Title
Predictive Optimization of Pharmaceutical Efficacy

Author
Wang, Hann

Publication Date
2014

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Predictive Optimization of Pharmeceutical Efficacy

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mechanical Engineering

by

Hann Wang

2014

Abstract of the Dissertation

Predictive Optimization of Pharmeceutical Efficacy

by

Hann Wang

Doctor of Philosophy in Mechanical Engineering

University of California, Los Angeles, 2014

Professor Chih-Ming Ho, Chair

Drug combinations significantly expanded the opportunity space of druggable genome in cancer therapeutics, but the discovery of novel combinations is still limited by the capacity of our current drug screening technology. To address the challenge, we introduced a data-driven search method called the Predictive Optimization of Pharmaceutical Efficacy, or PROPHECY, for the selection of drugs in combinatorial cancer therapeutics. The user provides the genetic profile, of cancer cell lines or primary cells, and PROPHECY will select optimal drug combinations from a comprehensive list of drugs to meet clinical objectives. The decision making is accomplished by in silico drug screening in which the sensitivities of a cell on different drug combinations are ranked. The predictive model of sensitivity is trained to recognize signatures of information spread in the protein-protein interaction network. Once a comprehensive dataset of drug screening experiment is supplied, the computer could automatically learn interactions between drug targets and disease genes in the information signatures, and infer sensitivity for unseen drug and cell line pairs. We showed that the prediction have high correlation with experimental data by cross validation performed on a dataset of 40,000 entries, which represents 100 cancer drugs applied on 450 cell lines. We also verified the applicability of PROPHECY by performing an in vitro experiment with 36

two drug pairs suggested by the program and a panel of 6 cell lines. PROPHECY not only predicted the sensitivity with high accuracy, but also discovered novel high efficacy combinations and reproduced existing drug combinations. Unlike currently predominant approach of reductionistic drug development, the prediction of drug efficacy is based on network view of proteomic scale data, so can accurately reflect modular activity of the proteome and elucidate target gene interactions in de novo drug combinations.

The dissertation of Hann Wang is approved.

Pei-Yu Chiou

Yong Chen

Ren Sun

Chih-Ming Ho, Committee Chair

University of California, Los Angeles

2014

To my parents . . .

who—among so many other things—

retained my curiosity and joy of research by never forcing me to study

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Prof. Chih-Ming Ho by introducing me the exciting field of combination therapy, and the continual support and guidance he gave me throughout my years at UCLA. I am especially blessed with the freedom Prof.Ho offered us to explore new frontier in science and engineering, and the encouragement he gave us to foster truely original research. I would like to thank the members of my committee, Prof. Benjamin Wu, Prof. Ren Sun, Prof. Yong Chen, and Prof. Pei-Yu Chiou for their friendly guidance, thought-provoking suggestions, collegiality each of them offered to me over the years and for their extreme patience in the face of numerous obstacles. I would like to give special thanks to the PROPHECY team member Thet Phyo Wei, Athena Huang, Meng-Huan Wu, Alejandro Schuler, Bryan Lee, Jing-Yao Chen, and Zhao Yang, this project would not had been as fruitful if not for the inspiration that spark from our endless discussion and dedication they put into making things happen. I would like to thank my fellow doctoral students for their support, feedback, and friendship.

Vita

2002–2006	B.S.E., Mechanical Engineering Department, National Taiwan University.
2008–2010	M.S.E., Mechanical Engineering Department, Johns Hopkins University.
2010–present	Graduate Student Researcher, Mechanical and Aerospace Engineering Department, UCLA.

Publications

Hann Wang, Aleidy Silva, and Chih-Ming Ho. "When Medicine Meets Engineering Paradigm Shifts in Diagnostics and Therapeutics."

CHAPTER 1

Introduction

1.1 Overview

In recent years, biology has under gone a major make over, especially as our technologies for data collection became faster and more high-throughput, and our computational power has grown exponentially in the past decade. New method needs to be introduced in order to tackle this explosion of information. Biologist today can no longer rely on simple over the counter software to solve the problem at hand. As we are crossing the "Excel barrier", meaning that the number of rows of data that we need to process exceed the capacity of excel, more sophisticated tools need to be used.

The main theme of this dissertation is to provide a method to process the massive amount of biological data we have now and to generate knowlege automatically. The original idea here came from my first two years of experience working on the optimization of combination drugs. During these two years, I worked on feedback systems control (FSC), a platform technology for dosage optimization of combination therapy invented by Prof. Chih-Ming Ho. [WYS08] FSC integrates experimental design and model fitting method to guide the search for optimal dosage in a multidimensional dose response surface. The technology allows the rapid search of optimal dosing of 5 to 15 drugs, and can avoid bruit force experiment which requires millions of experiment to test all possible dosages. Despite FSC is a robust and quantitative way for a multi drug design, the selection

of the initial drug library still relies on knowledge and experience of pharmacologist and physicians. Knowledge-based techniques for drug search has historically been an important mean to gain insight into the underlying mechanism, and made important contributions to our understanding of biology. However, in view of the real size of the possible drug combinations that can be generated from existing drugs and chemicals, knowledge based approach for drug selection will be biased toward existing biomedical practices. Furthermore, the number of possible pathways and pathway crosstalks needs to be considered in a complex network system is beyond the grasp of human brain. Thus, a computational method that can extract meaningful information and quantitatively guide the drug selection process is highly desirable.

To construct a small scale library of 5 to 10 drugs from hundred and thousands of chemicals, we developed a computational screening method called Predictive Optimization of Pharmaceutical Efficacy, or the PROPHECY, which took advantage of the currently available public genomic and proteomic data to accelerate the selection process. The selection of combination drugs is based on ranking of the predicted sensitivity, which is infered from the mutation profile of the cell being treated and the target set of the drug combination.

The inference of sensitivity is based on a statistical model trained with machine learning methods in a process very similar to the training of an email spam filter. To construct a email spam filter, the computer has to be exposed to a large amount of email samples and the machine learning algorithm will figure out a statistical model that best describe what is a spam. In our case, we need to aggregate a large amount of training cases of dose response experiment in order for the computer to find out the mapping between the input, which is mutation profile of the diseased cell and target set of combination drugs, and the output, the sensitivity of the cell toward the combination. The data is aggregated from a number of publicly available databases, which includes data for the genomic profiles of the cell lines,

drug-target information, dose response dataset, and protein protein interaction network. One key finding in our inference process is the use of diffusion signature in the network, which can greatly enhance the accuracy of the statistical model.

The dissertation is divided into 7 chapters. In chapter 1, we gave a general overview of the project, and then review the key concepts in the literature that have inspired the work and laid down the ground work. In chapter 2, we introduced the basics of tumor biology and a flavor of post Darwinian biology, which is a convinient mental instrument to guide the thinking process. In chapter 3, we provided the complete framework of PROPHECY, and all things considered in the algorithms is given unambigously in mathematical formula. In chapter 4, we discussed graphs and networks and how the network signature is extracted. In chapter 5, we discussed the regression methods that used to build the machine learning model. In chapter 6, we examined the data that were extracted from the databases, and specifically the way we assembled the data. In chapter 7, we provided a full proof of the model by both numerical and physical experiment. The network signature's specificity in classification was demonstrated by clustering method. We then provided detailed characterization on the quality of the regression result. Finally, we discussed a physical drug combination experiment that provided solid evidence of the discerning power of PROPHECY.

In this chapter, we began by reviewing existing methods for drug combination selection that attack the problem from different angles. We hope that through exposing readers from these different views we can help the reader understand the complexity involved in combinatorial drug selection and also the various assumptions made on the inception of the works. We then briefly review a few methods that were used to make predictions on drug response. Finally, we will introduce the concept in network medicine and pharmacology, a booming field that provide a wealth of novel idea that inspired many aspects of this project.

1.2 Selection of Combination Drugs

The predominant strategy in curing cancer is to eliminate as many cancer cells as possible, while sparing normal cell in the human body. However, the traditional chemotherapy attacks regular bioprocesses which also exist in regular cells, exploiting only the fact that the cancer cells divide more rapidly. So, the toxicity profile for chemotherapy is usually hard to manage. After 70th when biologicals gradually became practical, there immerge many novel targeted therapeutics which target the proteins that only expressed in cancer cell. However, targeted therapy only gain limited success in treating cancer, mainly due to the redundancy and robustness present in the cellular network. So, in order to better regulate the complex biological system, combination therapy, a neo-classic approach that reimagine the possibility of mixing chemicals, gradually regain momentum as researchers discover that multiple perturbation is more likely to better regulate the underlying system.

1.2.1 Key advantages of combination drugs

Combination is more selective toward cancer cell Recent studies on network theory suggested that multiple weak hits can disrupt the signal transduction of the network and can more effectively damage the integrity of the complex system, and systems biology studies on yeast and E. coli network supported the result. [ACP05] Researchers also found that two drug combinations will have the potential to induce "synthetic lethality" in cancer. Synthetic lethality is the sensitization of cancer cells by inhibiting two genes which are essential to the growth advantage of cancer cells but non-essential in the physiology of regular cells. [Kae05]

Combination therapy can solve the crisis of target depletion The current preclinical cost of developing a new drug is 1.5 billion dollars. The R&D cost of the pharmaceutical industry has grown exponentially, but the number of the approval from FDA remains constant. [Mun09] We argue that the surging cost came partially due to the fact that new therapeutic target has come to a depletion after the human genome project. Nearly all druggable targets have been discovered. [OAH06] However, facing the new situation of target depletion, the current industry still respond it with the old thinking of targeted therapy. More high-throughput experiment was wasted, exploiting the same mechanism because each chemical has its own defined target. One obvious way to solve this crisis is to introduce the thinking of combination therapy. So, a single target of intervention can be evolved to multiple target and thus introduce a larger possibility space for perturbation.

1.2.2 Difficulty of identifying and optimizing a combination drug

The problem of finding combination drug can be broken down into two parts. First, a set of chemicals needs to be identified to narrow down the search space from a large library of chemicals. Second, the narrowed downed library of drugs can then be optimized for best dosage combination and dosing schedule. Our current effort focus on solving the first problem, and we will briefly review the second problem as there are a number of valuable lessons which can be learned from solving the dosing problem. The scheduling of drugs is beyond the scope of this dissertation, as it is still an open ended question in the field. For general review for the strategy of choosing combination drugs, there are a handful of good reviews that the readers can refer to. [ZLK07, FCD10, ABW12, Azm12, BGL11, Kit07a, ABW12]

1.2.3 Approaches for combination drug optimization

The approaches for assembling combination drug is difficult to categorize, as the assumptions on the model systems, physical and mathematical tools, and the objectives of the studies have a wide range of spectrum. As such, we listed some of the most promising approaches in the field in the hope of being exhaustive. However, the optimization problem in terms of chemoinformatics is beyond the scope of this dissertation, so it is not enlisted in our review.

Combination based on properties of existing combinations Drugs can usually be accurately classified based on their properties including their set of targets and the therapeutic indications. This class of methods exploit the fact that certain properties are often enriched in successful combinations, and can perceivablly be used to predict new combinations. Zhao et al. studied the feature pairs that arises in FDA approved combination, and found that one of the feature dictates that certain protein pairs are repeatedly being targeted in approved combinations, and the features identified can be used to narrow down the search space for combinations. [ZIZ11] Yildirim et al. constructed the drug-target network by linking approved drugs with known therapeutic target. They found that strong local clusters were formed in the network, indicating that many drugs share the same targets and are follow on drugs. They also observed that the trend of drug development is going toward a more functionally diverse polypharmacology. [YGC07] Vazquez et al. focused on uncovering drug combinations which can eliminate a heterogeneous population of cancer cells. [Vaz09] They found the drug by identifying the minimum set of drugs which can cover more of the lethal targets of the heterogeneous cancer population, and thus transform the selection problem into a minimum hitting set problem. Their method can help the design for dealing with heterogeneous population of cancer cells, but ignores the interaction of drug targets in the context of molecular network.

Bruit force based approach for combination search Although knowledge based approach can leverage our knowledge for biology and greatly accelerate the discovery of new drugs, the complex behavior and the context of biology is often ignored or stripped from the whole picture. So, knowledge based approach often produce less than optimal outcomes at later experimentation of the system, generating unexpected result. A robust method which can reflect the biological context is therefore preferred. Bruit force method refers to the exhaustive experimentation of the entire possibility space. Given the explosion of dosage combinations, bruit force method is limited to study pairwise combination in general. Borisy et al. exploit this approach by applying the bruit force search on multiple biological systems and several hundred compounds. [BEH03] The synergy between drug pairs are then calculated which cannot otherwise been deduced from single drug study. Even though bruit force approach can hardly be extended to multi-drug scenario, it is currently the only experimental way to reliably probe a large set of initial chemicals, ranges from tens to hundreds of compound.

Network pharmacology Network pharmacology concerns the combination of network biology and polypharmacology. [Hop08, SFS09] As mentioned earlier, it was shown that multiple perturbations in the network can be more effective than single targeted inhibition, as demonstrated by systems biology studies and network theory. The challenge, however, is how to leverage the knowledge of network properties to find out the principles that consititute effective perturbation of the network. This new line of study sits in a very special place amidst all spectrum of method. The network pharmacology take advantage of the current advances in omics technology, and use the most condensed form of the omic information, the network map to help the development. Also, the information volume has been pushed to an unprecedented scale. A comprehensive review of the subject can be found in [CKK13]. Wang et al. examined the location of targeted proteins in the

genetic regulatory network, and found that drug targets of combination therapy tend to be neighbors in interacting pathways in the network. [WXS12]

[JZM09, KSB07]

Modeless approach to search for combination on a response surface Feedback system control (FSC) is a method of phenotypic screening that leverages the power of stochastic search algorithm to optimize the dosages of combination therapy. [WYS08, SUY09, HH10, TVH11, AYF11, YAF11, VTH11, WBH11] It has been demonstrated successfully at a number of biological systems including cancer, viral infection, stem cell culture, and herbal drugs. At the core of method, a complex response surface is assumed to exist as a manifold in the multidimensional space of the dosages of the chemicals involved. The response surface represents the objective function which is determined by the goal of the therapy.

Combination search based on statistical model of the response surface For modeless approaches of optimization, as it is iterative, it still takes considerable time to perform the experiment and find the global optimal dosage, making the method unsuitable for applications such as animal studies or personalized medicine. In order to over come the difficulty of iterative searching, the modeless method is improved by incorporating our knowledge of the underlying model. As most of the complex systems are evolved to adapt to multiple stimuli, and become robust in the transition of states, the response surface should appear to be smooth in the multidimensional dosage space. By using the assumption of robustness, we can approximate the response surface by using simple mathematical function such as polynomials or nonparametric methods such as Gaussian process. Special sampling methods, such as orthogonal array design or Latin hypercube design are then applied to probe the surface. Statistic model based approaches can greatly reduce the time required to optimize the

drugs, so is gradually becoming the default choice for combinatorial optimization. [DXH13, DSS11, JDX13, WSH13, YZD13, HDM13, PNV13] The approach we proposed in this dissertation drew inspiration from this approach by building statistical model of the phenomenon, but the way the model is built is significantly different from this approach.

Combination optimization based on systems biology approach The systems biology approach is closely related to the network approach, but different in the scope and detailed mathematical modeling. The systems biology approach considers more detailed maps of molecular network and how the wiring, feedback, feedforward loops, cross-talk were constructed in the molecular network. The approach also usually required detailed knowledge of the reaction constants involved in the chemical processes of the network.

Yang et al. simulated the dynamics of arachidonic acid metabolic network under the influence of drugs, and find the drug targets that can best restore the network from the disease state to normal state. [YBO08] Wu et al. identified optimal combinations for type 2 diabetes by ranking the possible combination with a custom made score. [WZC10] They first generate a predicted gene expression level of the applied combination, and then use the gene expression level of the subnetwork to infer the possible symptoms and efficacy, which in turn is used to calculate the score. They successfully identified the combination of metformin and rosiglitazone, which is a clinically approved drug that used to treat type 2 diabetes. There are many excellent reviews on this type which can be found at [FSN06, BIB10, CSE12].

1.3 Prediction of drug response

Making predictions on the drug response often requires building a mathematical model. There are a few types of prediction varying in complexity, as shown in figure 1.1. Here, we are focusing our attention on machine learning type of model that makes pointwise predictions (instead of dynamic output) based on training data.

Building a machine learning model means building a mapping between a set of predictors (or input, independent variable) and a label (or output, dependent variable). The label can be a numerical value or class labels. When the label is numerical value the learning problem is called regression, and in the case of class labels the learning problem is called classification. Throughout this dissertation, we will refer to the predictors and labels very often, and sometimes interchanged with other names.

1.3.1 Feature Selection

Feature selection is the process of extracting meaningful quantities from the predictors, and features are the actual input that a machine learning model takes to map to labels. Often times, predictors can be used directly as the feature for the learning algorithm. For example, the length of the spring in a spring mass system can be used directly as the training data to fit the model. However, in many real world applications, when the predictors have nonlinear interactions or does not directly cause the label, feature extraction is often required to process the data for better fitting. A good feature is reflective of the underlying mechanism and often presents high linearity with the output. In any model fitting or machine learning scenario, feature selection represent if not the most important part of the whole analysis. In other words, feature selection can either make or break a fitting algorithm.

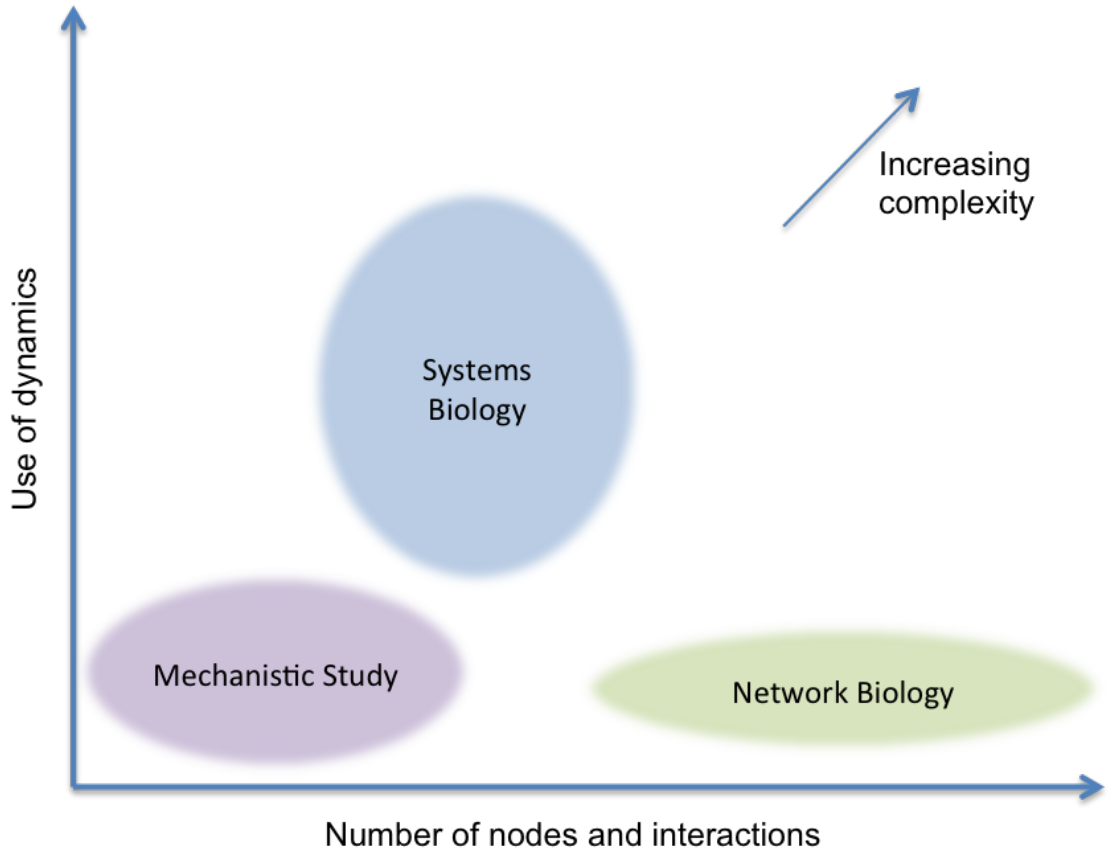


Figure 1.1: The complexity of predictive models. Network biology usually use fixed point prediction, while systems biology provide dynamical simulations. However, the scope of network biology can usually encompass the entire proteom while systems biology can only simulate a subset of the network.

The raw data for the prediction of drug response span from macroscopic trait of an organism, organ, tissue, cell and all the way to molecular signature. The macroscopic features for a disease includes lineage [NNC00], race [JSG05], gender[HGH06], age [GLP95], PET scan [WOB01], MRI images [PGL05, TBT04, DDG02], response of in vitro assay [AGS80, KYM08]. Recently, due to the advances in several technologies in molecular biology, especially those of microarray, and next generation sequencing, larger and more reliable data set started to emerge. Common molecular signatures includes gene mutation profile, DNA

copy number variation, gene expression profile [PDB06, LCP06, BCS12, GEH12], epigenetic signals [JB02], proteomic profile [HPF08, NWG11], post translational modifications [MJ03].

1.3.2 Model fitting Algorithm

The predominent method in molecular level generally exploit the gene expression signature as predictors of the drug response. Here we review some of the most important models through out the years, and the goal in which they intended to achieve. The eventual goal is important as each study might have different application in mind when they developed the method and thus different assumptions and expeirmental collection methods.

connectivity map The connectivity map is a plateform for finding the similarity between the gene expression signature of disease, and drug perturbation. [LCP06] At the core of connectivity map was their intention to capture the trantient perturbation of gene expression caused by drug action. The best molecule or set of molecules that has the opposite perturbation compared with disease can then be used presumably to reverse the disease phenotype. Connectivity map serves as an interesting idea that capture the transient nature of perturbation. In the dynamical systems view, this is analogous to identifying the "force" that is applied on a state, and the state is the current expression level of the cell. However, the method does not take into account the genetic profile of the disease, which in the case of cancer, is vital to define the malfunctioning parts or i.e. the epigenetic landscape of the system. [1] So, the method will have limited use for sensitivity prediction of cancer drug combinations, as it does not take into account the long term and high order interaction of drug actions with disease genes.

Sirota et al. applied similar approach as the connectivity map for drug reposi-

[1] Please refer to chapter 2 for epigenetic landscape.

tioning. [SDK11] This type of method is strong in illucidating the mechanism of action of the drug. However, it is not clear how it can be applied in combinatorial drug scenario as the gene expression pattern will change unlinearly with the application of combinations due to principles of enzymatic reactions.

Drug sensitivity biomarker discovery through elastic net regression In a back to back publication by Barretina and Garnett et al., they took the scale of genomic study to the next level by integrating various genetic profile and by collecting the signatures from 947 and 368 cell lines each.[BCS12, GEH12] The goal of this study is to identify biomarkers that correlate with drug sensitivity. This work has a significant difference with our work as it is not applicable to the comparison between different drugs as in the scenario when combination drug needs to be searched from a library of drugs. It is geared toward distinguishing the difference of sensitivity between cell lines for a given drug. Nontheless, the study inspired the use of signatures representing gene mutations in PROPHECY, and the comprehensive database of mutations were borrowed as training data here.

One thing worth notice is the gene expression signature from connectivity map is not taken for each chemicals, so it has different meaning in terms of its role in a dynamical system. The gene expression level here respresents the initial condition of the cell instead of the force that exerts on the cell lines.

1.4 Network medicine and network pharmacology

Biological networks are an abstraction of the collection of interactions between any component that are related to bio processes. The components can include inter- or intra cellular components, and also can be extended to metabolite and chemicals. A comprehensive review can be found in [PWS08, GOB10, LHM09]. The section here is intended to review the aspect of biological networks that are

relevant to combination drug selection.

1.4.1 Protein-Protein interaction network

Protein-Protein interaction network (PPI network) is a graph structure on which the linkages between all types of proteins in an organism are mapped. The information in the PPI network is usually global and on a proteomic scale, so it is also called the interactome. The nodes in a PPI network are proteins, and currently the edges are represented probabilistically as confidence scores. This type of probabilistic linkages can not only represent direct physical interactions between two types of proteins, but also other function interplay including interaction affinity, co-expression level, co-localization in cellular compartment, and regulatory control of one protein by the other. [DF10, JPG11, SW11, VCB11] Current PPI networks have many versions and hosted by different databases. Each of these databases have different set of methods that is used to infer the linkages between proteins. Common methods for linkage assesment includes protein binding assays, co-expression analysis from microarray data, and text mining from the literature.

PPI network is a central piece of information in the construction of PROPHECY, as protein is highly relevant with disease pathogenesis and therapeutics. [NK10] Understanding the structure of the network can help us extend the perturbation of disease genes and chemicals from the causal nodes to the neighbring nodes and then the entire network. Therefore, the limited view of "one gene, one functionality" can be scaled up to the proteomic scale.

Proteins are the central machinery that operates most of the vital functions in a cell. Vast majority of genetic diseases are associated with the regulation or functional change of proteins; most of these diseases are oligogenic or polygenic. The vast majority of drug molecules also target proteins instead of other biomolecules.

There are other types of biological network in use now, including genetic net-

work, signaling network, and metabolic network. Why did we choose PPI network intead of the others? There are several major advantages of the PPI network. First, the structure of PPI network is relatively immutable compare to other network. The linkage of PPI network represent physical processes that will always hold true for all cell types in human body. The weight of the linkage might change due to mutation of the corresponding genes or post translational modification of the protein, but it is relatively small compare to changes in other network, such as rearrangement of the linkages, which is not uncommon in genetic network. Secondly, the topology of the PPI network were shown to be related to disease genes and prominent drug targets. Lastly, PPI network is in proteomic scale, so cover a wide range of cellular processes.

However, there are potential pitfalls associated with the data quality of the PPI network. The data completeness of PPI network is a challenging issue, as can be seen from the new interactions that were discovered historically each year. We cope with the data imcompleteness by using STRING, the most comprehensive database known to our knowledge. There are also errors that plague the structure of the PPI network, including false positives, sampling biases, compartmentalization in the cell, and coupling with other bio processes, to name a few. Error in PPI network is common and the influence can be minimized by choosing a feature extraction method that is insensitive to false positives. PageRank was selected as the method of feature extraction in PROPHECY, and was proven to be insensitive to modification or errors in the network.

1.4.2 Network-based tools for the prediction of disease genes

The predominent method for the discovery of disease genes are through linkage mapping or genome wide association study. However, the identification of the causal genes or disease associated genes remains challenging. Thus, a set of network-based methods were developed to assit the discovery of disease genes by

taking advantage of the knowledge we have for biological network. The methods can be categorized into linkage methods, disease module-based methods and diffusion-based methods, [KBH08, VMR10] among which diffusion-based methods have the best performance. [BGL11, NK10] Later on in the construction of PROPHECY, we extend the idea of network diffusion into feature extraction in machine learning, and proved that network diffusion can uncover extra information useful for model fitting of treatment sensitivity.

1.4.3 Network tolerance study

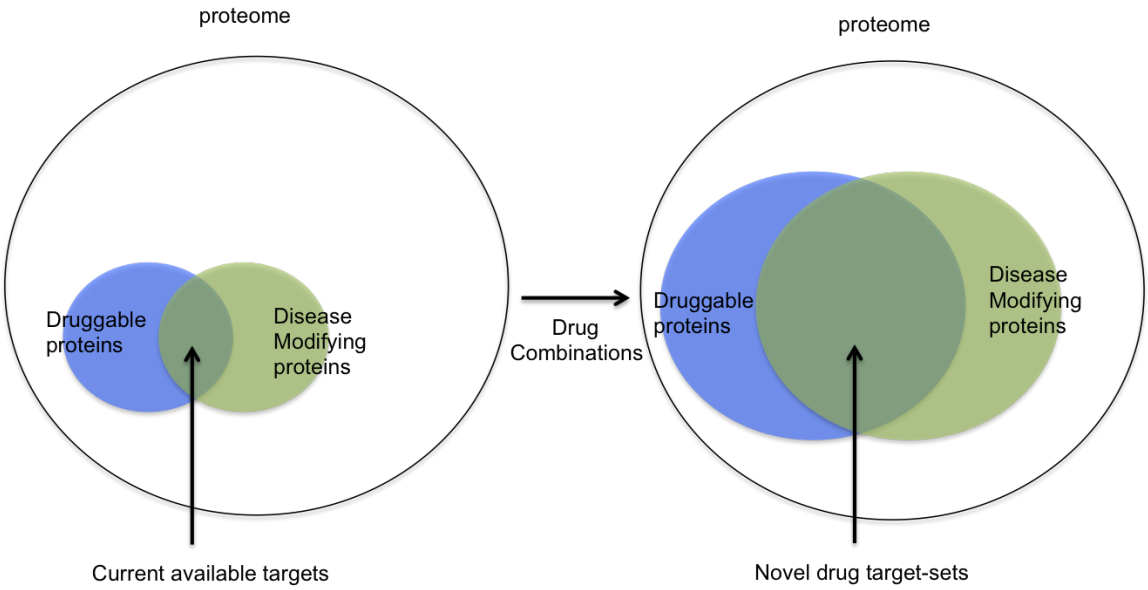


Figure 1.2: Venn diagram showing the expansion of target space. Drug combinations can effectively provide more protein target sets available; due to the lower toxicity of weak hits, combinations can also increase the druggable target sets. Therefore, the possible drug target sets can be significantly increased.

