. Margolin AA, Nemenman I, Basso K, et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*. 2006;7(1):S7. doi:10.1186/1471-2105-7-S1-S7

147. Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R. DREAM3: Network Inference Using Dynamic Context Likelihood of Relatedness and the Inferelator. *PLoS One*. 2010;5(3):e9803. https://doi.org/10.1371/journal.pone.0009803.

148. Olsen C, Meyer PE, Bontempi G. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *EURASIP J Bioinform Syst Biol*. 2009;2009(1):308959. doi:10.1155/2009/308959

149. Gómez-Vela F, Barranco CD, Díaz-Díaz N. Incorporating biological knowledge for construction of fuzzy networks of gene associations. *Appl Soft Comput*. 2016;42:144-155. doi:https://doi.org/10.1016/j.asoc.2016.01.014

150. Cavill R, Jennen D, Kleinjans J, Briedé JJ. Transcriptomic and metabolomic data integration. *Brief Bioinform*. 2016;17(5):891-901. doi:10.1093/bib/bbv090

151. WOLD H. Estimation of principal components and related models by iterative least squares. *Multivar Anal*. 1966:391-420. https://ci.nii.ac.jp/naid/20001378860/en/.

152. Lê Cao KA, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol*. 2008;7(1). doi:10.2202/1544-6115.1390

153. HOTELLING H. RELATIONS BETWEEN TWO SETS OF VARIATES*. *Biometrika*. 1936;28(3-4):321-377. doi:10.1093/biomet/28.3-4.321

154. Lê Cao K-A, González I, Déjean S. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*. 2009;25(21):2855-2856. doi:10.1093/bioinformatics/btp515

155. Singh A, Shannon CP, Gautier B, et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*. 2019;35(17):3055-3062. doi:10.1093/bioinformatics/bty1054

156. Conesa A, Prats-Montalbán JM, Tarazona S, Nueda MJ, Ferrer A. A multiway approach to data integration in systems biology based on Tucker3 and N-PLS. *Chemom Intell Lab Syst*. 2010;104(1):101-111. doi:https://doi.org/10.1016/j.chemolab.2010.06.004

157. Li W, Zhang S, Liu C-C, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*. 2012;28(19):2458-2466. doi:10.1093/bioinformatics/bts476

158. Löfstedt T, Hanafi M, Mazerolles G, Trygg J. OnPLS path modelling. *Chemom Intell Lab Syst*. 2012;118(Complete):139-149. doi:10.1016/j.chemolab.2012.08.009

159. Löfstedt T, Hoffman D, Trygg J. Global, local and unique decompositions in OnPLS for multiblock data analysis. *Anal Chim Acta*. 2013;791:13-24. doi:10.1016/j.aca.2013.06.026

160. Bouhaddani S El, Houwing-Duistermaat J, Salo P, Perola M, Jongbloed G, Uh H-W. Evaluation of O2PLS in Omics data integration. *BMC Bioinformatics*. 2016;17 Suppl 2(Suppl 2):11. doi:10.1186/s12859-015-0854-z

161. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*. 2009;8(1). doi:10.2202/1544-6115.1470

162. Lin D, Zhang J, Li J, Calhoun VD, Deng H-W, Wang Y-P. Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics*. 2013;14:245. doi:10.1186/1471-2105-14-245

163. Markovsky I. *Low Rank Approximation: Algorithms, Implementation, Applications*. Springer Publishing Company, Incorporated; 2011.

164. Jolliffe I. Principal Component Analysis BT - International Encyclopedia of Statistical Science. In: Lovric M, ed. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011:1094-1096. doi:10.1007/978-3-642-04898-2_455

165. Culhane AC, Perrière G, Higgins DG. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*. 2003;4(1):59. doi:10.1186/1471-2105-4-59

166. Meng C, Kuster B, Culhane AC, Gholami AM. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*. 2014;15:162. doi:10.1186/1471-2105-15-162

167. de Tayrac M, Lê S, Aubry M, Mosser J, Husson F. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics*. 2009;10(ii):32. doi:10.1186/1471-2164-10-32

168. Pagès J. ANÁLISIS FACTORIAL MÚLTIPLE: PRINCIPALES CARACTERÍSTICAS Y SU APLICACIÓN A DATOS SENSORIALES. *Rev Colomb Estadística*. 2004;27(1 SE-):1-26. https://revistas.unal.edu.co/index.php/estad/article/view/29982.

169. Chen J, Bushman FD, Lewis JD, Wu GD, Li H. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*. 2013;14(2):244-258. doi:10.1093/biostatistics/kxs038

170. Wang B, Mezlini AM, Demir F, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods*. 2014;11(3):333-337. doi:10.1038/nmeth.2810

171. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*. 2015;31(12):i268-75. doi:10.1093/bioinformatics/btv244

172. Li W, Liu C-C, Zhang T, Li H, Waterman MS, Zhou XJ. Integrative Analysis of Many Weighted Co-Expression Networks Using Tensor Computation. *PLOS Comput Biol*. 2011;7(6):e1001106. https://doi.org/10.1371/journal.pcbi.1001106.

173. Daemen A, Gevaert O, Ojeda F, et al. A kernel-based integration of genome-wide data for clinical decision support. *Genome Med*. 2009;1(4):39. doi:10.1186/gm39

174. Mariette J, Villa-Vialaneix N. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*. 2018;34(6):1009-1015. doi:10.1093/bioinformatics/btx682

175. Akavia UD, Litvin O, Kim J, et al. An integrated approach to uncover drivers of cancer. *Cell*. 2010;143(6):1005-1017. doi:10.1016/j.cell.2010.11.013

176. Zhu J, Sova P, Xu Q, et al. Stitching together Multiple Data Dimensions Reveals Interacting Metabolomic and Transcriptomic Networks That Modulate Cell Regulation. *PLOS Biol*. 2012;10(4):e1001301. https://doi.org/10.1371/journal.pbio.1001301.

177. Zhu J, Zhang B, Smith EN, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet*. 2008;40(7):854-861. doi:10.1038/ng.167

178. Huang S, Chaudhary K, Garmire LX. More is better: Recent progress in multi-

omics data integration methods. *Front Genet*. 2017;8(JUN):1-12. doi:10.3389/fgene.2017.00084

179. Zhang S, Liu C-C, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*. 2012;40(19):9379-9391. doi:10.1093/nar/gks725

180. Shen R, Mo Q, Schultz N, et al. Integrative subtype discovery in glioblastoma using iCluster. *PLoS One*. 2012;7(4):e35236. doi:10.1371/journal.pone.0035236

181. Mo Q, Wang S, Seshan VE, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci*. 2013;110(11):4245 LP - 4250. doi:10.1073/pnas.1208949110

182. Lock EF, Hoadley KA, Marron JS, Nobel AB. JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES. *Ann Appl Stat*. 2013;7(1):523-542. doi:10.1214/12-AOAS597

183. Ray P, Zheng L, Lucas J, Carin L. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*. 2014;30(10):1370-1376. doi:10.1093/bioinformatics/btu064

184. Chen J, Zhang S. Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics*. 2016;32(11):1724-1732. doi:10.1093/bioinformatics/btw059

185. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*. 2012;28(24):3290-3297. doi:10.1093/bioinformatics/bts595

186. Cho D-Y, Przytycka TM. Dissecting cancer heterogeneity with a probabilistic genotype-phenotype model. *Nucleic Acids Res*. 2013;41(17):8011-8020. doi:10.1093/nar/gkt577

187. Yuan Y, Savage RS, Markowetz F. Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput Biol*. 2011;7(10):e1002227. doi:10.1371/journal.pcbi.1002227

188. Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics*. 2013;29(20):2610-2616. doi:10.1093/bioinformatics/btt425

189. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010;26(12):237-245. doi:10.1093/bioinformatics/btq182

190. Bonnet E, Calzone L, Michoel T. Integrative multi-omics module network inference with Lemon-Tree. *PLoS Comput Biol*. 2015;11(2):e1003983. doi:10.1371/journal.pcbi.1003983

191. Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number,

methylation and expression data. *Bioinformatics*. 2011;27(6):887-888. doi:10.1093/bioinformatics/btr019

192. Aure MR, Steinfeld I, Baumbusch LO, et al. Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data. *PLoS One*. 2013;8(1):e53014. doi:10.1371/journal.pone.0053014

193. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*. 2002;18(suppl_1):S233-S240. doi:10.1093/bioinformatics/18.suppl_1.S233

194. Ruffalo M, Koyutürk M, Sharan R. Network-Based Integration of Disparate Omic Data To Identify "Silent Players" in Cancer. *PLOS Comput Biol*. 2015;11(12):e1004595. https://doi.org/10.1371/journal.pcbi.1004595.

195. Lanckriet GRG, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. *Bioinformatics*. 2004;20(16):2626-2635. doi:10.1093/bioinformatics/bth294

196. Seoane JA, Day INM, Gaunt TR, Campbell C. A pathway-based data integration framework for prediction of disease progression. *Bioinformatics*. 2014;30(6):838-845. doi:10.1093/bioinformatics/btt610

197. Jennings EM, Morris JS, Carroll RJ, Manyam GC, Baladandayuthapani V. Bayesian methods for expression-based integration of various types of genomics data. *EURASIP J Bioinform Syst Biol*. 2013;2013(1):13. doi:10.1186/1687-4153-2013-13

198. Chari R, Coe BP, Vucic EA, Lockwood WW, Lam WL. An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Syst Biol*. 2010;4:67. doi:10.1186/1752-0509-4-67

199. Ovaska K, Laakso M, Haapa-Paananen S, et al. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome Med*. 2010;2(9):65. doi:10.1186/gm186

200. You Z-H, Yin Z, Han K, Huang D-S, Zhou X. A semi-supervised learning approach to predict synthetic genetic interactions by combining functional and topological properties of functional gene network. *BMC Bioinformatics*. 2010;11:343. doi:10.1186/1471-2105-11-343

201. Kim D, Shin H, Song YS, Kim JH. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inform*. 2012;45(6):1191-1198. doi:10.1016/j.jbi.2012.07.008

202. Mankoo PK, Shen R, Schultz N, Levine DA, Sander C. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One*. 2011;6(11):e24709. doi:10.1371/journal.pone.0024709

203. Kim D, Li R, Lucas A, Verma SS, Dudek SM, Ritchie MD. Using knowledge-

driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. *J Am Med Inform Assoc*. 2017;24(3):577-587. doi:10.1093/jamia/ocw165

204. Hasan M Al, Zaki MJ. A Survey of Link Prediction in Social Networks BT - Social Network Data Analytics. In: Aggarwal CC, ed. Boston, MA: Springer US; 2011:243-275. doi:10.1007/978-1-4419-8462-3_9

205. Symeonidis P, Perentis C. Link Prediction in Multi-modal Social Networks BT - Machine Learning and Knowledge Discovery in Databases. In: Calders T, Esposito F, Hüllermeier E, Meo R, eds. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014:147-162.

206. Huang Z, Li X, Chen H. Link Prediction Approach to Collaborative Filtering. In: *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL '05. New York, NY, USA: Association for Computing Machinery; 2005:141–142. doi:10.1145/1065385.1065415

207. Hasan M Al, Chaoji V, Salem S, Zaki M. Link Prediction Using Supervised Learning. In: *In Proc. of SDM 06 Workshop on Link Analysis, Counterterrorism and Security*. ; 2006. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.1225.

208. Ideker T, Sharan R. Protein networks in disease. *Genome Res*. 2008;18(4):644-652. doi:10.1101/gr.071852.107

209. Buchanan M, Caldarelli G, De Los Rios P, Rao F, Vendruscolo M. *Networks in Cell Biology*.; 2010. https://ui.adsabs.harvard.edu/abs/2010ncb..book.....B.

210. Kann MG. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform*. 2007;8(5):333-346. doi:10.1093/bib/bbm031

211. Albert R. Scale-free networks in cell biology. *J Cell Sci*. 2005;118(21):4947-4957. doi:10.1242/jcs.02714

212. Barabási AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. *Nat Rev Genet*. 2004;5(2):101-113. doi:10.1038/nrg1272

213. Vidal M, Cusick ME, Barabási A-L. Interactome Networks and Human Disease. *Cell*. 2011;144(6):986-998. doi:https://doi.org/10.1016/j.cell.2011.02.016

214. Sulaimany S, Khansari M, Masoudi-Nejad A. Link prediction potentials for biological networks. *Int J Data Min Bioinform*. 2018;20(2):161-184. doi:10.1504/IJDMB.2018.093684

215. Kashima H, Kato T, Yamanishi Y, Sugiyama M, Tsuda K. Link Propagation: A Fast Semi-supervised Learning Algorithm for Link Prediction. In: *Proceedings of the 2009 SIAM International Conference on Data Mining (SDM)*. Proceedings. Society for Industrial and Applied Mathematics; 2009:1100-1111.

doi:doi:10.1137/1.9781611972795.94

216. Kashima H, Abe N. A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction. In: *Sixth International Conference on Data Mining (ICDM'06)*. ; 2006:340-349. doi:10.1109/ICDM.2006.8

217. Shojaie A. Link Prediction in Biological Networks using Multi-Mode Exponential Random Graph Models. *11th Work Min Learn with Graph*. 2013. http://snap.stanford.edu/mlg2013/submissions/mlg2013_submission_24.pdf.

218. Crichton G, Guo Y, Pyysalo S, Korhonen A. Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. *BMC Bioinformatics*. 2018;19(1):176. doi:10.1186/s12859-018-2163-9

219. Katz L. A new status index derived from sociometric analysis. *Psychometrika*. 1953;18(1):39-43. doi:10.1007/BF02289026

220. Cannistraci CV, Alanis-Lobato G, Ravasi T. From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. *Sci Rep*. 2013;3. doi:10.1038/srep01613

221. Lu Y, Guo Y, Korhonen A. Link prediction in drug-target interactions network using similarity indices. *BMC Bioinformatics*. 2017;18(1):1-9. doi:10.1186/s12859-017-1460-z

222. Kovács IA, Luck K, Spirohn K, et al. Network-based prediction of protein interactions. *Nat Commun*. 2019;10(1):1240. doi:10.1038/s41467-019-09177-y

223. Pech R, Hao D, Lee Y-L, Yuan Y, Zhou T. Link prediction via linear optimization. *Phys A Stat Mech its Appl*. 2019;528:121319. doi:https://doi.org/10.1016/j.physa.2019.121319

224. Chen Y, Wang W, Liu J, Feng J, Gong X. Protein Interface Complementarity and Gene Duplication Improve Link Prediction of Protein-Protein Interaction Network. *Front Genet*. 2020;11:291. doi:10.3389/fgene.2020.00291

225. Yuen HY, Jansson J. Better Link Prediction for Protein-Protein Interaction Networks. In: *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. ; 2020:53-60. doi:10.1109/BIBE50027.2020.00017

226. Li J, Peng X, Wang J, Zhao N. A Method for Improving the Accuracy of Link Prediction Algorithms. Chakraborti A, ed. *Complexity*. 2021;2021:8889441. doi:10.1155/2021/8889441

227. Koutrouli M, Karatzas E, Paez-Espino D, Pavlopoulos GA. A Guide to Conquer the Biological Network Era Using Graph Theory. *Front Bioeng Biotechnol*. 2020;8(January):1-26. doi:10.3389/fbioe.2020.00034

228. Kunegis J, Fay D, Bauckhage C. Spectral evolution in dynamic networks. *Knowl Inf Syst*. 2013;37(1):1-36. doi:10.1007/s10115-012-0575-9

229. Guns R. Link Prediction BT - Measuring Scholarly Impact: Methods and Practice. In: Ding Y, Rousseau R, Wolfram D, eds. Cham: Springer International Publishing; 2014:35-55. doi:10.1007/978-3-319-10377-8_2

230. Lichtenwalter RN, Chawla N V. LPmade: Link prediction made easy. *J Mach Learn Res*. 2011;12:2489-2492.

231. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux G, Vaught T, Millman J, eds. *Proceedings of the 7th Python in Science Conference*. Pasadena, CA USA; 2008:11-15.

232. Abelin JG, Patel J, Lu X, et al. Reduced-representation Phosphosignatures Measured by Quantitative Targeted MS Capture Cellular States and Enable Large-scale Comparison of Drug-induced Phenotypes * □. 2016:1622-1641. doi:10.1074/mcp.M116.058354

233. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science (80-)*. 2006;313(5795):1929 LP - 1935. doi:10.1126/science.1132939

234. Corsello SM, Bittker JA, Liu Z, et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat Med*. 2017;23(4):405-408. doi:10.1038/nm.4306

235. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2017;46(D1):D1074-D1082. doi:10.1093/nar/gkx1037

236. Kim S, Chen J, Cheng T, et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res*. 2021;49(D1):D1388-D1395. doi:10.1093/nar/gkaa971

237. Virtanen P, Gommers R, Oliphant TE, et al. {SciPy} 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat Methods*. 2020;17:261-272. doi:10.1038/s41592-019-0686-2

238. Baldi P, Nasr R. When is chemical similarity significant? The statistical distribution of chemical similarity scores and its extreme values. *J Chem Inf Model*. 2010;50(7):1205-1222. doi:10.1021/ci100010v

239. Landrum G. RDKit: Open-Source Cheminformatics Software. 2016. https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4.

240. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci*. 2002;42(6):1273-1280. doi:10.1021/ci010132r

241. Johnson DS. The Prize Collecting Steiner Tree Problem : Theory and Practice. 2000.

242. Liu Z, Wu C, Miao H, Wu H. Original article RegNetwork : an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. 2015;(September). doi:10.1093/database/bav095

243. Hristov BH, Singh M. Network-Based Coverage of Mutational Profiles Report Network-Based Coverage of Mutational Profiles Reveals Cancer Genes. *Cell Syst*. 2017;5(3):221-229.e4. doi:10.1016/j.cels.2017.09.003

244. Menche J, Sharma A, Kitsak M, et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*. 2015;347(6224):1257601. doi:10.1126/science.1257601

245. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*. 2009;37(1):1-13. doi:10.1093/nar/gkn923

246. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211

247. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. *Nat Genet*. 2006;38(5):500-501. doi:10.1038/ng0506-500

248. Pinzi L, Benedetti R, Altucci L, Rastelli G. Design of Dual Inhibitors of Histone Deacetylase 6 and Heat Shock Protein 90. *ACS Omega*. 2020;5(20):11473-11480. doi:10.1021/acsomega.0c00559

249. Li X, Yin S, Sakr W, Sarkar F, Heath E, Sheng S. Maspin inhibits HDAC1-mediated Hsp90 deacetylation. *Cancer Res*. 2008;68(9 Supplement):327 LP - 327. http://cancerres.aacrjournals.org/content/68/9_Supplement/327.abstract.

250. Krämer OH, Mahboobi S, Sellmer A. Drugging the HDAC6-HSP90 interplay in malignant cells. *Trends Pharmacol Sci*. 2014;35(10):501-509. doi:10.1016/j.tips.2014.08.001

251. Chai RC, Vieusseux JL, Lang BJ, et al. Histone deacetylase activity mediates acquired resistance towards structurally diverse HSP90 inhibitors. *Mol Oncol*. 2017;11(5):567-583. doi:10.1002/1878-0261.12054

252. Glozak MA, Sengupta N, Zhang X, Seto E. Acetylation and deacetylation of non-histone proteins. *Gene*. 2005;363:15-23. doi:https://doi.org/10.1016/j.gene.2005.09.010

253. Zhang J, Xu E, Chen X. TAp73 protein stability is controlled by histone deacetylase 1 via regulation of Hsp90 chaperone function. *J Biol Chem*. 2013;288(11):7727-7737. doi:10.1074/jbc.M112.429522

254. Kovacs JJ, Murphy PJM, Gaillard S, et al. HDAC6 regulates Hsp90 acetylation and chaperone-dependent activation of glucocorticoid receptor. *Mol Cell*. 2005;18(5):601-607. doi:10.1016/j.molcel.2005.04.021

255. Iannuccelli M, Micarelli E, Surdo P Lo, et al. CancerGeneNet: linking driver genes to cancer hallmarks. *Nucleic Acids Res*. 2020;48(D1):D416-D421. doi:10.1093/nar/gkz871

256. Zhou Y, Zhu Y, Dong X, et al. Exosomes Derived from Pancreatic Cancer Cells Induce Osteoclast Differentiation Through the miR125a-5p/TNFRSF1B Pathway. *Onco Targets Ther*. 2021;14:2727-2739. doi:10.2147/OTT.S282319

257. Malvaez M, McQuown SC, Rogge GA, et al. HDAC3-selective inhibitor enhances extinction of cocaine-seeking behavior in a persistent manner. *Proc Natl Acad Sci U S A*. 2013;110(7):2647-2652. doi:10.1073/pnas.1213364110

258. Yu X, Yang F, Jiang H, Fan L. RGFP966 Suppresses Tumor Growth and Migration Through Inhibition of EGFR Expression in Hepatocellular Carcinoma Cells in vitro. *Drug Des Devel Ther*. 2020;14:121-128. doi:10.2147/DDDT.S234871

259. Quintás-Cardama A, Vaddi K, Liu P, et al. Preclinical characterization of the selective JAK1/2 inhibitor INCB018424: therapeutic implications for the treatment of myeloproliferative neoplasms. *Blood*. 2010;115(15):3109-3117. doi:10.1182/blood-2009-04-214957

260. Verstovsek S, Kantarjian H, Mesa RA, et al. Safety and efficacy of INCB018424, a JAK1 and JAK2 inhibitor, in myelofibrosis. *N Engl J Med*. 2010;363(12):1117-1127. doi:10.1056/NEJMoa1002028

261. Clark JD, Flanagan ME, Telliez J-B. Discovery and development of Janus kinase (JAK) inhibitors for inflammatory diseases. *J Med Chem*. 2014;57(12):5023-5038. doi:10.1021/jm401490p

262. Changelian PS, Flanagan ME, Ball DJ, et al. Prevention of organ allograft rejection by a specific Janus kinase 3 inhibitor. *Science*. 2003;302(5646):875-878. doi:10.1126/science.1087061

263. Geyer HL, Tibes R, Mesa RA. JAK2 inhibitors and their impact in myeloproliferative neoplasms. *Hematology*. 2012;17 Suppl 1:S129-32. doi:10.1179/102453312X13336169156375

264. Wernig G, Kharas MG, Okabe R, et al. Efficacy of TG101348, a selective JAK2 inhibitor, in treatment of a murine model of JAK2V617F-induced polycythemia vera. *Cancer Cell*. 2008;13(4):311-320. doi:10.1016/j.ccr.2008.02.009

265. Haura EB, Ricart AD, Larson TG, et al. A phase II study of PD-0325901, an oral MEK inhibitor, in previously treated patients with advanced non-small cell lung cancer. *Clin cancer Res an Off J Am Assoc Cancer Res*. 2010;16(8):2450-2457. doi:10.1158/1078-0432.CCR-09-1920

266. Bian Y, Han J, Kannabiran V, et al. MEK inhibitor PD-0325901 overcomes resistance to CK2 inhibitor CX-4945 and exhibits anti-tumor activity in head and neck cancer. *Int J Biol Sci*. 2015;11(4):411-422. doi:10.7150/ijbs.10745

267. Baselga J, Campone M, Piccart M, et al. Everolimus in postmenopausal hormone-receptor-positive advanced breast cancer. *N Engl J Med*. 2012;366(6):520-529. doi:10.1056/NEJMoa1109653

268. Royce ME, Osman D. Everolimus in the Treatment of Metastatic Breast Cancer. *Breast Cancer (Auckl)*. 2015;9:73-79. doi:10.4137/BCBCR.S29268

269. Lee L, Ito T, Jensen RT. Everolimus in the treatment of neuroendocrine tumors: efficacy, side-effects, resistance, and factors affecting its place in the treatment sequence. *Expert Opin Pharmacother*. 2018;19(8):909-928. doi:10.1080/14656566.2018.1476492

270. Pusceddu S, Verzoni E, Prinzi N, et al. Everolimus treatment for neuroendocrine tumors: latest results and clinical potential. *Ther Adv Med Oncol*. 2017;9(3):183-188. doi:10.1177/1758834016683905

271. Yao JC, Oh D-Y, Qian J, et al. Everolimus for the treatment of advanced gastrointestinal or lung nonfunctional neuroendocrine tumors in East Asian patients: a subgroup analysis of the RADIANT-4 study. *Onco Targets Ther*. 2019;12:1717-1728. doi:10.2147/OTT.S182259

272. Zhi X, Chen W, Xue F, et al. OSI-027 inhibits pancreatic ductal adenocarcinoma cell proliferation and enhances the therapeutic effect of gemcitabine both in vitro and in vivo. *Oncotarget*. 2015;6(28):26230-26241. doi:10.18632/oncotarget.4579

273. Bayat Mokhtari R, Homayouni TS, Baluch N, et al. Combination therapy in combating cancer. *Oncotarget*. 2017;8(23):38022-38043. doi:10.18632/oncotarget.16723

274. Hanahan D, Bergers G, Bergsland E. Less is more, regularly: metronomic dosing of cytotoxic drugs can target tumor angiogenesis in mice. *J Clin Invest*. 2000;105(8):1045-1047. doi:10.1172/JCI9872

275. Wang X, Zhang H, Chen X. Drug resistance and combating drug resistance in cancer. *Cancer Drug Resist*. 2019;2(2):141-160. doi:10.20517/cdr.2019.10

276. Lippert TH, Ruoff H-J, Volm M. Intrinsic and acquired drug resistance in malignant tumors. The main reason for therapeutic failure. *Arzneimittelforschung*. 2008;58(6):261-264. doi:10.1055/s-0031-1296504

277. Kelderman S, Schumacher TNM, Haanen JBAG. Acquired and intrinsic resistance in cancer immunotherapy. *Mol Oncol*. 2014;8(6):1132-1139. doi:10.1016/j.molonc.2014.07.011

278. Alomar MJ. Factors affecting the development of adverse drug reactions (Review article). *Saudi Pharm J SPJ Off Publ Saudi Pharm Soc*. 2014;22(2):83-94. doi:10.1016/j.jsps.2013.02.003

279. Dent P, Curiel DT, Fisher PB, Grant S. Synergistic combinations of signaling pathway inhibitors: mechanisms for improved cancer therapy. *Drug Resist Updat*

Rev Comment Antimicrob Anticancer Chemother. 2009;12(3):65-73. doi:10.1016/j.drup.2009.03.001

280. Zou J, Ji P, Zhao Y-L, et al. Neighbor communities in drug combination networks characterize synergistic effect. *Mol Biosyst*. 2012;8(12):3185-3196. doi:10.1039/c2mb25267h

281. Cheng F, Kovács IA, Barabási AL. Network-based prediction of drug combinations. *Nat Commun*. 2019;10(1). doi:10.1038/s41467-019-09186-x

282. Tamborero D, Rubio-Perez C, Deu-Pons J, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med*. 2018;10(1):25. doi:10.1186/s13073-018-0531-8

283. Siu KT, Eda H, Santo L, et al. Effect of the BET Inhibitor, Cpi-0610, Alone and in Combination with Lenalidomide in Multiple Myeloma. *Blood*. 2015;126(23):4255. doi:10.1182/blood.V126.23.4255.4255

284. Tania Díaz, Vanina Rodríguez, Ester Lozano, et al. The BET bromodomain inhibitor CPI203 improves lenalidomide and dexamethasone activity in in vitro and in vivo models of multiple myeloma by blockade of Ikaros and MYC signaling. *Haematologica*. 2017;102(10 SE-Articles):1776-1784. doi:10.3324/haematol.2017.164632

285. Moros A, Rodríguez V, Saborit-Villarroya I, et al. Synergistic antitumor activity of lenalidomide with the BET bromodomain inhibitor CPI203 in bortezomib-resistant mantle cell lymphoma. *Leukemia*. 2014;28(10):2049-2059. doi:10.1038/leu.2014.106

286. Fiskus W, Verstovsek S, Manshouri T, et al. Heat shock protein 90 inhibitor is synergistic with JAK2 inhibitor and overcomes resistance to JAK2-TKI in human myeloproliferative neoplasm cells. *Clin Cancer Res*. 2011;17(23):7347-7358. doi:10.1158/1078-0432.CCR-11-1541

287. Bae D, Choi Y, Lee J, et al. M-134, a novel HDAC6-selective inhibitor, markedly improved arthritic severity in a rodent model of rheumatoid arthritis when combined with tofacitinib. *Pharmacol Reports*. 2021;73(1):185-201. doi:10.1007/s43440-020-00188-x

288. Liang X, Liu H, Zhang Y. Novel-targeted therapy for hematological malignancies with JAK and HDAC dual inhibitors. *Future Med Chem*. 2019;11(15):1849-1852. doi:10.4155/fmc-2019-0168

289. Liang X, Zang J, Li X, et al. Discovery of Novel Janus Kinase (JAK) and Histone Deacetylase (HDAC) Dual Inhibitors for the Treatment of Hematological Malignancies. *J Med Chem*. 2019;62(8):3898-3923. doi:10.1021/acs.jmedchem.8b01597

290. Yao L, Ramanujulu PM, Poulsen A, Ohlson S, Dymock BW. Merging of ruxolitinib and vorinostat leads to highly potent inhibitors of JAK2 and histone

deacetylase 6 (HDAC6). *Bioorg Med Chem Lett*. 2018;28(15):2636-2640. doi:10.1016/j.bmcl.2018.06.037

291. Hu W-H, Chan GK-L, Duan R, et al. Synergy of Ginkgetin and Resveratrol in Suppressing VEGF-Induced Angiogenesis: A Therapy in Treating Colorectal Cancer. *Cancers (Basel)*. 2019;11(12). doi:10.3390/cancers11121828

292. Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*. 2013;41(D1):D955-D961. doi:10.1093/nar/gks1111

293. Wang XZ, Beebe JR, Pwiti L, Bielawska A, Smyth MJ. Aberrant sphingolipid signaling is involved in the resistance of prostate cancer cell lines to chemotherapy. *Cancer Res*. 1999;59(22):5842-5848.

294. Jaworski T, Banach-Kasper E, Gralec K. GSK-3β at the Intersection of Neuronal Plasticity and Neurodegeneration. *Neural Plast*. 2019;2019:4209475. doi:10.1155/2019/4209475

295. Smillie KJ, Cousin MA. The Role of GSK3 in Presynaptic Function. *Int J Alzheimers Dis*. 2011;2011:263673. doi:10.4061/2011/263673

296. Takashima A. GSK-3 is essential in the pathogenesis of Alzheimer's disease. *J Alzheimers Dis*. 2006;9(3 Suppl):309-317. doi:10.3233/jad-2006-9s335

297. Gould TD, Zarate CA, Manji HK. Glycogen synthase kinase-3: a target for novel bipolar disorder treatments. *J Clin Psychiatry*. 2004;65(1):10-21.

298. Kozlovsky N, Belmaker RH, Agam G. GSK-3 and the neurodevelopmental hypothesis of schizophrenia. *Eur Neuropsychopharmacol J Eur Coll Neuropsychopharmacol*. 2002;12(1):13-25. doi:10.1016/s0924-977x(01)00131-6

299. Mussmann R, Geese M, Harder F, et al. Inhibition of GSK3 Promotes Replication and Survival of Pancreatic Beta Cells *. *J Biol Chem*. 2007;282(16):12030-12037. doi:10.1074/jbc.M609637200

300. Baarsma HA, Engelbertink LHJM, van Hees LJ, et al. Glycogen synthase kinase-3 (GSK-3) regulates TGF-β_1-induced differentiation of pulmonary fibroblasts. *Br J Pharmacol*. 2013;169(3):590-603. doi:10.1111/bph.12098

301. Campa VM, Baltziskueta E, Bengoa-Vergniory N, et al. A screen for transcription factor targets of glycogen synthase kinase-3 highlights an inverse correlation of NFκB and androgen receptor signaling in prostate cancer. *Oncotarget*. 2014;5(18):8173-8187. doi:10.18632/oncotarget.2303

302. Ammanamanchi S, Brattain MG. Restoration of Transforming Growth Factor-β Signaling through Receptor RI Induction by Histone Deacetylase Activity Inhibition in Breast Cancer Cells *. *J Biol Chem*. 2004;279(31):32620-32625. doi:10.1074/jbc.M402691200

303. Mazzio EA, Soliman KFA. Whole-transcriptomic Profile of SK-MEL-3 Melanoma

Cells Treated with the Histone Deacetylase Inhibitor: Trichostatin A. *Cancer Genomics Proteomics*. 2018;15(5):349-364. doi:10.21873/cgp.20094

304. Hu C, Dadon T, Chenna V, et al. Combined Inhibition of Cyclin-Dependent Kinases (Dinaciclib) and AKT (MK-2206) Blocks Pancreatic Tumor Growth and Metastases in Patient-Derived Xenograft Models. *Mol Cancer Ther*. 2015;14(7):1532-1539. doi:10.1158/1535-7163.MCT-15-0028

305. Criscitiello C, Viale G, Esposito A, Curigliano G. Dinaciclib for the treatment of breast cancer. *Expert Opin Investig Drugs*. 2014;23(9):1305-1312. doi:10.1517/13543784.2014.948152

306. Chen X-X, Xie F-F, Zhu X-J, et al. Cyclin-dependent kinase inhibitor dinaciclib potently synergizes with cisplatin in preclinical models of ovarian cancer. *Oncotarget*. 2015;6(17):14926-14939. doi:10.18632/oncotarget.3717

307. Lin S-F, Lin J-D, Hsueh C, Chou T-C, Wong RJ. A cyclin-dependent kinase inhibitor, dinaciclib in preclinical treatment models of thyroid cancer. *PLoS One*. 2017;12(2):e0172315. doi:10.1371/journal.pone.0172315

308. Bryja V, Červenka I, Čajánek L. The connections of Wnt pathway components with cell cycle and centrosome: side effects or a hidden logic? *Crit Rev Biochem Mol Biol*. 2017;52(6):614-637. doi:10.1080/10409238.2017.1350135

309. Rasmussen ML, Ortolano NA, Romero-Morales AI, Gama V. Wnt Signaling and Its Impact on Mitochondrial and Cell Cycle Dynamics in Pluripotent Stem Cells. *Genes (Basel)*. 2018;9(2). doi:10.3390/genes9020109

310. Moffat JG, Rudolph J, Bailey D. Phenotypic screening in cancer drug discovery — past, present and future. *Nat Rev Drug Discov*. 2014;13(8):588-602. doi:10.1038/nrd4366

311. Galluzzi L, Vacchelli E, Bravo-San Pedro J-M, et al. Classification of current anticancer immunotherapies. *Oncotarget*. 2014;5(24):12472-12508. doi:10.18632/oncotarget.2998

312. Espinosa E, Zamora P, Feliu J, González Barón M. Classification of anticancer drugs—a new system based on therapeutic targets. *Cancer Treat Rev*. 2003;29(6):515-523. doi:https://doi.org/10.1016/S0305-7372(03)00116-6

313. Liu Z, Delavan B, Roberts R, Tong W. Lessons Learned from Two Decades of Anticancer Drugs. *Trends Pharmacol Sci*. 2017;38(10):852-872. doi:https://doi.org/10.1016/j.tips.2017.06.005

314. Sun J, Wei Q, Zhou Y, Wang J, Liu Q, Xu H. A systematic analysis of FDA-approved anticancer drugs. *BMC Syst Biol*. 2017;11(5):87. doi:10.1186/s12918-017-0464-7

315. Janku F, Yap TA, Meric-Bernstam F. Targeting the PI3K pathway in cancer: are we making headway? *Nat Rev Clin Oncol*. 2018;15(5):273-291.

doi:10.1038/nrclinonc.2018.28

316. Farhan M, Wang H, Gaur U, Little PJ, Xu J, Zheng W. FOXO Signaling Pathways as Therapeutic Targets in Cancer. *Int J Biol Sci*. 2017;13(7):815-827. doi:10.7150/ijbs.20052

317. Panda AK, Chakraborty D, Sarkar I, Khan T, Sa G. New insights into therapeutic activity and anticancer properties of curcumin. *J Exp Pharmacol*. 2017;9:31-45. doi:10.2147/JEP.S70568

318. Sun X, Bao J, Shao Y. Mathematical Modeling of Therapy-induced Cancer Drug Resistance: Connecting Cancer Mechanisms to Population Survival Rates. *Sci Rep*. 2016;6(1):22498. doi:10.1038/srep22498

319. Johnson SF, Cruz C, Greifenberg AK, et al. CDK12 Inhibition Reverses De Novo and Acquired PARP Inhibitor Resistance in BRCA Wild-Type and Mutated Models of Triple-Negative Breast Cancer. *Cell Rep*. 2016;17(9):2367-2381. doi:10.1016/j.celrep.2016.10.077

320. Fabbro D, Manley PW, Jahnke W, et al. Inhibitors of the Abl kinase directed at either the ATP- or myristate-binding site. *Biochim Biophys Acta - Proteins Proteomics*. 2010;1804(3):454-462. doi:https://doi.org/10.1016/j.bbapap.2009.12.009

321. Griffin M, Scotto D, Josephs DH, et al. BRAF inhibitors: resistance and the promise of combination treatments for melanoma. *Oncotarget; Vol 8, No 44*. 2017. https://www.oncotarget.com/article/19836/text/.

