T.R.

ANTALYA BİLİM UNIVERSITY

INSTITUTE OF POSTGRADUATE EDUCATION

DISSERTATION MASTER'S PROGRAM OF ELECTRICAL AND COMPUTER ENGINEERING

A COMPUTATIONAL APPROACH FOR PRIORITIZATION OF PATIENT-SPECIFIC CANCER DRIVERS

DISSERTATION

Prepared By

Ahmed Amine TALEB BAHMED

181212004

ANTALYA–2020

T.R.

ANTALYA BİLİM UNIVERSITY

INSTITUTE OF POSTGRADUATE EDUCATION

DISSERTATION MASTER'S PROGRAM OF ELECTRICAL AND COMPUTER ENGINEERING

A COMPUTATIONAL APPROACH FOR PRIORITIZATION OF PATIENT-SPECIFIC CANCER DRIVERS

DISSERTATION

Prepared By

Ahmed Amine TALEB BAHMED

181212004

Dissertation Advisors

Assoc. Prof. Dr. Hilal KAZAN

Prof. Dr. Cesim ERTEN

ANTALYA–2020

APPROVAL/NOTIFICATION FORM

ANTALYA BILIM UNIVERSITY
INSTITUTE OF POST-GRADUATE EDUCATION

Ahmed Amine TALEB BAHMED, a M.Sc. student of Antalya Bılım University, Institute of Post Graduate Education, Electrical and Computer Engineering owning student ID 181212004, successfully defended the thesis/dissertation entitled ”A computational approach for prioritization of Patient-Specific Cancer Drivers”, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Academic Tittle, Name-Surname, Signature

Thesis Advisor: Assoc. Prof. Dr. Hilal KAZAN ,.....

Thesis Co-Advisor: Prof. Dr. Cesim ERTEN ,.....

Jury Member: ,.....

Jury Member: ,.....

Jury Member: ,.....

Director of The Institute: ,.....

Date of Submission : ……..

Date of Defence : …….....

"*Sometimes you find out what you are supposed to be doing by doing the things you are not supposed to do.*"

Oprah Winfrey.

ABSTRACT

A COMPUTATIONAL APPROACH FOR PRIORITIZATION OF PATIENT-SPECIFIC CANCER DRIVERS

A major challenge in cancer genomics is to distinguish the driver mutations that are causally linked to cancer from passenger mutations that are neutral and do not contribute to cancer development. The identification of these driver genes could lead to the development of therapies. Numerous methods have been proposed for this problem; however, the majority of these methods provide a single driver gene list for the entire cohort of patients. On the other hand, mutational profiles of cancer patients show a high degree of mutational heterogeneity. As such, because the set of driver genes can be distinct for each patient, a more ideal approach is to identify patient-specific drivers. The results from such an approach can lead to the development of personalized treatments and therapies.

In this thesis, we develop a computational approach that integrates genomic data, biological pathways, and protein connectivity information to identify patient-specific cancer driver genes. We construct a bipartite graph that relates specific mutated genes and various outliers for each specific patient. For each patient, we rank the mutated genes based on a convex combination of two terms. The first term is a weighted scoring of the number of connections to outlier genes of that patient as well as the outlier genes of other patients. The second term incorporates the co-occurrences of a mutated gene and an outlier gene within the same pathway. We compare our method against state-of-the-art patient-specific cancer gene prioritization methods on patients and cell line data for colon, lung, and head-neck cancer. We define novel reference gene sets for evaluation of results obtained from cell line data by utilizing drug sensitivity datasets. Furthermore, we propose and discuss alternative approaches for evaluating the recovery of known cancer drivers when patient-specific drivers are provided. Overall, we show that our method can better recover known and rare cancer genes based on various reference compared to other approaches. Additionally, we demonstrate the importance of pathway coverage in the identification and ranking of driver genes.

Keywords: Driver Genes Prioritization, Patient-Specific, Protein-Protein Interactions Network, Biological Pathways, Cell Lines, Cancer.

DEDICATION AND ACKNOWLEDGMENT

First, I thank God Almighty for everything including the success of finishing my MSc dissertation. Then, I would like to express my sincere gratitude to my Advisors Prof. Dr. Cesim Erten, and Assoc. Prof. Dr. Hilal Kazan for their consistent support and guidance during the running of this study. They continuously encouraged me and were always willing and enthusiastic to assist in any way they could throughout the research project.

Further, many thanks to all participants that took part in the study and enabled this research to be possible, to my labmates Aissa HOUDJEDJ, Ilyes BAALI, Rafsan AHMED, and Yacine MAROUF for their help, constant support and unforgettable time we spent together.

Finally, my greatest appreciation and thanks go to both my parents, brothers, sister, and my fiancée for their unlimited love, supplications, and support throughout my life. I wish I was worthy of your patience, and support.

Contents

INSTITUTE OF POSTGRADUATE EDUCATION ELECTRICAL AND COMPUTER ENGINEERING MASTER OF SCIENCE PROGRAM WITH THESIS

ACADEMIC DECLARATION

I hereby declare that this master's thesis titled "A computational approach for prioritization of Patient-Specific Cancer Drivers" has been written by myself under the academic rules and ethical conduct of the Antalya Bilim University. I also declare that the work attached to this declaration complies with the university requirements and is my work. I also declare that all materials used in this thesis consist of the mentioned resources in the reference list. I verify all these with my honor.

31/08/2020

Ahmed Amine TALEB BAHMED

List of Figures

CHAPTER 1

1. Introduction

1.1. Problem Statement

Cancer is a complex disease caused by somatic mutations that lead to uncontrolled growth, which can cause abnormal proliferation and tumor development.

A major challenge in cancer biology is to identify the driver mutations that are causally linked to cancer since most somatic mutations do not lead to cancer. The identification of these driver genes could lead to the development of therapies. This problem is called driver mutation (gene) identification.

Several methods have been proposed for this problem; however, the majority of these methods provide a single driver gene list for a cohort of patients. On the other hand, mutational profiles of cancer patients show a high degree of mutational heterogeneity. Considering the mutational heterogeneity problem and also in terms of personalized medicine and therapies, a more ideal approach is to identify patient-specific drivers.

Our proposed method integrates multi-omic datasets to identify patient-specific drivers and has several contributions in terms of the datasets utilized, the integration with network and pathway information.

Genes identified by our method will provide insight into cancer initiation mechanisms and will serve as potential new drug targets.

1.2. Thesis Organization

The thesis is organized as follows. First, in Chapter 2, we review relevant facts about cancer, types of somatic mutations, and experimental methods for studying it. We also review the state-of-the-art methods for discovering patient-specific cancer driver mutations. Chapter 3 presents the algorithmic and mathematical background behind our proposed method. Chapter 4 presents the results of using our model on different data from TCGA and CCLE, compared to established state-of-the-art methods. Finally, in Chapter 5, we summarize the thesis.

CHAPTER 2

2. Background

2.1. Biological Background

2.1.1. Genome Variation

All body cells come from the mitotic cell division of the first fertilized egg. During the division process, some alterations occur in DNA bases compared to the first fertilized egg.

There are two types of mutations, somatic and germline [1]. Germline mutations are those carried to the children by the parents, they occur in sperm and egg cells. These mutations are fairly responsible for the discrepancy of the human population. The other type of alteration is somatic. These are not passed to the next generations but still have an effect on the organism on which they occur. Some of these mutations may lead to cancer.

There are classes of genetic sequence alterations, from a single base in a genome to entire chromosome arms, Single Nucleotide Variants (SNVs), INDELs (INsertions and DELetions) and Copy Number Variants (CNVs) are part of them. Single Nucleotide Variations consist of a single nucleotide at an exact location in the genome. INDELs are the insertions or deletions of genome bases. Copy Number Variations are a

special type of structural variant (move, copy, or delete entire regions of bases to whole chromosome arms of DNA). Since a healthy human cell has two copies of each chromosome (diploid) its copy number of a region of DNA is two. CNVs occur when a region of DNA is either amplified or deleted.

For this thesis, we will use the output of the SNV and INDELs mutation calling algorithms as the input data to our methods.

2.1.2. Cancer

Most cancers are mutation-driven. Human cells grow and divide to form new cells, which are needed. New cells are supposed to take the place of old and damaged ones. However, cancer occurs when abnormal cells, old or damaged cells survive and divide indefinitely when they should die.

Tumors are created by these abnormal cells that can divide endlessly, and disseminate into nearby tissues. A malignant tumor, can spread to nearby tissues, or even travel to distant organs in the body, where it forms new tumors far from the original organ, through the blood or lymph system. It is a complex and heterogeneous disease to which all body organs can be affected.

2.1.3. Driver and Passenger Mutations in Cancer

Considering the fact that only a subset of somatic mutations leads to cancer, the question that arises is: Which specific mutation or mutations are implicated in the contribution of tumor development?

Most somatic mutations within the cancer cell are passenger mutations that are not directly involved in tumor development [1], [2]. Thus, it is not clear whether a given mutation in a patient's tumor is a driver mutation or a passenger.

Driver mutations disturb normal cell control of proliferation, differentiation, and death. They provide survival and growth advantage, leading to clonal proliferation of these mutated cancerous cells. However, passenger mutations, which are the majority of

the somatic mutations, do not provide any of the perks of the driver mutations.

Cancer genomes can carry up to thousands of mutations that include both driver mutations and passenger mutations [3]. Two different patients can have completely different sets of driver mutations, even though they have the same cancer type.

Thus the identification of patient-specific driver genes is difficult due to the large number of passenger mutations that coexist in the same cancer genome. The identification of driver genes is critical for understanding how cancer develops and for developing personalized therapies.

2.1.4. Protein-protein Interactions

In all biologic processes in a living organism in vivo, we find Protein-Protein Interaction (PPI).

Protein-protein interactions contribute to cellular functions and biological processes in all organisms [4]: structural proteins need to interact in order to shape organelles, molecular machines hold together by protein-protein interactions. Numerous computational techniques and physiochemical experiments have been employed to detect PPIs. However, these methods are computationally expensive and time-consuming. Protein-protein interactions are defined as physical contacts that occur in a cell between proteins. The network consists of multiple nodes (proteins) where edges correspond to the interactions between these proteins. It is estimated that the current human PPI catalog cover around 25% of all possible interactions [5].

2.1.5. Biological Pathways

For our body to stay alive and develop, all of its elements from organs to cells to genes must work in coordination. Cells contain thousands of molecules, mostly proteins that work together to accomplish missions like responding to the environment around the body and produce energy. These cells are endlessly receiving chemical cues from both inside and outside the body, such as infections. Biological pathways send and receive signals by the cells as a response to these stimuli. To accomplish their assigned

tasks, the proteins (components of the cells) that make up the biological pathways interact with each other (protein-protein interaction), as well as with the signals.

A biological pathway is a sequence of interactions between molecules in a cell. Pathways control the flow of information, energy and biochemical compound in the cell and the ability of the cell to change its behavior in response to stimuli. It can trigger the production of a new molecule, like fat and protein, alarm the presence of a hormonal signal or even motivate a cell to move.

The biological pathways can take actions over short or long distances. Like the previous example, an infection received after a scratch on an arm, some cells send signals to nearby cells to attack stranger organisms, and repair the localized damage. Other cells produce substances that travel through the blood to distant cells, like the epinephrine, when someone is on a flight, to regulate his/her heart rate and metabolism.

Biological pathways do not always function faultlessly. Most diseases such as cancer, or diabetes are caused by these faulty pathways, which the body couldn't correct. There are various types of biological pathways, the most common are involved in metabolism, gene expression, gene regulation, and signal transduction.

Metabolic pathways are responsible for the chemical reactions that happen in the cells to endure life. These reactions help the cells to develop, reproduce, and keep their structure. As transforming food into energy molecules that keep the living organism alive and healthy.

Gene-regulation pathways turn genes on and off. It is indispensable because genes are in charge of protein production, which are the keystone needed to carry out nearly every task in our bodies. Our body is made up of proteins, helps it to move, and defend itself against the strange organisms.

2.2. Computational Background

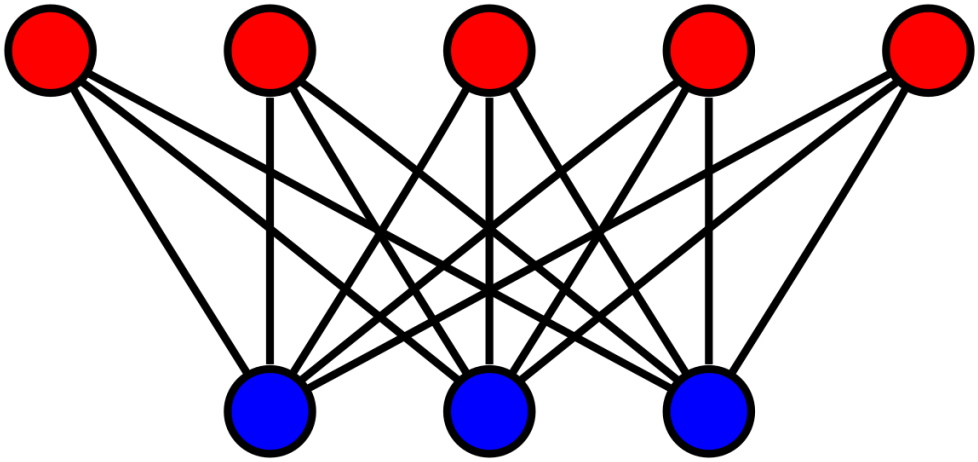

Figure 2.1: A representation of a bipartite graph

2.2.1. Bipartite Graph

A bipartite graph is a graph whose vertices can be divided into two disjoint and independent sets $\boldsymbol{U}$ and $\boldsymbol{V}$ such that every edge connects a vertex in U to one in V. A bipartite graph is a special case of a k-partite graph with k = 2. Equivalently, a cyclic graph is bipartite if and only if all its cycles are of even length [6].

The Figure 2.1 shows a bipartite graph, with vertices in each graph colored based on the disjoint set they belong to.

2.2.2. Cohort Level Methods for Driver Gene Analysis

2.2.2.1. DriverNet

DriverNet [7] is one of the first cohort-level driver gene detection methods to combine both genomic aberrations and gene expression. DriverNet discovers driver genes by evaluating their effect on gene expression. This algorithm uses a mutation data matrix, a gene expression data matrix, and an influence graph.

The binary gene mutation data matrix $\boldsymbol{M}$ contains genes in rows and patients in columns. And $M(i,j) = 1$ if gene i is mutated in patient j. The gene expression

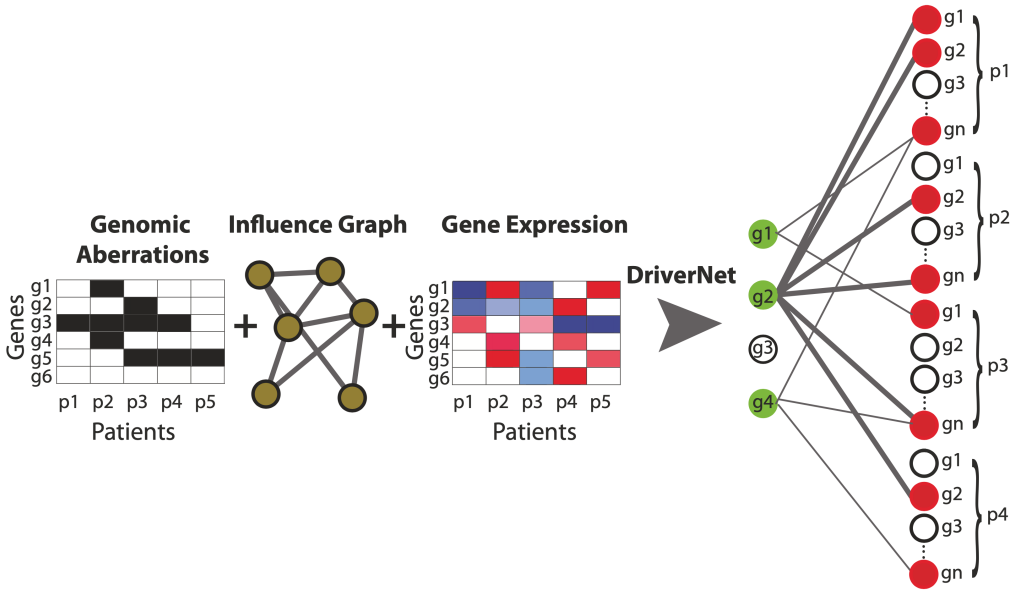

Figure 2.2: Schematic representation of the DriverNet approach. Green nodes on the left partition of the bipartite graph correspond to aberrated genes and nodes on the right represent the outlying expression status for each patient where red indicates outlying patient-gene events from the gene expression matrix.

data matrix ***G*** contains genes in rows and patients in columns like ***M***, and *G(i,j)* have the expression value of gene *i* in patient *j*. The differentially expressed genes matrix is derived from *G*, as a binary matrix ***G'(i,j)***, where *G'(i,j) = 1*, if gene i is over-expressed or under-expressed in patient *j* compared to the normal population. The influence graph is a square adjacency matrix ***I***, where *I(i,j)* = 1, if *i* has a direct edge to *j* in the network. The influence graph is a mixture of multiple preexisting biological knowledge, including protein-protein interaction, gene coexpression and others.