Recently, theoretical study on complex network systems has revealed a robustness against random errors or attacks. [AJB00] The conclusion challenges the current practice of high-throughput screening for targeted therapy, as cellular network can often bypass the single target attack by the built in redundency

in the neighboring pathways. [Kit07b] It is also found that complex system are vulnerable to attack on hubs in the network. However, hubs are generally not druggable target as they usually encodes essential proteins in the network; an attack on these node will result in systemic failure, diseased cell or regular cell alike.

A multiple target approach is found to be more effective in modulating disease. [CAP05] It was found that multiple weak hits on the network can more effectively decrease network efficiency, the information carrying capability, and better control information flow than single targets. In addition, weak hits on the network does not completely disable the function of essential genes, so is less toxic compare with single target strategy. A combination approach that attack multiple nodes in the network can dramatically increase the druggable proteome, but also pose challenge to the search methods as the search space of combinatorial problem is significantly larger than single target problem.

Therefore, to facilitate in silico assembly of drug combinations, new computational methods need to address the topic of multitarget design, particularly aiming at prediction of network efficiency. In PROPHECY, the inclusion of multitarget effect in the network is integrated from the beginning when we started to design the software. Even for single drug or targeted therapy, the predictors of drugs in PROPHECY includes all possible bindings of proteins and genes where data is available.

1.5 Measurement of Combination efficacy

According to [GBP95], the most highly used reference models for synergism are Loewe additivity and Bliss independence. We will briefly introduce these two measurement, and then introduce the combination index of Chou and Talalay. Although these models are all controversial, there is by far no consensus in the

literature which one shall be the golden standard. We still enlisted them because they inspired some of the techniques later on applied in the model fitting, and they can be a good reference for later on to interpret the mechanistic result of the predicted combinations. Note that, in an actual toxicology study, usually multiple methods will be used to evaluate the synergism or antagonism of the proposed therapy. [GJ07] The readers are encouraged to dig into the topic of synergism. A number of good reviews on this topic can be found in the literature. [FSN06, KBS05, GBP95, ZLK07, ZLK07].

Bliss independence The Bliss independence is applicable when two compound is independent, meaning that they take effect by binding to completely different targets. The efficacy for two independent reagents can be calculated as

$$E(x, y) = E(x) + E(y) - E(x)E(y) \tag{1.1}$$

where $E(x, y)$ is the effect of the combined drug, $E(x)$ and $E(y)$ is the effect for individual drugs at x and y concentration when applied singly.

If the combined effect of the combination drug is higher than what is predicted in equation 1.1, the combination is considered synergistic, if equal, additive, and if lower, antagonistic. The resulting formula is simple for Bliss additivity, but the restriction for independence is high, as many combination drugs share common targets.

Loewe additivity The interesting thing about Loewe additivity is that its assumptions are exactly the opposite compared with Bliss independence. In Loewe additivity, it is assumed that all of the drugs acts on exactly the same set of targets, but can exhibit different potency. We will not list the equations for Loewe additivity here as it is equivalent to the combination index model of Chou and Talalay in the mutually exclusive drug case.

Combination index Combination index is a concept arises when a researcher needs to compare the efficacy of individual drugs and a combination of them. Complication exists due to the fact that the dose-response relationship is usually nonlinear, and more often follows a sigmoidal curve. So, it is non-obvious how one compare the dosages of one drug and a drug combination. One intuitive way of solving the problem is to look at a simpler system, namely the reaction of enzymatic inhibition. Chou and Talalay found the medium effect equation, which under arithmetic manipulation can represent multiple enzymatic equations. [CT84] The medium effect equation write as follows

$$\frac{f_a}{f_u} = (\frac{D}{D_m})^m \tag{1.2}$$

where f_a and f_u are the portion of the system effected and unaffected, D is the dose, and D_m is the median effect or the IC_{50} of the inhibitor. For multiple enzyme system, the median effect becomes,

$$[\frac{(f_a)_n}{(f_u)_n}]^{\frac{1}{m}} = \sum_{i=1}^{n} [\frac{(f_a)_i}{(f_u)_i}]^{\frac{1}{m}} \tag{1.3}$$

where $(f_a)_n$ and $(f_a)_i$ are the fraction affected for n drug combination and ith individual drug respectively. Given the multi-inhibitor form, we can define the combination index as

$$CI = \sum_{i=1}^{n} \frac{D_i}{D_{x,i}} \tag{1.4}$$

where D_i is the dosage of the ith drug applied, and $D_{x,i}$ is the dosage of ith drug when applied singly can produce the effect x as the combination drug. When

$$\begin{aligned} CI < 1 &\quad \text{the combination is synergistic,} \\ CI = 1 &\quad \text{the combination is additive,} \\ CI > 1 &\quad \text{the combination is antagonistic.} \end{aligned}$$

This can be deduced from the multi-inhibitor form of the median effect equation. While $CI = 1$, the median effect equation will describe the isobologram of two drug, and thus the pure additive effect of the drugs.

Combination index is relevant to our discovery, as it reveal the linearity of the effect of enzymatic systems under proper transformation. Thus, we applied logit transformation for drug sensitivity in our later model, and gain huge effect to linearizing the dependent variable. Under this linearization, we can show that many of the effect of genes that convey the growth advantage of tumor cells can be regarded as additive.

CHAPTER 2

Tumor Biology

PROPHECY is a plateform for the prediction of drug response of cancer therapeutics. We need to understand tumor biology in terms of the cause of cancer, how cancer convey growth advantage and drug resistance, and the working principle of cancer therapeutics in order to generate a model that takes most of the important predictors into account. We pay special attention to the observables that are related to these functional advantages as they are the main factors that can be used to distinguish therapeutic outcome.

We seperate the predictors (or observables) of PROPHECY into two main categories, including intrinsic predictors that unabiguously point to a disease phenotype, and extrinsic predictors that pertains to the perturbations that are introduced by therapeutics. Intrinsic predictors serves two major purposes; they can point to the functional gain of the disease which will effect the response of disease to treatment, and they serve as signatures to tell different cancer types apart. We selected the oncogene profile as the intrisic predictors, and we will review the reason behind our selection. Notice that many other predictors such as gene expression levels and epigenetic status are often used as predictors in contrast to our selection. For extrinsic predictors, we used drug targets and their specificity as an indicator. In later chapters, we will show how the combination of intrinsic and extrinsic predictors can provide powerful insight that can accurately determine the outcome of cancer therapy.

Cancer is a genetic disease, as pointed out in Bert Vogelstein's seminal paper

[VK04]. In other words, the genetic signature of cancer will unambiguously define the characteristics of the cell and thus determine its response to external stimulus.

2.1 The evolutionary process of cancer

The pathogenesis of cancer is a evolutionary process of cells that take place in the microenvironment of a multicellular organism. Cells aquire random genetic alterations from various sources and produced a diverse set of phenotypes that eventually undergo natural selection. Those cells that acquired the hallmarks of cancer will eventually proliferate, invade organs and finally metastisize. Due to the random nature of genetic alteration, the cancer cells in a human body will present a spectrum of genetic alterations, which give cancer cells different capability to proliferate. This heterogeniety of cancer cells post a great challenge to the design of therapy, and another confounding factor for the objectives of the design. Much of the relapse of therapy also follows this evolutionary process, in which a small number of cancer cells that bear drug resistance survived the therapy and come back after they repopulated the organs. This evolutionary process is powerful and can be stopped by reducing the probability of having the phenotype to overcome therapy. Combination therapy has a higher selective pressure for cancer cells to overcome, so will lower the risk for relapse. A comprehensive review of the evolutionary process of cancer can be found at [SCF09].

2.2 Cancer genes and the growth advantage they convey

Cancer genes can be categorized into three types: the oncogene, the tumor suppressor genes, and stability genes. [VK04] Oncogenes and tumor suppressor genes when mutated can convey growth advantages including the activation of genes that drives the cell cycle, and inhibition of apoptotic signaling and increased

supply of nutrient through angiogenesis. [HW11] When the stability genes are mutated, they do not directly contribute to tumor proliferation. Instead, the stability genes disrupt the regular safety mechanism which protect the organism from gene alteration, thereby, increasing the risk of tumor formation.

The exact mechanism through which the cancer cells gain their growth advantage

2.3 Somatic mutations of cancer cells

Cancer cells are the direct descendants of the fertilized eggs of the patient and therefore carries a copy of its diploid genome. However, just like other lineages, cancer cell carries a set of differences from its progeniter cells which give cancer cell selective advantage. We call these differences somatic mutations in contrast with the germline mutations that are inherited from parents. Germline mutations on important genes often result in embryonic lethality and thus is not included in our discussion. [BGL11]

There are several common ways in which cell obtain genetic lesion. Cells acquire mutations during mitosis due to the built in error rate of the DNA polymerase, exposure to radiation such as UV light or X-ray or mutagen of both external and internal origins. Cancer cell also inherite the mutations coming from its progenitor cells.

2.4 Nonlinear effect of mutations

Traditional oncology work sepetate cancer mutations into two major categories: the driver mutations and passenger mutations. Driver mutations are those genes which mutated will convey growth advantage to the tumor cell, while passenger mutations are those mutations that does not directly contribute to growth but

carried on with other genes nontheless. This type of thinking was very popular in the literature, and associate an causal mutation with the resulting phenotype. However, it has very limited use as a slight change in the state of the network or the introduction of an external stimulus will alter the resulting phenotype easily. Thus, this type of classification is imprecise and misleading, and the function of each gene is actually dependent on the context of the network dynamics.

Function of genes are context dependent The function of each gene on the resulting phenotype usually dependant on the context of the network. One of the very first example found on this dependance was the affect of Jun-activated kinase (JNK) on apoptosis. [JAG05] It is found that the function of JNK can be apoptotic or anti-apoptotic depending on the signaling recieved from EGF and TNF, and the phosphorylation status of JNK along is not sufficient in deteriming its function. This means the resulting phenotype of perturbation on JNK will depend on perturbation on other parts of the network. Therefore, it is necessary to supply information on all of the perturbation in order to predict the final phenotype.

2.5 Epigenetic landscape

As we come to know that the complex network nature of biological system naturally exhibits, a simple mendelian picture that maps one gene to one phenotype is no longer suffice. With the aid of network theory and systems biology, we begin to appreciate the dynamics that governs the phenotype of the system. Waddington introduced the concept of "epigenetic landscape", that coincides with the dynamic theory of complex systems, and can unambiguously and quantitatively explain the observed phenomenon, which used to perplex reductionistic biologist. [Hua12]

The basics of epigenetic landscape The basic framework started with the notion that we can keep track of the state of the cell by keeping track of the quantity of all gene products. Then, the dynamics can be described by the change of gene product with respect to time. The change of state is governed by the wiring diagram, or network structure, which describes the basic biochemical relationship such as inhibition and activation between genes. A clear distinction must be made between the network structure and network dynamics, as the term is used interchangablly in some literature. However, the structure of the network usually remains static for an organism while the states of the network nodes may vary drastically. If we plot the potential energy of the genes in a multidimensional "state" space of the genes, the resulting manifold is what we called the epigenetic landscape.

The linkage between phenotypes and epigenetic landscape Each point in the epigenetic landscape is a manifastation of a particular phenotype. We would expect that there will be a continuous spectrum of cell types exist in an organism, but the reality is that there are only a finite number of cell types in the landscape. Why is the finite number of state the case? This can be explained well by the attractor states in the landscape. Due to the complex dynamic theory, the epigenetic landscape is full of attractor states that are stable basins that posses lower local energy and thus can attract near by states. The process is very similar to the situation when placing a ball on the edge of a sink, and eventually the ball can only be stable at the bottom of the sink.

External stimulous such as growth factor or drug will present as network perturbations which forces the state of the cell to move from one point to another in the landscape. If the perturbation is strong enough, the cell will be pushed away from its orginal state and finally rest in another stable state. [HEB05]

The static nature of the molecular network structure Just like the network structure, the epigenetic landscape remains mostly static. It is under genetic alterations that will result in the rewiring of the network. For instance, if an inhibitor of an enzyme lost its activity due to mutation in the functional moiety, the inhibitory linkage from the inhibitor to the enzyme will be deleted. Interestingly, even though the network structure changed, it appears that the entire epigenetic landscape will be changed due to the reorganization of the system. However, large scale simulation of a genetic regulatory network suggested otherwise. People found that the rewiring result from mutation will actually contribute to the local change of a small area in the epigenetic landscape. [HEK09] This can be seen from the fact that many normal cells in the human body also acquired somatic mutation such as cancer, but remains as healthy phenotype nontheless.

It is hypothesized that cancer emerged as cells reached stable cancerous states called cancer attractors. These stable states are usually unreachable by normal developmental path. However, the rewiring of network caused by mutation will change the barrier between attractors and make these cancer attractors more accessible to external stimuli, leading to tumor genesis. The construct of cancer attractors can handsomely explain the reason of finite phenotype exist in the system, despite the fact that there are astronomical amount of possible combinations of genetic lesions. There are only finite amount of cancer attractors produced by the complex network of the cell, and thus finite number of phenotype.

The theory of epigenetic landscape is gaining momentum and started to be accepted by mainstream biologist, as recently demonstrated by the stem cell research. [FK12]

2.6 External and internal perturbation

Under the framework of epgenetic ladscape, we can then define the influence of external and internal perturbation clearly. External perturbations are forces that directly cause the system to move from one point to another in the state space by doing work. So, we can categorize the temporal inhibition of drug molecule as an external perturbation, as the action is reversible and the circuitry of the network remains untarnished after the administration of the inhibition. We can also categorize the genetic lesions, caused by drug or by random process as internal perturbation as they change the circuitry directly. Notice, however, the genetic lesion does not directly cause the network to move from one state by perturbing the state, but by changing the epigenetic landscape of the system and thereby alter the accessibility of certain network states.

2.6.1 Trade-off of information, limitations

In systems biology approach, the parameter space that is required to described the dynamics is huge, and the availability of the parameters is a huge challenge to create the model. Even if the parameters were estimated to be the optimal set that fits the data, it still beg the question of the explanatory power of the model on whether the data is being simulated or memorized in the model. [Gun10] Under the current stage of development of the omics technology, it is still out of hand to try to simulate the dynamics of the entire cellular network of human. In essense, we need to consider the available data to us and select the ones that can be obtained economically.

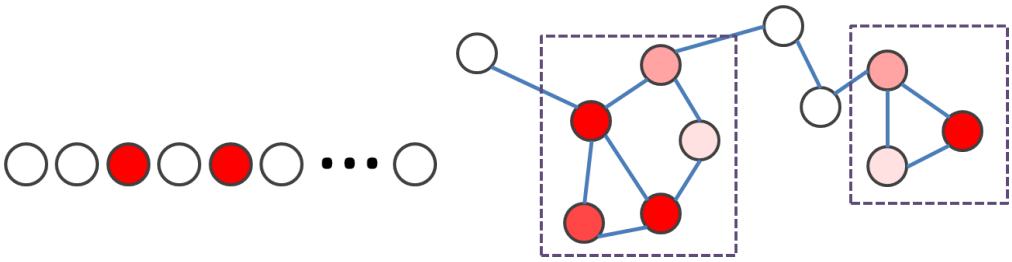
Currently, the protein-protein interaction network is available, containing around 22,000 nodes and 1 million linkages. However, detailed parameters is still lacking, and false positives still prevalant in the network, prohibiting a systems biology approach to model the dynamics of the whole network. We therefore choose to

work with static systems which only concern the end state of the network. End point measurement is a convenient choice, as the most common phenotypic experiment data is the dose-response curve, the end point measurement of drug screening experiment. The dose response curve present a simplified expeirmental model system, which exclude the microenvironment in the organism, but preserve the context of network dynamics and the external stimuli from drugs. The dose response has been historically used as an effective screening experimental model. Recent data shows that most of the novel targets were discovered through phenotypic type of screening, as opposed to the much anticipated molecular target screening approach. [ZTM13]

2.6.2 Network diffusion to mimic the convergence of cancer attractor

The goal for our model fitting is to have all perturbations available to our knowledge mapped to drug sensitivity.

Supposed that similar disease phenotype will result from different set of mutations, there must be some similar changes of functions in the network that causes the same phenotype. We thus hypothesize that it is the disruption of the states in the functional module in the network that causes the resulting phenotype. Now the question is how do you measure the disruption of states in a functional module? It will becomes clearer if you look at the information spread in the network. Molecular machinary in a cell rarely carry out a function by itself. Rather, the function is a manifestation of the interaction between the molecules, starting from the molecules that were disturbed and then propergate throughout the entire network. If the disruption of state is measured as probability density that goes from one node to the other, you can quantitatively evaluate the probability that certain nodes got disrupted by the original perturbation, either external or internal. In later chapter of graph theory, we will go into the diffusion of probability in detail. The conclusion from the diffusion study is that the information spread can



(a) (b)

Figure 2.1: Difference between gene centric and network centric predictors. (a) mutation-based predictors, the features is usually represented by a boolean vector where the mutated genes are marked as 1 and others 0. This type of representation implies the mendilian view of genetics where genes are discretized blocks and each gene serve a specific function. (b)network-diffusion-based predictors. The result of network diffusion will correlate with the local network structure, and the part high lighted by dash boxes shows the distribution of probability density hovers around the local clusters, demonstrating that the diffusion scheme can find out local functional groups in the network.

mimic the diruption of modules in the network, as the propagation of probability in the network has the property that the local probability will be higher in a local module. [VMR10, KBH08] This property of localization can be used to explain why different sets of mutations could have similar phenotypes because they share similar pattern of information propagation in the network.

2.6.3 Navigating the cancer attractors

Now that we translate the signature of perturbation from a limited set of nodes to the information spread in the entire network, we can then repose the problem of model finding cancer attractors.

Now we turn our attention to using these network signatures to make predictions about the drug sensitivity. What does the model tell us exactly? We

mapped the disease genes onto a protein-protein network, and then through diffusion we find the functional modules that causes the disease and gave cancer cell the growth advantage. The diffusion process is a method to help us approximate the influence of rewiring on functional modules which are better signatures to predict the basins of the epigenetic landscape and the phenotype of the cell.

We also mapped the drug targets onto the protein-protein network and finds the diffusion pattern in the network. In this case, however, the diffusion shows the perturbation of the state of each module. So, essentially, the model we created is a road map in which we can find the sensitivity of the cell line given the location of the basin and direction of perturbation. Finally, we can use this model and apply optimization technique, and we can find the best set of perturbations that can lead the cells to the desired phenotype.

CHAPTER 3

Framework of PROPHECY

3.1 Systems overview

The current PROPHECY platform is a predictive model of combinatorial drug sensitivity that is capable of making prediction based on the list of drug targets and disease genes. We reason that the disease phenotype is a result of series of interaction of proteins that propagate from the initial perturbation of several disease genes to the entire network. Similar process also applied to the response of drug perturbation. Unlike most of the existing models which only take the initial perturbations into account, we want to include the effect of all other genes into the picture.

Therefore, in order to measure quantitatively the propagation of information onto every protein in the interactome, we use random walker scheme or equivalently network diffusion to determine a probability signature that measures quantitatively the proximity of every proteins and the perturbed nodes in the network. The signature is central to the information gain and the quality of the PROPHECY model because the signature encodes information of the structure of the network and functional modules. We then feed a large data set of information pattern and example drug sensitivity to a machine learning algorithm, which in essence will determine how each piece of perturbation of one protein, and the interaction of these perturbation will contribute to the drug sensitivity. Equipped with the learned knowledge of how the network signature of perturbation will af-

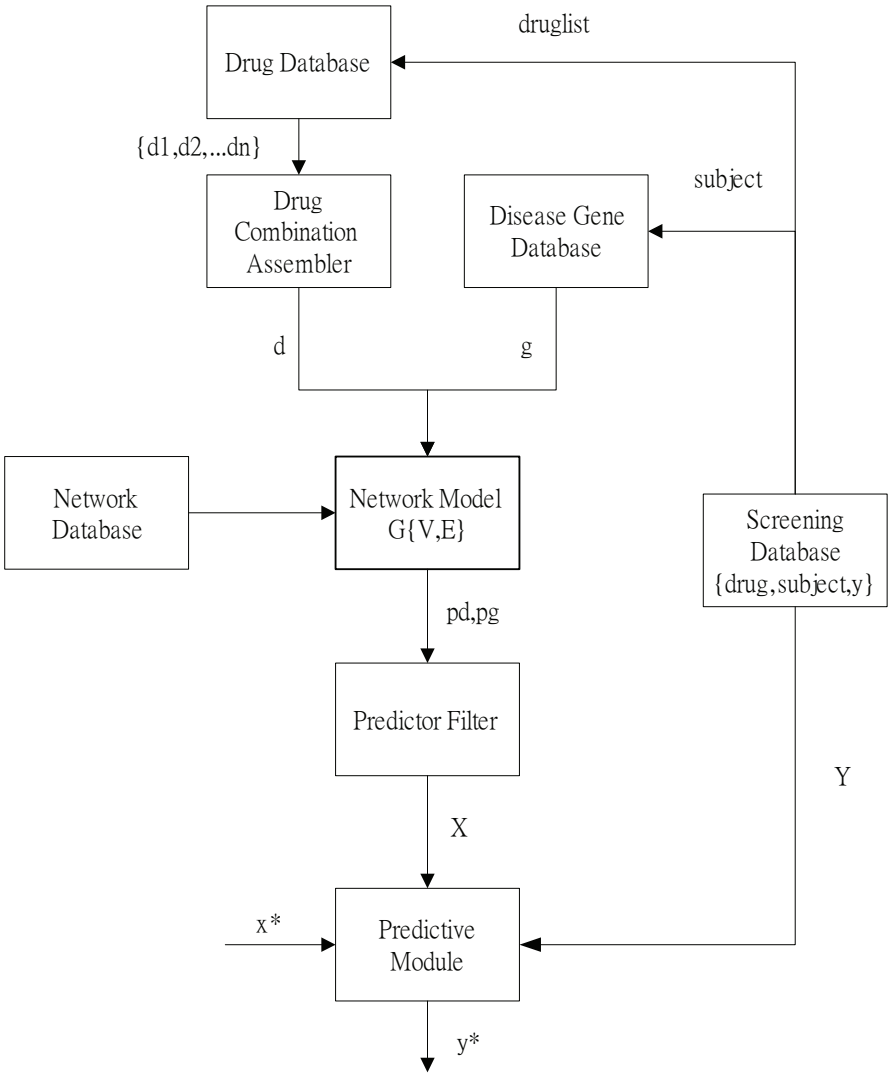


Figure 3.1: Predictive Module of PROPHECY.

fect the sensitivity, the model was able to make accurate predictions. The model can even predict unseen drugs or combinations, due to the fact that the model learned interaction of perturbation signature instead of memorizing the effect of each drug.

Notice that the PROPHECY is a general conceptual framework that is not limited to prediction of drug sensitivity, and can be easily generalized to predict other phenotype as well, such as the expression level of certain gene, or the categorical outcome of certain genetic disease.

The PROPHECY method consist of 4 integral components, the training databases, the network model, predictor filter, and the predictive module. All of the components are essential to the success of the prediction. We reviewed in detail the data source of each component, the mathematical structures that defined the components, and how these components are linked in this chapter.

3.2 Screening database

The drug screening database contains the experience universe of the predictive machine. If we were to compare the predictive machine to a medical doctor, the screening database will be equivalant to the doctors first exposure to an unknown disease. For the predictive machine to learn how to treat the disease, the machine has to be exposed to enough training cases in order for the machine to find patterns inside. In the machine learning literature, a rule of thumb is that for every predictor, there has to be at least 5 data points in order to have a reasonable fit. The most publicly available drug screening dataset to our knowledge is COSMIC, so we used COSMIC to train PROPHECY. [SHS10]

COSMIC contains the drug screening data of 150 cancer drugs across 900 cell lines, adding up to around 40,000 data points. The sensitivity data is supplied in multiple formats, including the fitted sensitivity, and IC_x data, and parameters

of the fitted curve. However, the raw fluorescent data from the screening is not available.

The most common way of quantitatively expressing the effect of the treatment is through the dose response curve, where the x axis is the log of the concentration of the single drug being used in the study and y axis the viability of the cells or the optical density in the raw read out. In COSMIC, every compound was screened at 9 concentrations with a 2-fold serial dilution, which spans a 256-fold range of the drug. The cells were incubated for 72 hours and treated with fluorescence-based assay for the measurement of viability. The same procedure was slightly modified and then adapted to our subsequent study on combination drugs.

Bayesian sigmoid model for the dose-response curve Notice that the raw data is rarely directly used for the interpretation of sensitivity. Usually the raw read out is contaminated with noise from the optical instrumentation and also from experiment, so it is desirable to do a curve fitting in order to filter out the effect of the noise and secondly incorporate our prior knowledge of the model. In COSMIC and our study, we used Bayesian sigmoid model to fit the dose response curve, in which biochemically the cells naturally follows when subjected to stimuli from external chemicals [CT84]. A generalized sigmoidal curve is fitted as follows.

The mean intensity X_{IC} is

$$E(x_{IC}) = I_{min} + \frac{I_{max} - I_{min}}{(1 + e^{\beta(IC-\alpha)})^f} \tag{3.1}$$

where I_{max} and I_{min} are the mean intensities of the positive and negative controls, α and β are the scale and gradient response, f is a shape factor, and IC is the log concentration of the drug. We assume that x_{IC} follows a gamma distribution, as the optical density in the reading must be non-negative. So, it follows that

$$p(x_{IC}) = \frac{B^{E(x_{IC}/B)}}{\Gamma(E(x_{IC}/B))} x_{IC}^{E(x_{IC})/B-1} e^{-Bx_{IC}} \tag{3.2}$$

Curve-fitting algorithm The major point of the curve fitting algorithm is to determine the coefficients presented in the sigmoidal curve, and to estimate the variance of the curve. According to Bayesian parameter inference, the posterior probability of the parameters given the data points is

$$p(w|\mathbf{X}, \mathbf{y}, \theta) = \frac{p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w}|\theta)}{p(\mathbf{y}|\mathbf{X}, \theta)} \tag{3.3}$$

where $\mathbf{w}$ is the vector of coefficients in the model, $\mathbf{X}$ the input design matrix, $\mathbf{y}$ the output vector, and θ the hyper parameters used in the prior distribution of the parameters and noise model. Since the hyper parameters also has its own distribution. We also have to infer this by marginalize out the information in $\mathbf{w}$. We have

$$p(\theta|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\theta, \mathbf{X})p(\theta)}{p(\mathbf{y}|\mathbf{X})}, \tag{3.4}$$

where we define the likelihood function here as

$$L(\theta) = p(\mathbf{y}|\theta, \mathbf{X}) = \int p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w}|\theta)d\mathbf{w} \tag{3.5}$$

When the distribution has a single mode, maximum likelihood solution is usually a good approximation to the posterior mean model. However, when the distribution has multiple modes, we will have to calculate the posterior mean to be representative of all possibility. We can use a stochastic optimization solver to calculate the maximum likelyhood solution of the hyper parameter.

We can calculate the likelihood function as

$$p(y|w, x) = \prod_{i=1}^{9} \Gamma(\frac{1}{1 + e^{\beta(x_i - \alpha)}}, B) \tag{3.6}$$

where $\Gamma(.,.)$ represent gamma distribution.

Maximum likelihood solution of hyper parameters While a cell line is sensitive to one drug, the viability of the cell line will drop sharply at low dosage, and appears to stay constant for a drug resistant cell. So, an intuitive way to

define sensitivity is to use one minus the area under curve (AUC) of the dose response curve. The AUC is defined as the integral of the sigmoid curve and normalized by the range of concentration applied, so the maximum possible value of AUC is one and the minimum is zero.

The way the data is processed in COSMIC allows the sensitivity to be expressed in a single numeric value, and this value can be compared across cell lines and drugs, making it suitable for machine learning purpose.

3.3 Drug Database

In the framework of PROPHECY, the dose response of a cell is interpreted as the response of a complex network to external stimuli. So, unlike traditional analysis where only the main drug target is considered, we include all known interactions between a drug and gene/proteins in our subsequent analysis. Actually, there are many studies that points to the fact that the off target effect or the non-specificity of the drugs ultimately determines the efficacy of the drug as mentioned in chapter 1. [ACP05, Cse04, CAP05, CKK13]

The data mining of all possible linkage between gene and proteins to drugs is a major undertaking itself. Fortunately, there are a number of databases that can be used for this purpose. We choose to use STITCH, a comprehensive database that documented over 300,000 small molecules and 2.6 million proteins from across 1133 organisms. [KSP14, KSF12, KSF10, KMC08] In order to maximize the information of known chemical-protein binding, STITCH integrates the following data source for chemical protein interactions. Experimental evidence of direct chemical-protein binding is obtained from the PDSP K_i Database and the protein data bank (PDB). Interactions between metabolites and proteins are selected from pathway databases including KEGG, Reactome, and NCI-Nature Pathway Interaction Database, and drug-target interactions from DrugBank and MATA-

DOR. Text mining is also performed on MEDLINE and OMIM to yield additional evidence by co-occurrence or natural language processing. The confidence score is then assigned to each interaction based on the level of significance and certainty of the interaction.

We extracted all drugs that were included in our screening database, and the corresponding drug targets that can be mapped to the human protein-protein network. The set of drugs Ω is represented as

$$\Omega = \{\mathbf{d}_i \in \mathbb{R}^{n\times 1} | \, i = 1, \cdots, N_d\} \tag{3.7}$$

, where n is the number of proteins in the protein-protein network, $\mathbf{d}_i$ is a vector with nonzero entries being the confidence score associated with the drug targets of drug i. N_d is the number of drugs in the library of chemicals.