322. Menden MP, Wang D, Guan Y, et al. A cancer pharmacogenomic screen powering crowd-sourced advancement of drug combination prediction. *bioRxiv*. January 2018:200451. doi:10.1101/200451

323. Menden MP, Wang D, Mason MJ, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun*. 2019;10(1):2674. doi:10.1038/s41467-019-09799-2

324. Celebi R, Bear Don't Walk O, Movva R, Alpsoy S, Dumontier M. In-silico Prediction of Synergistic Anti-Cancer Drug Combinations Using Multi-omics Data. *Sci Rep*. 2019;9(1):8949. doi:10.1038/s41598-019-45236-6

325. Li H, Li T, Quang D, Guan Y. Network Propagation Predicts Drug Synergy in Cancers. *Cancer Res*. 2018;78(18):5446 LP - 5457. doi:10.1158/0008-5472.CAN-18-0740

326. Sharma A, Rani R. An integrated framework for identification of effective and synergistic anti-cancer drug combinations. *J Bioinform Comput Biol*. 2018;16(05):1850017. doi:10.1142/S0219720018500178

327. Zhong L, Li Y, Xiong L, et al. Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduct Target Ther*.

2021;6(1):201. doi:10.1038/s41392-021-00572-w

328. Lv S, Xu Y, Chen X, et al. Prioritizing cancer therapeutic small molecules by integrating multiple OMICS datasets. *OMICS*. 2012;16(10):552-559. doi:10.1089/omi.2012.0005

329. Shameer K, Johnson KW, Glicksberg BS, et al. Prioritizing Small Molecule as Candidates for Drug Repositioning using Machine Learning. *bioRxiv*. January 2018:331975. doi:10.1101/331975

330. Gupta S, Chaudhary K, Kumar R, et al. Prioritization of anticancer drugs against a cancer using genomic features of cancer cells: A step towards personalized medicine. *Sci Rep*. 2016;6(1):23857. doi:10.1038/srep23857

APPENDICES

APPENDIX A

TABLE OF DATA INTEGRATION TOOLS SUMMARIZED IN SECTION 2.2.4

Current summary of data integration tools (adapted from Huang et al.,2017 [178])

Name	Data Type	Output	Statistical Method	FS method	Reference
Joint NMF	Multi-data	Subset of genes (modules)	MFA	NA	[172,179]
iCluster	EXP, CNV	Cluster	MFA	L1 penalty	[180]
iCluster+	Multi-data	Cluster	MFA	L1 penalty	[181]
JIVE	Multi-data	Shared and unique factors	MFA	L1 penalty	[182]
Joint Bayes Factor	EXP, MET, CNV	Shared and unique factors	MFA	Student-t sparseness promoting prior	[183]
ssCCA	Sequence data	Operational taxonomic unit and cluster	CCA	L1 penalty	[169]
CCA sparse Group	Two types of data	Group of features with weights	CCA	L1 penalty	[162]
sMBPLS	Multi-data	Group of features as modules	PLS	L1 penalty	[157]
SNPLS	EXP, drug response, gene network info.	Gene-drug co-module	PLS	Network-based penalty	[184]
MDI	Multi-data	Cluster	Bayesian	NA	[185]

Prob_GBM	EXP, CNV, miRNA, SNP	Cluster	Bayesian	NA	[186]
PSDF	EXP, CNV	Cluster	Bayesian	Binary indicator-likelihood of feature	[187]
BCC	EXP, MET, miRNA, proteomics	Cluster	Bayesian	NA	[188]
CONEXIC	EXP, CNV	Groups of genes associated with modulators	Bayesian	NA	[175]
PARADIGM	Multi-data	Gene score and significance in each pathway	pathway networks	NA	[189]
SNF	EXP, MET, miRNA	Cluster	SNF	NA	[170]
Lemon-Tree	EXP, CNV/miRNA/ methyl	Association network graphics	module network	NA	[190]
rMKL-LPP	Multi-data	Cluster	MKL	Dimension reduction metric Locality Preserving Projections (LPP)	[171]
CNAmet	EXP, MET, CNV	Scores and p-values of genes	Multi-step analysis	NA	[191]
iPAC	EXP, CNV	Subset of genes	Multi-step analysis	Multiple filtering steps including common aberrant genes, in-cis correlation and in-trans functionality	[192]
ATHENA	EXP, CNV, MET, miRNA	Final model with patient index	Grammatical Evolution	Neural Networks	[21]

			Neural Networks (GENN)		
jActiveModules	EXP, PPI, protein-DNA interactions	Subnetwork (network hotspots)	Network simulated annealing	NA	[193]
Network Propagation	Gene expression, mutation, PPI	Propagated network relative to differential expression of gene	Network	NA	[194]
SDP/SVM	EXP, protein sequence, protein interactions, hydropathy profile	Linear classifier based on combination of kernels	SDP/SVM	Recommends CCA	[195]
FSMKL	EXP, CNV, Clinic feature (ER status)	Linear classifier based on combination of kernels	MKL	SimpleMKL (gradient descent method)	[196]
iBAG	Multi-data	Subset of genes	Multi-step analysis	Bayesian	[197]
MCD	MET, CNV, LoH	Subset of genes	Multi-step analysis	NA	[198]
Anduril	EXP,MET, miRNA,exon, aCGH,SNP	Comprehensive report	Multi-step analysis	NA	[199]
GeneticInterPred	EXP, PPI, protein complex data	Genetic interaction Labels	Graph integration	NA	[200]
Graph-based Learning	EXP, CNV, MET, miRNA	Patient scores for classification purpose	Graph integration	NA	[201]
CoxPath	EXP, CNV, MET, miRNA	Prognosis index for each patient	Multi-step analysis	L1 penalty	[202]

MKGI	EXP, CNV, MET, miRNA	Final model with patient index	Grammatical Evolution Neural Networks (GENN)	Neural Networks	[203]

FS Method: Feature Selection Method,
EXP: Expression,
CNV: Copy Number Variation,
MET: DNA Methylation,
SNP: Single Nucleotide Polymorphism,
aCGH: Array Comparative Genomic Hybridization,
PPI: Protein-Protein Interaction,
LoH: Loss of Heterozygosity,
MFA: Matrix factorization Analysis,
CCA: Canonical Correlation Analysis,
PLS: Partial Least Squares,
SNF: Similarity Network Fusion
MKL: Multiple-Kernel Learning

APPENDIX B

TABLE OF SIGNIFICANCE THRESHOLDS USED FOR SEED PROTEIN PREPARATION AND THE NUMBER OF SEED PROTEINS COLLECTED IN ALL DRUG AND CELL LINE CONDITIONS WITHIN NETWORK RECONSTRUCTION PROTOCOL

	A375		***A549***		***MCF7***		***PC3***		***YAPC***		***NPC***	
DRUG	**L1000 threshold**	**number of seed proteins**	**L1000 threshold**	**number of seed proteins**	**L1000 threshold**	**number of seed proteins**	**L1000 threshold**	**number of seed proteins**	**L1000 threshold**	**number of seed proteins**	**L1000 threshold**	**number of seed proteins**
4,5,6,7-tetrabromobenzotriazole	p<0.001	63	p<0.01	42	p<0.001	79	p<0.01	43	NA	NA	NA	NA
alpelisib	p<0.001	10	p<0.001	29	NA	NA	p<0.01	43	NA	NA	NA	NA
alvocidib	p<0.001	87	p<0.001	86	p<0.001	97	p<0.001	73	p<0.001	203	p<0.001	93
AR-A014418	p<0.001	8	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
belinostat	p<0.001	168	p<0.001	112	p<0.001	214	p<0.001	155	p<0.001	208	NA	NA
BIX-01294	p<0.001	152	p<0.001	44	p<0.001	190	p<0.001	155	NA	NA	NA	NA
BIX-01338	p<0.001	103	p<0.001	37	p<0.001	178	p<0.001	153	NA	NA	NA	NA
BMS-345541	p<0.001	14	p<0.01	18	NA	NA	p<0.01	34	NA	NA	NA	NA
CC-401	p<0.001	21	p<0.001	19	NA	NA	NA	NA	NA	NA	NA	NA
CHIR-99021	p<0.001	35	NA	NA	p<0.001	33	p<0.01	45	NA	NA	NA	NA

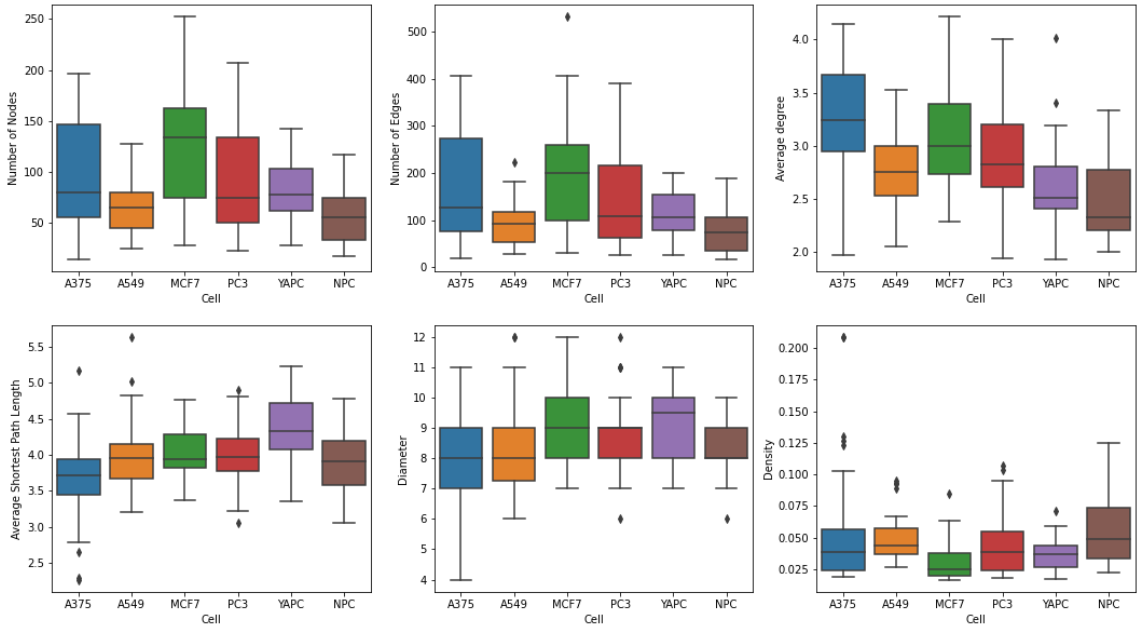
curcumin	p<0.001	70	p<0.005	39	p<0.001	61	p<0.01	36	p<0.1	30	NA	NA
decitabine	p<0.001	131	NA	NA	p<0.001	161	p<0.001	132	NA	NA	NA	NA
dexamethasone	p<0.001	36	p<0.001	31	p<0.005	57	p<0.1	27	NA	NA	NA	NA
dinaciclib	p<0.001	119	p<0.001	86	p<0.001	122	p<0.001	100	p<0.001	191	p<0.001	136
entinostat	p<0.001	170	p<0.001	41	p<0.001	209	p<0.001	158	p<0.001	79	NA	NA
epz004777	p<0.001	130	p<0.001	34	p<0.001	152	p<0.001	136	p<0.001	43	NA	NA
EPZ-005687	p<0.001	139	NA	NA	p<0.001	181	p<0.001	144	NA	NA	NA	NA
etoposide	p<0.001	104	p<0.005	40	p<0.001	67	p<0.1	13	NA	NA	NA	NA
everolimus	p<0.001	25	p<0.005	28	NA	NA	NA	NA	NA	NA	NA	NA
EX-527	p<0.001	143	p<0.001	49	p<0.001	173	p<0.001	110	NA	NA	NA	NA
exifone	p<0.001	47	NA	NA	NA	NA	p<0.1	21	NA	NA	NA	NA
geldanamycin	p<0.001	171	p<0.001	60	p<0.001	172	p<0.001	130	NA	NA	NA	NA
ginkgetin	p<0.001	49	NA	NA	p<0.001	63	p<0.1	31	NA	NA	NA	NA
GSK-J4	p<0.001	150	NA	NA	p<0.001	171	p<0.001	153	NA	NA	NA	NA
I-BET-151	p<0.001	126	p<0.001	46	p<0.001	140	p<0.001	161	NA	NA	NA	NA
I-BET-762	p<0.001	139	p<0.001	36	p<0.001	155	p<0.001	140	NA	NA	NA	NA
IKK-inhibitor-X	p<0.001	15	NA	NA	NA	NA	p<0.005	35	NA	NA	p<0.005	49
IPI-145	p<0.001	10	p<0.005	23	NA	NA	NA	NA	NA	NA	NA	NA
JQ1-(+)	p<0.001	143	p<0.001	64	p<0.001	141	p<0.001	153	NA	NA	NA	NA
KN-62	p<0.001	30	p<0.005	39	p<0.001	38	p<0.05	72	NA	NA	NA	NA
KN-93	p<0.001	60	NA	NA	p<0.001	55	p<0.1	40	NA	NA	NA	NA
KU-55933	p<0.001	41	NA	NA	NA	NA	p<0.05	49	NA	NA	NA	NA

lenalidomide	p<0.001	12	p<0.01	24	NA	NA	p<0.01	29	NA	NA	NA	NA
losmapimod	p<0.001	6	NA	NA	NA	NA	p<0.005	29	NA	NA	NA	NA
LY-294002	p<0.001	149	p<0.001	67	p<0.001	177	p<0.001	159	p<0.001	123	NA	NA
methylstat	p<0.001	158	p<0.005	38	p<0.001	147	p<0.001	152	p<0.001	162	NA	NA
niclosamide	p<0.001	62	p<0.001	55	NA	NA	p<0.001	53	NA	NA	NA	NA
nilotinib	p<0.001	11	p<0.005	37	NA	NA	NA	NA	NA	NA	p<0.01	21
olaparib	p<0.001	59	NA	NA	p<0.001	64	p<0.01	20	NA	NA	NA	NA
OSI-027	p<0.001	167	p<0.001	56	p<0.001	173	p<0.001	147	p<0.001	66	NA	NA
palbociclib	p<0.001	14	p<0.005	19	NA	NA	p<0.005	22	p<0.001	65	p<0.005	30
pazopanib	p<0.001	32	p<0.001	56	NA	NA	p<0.001	63	NA	NA	p<0.001	70
PD-0325901	p<0.001	26	p<0.005	38	NA	NA	p<0.005	20	p<0.001	90	p<0.005	52
pravastatin	p<0.001	28	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
pyrazolanthrone	p<0.001	31	NA	NA	NA	NA	p<0.005	56	NA	NA	p<0.001	40
resveratrol	p<0.001	150	p<0.001	62	p<0.001	180	p<0.001	148	NA	NA	NA	NA
RG-4733	p<0.001	12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
RGFP-966	p<0.001	46	NA	NA	p<0.001	32	p<0.01	34	NA	NA	NA	NA
rolipram	p<0.001	39	p<0.005	37	p<0.001	49	p<0.01	24	NA	NA	NA	NA
roscovitine	p<0.001	33	NA	NA	NA	NA	p<0.005	35	NA	NA	NA	NA
ruxolitinib	p<0.001	43	NA	NA	p<0.001	96	p<0.05	44	NA	NA	NA	NA
salermide	p<0.001	153	p<0.001	41	p<0.001	159	p<0.001	142	p<0.001	62	NA	NA
selumetinib	p<0.001	41	p<0.005	22	NA	NA	NA	NA	p<0.001	105	p<0.005	49
semagacestat	p<0.001	52	p<0.005	44	p<0.001	73	p<0.01	40	NA	NA	NA	NA