The algorithm is formulated in a bigraph where the nodes on the left represent the genomic aberrations *M* (represented in the figure as green nodes) and nodes on the right (outliers) are multiple sets of differentially expressed genes from *G'*, where each set represents a patient(represented as red nodes). Edge is drawn if for each patient P_k, a gene *i* in the left partition is mutated in P_k, and is known to have interaction with a gene *j* in the right (*I(i,j) = 1)* that is expressed abnormally in P_k.

DriverNet tries to find mutations in the left part that are highly connected to the outliers nodes in the right. DriverNet uses a greedy algorithm to solve the optimization problem. It chooses the mutation with the highest degree and removes it with the

outliers connected to it, and the edges between them, repeatedly. If multiple mutations have the same highest degree, one is chosen randomly. It stops when there are no mutations or outliers left to cover.

2.2.3. Personalized Methods for Driver Gene Analysis

2.2.3.1. DawnRank

DawnRank [8] considers mutated genes with higher connectivity in the gene regulatory network as more impactful, using a random walk approach where Google's PageRank [9] is applied to the PPI. Using each patient's gene expression data but the same gene network for all patients, potential cancer driver genes are ranked based on their impact on the perturbation of downstream genes in a large directed molecular interaction network.

DawnRank uses a directed graph represented in a binary adjacency matrix, in addition to a vector which contains the absolute differential expression. The rank of each gene is related to its in-degree (the incoming edges to the gene), and the damping factor of the gene $0 \leq d_j \leq 1$, d is a parameter representing the extent to which the ranking depends on the structure of the graph (the higher the d_i the higher dependency on the graph). To handle the zero-one gap problem, a dynamic damping factor was used, where each gene has its own damping factor. The damping factor slowly increases as the number of incoming edges increases to include more connectivity information into the ranking of the gene.

At convergence, the algorithm stops when the magnitude of the difference of the ranks between time $t + 1$ and the previous time point t falls below a small ϵ set to $\epsilon = 0.001$, or when no solution is present after 100 iterations.

2.2.3.2. SCS

Single-sample Controller Strategy (SCS) [10] is the first to use network control theory for personalized driver mutation discovery. It combines mutation data and expression data into a protein-protein interaction network for each patient, to attain the driver mutations in a personalized-sample manner. The idea in using network control theory is to detect the correct subset of network nodes (the mutated genes) which control the transition from the normal state to the disease state, which is presented by the differential expressed genes, based on individual omics data. Single-sample Controller Strategy tries to find the minimal set of mutated genes to control the maximal coverage of individual DEGs in a large directed PPI network. SCS uses binary matrices of SNV and CNV, mRNA profiles from paired normal tumor samples, for DEGs analysis, and a PPI network. SCS applies random Markov sampling on a bipartite graph constructed by control paths to identify driver genes and the corresponding paths.

SCS first gets the log2 fold-change of gene expression between the paired samples (tumor and normal). Every gene i for which $|log_2 FC(i)| > 1$ is considered as DEG for each patient, and is assigned +/-1 according to fold change direction. Using Random Walker with Restarter algorithm (RWR) to extract the mutation genes and their interactions from each patient. Which means calculate the probability to reach each gene from the individual mutations of each sample. The personalized network is then constructed using the individual mutation genes, alongside with the individual DEGs.

Paths are discovered using a bipartite graph, where there is a set of target genes (DEGs) connected by directed edges to their inbound neighbors. A bipartite graph is constructed using the paths obtained previously, containing the set of mutation genes, and the set of DEGs, edges are made based on the previous paths found. Weights are assigned to the consensus models according to the frequency of its edges, then genes are ranked according to their total weights of their consensus models.

2.2.3.3. Prodigy

The latest method that addressed the problem, PRODIGY (Personalized Ranking Of DrIver Genes [11], is a method that quantifies the impact of each mutated gene on every deregulated pathway using the prize collecting Steiner tree model (**PCST**).

Drivers are ranked by their aggregated impact on all deregulated pathways.

Prodigy evaluates the expression and mutation profiles of the patient along with data on known pathways and global PPI network. The algorithm assumes that driver mutations influence the deregulation of other genes in affected pathways. Exceptionally the drivers will have a connectedness to these pathways, and the method is designed to score these connections using a variation of the prize collecting Steiner tree problem.

Prodigy uses a binary matrix of SNVs/INDELs and mRNA expression profiles from healthy and tumor tissue samples. Alongside with two types of undirected interaction network, a global PPI network STRING (only highly reliable interactions were selected, those having a confidence score > 0.7), and a collection of pathways (Reactome, KEGG, NCI PID).

Prodigy first computes the differential expression genes (DEGs), genes with log2 FC(g) > 2. Prodigy uses **PCST** to score the influence of a mutation on a deregulated pathway, so the goal is to find the subtree that maximizes the sum of the weights of the nodes minus the edges cost. Prodigy uses pathways assuming that the influence of driver genes is disseminated along pathways and is manifested by DEGs. Every DEGs that belongs to the pathway has a positive score (**prize**) depending on its log_2 fold change value and every other node serving as intermediate nodes in the Steiner tree has a negative score (**penalty**) depending on its degree. In addition, edge weights are considered as penalties reflecting the PPI interaction reliability.

CHAPTER 3

3. Materials and Methods

Given the mutation data and gene expression profile from a cancer cohort, we aim to discover and rank the responsible driver genes in each individual. Our proposed method relies on a bipartite graph using multi-omic datasets, and multiple pathways, assuming that driver genes are disseminated along pathways.

3.1. Input Data

We perform evaluations on three cancer types: COAD, LUAD, and HNSC. Our evaluations utilize both patient data from The Cancer Genome Atlas (TCGA) project [12] and cell line data from Cancer Cell Line Encyclopedia (CCLE) [13]. We focus on these three cancer types in particular as they have the maximum number of cell lines in the CCLE project. The datasets we compile include 279 patients and 46 cell lines for COAD; 505 patients and 35 cell lines for LUAD; and 498 patients and 60 cell lines for HNSC. Datasets for TCGA patients are downloaded from TCGAbiolinks [14] and datasets for cell lines are downloaded from depmap.org [15]. The collection of pathways is retrieved from Prodigy's supplementary data, which contains 285 pathways from KEGG database [16].

For PPI, two networks are used : (i) a global PPI network taken from STRING [17] where only experimentally validated physical interactions with confidence score > 0.7 are included, that consists of 11,302 nodes, and 273,210 edges (ii) and a second PPI network constructed by SCS, which consists of 11,648 genes and 211,794 edges.

3.2. Data Preparation

We assume that the expression of a gene across the patients is distributed as a normal distribution. We deem the gene of interest to be an outlier gene for those patients where the expression value of the gene is outside 2 stds from the mean.

Among the evaluated methods, only SCS requires expression data from paired normal and tumor patients. To evaluate all the employed methods on a larger dataset, we calculate differentially expressed genes for SCS from unpaired gene expression data using Prodigy's method of identifying differentially expressed genes.

3.3. The Algorithm

Our model extends DriverNet's bipartite graph construction such that a personalized bipartite graph is created for each patient.

Let $P = \{P_1, P_2, \ldots, P_r\}$ denote the set of functional pathways. Let $G = (V, E)$ denote the PPI network, where V denotes the set of nodes corresponding to the proteins, and E denotes the set of edges corresponding to pairwise protein interactions. Let $S = \{S^1, \ldots S^n\}$ denote the set of patients (patients). If a gene u is mutated in sample S^i, we create an instance of u denoted with u_m^i and denote the set of all such instances with M^i. Similarly if a gene v is an outlier for sample S_i, we create an instance of v denoted with v_o^i and denote the set of all such instances with O^i.

3.3.1. Graph Construction

We construct a bipartite graph B^i with the edge set E^i, for each sample S^i. B^i has two node partitions, the first of which is M^i. The second partition is $O^1 \cup O^2 \cup \ldots \cup O^n$. Note that a gene may appear multiple times in the second partition as instances of different outlier sets. For $u_m^i \in M^i, v_o^j \in O^j$, there exists an edge $(u_m^i, v_o^j) \in E^i$, if $(u, v) \in E$ and $u_m^j \in M^j$. In other words, an instance of a gene u mutated in sample S_i has a bipartite edge with an instance of a gene v determined to be an outlier in some sample S_j, if they interact in the PPI network and u is a mutated gene in S_j also.

A schematic view of the graphs is given in the Figure 3.1

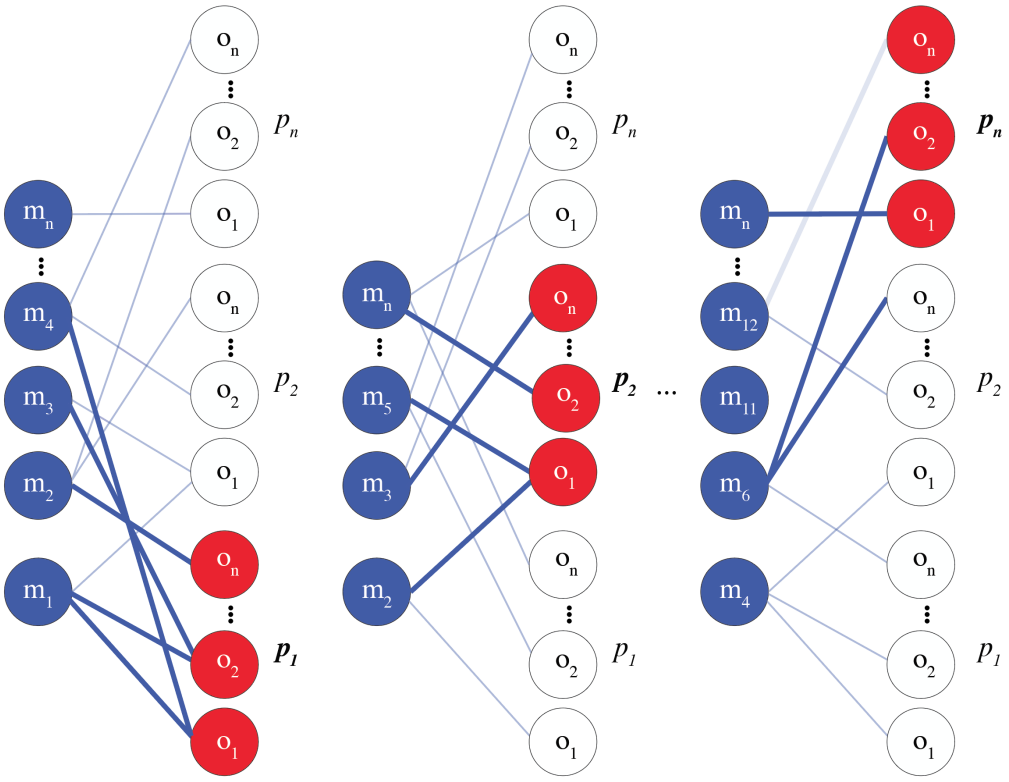

Figure 3.1: Patient-specific graphs

For each gene u mutated in S^i, we define a weight $w^i(u)$ based on the bipartite graph B^i and the set of pathways P. It is a convex combination of the *personalized score (ps)* of u, denoted with $ps^i(u)$ and the *pathway coverage score (pcs)* of u, denoted with $pcs(u)$. Thus $w^i(u) = \alpha_1 \times ps^i(u) + (1 - \alpha_1) \times pcs(u)$. The pathway coverage score $pcs(u)$ is simply the number of pathways $P_k \in P$ where $u \in P_k$.

The personalized score $ps^i(u)$ on the other hand, makes use of the personalized information of S^i itself, as well as the personalized information available from all other patients. The contribution of the former is denoted with $ps^i_{own}(u)$ and that of the latter with $ps^i_{other}(u)$. The personalized score then is a convex combination of these two scores, that is

$$\alpha_2 \times ps^i_{own}(u) + (1 - \alpha_2) \times ps^i_{other}(u).$$

To define both terms we use a combination of the relevant degrees in the bipartite graph B^i and the number of pathways common to a relevant pair of genes. Let $deg_{own}(u^i_m)$ be the number of edges $(u^i_m, v^i_o) \in E^i$. For a pair of nodes u, v, let $cp(u, v)$ denote the number of pathways $P_k \in P$ that includes both u and v. We define $cpcs_{own}(u^i_m)$,

the *common pathway coverage score (cpcs)* of u_m^i with respect to its own set of outliers O^i, as

$$\sum_{\forall v|(u_m^i,v_o^i)\in E^i} cp(u,v).$$

The score $ps_{own}^i(u)$ is then defined as $deg_{own}(u_m^i) + cpcs_{own}(u_m^i)$ normalized with max_{own}, which is the maximum value of $deg_{own}(x_m^i) + cpcs_{own}(x_m^i)$ obtained by any $x_m^i \in M^i$. Analogously, for the contribution of the information gathered from other patients in the cohort, let $avgdeg_{other}(u_m^i)$ be the number of edges $(u_m^i, v_o^j) \in E^i$ for $i \neq j$, averaged over n_{other} which denotes the number of other outlier sets inducing at least one edge with u_m^i in B_i, that is $|\{O^j | i \neq j \wedge \exists v_o^j((u_m^i, v_o^j) \in E^i)\}|$. Let $avgcpcs_{other}(u_m^i)$, the common pathway coverage score of u_m^i with respect to the other sets of outliers O^j for $i \neq j$ be defined as,

$$\frac{\sum_{\forall v|(u_m^i,v_o^j)\in E^i \wedge i\neq j} cp(u,v)}{n_{other}}.$$

The score $ps_{other}^i(u)$ is then defined as $avgdeg_{other}(u_m^i) + avgcpcs_{other}(u_m^i)$ normalized with max_{other}, which is the maximum value of $avgdeg_{other}(x_m^i) + avgcpcs_{other}(x_m^i)$ obtained by any $x_m^i \in M^i$.

A schematic view of the algorithm is given in Figure 3.2.

3.3.2. Prioritization of Drivers

Employing the weight assignments of the previous subsection we rank the genes in M^i in an adaptive manner. Gene u with the largest weight $w^i(u)$ is assigned the next available rank. Node u_m^i and all its nodes incident to it in B^i are removed from B^i and the weight assignment step is repeated with the new bipartite graph. This process of selecting the largest weight node followed by necessary node removals and reassignment of node weights is repeated until M^i becomes empty.