Notice that after the data processing of STITCH, we still cannot have exact information of all protein-chemical interaction. There are a number of error sources, including false positives, and false negatives in the databases, the inconsistency in the normalization scheme used in integrating multiple databases. Also, the databases derived from in vitro protein binding experiment such as protein array experiments might over estimate some interaction that might not have existed due to compartmentalization in cells. We can see in the later section that PROPHECY can actually reduce the presence of these noise sources indirectly by a couple of built in mechanism. First, the PageRank used to calculate the diffusion of information, is insensitive in terms of the starting probability and also the false positives linkages in the network. Secondly, the learning procedure in the machine learning, in analogous to regularization in regression, will reduce the effect associated with unimportant predictors.

3.4 Disease Gene Database

For the disease gene database, our goal is to source a comprehensive collection of causal mutations in the genome of the cell lines involved. Gene expression level is not considered due to the following reasons. First, the network model we used is partially derived from gene coexpression studies, so the gene expression level information can be argued to be already contained in the network linkage. Secondly, gene expression level is a transient snap shot of the current state of the cell, so even the same cell at different cell cycle might express different amount of genes and also might express differently when subjected to the same chemical stimuli. However, for genetic variant, the genome is well defined for every cell line, and remains stable during culturing or even treatment. The stability of the genome is the main reason that cell lines became standard choice as the in vitro model for disease. [VK04, BCS12]

If we consider the problem of drug response in terms of dynamical system and epigenetic landscape, the genomic profile of the cancer cells define the epigenetic landscape, and the subsequent terrain neighboring the state where the therapy is intended to drive it. [Hua12, BE10, CSE12, HEK09] So, the landscape itself should contain most of the information that is required to determine the driving force that needs to direct the cell to certain state. In this view, the network provides the prototype of the shape of the landscape, and the mutation profile gives the modification on the landscape.

Currently, the most comprehensive and consistent public database for genetic profile of cell lines is the Cancer Cell Line Encyclopedia (CCLE), which encompasses 947 human cancer cell lines and span 36 tumor lineages. [BCS12] We included the mutation profile of the cell lines which were determined by targeted massively parallel sequencing on ¿1,600 genes, 392 genes that effect 33 known cancer genes were collected by mass spectrometric genotyping, and DNA copy

number was measured using high-density single nucleotide polymorphism arrays. We then map all of the causal genes on to the protein-protein network, excluding the mutations of introns that do not effect the transcription or the gene product.

We represent diseases models as the set of cell lines, Γ, as

$$\Gamma = \{\mathbf{g}_i \in \mathbb{R}^{n\times 1} | \, i = 1, \cdots, N_c\}, \tag{3.8}$$

where $\mathbf{g}_i$ denotes a vector where the non zero entries are the confidence score of disease gene and zero elsewhere. The confidence score reflects the type of mutation the gene causes. Currently, we treated the disease gene vector as a Boolean vector so each entry is a logical representing whether the gene is mutated or not.

3.5 Network Model

Our intuition is that the interactions between nodes collectively will dictate most of the response of a cell to a treatment. Not only do drug targets need to be considered, but also the location of disease nodes. What we need then is a language to describe the level of impact when some nodes are affected, either by target nodes or disease nodes. While the states of cells do vary between cells or individuals, the network structure is almost invariant between cells or individuals, so we chose network as the universal language for the description of the response model.

The p-p network can be represented by a graph $G(V, E)$, where V denotes the set of nodes and E denotes the set of edges. There are n nodes and k edges in G. We can represent a undirected network with an adjacency matrix, $\mathbf{A}$, in which

$$\mathbf{A}_{ij} = \begin{cases} a_{ij} & \text{if } \{i, j\} \in E, \\ 0 & \text{else.} \end{cases}, \; a_{ij} \in (1, 0), \tag{3.9}$$

where a_{ij} represent a confidence score which link to evidence on this interaction.

3.6 Predictor Filter

The predictor filter is the component where In the predictor filter, the predictors of drug targets are assembled, and both the disease genes and drug targets are transformed to encode the information of the network.

3.7 Transformation of dependent variable

The range of sensitivity lies between zero and one. When doing any form of regression analysis, it is more nature to assume the range is unbound, or between $(-\infty, \infty)$. In order to change the range, it is customary in machine learning to apply the logit transformation to transform sensitivity s to another measurement s'. The logit transformation is written as

$$s' = logit(s) = \frac{1}{(1 + e^{-\alpha s})^f} \tag{3.10}$$

Notice the difference between the logit transformation here and the sigmoidal transformation we used to fit the dose response curve. We do not need to include a β factor in our calculation here. However, a shape factor, f, is still included, as this shape factor will influence the distribution of the sensitivity and potentially the effect the machine catches.

3.8 Decomposition in machine learning

The cells' response to external stimuli is determined largely by the topology of the network, as demonstrated in the network pharmacology studies [Hop08]. Our current invention is intended to mine the global structural information encoded by the biological network. Specifically, We want to see whether knowing a given set of disease genes and their position on a biological network will lead to a predictable

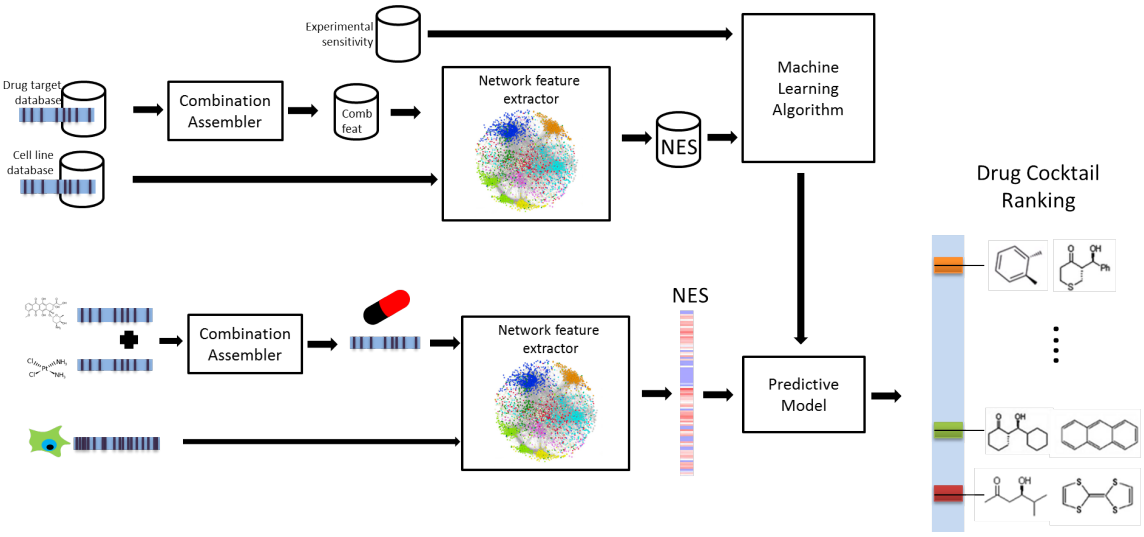


Figure 3.2: The block diagram of PROPHECY.

result of drug screening experiments.

We introduce Predictive Optimization of Pharmaceutical Efficacy, PROPHECY, to link the network properties of drugs and the genetic profile of a subject to the efficacy of a drug combination. The overall structure of PROPHECY is shown in figure 3.2. A high level overview will be provided in this section, while details will be given in later sections.

The training data for PROPHECY consists of four components, the screening database, drug database, disease gene database, and network database. The screening database contains the actual experimental data, which includes the experimental condition used, and results. There is no limitation for the experimental results chosen as long as it is related to drug efficacy. For example, area under curve (AUC) of drug sensitivity assay can be used as an indicator for drug efficacy. The efficacy of all experiments should all be represented in the same format in order for the fitting to work. The experimental data will eventually guide the program to find the interactions between network nodes. The purpose of the drug database is to link a drug to its targets. Similarly, the disease gene database relates a subject, which can be a cell line or the primary cells of a patient, to

its genetic profile. We then select the mutations that cause physical changes to the network and map the mutations of the subject to the protein-protein network used. The network databases we included are protein-protein interaction networks (PPI) that link molecular interactions in a undirected graph format. Note that the framework proposed in PROPHECY can be used not only in PPI, but also other biological network, including genetic network and signal transduction network. We chose PPI for the edge of the network is a representation of physical interaction, so is less prone to unexpected interactions stem from the interpretations of different dataset as occur in genetic network.

All databases are incorporated into a network model, and the network model will produce a preliminary input training set for the predictive module. The inputs encodes network as well as bioactivity information. At this point, the data will be too large for realistic prediction, so the input is passed through a predictor filter to filter out low information content data. Finally, the filtered input is send into predictive module to generate model for efficacy prediction.

CHAPTER 4

Graphs and Network diffusion

Graph theory is a branch of mathematics that deals with graph, which is consist of nodes and edges. Many modern computer science or engineering problems can be abstracted and recast as a graph problem. We applied the same type of abstraction in our method, in which the dynamics of interactions between elements in the molecular circuitry is simplified and reformulated into information spread in the network. As such, we are freed from the necessity to collect a large set of parameters used in the dynamical model and can still gain insight into the network interaction quantitatively by using the structure of the network.

In this chapter, we first reviewed the idea of network centrality, which can be seen as the first step towards understanding of how structural information of a network can be summarized in a nodal numerical value. Some pioneering work in terms of the relationship between network centrality and its biological relevance also inspired some of the works in this dissertation. We then introduced diffusion in the network, which we applied in PROPHECY as a measurement for modular excitation and information spread. Finally, we discussed spectral clustering as a powerful way for classification of cellular populations.

4.1 Network metrics

Network metrics are nodal scores that reflect characteristics of a node in relation to the geometry of the network. For instance, degree centrality is the nodal score

that shows how connected a particular node is. It has been proven that network metrics encodes rich information related to the response of the cell to external stimuli. Previous study showed that degree centrality, betweenness centrality, bridging centrality, can be related to how good a node can be used as a drug target.

Our intuition is that the interactions between nodes collectively will dictate most of the response of a cell to a treatment. Not only does drug targets needs to be considered, but also the location of disease nodes. What we need then is a language to describe the level of impact when some nodes are affected, either by target nodes or disease nodes. While the states of cells do vary between cells or individuals, the network structure is almost invariant between cells or individuals, so we chose network as the universal language for the description of the response model. The p-p network can be represented by a graph $G(V, E)$, where V denotes the set of nodes and E denotes the set of edges. There are n nodes and k edges in G. We can represent a undirected network with an adjacency matrix, $\mathbf{A}$, in which

$$\mathbf{A}_{ij} = \begin{cases} a_{ij} & \text{if } \{i,j\} \in E, \\ 0 & \text{else.} \end{cases}, \; a_{ij} \in (1,0), \tag{4.1}$$

where a_{ij} represent a confidence score which link to evidence on this interaction.

In the following paragraphs, we review some of the most highly studied centralities, and demonstrate that there is a strong relationship between the topological features of a network with functionality of the nodes.

4.1.1 Degree centrality

The degree of a node, v, is simply the number of edges which are connected to v. The degree for a node,v, denoted $degv$, satisfies

$$\sum_{i=1}^{n} deg(v_i) = 2E \tag{4.2}$$

Degree is the very first network centrality (or network metric) that had been extensively studied. At the beginning, the research direction was focused on using degree to find the organizational principle of biological network. [JTA00] Through the distribution of degree centralities, it is found that it follows power law distribution, meaning that the higher the degree a node has, the less probable it will occor in the network. It is also shown that high degree nodes are often essential genes, and thus the inhibitions of these nodes are highly toxic. [JMB01]

4.1.2 Betweenness centrality

Betweenness centrality measures the number of non-redundant pathways that pass through a node. So, a node with a high betweeness centrality can be regarded as the "bottleneck" in the network. The betweenness centrality of a node v is given by the following expression:

$$\Phi(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{4.3}$$

where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v. Joy et al. found that yeast network contains a large number of nodes with high betweenness centrality, and the number is higher than that could be explained by a scale free network. [JBI05] The unusually high betweenness can be explained by the high modularity in the yeast network, in which the function of high betweenness nodes is to pass information between modules. Yu et al. studied the dynamics of gene expression in regulatory network, and found that high betweenness centrality nodes are more likely to be essential in the network than high degree nodes. [YKS07]

4.1.3 Bridging centrality

Bridging centrality measures of a node's ability to connect modular subregions in the graph and its betweenness within these modules. The calculation of bridging

centrality is the product of the bridging coefficient and the betweenness centrality. A node with high bridging centrality would locate between more modules in the network and holds the global position of connecting nodes. The bridging coefficient of a node v can be expressed as

$$\Psi(v) = \frac{1}{d(v)} \sum_{i \in N(v), d(i)>1} \frac{\delta(i)}{d(i)-1}, \tag{4.4}$$

where $d(\cdot)$ is the degree function, $N(\cdot)$ is a function that returns the set of neighbors of a node, and $\delta(i)$ is the number of edges that leaves the direct neighbor node i. Then, the bridging centrality is defined as

$$C_{Br}(v) = R_{Psi}(v) \cdot R_{Phi}(v), \tag{4.5}$$

where $R_{Psi}(v)$ and $R_{Phi}(v)$ is the ranking of node v in terms of its betweenness centrality and bridging coefficient in the network respectively.

Hwang et al. showed that nodes with high bridging centrality could be nice candidate of drug targets, as these nodes has low lethality and can modulate the network quite effectively. [HZR08] The main contribution of this work is to show that the application of network topology can be applied to reveal the functionality of nodes and how these properties of the nodes can be an indicator for drug target selection.

4.2 Diffusion of information in the network

The centralities of a network were interesting metrics that linked the topological features to nodal function. As the literature shows that nodes have unusual centralities usually has unique function, our first thought was to use these centralities as predictors for a sensitivity model. However, the centralities itself does not provide extra information if we were to treat them as predictors in a predictive model because a two value input of zero and a centrality score is equivalent to zero and 1 in a regression model. We thus turn our attention into searching for a network

metric which depends on your initial nodal selection, and also encode modular information. In complex disease, the diseased phenotype is seldom the product of the loss of function in a single gene. Even for Mendelian disease, which has only one causal mutation, the phenotype is the product of the malfunction of the causal gene propagated through the network. As mentioned in the tumor biology chapter, the information propagated from the causal genes to the local modules and then to the entire network can be a powerful predictor to the phenotype. Here, we discribed two methods that quantitatively evaluate the propagation of information, namely, the diffusion kernel and PageRank.

4.2.1 Diffusion kernel

Diffusion kernel is a form of lazy random walk and was used to prioritize unkown disease genes. [KBH08] We can define the diffusion kernel, $\mathbf{K}$ of a graph G as

$$\mathbf{K} = e^{-\beta \mathbf{L}} \tag{4.6}$$

where β can be thought of as a rate constant that controls the magnitude of diffusion, and the matrix L is the laplacian of G, which is defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{4.7}$$

where $\mathbf{D}$ is a diagonal matrix with the diagonal terms holding the degrees of the nodes. By using K, the nodes j can be ranked by the score defined by

$$score(j) = \sum_{i \in \text{genetic lesion}} \mathbf{K}_{ij} \tag{4.8}$$

4.2.2 PageRank

We represent diseases models as the set of cell lines, Γ, as $\Gamma = \{\mathbf{g}_i \in \mathbb{R}^{n \times 1} | \, i = 1, \cdots, N_c\}$, where $\mathbf{g}_i$ denotes a vector where the non zero entries are the confidence score of disease gene and zero elsewhere. The confidence score reflects the type of

mutation the gene causes. N_c is the number of cell lines in the library of cell lines. Similarly, the set of drugs Ω is represented as $\Omega = \{\mathbf{d}_i \in \mathbb{R}^{n\times 1} | \, i = 1, \cdots, N_d\}$, where $\mathbf{d}_i$ is a vector with nonzero entries being the confidence score associated with the drug targets of drug i. N_d is the number of drugs in the library of chemicals.

The first thing we want to do with the network is to find the propagation of information from drug targets and disease genes. Due to lack of detailed network directionality and reaction constants, we do not have the complete picture on how the actual dynamics of the network state will shift due to intervention of drugs and disease genes. What we do have is the linkage between nodes, and the fact that the state of the neighbors of the drugged (or mutated) nodes will be distorted. Therefore, we use random walker starting from the drugged (or mutated nodes) to model the propagation of information. The assumption is that we can see the most affected nodes by looking at the steady state distribution after the random walker was presented in the network for a long time. However, one thing worth notice is that the propagation of random walker will always reach a single steady state regardless of the starting nodes, as predicted by the Perron Frobenius Theorem. The theorem says the stationary distribution of the random walker will always converge to the eigenvector which has eigenvalue equals to 1. Therefore, in order to make the starting nodes matter, we will use random walk with restart. By putting the random walker back to the original nodes, the converged distribution will be analogous to the so called personalized PageRank [PBM99].

To model the random walk, we first define the state tansition matrix $\mathbf{W}$ as

$$\mathbf{W} \equiv \mathbf{D}^{-1}\mathbf{A}, \tag{4.9}$$

where $\mathbf{D}$ is the degree matrix.

To evaluate the effect of mutated nodes on the network, we do a random walk

where the initial distribution is

$$\mathbf{p}_{g,i}^0 = \frac{\mathbf{g}_i}{|\mathbf{g}_i|_1}, \tag{4.10}$$

and the same normalization with taxicab norm can be applied to drugged nodes, $\mathbf{d}_i$. For any initial distribution, $\mathbf{p}^0$, we can model the random walk process as

$$\mathbf{p}^{t+1} = (1-r)W\mathbf{p}^t + r\mathbf{p}^0, \tag{4.11}$$

where r is the teleportation constant. The steady state probability $\mathbf{p}^\infty$ is an evaluation of how one subset of nodes affect the nodes in the whole network. Note that the steady state probability, $\mathbf{p}^\infty$, is the solution of

$$\mathbf{p}^\infty = (1-r)W\mathbf{p}^\infty + r\mathbf{p}^0. \tag{4.12}$$

So, the solution of the PageRank for a given initial distribution is

$$p^\infty = r(\mathbf{I} - (1-r)\mathbf{W})^{-1}p^0. \tag{4.13}$$

We can use this to calculate the information propagation for a given initial probability distribution such as $\mathbf{p}_{g,i}^0$ and $\mathbf{p}_{d,i}^0$. In the later fitting stage, we will not use the initial distribution, but only the steady state distribution. Thus, we will drop the superscript ∞ and use $\mathbf{p}_{g,i}$ and $\mathbf{p}_{d,i}$ to denote the steady state distribution.

Transformation of graph The original network model has to be transformed to take into account the dangling nodes, which are the nodes that has only one incoming edge. Otherwise, the dangling nodes will absorb most of the random walk probability and the probability at these nodes be over amplified after infinite iterations. The way for transformation is to add outgoing edges from the dangling nodes to all other nodes.

Combination drugs One of the main advantage of expressing a drug as a vector $\mathbf{d}$ is that a drug combination can be expressed as the exact same format.

In fact, the program has no knowledge whether a given drug is a single drug or a combination; it just takes it as a set of drug targets. The only problem now is the case when there are overlaps between drugs, which are common between a family of drugs or non-specific drugs such as cytotoxic drugs. We combine a set of m drugs, $\{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_m\}$, by the following.

$$\mathbf{d} = 1 - \prod_i (1 - \mathbf{d}_i). \tag{4.14}$$

Now, we can drop the dependence of subscript, i, on single drugs. From now on, all drugs including drug combinations will be denoted as $\mathbf{d}_i$, where i denotes the internal bookkeeping of the drug/drug combination, but not the indexing of single drugs.

4.3 Spectral clustering of data points in different sets of feature space

Spectral clustering is a clustering method to classify populations distributed in a multidimensional feature space, and the method is particularly useful when the distribution of the population is unknown or is highly complex. Traditional clustering methods such as k-mean clustering use elliptical metrics to group the population of data points, so they do not work well when the group of data is non-convex. For example, if the population forms donut shapes in the feature space, traditional methods will be ineffective in separating the population. On the other hand, spectral clustering does not make assumption on the distribution of population, and in this sense the method is closely related to multidimensional scaling. Since we want to compare the distribution of cell lines residing in different feature space, and we have no prior knowledge on how the distribution might look like, we chose spectral clustering as the default method. Usually, clustering is introduced under unsupervised learning which is related to machine learning algorithm. Since spectral clustering is closely related to graph partition problem,

we insert this topic in the graph theory chapter.

4.4 Spectral clustering

The spectral clustering of N data points, $X = \{x_1, ..., x_N\}$, begins with a N by N similarity matrix S, in which S_{ij} is the pairwise similarity between points i and j. The similarity can be measured by various metric. One popular choice is the radial-kernel gram matrix, in which

$$S_{ij} = \exp(-\frac{|x_i - x_j|_2^2}{c}), \tag{4.15}$$

where c is a scale parameter that measures the length scale of the problem. With the similarity matrix, we can then construct an N by N adjacency matrix W of a similarity graph $G =< V, E >$. W is constructed so that we set up a threshold η, within which $W_{ij} = S_{ij}$, and zero else where. Now the problem for clustering is recast into a graph-partitioning problem, as can be seen that neighbors will form a local cluster or modules in the graph.

Lastly, we calculate the graph Laplacian L as

$$L = G - W, \tag{4.16}$$

where G is a diagonal matrix of nodal degree with $G_{ii} = \sum_{j \in N(i)} W_{ij}$. Then spectral clustering is the clustering of a matrix $Z^{N \times m}$, in which the columns corresponds to the eigenvectors which has the m smallest eigenvalues of L. For a detailed explanation why spectral clustering works, readers are encouraged to read [HTF09].

In spectral clustering, the free parameters $cand\eta$ need to be determined, as well as the number of clusters. To avoid the bias in selecting the number of clusters, we use hierarchical clustering to look at the distribution at all possible number of clusters.

4.4.1 Comparing the similarity of two hierarchical clusters

We introduced spectral clustering as a way to group cell lines without knowing how their distributions look like in a feature space. Cell lines which have similar therapeutic profiles will certainly cluster together if measured by a feature space of drug sensitivities. What if the drug sensitivity is not available? So, our goal is to find another feature space, which can be the cell lines feature of gene expression level or the personalized PageRank of gene mutations, so that the clustering of groups can be similar to the clusters that generated by drug sensitivity.

So in here, we introduced a method that can compare the similarity of clusters in two trees which was introduced by Fowlkes et al. [FM83] to compare the similarity between the hierarchical clusters generated by using different metrics. Let us assume that we have two hierarchical clusters of the same number of objects, n, which we label A_1 and A_2. We can then cut each tree at a particular cluster strength to produce $k = 2, ..., n-1$ clusters for each tree. Since we have no prior knowledge how the clusters in the two tree should map to each other, we may label the clusters of A_1 and A_2 arbitrarily from one to k. From this random labeling, we can calculate the similarity matrix

$$M = [m_{ij}], (i = 1, ..., k; j = 1, ..., k), \tag{4.17}$$

where m_{ij} is the number of common objects between the ith cluster of A_1 and the jth cluster of A_2. Note that most of the similarity measure developed exploit the concept of similarity matrix. [HA85, Bak74, Ran71] We followed the measure score proposed by Fowlkes et al. and define our measurement of similarity as

$$B_k = T_k / \sqrt{P_k Q_k} \tag{4.18}$$

where

$$T_k = \sum_{i=1}^{k} \sum_{j=1}^{k} m_{ij}^2 - n, \tag{4.19}$$

$$m_{i.} = \sum_{j=1}^{k} m_{ij}, \tag{4.20}$$

$$m_{.j} = \sum_{i=1}^{k} m_{ij}, \tag{4.21}$$

$$m_{..} = n = \sum_{i=1}^{k} \sum_{j=1}^{k} m_{ij}, \tag{4.22}$$

$$P_k = \sum_{i=1}^{k} m_{i.}^2 - n, \tag{4.23}$$

$$Q_k = \sum_{j=1}^{k} m_{.j}^2 - n. \tag{4.24}$$

For every value of k, B_k is calculated, and the similarity of the two trees can be given by plotting Bk versus k. When all of the clusters in each trees correspond completely, there will be exactly k nonempty elements in M, so $B_k = 1$. When all pair of objects appear in the same cluster in A_1 is assigned to different clusters in A_2, m_ij will be 0 or 1, so $B_k = 0$. The result of the clustering was shown in the chapter of result and discussion.

CHAPTER 5

Regression Methods

Classification and regression consistute the two pillars of supervised learning in machine learning. Classification concerns the mapping from a set of predictors to a discrete label, while regression maps predictors to a continuous number. Generalization from a single number to a vector of number is trivial.

5.1 Assemblage of design matrix from network metrices

Suppose we have n traning data, in which each one of them contain p observables, that can be used for model fitting. In the machine learning or statistics literature, we often organize the n training data into a p by n matrix X, which we calld the design matrix. In many fitting algorithm, design matrix is a convenient construct which allows us to calculate the objectives with simple matrix algebra.

We assemble the predictors we collected, namely the PageRank of oncogenes and drug targets, into a design matrix from a set of n data points as

$$\mathbf{X} = \begin{pmatrix} \mathbf{p}_{g,1}^T & \mathbf{p}_{d,1}^T & 1 \\ \mathbf{p}_{g,2}^T & \mathbf{p}_{d,2}^T & 1 \\ \vdots & \vdots & \vdots \\ \mathbf{p}_{g,n}^T & \mathbf{p}_{d,n}^T & 1 \end{pmatrix}, \tag{5.1}$$

where $\mathbf{p}_{g,1}^T$ is the transpose of the PageRank of oncogene from the first datapoint of cell line, and $\mathbf{p}_{d,1}^T$ is the transpose of the PageRank of drug targets from the first datapoint of drugs. The column of 1 is introduced for the fitting of intercept

at the later regression step, which is required for most of the kernel methods. Unfortunately, we will include too many predictors by doing so. So, we will introduce a cutoff probability, to discard those columns where there is no probability larger than cutoff, and use it as the new training matrix, $\mathbf{X}$. For the output y, it also requires transfomation in order to gain better fitting. Originally, the range of y is within $[0, 1]$, and needs to be transformed to $[-\infty, \infty]$. A sigmoidal transformation is introduced, such that

$$y = \frac{1}{(1 + \exp(-\hat{y}))^{\gamma}}, \tag{5.2}$$

where $\hat{y}$ is the transformed output and γ is a shape factor that describe the signoidal curve. The transformed output will be assembled as a output training vector as

$$\mathbf{y} = [\hat{y_1}, \hat{y_2}, \ldots, \hat{y_m}]^T. \tag{5.3}$$

5.2 Gaussian process

Gaussian process is a statistical distribution that describe the distribution of a function, in which every point in the function is jointly Gaussian. It is uniquely specified by its mean function m and covariance function k.

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})], \tag{5.4}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))], \tag{5.5}$$

and the Gaussian process can be written as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \tag{5.6}$$

5.2.1 Prediction with noisy observations

Here, we begin to construct the model of drug sensitivity. Suppose y_i is the logit transformed sensitivity of the ith dose response curve, and $\mathbf{x}_i$ the corresponding

features, $\mathbf{x_i} = (\mathbf{p}_{g,i}^T, \mathbf{p}_{d,i}^T, 1)^T$. To account for the random error that arised from screening experiment, We assume the transformed sensitivity to be

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \tag{5.7}$$

where $\epsilon_i \sim N(0, \sigma^2)$, and σ is the standard deviation of the error of transformed sensitivity. Under these model assumption, we can write the dataset as $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{y} = (y_1, ..., y_n)^T \in \mathcal{R}^n$, and $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathcal{R}^{n \times d}$. d is the dimension of $\mathbf{p}_g$, $\mathbf{p}_d$ combined plus one. The likelihood of the output is then

$$p(\mathbf{y}|\mathbf{f}) = N(\mathbf{f}, \sigma^2\mathbf{I}) \tag{5.8}$$

where $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_n))^T$, and the error terms $(\epsilon_1, ..., \epsilon_n)$ are assumed to be independent and identically distributed.

We assumed that the model follows a Gaussian process prior. Since all points in the function will be jointly Gaussian, we can write the probability distribution of the collection of data output $\mathbf{y}$ and the prediction $\mathbf{f}_*$ as

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right) \tag{5.9}$$

As many of the growth advantage and cell toxicity of drugs are additive and coupled with interactions between several genes, we described the model using a polynomial kernel, which is defined as

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + \sigma_0^2)^p \tag{5.10}$$

Alternatively, we also tried to used squared exponential kernel because it poses less restrictions on the expressiveness of the functional space it includes. Actually, the function space encoded by exponential kernel is infinite dimensional. However, in our later numerical experiment. The exponential kernel perform suboptimally compared with polynomial kernel, presumablly due to the strong additive effect of certain genes.