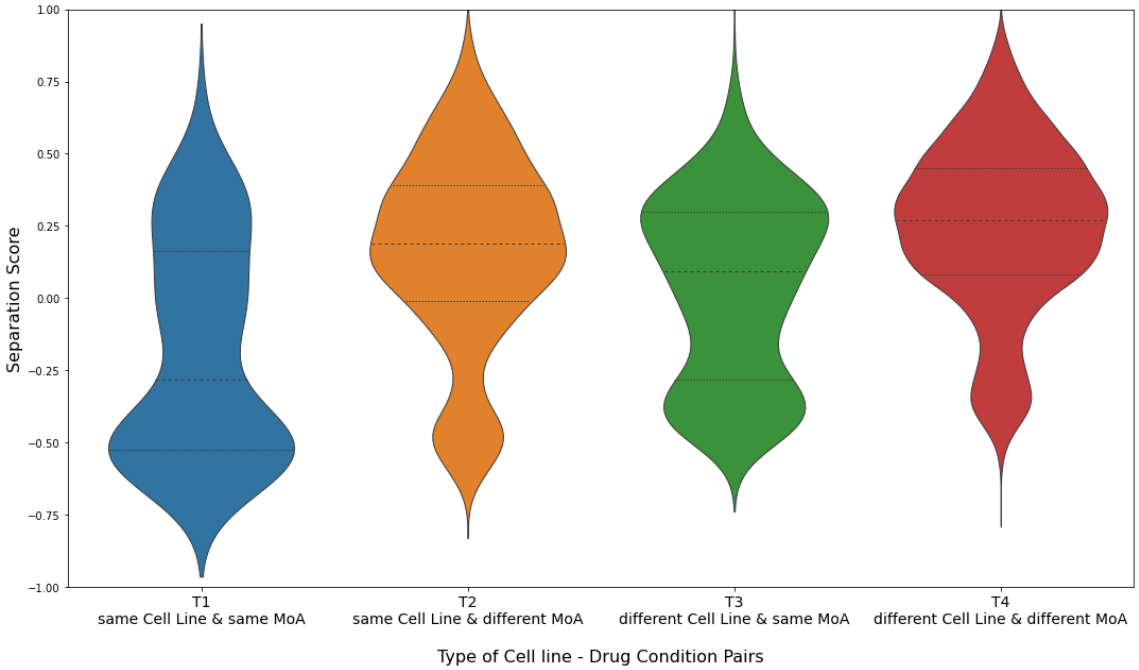
sirolimus	p<0.001	151	p<0.001	51	p<0.001	177	p<0.001	161	p<0.001	56	NA	NA
sotrastaurin	p<0.001	60	p<0.005	87	p<0.001	75	p<0.005	39	NA	NA	NA	NA
staurosporine	p<0.001	73	p<0.001	120	p<0.001	72	p<0.001	64	NA	NA	p<0.001	104
tacedinaline	p<0.001	155	NA	NA	p<0.001	183	p<0.001	151	p<0.001	47	NA	NA
tacrolimus	p<0.001	16	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
TG-101348	p<0.001	56	p<0.005	41	p<0.001	71	NA	NA	NA	NA	p<0.001	32
tofacitinib	p<0.005	19	p<0.1	40	NA	NA	p<0.1	27	NA	NA	p<0.1	21
tretinoin	p<0.001	34	p<0.1	41	p<0.001	99	p<0.005	43	p<0.001	34	NA	NA
trichostatin-a	p<0.001	172	p<0.001	80	p<0.001	169	p<0.001	168	NA	NA	NA	NA
UNC-0321	p<0.001	154	NA	NA	p<0.001	187	p<0.001	162	p<0.005	48	NA	NA
UNC-1215	p<0.001	161	p<0.005	59	p<0.001	177	p<0.001	160	NA	NA	NA	NA
VE-822	p<0.001	14	p<0.1	23	NA	NA	p<0.005	32	NA	NA	NA	NA
vemurafenib	p<0.001	56	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
verteporfin	p<0.001	24	p<0.001	42	NA	NA	p<0.001	55	NA	NA	p<0.005	42
vorinostat	p<0.001	28	NA	NA	NA	NA	p<0.001	31	NA	NA	p<0.001	31
zebularine	p<0.001	117	NA	NA	p<0.001	179	p<0.001	156	p<0.005	57	NA	NA

APPENDIX C

SUMMARY OF TOPOLOGICAL PROPERTIES OF RECONSTRUCTED NETWORKS PER CELL LINE



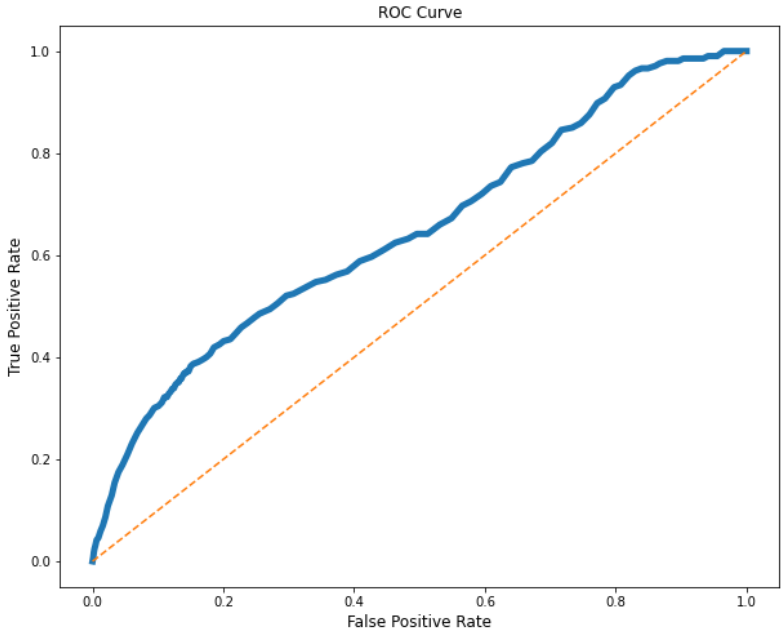
Appendix C – Figure 1. The topological properties of cell line networks: Number of nodes, number of edges, average degree, average shortest path length, diameter and density.



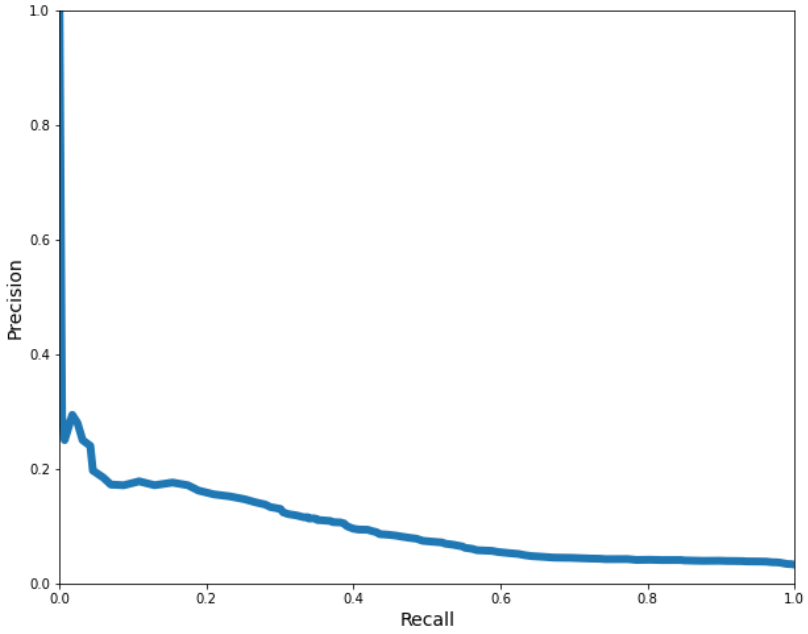
Appendix C – Figure 2. Distribution of separation scores based on network pair types.

COMPARISON OF NETWORK SEPARATION AGAINST LITERATURE-CURATED MECHANISM OF ACTIONS OF DRUGS

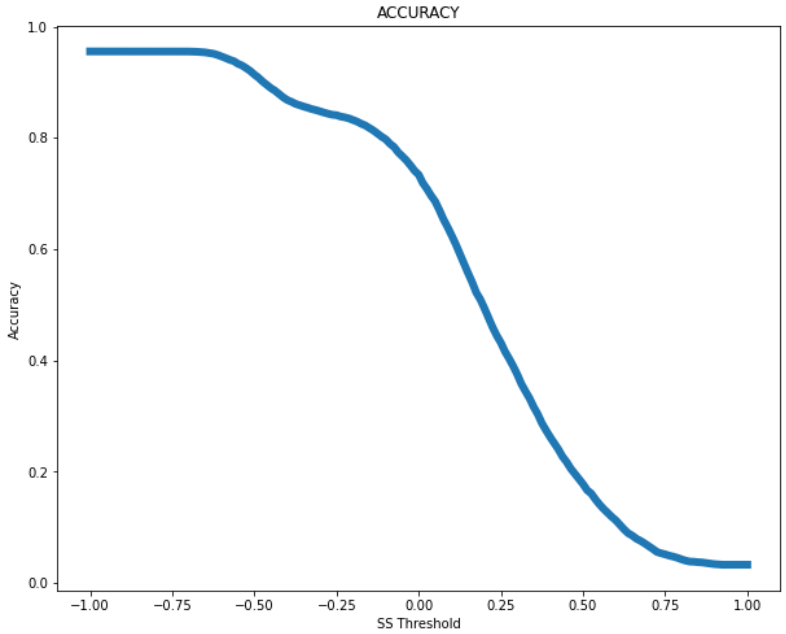
Drug pairs with same MoA and having overlapped networks are defined as true positives (TP), those with same MoA and having separate networks are defined as false negatives (FN), those with different MoA and having overlapped networks are defined as false positives (FP) and those with different MoA and having separate networks are defined as true negatives (TN). Separation scores lie between -1.0 and 1.0.



Appendix C – Figure 3. ROC curve of reconstructed networks based on literature-curated MoAs.



Appendix C – Figure 4. Precision – Recall curve for separation scores of network pairs.



Appendix C – Figure 5. Accuracy curve of reconstructed networks based on literature-curated MoAs.

APPENDIX D

TABLE OF HIGHEST RANKED FIRST 100 PREDICTED DRUG PAIRS

Drug1	Drug2	Cell	Separation Score
I-BET-762	lenalidomide	A375	0.69
salermide	lenalidomide	A375	0.68
rolipram	sirolimus	A549	0.67
geldanamycin	lenalidomide	A375	0.67
geldanamycin	tofacitinib	A375	0.66
BIX-01338	sotrastaurin	PC3	0.65
BIX-01338	semagacestat	PC3	0.65
methylstat	lenalidomide	A375	0.65
BIX-01294	tofacitinib	A375	0.64
BIX-01338	exifone	PC3	0.63
JQ1	tofacitinib	A375	0.63
curcumin	tofacitinib	A375	0.62
BIX-01294	ginkgetin	PC3	0.62
OSI-027	curcumin	A549	0.62
UNC-1215	lenalidomide	A375	0.62
ruxolitinib	roscovitine	A375	0.61
rolipram	OSI-027	A549	0.61
staurosporine	lenalidomide	A375	0.60
BIX-01294	IKK-inhibitor-X	A375	0.59
zebularine	pravastatin	A375	0.59

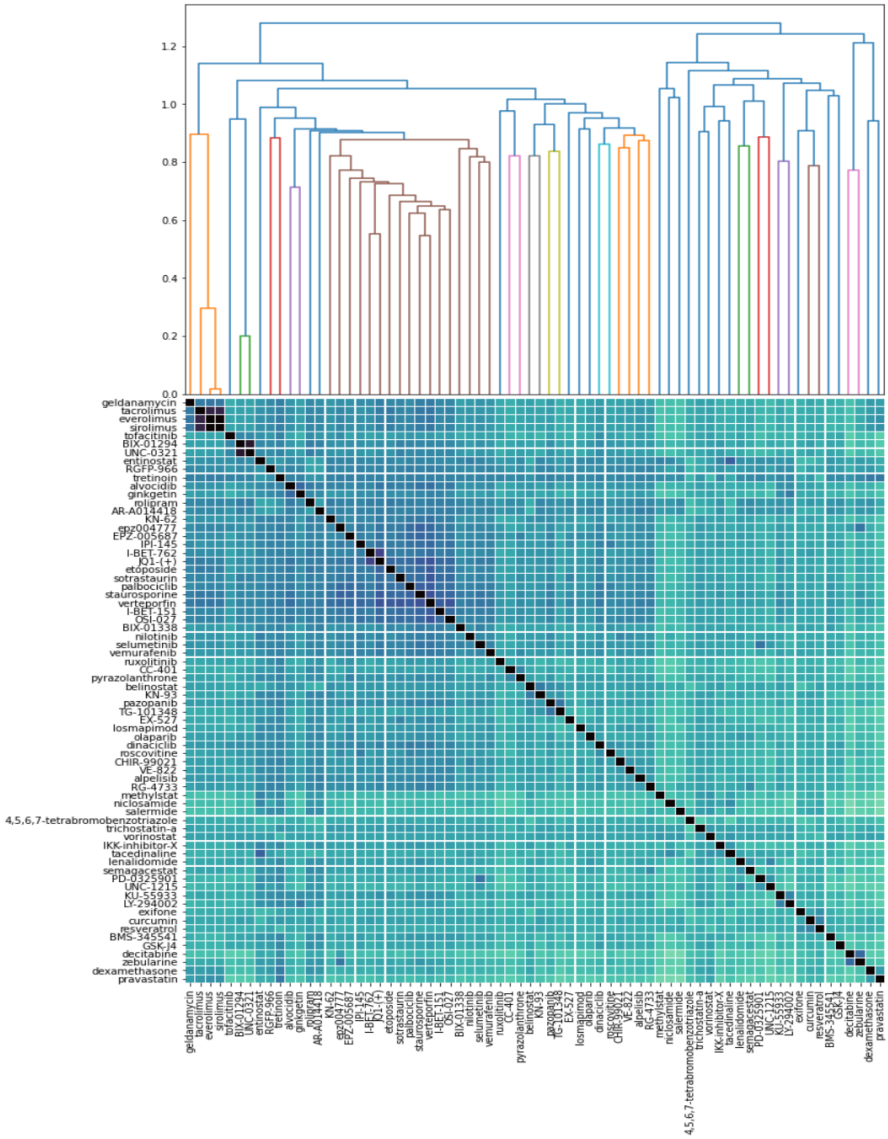
dexamethasone	sirolimus	A549	0.59
sirolimus	pravastatin	A375	0.59
curcumin	lenalidomide	A375	0.59
tacedinaline	rolipram	A375	0.58
BIX-01294	pravastatin	A375	0.58
salermide	tofacitinib	A375	0.58
tacedinaline	tofacitinib	A375	0.58
methylstat	ginkgetin	PC3	0.58
UNC-1215	ginkgetin	PC3	0.57
semagacestat	sirolimus	A549	0.57
JQ1	ginkgetin	PC3	0.57
UNC-1215	tofacitinib	A375	0.57
salermide	ginkgetin	PC3	0.57
alvocidib	curcumin	YAPC	0.56
trichostatin-a	tofacitinib	A375	0.56
BIX-01294	verteporfin	A375	0.56
epz004777	ginkgetin	PC3	0.56
UNC-0321	pravastatin	A375	0.56
entinostat	tofacitinib	A375	0.56
exifone	lenalidomide	A375	0.56
I-BET-762	tofacitinib	A375	0.56
OSI-027	pravastatin	A375	0.56
geldanamycin	IKK-inhibitor-X	A375	0.55
methylstat	curcumin	YAPC	0.55
dinaciclib	curcumin	YAPC	0.55
UNC-0321	rolipram	A375	0.55
belinostat	ginkgetin	PC3	0.55
tretinoin	sirolimus	A549	0.55
I-BET-151	tofacitinib	A375	0.55

BIX-01294	exifone	PC3	0.55
zebularine	ginkgetin	PC3	0.55
BIX-01338	curcumin	PC3	0.55
GSK-J4	tofacitinib	A375	0.54
BIX-01294	rolipram	A375	0.54
tretinoin	OSI-027	A549	0.54
zebularine	rolipram	A375	0.54
LY-294002	ginkgetin	PC3	0.54
epz004777	rolipram	A375	0.54
trichostatin-a	lenalidomide	PC3	0.54
JQ1	pravastatin	A375	0.54
OSI-027	IKK-inhibitor-X	A375	0.54
I-BET-151	ginkgetin	PC3	0.53
resveratrol	ginkgetin	PC3	0.53
UNC-0321	lenalidomide	PC3	0.53
sirolimus	KN-62	A549	0.53
etoposide	tofacitinib	A375	0.53
LY-294002	exifone	PC3	0.53
epz004777	pravastatin	A375	0.53
salermide	exifone	PC3	0.53
JQ1	exifone	PC3	0.52
LY-294002	pravastatin	A375	0.52
UNC-1215	semagacestat	PC3	0.52
salermide	IKK-inhibitor-X	A375	0.52
tacedinaline	ginkgetin	PC3	0.52
exifone	pazopanib	A375	0.52
staurosporine	lenalidomide	PC3	0.52
BIX-01294	sotrastaurin	PC3	0.52
epz004777	IKK-inhibitor-X	A375	0.52