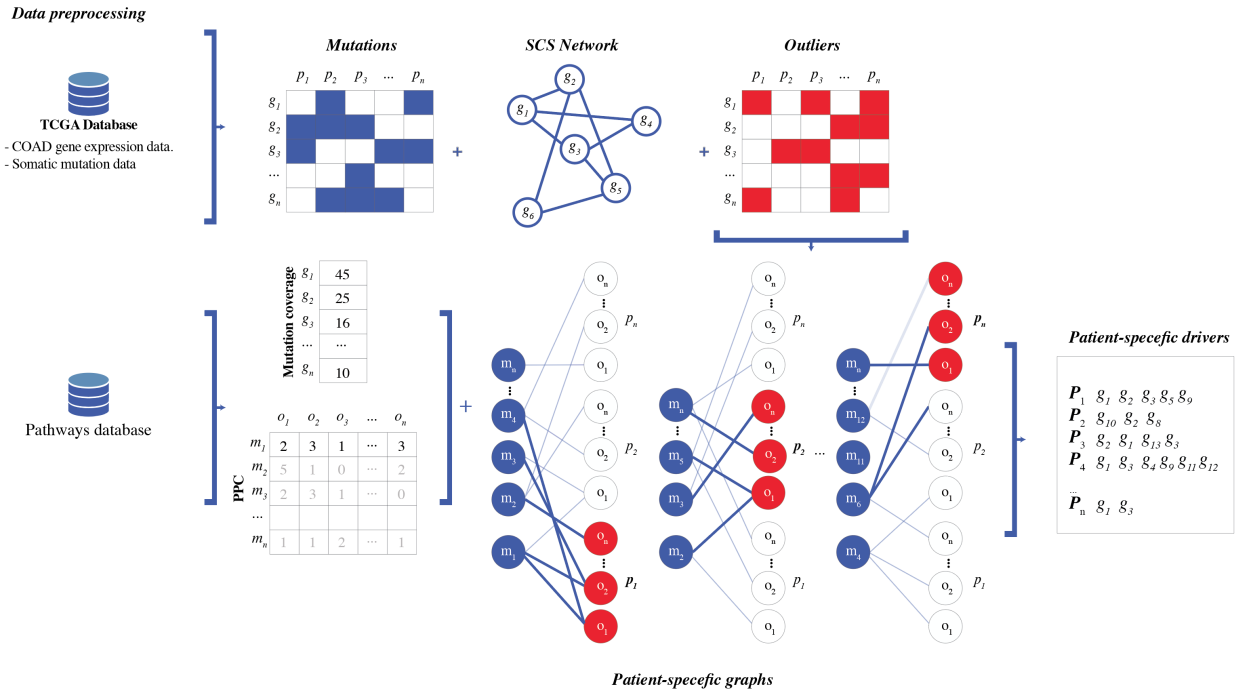

Figure 3.2: Patient-specific method Approach.

Algorithm 1 ps(u) Calculations

for $u_m^i \in M^i$ **do**
    $ps^i_{own}(u^i_m) \leftarrow deg_{own}(u^i_m) + cpcs_{own}(u^i_m)$
    $ps^i_{other}(u^i_m) \leftarrow avgdeg_{other}(u^i_m) + avgcpcs_{other}(u^i_m)$
return $ps^i_{own}(u^i_m), ps^i_{other}(u^i_m)$

Algorithm 2 Prioritize drivers

Require: graph G, Degs and PPCs
    calculate ps
for $u_m^i \in M^i$ **do**
        $max_{own} \leftarrow Max(ps^i_{own}(u^i_m))$
        $max_{other} \leftarrow Max(ps^i_{other}(u^i_m))$
        calculate $pcs(u^i_m)$
while $\exists u_m^i \in M^i$ **do**
        for $u_m^i \in M^i$ **do**
            $ps^i(u^i_m) = \alpha_2 \times ps^i_{own}(u^i_m) + (1-\alpha_2) \times ps^i_{other}(u^i_m).$
            $w^i(u^i_m) = \alpha_1 \times ps^i(u^i_m) + (1-\alpha_1) \times pcs(u^i_m)$
        $selected\_driver \leftarrow w^i(u^i_m))$
        delete neighbors($selected\_drive$)
        delete $selected\_drive$
        calculate ps

CHAPTER 4

4. Results and Discussion

We implemented the algorithm in Python, using the NetworkX library to create the bipartite graphs. The source code and the input data are available on Github.

4.1. Comparison to other methods

We compare our Models to Prodigy [11], and SCS [10] using the same input (COAD, LUAD, HNSC from TCGA and cell lines from CCLE) data and network. These methods were chosen since there aren't many methods that study specific-sample cancer prioritization of driver genes, and due to its data input similarity.

4.2. Validation

In order to evaluate the performance of our method, we use a well-studied cancer gene database, consisting of 723 genes, from the COSMIC Cancer Gene Census (CGC) [18]. CGC is a part of COSMIC, the catalog of somatic mutations in cancer datasets. CGC is used for the evaluation of methods that are run on both TCGA data and CCLE datasets. CGC contains cancer-associated genes that have different types of mutations: SNVs, translocations, amplifications. Since we only use SNVs and INDELs as input, we define the reference gene set of a patient as the set of genes that are mutated in that patient.

Apart from CGC, other reference lists are used for evaluating the results obtained from CCLE. We use the Genomics of Drug Sensitivity in Cancer (GDSC, [19]) database to retrieve the list of drugs to which the corresponding cell line shows strong sensitivity.

Namely, for each cell line we identify the drugs with z-scores < 0 where the z-score is provided by GDSC. It assesses whether a cell line is significantly sensitive or resistant to a drug by comparing the drug response curves of the drug of interest for all the cell lines. Once we identify the drugs, we also retrieve their target genes from the same database. We intersectthese target genes with CGC and with the set of mutated genes in the corresponding cell line to get the final reference list. We also compile a related reference list where we also include the neighbors of the target genes based on the PPI.

CGC reference list

$$reference = CGC \cap SNVs \tag{4.1}$$

Drug Target reference list:

$$reference = DrugTargets \cap CGC \tag{4.2}$$

Drug Targets and neighbor reference list:

$$reference = (DrugTargets + neighbors) \cap CGC \tag{4.3}$$

The performance of each method is measured by means of Precision, Recall and F1 score, that are computed with respect to reference gene sets.

Let $Drivers_i[k]$ be the set of top k genes for the patient *i*, then

$$Precision_i[k] = \frac{Drivers_i[k] \cap reference_i}{k} \tag{4.4}$$

$$Recall_i[k] = \frac{Drivers_i[k] \cap reference_i}{reference_i} \tag{4.5}$$

$$F1_i[k] = 2 * \frac{Precision_i[k] * Recall_i[k]}{Precision_i[k] + Recall_i[k]} \tag{4.6}$$

We use two different evaluation methods. The *Repetitive* used by Prodigy if an individual has less than N ranked genes, the last value for this patient would be duplicated so that all quality measure vectors for all patients are of length N. Together with *Common* methods, where every patient would have the same number of driver genes across all methods, in other words for each patient we take the minimum number of drivers across all the methods.

4.3. Results

Using three cohorts of cancer patients from TCGA: COAD, LUAD and HNSC (279, 505, and 498 patients, respectively), and Cell Lines (colon, lung, 45, and 33 patients respectively), we perform a comparison with two other methods, Prodigy [11], and SCS [10] along with our method using various values for our alpha parameters. We also include a simpler version of our model called the the Pathway Coverage Score Only (***PCS***) model, where we don't include the first part of the weight $w^i(u)$ formula, the *personalized score* of u ($ps^i(u)$) is not included in the calculations. We just take into consideration the the *pathway coverage score* of u ($pcs^i(u)$) ($w^i(u) = pcs(u)$).

4.3.1. Ranking of Driver Genes

4.3.1.1. Evaluations with respect to CGC reference list

We first compare the models based on their potential to recover the CGC genes. We compute the precision, recall, and F1 score for each patient and then compute the average value across the patients for each gene in the top 20. We then calculate the area under the curve (**AUC**) as the value over all the patients. Figure 4.1 shows the results obtained from all four models using the COAD cancer data.

We compare the driver genes obtained from each method to the CGC list, using *Repetitive* evaluation (Figure 4.1), and the *Common* evaluation (Figure 4.2). For each gene in the top *k*, the number of patients covered is represented in the figures.

Giving a higher weight to the *pathway coverage score* ($pcs^i(u)$) (a small α_1), and the specific patient's personalized score ($ps^i_{own}(u)$) (a large α_2), results in higher performance for our model, as shown in the following figures, the model with α_1 as 0.1 (Pathway Coverage Score as 0.9), and α_2 as 0.9, outputs the best ranking over all the methods for both evaluations *Repetitive* and *Common*.

The PCS Only model achieves a higher AUC value than the alternative methods, although it does not use any other data related to the patient rather than the mutation genes set and pathways. The second ranked method is our model, due to the integration of the PCS data to the equation, which gives it higher performance. Finally, Prodigy and SCS rank third and fourth respectively in both evaluation types.

In the *Repetitive* evaluation, our models' performance gets worse in precision after the top 9 genes, and this is due to the fact that our models output a larger set of driver genes for each patient than the alternative methods. For Prodigy and SCS, which usually output a small set of drivers, the precision value at the last output gene is repeated till 20 for most of the patients. As shown in Figure 4.1 after the 9^th ranking gene the curves start getting an almost straight line to the end of the curve. This is more clear in the recall and F1 score evaluation curves.

While in the *Common* evaluation our models' performance is kept higher along with all the top k genes since the number of drivers is the same for every patient overall models where the evaluation is fair.

These results show that our method can find driver genes for every patient separately, and accurately in colon cancer.

Figures 4.3 and 4.4 shows corresponding results for the lung cancer data. We use the same parameters for LUAD as in COAD in our method, giving a higher weight to the PCS and the patient's degree in the Alpha Score formula. As in the previous evaluation, our model reaches the best performance with CGC reference sets compared to the alternative methods Prodigy and SCS, but the PCS Only model still has a better performance.

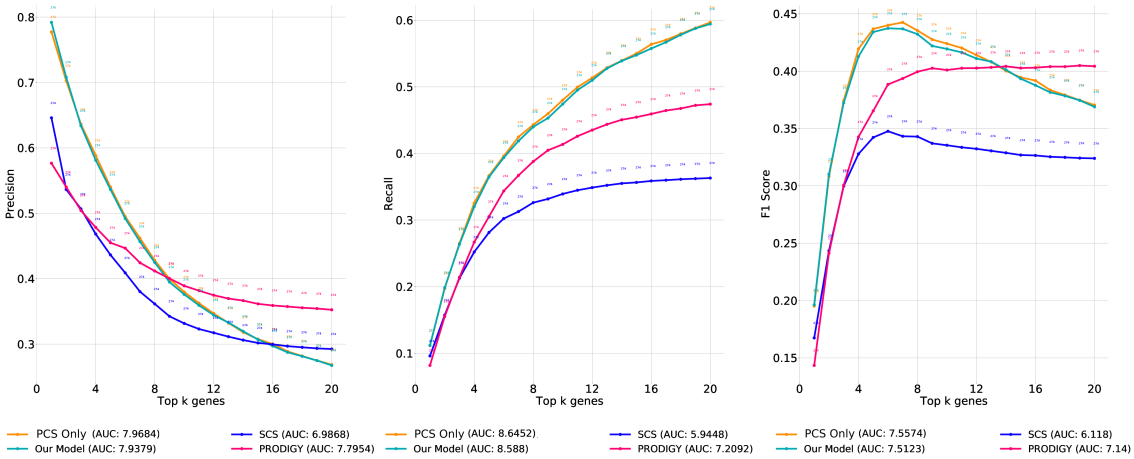

Figure 4.1: CGC Repetitive with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD data

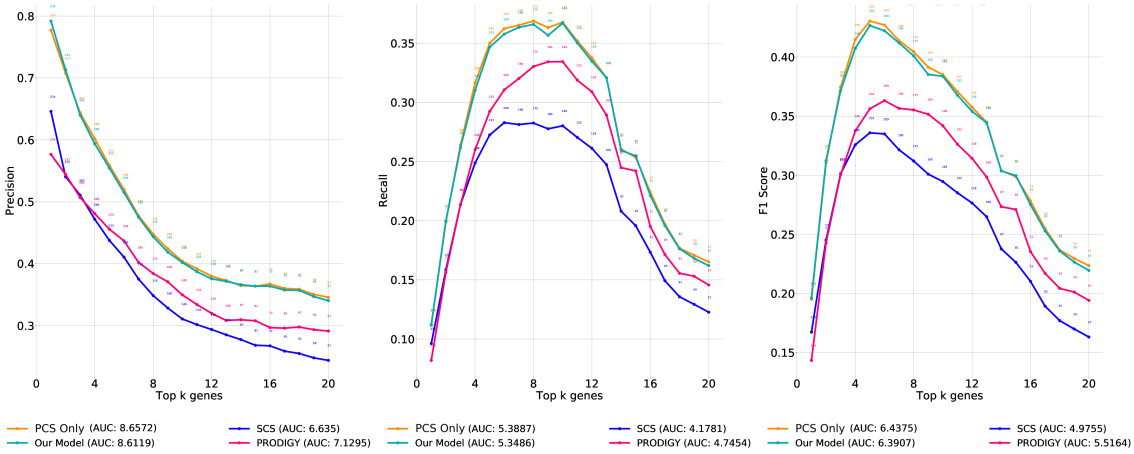

Figure 4.2: CGC Common with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD data

These results show that our method can find driver genes for every patient separately, and accurately in lung cancer.

Figures 4.5 and 4.6 present the corresponding results for the HNSC data. As in the previous cancer data, we use the same parameters in our method. Unlike the previous evaluation's results, Prodigy ranks first as per *Repetitive* precision AUC value, the PCS Only, and our model comes after respectively. Finally, SCS is the worst ranking method. However in the rest of the evaluations, the ranking is similar to the previous cancer data ranks, the PCS Only model has a higher performance, then our model comes next very close to the PCS model. Prodigy comes next, far from our model. Finally, SCS scores the worst between all the alternative models.

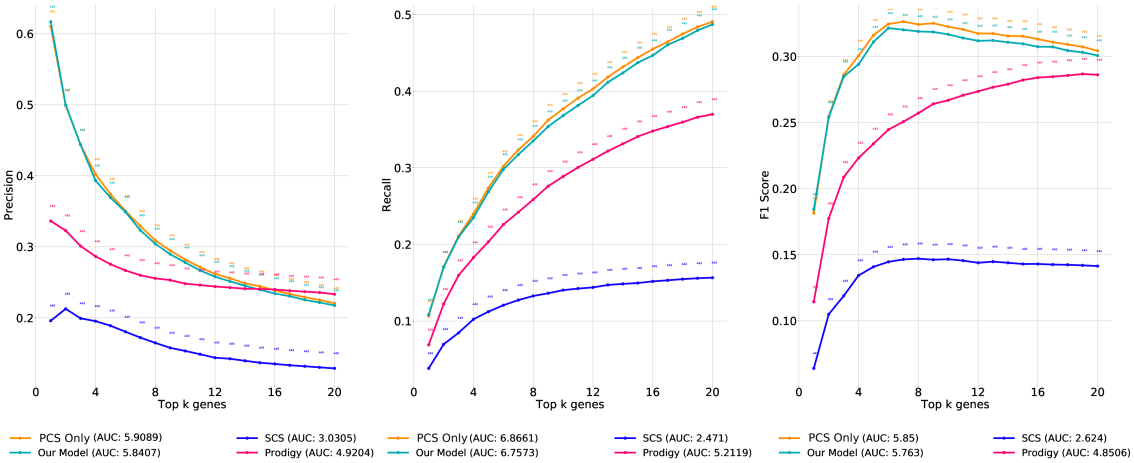

Figure 4.3: CGC Repetitive with respect to precision, recall and F1 scores calculated as an average across the cohort for LUAD

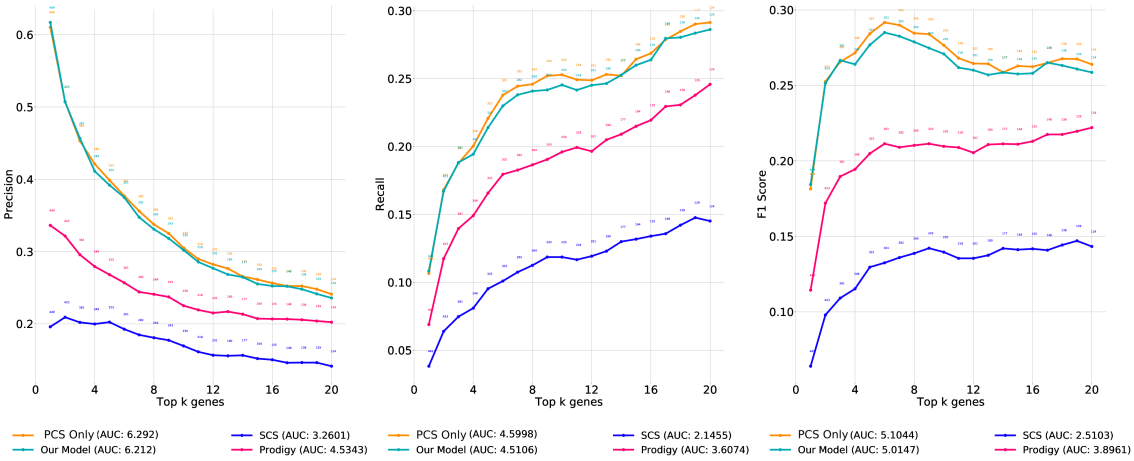

Figure 4.4: CGC Common with respect to precision, recall and F1 scores calculated as an average across the cohort for LUAD