To simplify the expression, we introduce a shorthand $K(X_*, X)$, a matrix where the i, jth element is $k(\mathbf{x}_{*,i}, \mathbf{x}_j)$ We can compute the conditional probability of $\mathbf{f}_*$ given the data, by applying the Bayes' Law, we have

$$\mathbf{f}_* \mid X, \mathbf{y}, X_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, cov(\mathbf{f}_*)), \tag{5.11}$$

where the posterior mean of $\mathbf{f}_*$, $\bar{\mathbf{f}}_*$, can be calculated as

$$\bar{\mathbf{f}}_* \triangleq \mathbb{E}[(\mathbf{f}_*) \mid X, \mathbf{y}, X_*] = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y}, \tag{5.12}$$

, and the point-wise covariance as

$$cov(\mathbf{f}_*) = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, X_*). \tag{5.13}$$

$\bar{\mathbf{f}}_*$ denotes the predictive mean, and $cov(\mathbf{f}_*)$ is the covariance matrix of the prediction. In the case of a single test point $\mathbf{x}_*$, we can write the mean and variance as

$$\bar{f}_* = K(x_*, X)[K(X, X) + \sigma_n^2 I]^{-1}\mathbf{y} \tag{5.14}$$

$$\mathbb{V}[f_*] = k(x_*, x_*) - K(x_*, X)(K(X, X) + \sigma_n^2 I)^{-1}K(X, x_*) \tag{5.15}$$

5.3 Transformation for better fitting

5.3.1 The mean and variance of prediction

In order to find out the mean and variance of prediction, it is better to consider the inference as random variables. The prediction is a transformation from the logit sensitivity X to the sensitivity Y, where $y = g(x)$, and the sigmoid function g is

$$y = g(x) = \frac{1}{1 + e^{-fx}}, \tag{5.16}$$

where f is the shape factor. The sigmoidal function has the following inverse, the logit function

$$x = g^{-1}(y) = -\frac{1}{f} ln(\frac{1-y}{y}) \tag{5.17}$$

The function of random variable is also a random variable. For a differentiable and increasing function, the differential and inverse are guaranteed to exist. Because g maps all $x \leq s \leq x + \Delta x$ to $y \leq t \leq y + \Delta y$:

$$\int_x^{x+\Delta x} f_X(s)\, ds = \int_y^{y+\Delta y} f_Y(t)\, dt \tag{5.18}$$

It follows that

$$f_Y(y) = f_X(x)\frac{dx}{dy} = \frac{f_X(x)}{g'(x)} = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))} \tag{5.19}$$

We can find the derivative of g for equation 5.16 as

$$\frac{d\, g(x)}{d\, x} = \frac{f e^{-fx}}{(1+e^{-fx})^2} \tag{5.20}$$

From the previous equations, we can derive the probability deisity of the transformation given $x \sim N(\mu, \sigma)$.

$$f_Y(y) = \frac{1}{\sqrt{2\pi} f y(y-1)\sigma} \exp\{-\frac{f\mu + \ln(\frac{1-y}{y})}{2f^2\sigma^2}\} \tag{5.21}$$

However, this equation does not have closed form solution for its mean and variance. Therefore, we need to calculate the approximate mean and variance with numerical simulation. For example, we can use a Metropolis-Hasting Algorithm to sample the pdf. Note that, a numerical integration for mean and variance does not work well here, since a large x usually generate excessively high probability density, making the numerical integration incorrect.

CHAPTER 6

Integration of Databases

In this chapter, we will review all the public databases that were used in generating predictions of PROPHECY, including STRING, STITCH, CCLE and COSMIC. For every database, we will first explain how the data we used from the database is constructed. Secondly, we will review the function of the dataset in PROPHECY. Finally, since most of the databases have different naming convention for the identification of genes and chemicals, we explained in detail about how the data from different databases was integrated.

6.1 STRING

STRING is a online public database that host a wealth of information related to the protein-protein interaction network in various organisms.[SLB00, VHJ03, VJS05, VJK07, JKS09, SFK11, FSF13] The protein-protein interaction included in STRING emcompass not only direct physical interaction, but also functional and predicted interaction. Indeed, direct physical interactions between proteins only account for less than 1% of the total possible interactions between proteins. Proteins can interact in various different ways. For example, several proteins can join together in a larger functional complex, in which two protein in the complex might not directly interact, to accomplish one modular function. Protein can also regulate another protein's transcription event, thereby influence the expression level of another protein. Some protein when activated can serve as a catalyst in the biocheimcal reaction triggered by other proteins. Therefore, including all of

the possible interactions, both physical and indirect, are crucial for comprehensive understanding on a systemic level.

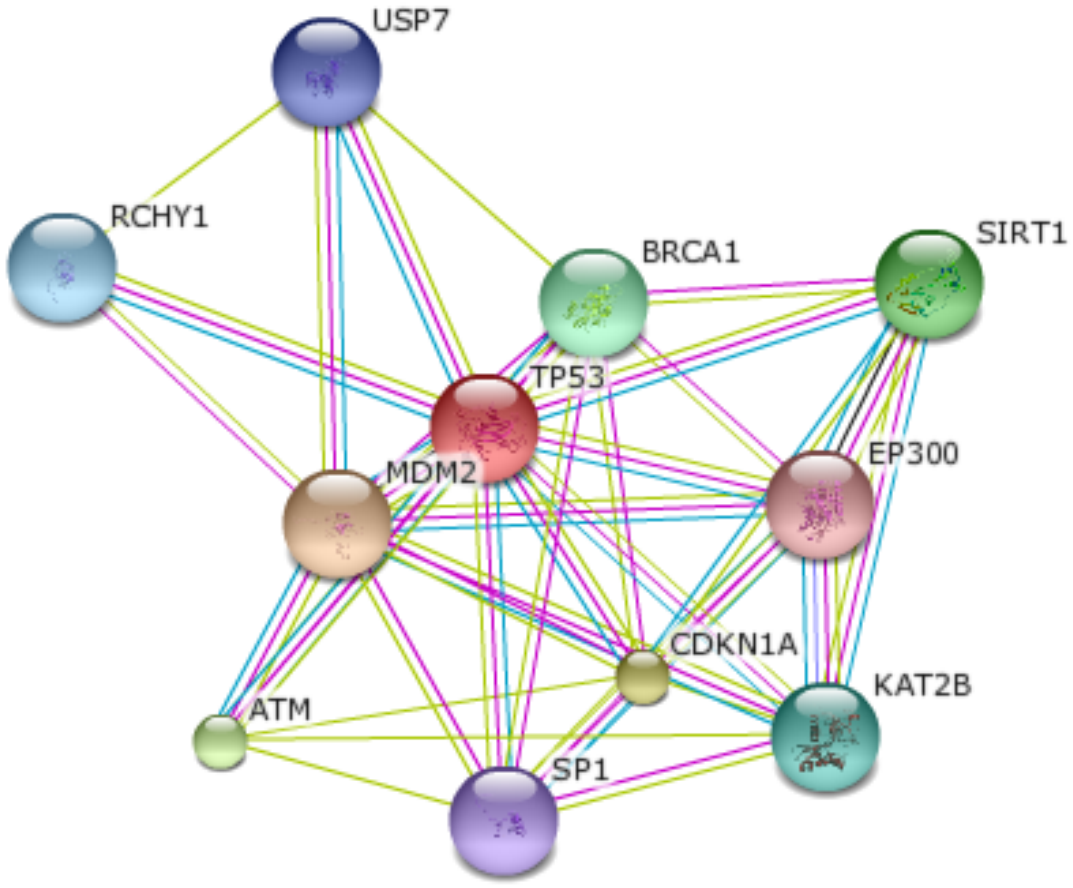


Figure 6.1: Sample query of the P53 gene from the STRING database. The color of the linkages shows the confidence score of association calculated from different type of data collection methods.

The major dataset we extracted from STRING is the homosepian protein protein interaction netwrok, which includes around 22,000 proteins and 1.5 million interactions. We downloaded the entire database and the schema of STRING and hosted the database on a PostgreSQL server. We relabeled the proteins with an internal numerical ID and stored in a dictionary file for the gene names and protein IDs.

For the edges in the network graph, we extract the confidence scores and store the scores in a sparse adjacency matrix. The confidence score is retained instead of using a boolean representation, as the weight fo the edges effects the information diffusion in the network, and is a valuable extra piece of information. The confidence score in STRING is a likelyhood score that is associated with the probability of two proteins having linkage by any one of the methods they used to mine data. Many datamining methods were used by STRING including genomic context, direct experimental result from high throughput protein binding assay, coexpression from microarray data, and text mining from the literature. A high, median, and low score for an edge is 700, 400, and 150. The score s is calculated from

$$s = 1 - \prod_i (1 - s_i), \tag{6.1}$$

where s_i is the ith type of score. Interestingly, we also apply this scoring scheme when combining drugs, and this shows to be very effective in evaluating the interaction when two drugs is applied on the same protein target.

6.2 STITCH

STITCH is a protein-chemical interaction database, which includes 390,000 chemicals and 3.6 million proteins from across 1133 organisms. [KMC08, KSF10, KSF12, KSP14] There were 367,000 high confidence level protein-chemical interactions in STITCH. The interaction data from STITCH is derived from sources including large scale expeirmental databases, manually curated database of interaction evidence with text mining from literature and predicted linkage.

6.3 Cancer Cell Line Encyclopedia (CCLE)

Cancer Cell Line Encyclopedia (CCLE) is a database of cancer cell line genome. [BCS12] It includes information of detailed genetic variation of 947 human cancer cell line which span 36 tumor subsypes. Mutational status of ¿1600 cells are determined by massive parallel sequencing, and germline mutations are removed. We gathered from CCLE the mutation status of 392 recurrent mutations of 33 cancer genes, and the copy number variations. We also obtained the messenger RNA expression data of each cell line from CCLE, in order to compare the disease gene mutation with gene expression. We mapped each genetic signatures onto the PPI network. Before mapping, silent and intron mutations are removed from the gene mutation list, since they do not effect the phenotype.

6.4 Genomics of Drug Sensitivity in Cancer (GDSC) of the COSMIC database

The Genomics of Drug Sensitivity in Cancer (GDSC) database is the largest database of cancer cell line drug sensitivity.[YSG13] The database host data of around 75,000 drug screening experiments, representing the therapeutic profile of 138 anticancer drugs, chemo or targeted therapy, across around 700 cell lines. Additionally, the database also offer data analysis for biomaker discovery purpose.

The screening experiment was designed following the same protocol. First the cell lines is seeded in microtiter plates, and incubated for 24 hours, and a serial dilution of drugs is applied to wells followed by another incubation of 72 hours. After 72 hours, fluorescent marker of cell viability is measured, and the intensity will be normalized to control and background to obtain the viability and the dose response curve. GDSC provided the dose response in terms of several fitted parameters. They use a home brewed parametric model of Bayesian smoothing,

which was reviewed in previous chapters in this dissertation. IC_x was provided along with other sigmoidal parameters including α, β, f, and B. The integrated sensitivity data from the dose response curve is also provided. In PROPHECY, every initial data point is consist of a drug name, a cell line name and a sensitivity.

The data from GDSC uses different format in terms of the identification of drugs or cell lines from other databases that we gathered, so additional mapping steps must be done in order to use the data. The drugs needs to be identified precisely to match the chemical formula for the purpose of integration with STITCH, from which the drug targets can be queried. We first matched the drug names from DrugBank, a database that houses information of FDA approved drugs and experimental drugs, to get the pubchem ID. Through the pubchem ID we can map the drugs from GDSC to STITCH. From all of the 138 drugs in GDSC, we were able to retrieve 100 drugs and find matches in STITCH.

For cell lines, due to the fact that CCLE holds more up to date and comprehensive data on the genetic variation in cell lines compared with COSMIC, the parent database of GDSC, we need to map the name of the cell ine to CCLE. Among the 700 cell lines in GDSC and the 947 cell lines in CCLE, we found an overlap of 400 cell lines.

6.5 Mapping of data across databases

Format of STRING Most of the mapping for genes and proteins goes back to STRING, which was the most comprehensive in terms of the scale of the genes and proteins included. In STRING, each protein has an interger ID, ENSG ID, and the Hugo Symbol of the gene (which is the same as the preferred name of the gene).

Format of STITCH STITCH comes from the same group which developed STRING and has the same protein naming convention, so the drugs in STITCH can be easily matched to their target proteins in STRING. In STITCH, the chemical IDs were given the form "CIDXXXXXXXXX", where the "'XXXXXXXXX' part is the PubChem compound ID.

Format of CCLE The cell line information in CCLE has the format *name_lineage*, e.g. MDAMB231_BREAST. The disease genes in CCLE are identified by Hugo Symbol. The genes in CCLE are matched to STRING's proteins by first matching the gene name directly. For those genes that remained unmatched, we used the HUGO Gene Nomenclature COmmittee (HGNC) Biomart to convert the gene names of CCLE into an ENSG ID, and match to the ENSG in STRING. For a few genes that remained unmatched, we checked if the ENSG ID can be found manually at ENsembl or STRING or HGNC's website, then match the query result to the ENSGs in STRING.

Format of GDSC The cell lines in GDSC has the format *name* with dashes, e.g. MDA-MB-231. These are matched to the cell lines in CCLE by removing the lineage from CCLE's names and the dashes from GDSC's names so the name of the cell lines can be matched.

The drug names need to be matched to STITCH's drugs by using the drug's PubChem compound ID, since our attempts to query by drug name did not work well. We manually searched the PubChem ID for each of the 140 drugs in GDSC. We ethen convert the integers formated PubChem IDs into the format of the STITCH chemical IDs and match to STITCH.

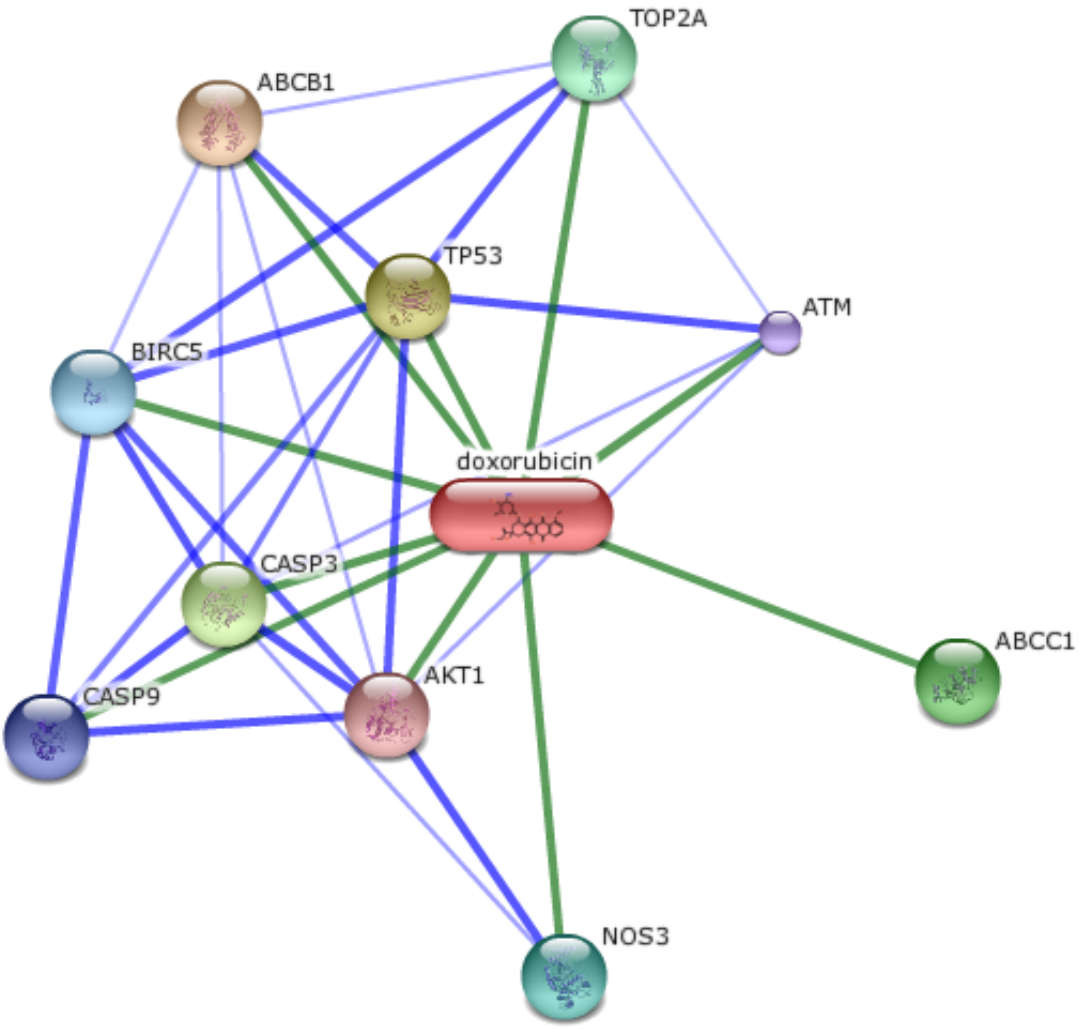


Figure 6.2: Sample query of doxorubicin from the STITCH database. Stronger associations are represented by thicker lines. Protein-protein interactions are shown in blue, chemical-protein interactions in green and interactions between chemicals in red.

CHAPTER 7

Result and Discussion

The main goal of this chapter was to give a summary of the evidence which supported the quality of predictions given by PROPHECY. We first demonstrated that the network diffusion signature of gene mutations is a powerful classification method that is capable of showing the drug resistance profile of cell lines by a comparative clustering analysis. Next, we compare the quality between the predictors and different regression method by various standard methods, and proved that PROPHECY has superior predictive power in every measure. Finally, a set of two drug combination experiment is performed to demonstrate the predictive power of PROPHECY on drug combinations. We showed that PROPHECY can predict the drug sensitivity with high accuracy even though it is never trained by combination drug data. The panel of 6 cell lines tested by the experiment also proved that PROPHECY is applicable to a wide range of genetic profiles. A set of high performing two drug combinations are identified by PROPHECY and then experimentally verified. The combinations discovered include novel drug combinations that have relavant biological indications, and approved combination that further confirm the strenghthen of the predictions.

The chapter were divided into sections representing major hypothesis made in the project. In each section, we first addressed the central questions we wanted to answer to prove the predictability, and then the rational behind the hypothesis. We provided the detailed numerical methods for simulations, data analysis and/or physical experiment performed to verify the hypothesis. Conclusion and discussion

were given at the end of each section.

7.1 Spectral clustering showed network diffusion signature as a powerful observable for classification

Clustering is a method used widely in biology to elucidate structures within a spectrum of populations without prior knowledge of how the population should be classified, or, in the terms of machine learning, the data points is unlabeled. For example, in a typical biomarker study, the vectors of gene expression level of several genes from different cell lines at regular culture condition can be used to classify the cell lines by clustering the cells. Clustering of cells often result in unexpected classification which is otherwise non-obvious without the aid of biomarkers. Many studies have found that lineages of cell lines are often less than optimal predictors compare to genetic profile or gene expression profile.

The purpose for clustering analysis is to prove that network diffusion is a good way for the classification of cell lines. Contemporary classifications of disease are often based on the pathology of the patient, and it had been proven ineffective by modern molecular diagnostic tools. [BGL11] Now that we have shown in a theoretical perspective that network diffusion can highlight the functional modules being affected by the disease. Can PageRank be further extended to search for cells that have similar response to therapy? If cell lines which have similar therapeutic profile can be clustered in the feature space, it can prove that the features are good indicators for classifying disease.

The major hypothesis we made is that the clustering made by PageRank would be more similar to that of therapeutic profile compared with clustering by other biomarkers. As introduced in chapter 4, we can measure the similarity between drug treatment and biomarker by the B_k score. Notice that we will select hierarchical clustering as the clustering method, as this method does not require the

user to determine the number of clusters k.

The first step toward the comparison is to perform hierarchical clustering of cell lines based on the drug response. We need a N_{CL} by N_{drug} design matrix to compute the cluster. For the drug response, however, not all of the drugs have complete data for all of the available drugs, so we need to fill in the missing data first with PROPHECY. 430 cell lines were matched to our database to have both genetic profile and dose response data. In total, 100 drugs were tested on the cell lines, so in total need 43,000 data points to assemble the design matrix. We only had 31,170 data points of sensitivity. So, excluding also the duplicated data points which same drugs and cell lines are used multiple times, another 12163 data points were to be generated from PROPHECY. After filling the data with PCA regression, the similarity matrix is completely filled.

Usually, hierarchical clustering is accomplished by directly calculating the distance in terms of the observations. This type of methods often uses spherical or elliptical metric to group data points, so they do not work well when the clusters in the feature space is non-convex. [HTF09] For example, if one group of data forms a donut structure in the multidimensional space, while another group sitting in the center of the hyper donut, the traditional clustering will not resolve this problem. So, instead, we use a method that closely related to spectral clustering, that is, we use the N_{CL} by N_{CL} similarity matrix to measure the distance for each observation as their relative distance to every other observations. By doing so, the hierarchical clustering can reveal the global structure of the grouping between observations.

In drug response case, the similarity matrix is computed based on the Pearson correlation of 100 drugs' response of cell line pairs. After the similarity matrix is computed, we use the matrices to cluster each case by average method. The resulting clustergram of drug response was illustrated in figure 7.1.

In the gene expression level case, we take the z score of all genes, and compute

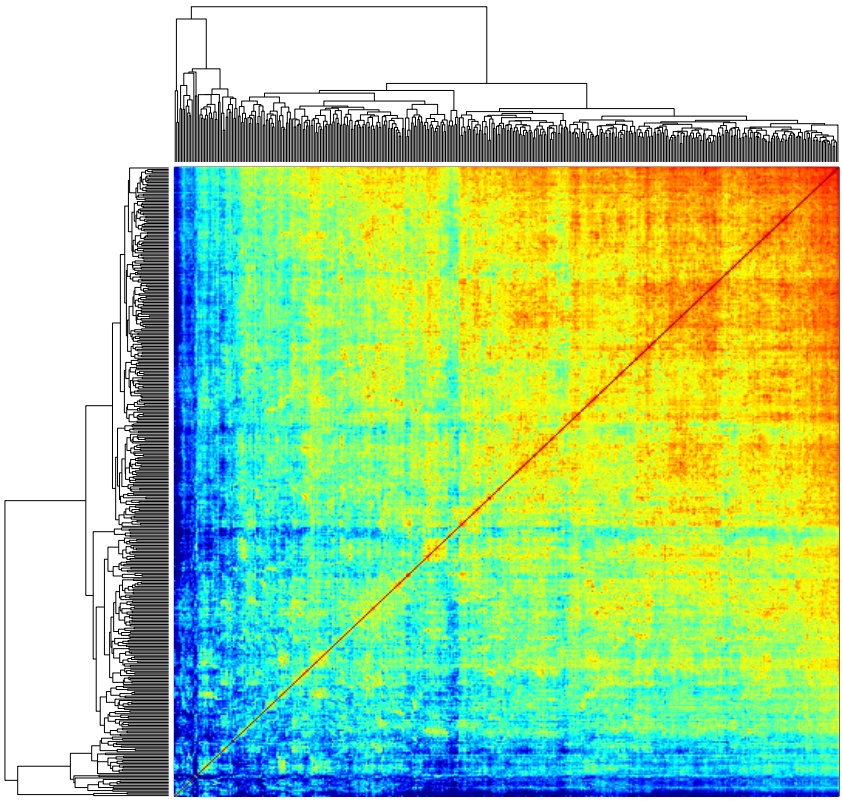


Figure 7.1: Hierarchical clustering of cell lines based on the similarity matrix of the Euclidean distance between cell lines in the feature space of drug sensitivity. We found that the clustergram forms clear checkerboard structure, indicating there is a clear distinction between group of cell lines. Also, lineages of cell lines are clearly not the major determinant in the grouping, many different lineages of cell lines were grouped together due to sharing of similar responses to treatments.

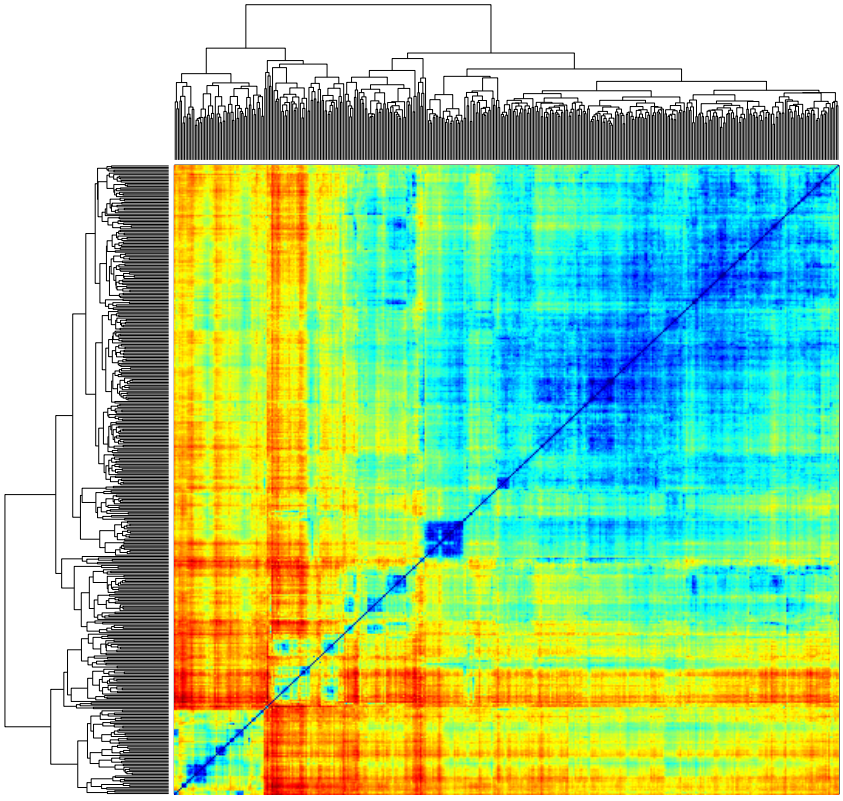


Figure 7.2: Hierarchical clustering of cell lines based on the Euclidean distance of cell lines in the feature space of gene expression level.

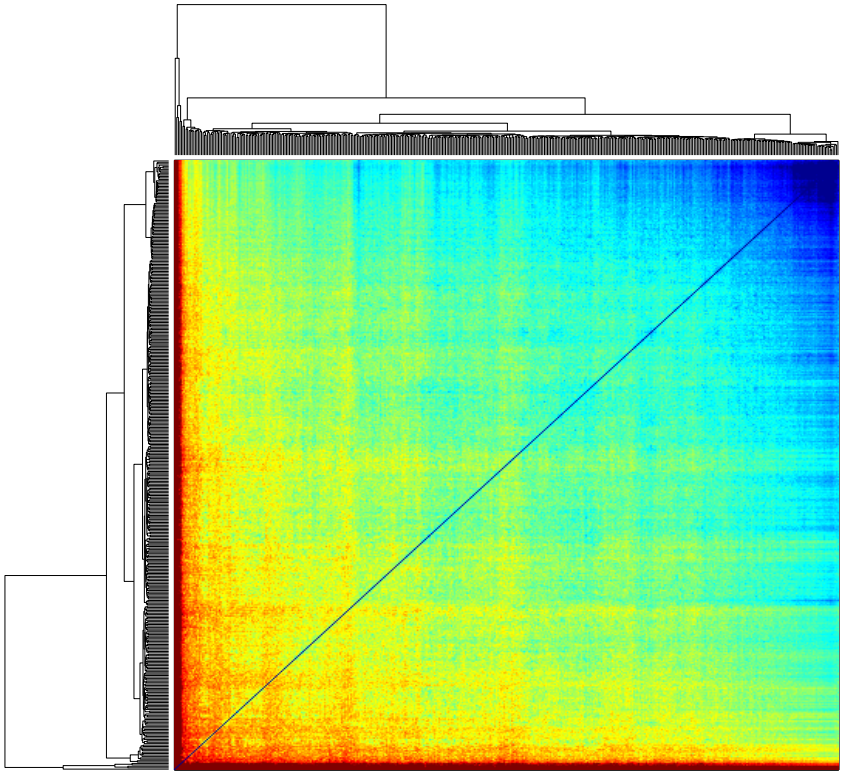


Figure 7.3: Hierarchical clustering of cell lines based on the Euclidean distance in the feature space of PageRank.

the Euclidean distance between the gene expression level vector of pairs of cell lines. The transformation to zscore is cutomary for the data analysis of gene expression level. The resulting clustergram of gene expression was illusterted in figure 7.2.