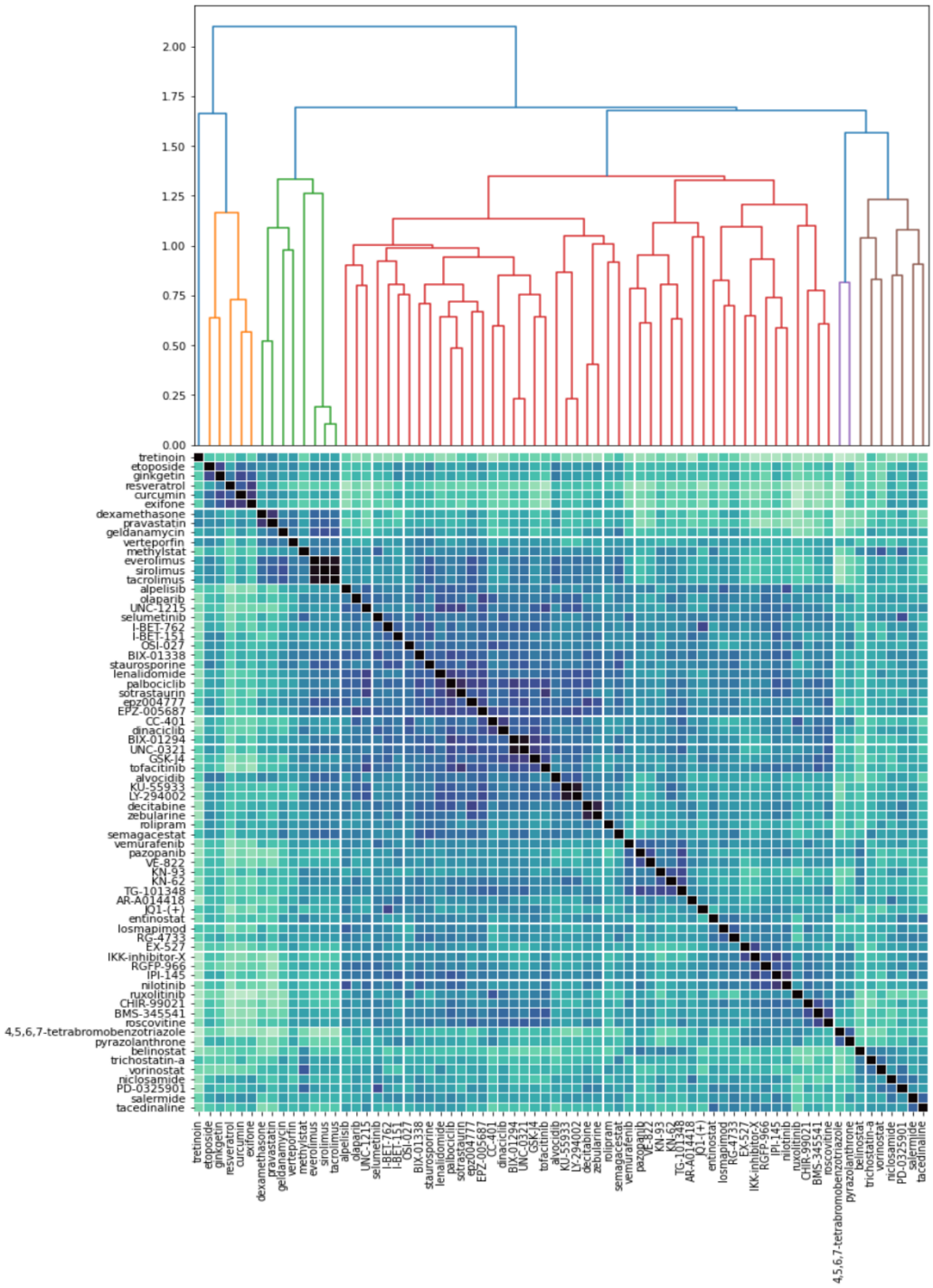
UNC-0321	verteporfin	A375	0.52
zebularine	exifone	PC3	0.52
OSI-027	ginkgetin	PC3	0.52
belinostat	tofacitinib	A375	0.52
geldanamycin	ginkgetin	PC3	0.52
UNC-1215	BMS-345541	PC3	0.52
GSK-J4	ginkgetin	PC3	0.51
UNC-0321	ginkgetin	PC3	0.51
roscovitine	lenalidomide	A375	0.51
olaparib	tofacitinib	A375	0.51
BIX-01338	lenalidomide	PC3	0.51
salermide	lenalidomide	PC3	0.51
salermide	verteporfin	A375	0.50
UNC-1215	lenalidomide	PC3	0.50
UNC-1215	pravastatin	A375	0.50
trichostatin-a	ginkgetin	PC3	0.50
trichostatin-a	sotrastaurin	PC3	0.50
decitabine	ginkgetin	PC3	0.50
etoposide	pravastatin	A375	0.50
JQ1	IKK-inhibitor-X	A375	0.50
UNC-1215	sotrastaurin	PC3	0.50

APPENDIX E

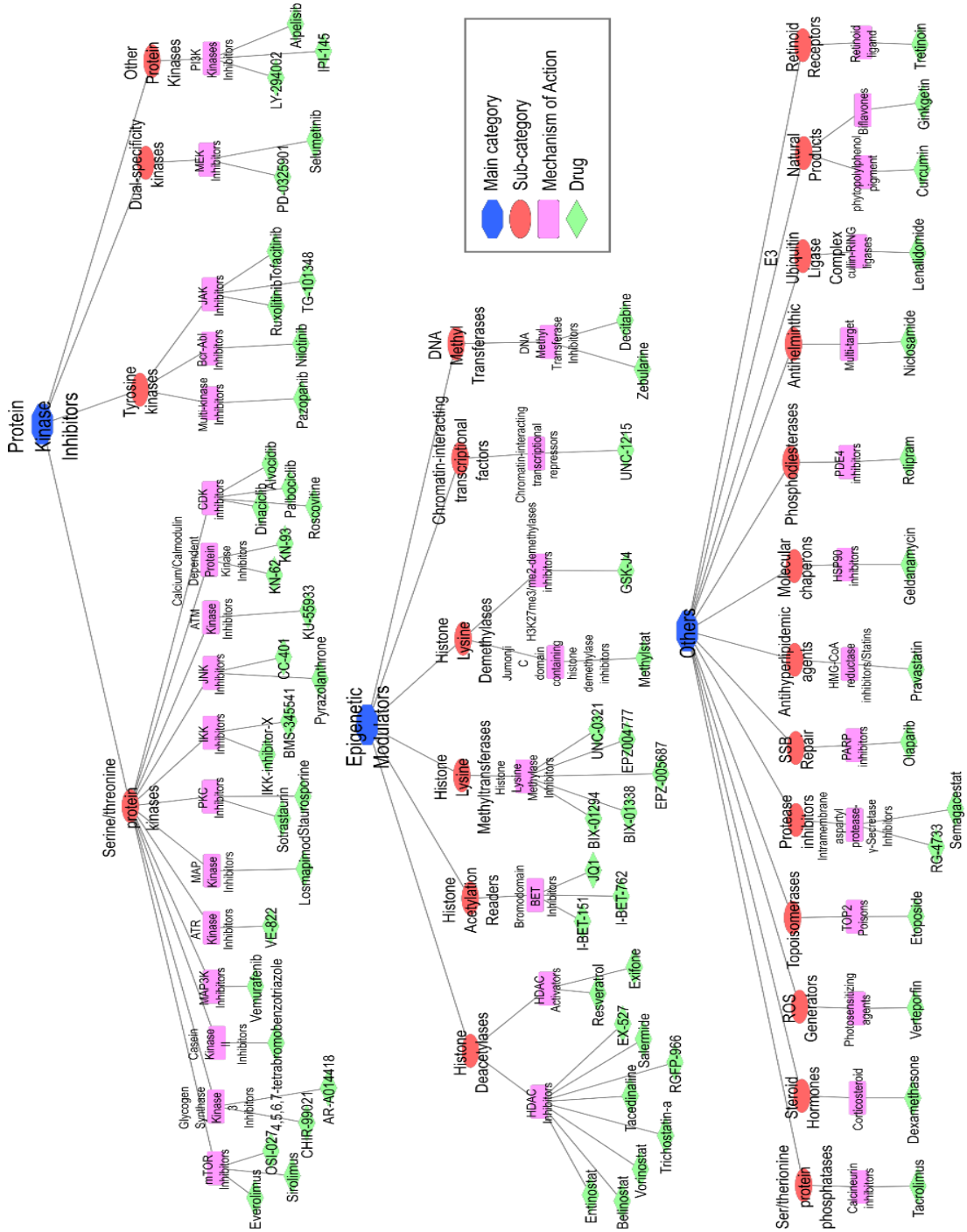
FIGURES THAT ARE NOT LEGIBLE IN THE MAIN TEXT



Appendix E- Figure 1. Tanimoto similarities of 70 small molecule drugs

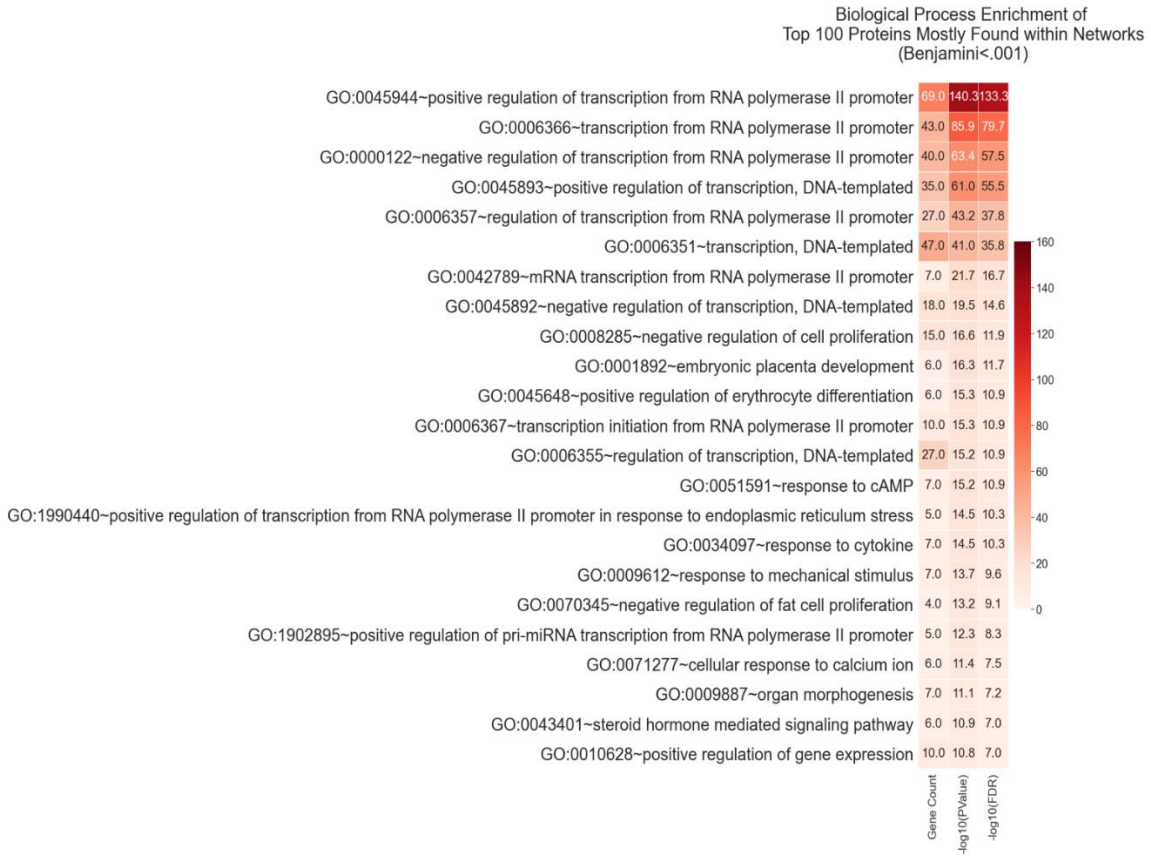


Appendix E- Figure 2. MACCS key distances of 70 small molecule drugs

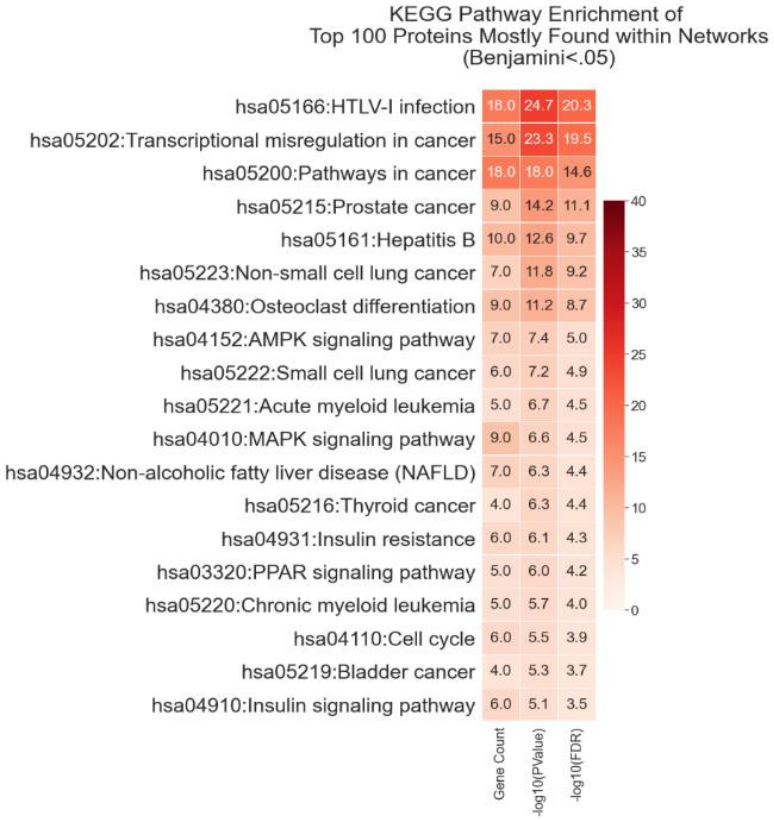


Appendix E- Figure 3. Drug groups classified based on mechanisms of action

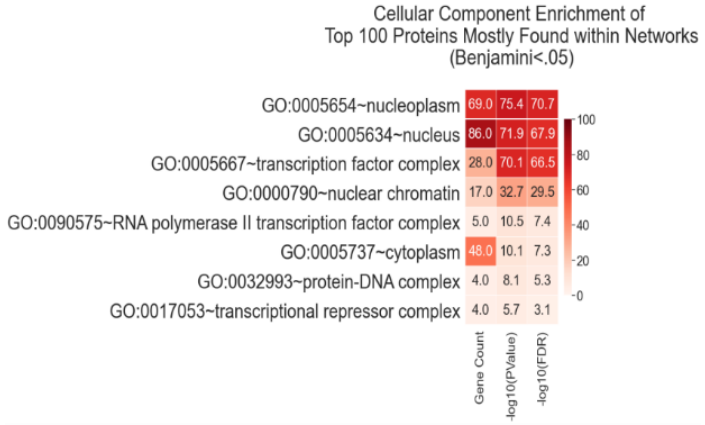
Appendix E- Figure 4. Functional analysis of mostly found 100 proteins – Molecular function enrichments



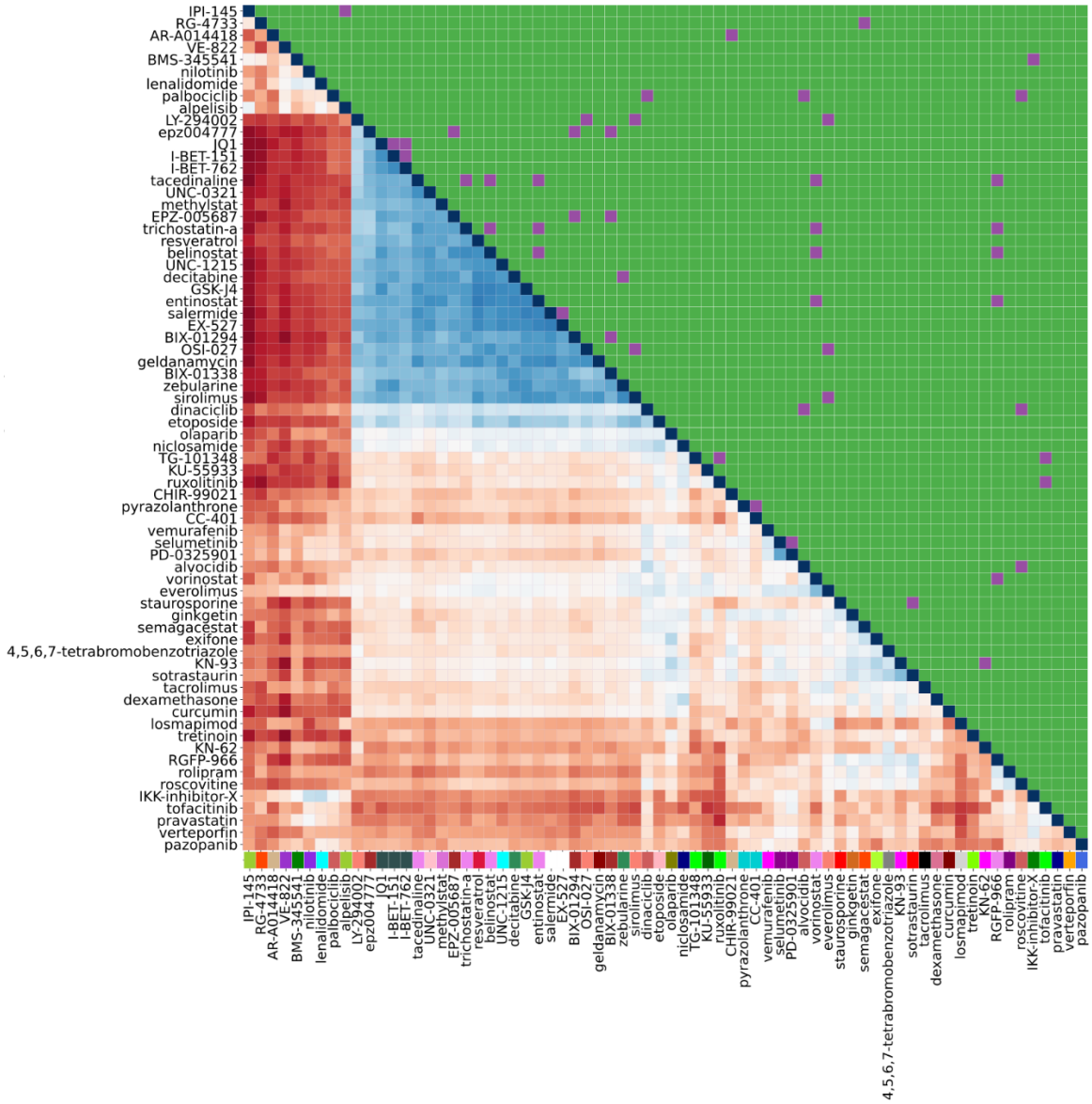
Appendix E- Figure 5. Functional analysis of mostly found 100 proteins – Biological process enrichments



