DISCOVERING CROSS-CANCER PATIENTS WITH A SEMI-SUPERVISED DEEP CLUSTERING APPROACH

by
DUYGU AY

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Master of Science

Sabancı University
August 2020

DISCOVERING CROSS-CANCER PATIENTS WITH A SEMI-SUPERVISED DEEP CLUSTERING APPROACH

Approved by:

Asst. Prof. Öznur Taştan Okan .....
(Thesis Supervisor)

Assoc. Prof. Cem İyigün .....

Asst. Prof. Kamer Kaya .....

Date of Approval: August 25, 2020

ABSTRACT

DISCOVERING CROSS-CANCER PATIENTS WITH A SEMI-SUPERVISED DEEP CLUSTERING APPROACH

DUYGU AY

Computer Science and Engineering, Master's Thesis, August 2020

Thesis Supervisor: Asst. Prof. Öznur Taştan Okan

Keywords: Cancer, Deep learning , Semi-supervised clustering, Patient similarity

In traditional medicine, the treatment decisions for a cancer patient are typically based on the patient's cancer type. The availability of molecular profiles for a large cohort of multiple cancer patients opens up possibilities to characterize patients at the molecular level. There have been reports of cases where patients with different cancers bear similarities. Motivated from these observations, in this thesis, we specifically focus on developing a method to discover cross-cancer patients. We define cross-cancer patients as those who have molecular profiles that bear a high level of similarity to other patient(s) diagnosed with a different cancer type and are not representative of their cancer type. To find cross-cancer similar patients, we develop a framework where we identify patients that co-cluster frequently when clustered based on their transcriptomic profiles. To solve the clustering problem, we propose a semi-supervised deep learning clustering in which the clustering task is guided by the cancer types of the patients and the survival times. The deep representation obtained in the network is used in the clustering module of DeepCrossCancer. Applying the method to nine different cancers from The Cancer Genome Atlas project using patient tumor gene expression data, we discover twenty patients similar to a patient or multiple patients in another cancer type. We analyze these patients in light of other genomic alterations. Our results find significant similarities both in mutation and copy number variations of the cross-cancer patients. The detection of cross-cancer patients opens up possibilities for transferring clinical decisions from one patient to another and expediting the investigation of novel cancer drivers shared among them. The method is available at `https://github.com/Tastanlab/DeepCrossCancer`.

ÖZET

YARI DENETIMLI DERIN KÜMELEME YAKLAŞIMIYLA ÇAPRAZ KANSER HASTALARININ BELIRLENMESI

DUYGU AY

Bilgisayar Bilimi ve Mühendisliği, Yüksek Lisans Tezi, Ağustos 2020

Tez Danışmanı: Asst. Prof. Öznur Taştan Okan

Anahtar Kelimeler: Kanser, Derin Öğrenme, Yarı Gözetimli Öbekleme, Hasta Benzerliği

Geleneksel tıpta, bir kanser hastasının tedavi kararları tipik olarak hastanın kanser türüne dayanır. Çok sayıda kanser hastasından oluşan geniş bir kohort için moleküler profillerin mevcudiyeti, hastaları moleküler düzeyde karakterize etmek için olanaklar sağlar. Farklı kanser hastalarının benzerlikler taşıdığı vakalar önceki çalışmalarda bildirilmiştir. Bu gözlemlerden motive olarak, bu tezde, özellikle çapraz kanser hastalarını keşfetmek için bir yöntem geliştirmeye odaklanıyoruz. Çapraz kanser hastalarını, farklı bir kanser türü ile teşhis edilen diğer hasta(lar) ile yüksek düzeyde benzerlik taşıyan ve kendi kanser türünü temsil etmeyen moleküler profillere sahip hastalar olarak tanımlıyoruz. Çapraz kanser benzeri hastaları bulmak için, transkriptomik profillerine göre kümelendiğinde sık sık birlikte kümelenen hastaları belirlediğimiz bir çerçeve geliştiriyoruz. Bu kümeleme problemini çözmek için, kümeleme görevinin hastaların kanser türleri ve hayatta kalma süreleri tarafından yönlendirildiği yarı denetimli bir derin öğrenme kümeleme yöntemi öneriyoruz. Bu yöntem ile elde edilen derin temsil, DeepCrossCancer'ın kümeleme modülünde kullanılır. Bu yöntemi, hasta tümör gen ekspresyon verilerinin kullanıldığı Kanser Genom Atlas projesinden dokuz farklı kansere uygulayarak, başka bir kanser türünde bir hastaya veya birden fazla hastaya benzer yirmi hasta keşfediyoruz. Bu hastaları diğer genomik değişikliklerin ışığında analiz ediyoruz. Sonuçlarımız, çapraz kanser hastalarının hem mutasyon hem de kopya sayısı varyasyonlarında önemli benzerlikler bulmaktadır. Çapraz kanser hastalarının tespiti, klinik kararların bir hastadan diğerine aktarılması ve aralarında paylaşılan yeni kanser sürücülerinin araştırılmasını hızlandırmak için olanaklar sağlar. Yöntem şu bağlantıda mevcuttur: `https://github.com/Tastanlab/DeepCrossCancer`.

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor Asst. Prof. Öznur Taştan Okan. It would not be possible to complete this thesis without her motivation, constant support, trust in me, and especially her understanding of everything. I also thank Asst. Prof. Kamer Kaya and Assoc. Prof. Cem İyigün for their presence in the thesis jury.

I thank my favorite lab friend Yasin for his help. I also thank all my friends at Sabancı during my graduate study; Simge, Polen, Elif, Ece, Pınar, Yunus, Hasan, Ömer, and others. We were always together in this difficult process and we always gave each other motivation.

In addition, I would like to thank my BFFs Perihan, Rana, Elif Cansu, Özge, Tuna, Ebrar, Hande, and Nurdan for their valuable supports and love. They were behind every decision I made. Especially, I would like to thank Rana, 5 years old roommate and one of my BFFs, for her emotional support. She was always with me while struggling with my master's degree.

Finally, I would like to thank my family for their support throughout my entire life. Most importantly, I'm very grateful to my sisters Müşerref and Fatoş, my brother-in-law İskender, and my little nephew Doruk for their lovely motivation.

...This thesis is dedicated to positive change in a world full of opportunity...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Cancer cells exhibit numerous genomic alterations compared to normal cells. These changes differ widely across patients, and patients diagnosed with the same cancer type typically bear different sets of molecular changes in their tumor cells. This heterogeneity sets significant challenges for designing effective diagnostic and treatment strategies that would work across all patients of a cancer type. With the large-scale cancer genome sequencing projects, it became possible to chart the tumor's molecular landscape. The molecular profiles for large cohorts of cancer patients have opened up possibilities for developing more precise diagnostic and therapeutic tools. Characterizing the alterations in the cancer cells also enhances the ability to understand the molecular underpinnings of tumor development and progression, which can also inform the clinical management. Here are two main research directions are undertaken that rely on the analysis of this molecular data that serves these goals.