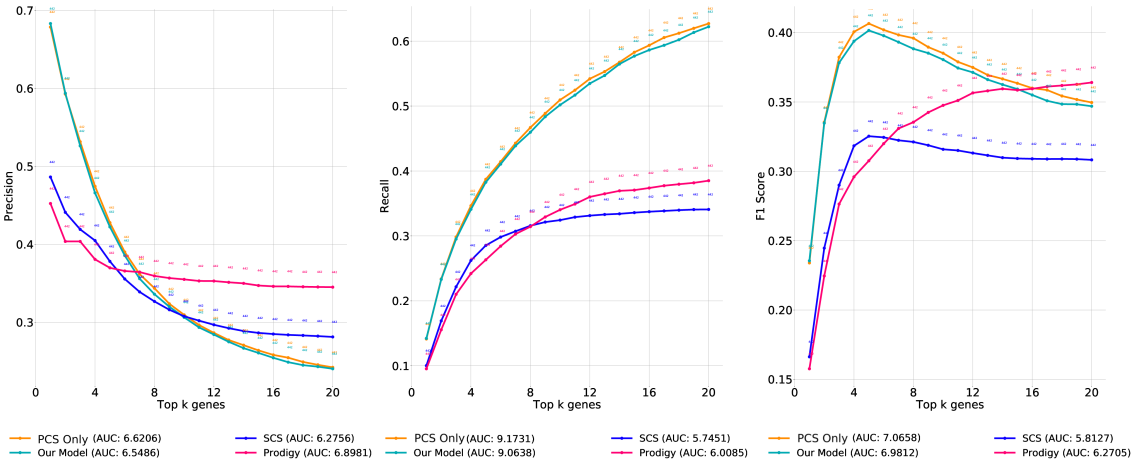

Figure 4.5: CGC Repetitive with respect to precision, recall and F1 scores calculated as an average across the cohort for HNSC

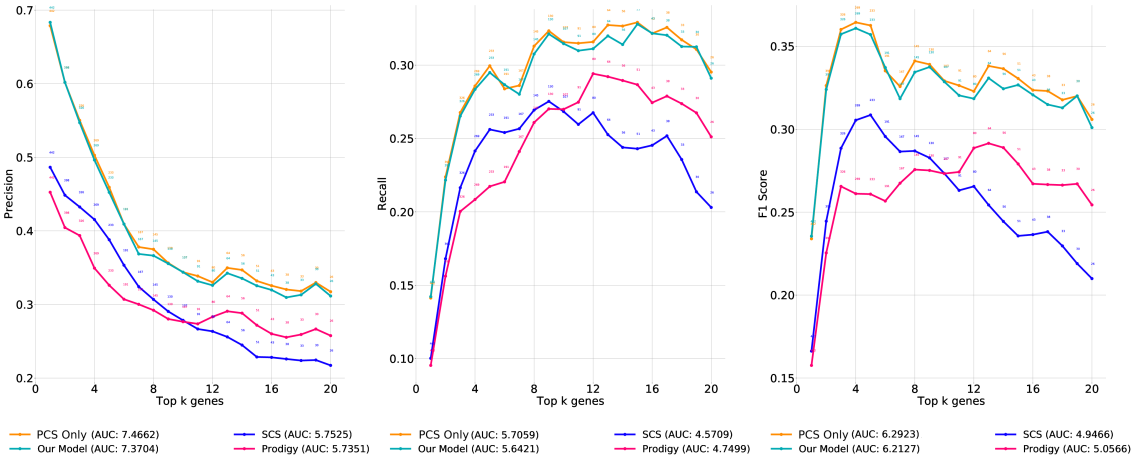

Figure 4.6: CGC Common with respect to precision, recall and F1 scores calculated as an average across the cohort for HNSC

4.3.1.2. Recovering Rare Driver Genes

In this evaluation, we identify the rare drivers, the mutated genes that have a mutation frequency of <2% in the cohort, and ranked in the top 20 driver lists of the patient. The reference list of the rare drivers differs from cancer to another, it's the intersection of CGC genes with the total rare genes of the cancer cohort, (COAD, LUAD, and HNSC have 258, 384, and 424 reference genes respectively).

$$reference = GGC \cap Rare\_SNVs \tag{4.7}$$

As in CGC evaluation, using the same model, giving a higher weight to the Pathway Coverage Score will result in a better performance for our model. The model with 0.1 as α_1 and 0.9 as α_2 (Pathway Coverage Score as 0.9), outputs the best driver ranks overall cancer types compared to the other possible parameters combinations.

Figures 4.7 and 4.8 shows analogous results for the colon cancer data. SCS ranks first in *Repetitive* precision and F1 Score, which is followed by Prodigy, our model, and PCS Only. Whereas in the *Repetitive* recall, Prodigy gets the best performance followed by our model and PCS Only respectively, and finally SCS ranks last.

In the *Common* evaluation, the ranks are different from the previous evaluations. Our model ranks the second best after PCS Only, by a small difference as per

AUC value, in the precision and F1 score evaluations followed by SCS and then Prodigy. On the other hand, SCS performs best in the recall evaluation, succeeded by Prodigy, our model and then PCS Only last.

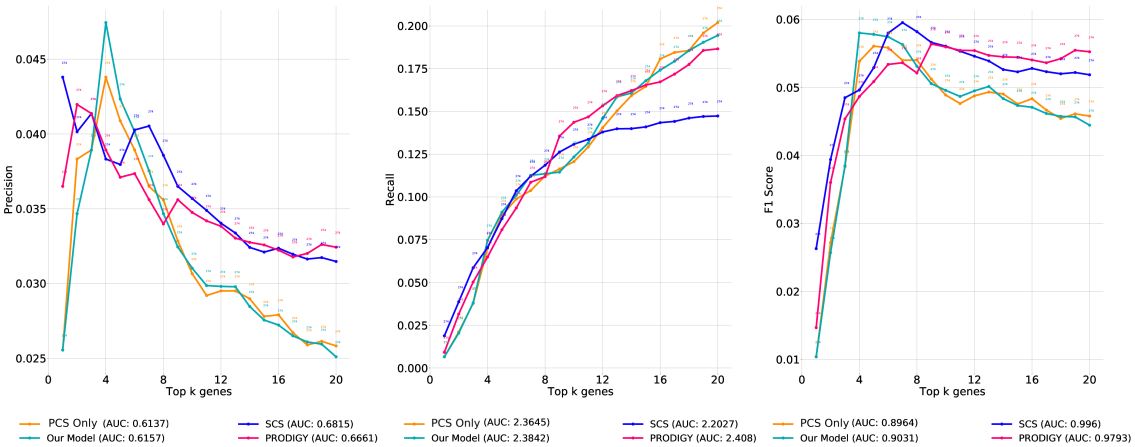

Figure 4.7: Rare drivers Repetitive evaluation with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD

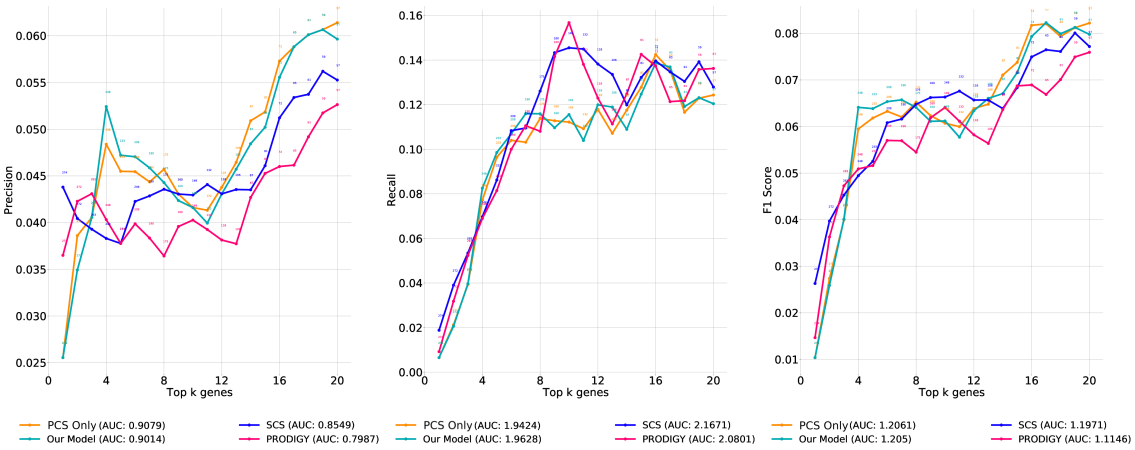

Figure 4.8: Rare drivers Common with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD

The figures 4.9 and 4.10 shows the corresponding results of the lung cancer data. Contrary to the previous cancer type results, the ranking of the methods are similar for *Repetitive* and *Common* evaluation schemes. Our model ranks as expected second after PCS Only, accompanied by Prodigy and then SCS at last in all ranking evaluations.

The results show that our model can identify rare driver genes in LUAD cancer effectively, and accurately.

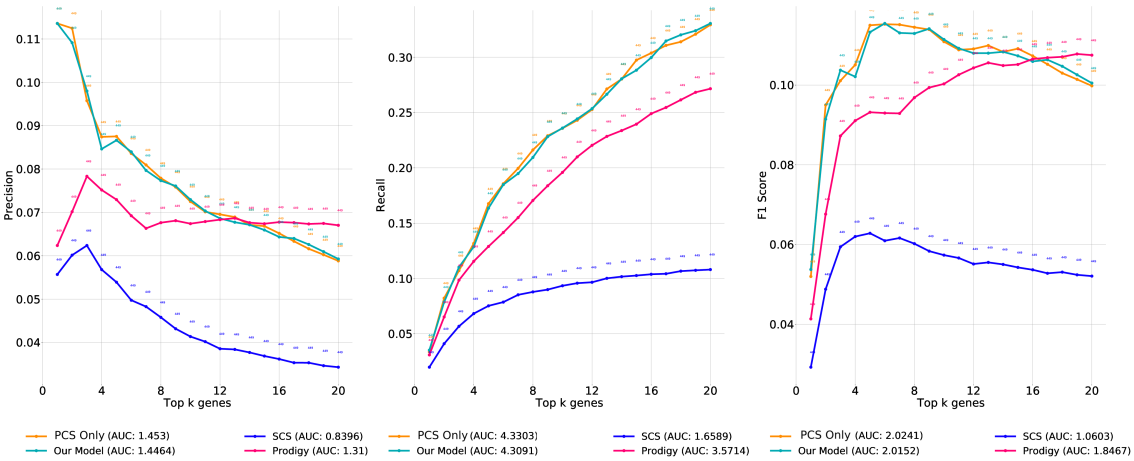

Figure 4.9: Rare drivers Repetitive evaluation with respect to precision, recall and F1 scores calculated as an average across the cohort for LUAD

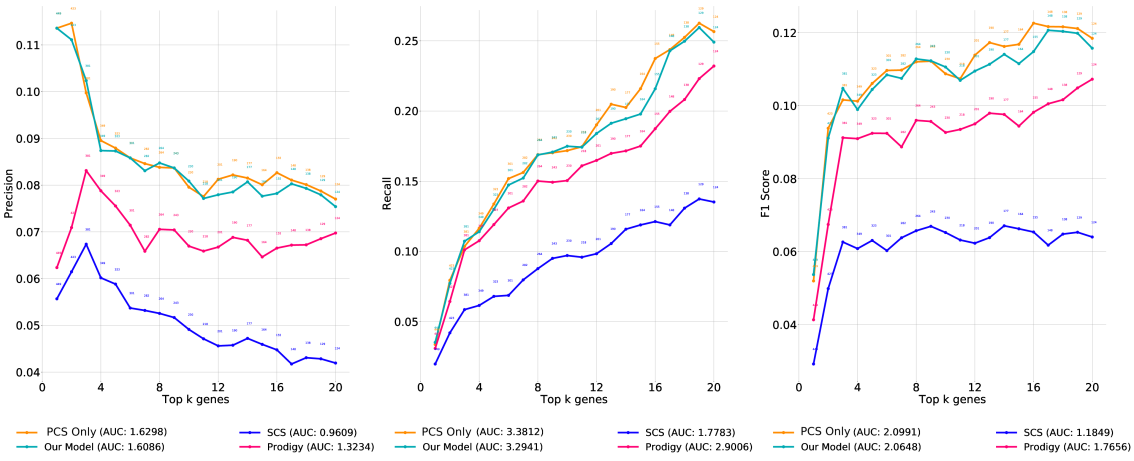

Figure 4.10: Rare drivers Common with respect to precision, recall and F1 scores calculated as an average across the cohort for LUAD

Lastly, the figures 4.11 and 4.12 shows the results on HNSC data. Starting with the *Repetitive* precision evaluation, Prodigy ranks first followed up by SCS accompanied closely by PCS Only then our model at last. Dissimilarly PCS Only model ranks the drivers best in all other evaluation types, followed by our model, succeeded by Prodigy and SCS respectively at last position.

These figures demonstrate that our model can distinguish rare driver genes more accurately than other methods generally.

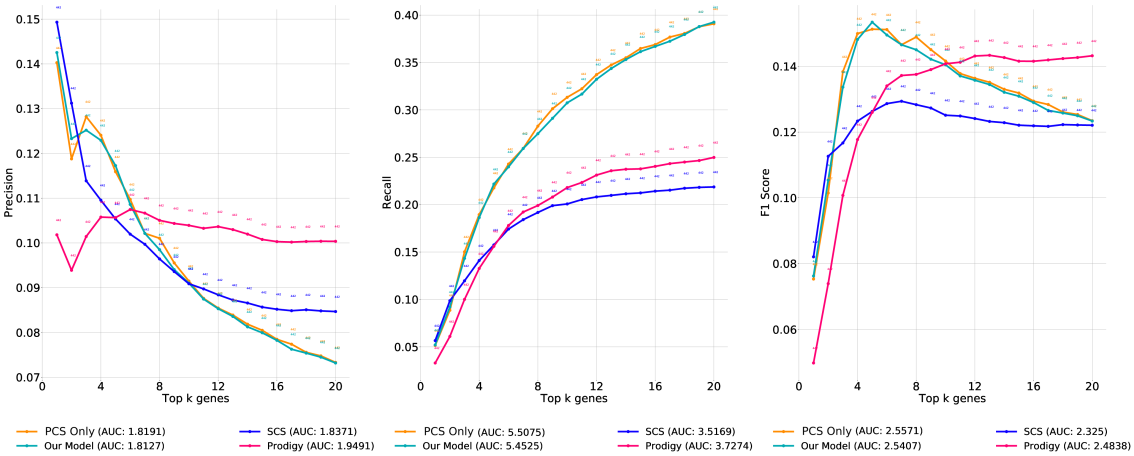

Figure 4.11: Rare drivers Repetitive evaluation with respect to precision, recall and F1 scores calculated as an average across the cohort for HNSC

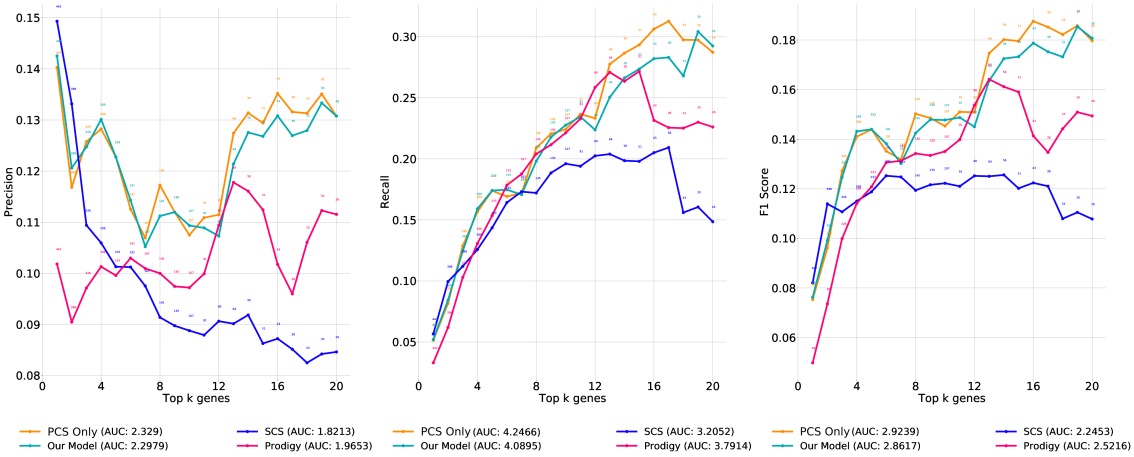

Figure 4.12: Rare drivers Common with respect to precision, recall and F1 scores calculated as an average across the cohort for HNSC

4.3.2. Evaluations on CCLE data

Cell lines are usually originated from a single *Common* ancestor cell. Human

cancer cell lines have been one of the pillars for cancer studies. The advantages that make cell lines useful as in vitro models are that they provide a homogenous population of cells. And they are relatively easy to grow in a short period. They are employed to study the biology of cancer and to test hypotheses to improve the efficacy of cancer treatment [20].