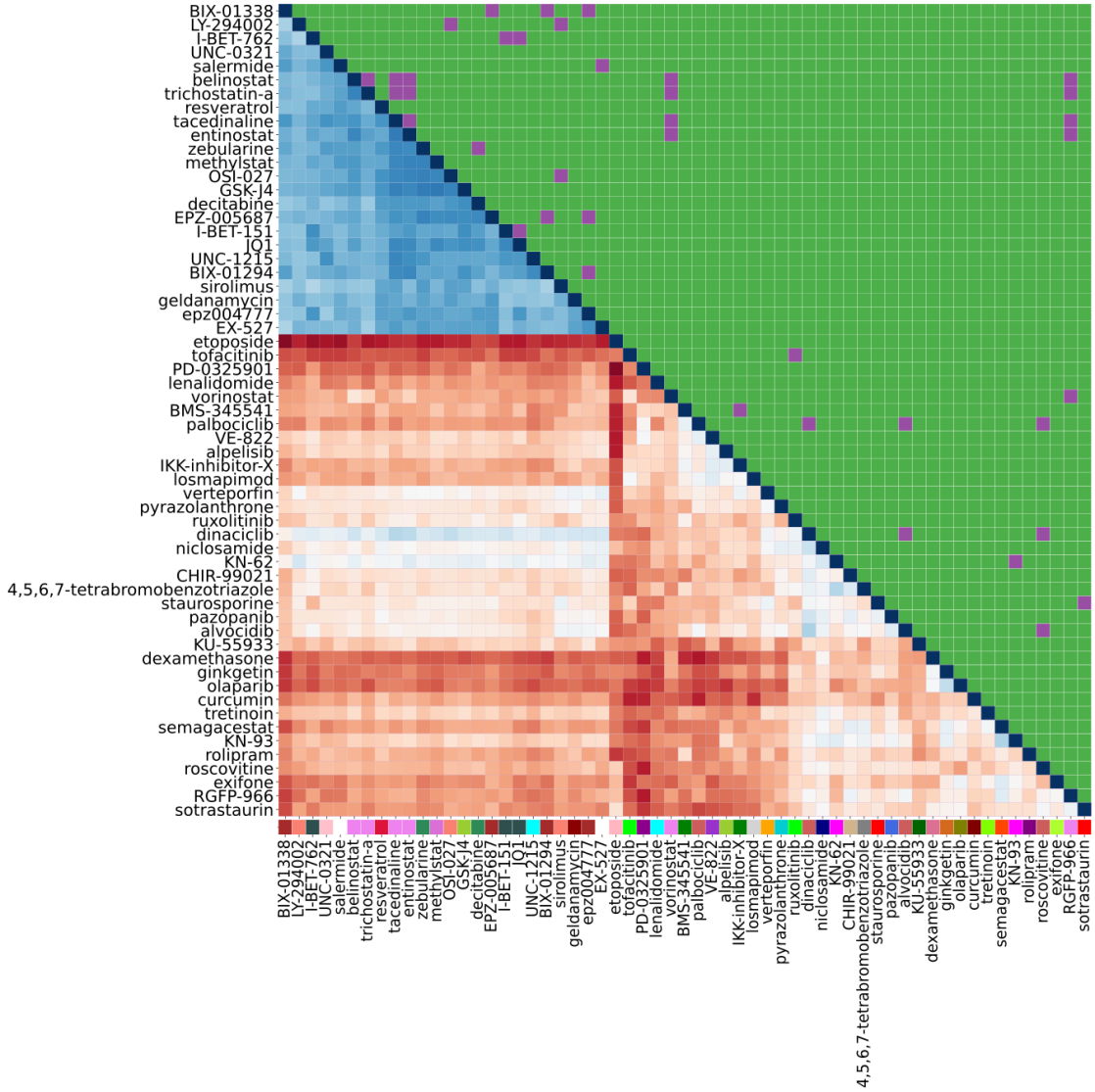
Appendix E- Figure 6. Functional analysis of mostly found 100 proteins – KEGG pathway enrichments



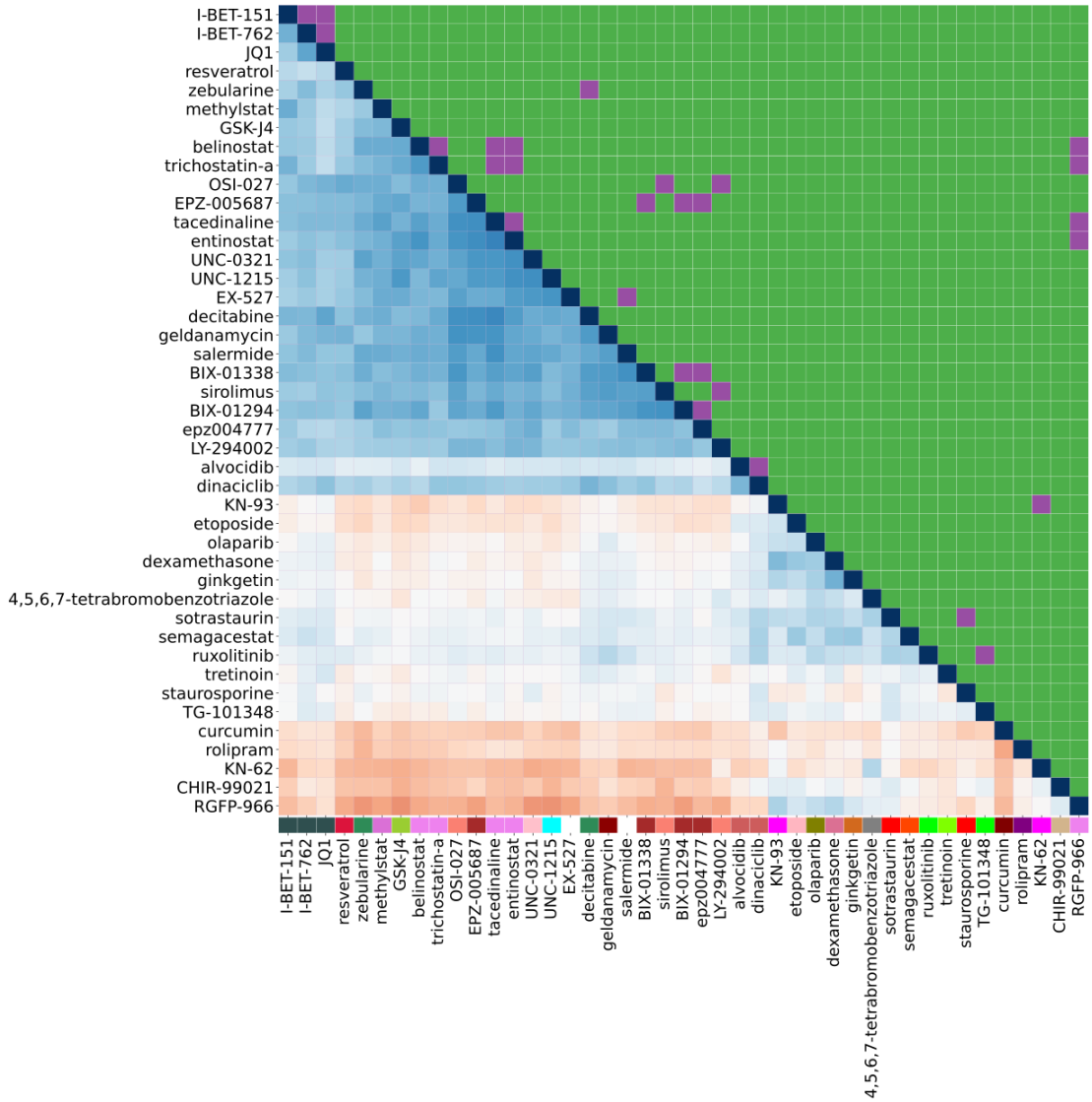
Appendix E- Figure 7. Functional analysis of mostly found 100 proteins – Cellular compartment enrichments



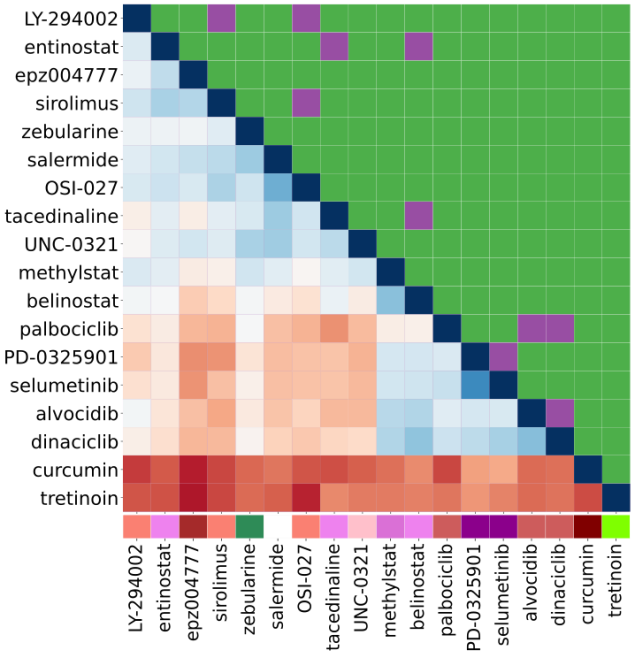
Appendix E- Figure 8. Separation score matrix of A375



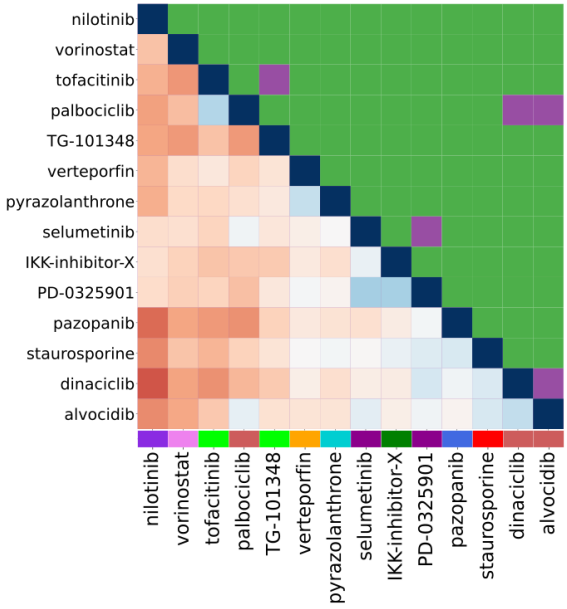
Appendix E- Figure 9. Separation score matrix of PC3



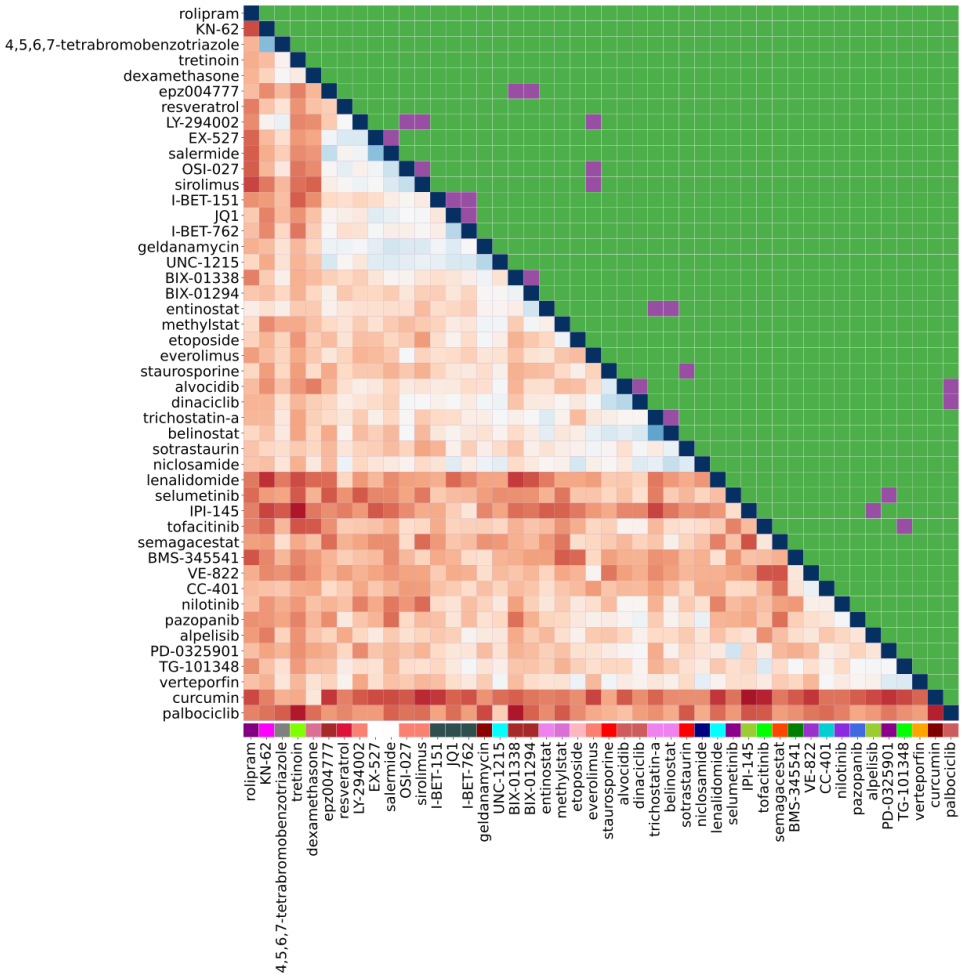
Appendix E- Figure 10. Separation score matrix of MCF7