In the first research direction, molecular subtypes of the same cancer type are sought after to dissect the heterogeneity observed within a cancer type. As these subtypes have different disease etiologies, responses to therapy, and clinical outcomes, the ultimate goal is to design treatment regimens tailored for each subgroup. The identification of breast cancer intrinsic molecular subtypes discovered almost two decades ago by analyzing gene expression profiles of cancer patients is an example of such an approach (Perou, Sørlie, Eisen, Van De Rijn, Jeffrey, Rees, Pollack, Ross, Johnsen, Akslen & others, 2000; Sotiriou, Neo, McShane, Korn, Long, Jazaeri, Martiat, Fox, Harris & Liu, 2003). These molecular subtypes have been used in the treatment of breast cancer patients. More recently, other subtypes have been suggested by analysis of recent larger cohorts of breast patients and other types of omic profiles (Ali, Rueda, Chin, Curtis, Dunning, Aparicio & Caldas, 2014). Similar molecular subtyping efforts have been undertaken for other cancer types (Abeshouse, Ahn, Akbani, Ally, Amin, Andry, Annala, Aprikian, Armenia, Arora & others, 2015; Al-

izadeh, Eisen, Davis, Ma, Lossos, Rosenwald, Boldrick, Sabet, Tran, Yu & others, 2000; Network & others, 2011,1,1; Tepeli, Ünal, Akdemir & Tastan, 2020; Verhaak, Hoadley, Purdom, Wang, Qi, Wilkerson, Miller, Ding, Golub, Mesirov & others, 2010; Yeoh, Ross, Shurtleff, Williams, Patel, Mahfouz, Behm, Raimondi, Relling, Patel & others, 2002).

The second main research direction is to conduct a pan-cancer analysis on cancer patient-derived molecular data spearheaded by the Pan-Cancer consortium (Weinstein, Collisson, Mills, Shaw, Ozenberger, Ellrott, Shmulevich, Sander, Stuart, Network & others, 2013). The goal here is to reclassify human tumor types based on their molecular similarity and get a unified view of multiple types of cancer on commonalities and differences with the ultimate goal of improving patient outcomes. To this end, the Pan-Cancer Genome Atlas project performed an integrative molecular analysis using multiple types of omic data from 33 different tumor types (Hoadley, Yau, Hinoue, Wolf, Lazar, Drill, Shen, Taylor, Cherniack, Thorsson & others, 2018) and using a clustering approach, (Shen, Mo, Schultz, Seshan, Olshen, Huse, Ladanyi & Sander, 2012; Shen, Olshen & Ladanyi, 2009) and arrived at 28 distinct molecular subtypes. In other diseases too, such global cross-disorder analysis has been conducted. A study by the Psychiatric Genomic Consortium (PGC) Cross-Disorders Group provided the first genome-wide evidence that risk loci are shared between five psychiatric disorders (autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia) treated as distinct categories in clinical practice (Cross-Disorder Group of the Psychiatric Genomics Consortium and others, 2013). With the findings of cross-disorder genetic risk factors, a recent study of the PsychENCODE Consortium set out to decipher the molecular mechanisms underlying psychiatric disorders by using gene expression data (Wang, Liu, Warrell, Won, Shi, Navarro, Clarke, Gu, Emani, Yang & others, 2018).

In this work, we focus on a third strategy to facilitate patient-specific clinical decisions to identify cross-cancer patients, which bear highly molecular similarity to a single or multiple cancer patients in another cancer type. This approach is different from the subtype discovery efforts aforementioned because it analyzes patients across cancers. It is also different from the pan-cancer analysis approach because instead of finding global similarities across a group of patients, it seeks more patient-specific similarities for a single patient that could be missed in a pan-cancer study to the small group size. The benefit of such an approach is two folds. If there are actionable genomic events, the detection of cross-cancer patients opens up possibilities for transferring clinical decisions from one patient to the other one immediately. Secondly, patients in different cancer types with these unexpected molecular similarities

can discover novel cancer-driving mechanisms.

Cross-cancer patient genomic similarities have been reported in the literature. A TCGA analysis had revealed that the subtype of breast cancer – the basal-like – bear extensive molecular similarities to high-grade serous ovarian cancer, which has been hard to treat (Network & others, 2012). Similarly, the results on endometrial carcinomas of TCGA demonstrated that 25% of 373 tumors studied that have been classified as high-grade endometriosis by pathologists have molecular similarities to uterine serous carcinomas (Levine, Network & others, 2013). In this work, we will even look into finer similarities.

The contributions in this thesis are three folds. The first contribution is that we develop a novel method to identify cross-cancer patients, patients with high molecular similarities to patients other than their cancer type. This method takes the transcriptomic data of tumors biopsied from patients from multiple cancer types and returns cross-cancer patients. The method relies on repeatedly clustering patients and finding patients that always co-cluster. The clustering step is based on a semi-supervised clustering approach. The second contribution of this thesis is in this clustering step. We extend an existing deep learning-based clustering method (Chen, Yang, Goodison & Sun, 2020) by adding a survival module. Our proposed model is trained to achieve three tasks jointly: cancer type classification, survival prediction, and clustering of patients. Although the ultimate aim is to reach good clusters of the patients, solving the auxiliary tasks of cancer type classification and the survival prediction serve to learn a good representation of the patients. The third contribution of this thesis is that upon applying the model on the Cancer Genome Atlas project data, we identify 20 cross-cancer patients. We inspect these patients in the light of other genomic data available such as somatic mutations and copy number variations and find interesting common genomic events. These are hypotheses to be tested for further experimental verification.

The outline of this thesis is as follows: In Chapter 2, we first review the related work and provide background information to understand the model presented and tools used. Then in Chapter 3, we introduce our novel model for identifying cross-cancer patients. In Chapter 4, we provide empirical results obtained with our proposed algorithms and provide a detailed analysis of the identified cross-cancer similarities using complementary omics data of the patients. We conclude our work and discuss future work in Chapter 5.

Chapter 2

RELATED WORK AND BACKGROUND

In this chapter, we review the related work elaborately. First, we will cover techniques for cancer subtype identification. Clustering analyzes are widely used to identify novel subtypes of cancer. We will not cover all the clustering techniques for cancer subtype identification since our main aim is not finding cancer subtypes. Next, we will elaborate on the second main research direction which is pan-cancer anaylsis. In Section 2.3, the studies on patient similarity will be analyzed. In Section 2.4, we will cover deep clustering methods for clustering cancer patients. Finally, the study of Deep SHAP for the interpretation of deep learning models will be analyzed.