The CCLE project is a collaboration between the Broad Institute and the Novartis Institutes for Biomedical Research. Its goal is to conduct a detailed genetic and pharmacologic characterization of a large panel of human cancer models.

4.3.2.1. Drug Targets Reference List

In this evaluation, we define the reference list by intersecting the *CGC* genes with the *Drug Targets* genes.

Figures 4.13 and 4.14 show the results obtained from colon cancer cell lines. Our model has the highest AUC, followed by the PCS Only model, Prodigy, and SCS.

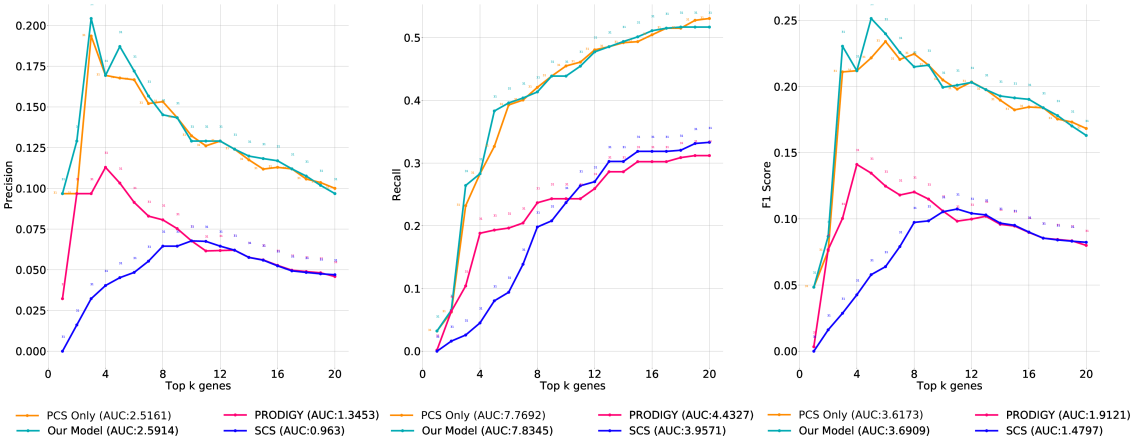

Figure 4.13: Drug Targets Repetitive average precision, recall, F1 across all cell lines for COAD

The figures 4.15 and 4.16 illustrate the results on the lung cancer cell lines. Our model has the highest AUC, ranking first followed by the PCS Only model, Prodigy respectively then SCS has the worst results in *Repetitive* precision and F1 score evaluations. On the other hand, SCS has a better AUC then Prodigy in the *Repetitive* recall coming after our model and PCS Only respectively.

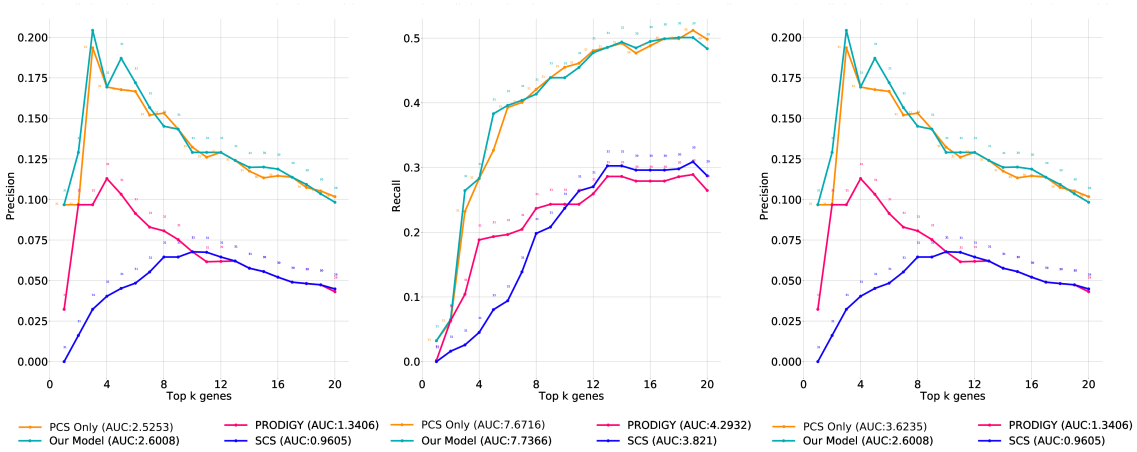

Figure 4.14: Drug Targets Common average precision, recall, F1 across all cell lines for COAD

On the other hand in the *Common* evaluations, our model achieves higher results per AUC value, followed by PCS Only. Prodigy scores a higher AUC in the precision evaluation then SCS, while in recall and F1 score SCS performs better.

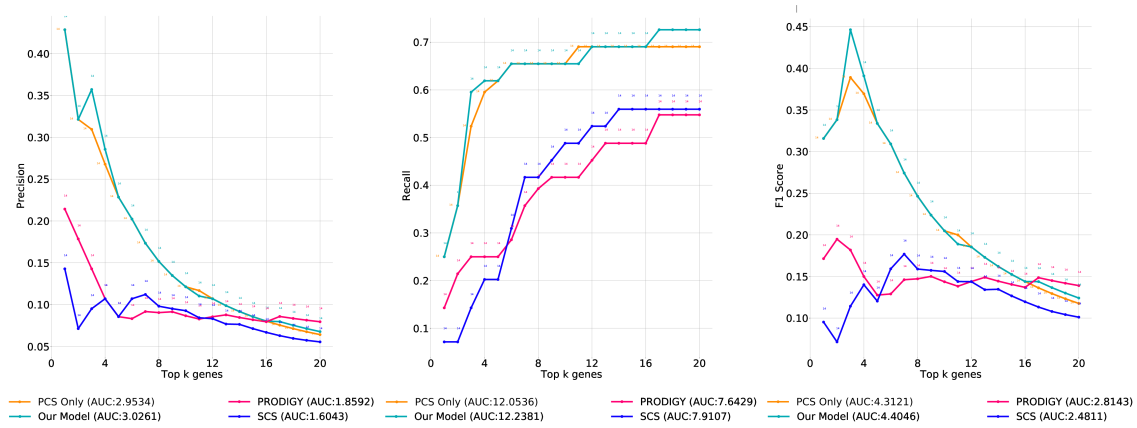

Figure 4.15: Drug Targets Repetitive average precision, recall, F1 across all cell lines for LUAD

4.3.2.2. Drug Targets and Neighbors Reference List

In this evaluation, we take as a reference list the intersection of the *CGC* genes with the *Drug Targets* genes in addition to their first mutated neighbors.

Using the same model against the other methods, we got the following results shown in Figures 4.17 and 4.18 evaluating the colon cell lines. Our model has the best performance in all evaluations followed by the PCS Only model. Prodigy comes

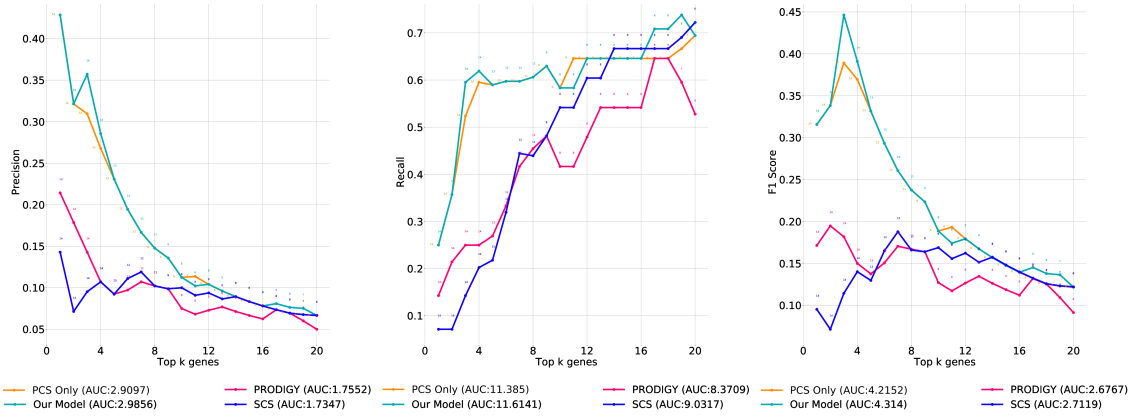

Figure 4.16: Drug Targets Common average precision, recall, F1 across all cell lines for LUAD

third before SCS in precision evaluation in both *Repetitive* and *Common*. However, SCS performs better than Prodigy in recall and F1 score in both *Repetitive* and *Common*.

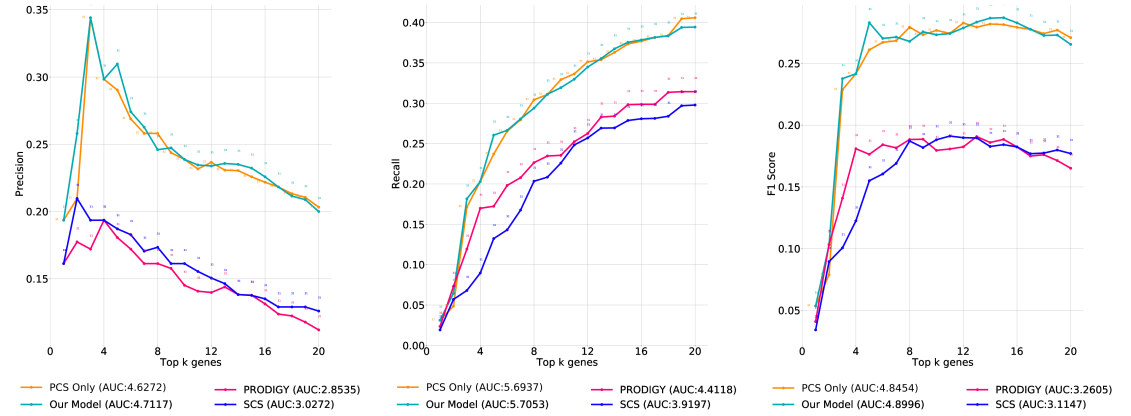

Figure 4.17: Drug Targets and Neighbors Repetitive average precision, recall, F1 across all cell lines for COAD

The Figures 4.19 and 4.20 illustrates the evaluation of the lung cancer cell lines data. Our model has the best performance in all evaluations followed by the PCS Only model. Prodigy and SCS rank third and fourth, respectively in terms of recall defined based on *Repetitive* evaluation scheme. SCS performs better than Prodigy in the rest of both *Repetitive* and *Common* evaluations.

Overall, our model outperforms PCS Only, Prodigy, and SCS in terms of Precision, Recall, and F1 for both *Repetitive* and Common evaluations, using various reference lists in CCLE data.

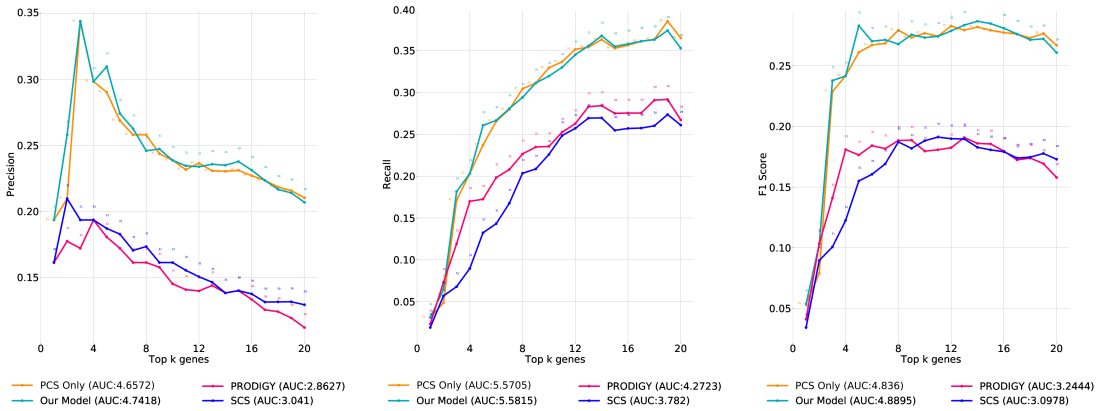

Figure 4.18: Drug Targets and Neighbors Common average precision, recall, F1 across all cell lines for COAD

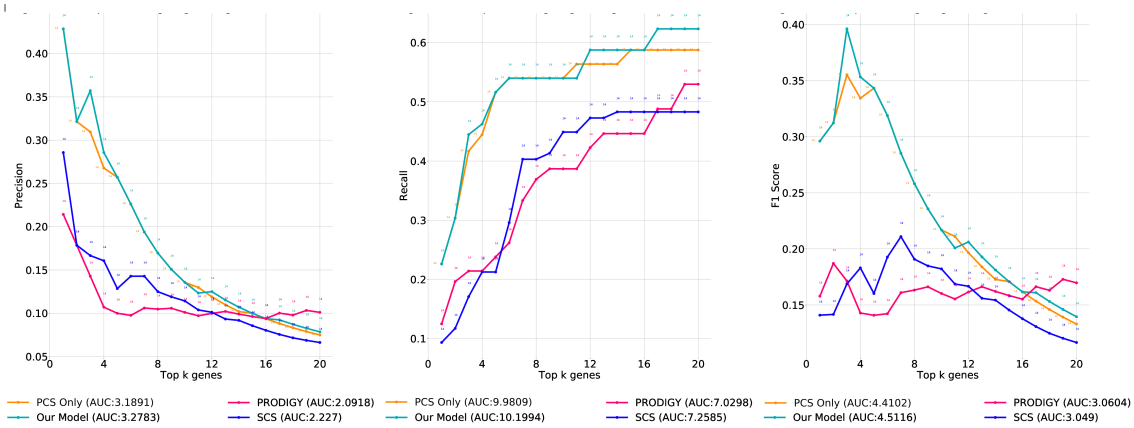

Figure 4.19: Drug Targets and Neighbors Repetitive average precision, recall, F1 across all cell lines for LUAD

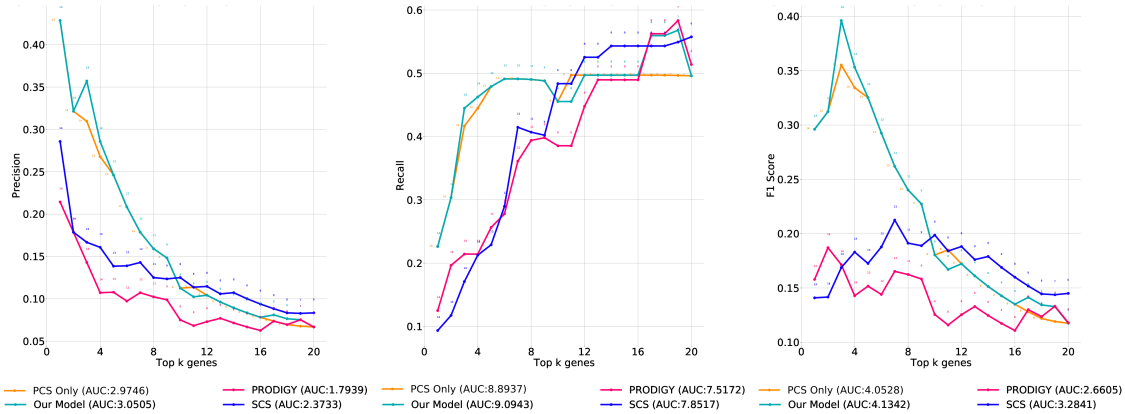

Figure 4.20: Drug Targets and Neighbors Common average precision, recall, F1 across all cell lines for LUAD

CHAPTER 5

5. Conclusion

5.1. Conclusion

In this thesis, we propose a novel approach to identify patient-specific driver genes. An important contribution of our method is the discovery of the effectiveness of pathway coverage on the identification of patient-specific driver genes identification. Our method ranks mutated genes by their occurrence in the pathways collection, and their effect on dysregulated genes. Bipartite graphs are created to represent the relations between the mutated genes (drivers) and dysregulated genes (outliers), as well as the common pathway coverage in each patient. Using various datasets and evaluations, we show that the use of pathway coverage score and common pathway coverage improve the identification of patient-specific cancer drivers. Furthermore, our model outperforms the alternative methods in most evaluations in recovering known cancer reference genes. Another contribution of our model is the use of cell lines data from CCLE. To evaluate results obtained on cell lines, we introduce a novel reference list based on drug sensitivity data from GDSC. In conclusion, we propose a model that effectively combines genomic data with pathway and connectivity information and show its superiority in identifying patient-specific driver genes.