In the PageRank case, the similarity matrix is computed based on the Euclidean distance between the PageRank of pairs of cell lines, and the result is shown in figure 7.3. On the surface, we can check the similarity by looking at the checkerboard pattern of each clustergram. The shading produced by PageRank

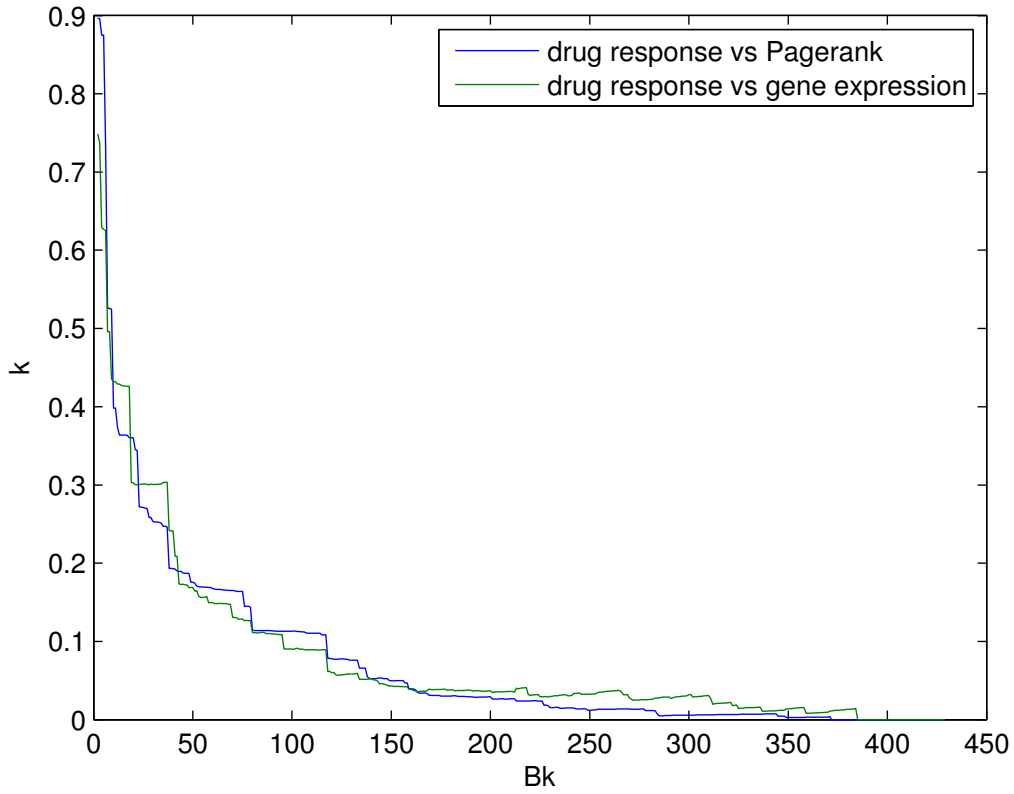


Figure 7.4: Bk versus the number of clusters, k. PageRank showed better performance than gene expression in multiple ranges of k.

appears to be closer to that of the drug response than the gene expression. However, this type of visual discrimination can be misleading at times, so we must use a more quantitative way to measure the similarity of two clustering.

In order to compare the quality of the clustering, we use the method mentioned in the paper, "a method for comparing two hierarchical clusterings" to calculate the similarity score Bk corresponding to each number of clusters k. [FM83] We treat the drug profile clustering as the gold standard clustering, and compare PageRank and gene expression against it. We found that the PageRank is more similar to drug response clustering at low k, and similar to gene expression at middle k (10 - 150) demonstrating that PageRank is a good measure for the drug

sensitivity. Given that measuring mutation is cheaper than measuring gene expression with current technology, we propose that PageRank of cell line mutation profile is an excellent candidate for both preclinical and clinical diagnostics.

7.2 Analysis of the quality of the sensitivity model

The central function of PROPHECY is to predict the sensitivity of the cell line based on the genetic profile of the cell and the drug target association. The sensitivity is a continuous number in the range $[0, 1]$, and hence a regression analysis is appropriate. The central question we want to answer is whether the prediction made by the regression is close enough to the true value. How do we measure the closeness? How do we get testing examples? There are several metric we can use to determine the quality of certain regression analysis. The first one is the Pearson correlation between the prediction and the experimental result. Usually, whether certain value of the Pearson correlation is good or bad depends on the system under investigation. Other measurement of the quality of regression algorithm is usually relative, which means that the performance of the algorithm needs to be compare to other methods judging by certain benchmark functions.

7.2.1 Cross validation on the single drug dataset

Cross validation is a technique widely used to determine the quality of the fitting of regression. [Bis06, Ras06] In cross validation, we divide all of the available data points in to training set and testing set. The training set is used for model fitting, and then the testing set will be compared against the training set to measure the Pearson correlation between the prediction and the testing set. Due to the fact that the testing set is disjoint from the training set and the regression had never seen the testing set before, the predictability of the method itself can be measured quite accurately. There are several versions of cross validation, depending on how

the data set is divided. Leave one out cross validation is suitable in the cases when the training data is expensive or hard to obtain, so all of the data are used to train the model except for one testing case is left out. The model is then trained N times and the data collected to analyze the Pearson correlation. However, in our case, we have around 30,000 data points and we are computing a highly nonlinear kernel function, so leave one out cross validation is too expensive to compute. Instead, we resort to k fold cross validation, in which the data points are randomized and then divided into k disjoint subsets. Each time, one subset is used as testing set while all of the other training set.

We computed a 3 fold cross validation for the single drug case from the COSMIC database. We can then collect all of the predictions and the test cases, essentially the whole dataset, and plot the correlation plot as in figure 7.5. The correlation coefficient is 76% in the case of Gaussian process regressor with PageRank predictor, which is 20% higher compared with the average of 50% that is reported in the literature by using elastic net regression and genetic signature as predictors. [BCS12]

7.2.2 Compare information gain by predictors: PageRank vs Gene scores

To figure out whether it is the machine learning model or the difference in predictor which contributed to the higher correlation (and thus a better learning), we did a 3 fold cross validation on the Gaussian process model with both the gene score predictors and PageRank predictors. Our rational is that the epistasis of the predictors should be important, so only a machine learning with polynomial kernel has the capability to take interaction terms into account. The result of the comparison is demonstrated in a box plot as shown in figure 7.6. It is shown that by switching to PageRank there was a 20% information gain in the system, which translates significantly when applied the algorithm for drug selection.

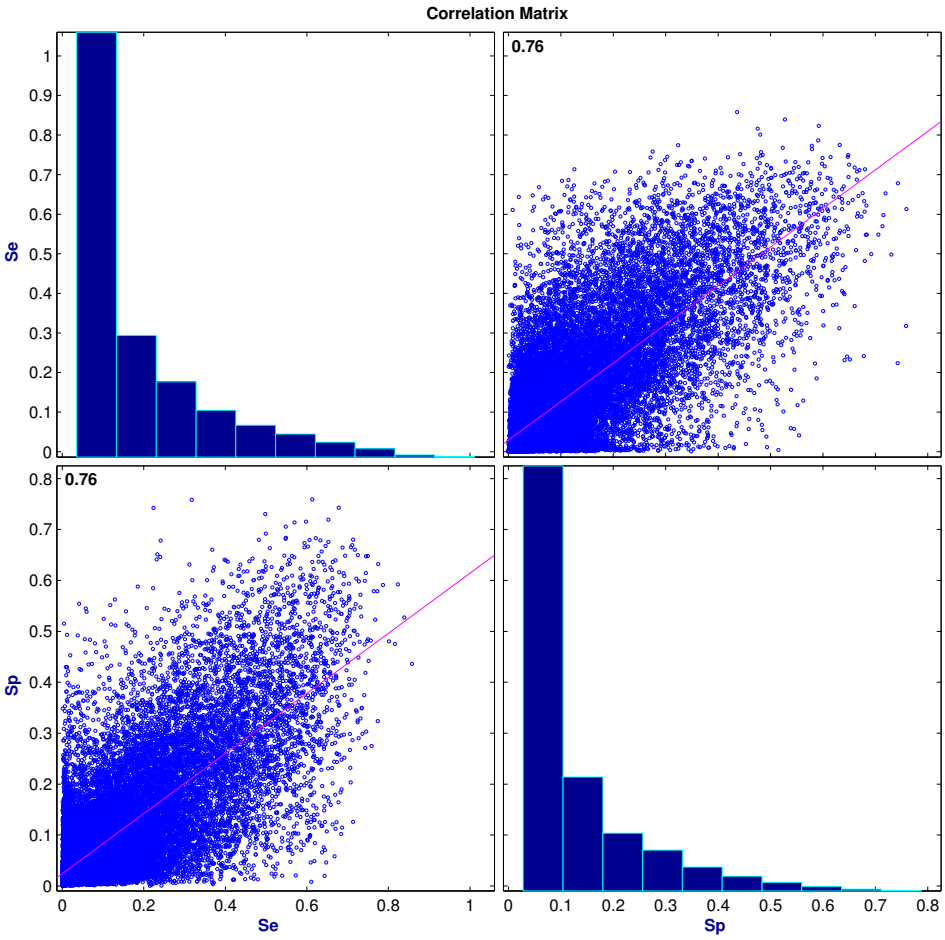


Figure 7.5: Correlation plot of the result of 3 fold cross validation of the Gaussian process predictor on 40,000 entries of sensitivity data. The total predicted sensitivity showed 76 % correlation with experimental data, which exceed the best prediction reported in the literature. [BCS12] The histogram of experiment and model showed similar trend, indicating that the model is capturing the underlying mechanisms.

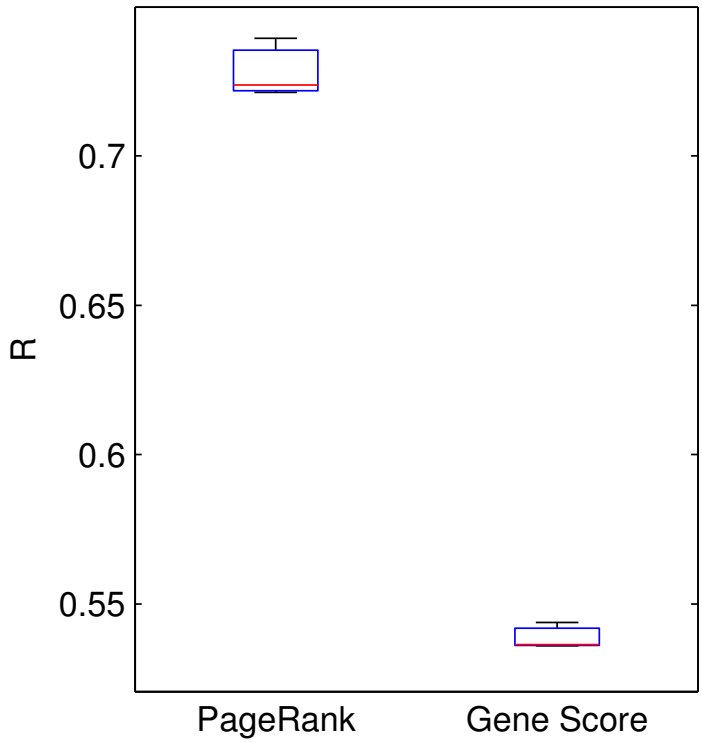


Figure 7.6: Box plot of the Pearson correlation coefficient of the result of cross validation from the PageRank model and the gene score model.

7.2.3 Learning rate of PROPHECY

The learning rate of a regression method is also an important measure of the performance. We are only training the machines with single drug data for preliminary test. However, the space of combination drugs is supposedly much larger than that of the single drug. For the regression method to perform well in predicting unseen combinations, it must pick up the epistasis of the PageRank very quickly. So, in order to see the learning rate, we plot the number of training points against the Pearson correlation coefficient, the result is shown in figure 7.7. We can see that the machine learning algorithm quickly learned the interactions between the PageRank predictor, and saturated at around 5000 data points, which comprises around 12% of the total possible training points.

To see the how the trend of the curve compare to exponential growth, we further plot the curve on a semilog scale, the result can be seen on figure 7.8. We can see that the learning rate grew exponentially at first and then saturated afterward.

7.2.4 Compare the intragroup correlation between regression methods

There are two major use of prediction in terms of therapeutic purpose. One way is to use it to predict the therapeutic response of different genetic signature, or different cell populations, given a particular drug. The second way is to select the best drug or combination of drugs given a particular genetic profile. In order to see the characteristics of PROPHECY in terms of both task, we collect the data of correlation coefficient by separating the training set into group of cell lines but different drugs or groups of drugs but different cell lines. We built the model with all training data, and then made prediction for one drug with a panel of cell lines, and calculate the Pearson correlation coefficient. We first plot the result from Gaussian process regression and from PCA regression, which is shown in figure

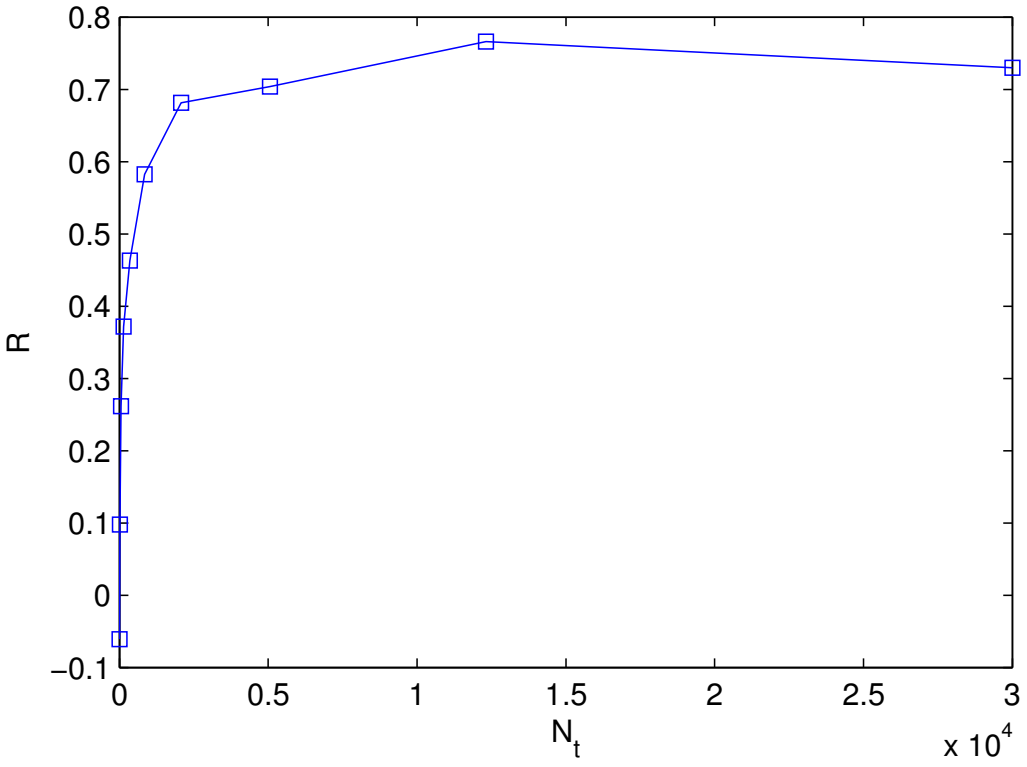


Figure 7.7: Number of randomized training point, N_t vs the Pearson correlation coefficient, R. This can demonstrate the learning speed of PROPHECY. This is PROPHECY trained with PageRank as independent variable and Gaussian process with polynomial kernel where $p = 3$.

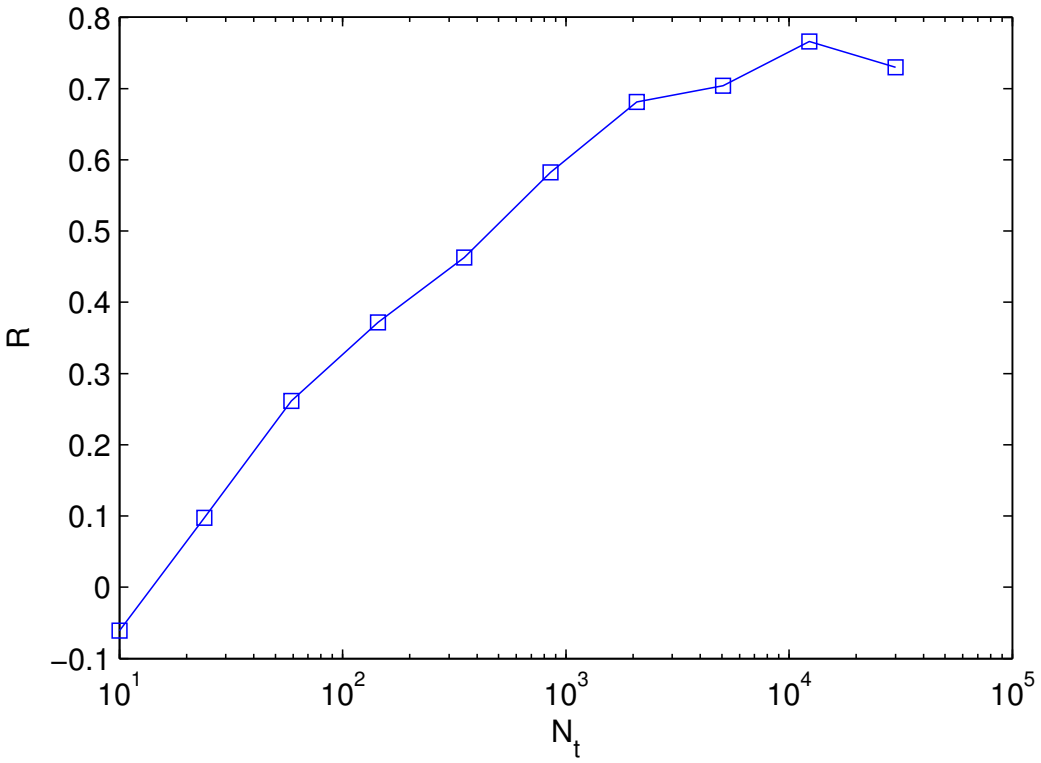


Figure 7.8: Number of randomized training point, N_t vs the Pearson correlation coefficient, R on a semilog scale. This can demonstrate at which point does the learning speed of PROPHECY follows exponential growth. This is PROPHECY trained with PageRank as independent variable and Gaussian process with polynomial kernel where $p = 3$.

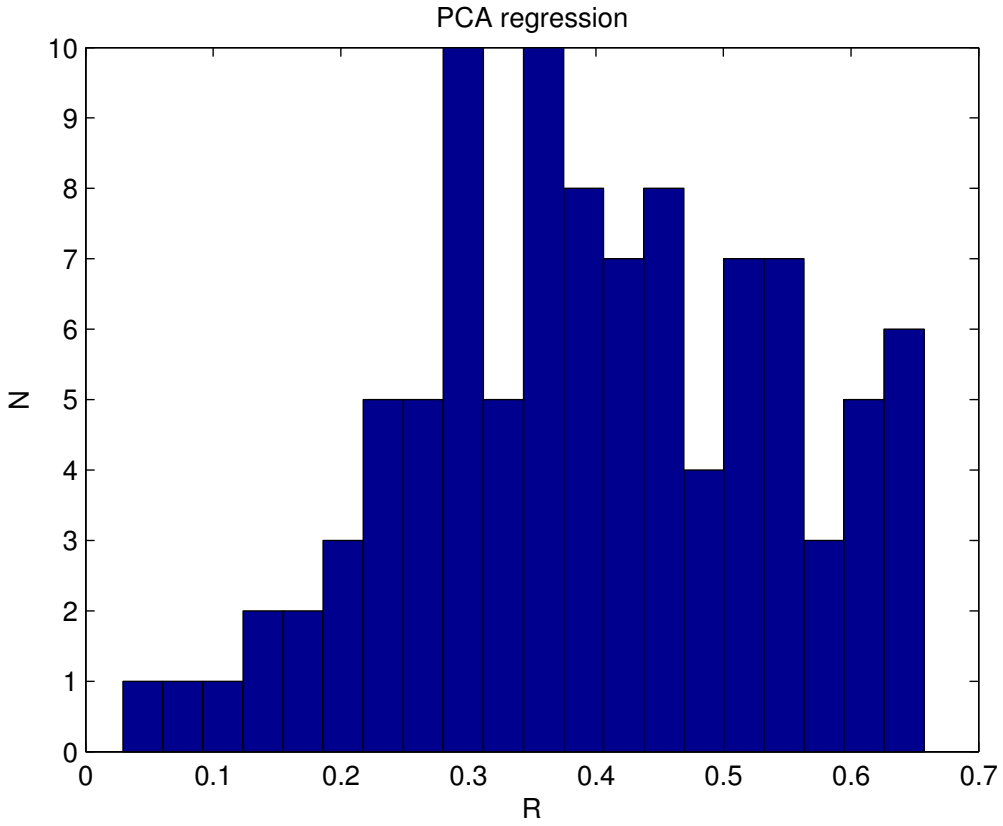


Figure 7.9: Histogram of correlation coefficient as separated by cell lines with PCA regression. The average correlation coefficient is 38%.

7.10 and figure 7.9.

We then plot the histogram of Pearson correlation coefficient of each cell lines with a spectrum of drugs. The resulting histogram is shown in figure 7.11. We found that the performance seems much higher in average compare with seperation by drugs. The reason is that the algorithm has better knowledge of the efficacy of the drugs, since each drug is trained with around 400 examples while for each cell line it is trained with only around 70 example of drugs. So, depending on the quality of the dataset we have, it will also affect the predictive power of the regression model.

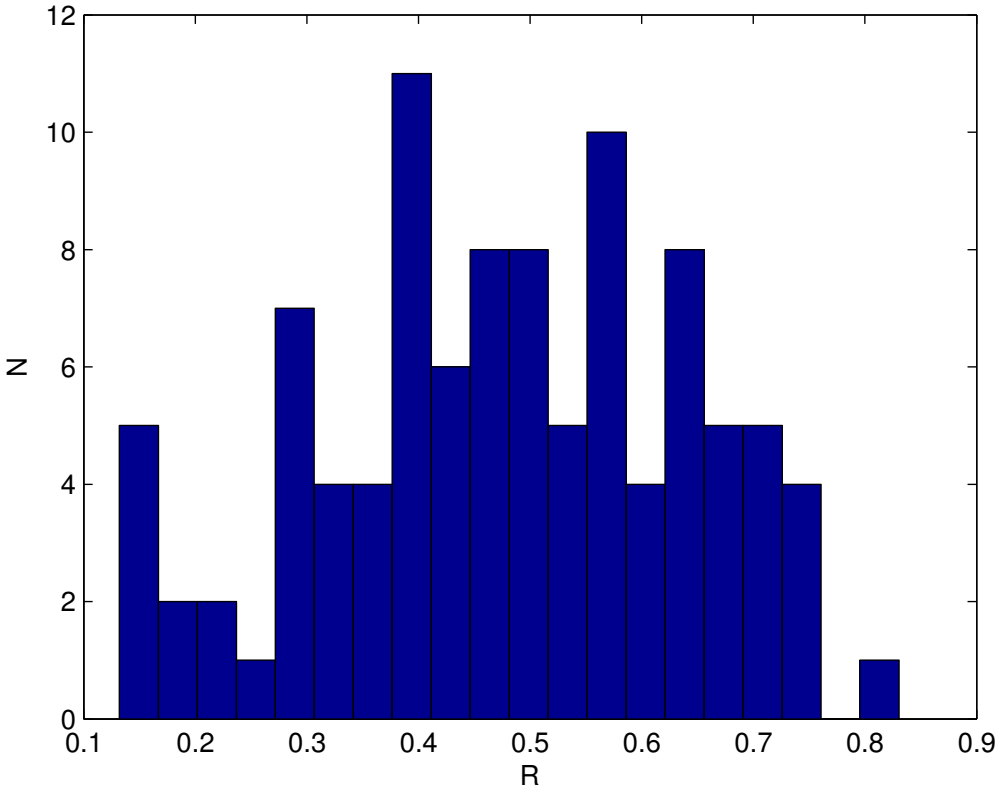


Figure 7.10: Histogram of correlation coefficient as separated by cell lines with Gaussian Process regression. The average correlation coefficient is 54.4%. It has in average higher correlation than PCA regression, but some drugs has less than optimum performance. We reason that it is due to those cell lines which have genetic profile separated far away from other cell lines, as shown by multidimensional scaling study of the cell lines.

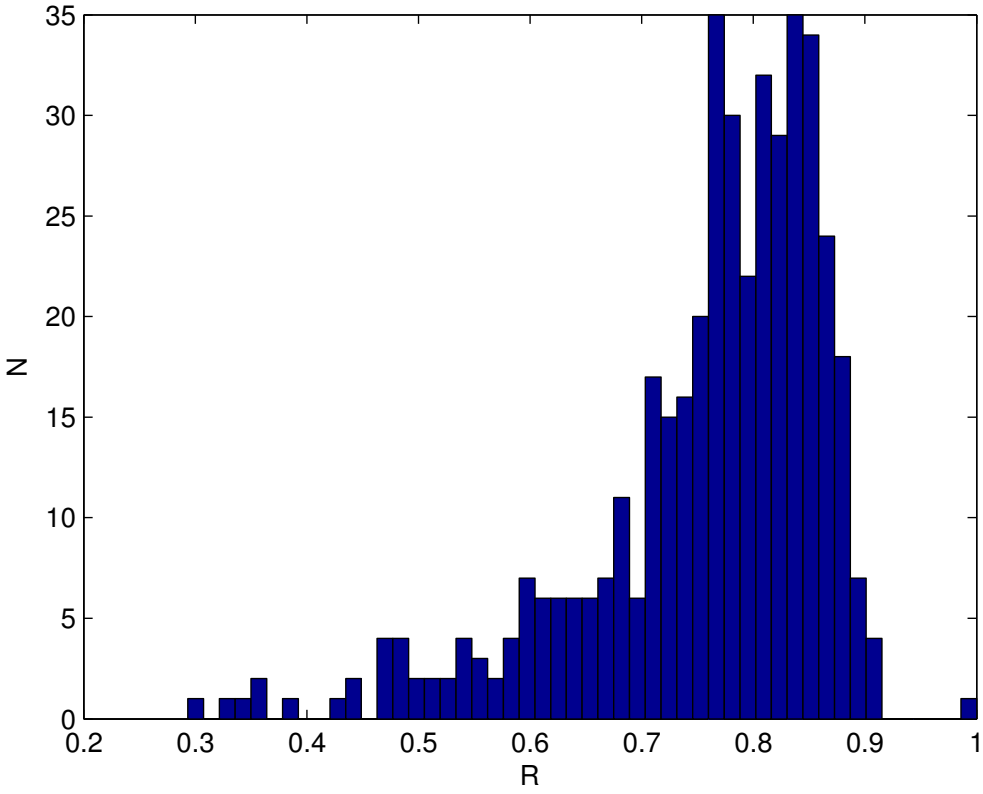


Figure 7.11: Histogram of correlation coefficient for each cell line as calculated with a profile of drugs by Gaussian Process regression.

7.2.5 Reciever operating characteristics curves by different predictors

One other way to compare the predictability of two learner is to plot the receiver operating characteristics curve (ROC curve). In a classification setting, the learner divides the population of all training set into two populations. The population is determined by comparing the prediction with a threshold value. If the prediction is higher than threshold, the data point is accepted, and vice versa. We can then calculate the sensitivity or true positive rate of the model prediction as

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{7.1}$$

where TPR, TP, P, and FN denotes true positive rate, true positives, positive cases, and false negatives. Also, we can calculate the fall-out or false positive rate as

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{7.2}$$

where FPR, FP, N and TN denotes false positive rate, false positives, negative cases, and true negative.

The ROC curve compare the false positive rate of a regressor against the true positive rate with respect to different threshold value. When subjected to random guesses, the ROC curve will appears to be a straight line across the diagonal; any classifier better than random guess will result in a curve on the upper left of the diagonal. The better the predictor the closer the curve will be to the top left corner, as it shows that the predictor will have very little false positives while having very high true positives.

The result of the calculated ROC curve with the threshold α, in this case the sensitivity, is shown in figure 7.12. We compare the ROC curve of the PageRank and gene score predictors both trained with Gaussian process regression with polynomial kernel with $p = 3$. Clearly, PageRank outperforms the genetic profile.

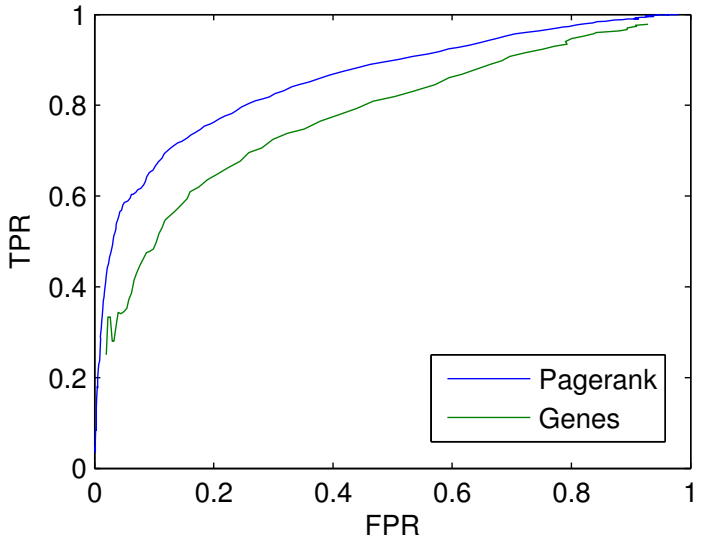


Figure 7.12: The reciever operating characteristics curve (ROC curve) with different predictors trained by the Gaussian process model.

7.3 Experimental verification of PROPHECY, a two drug combination verification

In order to demonstrate that PROPHECY can accurately predict drug sensitivity of a diverse spectrum of cell lines, we performed a combination serial dilution experiment with 3 lineages of 6 cell lines, including MDA-MB-231(breast), MDA-MB-468(breast), KG1(hematopoietic and lymphoid tissue), K562(hematopoietic and lymphoid tissue), A549(lung), and NCI-H522(lung). We trained the PROPHECY model with 30,000 data points of single drugs, and use the model to predict the result of all 2 way combinations possible for the given 100 drug library for each cell line, resulting in 29,700 sensitivity scores. We then separate the sensitivity of each cell line into groups of high, medium, and low sensitivity, and randomly pick 2 combinations from each group. Each combination was assembled from the highest concentration used in COSMIC and then serial diluted, added to assay plates. We then measured the sensitivity by integrating the dose response curve

of the combinations.

In this experiment, we then select a initial drug library of 30 drugs, which include a spectrum of drug from low activity to high activity, to demonstrate that PROPHECY can not only predict useful result, but also low sensitivity result. We then use the program to generate the projected sensitivity of the 435 drugs, by using the model trained with all 30,000 cosmic drug sensitivity dataset. Note that the program was never trained by any drug combination before.