2.1 Techniques for Cancer Subtype Identification

A number of different clustering methods have been used in the context of genomic studies. The most widely known clustering techniques for the identification of cancer subtypes are K-means clustering, Hierarchical clustering, Consensus clustering, and Non-negative matrix factorization(NMF). In this section, we will briefly explain these clustering methods.

2.1.1 K-means Clustering

K-means Clustering (Lloyd, 1982) is one of the most used clustering algorithm due to its simplicity. Given the specified number of clusters K, the centroids of each cluster are initialized by randomly selecting K data points after shuffling the dataset.

Then, the algorithm iterates between two steps: all data points are assigned to the nearest centroids by calculating the sum of the squared distance between data points and all centroids, and compute the new centroids by taking the average of all data points in each cluster. The algorithm stops when the assignment of data points or the centroids is no further changing. In the studies of clustering cancer data, K-means clustering algorithms have been used successfully. It is proven by a comparison study (de Souto, Costa, de Araujo, Ludermir & Schliep, 2008). They compared seven different types of clustering algorithms on the analysis of 35 cancer gene expression data: hierarchical clustering with single, complete, average linkage, k-means, a mixture of multivariate Gaussians, spectral clustering, and shared nearest neighbor-based clustering. In this study, k-means has been reported as one of the best algorithm in terms of recovering the actual structure of data sets despite its disadvantages such as non-deterministic feature. K-means is widely used in gene expression data analysis (Quackenbush, 2001; Slonim, 2002).

2.1.2 Hierarchical Clustering

Hierarchical clustering (Johnson, 1967) is a clustering algorithm that produces a nested set of clusterings in a tree-like hierarchy and then creates actual clusters by cutting the dendrogram at a certain height. The tree is split into several branches, and the data points in each branch form a cluster. There are two types of hierarchical clustering: Agglomerative and divisive. The agglomerative technique is also known as the bottom-up technique. The clusters are formed from the bottom starting with individual data points and merge the closest data points until the desired number of clusters formed. The divisive technique (top-down) is not much used for cancer subtype identification. In this technique, the clusters are formed from the top starting with whole data as one cluster and it is divided into the desired number of clusters according to the dissimilarity of data points.

The choice of distance metrics is important in the hierarchical clustering method and it is named as linkage techniques. Single, complete, and average linkage methods are the common ones. In the single linkage, the similarity between clusters is calculated between two closest data points one from a cluster and one from another cluster. Complete linkage works opposite to the single linkage method. The similarity is calculated between two farthest points. In the average linkage method, the distances between all data points between two clusters are calculated and the average of them is taken.

Hierarchical clustering has been used successfully in the studies of tumor subtype

identification. (Bhattacharjee, Richards, Staunton, Li, Monti, Vasa, Ladd, Beheshti, Bueno, Gillette & others, 2001) identified distinct subclasses of lung adenocarcinoma by applying hierarchical clustering to gene expression data. Distinct types of diffuse large B-cell lymphoma were identified by hierarchical clustering on the gene expression data (Alizadeh et al., 2000). There are also other studies that have been used the hierarchical clustering method for tumor subtype identification in other diseases (Beer, Kardia, Huang, Giordano, Levin, Misek, Lin, Chen, Gharib, Thomas & others, 2002; Eisen, Spellman, Brown & Botstein, 1998).

2.1.3 Consensus Clustering

Consensus clustering (Monti, Tamayo, Mesirov & Golub, 2003) relies on cluster ensembles that aggregate clustering information coming from multiple iterations of algorithms based on the resampling of the dataset. First, the subset of samples is selected and K-means clustering is performed. The original data is classified based on the results after every iteration. The results of each iteration are combined in a consensus matrix that shows the pairwise similarity of samples, how many times were they in the same cluster. The consensus matrix can be used by any clustering algorithm that uses input as a similarity matrix such as spectral clustering.

Since consensus clustering is less sensitive to noise or outliers in the data, it enables us to obtain biologically robust clusters. By consensus clustering on gene microarray data, distinct clear cell renal cell carcinoma subtypes are revealed (Brannon, Reddy, Seiler, Arreola, Moore, Pruthi, Wallen, Nielsen, Liu, Nathanson & others, 2010). Damrauer, Hoadley, Chism, Fan, Tiganelli, Wobker, Yeh, Milowsky, Iyer, Parker & others (2014) identified two intrinsic, molecular subsets of high-grade bladder cancer by performing consensus clustering on gene expression data. Another study is that three subtypes of gastric cancer are obtained by using consensus hierarchical clustering with iterative feature selection on gene expression patterns (Lei, Tan, Das, Deng, Zouridis, Pattison, Chua, Feng, Guan, Ooi & others, 2013).

2.1.4 Non-Negative Matrix Factorization

The non-negative matrix factorization (NMF) (Lee & Seung, 2001) is a dimension reduction method and used for clustering and classification. Consider a matrix X with dimension n by m which is a non-negative matrix. This matrix is factorized into two non-negative matrices W and H in the way that W is n by K and H is K by m where n, m and K represent the number of samples, number of genes

and number of clusters respectively. To achieve this, we need to solve the following minimization problem.

$$\min_{\boldsymbol{W}\geq 0,\boldsymbol{H}\geq 0} f(\boldsymbol{W},\boldsymbol{H}) = \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\|^2$$

NMF has been used successfully in discovering molecular profiles for high-dimensional genomic data. Brunet, Tamayo, Golub & Mesirov (2004) found meaningful cancer subtypes in a leukemia study by applying NMF to gene expression dataset. NMF is also used for integrative analysis of multiple types of genomic data in cancer subtype identification (Zhang, Liu, Li, Shen, Laird & Zhou, 2012).

2.2 Pan-cancer Analysis

The other main research direction is to conduct a pan-cancer analysis on cancer patient-derived molecular data. Many studies are designed for finding subgroups of the same disease type although patients might bear key similarities across cancers. However, pan-cancer researches have become possible by integrating the datasets around a specific type of cancer into a single analysis. We presented the related two works in the following sections. These studies generally focus on similarities across cancer rather than focusing on patient similarity within the same cancer type. This motivated us to find patient-specific similarities across cancers.

2.2.1 Network-based Pan-cancer Stratification Approach

Network-based Stratification (NBS) is proposed by Hofree, Shen, Carter, Gross & Ideker (2013) to integrate somatic tumor genomes with gene networks. They used somatic mutation profiles of patients indicating binary states on genes $(0, 1)$ in which the state of 1 occurs if any mutation of a gene occurred in the patient otherwise the state is 0. For each patient, they constructed a gene interaction network and applied network propagation technique (Vanunu, Magger, Ruppin, Shlomi & Sharan, 2010) to spread the effect of mutations to the neighboring genes by smoothing the states on genes. Following the network smoothing, NMF is applied to find the subgroups of ovarian, uterine, and lung cancer according to high network connectivity. These steps are repeated for N different subsamples to obtain robust clusters. The results are aggregated by constructing a patient-by-patient co-occurrence matrix and consensus clustering is applied on the matrix.