Appendix E- Figure 11. Separation score matrix of YAPC

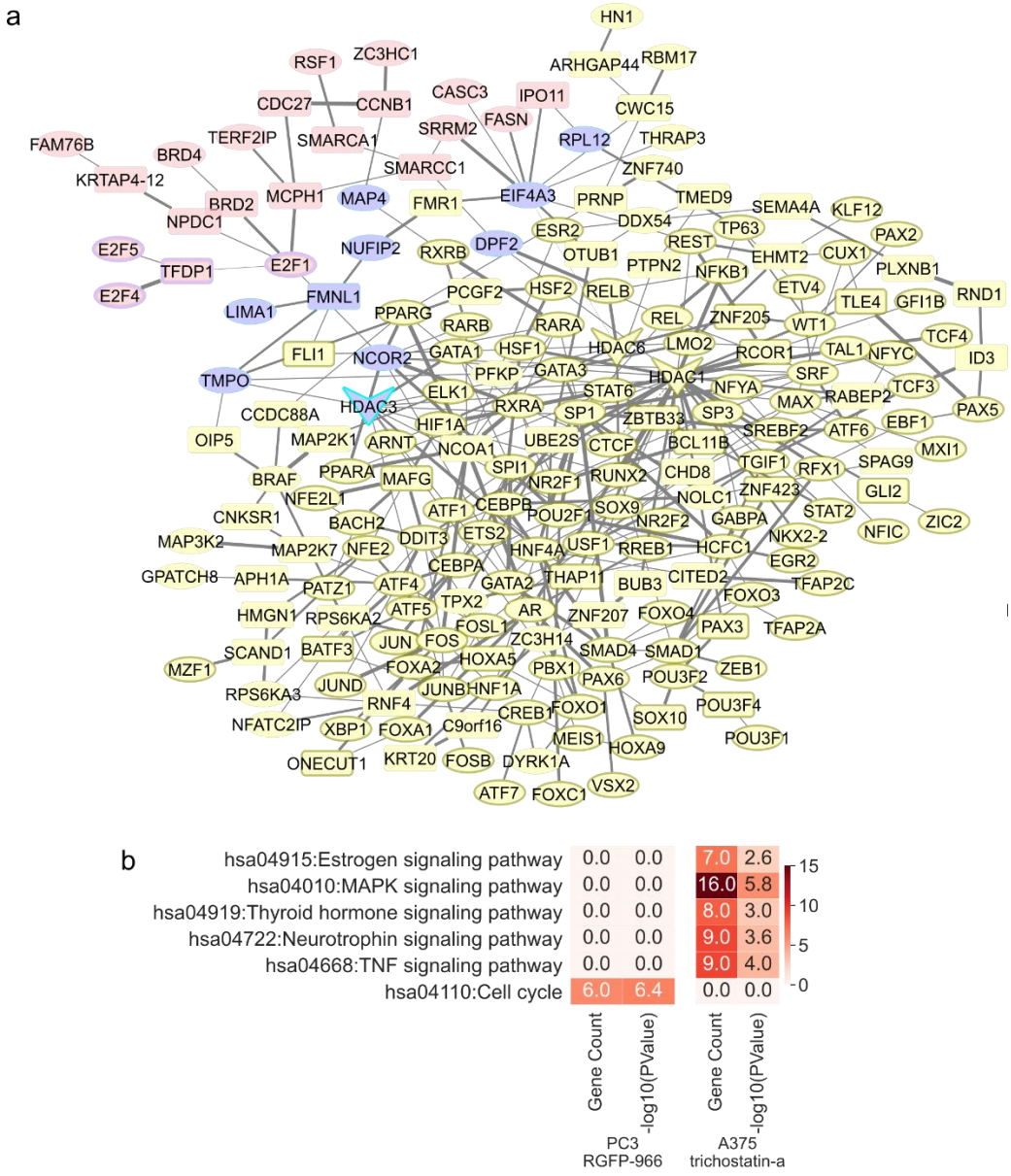


Appendix E- Figure 12. Separation score matrix of NPC

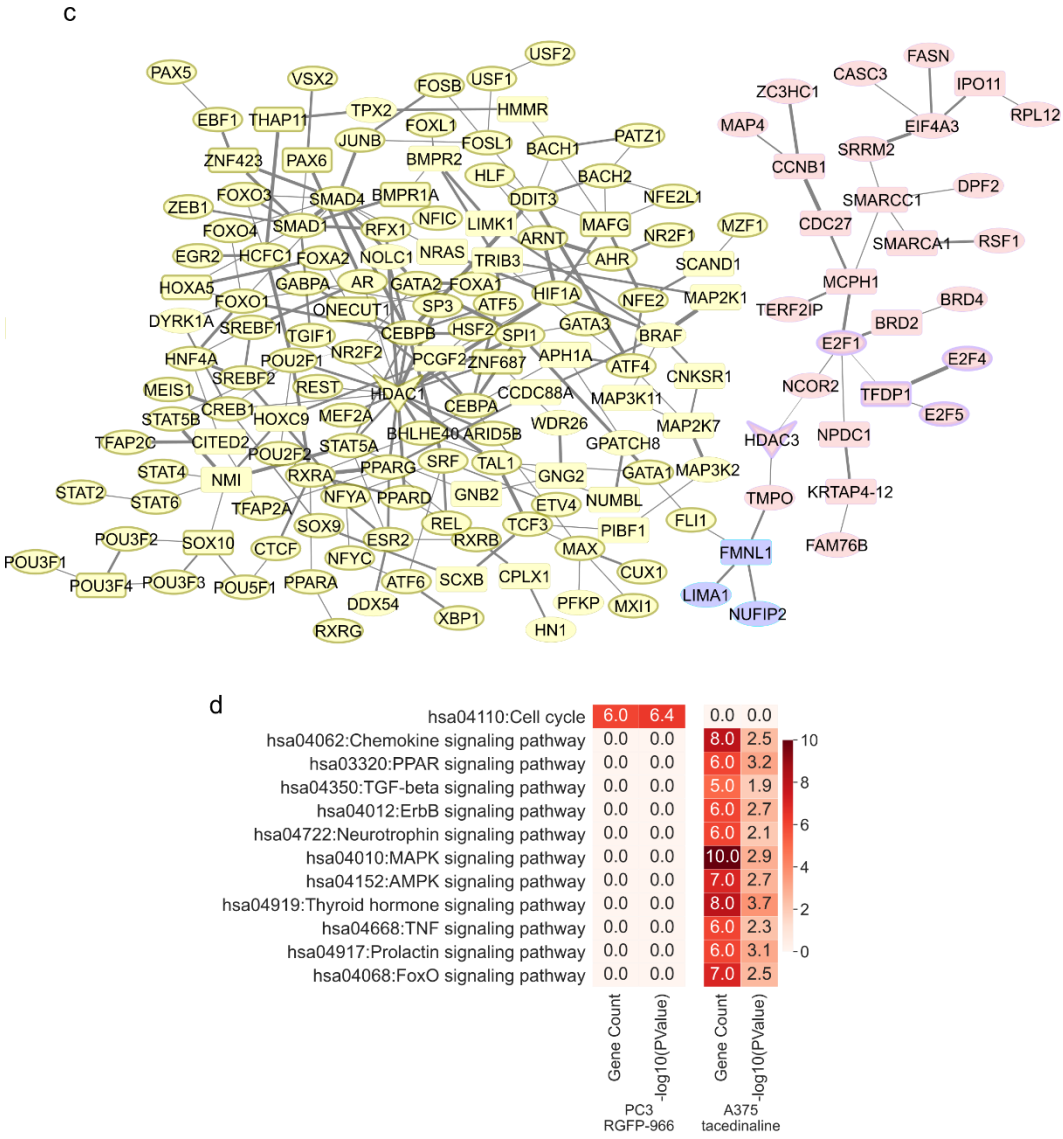


Appendix E- Figure 13. Separation score matrix of A549

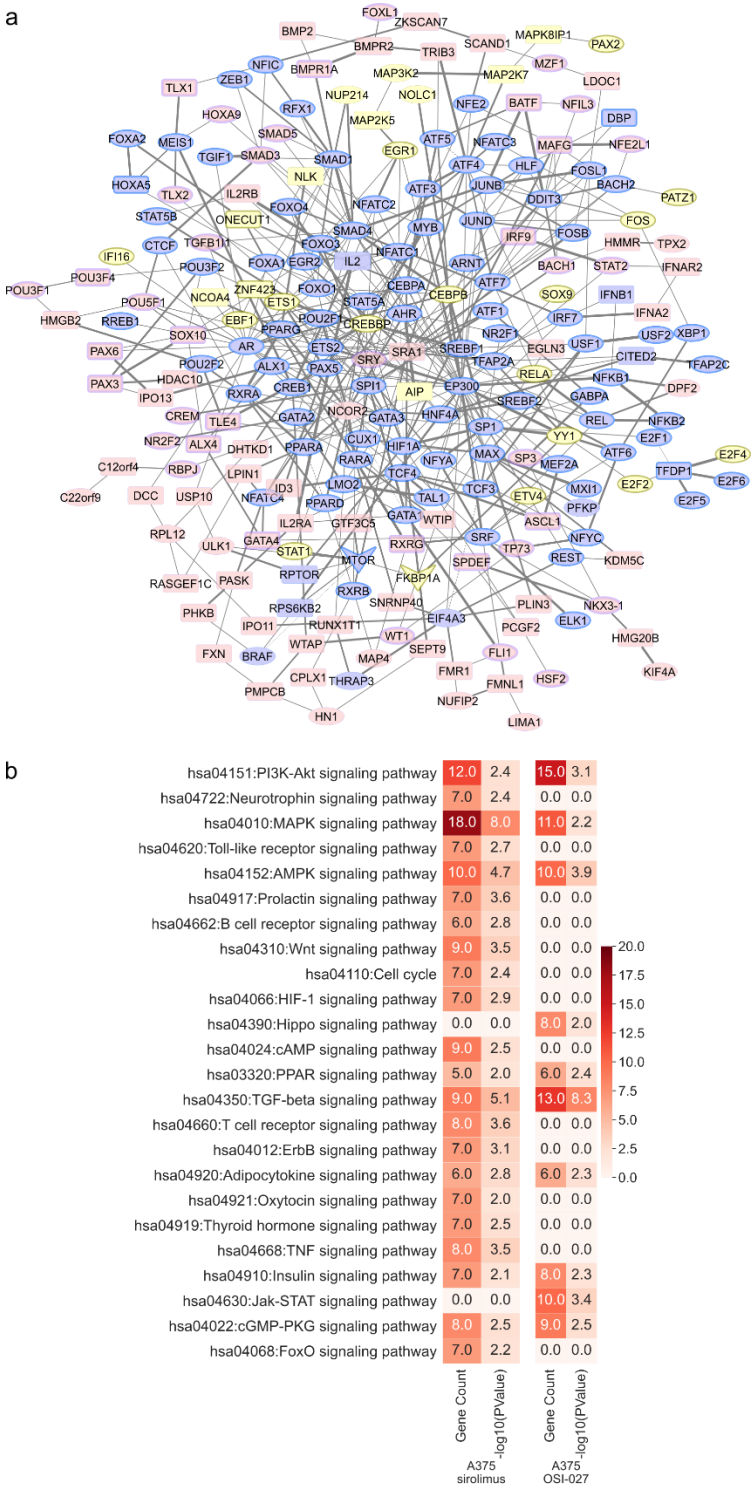
Appendix E- Figure 14. Color code of the separation score matrices



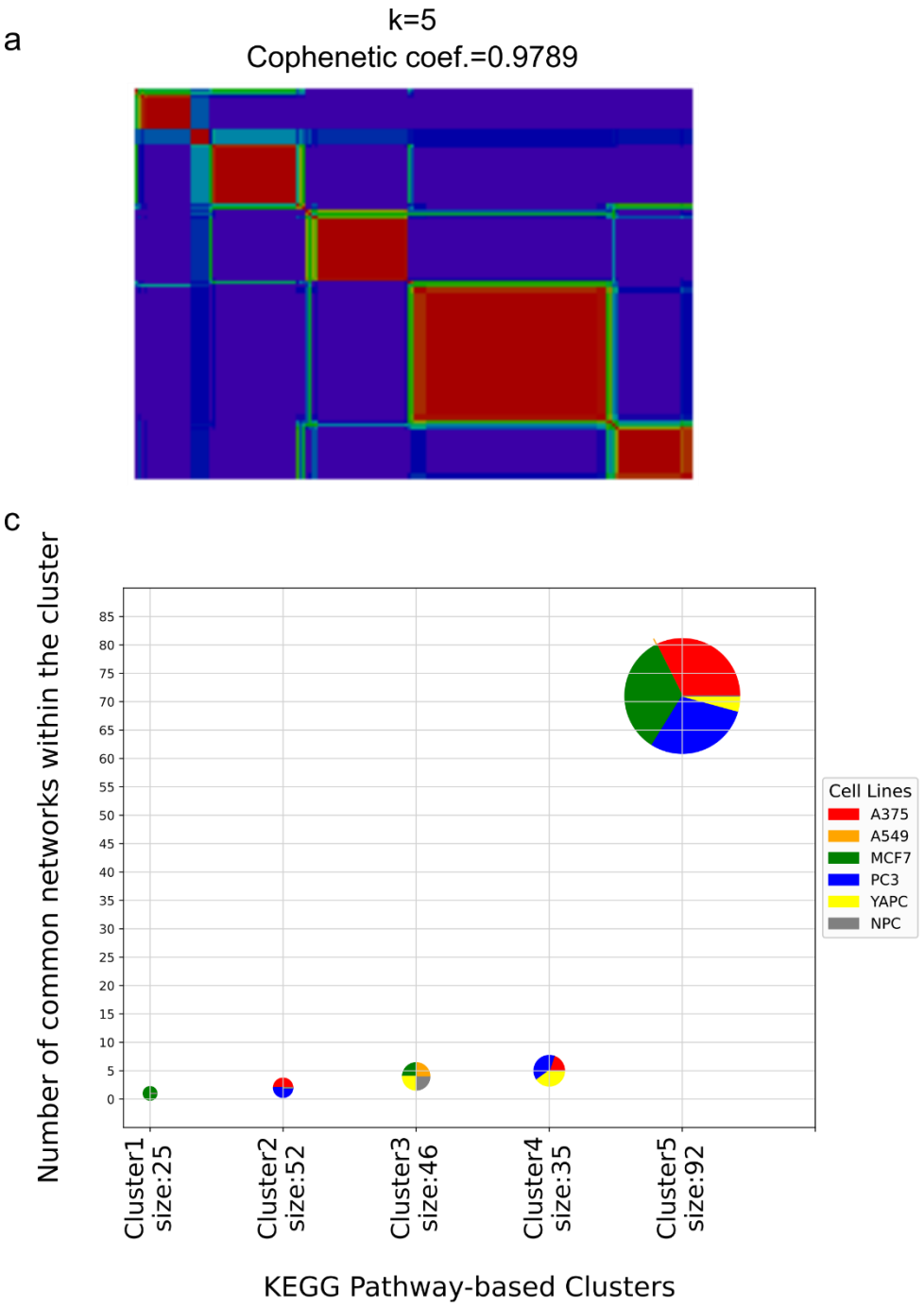
Appendix E- Figure 15. Merged network of two cell line – drug conditions and their associated pathway enrichments **a-b)** PC3-RGFP-966 and A275-Trichostatin-a



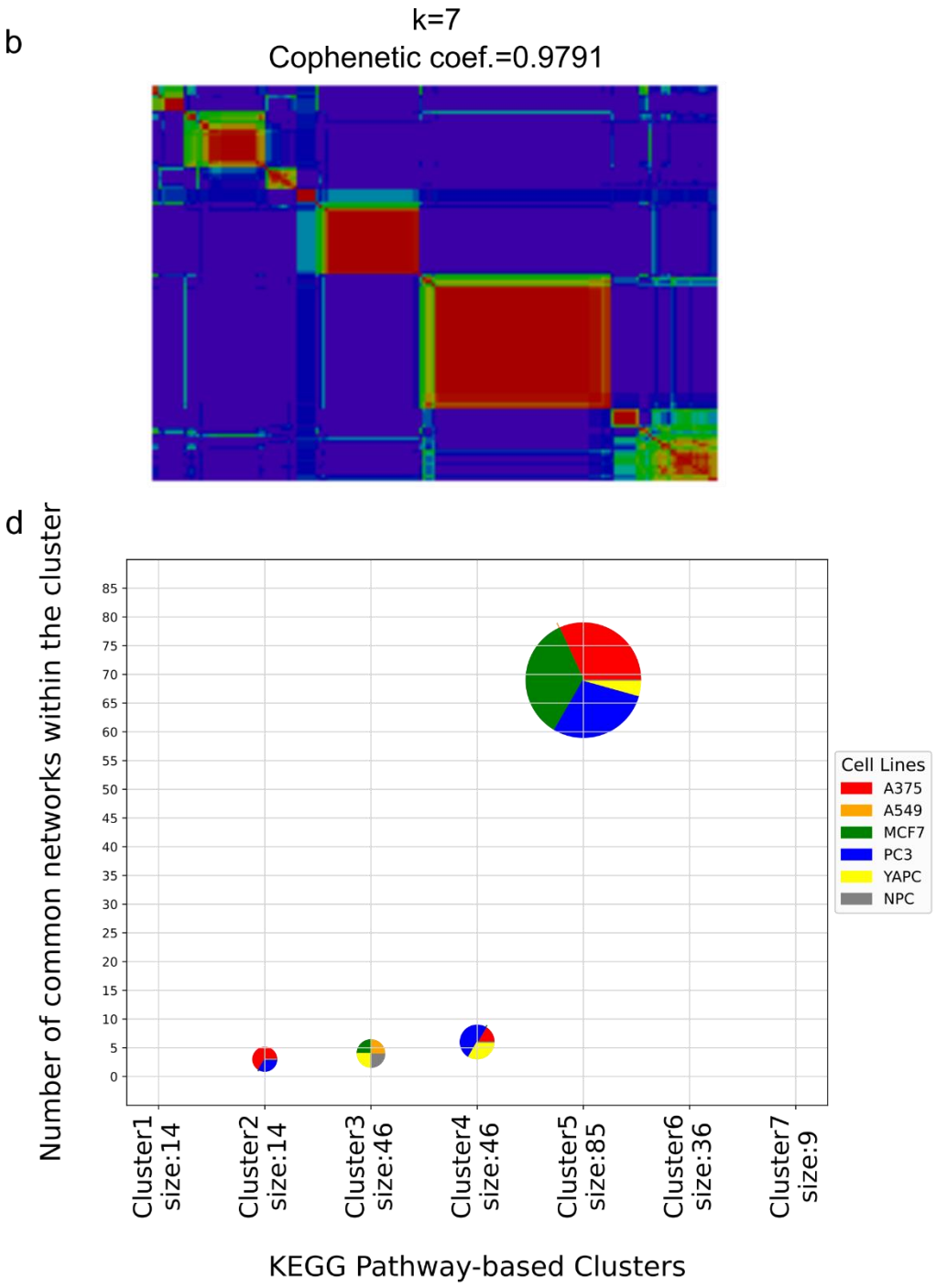
Appendix E- Figure 16. Merged network of two cell line – drug conditions and their associated pathway enrichments **c-d)** PC3-RGFP-966 and A375-Tacedinaline