7.3.1 Design of experiment

All cell culture is done as specified in the ATCC guideline. We harvest the cells one day before plating the drugs onto microtiter plates. We seeded 10,000 cells along with 75 μl of medium into each well of a 96 well plate, and incubated for 24 hours. The 96 well plates we used is with dark wells and transparent bottom to prevent interference while reading the plates in a spectrophotometer. We then prepared the combination drugs using a Hamilton STARLet system liquid handling robot.

We set all the ratio of the 2 way drug combinations to be the ratio of the maximum concentration of the single drugs. Each drug combinations is 2 fold serial diluted with 9 stages plus one medium only stage. We plated the wells of the edges of the 96 well plate with medium only, and those wells serve as background control. Each row contains one combination drug, from the highest dosage to the lowest, while the last column contains the blank control. After 24 hours of seeding cells, we then plate the prepared drugs each well with 25 μl of medium mixed with appropriate amount of drugs. The cells are then incubated for 72 hours for the drugs to take effect. After 72 hours, we added resazurine to determine the viability of the cells. After 4 hours of extra incubation, we place the plates in a spectrophotometer to measure the fluorescence of the reduced form of resazurine, resorufin.

7.3.2 Result of the two drug experiment

After the experiment, we use the standard 1 dimensional dose-response curve of experimental data to calculate the sensitivity. We found that the calculated experimental sensitivity produced a 70% correlation with the model prediction provided by PROPHECY. This demonstrate that the model actually learned the contribution of sensitivity from drug target, and that is why the model can accurately predict useful result even though it has never seen combination before. Nevertheless, it is worth to note that there might be unexpected interactions generated from feature pairs that is never seen by the program. So, the program and experiment can be executed iteratively, in order to allow discovery of unexpected interactions, and PROPHECY will have a net gain in network interaction knowledge.

Interestingly, we also did experiment corresponding to different ratio mix of the 2 way drug combinations. However, the correlation of the prediction drops slightly. We hypothesize that it is due to the fact that the ratio change requires a nonlinear transformation in the logit space, so the effect of ratio needs to be further investigated to have a better theory.

The 2 way drug experiment has identified multiple drug combinations that has been under investigation, proving that PROPHECY is predictive. For KG1 cell line, we identified the combination etoposide and vorinostat to be the most potent, as shown in figure 7.15, and this combination is currently under preclinical investigation. [SNT09] For K562 cell, as shown in figure 7.16, we found the combination of parthenolide and vinblastine to be effective, agreeing with the study of Dai et al.. [DGC10] We also found out novel combinations for K562 cells such as the combination of DMOG, a hypoxia inducible factor, and vinblastine. We found that the addition of DMOG will sensitize the cell toward vinblastine. Other novel combinations include 17-AAG and thapsigargin for K562 cells, AZD6244 and Dox-

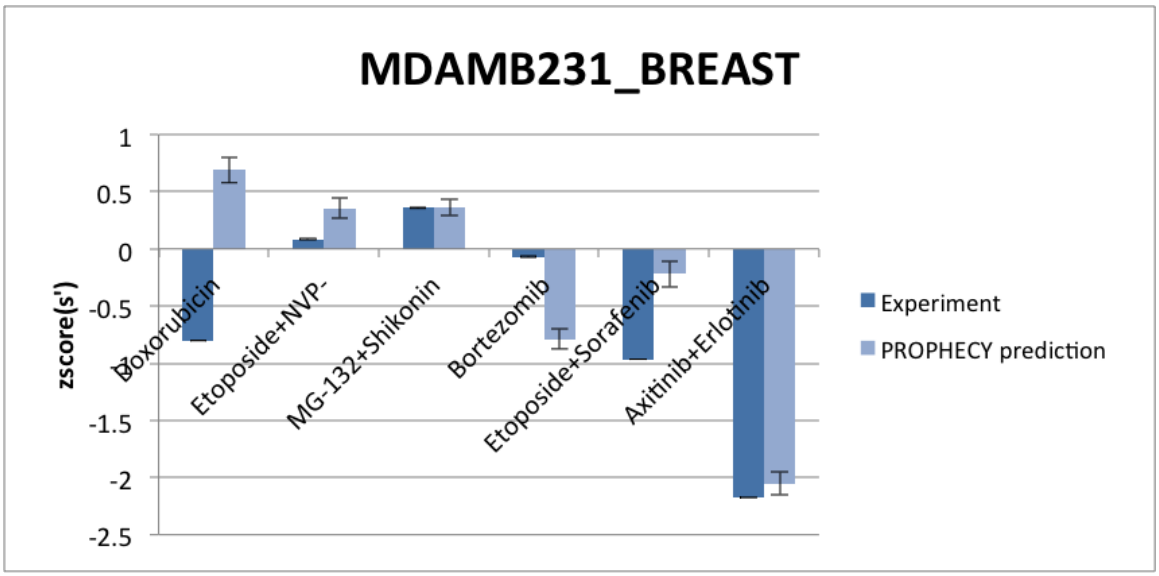


Figure 7.13: Bargraph comparison of 2 drug combination experiment of MDA-MB-231. The y axis is the zscore of the logit transformed sensitivity, which is a way of normalizing the data to compare sensitivities that had slightly different scales. For the experiment cases, the zscore is calculated from all 36 experiment done. For Prophecy prediction, the zscore is calculated based on predictions generated from all possible 2 drug combinations.

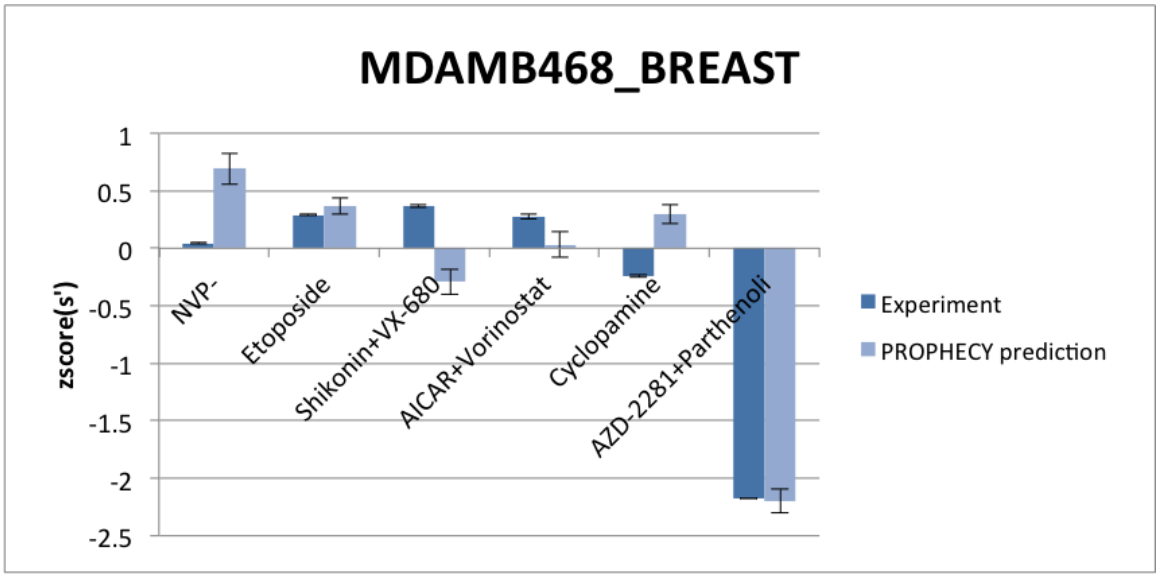


Figure 7.14: Bargraph comparison of 2 drug combination experiment of MDA-MB-468.

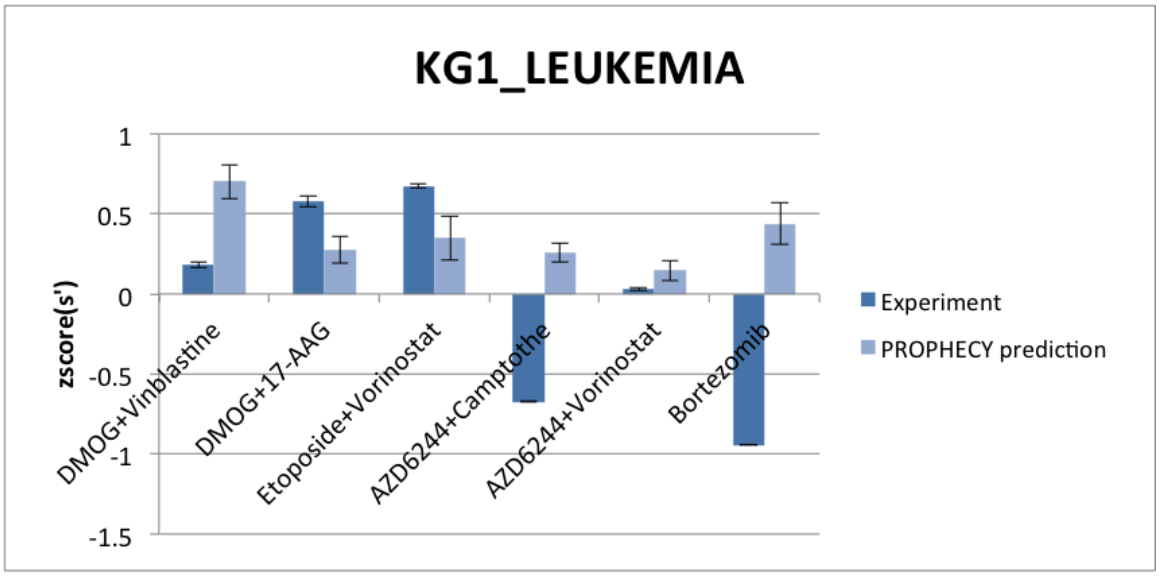


Figure 7.15: Bargraph comparison of 2 drug combination experiment of KG1.

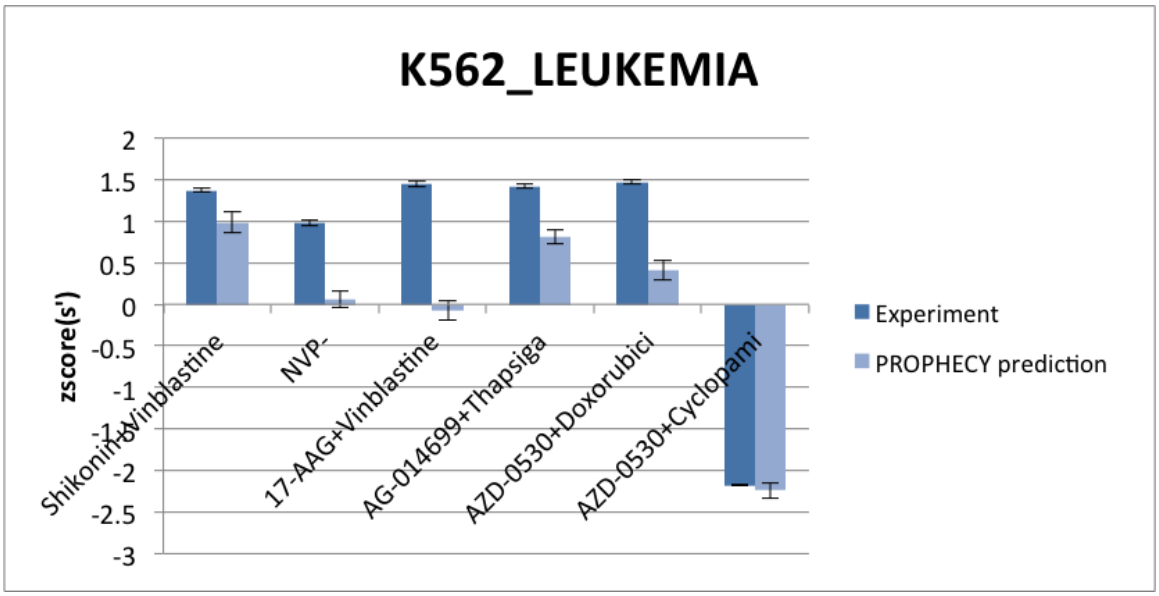


Figure 7.16: Bargraph comparison of 2 drug combination experiment of K562.

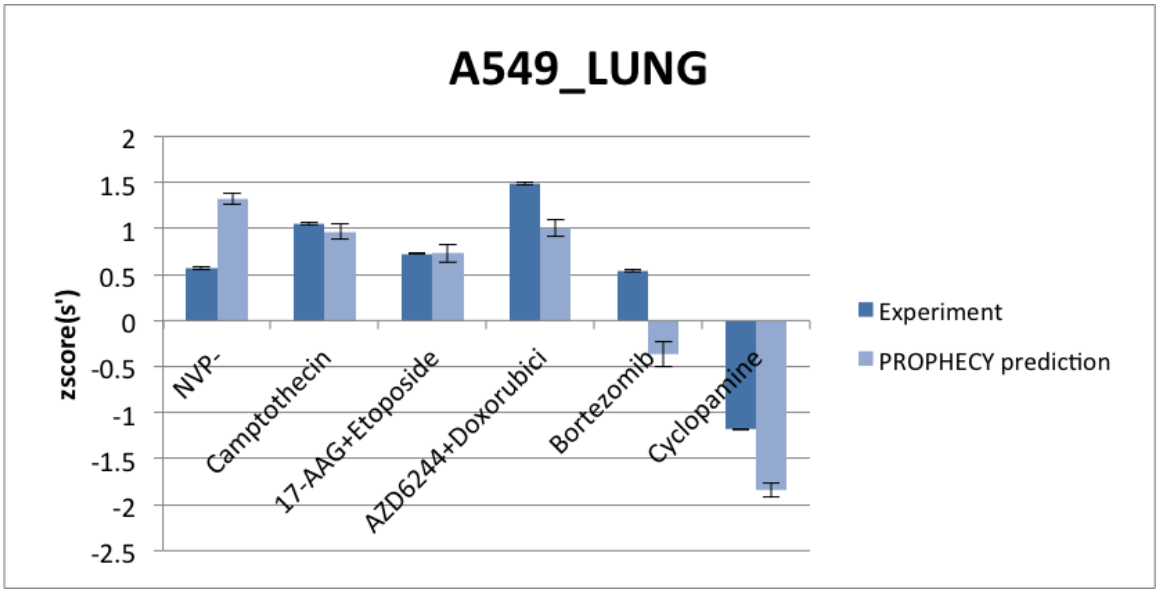


Figure 7.17: Bargraph comparison of 2 drug combination experiment of A549.

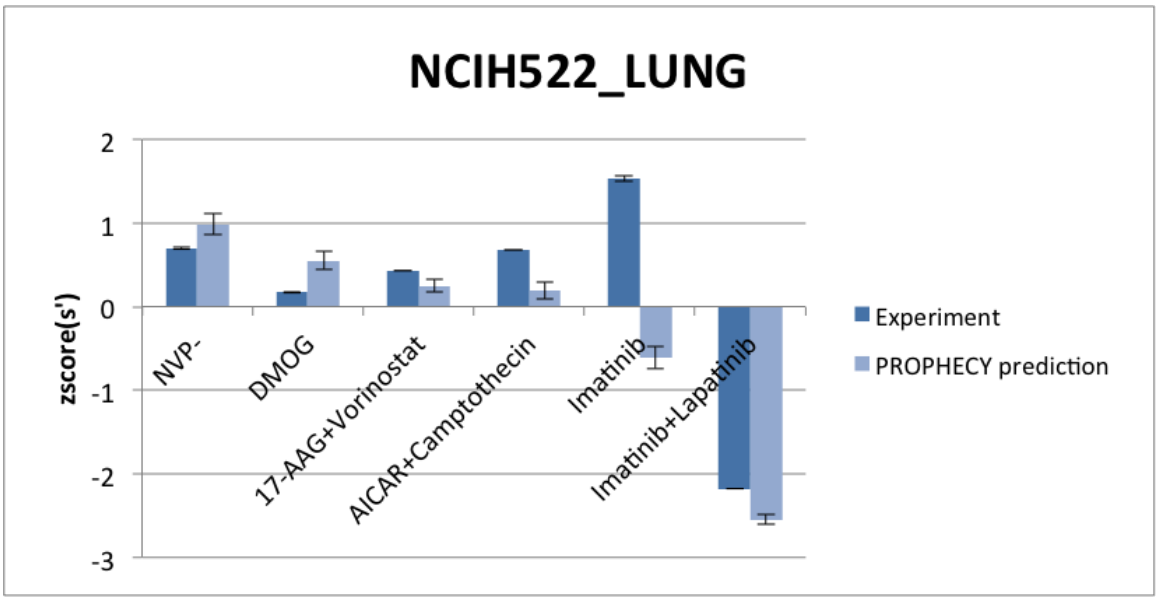


Figure 7.18: Bargraph comparison of 2 drug combination experiment of NCIH522.

orubicin for A549 cells (shown in figure 7.17). Furthermore, we found that the combination of imatinib and thapsigargin, which is found to be effective toward gastrointestinal cancer, will be effective toward NCI H522 cells, a lung cancer cell (shown in figure 7.18). [JNT06]

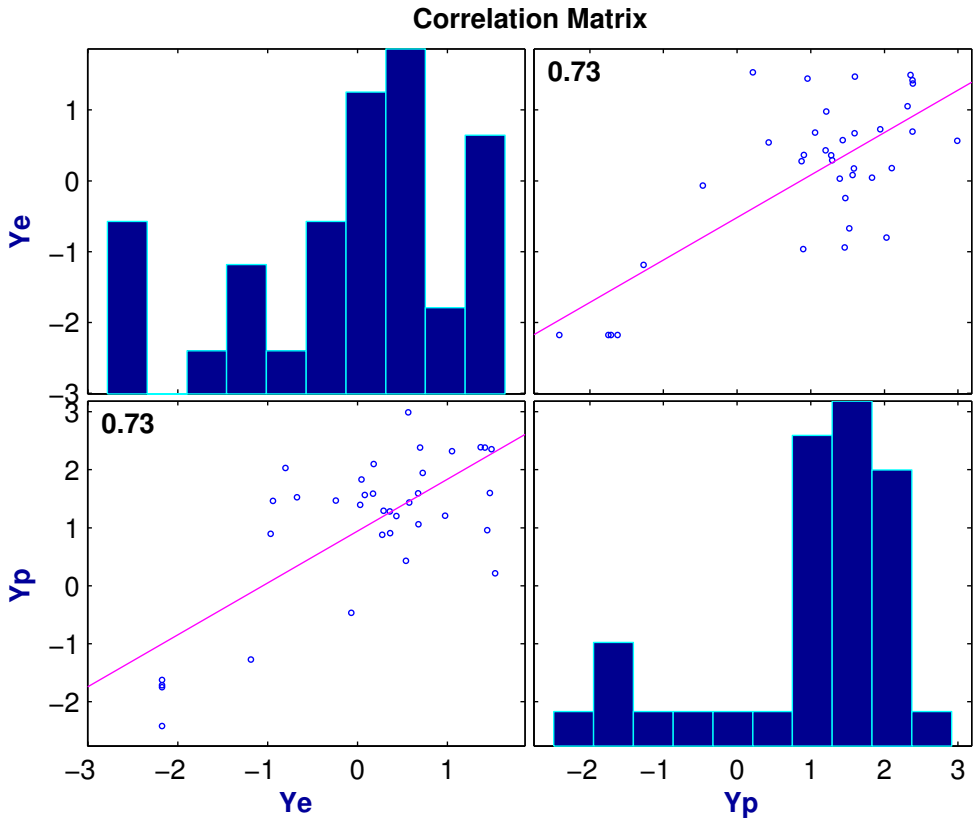


Figure 7.19: Correlation plot of the prediction versus output of 2 drug combination experiment.

Table 7.1: 2 drug combinations applied on breast cancer cell lines.

	Cell Line	Drug 1	Drug 2	Sensitivity	STD
1	KG1	DMOG	Vinblastine	0.048957224	0.027832123
2	KG1	DMOG	17-AAG	0.104813468	0.060189363
3	KG1	Etoposide	Vorinostat	0.125463138	0.020248064
4	KG1	AZD6244	Camptothecin	0.009357624	0.001126763
5	KG1	AZD6244	Vorinostat	0.033632856	0.01663079
6	KG1	Bortezomib	CHIR-99021	0.000645865	6.42E-05
7	K562	Shikonin	Vinblastine	0.462886938	0.035620256
8	K562	NVP-BEZ235	Vorinostat	0.277815199	0.0578719
9	K562	17-AAG	Vinblastine	0.510909812	0.060077231
10	K562	AG-014699	Thapsigargin	0.499621278	0.040552218
11	K562	AZD-0530	Doxorubicin	0.5196122	0.047307594
12	K562	AZD-0530	Cyclopamine	0.04338182	0.006651824

Table 7.2: 2 drug combinations applied on leukemia cell lines.

	Cell Line	Drug 1	Drug 2	Sensitivity	STD
1	KG1	DMOG	Vinblastine	0.048957224	0.027832123
2	KG1	DMOG	17-AAG	0.104813468	0.060189363
3	KG1	Etoposide	Vorinostat	0.125463138	0.020248064
4	KG1	AZD6244	Camptothecin	0.009357624	0.001126763
5	KG1	AZD6244	Vorinostat	0.033632856	0.01663079
6	KG1	Bortezomib	CHIR-99021	0.000645865	6.42E-05
7	K562	Shikonin	Vinblastine	0.462886938	0.035620256
8	K562	NVP-BEZ235	Vorinostat	0.277815199	0.0578719
9	K562	17-AAG	Vinblastine	0.510909812	0.060077231
10	K562	AG-014699	Thapsigargin	0.499621278	0.040552218
11	K562	AZD-0530	Doxorubicin	0.5196122	0.047307594
12	K562	AZD-0530	Cyclopamine	0.04338182	0.006651824

Table 7.3: 2 drug combinations applied on lung cancer cell lines.

	Cell Line	Drug 1	Drug 2	Sensitivity	STD
1	A549	NVP-BEZ235	Thapsigargin	0.11964675	0.024164239
2	A549	Camptothecin	Doxorubicin	0.251544057	0.021385915
3	A549	17-AAG	Etoposide	0.096486438	0.003286385
4	A549	AZD6244	Doxorubicin	0.466886718	0.029986987
5	A549	Bortezomib	NSC-87877	0.104816587	0.019624957
6	A549	Cyclopamine	Nilotinib	0.001926323	0.001479996
7	NCIH522	NVP-BEZ235	Thapsigargin	0.125578921	0.016641796
8	NCIH522	DMOG	Camptothecin	0.038271644	0.014182584
9	NCIH522	17-AAG	Vorinostat	0.035324179	0.003688548
10	NCIH522	AICAR	Camptothecin	0.102978529	0.009278063
11	NCIH522	Imatinib	Thapsigargin	0.467363766	0.059965251
12	NCIH522	Imatinib	Lapatinib	0	0.000312229

APPENDIX A

Dose response of the 2 drug combinations experiment

The 2 drug combination experiment is carried out on a platform of 6 cell lines. Cell lines were cultured individually in the medium specified by ATCC for at least 3 passages before running the dose response. Cells were tripsinized one day before plating drugs and seeded into 96 well plate format by Hamilton STARlet. The wells on the edges of the 96 well plate were added with 100 μl of medium to serve as backgroun control group, while only the internal 60 wells were used for assay purpose. Each well contains 75 μl of medium and 10,000 cells. On the 2nd day, 25 μl of drugs disolved in medium with appropriate concentrations were added into each well. The 2 drugs were mixed in a well of a 48 well plate, and then serial diluted to 9 different concentrations plus one medium only well. The drugs are then transfered to the 96 well plate of cells, with which each raw contains one drug combination with 9 concentrations and 1 blank control well. After 72 hours of incubation, 20 μl of 1 mM resazurin solution were added to all wells in the 96 well plates, and the fluorescent intensity is measured after 4 hours of incubation, with excitation 560 nm and emission 590 nm.

Due to limitation on the experiment capacity, each datapoint is only gathered once, and data is analyzed with bayesian regression to fit to a sigmoidal curve as described in [BCS12]. The resulting curve is shown in figure A.1 and onward, where the dots are experimental data points and the solid lines are the fitted function. The sensitivity can be integrated from the resulting posterior mean curve.

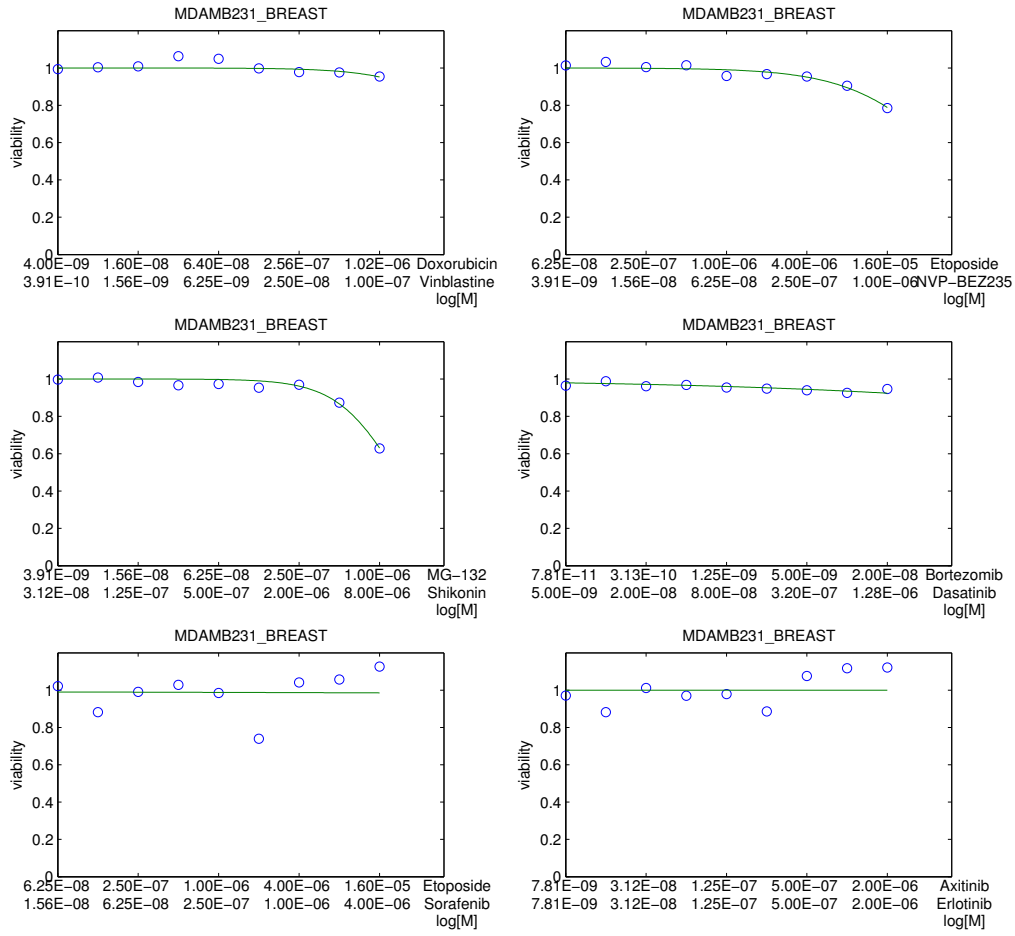


Figure A.1: Dose response of MDA-MB-231.

The noise level of the sensitivity can be determined by sampling the posterior distribution, do integrations on each curve and take the standard deviation of the samples. The results were shown in figure 7.13 and onward.

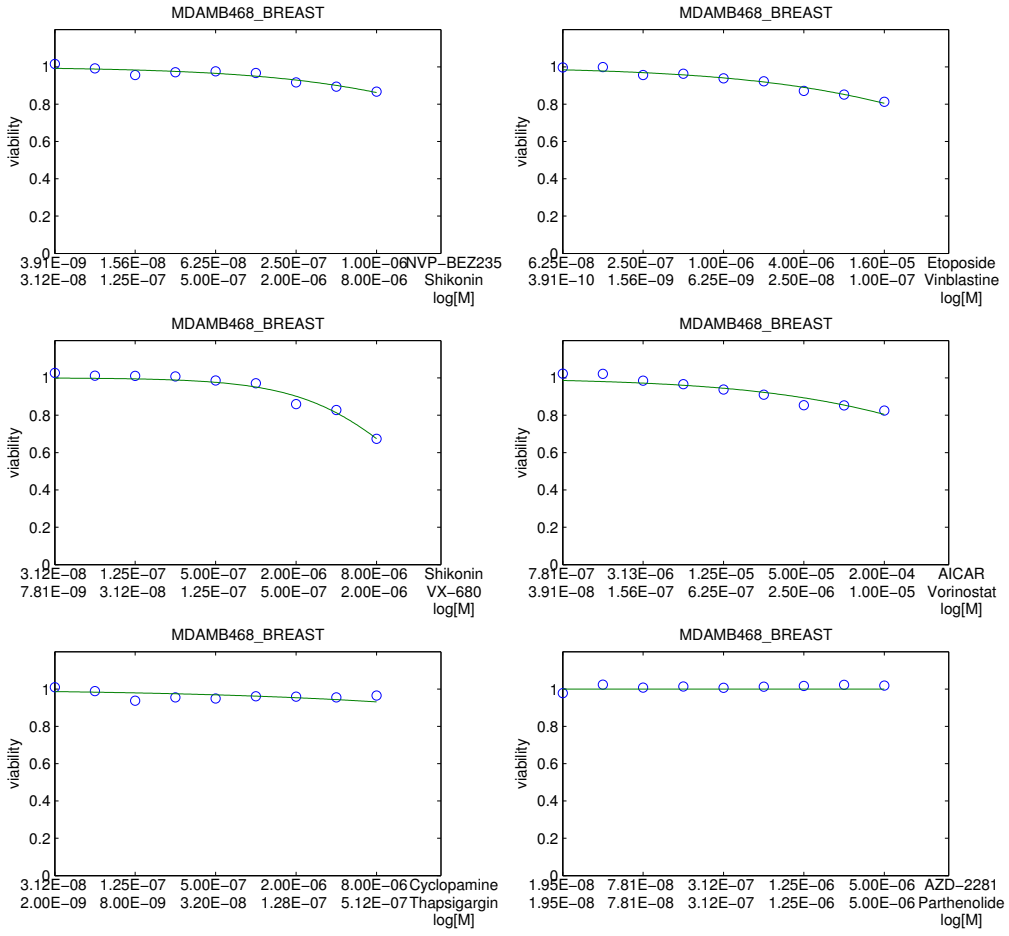


Figure A.2: Dose response of MDA-MB-468.