NBS is used in the study that finds cross-cancer indications in 12 cancer types since it is suitable for integrating multiple data types into a single analysis (Liu & Zhang, 2015). The 12 cancer types include bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon and rectum adenocarcinoma (COAD, READ), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear-cell carcinoma (KIRC), acute myeloid leukemia (LAML), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), and uterine corpus endometrioid carcinoma (UCEC). They used the selected functional events (SFEs) binary data of copy number alterations, somatic mutations, and DNA hyper-methylation data types. After transforming the functional genetic changes to genes, they projected the binary data to gene interaction networks and applied the NBS approach. They obtained the 9 pan-cancer subgroups that imply important cross-cancer commonalities without considering the primary tumor organ information. For example, LAML and UCEC were clustered in the same pan-cancer group, and the subsets of GBM, BLCA, LUSC, and HNSC tumors dropped in the same pan-cancer group. They easily show pan-cancer heterogeneity with subgroup-specific gene network characteristics and biological functions.

2.2.2 Pan-cancer Atlas Integrative Analysis

Recently, Hoadley et al. (2018) conducted the most comprehensive cross-cancer analysis up to date. They run iCluster (Shen et al., 2009) on the datasets of chromosome-arm-level aneuploidy, DNA hypermethylation, mRNA, and miRNA expression levels and reverse-phase protein arrays of approximately $10,000$ patient samples from 33 cancer types. iCluster works with the appealing approach of integrative cluster analysis, which includes the variable selection feature. The approach combines multiple data types from the same patient samples simultaneously through a latent variable explaining the correlations across the data types. Hoadley et al. (2018) identified 28 distinct molecular subtypes from 33 different tumor types. By dominating molecular classification, organ and cell-of-origin patterns affect iCluster groupings. They also rationalised several pan-cancer analysis that are based on organ systems such as, pan-gastrointestinal (Liu, Sethi, Hinoue, Schneider, Cherniack, Sanchez-Vega, Seoane, Farshidfar, Bowlby, Islam & others, 2018), pan-gynecological (Berger, Korkut, Kanchi, Hegde, Lenoir, Liu, Liu, Fan, Shen, Ravikumar & others, 2018), pan-kidney (Ricketts, De Cubas, Fan, Smith, Lang, Reznik, Bowlby, Gibb, Akbani, Beroukhim & others, 2018), and pan-squamous (Campbell, Yau, Bowlby, Liu, Brennan, Fan, Taylor, Wang, Walter, Akbani & others, 2018). Results show that there exists genomic, epigenomic, and transcriptomic similarities and differences across

cancer types.

2.3 Patient Similarity Tools

Discovering the similarity between patients is important for the development of personalized patient care for precision medicine. Each patient has unique data and can be different from other patients. For example, a patient's clinical outcome is revealed as breast cancer but somehow the patient might be similar to a patient from kidney cancer in terms of genomics profile. Patient similarity tools can reveal these types of patients and interpret this similarity.

2.3.1 Patient Similarity Networks

Patient similarity network (PSN) (Pai & Bader, 2018) is a recently developed framework that is used for clustering and classification by integrating multiple data types. In PSN, patients are connected according to their similarities for each data feature, e.g. age, sex, mutation status. Each node in the graph represents a patient, and edges between patients represent pairwise similarity for one feature. The thickness of edges shows the amount of similarity between patients.

PSN frameworks have lots of advantages in terms of interpretability, handling with heterogeneous data, processing missing information, and protecting patient privacy. PSNs are easily interpretable because it represents the data into networks where the decision boundaries can be visible. By looking at the graphs, we can easily found similar patients to an indexed patient. Secondly, PSNs can discover latent factors by integrating multiple data types. Any data type can be converted into a network by determining a similarity measure. Therefore, it will be easy to handle missing information. If patient information is missing in terms of clinical data type, we can still use that patients' information in other data types by integrating the networks. PSNs also help in protecting patient privacy. With PSNs, we do not need to store the raw data. It can be stored as graphs.

PSNs have been used in studies of disease subtype identification. The first example of that is the subgroup identification of patients with type 2 diabetes (Li, Cheng, Glicksberg, Gottesman, Tamler, Chen, Bottinger & Dudley, 2015). It provides the utility and promise of applying the precision medicine paradigm. The authors built a patient-patient similarity network based on 73 clinical features such as laboratory tests and gender from electronic medical records. As for similarity distance measure

metrics, singular value decomposition and cosine similarity were been used. The authors demonstrated that identified patient clusters are improved on different comorbidities and biological pathways by means of medical records and genotype data on the same people.

2.3.1.1 Similarity Network Fusion

Similarity network fusion (SNF) (Wang, Mezlini, Demir, Fiume, Tu, Brudno, Haibe-Kains & Goldenberg, 2014) is a clustering algorithm that uses PSNs. PSNs are constructed for each input data type, e.g. mRNA expression, DNA methylation, and miRNA expression. As the similarity measure metrics, they used euclidean distance scaled by exponential similarity kernel for continuous variables, the chi-squared distance for discrete variables, and agreement-based measure for binary variables. The constructed PSNs are then combined by growing repeatedly the weights of the edges consistent with the other PSNs and reducing the weights of the edges containing on just some of the PSNs, but not on all of them. This process continues until it converges to a single similarity network and the network summarizes the similarity between the samples in all data types. Finally, this network is cut into highly interconnected groups by spectral clustering.

Wang et al. (2014) proposed SNFs to identify patient subgroups in five tumors by integrating mRNA expression, DNA methylation, and miRNA expression. They demonstrated that SNFs outperform other approaches in terms of clinically distinct subgroup identification and the algorithm runs consistently fast no matter how many genes the input data includes. With the development of SNFs, they have been used in various studies of tumor subtype identification. The subtypes of medulloblastoma have been identified with SNFs by integrating DNA methylation and gene expression data (Cavalli, Remke, Rampasek, Peacock, Shih, Luu, Garzia, Torchia, Nor, Morrissy & others, 2017). The Cancer Genome Atlas Research Network has been identified proteomic subtypes of pancreatic cancer by applying SNF on RNA, DNA methylation, and miRNA expression data (Raphael, Hruban, Aguirre, Moffitt, Yeh, Stewart, Robertson, Cherniack, Gupta, Getz & others, 2017).

2.4 Deep Clustering Methods

Multi-omic data, which is made up of high-dimensional and complex structure, has made itself no longer applicable for conventional machine learning algorithms. Fortunately, deep learning can overcome these challenges. Deep clustering methods are

used to transform inputs into a new feature representation. The most widely known deep clustering techniques are Multi-layer Neural Networks, Deep Belief Networks, Autoencoders, and other deep learning architectures. In this section, we will briefly explain these deep clustering methods.