5.2. Future Work

One main direction is to improve the personalized score formula such that our method

performs better than PCS only model in a larger number of evaluations. Another direction is to utilize more recent and accurate input data. Within this context, we can use tissue-specific PPI and pathways depending on the cancer type. Currently, we use "bulk" mutation and gene expression data. Bulk approaches measure properties of tumor samples which are mixtures of multiple cell populations such as immune, stroma, and normal cells [21] [22]. As such, bulk measurements ignore the intra-tumor heterogeneity present in tumor samples. One strategy is to use the recently proposed deconvolution algorithms [23] to infer gene expression and mutation profiles specific to tumor cells to get rid of the effect of other accompanying non-tumor cells. A further future direction is to add another layer for evaluation of results obtained from TCGA data where patients are mapped to cell lines. We can use one of the recently developed approaches to construct this mapping [24] [25] [26] [27] [28]. Then, each patient can have its own reference gene set which is defined based on the drug targets approach that we use for cell lines. This would lead to a more accurate definition of patient-specific reference gene sets for evaluating patient-specific driver ranking methods. Lastly, one straightforward future step is to extend the evaluation to other cancer types in TCGA data.

Appendix A

Different Alpha values comparison on TCGA data

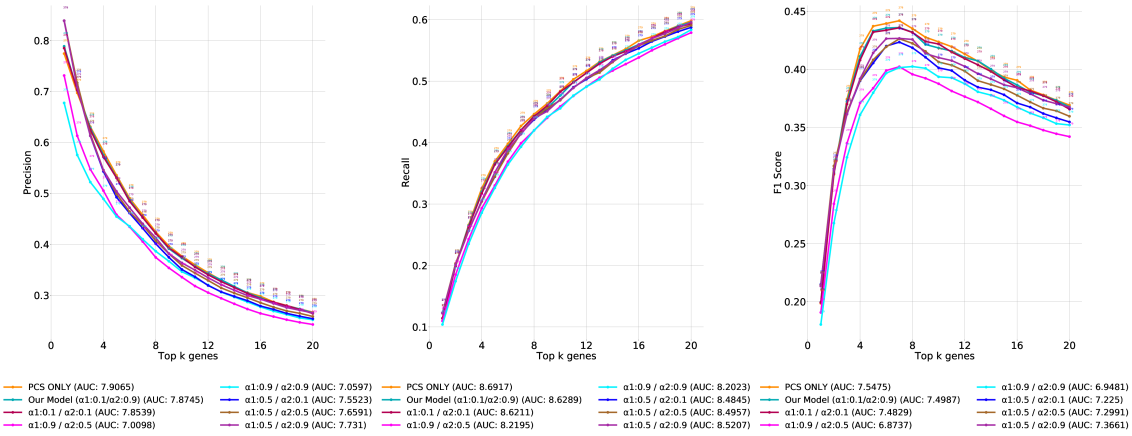

Figure A.1: CGC Repetitive with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD data different alpha values

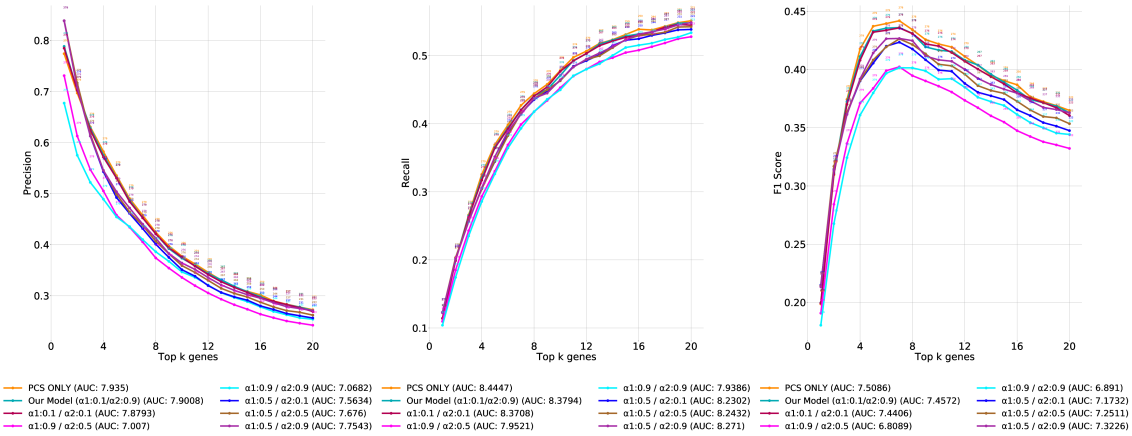

Figure A.2: CGC Common with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD data different alpha values

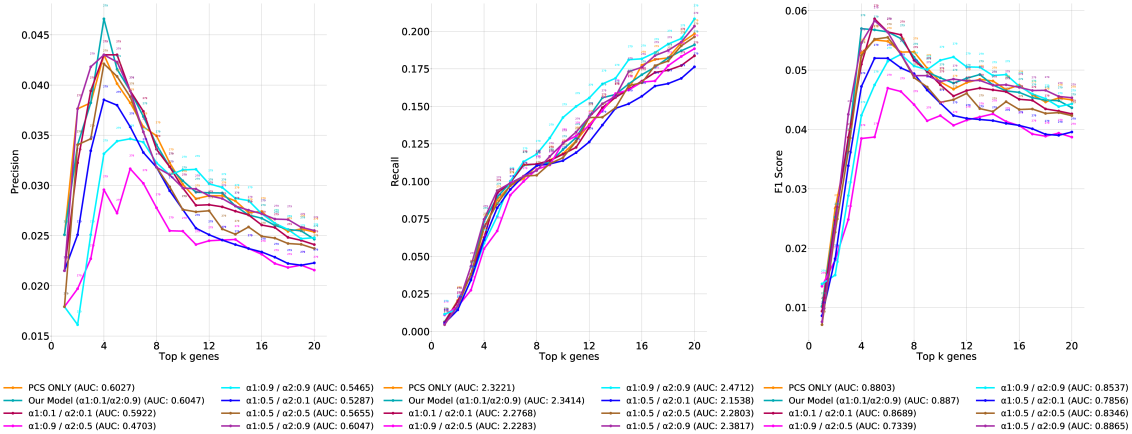

Figure A.3: CGC Repetitive with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD data different alpha values

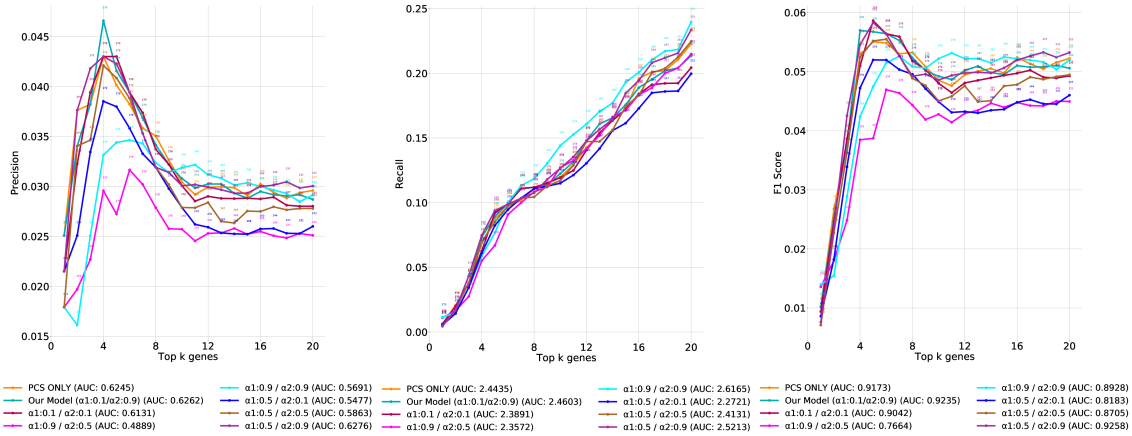

Figure A.4: CGC Common with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD data different alpha values

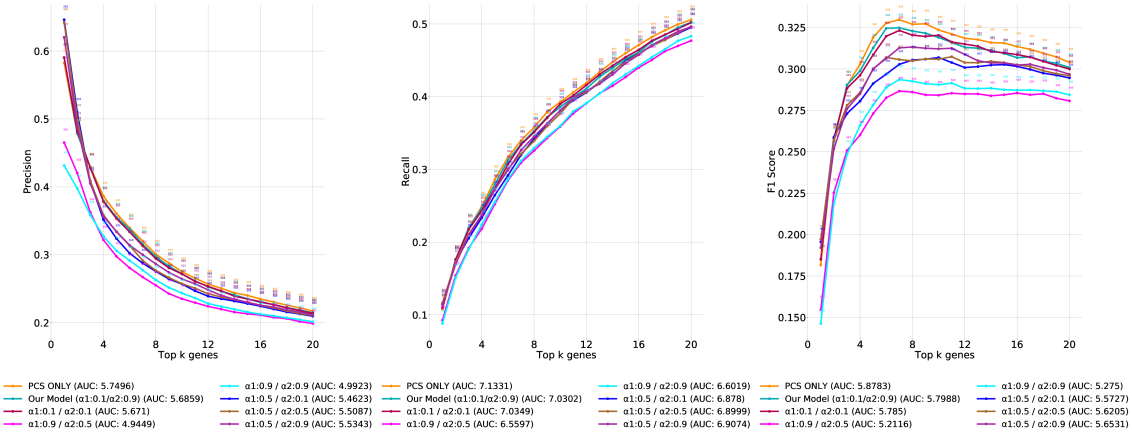

Figure A.5: CGC Repetitive with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD data different alpha values

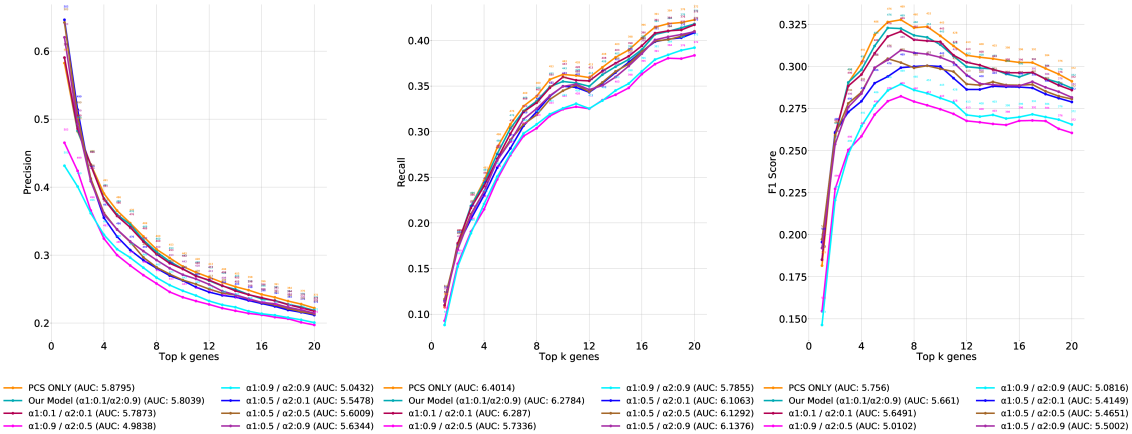

Figure A.6: CGC Common with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD data different alpha values

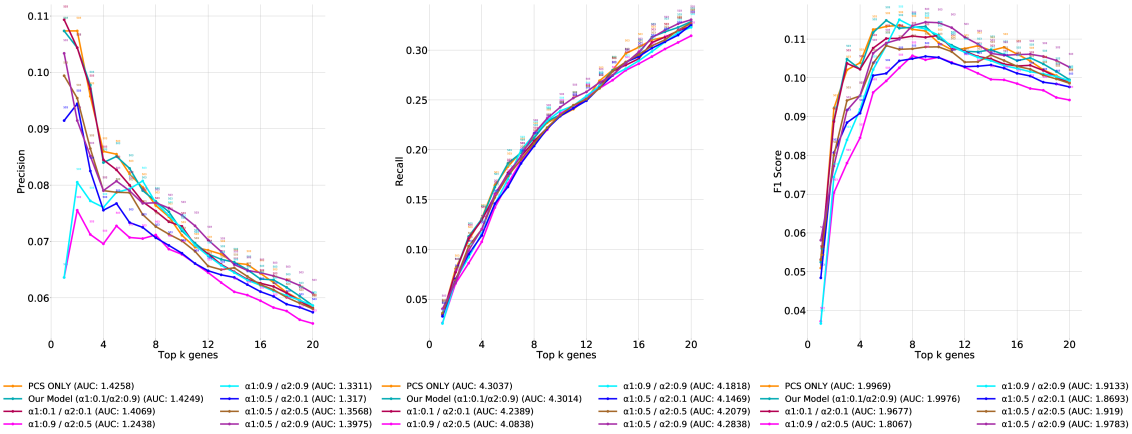

Figure A.7: CGC Repetitive with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD data different alpha values

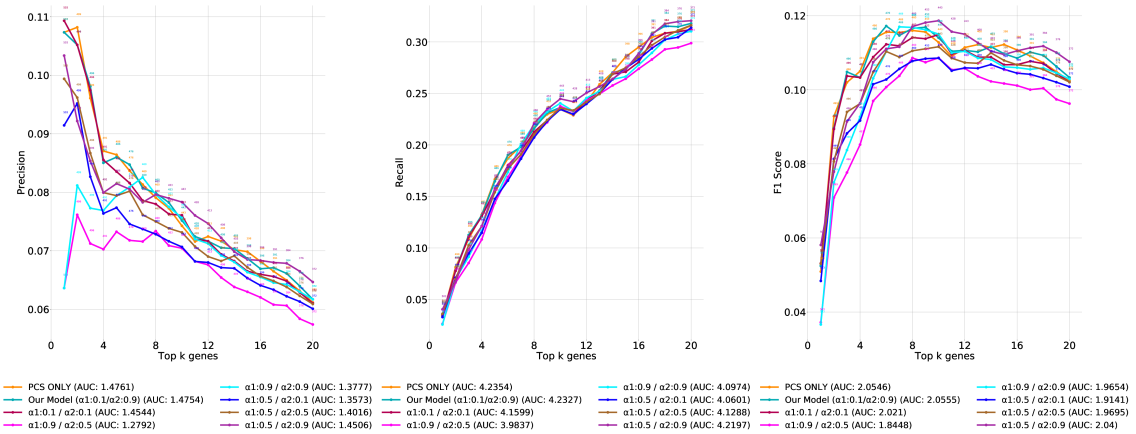

Figure A.8: CGC Common with respect to precision, recall and F1 scores calculated as an average across the cohort for COAD data different alpha values