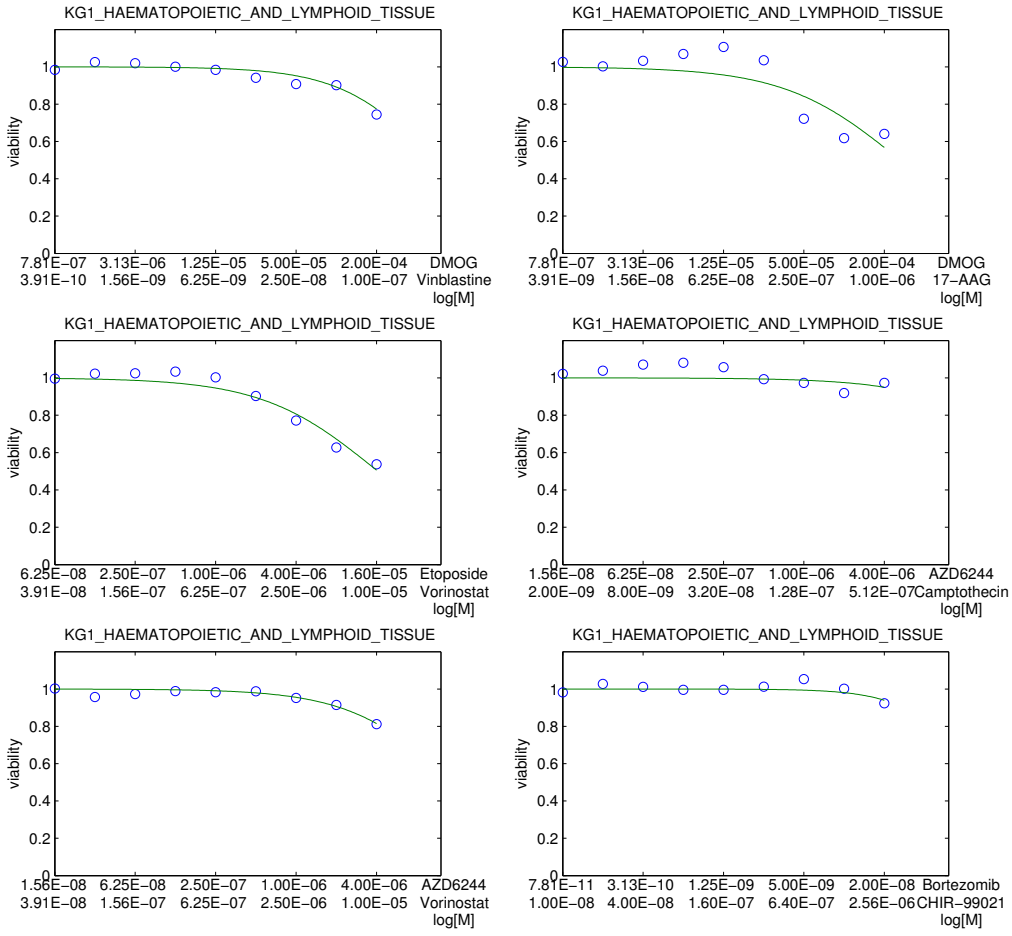


Figure A.3: Dose response of KG1.

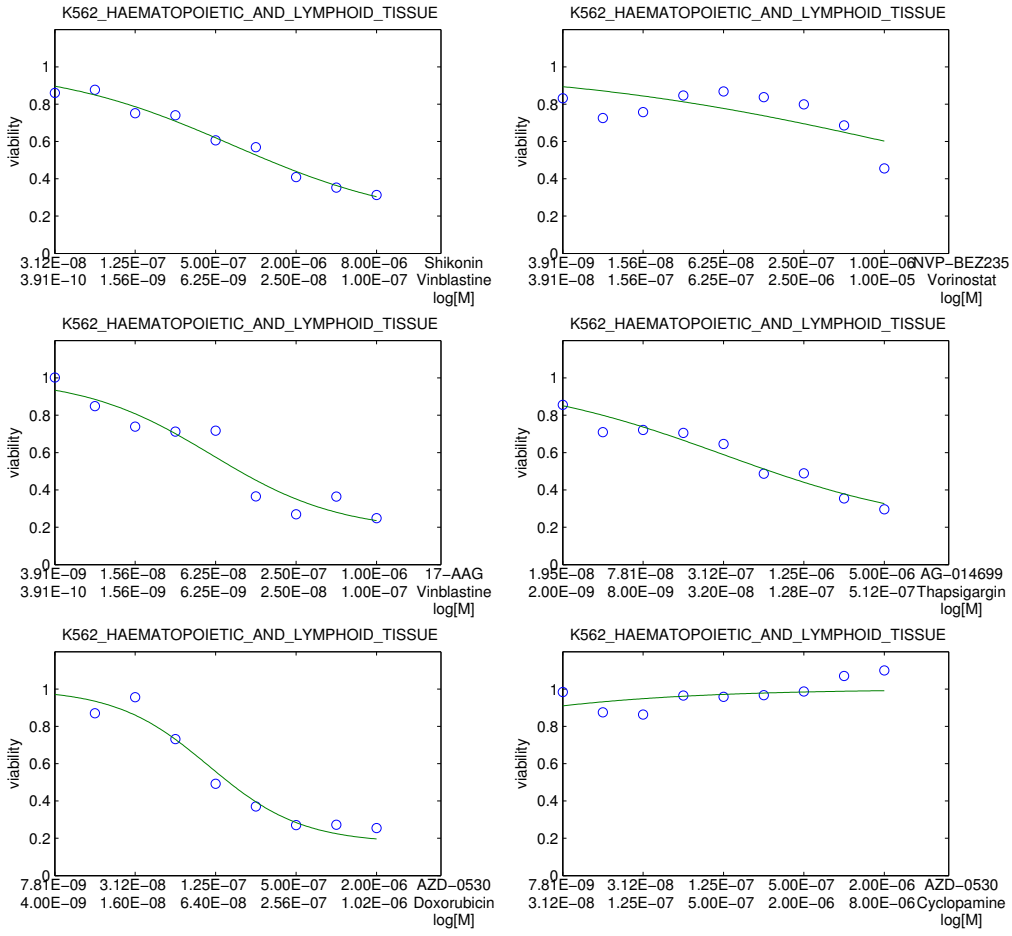


Figure A.4: Dose response of K562.

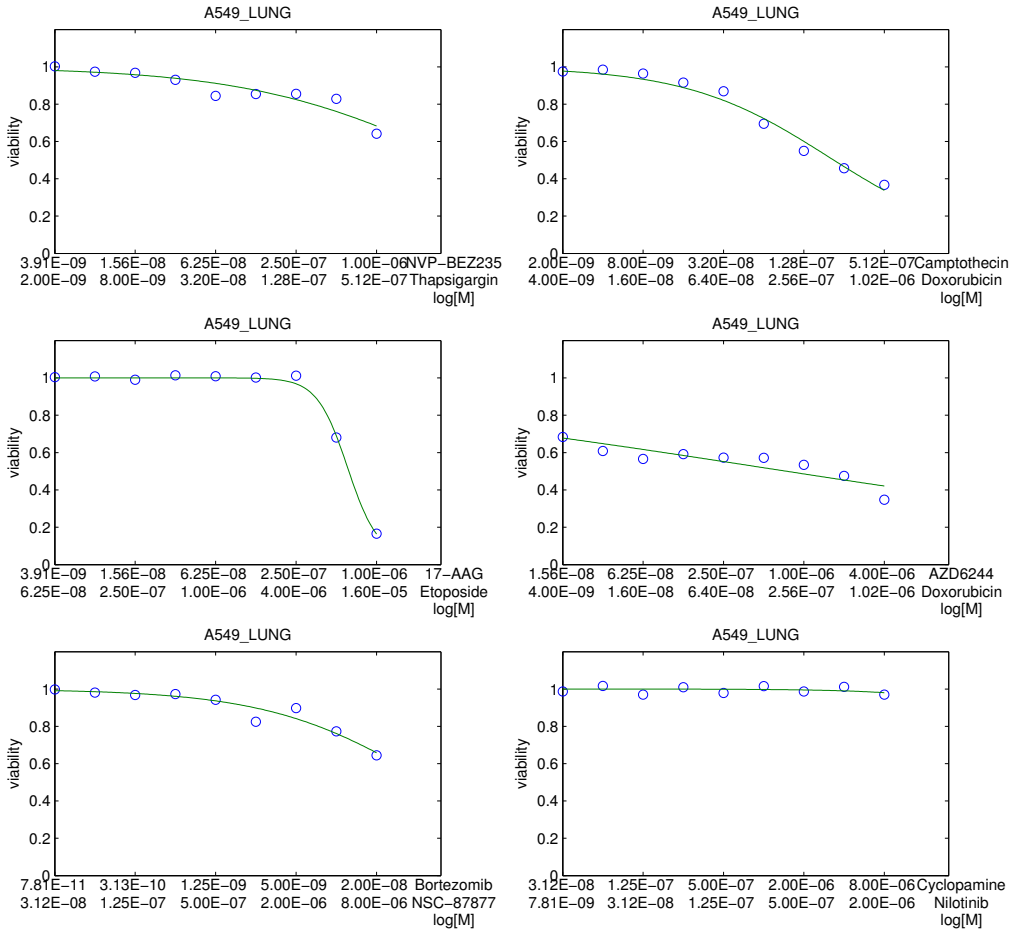


Figure A.5: Dose response of A549.

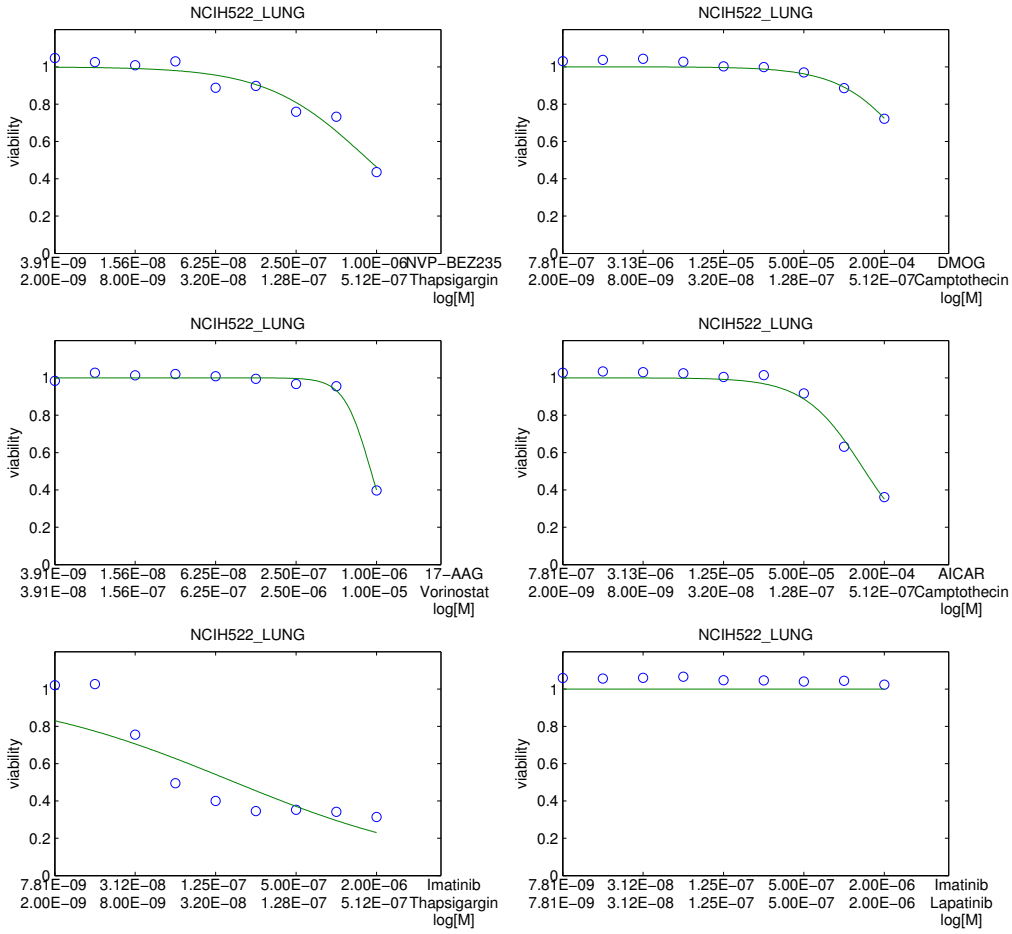


Figure A.6: Dose response of H522.

APPENDIX B

Specifications in Prophecy

Table B.1: List of cell lines trained in Prophecy

Index	Cell Line
1	22RV1_PROSTATE
2	2313287_STOMACH
3	5637_URINARY_TRACT
4	639V_URINARY_TRACT
5	647V_URINARY_TRACT
6	697_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
7	769P_KIDNEY
8	8305C_THYROID
9	8505C_THYROID
10	8MGBA_CENTRAL_NERVOUS_SYSTEM
11	A101D_SKIN
12	A172_CENTRAL_NERVOUS_SYSTEM
13	A204_SOFT_TISSUE
14	A2058_SKIN
15	A253_SALIVARY_GLAND
16	A2780_OVARY
17	A375_SKIN
18	A3KAW_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
19	A498_KIDNEY
20	A4FUK_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
21	A549_LUNG
22	A673_BONE
23	A704_KIDNEY
24	ABC1_LUNG
25	ACHN_KIDNEY
26	AGS_STOMACH
27	AM38_CENTRAL_NERVOUS_SYSTEM
28	AN3CA_ENDOMETRIUM
29	ASPC1_PANCREAS
30	AU565_BREAST
31	BCPAP_THYROID
32	BECKER_CENTRAL_NERVOUS_SYSTEM
33	BEN_LUNG
34	BFTC905_URINARY_TRACT
35	BFTC909_KIDNEY
36	BHY_UPPER_AERODIGESTIVE_TRACT
37	BL41_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
38	BL70_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
39	BT20_BREAST
40	BT474_BREAST
41	BT549_BREAST
42	BV173_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
43	BXPC3_PANCREAS

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
44	C2BBE1_LARGE_INTESTINE
45	C32_SKIN
46	C3A_LIVER
47	CA46_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
48	CAKI1_KIDNEY
49	CAL120_BREAST
50	CAL12T_LUNG
51	CAL148_BREAST
52	CAL27_UPPER_AERODIGESTIVE_TRACT
53	CAL33_UPPER_AERODIGESTIVE_TRACT
54	CAL51_BREAST
55	CAL54_KIDNEY
56	CAL62_THYROID
57	CAL851_BREAST
58	CALU1_LUNG
59	CALU3_LUNG
60	CALU6_LUNG
61	CAMA1_BREAST
62	CAOV3_OVARY
63	CAOV4_OVARY
64	CAPAN1_PANCREAS
65	CAPAN2_PANCREAS
66	CAS1_CENTRAL_NERVOUS_SYSTEM
67	CCFSTTG1_CENTRAL_NERVOUS_SYSTEM
68	CFPAC1_PANCREAS

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
69	CGTHW1_THYROID
70	CHAGOK1_LUNG
71	CHL1_SKIN
72	CHP212_AUTONOMIC_GANGLIA
73	CMK115_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
74	COLO205_LARGE_INTESTINE
75	COLO668_LUNG
76	COLO679_SKIN
77	COLO680N_OESOPHAGUS
78	COLO684_ENDOMETRIUM
79	COLO741_SKIN
80	COLO792_SKIN
81	COLO829_SKIN
82	CORL23_LUNG
83	CORL279_LUNG
84	CORL88_LUNG
85	CPCN_LUNG
86	CW2_LARGE_INTESTINE
87	DAOY_CENTRAL_NERVOUS_SYSTEM
88	DAUDI_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
89	DB_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
90	DBTRG05MG_CENTRAL_NERVOUS_SYSTEM
91	DEL_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
92	DETROIT562_UPPER_AERODIGESTIVE_TRACT
93	DKMG_CENTRAL_NERVOUS_SYSTEM

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
94	DMS114_LUNG
95	DMS153_LUNG
96	DMS273_LUNG
97	DMS53_LUNG
98	DMS79_LUNG
99	DOHH2_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
100	DU145_PROSTATE
101	DU4475_BREAST
102	DV90_LUNG
103	EB2_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
104	ECC10_STOMACH
105	ECC12_STOMACH
106	ECGI10_OESOPHAGUS
107	EFE184_ENDOMETRIUM
108	EFM192A_BREAST
109	EFO21_OVARY
110	EFO27_OVARY
111	EHEB_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
112	EM2_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
113	EPLC272H_LUNG
114	ESS1_ENDOMETRIUM
115	EVSAT_BREAST
116	FADU_UPPER_AERODIGESTIVE_TRACT
117	FTC133_THYROID
118	G361_SKIN

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
119	G401_SOFT_TISSUE
120	G402_SOFT_TISSUE
121	GAMG_CENTRAL_NERVOUS_SYSTEM
122	GCIY_STOMACH
123	GCT_SOFT_TISSUE
124	GDM1_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
125	GI1_CENTRAL_NERVOUS_SYSTEM
126	GMS10_CENTRAL_NERVOUS_SYSTEM
127	H4_CENTRAL_NERVOUS_SYSTEM
128	HCC1143_BREAST
129	HCC1187_BREAST
130	HCC1395_BREAST
131	HCC1419_BREAST
132	HCC1500_BREAST
133	HCC1588_LUNG
134	HCC1806_BREAST
135	HCC1954_BREAST
136	HCC2157_BREAST
137	HCC2218_BREAST
138	HCC38_BREAST
139	HCC70_BREAST
140	HCT116_LARGE_INTESTINE
141	HCT15_LARGE_INTESTINE
142	HDLM2_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
143	HDMYZ_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE

Continued on next page

Table B.1 – *Continued from previous page*

Index	Cell Line
144	HEL9217_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
145	HGC27_STOMACH
146	HH_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
147	HOS_BONE
148	HPAFII_PANCREAS
149	HS578T_BREAST
150	HSC2_UPPER_AERODIGESTIVE_TRACT
151	HSC3_UPPER_AERODIGESTIVE_TRACT
152	HSC4_UPPER_AERODIGESTIVE_TRACT
153	HT_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
154	HT115_LARGE_INTESTINE
155	HT1376_URINARY_TRACT
156	HT144_SKIN
157	HT29_LARGE_INTESTINE
158	HTK_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
159	HUCCT1_BILIARY_TRACT
160	HUH7_LIVER
161	HUPT3_PANCREAS
162	HUPT4_PANCREAS
163	IALM_LUNG
164	IGROV1_OVARY
165	IPC298_SKIN
166	ISTMES1_PLEURA
167	J82_URINARY_TRACT
168	JVM2_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE

Continued on next page

Table B.1 – *Continued from previous page*

Index	Cell Line
169	JVM3_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
170	K562_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
171	KALS1_CENTRAL_NERVOUS_SYSTEM
172	KARPAS299_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
173	KASUMI1_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
174	KE37_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
175	KG1_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
176	KLE_ENDOMETRIUM
177	KM12_LARGE_INTESTINE
178	KMH2_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
179	KNS42_CENTRAL_NERVOUS_SYSTEM
180	KNS62_LUNG
181	KP4_PANCREAS
182	KPNRTBM1_AUTONOMIC_GANGLIA
183	KPNYN_AUTONOMIC_GANGLIA
184	KS1_CENTRAL_NERVOUS_SYSTEM
185	KU1919_URINARY_TRACT
186	KU812_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
187	KURAMOCHI_OVARY
188	KYSE140_OESOPHAGUS
189	KYSE150_OESOPHAGUS
190	KYSE180_OESOPHAGUS
191	KYSE270_OESOPHAGUS
192	KYSE410_OESOPHAGUS
193	KYSE450_OESOPHAGUS

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
194	KYSE510_OESOPHAGUS
195	KYSE520_OESOPHAGUS
196	KYSE70_OESOPHAGUS
197	L363_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
198	L428_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
199	L540_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
200	LC1F_LUNG
201	LCLC103H_LUNG
202	LCLC97TM1_LUNG
203	LK2_LUNG
204	LNCAPCLONEFGC_PROSTATE
205	LOUCY_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
206	LOVO_LARGE_INTESTINE
207	LOXIMVI_SKIN
208	LP1_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
209	LS1034_LARGE_INTESTINE
210	LS123_LARGE_INTESTINE
211	LS411N_LARGE_INTESTINE
212	LS513_LARGE_INTESTINE
213	LU65_LUNG
214	LXF289_LUNG
215	MC116_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
216	MCF7_BREAST
217	MDAMB134VI_BREAST
218	MDAMB231_BREAST

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
219	MDAMB361_BREAST
220	MDAMB415_BREAST
221	MDAMB453_BREAST
222	MDAMB468_BREAST
223	MEG01_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
224	MELHO_SKIN
225	MELJUSO_SKIN
226	MEWO_SKIN
227	MFE280_ENDOMETRIUM
228	MFE296_ENDOMETRIUM
229	MG63_BONE
230	MHHCALL2_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
231	MHHES1_BONE
232	MHHNB11_AUTONOMIC_GANGLIA
233	MIAPACA2_PANCREAS
234	MKN1_STOMACH
235	MKN74_STOMACH
236	MOLT13_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
237	MOLT16_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
238	MONOMAC6_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
239	MPP89_PLEURA
240	MSTO211H_PLEURA
241	MV411_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
242	NALM6_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
243	NB1_AUTONOMIC_GANGLIA

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
244	NCIH1048_LUNG
245	NCIH1092_LUNG
246	NCIH1299_LUNG
247	NCIH1355_LUNG
248	NCIH1436_LUNG
249	NCIH1437_LUNG
250	NCIH1563_LUNG
251	NCIH1573_LUNG
252	NCIH1581_LUNG
253	NCIH1618_LUNG
254	NCIH1623_LUNG
255	NCIH1648_LUNG
256	NCIH1650_LUNG
257	NCIH1651_LUNG
258	NCIH1666_LUNG
259	NCIH1693_LUNG
260	NCIH1694_LUNG
261	NCIH1703_LUNG
262	NCIH1734_LUNG
263	NCIH1755_LUNG
264	NCIH1792_LUNG
265	NCIH1793_LUNG
266	NCIH1838_LUNG
267	NCIH196_LUNG
268	NCIH1975_LUNG

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
269	NCIH2009_LUNG
270	NCIH2029_LUNG
271	NCIH2030_LUNG
272	NCIH2052_PLEURA
273	NCIH2081_LUNG
274	NCIH2087_LUNG
275	NCIH209_LUNG
276	NCIH2122_LUNG
277	NCIH2126_LUNG
278	NCIH2141_LUNG
279	NCIH2170_LUNG
280	NCIH2171_LUNG
281	NCIH2196_LUNG
282	NCIH2227_LUNG
283	NCIH2228_LUNG
284	NCIH226_LUNG
285	NCIH2291_LUNG
286	NCIH23_LUNG
287	NCIH2342_LUNG
288	NCIH2347_LUNG
289	NCIH2405_LUNG
290	NCIH2452_PLEURA
291	NCIH358_LUNG
292	NCIH441_LUNG
293	NCIH446_LUNG

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
294	NCIH460_LUNG
295	NCIH520_LUNG
296	NCIH522_LUNG
297	NCIH524_LUNG
298	NCIH526_LUNG
299	NCIH596_LUNG
300	NCIH650_LUNG
301	NCIH661_LUNG
302	NCIH69_LUNG
303	NCIH716_LARGE_INTESTINE
304	NCIH727_LUNG
305	NCIH747_LARGE_INTESTINE
306	NCIH810_LUNG
307	NCIH82_LUNG
308	NCIH838_LUNG
309	NCIH889_LUNG
310	NCIN87_STOMACH
311	NMCG1_CENTRAL_NERVOUS_SYSTEM
312	NUGC3_STOMACH
313	OAW28_OVARY
314	OAW42_OVARY
315	OCIAML2_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
316	OE19_OESOPHAGUS
317	OE33_OESOPHAGUS
318	ONS76_CENTRAL_NERVOUS_SYSTEM

Continued on next page

Table B.1 – *Continued from previous page*

Index	Cell Line
319	OPM2_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
320	OSRC2_KIDNEY
321	OVCAR4_OVARY
322	OVCAR8_OVARY
323	P12ICHIKAWA_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
324	P31FUJ_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
325	PANC0327_PANCREAS
326	PANC0813_PANCREAS
327	PANC1005_PANCREAS
328	PC14_LUNG
329	PC3_PROSTATE
330	PF382_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
331	PSN1_PANCREAS
332	RAJI_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
333	RCC10RGB_KIDNEY
334	RCM1_LARGE_INTESTINE
335	RD_SOFT_TISSUE
336	REH_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
337	RERFLCMS_LUNG
338	RKO_LARGE_INTESTINE
339	RL_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
340	RL952_ENDOMETRIUM
341	RMGI_OVARY
342	RPMI7951_SKIN
343	RPMI8226_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE

Continued on next page

Table B.1 – *Continued from previous page*

Index	Cell Line
344	RPMI8402_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
345	RS411_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
346	RT11284_URINARY_TRACT
347	RT4_URINARY_TRACT
348	S117_SOFT_TISSUE
349	SAOS2_BONE
350	SBC5_LUNG
351	SCC15_UPPER_AERODIGESTIVE_TRACT
352	SCC25_UPPER_AERODIGESTIVE_TRACT
353	SCC4_UPPER_AERODIGESTIVE_TRACT
354	SCC9_UPPER_AERODIGESTIVE_TRACT
355	SCLC21H_LUNG
356	SF126_CENTRAL_NERVOUS_SYSTEM
357	SF295_CENTRAL_NERVOUS_SYSTEM
358	SH4_SKIN
359	SHP77_LUNG
360	SIGM5_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
361	SIMA_AUTONOMIC_GANGLIA
362	SJRH30_SOFT_TISSUE
363	SJSA1_BONE
364	SKCO1_LARGE_INTESTINE
365	SKHEP1_LIVER
366	SKLMS1_SOFT_TISSUE
367	SKLU1_LUNG
368	SKM1_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
369	SKMEL1_SKIN
370	SKMEL2_SKIN
371	SKMEL24_SKIN
372	SKMEL28_SKIN
373	SKMEL3_SKIN
374	SKMEL30_SKIN
375	SKMEL5_SKIN
376	SKMES1_LUNG
377	SKNAS_AUTONOMIC_GANGLIA
378	SKNDZ_AUTONOMIC_GANGLIA
379	SKNFI_AUTONOMIC_GANGLIA
380	SKOV3_OVARY
381	SKUT1_SOFT_TISSUE
382	SNGM_ENDOMETRIUM
383	SNU387_LIVER
384	SNU423_LIVER
385	SNU449_LIVER
386	SNU475_LIVER
387	ST486_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
388	SUDHL10_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
389	SW1088_CENTRAL_NERVOUS_SYSTEM
390	SW1116_LARGE_INTESTINE
391	SW1417_LARGE_INTESTINE
392	SW1463_LARGE_INTESTINE
393	SW1573_LUNG

Continued on next page

Table B.1 – *Continued from previous page*

Index	**Cell Line**
394	SW1710_URINARY_TRACT
395	SW1783_CENTRAL_NERVOUS_SYSTEM
396	SW1990_PANCREAS
397	SW480_LARGE_INTESTINE
398	SW620_LARGE_INTESTINE
399	SW780_URINARY_TRACT
400	SW837_LARGE_INTESTINE
401	SW900_LUNG
402	SW948_LARGE_INTESTINE
403	T47D_BREAST
404	T84_LARGE_INTESTINE
405	T98G_CENTRAL_NERVOUS_SYSTEM
406	TE1_OESOPHAGUS
407	TE10_OESOPHAGUS
408	TE159T_SOFT_TISSUE
409	TE4_OESOPHAGUS
410	TE5_OESOPHAGUS
411	TE617T_SOFT_TISSUE
412	TE8_OESOPHAGUS
413	TE9_OESOPHAGUS
414	TGBC11TKB_STOMACH
415	THP1_HAEMATOPOIETIC_AND_LYMPHOID_TISSUE
416	TYKNU_OVARY
417	U2OS_BONE
418	U87MG_CENTRAL_NERVOUS_SYSTEM

Continued on next page

Table B.1 – *Continued from previous page*

Index	Cell Line
419	UACC257_SKIN
420	UACC62_SKIN
421	UACC812_BREAST
422	UACC893_BREAST
423	UMUC3_URINARY_TRACT
424	VMCUB1_URINARY_TRACT
425	VMRCRCZ_KIDNEY
426	WM115_SKIN
427	YAPC_PANCREAS
428	YH13_CENTRAL_NERVOUS_SYSTEM
429	YKG1_CENTRAL_NERVOUS_SYSTEM
430	ZR7530_BREAST

Table B.2: List of drugs trained in Prophecy

Index	Drug
1	PD-173074
2	Bicalutamide
3	Embelin
4	Metformin
5	MS-275
6	Pyrimethamine
7	Shikonin
8	Imatinib
9	Vorinostat
10	MitomycinC

Continued on next page

Table B.2 – *Continued from previous page*

Index	**Cell Line**
11	Doxorubicin
12	Paclitaxel
13	Etoposide
14	AICAR
15	S-Trityl-L-cysteine
16	Bexarotene
17	Cisplatin
18	Camptothecin
19	Gefitinib
20	CEP-701
21	Methotrexate
22	BIRB0796
23	Tipifarnib
24	Roscovitine
25	SB216763
26	Erlotinib
27	Lapatinib
28	Sorafenib
29	Lenalidomide
30	Elesclomol
31	Bortezomib
32	Cyclopamine
33	ATRA
34	Thapsigargin
35	EpothiloneB

Continued on next page

Table B.2 – *Continued from previous page*

Index	**Cell Line**
36	MG-132
37	DMOG
38	CGP-60474
39	Nilotinib
40	Dasatinib
41	KU-55933
42	Rapamycin
43	GNF-2
44	Bosutinib
45	Sunitinib
46	PD-0332991
47	Parthenolide
48	NSC-87877
49	Bleomycin
50	VX-680
51	Axitinib
52	17-AAG
53	Vinblastine
54	PAC-1
55	Temsirolimus
56	CI-1040
57	A-770041
58	LFM-A13
59	PD-0325901
60	ZM-447439

Continued on next page

Table B.2 – *Continued from previous page*

Index	**Cell Line**
61	AG-014699
62	CHIR-99021
63	BX-795
64	AUY922
65	Pazopanib
66	AZD6244
67	A-443654
68	BIBW2992
69	AZD-0530
70	VX-702
71	BMS-536924
72	PHA-665752
73	SB590885
74	NU-7441
75	BI-2536
76	Nutlin-3a
77	TW37
78	JNKInhibitorVIII
79	OSI-906
80	AMG-706
81	PF-562271
82	WH-4-023
83	ABT-888
84	NVP-BEZ235
85	NVP-TAE684

Continued on next page

Table B.2 – *Continued from previous page*

Index	**Cell Line**
86	GSK269962A
87	Z-LLNle-CHO
88	GDC0941
89	AZD-2281
90	PLX4720
91	Midostaurin
92	GDC-0449
93	BMS-754807
94	AP-24534
95	MK-2206
96	ABT-263
97	GSK-650394
98	BI-D1870
99	RDEA119
100	RO-3306

References

[ABW12] Bissan Al-Lazikani, Udai Banerji, and Paul Workman. "Combinatorial drug therapy for cancer in the post-genomic era." *Nature biotechnology*, **30**(7):679–692, 2012.