2.4.1 Multi-layer Neural Networks

Multi-layer Neural Network (MLP) is a classic feed-forward artificial neural network. MLP is made up of layers of neurons that are the core processing units. The neurons in each layer are connected with the neurons in the previous layers, and each connection has its own weight. MLP consists of an input layer, hidden layers, and an output layer. The input layer represents the feature matrix of the input data, hidden layers apply linear or non-linear transformations with activation functions. The output layer predicts the labels of data in the case of supervised learning. In the context of clustering, MLP is used for feature representation, especially with high-dimensional datasets. These learned features can be used for clustering in the case of unsupervised learning.

2.4.2 Deep Belief Networks

Deep Belief Networks (DBNs) (Hinton, Osindero & Teh, 2006) are generative models that contain both undirected layers and directed layers. DBNs are composed of a stack of Restricted Boltzmann machines (RBMs) (Hinton, 2012), where the hidden layer of one RBM is the visible layer one above it. A DBN is identical to an MLP in terms of network structure. But, they differ in the training process, where a DBN has trained two layers at a time, and every two layers act like an RBM. The output of the two layers is the input of the next two layers. The training process continues until the output layer. The most important thing related to DBNs is that each RBM layer learns the entire input. After the training process, the DBN is fine-tuned with the respective loss functions.

Liang, Li, Chen & Zeng (2014) proposed a multi-model deep belief network approach to discover subtypes of ovarian cancer by using gene expression, methylation, and miRNA data. They constructed separate hidden layers, each hidden layer has input from one data type, and the layers above gets the input from all the hidden layers. Then, they used a joint latent model to fuse common features from multiple data types and they applied the Contrastive Divergence (CD) learning algorithm in an unsupervised way. They discovered clinically distinctive eight subtypes of ovarian

cancer.

2.4.3 Autoencoders

Recent techniques use autoencoder which is a very common deep learning method as unsupervised learning for dimensionality reduction. Autoencoder (Bengio, Lamblin, Popovici & Larochelle, 2007) is a type of feed forward neural networks that consists of two parts: Encoder and Decoder. The input $\mathbf{x}$ is encoded to the representation layer $\mathbf{y}$ which is a bottleneck including a compressed information in the low dimensional space through mapping $\mathbf{y} = f_\theta(\mathbf{x})$. The decoder part reconstructs the input by minimizing the reconstruction error, $L(\mathbf{x}, \hat{\mathbf{x}})$, between the input and the output which is $\hat{\mathbf{x}} = g_{\theta'}(\mathbf{y})$. The minimization problem is formulated as follows (Vincent, Larochelle, Bengio & Manzagol, 2008):

$$\begin{aligned}\theta^\star, \theta'^\star &= \arg\min_{\theta,\theta'} \frac{1}{n} \sum_{i=1}^{n} L\left(\mathbf{x}^{(i)}, \hat{\mathbf{x}}^{(i)}\right) \\ &= \arg\min_{\theta,\theta'} \frac{1}{n} \sum_{i=1}^{n} L\left(\mathbf{x}^{(i)}, g_{\theta'}\left(f_\theta\left(\mathbf{x}^{(i)}\right)\right)\right)\end{aligned}$$

where n is the number of samples, θ is the weight between the input and bottleneck, and θ' is the weight between the bottleneck and reconstructed input.

Autoencoder is a powerful method for feature extraction and provides a new way of clustering by capturing non-linear structures on the representation layer. It is used for clustering by throwing away the decoder part. The raw data is encoded to the representation layer that outputs transformed data on a low dimensional space and a clustering algorithm, e.g. K-means, can be applied to the transformed data.

Multi-omic data types can be integrated and an autoencoder framework can be applied to the integrative dataset by achieving dimensionality reduction and capturing latent factors on the dataset. Recent studies focus on unsupervised and semi-supervised deep learning methods for cancer subtype identification by using multi-omic data. Chaudhary, Poirion, Lu & Garmire (2018) built an autoencoder framework that takes the integrated input of RNA sequencing (RNA-Seq), miRNA sequencing (miRNA-Seq), and methylation data for discovering robust survival subgroups of hepatocellular carcinoma (HCC). To achieve this, they did survival-associated feature selection on the bottleneck of autoencoder by using univariate Cox-PH models and applied K-means clustering on the new dataset. They identified two subgroups that have significant survival differences. Another study is about discovering the subtypes of high-risk neuroblastoma. To achieve this, they designed

an autoencoder framework by using gene expression and copy number alteration data and apply K-means on the bottleneck layer (Zhang, Lv, Jin, Cheng, Fu, Yuan, Tao, Guo, Ni & Shi, 2018).

2.4.4 Other Deep Learning Architectures

Convolutional neural networks (CNNs) and Generative Adversarial Network (GAN) are also used for deep clustering.

Convolutional neural networks (CNNs) (Krizhevsky, Sutskever & Hinton, 2012) were inspired by the organization of the animal visual cortex. Instead of neurons being connected to every neuron in the previous layer, they are only connected to neurons close to it, and every neuron uses the same weights. It is widely applied for image processing problems, and it treats input data in a spatial manner. They can be trained with a clustering loss.

Generative Adversarial Network (GAN) (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville & Bengio, 2014) is a type of deep generative model. A GAN is trained with two networks: artificial data samples that resemble training data and discriminative network that distinguishs between the artificial and the actual model. In the context of unsupervised learning, it learns the feature representation and conducts a specific clustering task.

2.5 Interpretation of Deep Learning Models: Deep SHAP

Deep learning models suffer from inherent challenges in determining the features to be used to predict labels. Because the information is overloaded in the neurons after the application of the activation functions which add non-linearity. Although the weights give an explanation about the input, the non-linearity makes it very difficult to decode.

Lundberg & Lee (2017) proposed Deep SHAP (Shapley Additive exPlanations) to interpret black-box deep learning models by using Shapley values in cooperative game theory. Deep SHAP is the combination of DeepLIFT (Deep Learning Important FeaTures) and SHAP methods. The main idea behind DeepLIFT is to explain the difference from some reference value of the output in terms of the difference from the reference value of the inputs (Shrikumar, Greenside & Kundaje, 2017).

DeepLIFT assigns contributions $C_{\Delta x_i \Delta t}$ to each input x_i such that:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t$$

where t is the output value, Δt is the difference from reference output, $\Delta x_i = x_i - r_i$, and r is the reference input. Deep SHAP uses Shapley values with the extension of DeepLIFT. Shapley values explain the contribution of an input feature to the difference between the predicted value and the average prediction value. From the Shapley values, Deep SHAP computes the contributions $C_{\Delta x_i \Delta t}$ of each input. The SHAP value evaluates the difference between the output made by including an indexed feature and the output made by all the combinations of features other than the indexed feature. There are a base value and an output value. The base value is just the value if any feature were not known. The output value is the prediction value of the actual model. SHAP values explain the contribution value of each feature to go from the base value to the output value.