Appendix B

Different Alpha values comparison on CCLE data

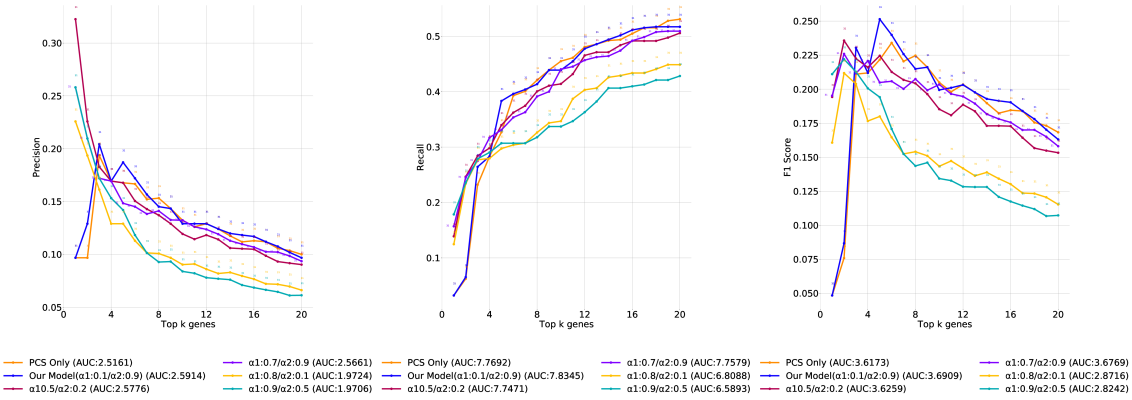

Figure B.1: Drug Targets Repetitive average precision, recall, F1 across all cell lines for COAD with different alpha values

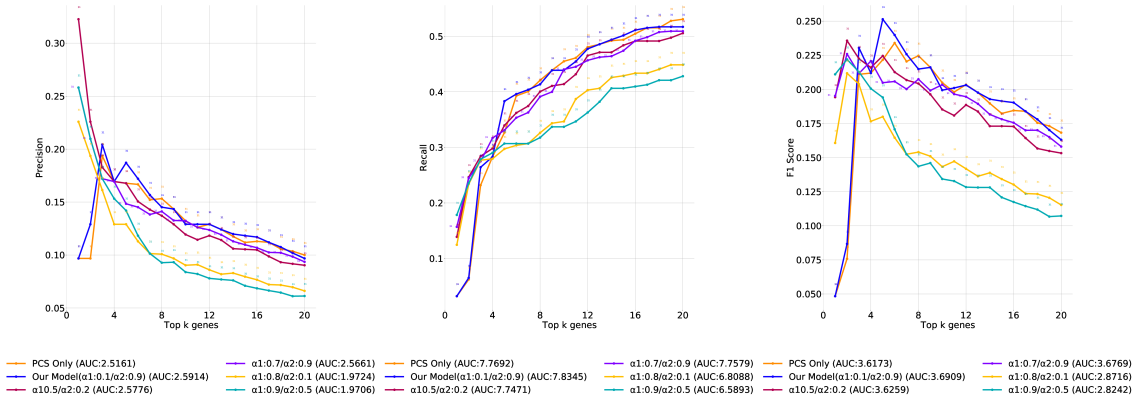

Figure B.2: Drug Targets Common average precision, recall, F1 across all cell lines for COAD with different alpha values

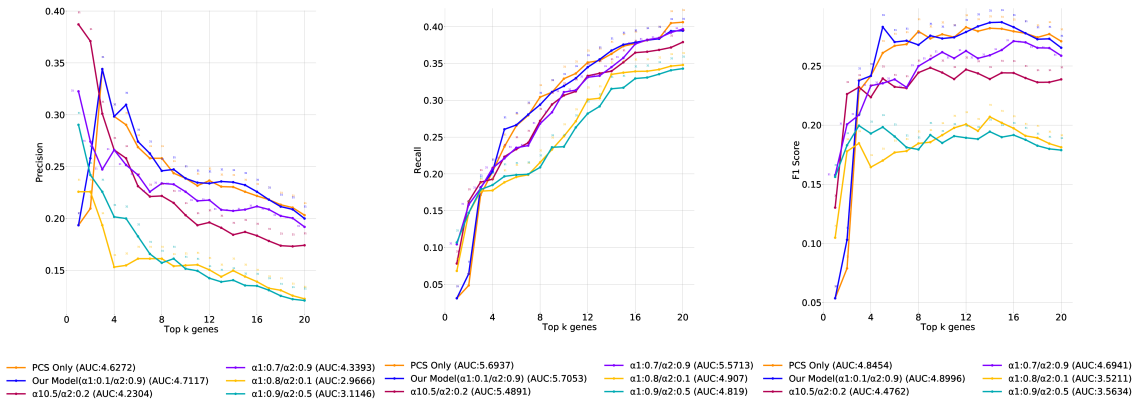

Figure B.3: Drug Targets and Neighbors Repetitive average precision, recall, F1 across all cell lines for COAD with different alpha values

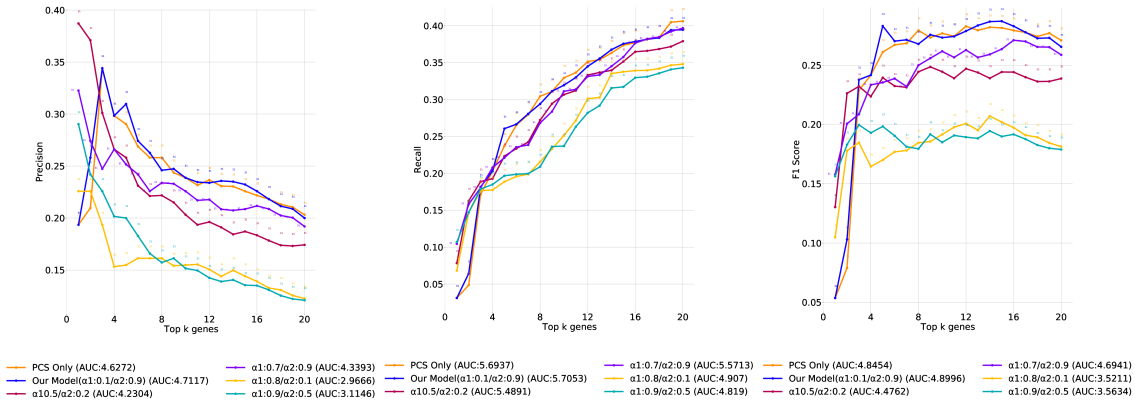

Figure B.4: Drug Targets and Neighbors Common average precision, recall, F1 across all cell lines for COAD with different alpha values

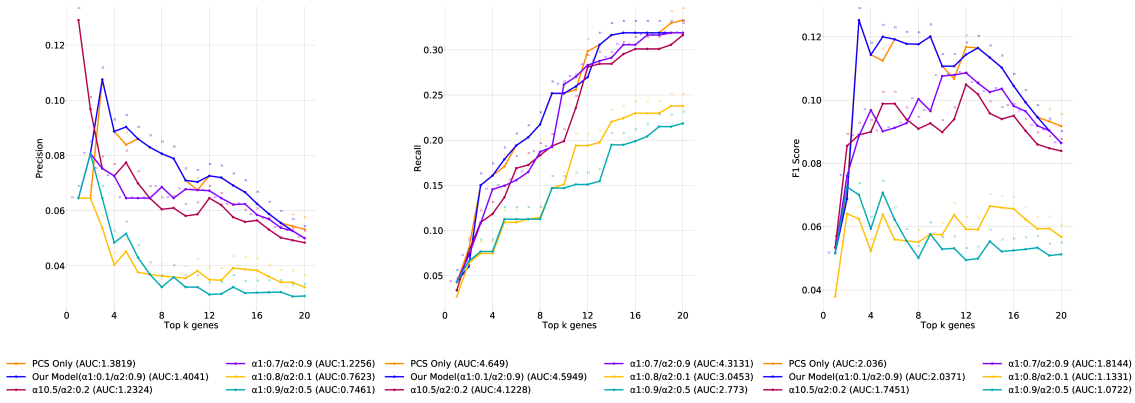

Figure B.5: Drug Targets Repetitive average precision, recall, F1 across all cell lines for COAD with different alpha values

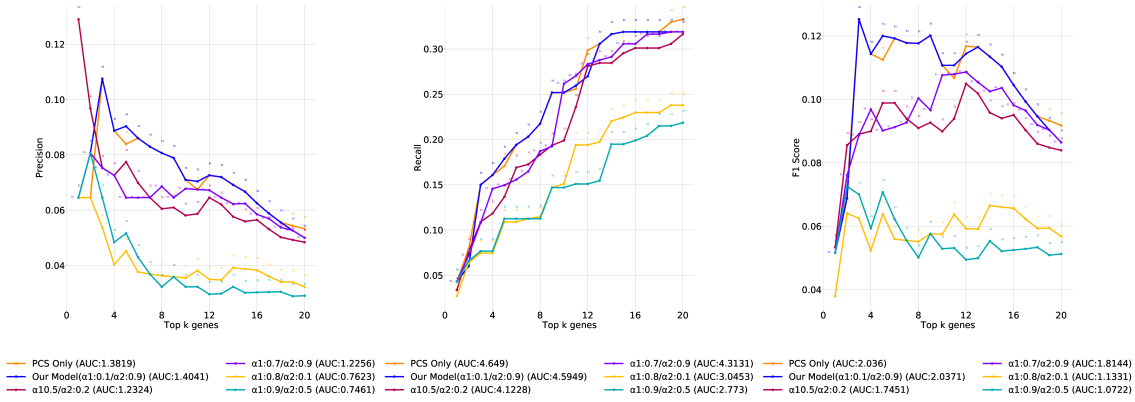

Figure B.6: Drug Targets Common average precision, recall, F1 across all cell lines for COAD with different alpha values

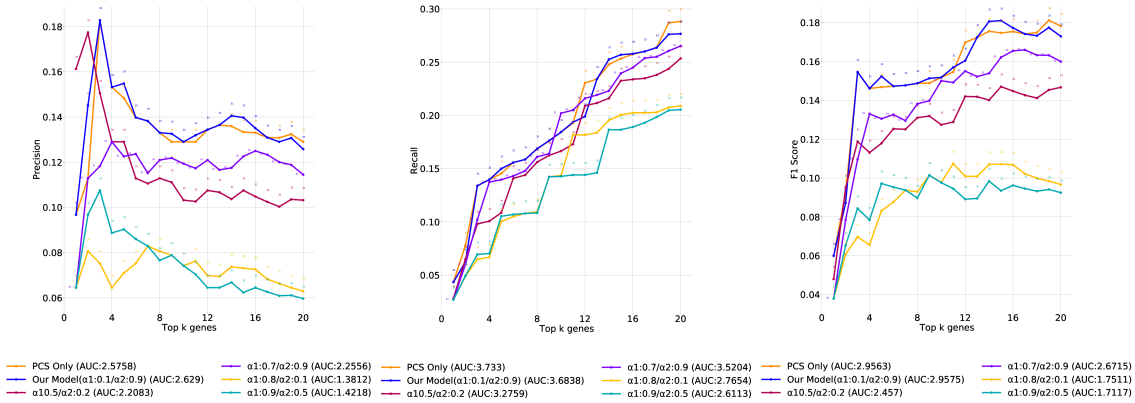

Figure B.7: Drug Targets and Neighbors Repetitive average precision, recall, F1 across all cell lines for COAD with different alpha values

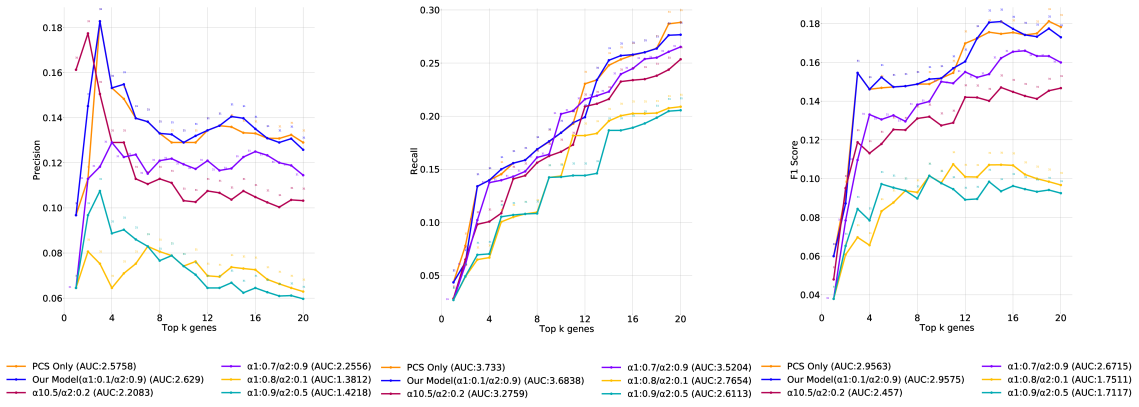

Figure B.8: Drug Targets and Neighbors Common average precision, recall, F1 across all cell lines for COAD with different alpha values

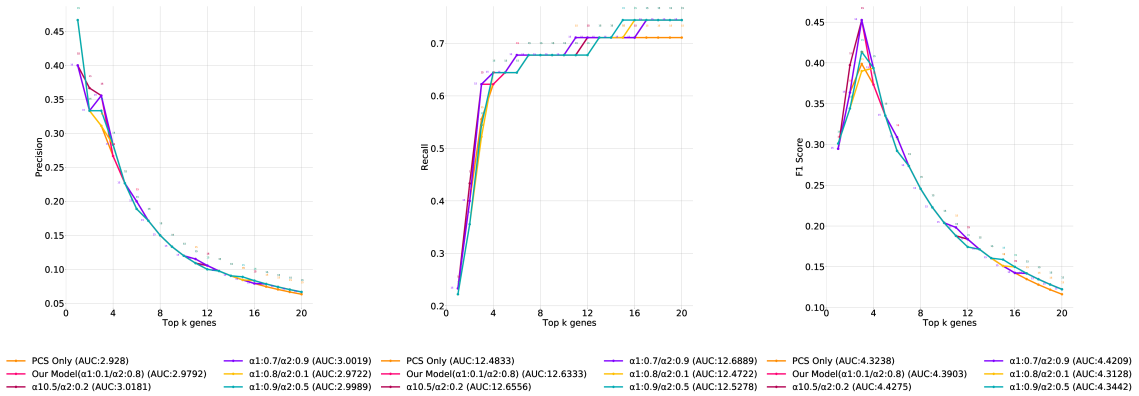

Figure B.9: Drug Targets Repetitive average precision, recall, F1 across all cell lines for LUAD with different alpha values

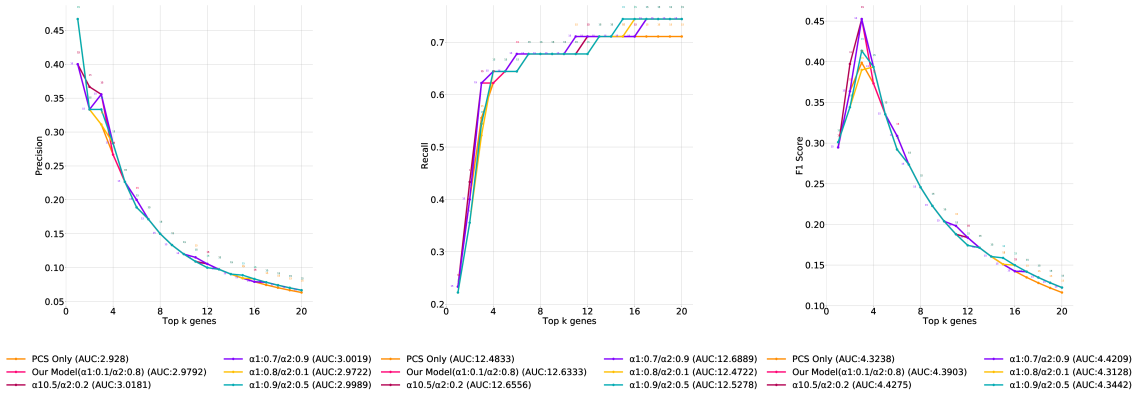

Figure B.10: Drug Targets Common average precision, recall, F1 across all cell lines for LUAD with different alpha values