[ACP05] Vilmos Ágoston, Péter Csermely, and Sándor Pongor. "Multiple weak hits confuse complex systems: a transcriptional regulatory network as an example." *Physical Review E*, **71**(5):051909, 2005.

[AGS80] DavidS Alberts, HS George Chen, Barbara Soehnlen, SydneyE Salmon, EarlA Surwit, Laurie Young, and ThomasE Moon. "In-vitro clonogenic assay for predicting response of ovarian cancer to chemotherapy." *The Lancet*, **316**(8190):340–342, 1980.

[AJB00] Réka Albert, Hawoong Jeong, and Albert-László Barabási. "Error and attack tolerance of complex networks." *nature*, **406**(6794):378–382, 2000.

[AYF11] Ibrahim Al-Shyoukh, Fuqu Yu, Jiaying Feng, Karen Yan, Steven Dubinett, Chih-Ming Ho, Jeff S Shamma, and Ren Sun. "Systematic quantitative characterization of cellular responses induced by multiple signals." *BMC systems biology*, **5**(1):88, 2011.

[Azm12] Asfar S Azmi. *Systems biology in cancer research and drug discovery.* Springer, 2012.

[Bak74] Frank B Baker. "Stability of two hierarchical grouping techniques Case I: Sensitivity to data errors." *Journal of the American Statistical Association*, **69**(346):440–445, 1974.

[BCS12] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity." *Nature*, **483**(7391):603–607, 2012.

[BE10] María Berdasco and Manel Esteller. "Aberrant epigenetic landscape in cancer: how cellular identity goes awry." *Developmental cell*, **19**(5):698–711, 2010.

[BEH03] Alexis A Borisy, Peter J Elliott, Nicole W Hurst, Margaret S Lee, Joseph Lehár, E Roydon Price, George Serbedzija, Grant R Zimmermann, Michael A Foley, Brent R Stockwell, et al. "Systematic discovery of multicomponent therapeutics." *Proceedings of the National Academy of Sciences*, **100**(13):7977–7982, 2003.

[BGL11] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. "Network medicine: a network-based approach to human disease." *Nature Reviews Genetics*, **12**(1):56–68, 2011.

[BIB10] Aislyn DW Boran, Ravi Iyengar, AD Boran, R Iyengar, et al. "Systems approaches to polypharmacology and drug discovery." *Current opinion in drug discovery & development*, **13**(3):297, 2010.

[Bis06] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[CAP05] Péter Csermely, Vilmos Agoston, and Sandor Pongor. "The efficiency of multi-target drugs: the network approach might help drug design." *Trends in Pharmacological Sciences*, **26**(4):178–182, 2005.

[CKK13] Peter Csermely, Tamás Korcsmáros, Huba JM Kiss, Gábor London, and Ruth Nussinov. "Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review." *Pharmacology & therapeutics*, **138**(3):333–408, 2013.

[Cse04] Peter Csermely. "Strong links are important, but weak links stabilize them." *Trends in biochemical sciences*, **29**(7):331–334, 2004.

[CSE12] Pau Creixell, Erwin M Schoof, Janine T Erler, and Rune Linding. "Navigating cancer network attractors for tumor-specific therapy." *Nature biotechnology*, **30**(9):842–848, 2012.

[CT84] Ting-Chao Chou and Paul Talalay. "Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors." *Advances in enzyme regulation*, **22**:27–55, 1984.

[DDG02] Andrzej Dzik-Jurasz, Claudia Domenig, Mark George, Jan Wolber, Anwar Padhani, Gina Brown, and Simon Doran. "Diffusion MRI for prediction of response of rectal cancer to chemoradiation." *The Lancet*, **360**(9329):307–308, 2002.

[DF10] Javier De Las Rivas and Celia Fontanillo. "Protein–protein interactions essentials: key concepts to building and analyzing interactome networks." *PLoS computational biology*, **6**(6):e1000807, 2010.

[DGC10] Yun Dai, Monica L Guzman, Shuang Chen, Li Wang, Sin-Kei Yeung, Xin-Yan Pei, Paul Dent, Craig T Jordan, and Steven Grant. "The NF (Nuclear factor)-κB inhibitor parthenolide interacts with histone deacetylase inhibitors to induce MKK7/JNK1-dependent apoptosis in human acute myeloid leukaemia cells." *British journal of haematology*, **151**(1):70–83, 2010.

[DSS11] Xianting Ding, David Jesse Sanchez, Arash Shahangian, Ibrahim Al-Shyoukh, Genhong Cheng, and Chih-Ming Ho. "Cascade search for HSV-1 combinatorial drugs with high antiviral efficacy and low toxicity." *International journal of nanomedicine*, **7**:2281–2292, 2011.

[DXH13] Xianting Ding, Hongquan Xu, Chanelle Hopper, Jian Yang, and Chih-Ming Ho. "Use of fractional factorial designs in antiviral drug studies." *Quality and Reliability Engineering International*, **29**(2):299–304, 2013.

[FCD10] Jacob D Feala, Jorge Cortes, Phillip M Duxbury, Carlo Piermarocchi, Andrew D McCulloch, and Giovanni Paternostro. "Systems approaches and algorithms for discovery of combinatorial therapies." *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, **2**(2):181–193, 2010.

[FK12] Chikara Furusawa, Kunihiko Kaneko, et al. "A dynamical-systems view of stem cell biology." *Science*, **338**(6104):215–217, 2012.

[FM83] Edward B Fowlkes and Colin L Mallows. "A method for comparing two hierarchical clusterings." *Journal of the American statistical association*, **78**(383):553–569, 1983.

[FSF13] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, et al. "STRING v9. 1: protein-protein interaction networks, with increased coverage and integration." *Nucleic acids research*, **41**(D1):D808–D815, 2013.

[FSN06] Jonathan B Fitzgerald, Birgit Schoeberl, Ulrik B Nielsen, and Peter K Sorger. "Systems biology and combination therapy in the quest for clinical efficacy." *Nature chemical biology*, **2**(9):458–466, 2006.

[GBP95] William R Greco, Gregory Bravo, and John C Parsons. "The search for synergy: a critical review from a response surface perspective." *Pharmacological reviews*, **47**(2):331–385, 1995.

[GEH12] Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. "Systematic identification of genomic markers of drug sensitivity in cancer cells." *Nature*, **483**(7391):570–575, 2012.

[GJ07] Matteo Goldoni and Carolina Johansson. "A mathematical approach to study combined effects of toxicants¡ i¿ in vitro¡/i¿: Evaluation of the Bliss independence criterion and the Loewe additivity model." *Toxicology in vitro*, **21**(5):759–769, 2007.

[GLP95] R Grant, BC Liang, MA Page, DL Crane, HS Greenberg, and L Junck. "Age influences chemotherapy response in astrocytomas." *Neurology*, **45**(5):929–933, 1995.

[GOB10] Nils Gehlenborg, Seán I O'Donoghue, Nitin S Baliga, Alexander Goesmann, Matthew a Hibbs, Hiroaki Kitano, Oliver Kohlbacher, Heiko Neuweger, Reinhard Schneider, Dan Tenenbaum, and Anne-Claude Gavin. "Visualization of omics data for systems biology." *Nature methods*, **7**(3 Suppl):S56–68, March 2010.

[Gun10] Jeremy Gunawardena. "Models in systems biology: the parameter problem and the meanings of robustness." *Elements of computational systems biology*, **1**, 2010.

[HA85] Lawrence Hubert and Phipps Arabie. "Comparing partitions." *Journal of classification*, **2**(1):193–218, 1985.

[HDM13] Yoshitomo Honda, Xianting Ding, Federico Mussano, Akira Wiberg, Chih-ming Ho, and Ichiro Nishimura. "Guiding the osteogenic fate of mouse and human mesenchymal stem cells through feedback system control." *Scientific reports*, **3**, 2013.

[HEB05] Sui Huang, Gabriel Eichler, Yaneer Bar-Yam, and Donald E Ingber. "Cell fates as high-dimensional attractor states of a complex gene regulatory network." *Physical review letters*, **94**(12):128701, 2005.

[HEK09] Sui Huang, Ingemar Ernberg, and Stuart Kauffman. "Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective." In *Seminars in cell & developmental biology*, volume 20, pp. 869–876. Elsevier, 2009.

[HGH06] PJ Hesketh, SM Grunberg, J Herrstedt, R de Wit, RJ Gralla, AD Carides, A Taylor, JK Evans, and KJ Horgan. "Combined data from two phase III trials of the NK1 antagonist aprepitant plus a 5HT3 antagonist and a corticosteroid for prevention of chemotherapy-induced nausea and vomiting: effect of gender on treatment response." *Supportive care in cancer*, **14**(4):354–360, 2006.

[HH10] Dean Ho and Chih-Ming Ho. "System control-mediated drug delivery towards complex systems via nanodiamond carriers." *International Journal of Smart and Nano Materials*, **1**(1):69–81, 2010.

[Hop08] Andrew L Hopkins. "Network pharmacology: the next paradigm in drug discovery." *Nature chemical biology*, **4**(11):682–690, 2008.

[HPF08] Samir M Hanash, Sharon J Pitteri, and Vitor M Faca. "Mining the plasma proteome for cancer biomarkers." *Nature*, **452**(7187):571–579, 2008.

[HTF09] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

[Hua12] Sui Huang. "The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology?" *Bioessays*, **34**(2):149–157, 2012.

[HW11] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation." *Cell*, **144**(5):646–674, 2011.

[HZR08] WC Hwang, A Zhang, and M Ramanathan. "Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery." *Clinical Pharmacology & Therapeutics*, **84**(5):563–572, 2008.

[JAG05] Kevin A Janes, John G Albeck, Suzanne Gaudet, Peter K Sorger, Douglas A Lauffenburger, and Michael B Yaffe. "A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis." *Science*, **310**(5754):1646–1653, 2005.

[JB02] Peter A Jones and Stephen B Baylin. "The fundamental role of epigenetic events in cancer." *Nature reviews genetics*, **3**(6):415–428, 2002.

[JBI05] Maliackal Poulo Joy, Amy Brock, Donald E Ingber, and Sui Huang. "High-betweenness proteins in the yeast protein interaction network." *BioMed Research International*, **2005**(2):96–103, 2005.

[JDX13] Jessica Jaynes, Xianting Ding, Hongquan Xu, Weng Kee Wong, and Chih-Ming Ho. "Application of fractional factorial designs to study drug combinations." *Statistics in medicine*, **32**(2):307–318, 2013.

[JKS09] Lars J Jensen, Michael Kuhn, Manuel Stark, Samuel Chaffron, Chris Creevey, Jean Muller, Tobias Doerks, Philippe Julien, Alexander Roth, Milan Simonovic, et al. "STRING 8a global view on proteins and their functional interactions in 630 organisms." *Nucleic acids research*, **37**(suppl 1):D412–D416, 2009.

[JMB01] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. "Lethality and centrality in protein networks." *Nature*, **411**(6833):41–42, 2001.

[JNT06] Toufeng Jin, Hajime Nakatani, Takahiro Taguchi, Hiroshi Sonobe, Norihito Morimoto, Takeki Sugimoto, Toyokazu Akimori, Takumi Nakano, Tsutomu Namikawa, Takehiro Okabayashi, et al. "Thapsigargin enhances cell death in the gastrointestinal stromal tumor cell line, GIST-T1, by treatment with Imatinib (Glivec)." *JOURNAL OF HEALTH SCIENCE-TOKYO-*, **52**(2):110, 2006.

[JPG11] Matthew Jessulat, Sylvain Pitre, Yuan Gui, Mohsen Hooshyar, Katayoun Omidi, Bahram Samanfar, Le Hoa Tan, Md Alamgir, James Green, Frank Dehne, et al. "Recent advances in protein-protein interaction prediction: experimental and computational methods." *Expert opinion on drug discovery*, **6**(9):921–935, 2011.

[JSG05] J Milburn Jessup, Andrew Stewart, Frederick L Greene, and Bruce D Minsky. "Adjuvant chemotherapy for stage III colon cancer: implications of race/ethnicity, age, and differentiation." *Jama*, **294**(21):2703–2711, 2005.

[JTA00] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. "The large-scale organization of metabolic networks." *Nature*, **407**(6804):651–654, 2000.

[JZM09] Jia Jia, Feng Zhu, Xiaohua Ma, Zhiwei W Cao, Yixue X Li, and Yu Zong Chen. "Mechanisms of drug combinations: interaction and network perspectives." *Nature reviews Drug discovery*, **8**(2):111–128, 2009.

[Kae05] William G Kaelin. "The concept of synthetic lethality in the context of anticancer therapy." *Nature reviews cancer*, **5**(9):689–698, 2005.

[KBH08] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. "Walking the interactome for prioritization of candidate disease genes." *The American Journal of Human Genetics*, **82**(4):949–958, 2008.

[KBS05] Curtis T Keith, Alexis A Borisy, and Brent R Stockwell. "Multicomponent therapeutics for networked systems." *Nature Reviews Drug Discovery*, **4**(1):71–78, 2005.

[Kit07a] Hiroaki Kitano. "A robustness-based approach to systems-oriented drug design." *Nature reviews Drug discovery*, **5**(3):202–210, 2007.

[Kit07b] Hiroaki Kitano. "Towards a theory of biological robustness." *Molecular systems biology*, **3**(1), 2007.

[KMC08] Michael Kuhn, Christian von Mering, Monica Campillos, Lars Juhl Jensen, and Peer Bork. "STITCH: interaction networks of chemicals and proteins." *Nucleic acids research*, **36**(suppl 1):D684–D688, 2008.

[KSB07] Tamás Korcsmáros, Máté S Szalay, Csaba Böde, István A Kovács, and Péter Csermely. "How to design multi-target drugs: target search options in cellular networks." 2007.

[KSF10] Michael Kuhn, Damian Szklarczyk, Andrea Franceschini, Monica Campillos, Christian von Mering, Lars Juhl Jensen, Andreas Beyer, and Peer Bork. "STITCH 2: an interaction network database for small molecules and proteins." *Nucleic acids research*, **38**(suppl 1):D552–D556, 2010.

[KSF12] Michael Kuhn, Damian Szklarczyk, Andrea Franceschini, Christian von Mering, Lars Juhl Jensen, and Peer Bork. "STITCH 3: zooming in on protein–chemical interactions." *Nucleic acids research*, **40**(D1):D876–D880, 2012.

[KSP14] Michael Kuhn, Damian Szklarczyk, Sune Pletscher-Frankild, Thomas H Blicher, Christian von Mering, Lars J Jensen, and Peer Bork. "STITCH 4: integration of protein–chemical interactions with user data." *Nucleic acids research*, **42**(D1):D401–D407, 2014.

[KYM08] Hyun-Ah Kim, Cha-Kyong Yom, Byung-In Moon, Kuk-Jin Choe, Sun-Hee Sung, Woon-Sup Han, Hye-Young Choi, Hye-Kyoung Kim, Heung-Kyu Park, Sung-Ho Choi, et al. "The use of an in vitro adenosine triphosphate-based chemotherapy response assay to predict chemotherapeutic response in breast cancer." *The Breast*, **17**(1):19–26, 2008.

[LCP06] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease." *science*, **313**(5795):1929–1935, 2006.

[LHM09] Nicolas Le Novere, Michael Hucka, Huaiyu Mi, Stuart Moodie, Falk Schreiber, Anatoly Sorokin, Emek Demir, Katja Wegner, Mirit I Aladjem, Sarala M Wimalaratne, et al. "The systems biology graphical notation." *Nature biotechnology*, **27**(8):735–741, 2009.

[MJ03] Matthias Mann and Ole N Jensen. "Proteomic analysis of post-translational modifications." *Nature biotechnology*, **21**(3):255–261, 2003.

[Mun09] Bernard Munos. "Lessons from 60 years of pharmaceutical innovation." *Nature Reviews Drug Discovery*, **8**(12):959–968, 2009.

[NK10] Saket Navlakha and Carl Kingsford. "The power of protein interaction networks for associating genes with diseases." *Bioinformatics*, **26**(8):1057–1063, 2010.

[NNC00] Catherine L Nutt, Mark Noble, Ann F Chambers, and J Gregory Cairncross. "Differential expression of drug resistance genes and chemosensitivity in glial cell lineages correlate with differential response of oligo-

dendrogliomas and astrocytomas to chemotherapy." *Cancer research*, **60**(17):4812–4818, 2000.

[NWG11] Nagarjuna Nagaraj, Jacek R Wisniewski, Tamar Geiger, Juergen Cox, Martin Kircher, Janet Kelso, Svante Pääbo, and Matthias Mann. "Deep proteome and transcriptome mapping of a human cancer cell line." *Molecular systems biology*, **7**(1), 2011.

[OAH06] John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. "How many drug targets are there?" *Nature reviews Drug discovery*, **5**(12):993–996, 2006.

[PBM99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. "The PageRank citation ranking: Bringing order to the web." 1999.

[PDB06] Anil Potti, Holly K Dressman, Andrea Bild, Richard F Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J Kelley, Rebecca Petersen, et al. "Genomic signatures to guide the use of chemotherapeutics." *Nature medicine*, **12**(11):1294–1300, 2006.

[PGL05] Savannah C Partridge, Jessica E Gibbs, Ying Lu, Laura J Esserman, Debasish Tripathy, Dulcy S Wolverton, Hope S Rugo, E Shelley Hwang, Cheryl A Ewing, and Nola M Hylton. "MRI measurements of breast tumor volume predict response to neoadjuvant chemotherapy and recurrence-free survival." *American Journal of Roentgenology*, **184**(6):1774–1781, 2005.

[PNV13] Mijung Park, Marcel Nassar, and Haris Vikalo. "Bayesian Active Learning for Drug Combinations." 2013.

[PWS08] Georgios a Pavlopoulos, Anna-Lynn Wegener, and Reinhard Schneider. "A survey of visualization tools for biological network analysis." *BioData mining*, **1**:12, January 2008.

[Ran71] William M Rand. "Objective criteria for the evaluation of clustering methods." *Journal of the American Statistical association*, **66**(336):846–850, 1971.

[Ras06] Carl Edward Rasmussen. "Gaussian processes for machine learning." 2006.

[RHR01] Marylyn D Ritchie, Lance W Hahn, Nady Roodi, L Renee Bailey, William D Dupont, Fritz F Parl, and Jason H Moore. "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer." *The American Journal of Human Genetics*, **69**(1):138–147, 2001.

[SCF09] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. "The cancer genome." *Nature*, **458**(7239):719–724, 2009.

[SDK11] Marina Sirota, Joel T Dudley, Jeewon Kim, Annie P Chiang, Alex A Morgan, Alejandro Sweet-Cordero, Julien Sage, and Atul J Butte. "Discovery and preclinical validation of drug indications using compendia of public gene expression data." *Science translational medicine*, **3**(96):96ra77–96ra77, 2011.

[SFK11] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, et al. "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored." *Nucleic acids research*, **39**(suppl 1):D561–D568, 2011.

[SFS09] Eric E Schadt, Stephen H Friend, and David A Shaywitz. "A network view of disease and compound screening." *Nature reviews Drug discovery*, **8**(4):286–295, 2009.

[SHS10] Sreenath V Sharma, Daniel A Haber, and Jeff Settleman. "Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents." *Nature Reviews Cancer*, **10**(4):241–253, 2010.

[SLB00] Berend Snel, Gerrit Lehmann, Peer Bork, and Martijn A Huynen. "STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene." *Nucleic acids research*, **28**(18):3442–3444, 2000.

[SNT09] Ken Shiozawa, Takeo Nakanishi, Ming Tan, Hong-Bin Fang, Wen-chyi Wang, Martin J Edelman, David Carlton, Ivana Gojo, Edward A Sausville, and Douglas D Ross. "Preclinical studies of vorinostat (suberoylanilide hydroxamic acid) combined with cytosine arabinoside and etoposide for treatment of acute leukemias." *Clinical cancer research*, **15**(5):1698–1707, 2009.

[SUY09] Chien-Pin Sun, Takane Usui, Fuqu Yu, Ibrahim Al-Shyoukh, Jeff Shamma, Ren Sun, and Chih-Ming Ho. "Integrative systems control approach for reactivating Kaposis sarcoma-associated herpesvirus (KSHV) with combinatory drugs." *Integrative Biology*, **1**(1):123–130, 2009.

[SW11] Mihaela E Sardiu and Michael P Washburn. "Building protein-protein interaction networks with proteomics and informatics tools." *Journal of Biological Chemistry*, **286**(27):23645–23651, 2011.

[TBT04] Rebecca J Theilmann, Rebecca Borders, Theodore P Trouard, Guowei Xia, Eric Outwater, James Ranger-Moore, Robert J Gillies, and Alison

Stopeck. "Changes in water mobility measured by diffusion MRI predict response of metastatic breast cancer to chemotherapy." *Neoplasia*, **6**(6):831–837, 2004.

[TVH11] Hideaki Tsutsui, Bahram Valamehr, Antreas Hindoyan, Rong Qiao, Xianting Ding, Shuling Guo, Owen N Witte, Xin Liu, Chih-Ming Ho, and Hong Wu. "An optimized small molecule inhibitor cocktail supports long-term maintenance of human embryonic stem cells." *Nature communications*, **2**:167, 2011.

[Vaz09] Alexei Vazquez. "Optimal drug combinations and minimal hitting sets." *BMC systems biology*, **3**(1):81, 2009.

[VCB11] Marc Vidal, Michael E Cusick, and Albert-Laszlo Barabasi. "Interactome networks and human disease." *Cell*, **144**(6):986–998, 2011.

[VHJ03] Christian Von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. "STRING: a database of predicted functional associations between proteins." *Nucleic acids research*, **31**(1):258–261, 2003.

[VJK07] Christian Von Mering, Lars J Jensen, Michael Kuhn, Samuel Chaffron, Tobias Doerks, Beate Krüger, Berend Snel, and Peer Bork. "STRING 7recent developments in the integration and prediction of protein interactions." *Nucleic acids research*, **35**(suppl 1):D358–D362, 2007.

[VJS05] Christian Von Mering, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A Huynen, and Peer Bork. "STRING: known and predicted protein–protein associations, integrated and transferred across organisms." *Nucleic acids research*, **33**(suppl 1):D433–D437, 2005.

[VK04] Bert Vogelstein and Kenneth W Kinzler. "Cancer genes and the pathways they control." *Nature medicine*, **10**(8):789–799, 2004.

[VMR10] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. "Associating genes and protein complexes with disease via network propagation." *PLoS computational biology*, **6**(1):e1000641, 2010.

[VTH11] Bahram Valamehr, Hideaki Tsutsui, Chih-Ming Ho, and Hong Wu. "Developing defined culture systems for human pluripotent stem cells." *Regenerative medicine*, **6**(5):623–634, 2011.

[WBH11] Fang Wei, Bin Bai, and Chih-Ming Ho. "Rapidly optimizing an aptamer based BoNT sensor by feedback system control (FSC) scheme." *Biosensors and Bioelectronics*, **30**(1):174–179, 2011.

[WOB01] Wolfgang A Weber, Katja Ott, Karen Becker, Hans-Joachim Dittler, Hermann Helmberger, Norbert E Avril, Günther Meisetschläger, Raymonde Busch, Jörg-Rüdiger Siewert, Markus Schwaiger, et al. "Prediction of response to preoperative chemotherapy in adenocarcinomas of the esophagogastric junction by metabolic imaging." *Journal of Clinical Oncology*, **19**(12):3058–3065, 2001.

[WSH13] Hann Wang, Aleidy Silva, and Chih-Ming Ho. "When Medicine Meets EngineeringParadigm Shifts in Diagnostics and Therapeutics." *Diagnostics*, **3**(1):126–154, 2013.

[WXS12] Yin-Ying Wang, Ke-Jia Xu, Jiangning Song, and Xing-Ming Zhao. "Exploring drug combinations in genetic interaction network." *BMC bioinformatics*, **13**(Suppl 7):S7, 2012.

[WYS08] Pak Kin Wong, Fuqu Yu, Arash Shahangian, Genhong Cheng, Ren Sun, and Chih-Ming Ho. "Closed-loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm." *Proceedings of the National Academy of Sciences*, **105**(13):5105–5110, 2008.

[WZC10] Zikai Wu, Xing-Ming Zhao, and Luonan Chen. "A systems biology approach to identify effective cocktail drugs." *BMC systems biology*, **4**(Suppl 2):S7, 2010.

[YAF11] Fuqu Yu, Ibrahim Al-Shyoukh, Jiaying Feng, Xudong Li, Chia Wei Liao, Chih-Ming Ho, Jeff S Shamma, and Ren Sun. "Control of Kaposi's Sarcoma-Associated Herpesvirus Reactivation Induced by Multiple Signals." *PloS one*, **6**(6):e20998, 2011.

[YBO08] Kun Yang, Hongjun Bai, Qi Ouyang, Luhua Lai, and Chao Tang. "Finding multiple target optimal intervention in disease-related molecular network." *Molecular Systems Biology*, **4**(1), 2008.

[YGC07] Muhammed A Yıldırım, Kwang-Il Goh, Michael E Cusick, Albert-László Barabási, and Marc Vidal. "Drugtarget network." *Nature biotechnology*, **25**(10):1119–1126, 2007.

[YKS07] Haiyuan Yu, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. "The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics." *PLoS computational biology*, **3**(4):e59, 2007.

[YSG13] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, et al. "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells." *Nucleic acids research*, **41**(D1):D955–D961, 2013.

[YZD13] Hui Yu, Wendy L Zhang, Xianting Ding, Ken YZ Zheng, Chih-Ming Ho, Karl WK Tsim, and Yi-Kuen Lee. "Optimizing Combinations of Flavonoids Deriving from Astragali Radix in Activating the Regulatory Element of Erythropoietin by a Feedback System Control Scheme." *Evidence-Based Complementary and Alternative Medicine*, **2013**, 2013.

[ZIZ11] Xing-Ming Zhao, Murat Iskar, Georg Zeller, Michael Kuhn, Vera Van Noort, and Peer Bork. "Prediction of drug combinations by integrating molecular and pharmacological data." *PLoS computational biology*, **7**(12):e1002323, 2011.

[ZLK07] Grant R Zimmermann, Joseph Lehar, and Curtis T Keith. "Multi-target therapeutics: when the whole is greater than the sum of the parts." *Drug discovery today*, **12**(1):34–42, 2007.

[ZTM13] Wei Zheng, Natasha Thorne, and John C McKew. "Phenotypic screens as a renewed approach for drug discovery." *Drug discovery today*, **18**(21):1067–1073, 2013.