Chapter 3

METHODS

This chapter describes the methodology to discover cross-cancer patients given a set of cancer patients from multiple cancer types. First, we will define the problem computationally. Next, we will explain the steps of the methodology in detail. We further provide details on how we use the interpretability tools to identify the predictive molecular features.

3.1 Problem Formulation

We define the cross-cancer patients as patients diagnosed with one cancer type but bear molecular similarities to patients diagnosed with another cancer type. Consider a set of n patients diagnosed with m different cancer types. Let y_i denote the cancer type for the patient i, where $y_i \in \{1, \cdots, m\}$. We deem patient i a cross-cancer patient if it follows the following conditions:

1.1 There is at least one patient j where i and j always co-clusters over multiple runs of clustering and for which $y_i \neq y_j$.

1.2 Patient i is closer to y_j members than its own cancer type y_i members.

Note for a pair of similar patients (i, j), not necessarily both of the patients will be cross-cancer patients due to the second condition above. While i is a cross-cancer patient, j might not be a cross-cancer patient because it represents its own cancer type.

To solve this problem, we propose DeepCrossCancer, which is composed of two main steps. The first step involves clustering the patients diagnosed with different cancer types using their molecular profiles and additional clinical annotation. We solve

the clustering step using a deep learning-based, semi-supervised clustering method, which we develop herein. The second step involves repeating this clustering procedure multiple times and identifying the patient samples consistently co-clustered with the patient sample(s) from another cancer type and finding the cross-cancer patients in the set of similar patient pairs. In the next section, we detail these steps.

3.2 Step 1 - Semi-supervised Deep Clustering

3.2.1 Preliminaries

We want to cluster n patient tumor samples using the samples' molecular profiles and additional information about the patient. We will use the terms patient and the patient's tumor sample interchangeably throughout the text. We denote i-th patient's feature vector with $\mathbf{x}_i \in \mathbb{R}^d$. In this work, the features are the gene expression data and the patient's age group. These features can be extended to incorporate other types of molecular and clinical information.

In the clustering step, the m patients presented with the feature vector $\mathbf{x}_i \in X$ are grouped into k disjoint clusters, each of which is represented by a centroid $\mathbf{u}_j, j \equiv 1, \ldots, k$. $\mathbf{U}$ will denote the centroid matrix, where j-th column is the cluster centroid, $\mathbf{u}_j$. We will denote the cluster assignment of the i-th example with the k-dimensional vector $\mathbf{q}_i$, where $q_{ij} = 1$ if the i-th example belongs to the j-th cluster and 0 otherwise.

We learn a representation of the patients that can successfully predict the cancer type of a given patient and the survival time. Thus, classification and survival prediction tasks are solved jointly. Every sample is associated with a class label that denotes the diagnosed cancer type of the patient i, $y_i \in \{1, \cdots, m\}$, where m is the number of cancer types. We denote the patient's survival time with the i-th sample as t_i and the patient's survival status with c_i. c_i is 1 if the patient passes away and 0 if it is censored. Censored refers to the cases for which the patient's passing does not take place within the observation window.

The cancer patient data, D, can be summarized as $\mathcal{D} = \{\mathbf{x}_i, y_i, t_i, c_i, \mathbf{q}_i\}_{i=1}^{n}$. Here, the cluster membership vector $\mathbf{q}$ is unobserved. The problem we aim to solve in clustering is to uncover these assignments $\mathbf{q}$.

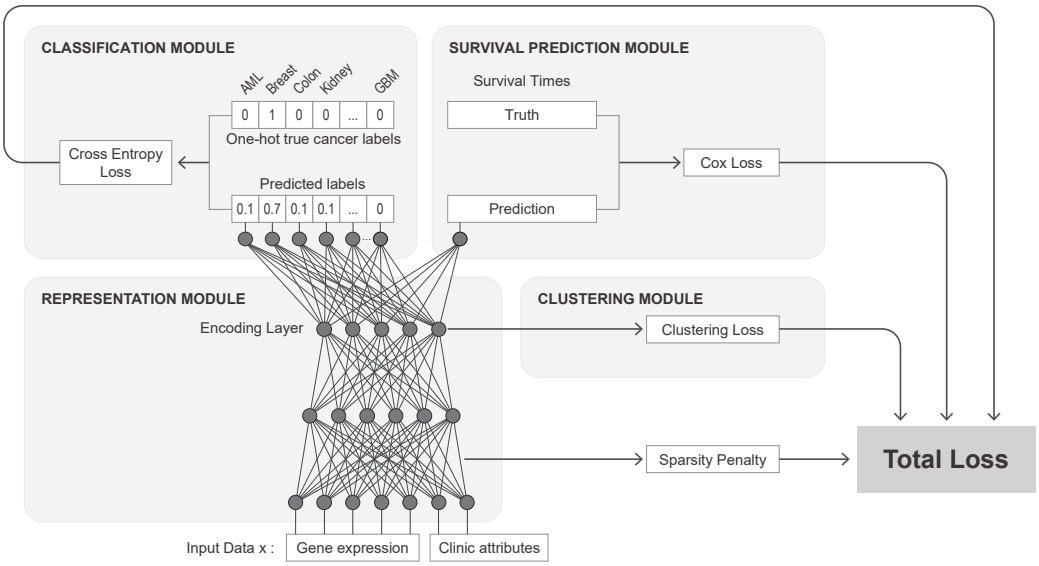


Figure 3.1 **Overview of DeepCrossCancer clustering network.** The network consists of four main components: representation, classification, survival prediction, and clustering modules. The representation module applies a nonlinear transformation on the input data and maps them into a lower-dimensional representation on the encoding layer. The representation module is guided with the classification and survival modules. The clustering module uses the representation provided in the encoding layer to group patients into k clusters.

3.2.2 DeepCrossCancer Clustering Architecture

DeepCrossCancer's network structure consists of an input representation module, a classification module, a survival module, and a clustering module (see Figure 3.1). While the classification module aims to categorize the patients into the correct cancer type, the survival module aims to predict the survival times of the patients accurately. The representation module takes the input, forms a nonlinear transformation of the data using multiple hidden layers, and projects it into a lower-dimensional space in the last hidden layer. This encoding layer is connected to the output layer for the classification and survival prediction. The deep encoding of the inputs is used in the clustering module. In this way, DeepCrossCancer clusters the patients with a learned representation of the inputs that can achieve classification and survival prediction. The network is optimized to accomplish these tasks jointly, in which we provide the details in Section 3.2.3.