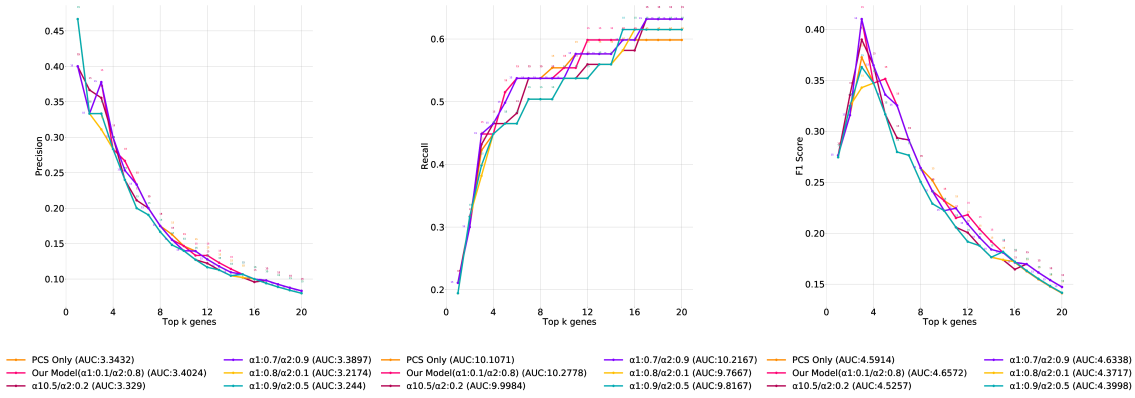

Figure B.11: Drug Targets and Neighbors Repetitive average precision, recall, F1 across all cell lines for LUAD with different alpha values

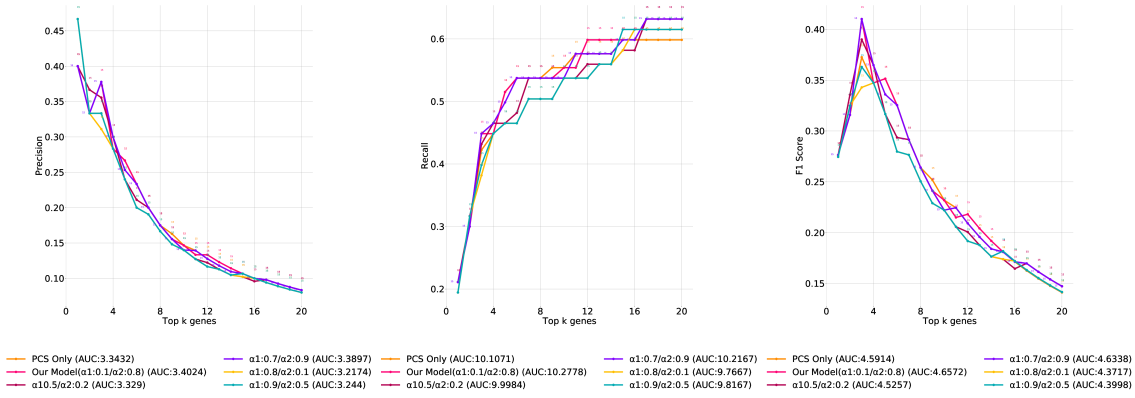

Figure B.12: Drug Targets and Neighbors Common average precision, recall, F1 across all cell lines for LUAD with different alpha values

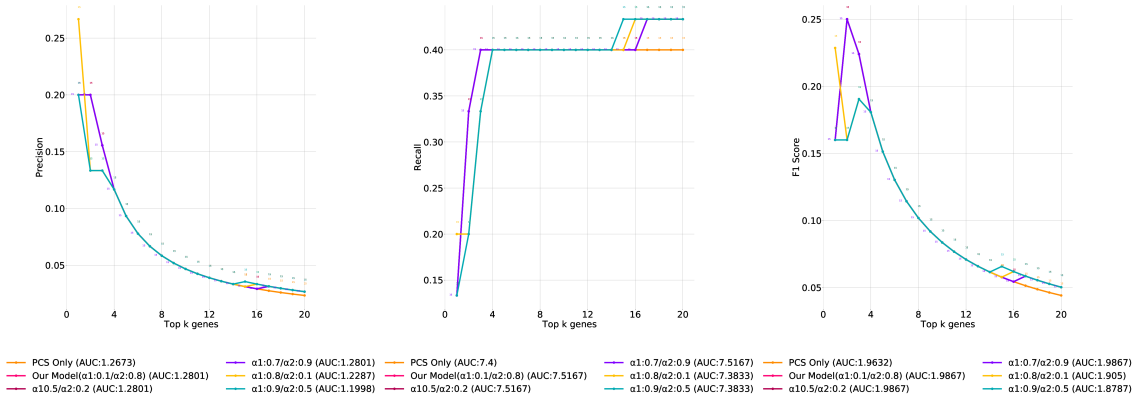

Figure B.13: Drug Targets Repetitive average precision, recall, F1 across all cell lines for LUAD with different alpha values

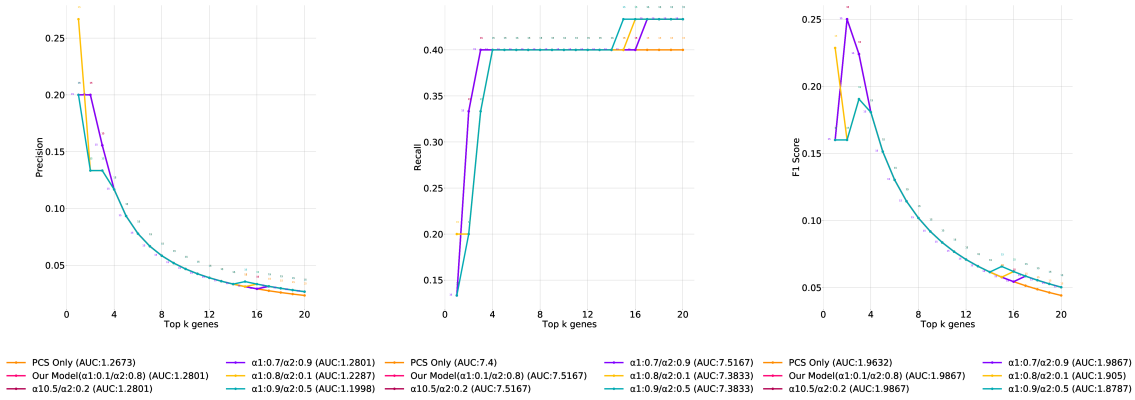

Figure B.14: Drug Targets Common average precision, recall, F1 across all cell lines for LUAD with different alpha values

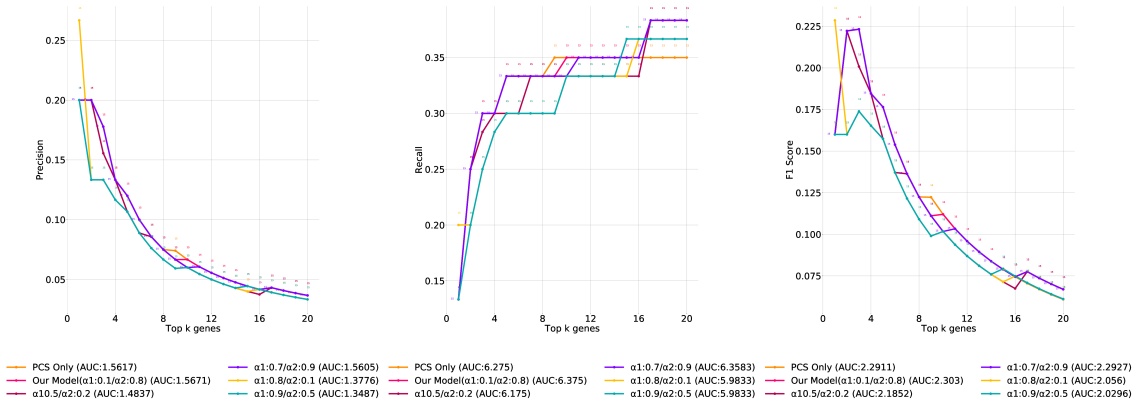

Figure B.15: Drug Targets and Neighbors Repetitive average precision, recall, F1 across all cell lines for LUAD with different alpha values

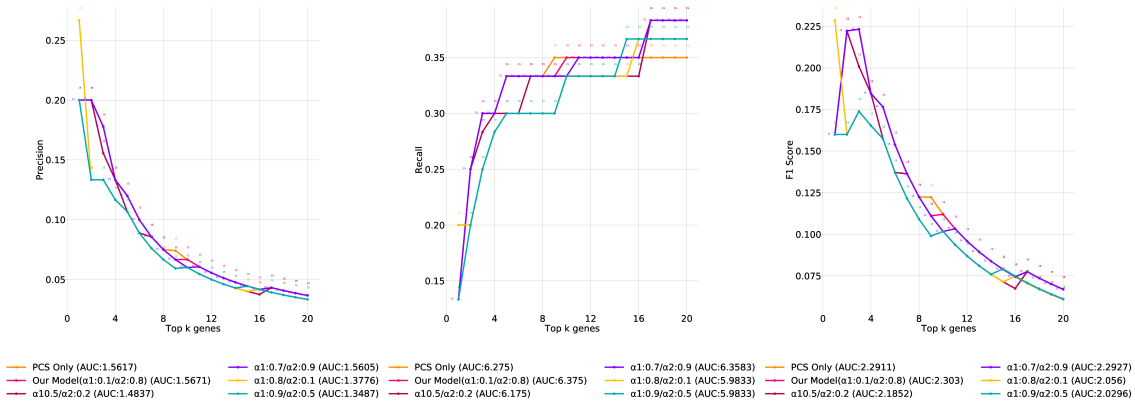

Figure B.16: Drug Targets and Neighbors Common average precision, recall, F1 across all cell lines for LUAD with different alpha values

Bibliography

[1] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, pp. 719–724, 04 2009.

[2] L. A. Garraway and E. S. Lander, "Lessons from the cancer genome.," *Cell*, vol. 153, pp. 17–37, Mar 2013.

[3] M. R. Stratton, "Exploring the genomes of cancer cells: progress and promise.," *Science*, vol. 331, pp. 1553–1558, Mar 2011.

[4] M. Shatnawi, "Chapter 6 - review of recent protein-protein interaction techniques," in *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology* (Q. N. Tran and H. Arabnia, eds.), Emerging Trends in Computer Science and Applied Computing, pp. 99 – 121, Boston: Morgan Kaufmann, 2015.

[5] J. Rolland, F. L. Condamine, F. Jiguet, and H. Morlon, "Faster speciation and reduced extinction in the tropics contribute to the mammalian latitudinal diversity gradient.," *PLoS Biol*, vol. 12, p. e1001775, Jan 2014.

[6] S. Skiena, *Implementing discrete mathematics: combinatorics and graph theory with Mathematica, Steven Skiena. Pp 334. 1990. ISBN 0-201-50943-1 (Addison-Wesley)*. Addison-Wesley, 1991.

[7] A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, D. G. Huntsman, C. Caldas, S. A. Aparicio, and S. P. Shah, "Drivernet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer," *Genome Biology*, vol. 13, no. 12, p. R124, 2012.

[8] J. P. Hou and J. Ma, "Dawnrank: discovering personalized driver genes in cancer," *Genome Medicine*, vol. 6, no. 7, p. 56, 2014.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[10] W.-F. Guo, S.-W. Zhang, L.-L. Liu, F. Liu, Q.-Q. Shi, L. Zhang, Y. Tang, T. Zeng, and L. Chen, "Discovering personalized driver mutation profiles of single samples in cancer by network control strategy," *Bioinformatics*, vol. 34, pp. 1893–1903, 01 2018.

[11] G. Dinstag and R. Shamir, "PRODIGY: personalized prioritization of driver genes," *Bioinformatics*, vol. 36, pp. 1831–1839, 11 2019.

[12] TCGA, "The cancer genome atlas research network." https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga, 2019.

[13] CCLE, "Cancer cell line encyclopedia." https://depmap.org/portal/, 2019.

[14] "BioinformaticsFMRP/TCGAbiolinks," July 2020. original-date: 2015-03-02T14:56:03Z.

[15] DepMap, "Depmap 20q1 public." https://depmap.org/portal/, 2020.

[16] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Research*, vol. 40, pp. D109–D114, 11 2011.

[17] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering, "STRING v10: protein–protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, pp. D447–D452, 10 2014.

[18] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, "The cosmic cancer gene census: describing genetic dysfunction across all human cancers," *Nature Reviews Cancer*, vol. 18, no. 11, pp. 696–705, 2018.

[19] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, D. A. Haber, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett, "Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells.," *Nucleic Acids Res*, vol. 41, pp. D955–61, Jan 2013.

[20] J.-P. Gillet, S. Varma, and M. M. Gottesman, "The Clinical Relevance of Cancer Cell Lines," *JNCI: Journal of the National Cancer Institute*, vol. 105, pp. 452–458, 02 2013.

[21] A. Marusyk, V. Almendro, and K. Polyak, "Intra-tumour heterogeneity: a looking glass for cancer?," *Nat Rev Cancer*, vol. 12, pp. 323–334, Apr 2012.

[22] V. K. Yadav and S. De, "An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples.," *Brief Bioinform*, vol. 16, pp. 232–241, Mar 2015.

[23] X. L. Peng, R. A. Moffitt, R. J. Torphy, K. E. Volmar, and J. J. Yeh, "De novo compartment deconvolution and weight estimation of tumor samples using decoder," *Nature Communications*, vol. 10, Oct 2019.

[24] Q. Liu, M. J. Ha, R. Bhattacharyya, L. Garmire, and V. Baladandayuthapani, "Network-based matching of patients and targeted therapies for precision oncology*," *bioRxiv*, 2019.

[25] P. Geeleher, N. J. Cox, and R. S. Huang, "Clinical drug response can be predicted using baseline gene expression levels and in vitrodrug sensitivity in cell lines," *Genome Biology*, vol. 15, no. 3, p. R47, 2014.

[26] J.-K. Lee, Z. Liu, J. K. Sa, S. Shin, J. Wang, M. Bordyuh, H. J. Cho, O. Elliott, T. Chu, S. W. Choi, D. I. S. Rosenbloom, I.-H. Lee, Y. J. Shin, H. J. Kang, D. Kim, S. Y. Kim, M.-H. Sim, J. Kim, T. Lee, Y. J. Seo, H. Shin, M. Lee, S. H. Kim, Y.-J. Kwon, J.-W. Oh, M. Song, M. Kim, D.-S. Kong, J. W. Choi, H. J. Seol, J.-I. Lee, S. T. Kim, J. O. Park, K.-M. Kim, S.-Y. Song, J.-W. Lee, H.-C. Kim, J. E. Lee, M. G. Choi, S. W. Seo, Y. M. Shim, J. I. Zo, B. C. Jeong, Y. Yoon, G. H. Ryu, N. K. D. Kim, J. S. Bae, W.-Y. Park, J. Lee, R. G. W. Verhaak, A. Iavarone, J. Lee, R. Rabadan, and

D.-H. Nam, “Pharmacogenomic landscape of patient-derived tumor cells informs precision oncology therapy.,” *Nat Genet*, vol. 50, pp. 1399–1411, Oct 2018.

[27] G. Jiang, S. Zhang, A. Yazdanparast, M. Li, A. V. Pawar, Y. Liu, S. M. Inavolu, and L. Cheng, “Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer,” *BMC Genomics*, vol. 17, no. 7, p. 525, 2016.

[28] H. Li, J. S. Wawrose, W. E. Gooding, L. A. Garraway, V. W. Y. Lui, N. D. Peyser, and J. R. Grandis, “Genomic analysis of head and neck squamous cell carcinoma cell lines and human tumors: A rational approach to preclinical model selection,” *Molecular Cancer Research*, vol. 12, no. 4, pp. 571–582, 2014.