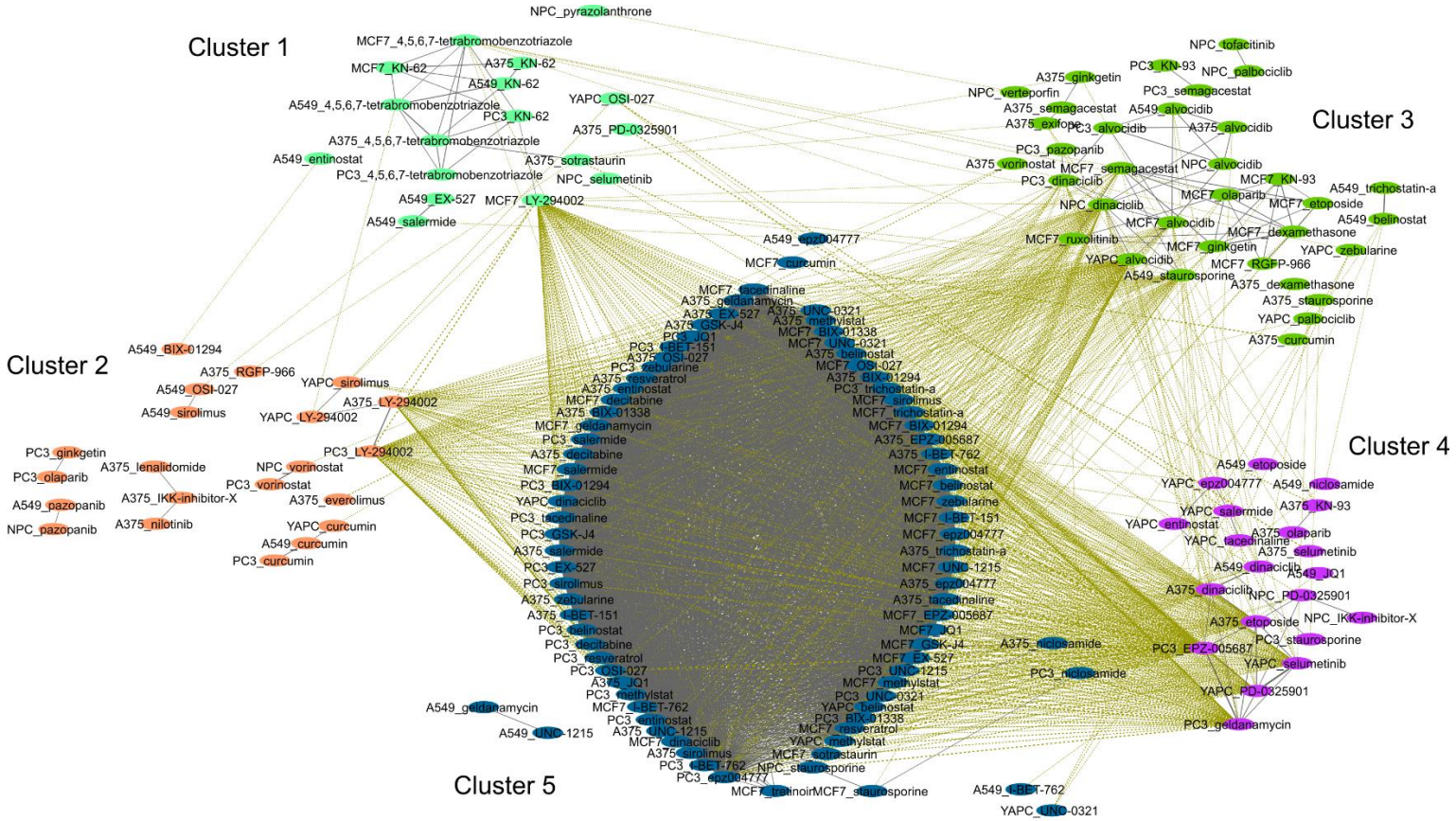
Appendix E- Figure 17. Network pairs (a) and their pathway enrichments (b) of A375-sirolimus and A375-OSI-027.



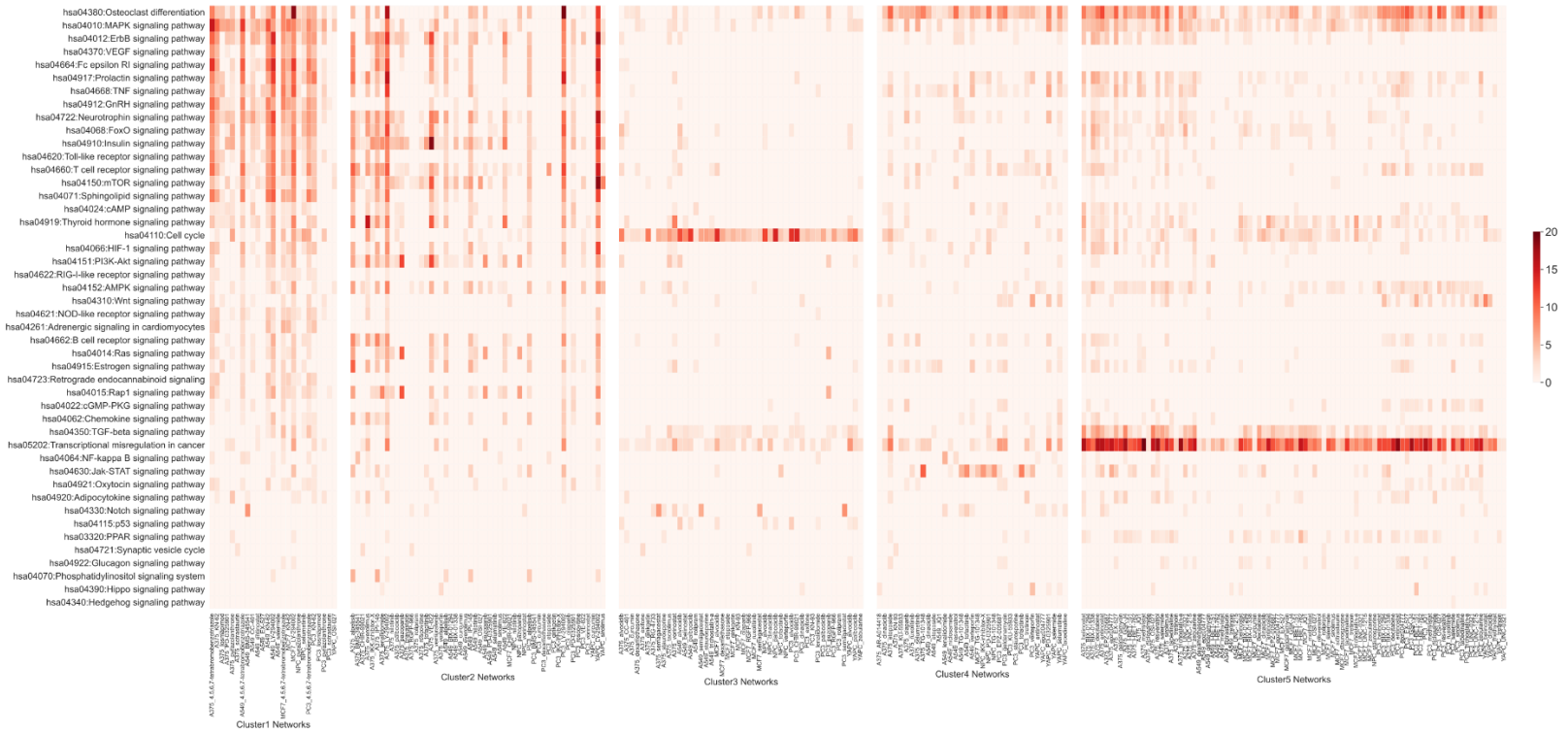
Appendix E- Figure 18. NMFConsensus Clustering Results for k=5 **a)** NMFConsensus output of five clusters **c)** Comparison of five clusters to the list of networks with negative pairwise separation scores (overlapping networks)



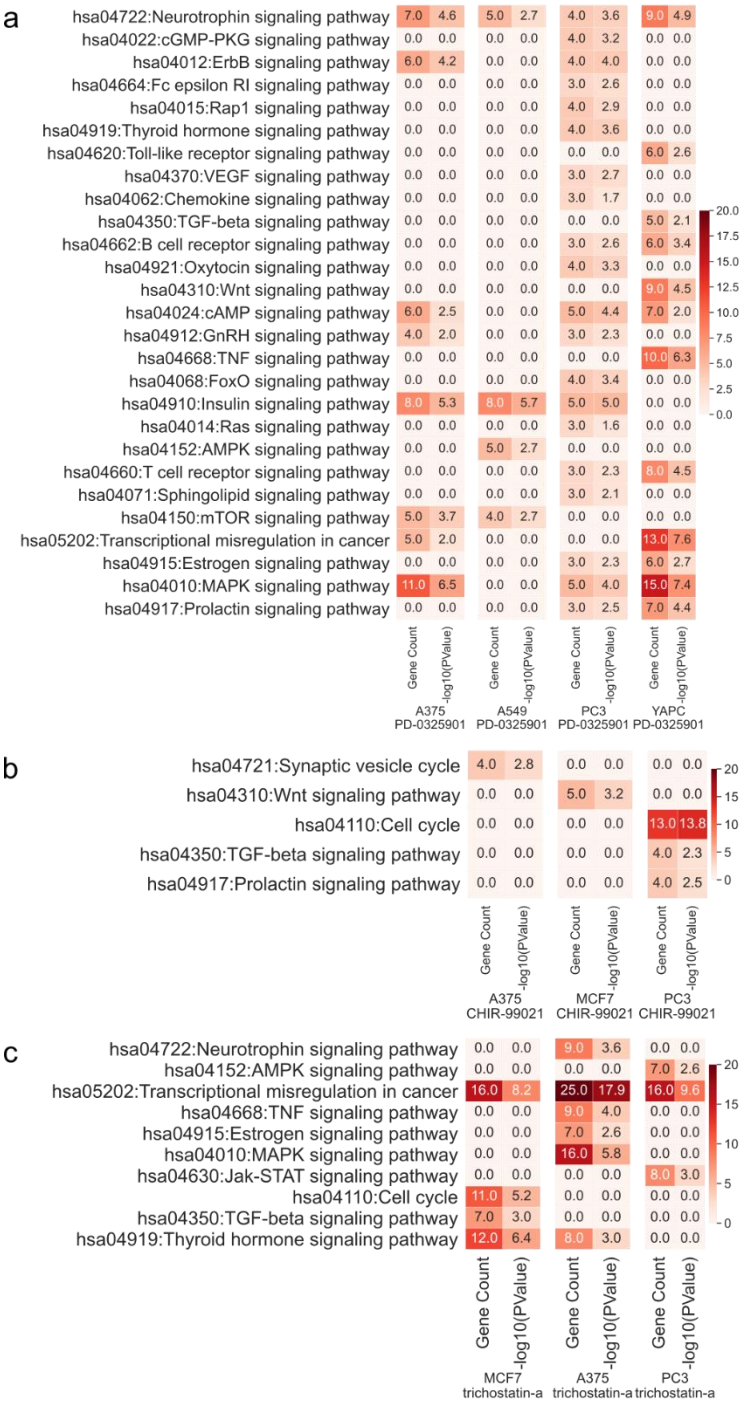
Appendix E- Figure 19. NMFConsensus Clustering Results for k=7. **b)** NMFConsensus output of seven clusters **d)** Comparison of seven clusters to the list of networks with negative pairwise separation scores (overlapping networks)



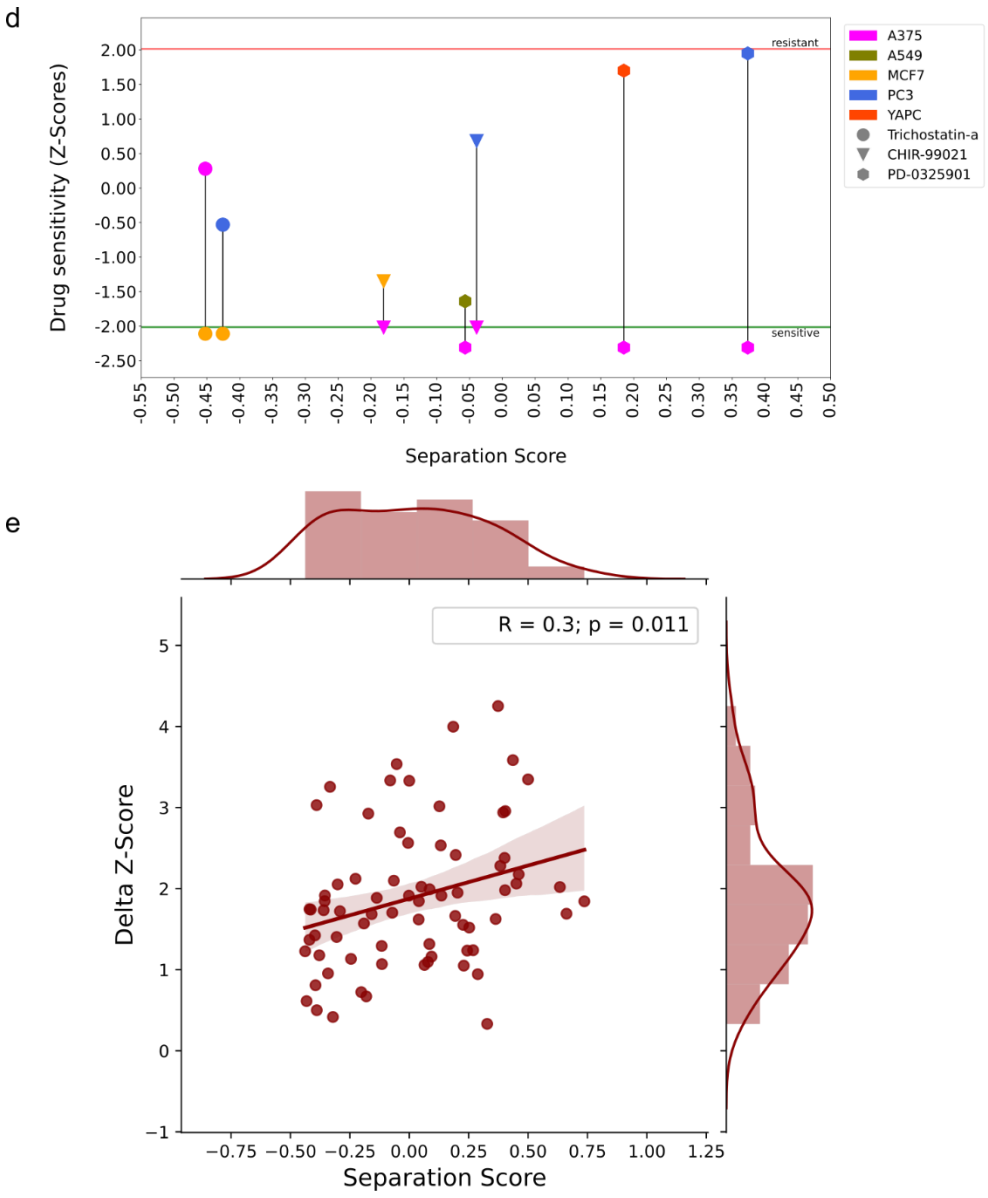
Appendix E- Figure 20. Relationships between drug-cell line pair networks and their clusters **a)** Network representation of KEGG clusters with edges representing overlapping cell line-drug pair networks.



Appendix E- Figure 21. Analysis of pathway-based clusters **b)** Heatmap of KEGG terms enriched in each cell line-drug pair network (KEGG terms on y-axis and cell line-drug pairs on x-axis). Networks are ordered by cluster number and by alphabetical order.



Appendix E- Figure 22. Analysis of drugs that cell lines are sensitive to. **a-c)** Signaling pathways enriched in drug networks per cell type: PD-0325901, Trichostatin-a, and CHIR-99021, respectively.



Appendix E- Figure 23. Analysis of drugs that cell lines are sensitive to. **d)** Separation scores for network pairs of each drug in sensitive cell lines and non-sensitive cell lines versus drug sensitivity plot. Drugs are mapped with different shapes, and cell lines are mapped with different colors. **e)** Regression plot of pairwise separation scores and drug sensitivity score differences for conditions in which one drug of two cell line-drug pair networks has a negative z-score on its corresponding cell line and the other drug has a positive z-score on its corresponding cell line and in which the separation score is higher than -0.45.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name	:	Ünsal Beyge, Şeyma
Nationality	:	Turkish (TC)
Date and Place of Birth	:	7 November 1987, Kırklareli
E-mail	:	seyma.beyge@metu.edu.tr unsalseyma@gmail.com
Address	:	Şehit Osman Avcı Mah. 1071 Malazgirt Cad. No:21/8 Etimesgut Ankara/Turkey

EDUCATION

Degree	Department	Institution	Year of Graduation
MS	Biotechnology	METU	2014
BS	Molecular Biology and Genetics	METU	2011

WORK EXPERIENCE

Enrollment	Place	Years
R&D Manager	TakTek Kimya Ar&Ge Ltd. Şti.	2015 – Present
Operations Analyst	Zypmedia, Inc.	2019 – 2020
R&D Officer	Bio-Dot Nano ve Biyoteknoloji Ltd. Şti.	2011 – 2017

PUBLICATIONS

1. Revision submitted. Beyge Ş.Ü., Tunçbağ N. Functional Stratification of Cancer Drugs through Integrated Network Similarity. *npj Syst Biol Appl*

2. Ünsal Ş., Beyge M. and Son Y. A., Building Dicer1 regulation network of mouse liver hepatocytes, 2012 7th International Symposium on Health Informatics and Bioinformatics, 2012, pp. 88-94, doi: 10.1109/HIBIT.2012.6209048.

FOREIGN LANGUAGES

Advanced English

TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences [ ]

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences [ ]

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics [ ]

Enformatik Enstitüsü / Graduate School of Informatics [X]

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences [ ]

YAZARIN / AUTHOR

Soyadı / Surname : Ünsal Beyge

Adı / Name : Şeyma

Bölümü / Department : Tıp Bilişimi

TEZİN ADI / TITLE OF THE THESIS (**İngilizce /** English) : Integrative Network Modelling of Drug Responses in Cancer for Revealing Mechanism of Action

TEZİN TÜRÜ / DEGREE: **Yüksek Lisans /** Master [ ] **Doktora /** PhD [X]

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. [X]
2. **Tez iki yıl süreyle erişime kapalı olacaktır. /** Secure the entire work for patent and/or proprietary purposes for a period of two year. * [ ]
3. **Tez altı ay süreyle erişime kapalı olacaktır. /** Secure the entire work for period of six months. * [ ]

* *Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir.*

A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature ..... **Tarih /** Date : 06/09/2021