The formulation and the architecture of the clustering network are built upon DeepType (Chen et al., 2020) which also uses representation, classification, and clustering modules. Different from DeepType, DeepCrossCancer's clustering network contains an additional survival module. Note also that the classification modules in these two

methods serve for two different purposes. While DeepType's classification module's goal is responsible for classifying the patients of the same cancer type into prior known subtypes, in DeepCrossCancer, the classification module focuses on classifying patients from multiple cancer types into the diagnosed cancer types of the patient.

The network used in DeepCrossCancer can be summarized as follows:

$$
\begin{aligned}
\mathbf{o}_1 &= \mathrm{ReLU}\left(\mathbf{W}_1\mathbf{X} + \mathbf{b}_1\right), \\
\mathbf{o}_1 &= \mathrm{ReLU}\left(\mathbf{W}_i\mathbf{o}_{i-1} + \mathbf{b}_i\right), 2 \leq i \leq M, \\
\hat{\mathbf{y}} &= \mathrm{softmax}\left(\mathbf{W}_{M+1}\mathbf{o}_M + \mathbf{b}_{M+1}\right), \\
\hat{\mathbf{h}} &= \mathrm{sigmoid}(\mathbf{W}_{M+1}\mathbf{o}_M + \mathbf{b}_{M+1}),
\end{aligned}
\tag{3.1}
$$

Here, the $\mathbf{W}_i$ is the weight matrix, $\mathbf{b}_i$ is the bias term and $\mathbf{o}_i$ is the output of the i-th layer. $\hat{\mathbf{y}}$ denotes the classification output and $\hat{\mathbf{h}}$ is the survival output. Θ denotes the learnable network parameters $\Theta = (\mathbf{W}, \mathbf{b})$. The RELU activation function (Nair & Hinton, 2010) is used in the hidden layers; while softmax activation and sigmoid activation functions are used for the classification and survival layers, respectively.

The network parameterized with Θ transforms points into a lower dimensional ($p \ll d$) *latent* feature space Z in $\mathbb{R}^p$ at the last hidden layer, $f_\Theta : X \to Z$. Instead of clustering directly in the original data space $X \in \mathbb{R}^d$, the clustering module uses this transformed representation of the inputs. The transformed data points, $\{z_i \in Z\}_{z=1}^n$ and the k cluster centers lie in $\{\mu_j \in Z\}_{j=1}^k$ in this latent feature space Z.

3.2.3 Network Optimization

The network is optimized jointly to achieve success in the three tasks using a joint supervised and unsupervised learning strategy. Supervised by the classification labels and the survival times, the network learns a representation that would lead to a latent space, Z, which will be useful in the unsupervised learning conducted for clustering. The network parameters are learned by minimizing an objective function that contains classification, clustering, survival losses, and a regularization term to enforce sparsity:

$$
\min_{\{\Theta, \mathbf{Q}, \mathbf{U}\}} L_{\text{classification}} + \alpha L_{\text{clustering}} + \beta L_{\text{survival}} + \lambda L_{\text{sparsity}}
\tag{3.2}
$$

$\mathbf{U}$ is the centroid matrix. Each column represents a cluster center and is hidden. The $\mathbf{Q}$ is the cluster membership matrix, $\mathbf{Q} = \left[\mathbf{q}^{(1)}, \cdots, \mathbf{q}^{(n)}\right]$, each row being one patient's assignment vector. The parameter λ is the regularization parameter that controls the model sparsity, and α and β are parameters that adjust the importance assigned to the clustering and survival losses relative to the classification loss.

We use the cross-entropy loss to quantify the discrepancy between the correct cancer type of the patient and the predicted cancer types of the patient as given below:

$$L_{\text{classification}} = -\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ji} \log \hat{y}_{ji} \tag{3.3}$$

We use the k-means (Lloyd, 1982) loss that quantifies the tightness of the clusters around their centroids:

$$L_{\text{clustering}} = \sum_{i=1}^{n} \|z_i - \mathbf{U}\mathbf{q}_i\|_2^2, \text{ subject to } \sum_{j=1}^{k} q_{ji} = 1, q_{ji} \in \{0,1\}, \forall j, \forall i, \tag{3.4}$$

The survival module follows the Cox partial likelihood model (Cox, 1972). For the prediction of the survival time of patients, we use the Cox loss as defined in (Katzman, Shaham, Cloninger, Bates, Jiang & Kluger, 2018) :

$$L_{\text{survival}} = \sum_{i:c^{(i)}=1} \left(\log \hat{\mathbf{h}}^{(i)} - \log \sum_{j:t^{(j)} \geq t^{(i)}} e^{\hat{\mathbf{h}}^{(j)}} \right) \tag{3.5}$$

Finally, as in (Chen et al., 2020) we also impose an $\ell_{2,1}$ regularization (Nie, Huang, Cai & Ding, 2010) on the weight matrix of the first layer to control the model complexity. The sparsity loss is defined as:

$$L_{\text{sparsity}} = \left\| \mathbf{W}_1^\top \right\|_{2,1} \tag{3.6}$$

The optimization problem should solve for $\mathbf{\Theta}$, the network parameters, and $\mathbf{U}$, the centroids of the clusters, and $\mathbf{Q}$, the assignment of the clusters simultaneously. Since they are coupled, as in DeepType, we employ an alternating minimization strategy. Initially, we ignore the clustering module by setting α to zero and pre-train the network to find an initial set of values for $\mathbf{\Theta}$ and the hyperparameters β and λ. We fix $\mathbf{\Theta}$ and calculate the transformed points, z_i $\forall i$, and using standard k-means algorithm finds the clusters; thus, $\mathbf{Q}$ and $\mathbf{U}$.

In the next step, we use the $\mathbf{Q}$ and $\mathbf{U}$ found in the previous step to optimize for θ

by minimizing the following loss:

$$\min_{\{\Theta\}} \quad L_{\text{classification}} + \alpha L_{\text{clustering}} + \beta L_{\text{survival}} + \lambda L_{\text{sparsity}} \tag{3.7}$$

We iterate these two steps alternatively until convergence. When training the network, we employ back-propagation by using the mini-batch based stochastic gradient descent method (Bottou, 2010).

3.3 Hyper-parameter Optimization

The loss function, as defined in Equation (3.2), is composed of different modules' losses. The trade-off parameters, α, β, and λ, needs to be optimized. Since a grid search strategy for these three parameters is computationally expensive, we first optimize β and λ by setting $\alpha = 0$. When optimizing β and λ, we use ten-fold stratified cross-validation on the training data. These procedures are described in Algorithms 1, 2 and 3.

Specifically, for each β value, we fix β and find the best λ value in each of the cycles of 10-fold cross-validation. For each fold, we pick the best λ by using the Talos optimization tool (Kotila, 2018). Talos is an open-source framework that performs hyperparameter optimization for Keras models. We use the random search with a probabilistic reduction optimization strategy provided in Talos. The strategy uses a probabilistic method to remove poorly performing parameter configurations from the search space by quantifying the decline in the specified reduction metric. We choose the reduction metric as the concordance index of survival time prediction. We obtain the average of the best λ values for each fold for a set β_l value, and we refer to this as λ_l^{avg} in Algorithm 1. In this 10-fold CV procedure to optimize λ, we also obtain the average classification error e_l^{avg} and the associated standard deviation of over the 10-folds σ_l (Line 6 in Algorithm 1). Using the one-standard-error rule (Hastie, Tibshirani & Friedman, 2009) the best λ_{l^*} value is picked for the β_l value. This procedure is repeated for each possible value of $\beta_l \in T = \{\beta_1, ..., \beta_L\}$. Next, we choose the optimal pair, (β^*, λ^*) using the one-standard-error (Steps 10-13 in Algorithm 1).

Once the optimal β and λ parameters are obtained, the deep learning model is pre-trained with these values and $\alpha = 0$. The pre-trained model m-th layer is used to transform the feature matrix $\mathbf{X}$ to $\mathbf{Z}$ and this is input to k-means algorithm to get the cluster centers, $\mathbf{U}$ and the cluster assignments $\mathbf{Q}$.

Secondly, we obtain the optimal β and λ, and train the entire model to find the optimal α for each number of clusters. Again by applying the one-standard-error rule (Hastie et al., 2009), we choose the optimal α values for each number of clusters. The pseudo-code of the proposed procedure is given in Algorithms 1, 2 and 3, and performs well in our numerical experience as shown in Figure 3.2.

Algorithm 1 Hyper-parameter optimization $(\mathcal{D}_{tr}, \mathbf{X}, \mathbf{Z}, A, B, T, k)$

Input: Training data $\mathcal{D}_{tr} = \{\mathbf{x}_i, y_i, t_i, c_i\}_{i=1}^{n_{tr}}$ (n_{tr} = the size of training data), $\mathbf{X}$, feature matrix, where i-th row is patient i's feature vector, $\mathbf{Z}$, transformed feature matrix at the m-th layer of the network, $A = \{\alpha_1, ..., \alpha_J\}$, $B = \{\lambda_1, ..., \lambda_L\}$, $T = \{\beta_1, ..., \beta_L\}$, number of clusters k .

Output: Optimized parameters $\alpha^*, \lambda^*, \beta^*$.

1: ———— **Optimize** β, λ ————
2: $\alpha \leftarrow 0$;
3: $E \leftarrow \emptyset$; // The set of average errors and standard deviations for each β in B
4: **for** $l = 1$ to L **do**
5: $\beta = \beta_l$;
6: $(e_l^{\text{avg}}, \sigma_l, \lambda_l^{\text{avg}}) =$ `OptimizeLambdawithTalosCV`$(\mathcal{D}_{tr}, B, \beta, \alpha)$; // 10-fold
7: $E = E \cup \{(e_l^{\text{avg}}, \sigma_l)\}$;
8: **end for**
9: ———— **Apply one-standard error rule** ————
10: Find the minimum avg classification error e_0 and one standard error σ_0 in E;
11: $l^* = \arg\max_{1 \le l \le L} l$, subject to $e_l^{\text{avg}} \le e_0 + \sigma_0$; // one-standard-error rule (Hastie et al., 2009)
12: $\beta^* = \beta_{l^*}$;
13: $\lambda^* = \lambda_{l^*}^{\text{avg}}$;
14: ————
15: $f_\Theta =$ `TrainNetwork` $(D_{tr}; \lambda^*, \beta^*, \alpha)$;
16: $\mathbf{Z} = f_\Theta(\mathbf{X})$;
17: $(\mathbf{Q}_0, \mathbf{U}_0) =$ `k-means` $(\mathbf{Z}, k)$;
18: $\alpha^* =$ `OptimizeAlpha`$(\mathcal{D}_{tr}, \mathbf{Q}_0, \mathbf{U}_0, A, \beta^*, \lambda^*, k)$;
19: **return** $(\lambda^*, \beta^*, \alpha^*)$

Algorithm 2 `OptimizeLambdawithTalosCV`

Input: $\mathcal{D}_{tr}$, $B = \{\lambda_1, ..., \lambda_L\}$, β, $\alpha = 0$.

Output: Average classification error e^{avg}, standard deviation of classification errors σ, average of optimal λ values λ^{avg}.

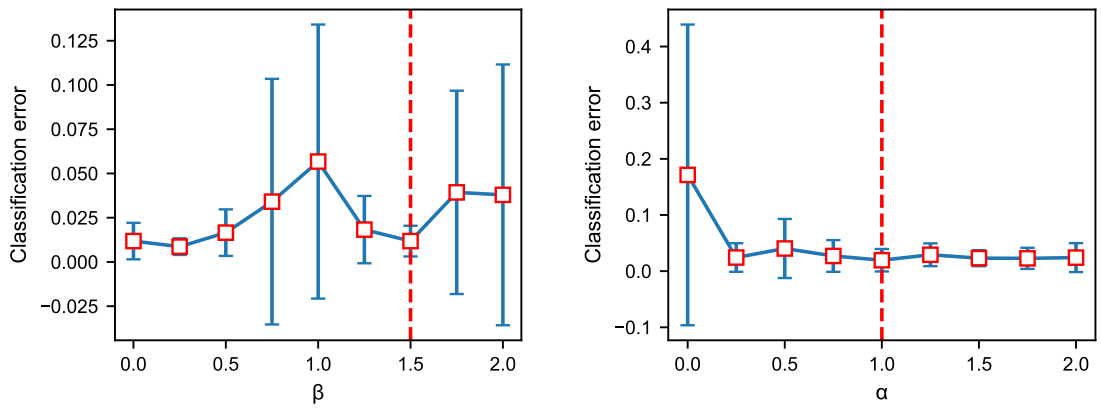
Randomly partition $\mathcal{D}_{tr}$ into ten folds;
for i= 1 **to** 10 **do**
    $(e_i, \lambda_i) =$ `OptimizeLambdawithTalos` $(\mathcal{D}_{tr}^{(i)}, B, \beta, \alpha)$;
    // gets optimal λ_i for fold i with Talos (Kotila, 2018) and the associated error.
end for
Compute average classification error e^{avg};
Compute standard deviation of classification errors σ;
Compute average of optimal lambda values λ^{avg};
return $(e^{\text{avg}}, \sigma, \lambda^{\text{avg}})$

Algorithm 3 `OptimizeAlphawithCV`

Input: $\mathcal{D}_{tr}$, $A = \{\alpha_1, ..., \alpha_J\}$, number of clusters k, β^* best β value, λ^* best λ value, $\mathbf{Q}_O$ cluster assignments obtained with the pretrained model, $\mathbf{U}_0$ cluster centroids obtained with the pretrained model.

Output: Best parameter α^*.

for $j = 1$ to J **do**
    $\alpha = \alpha_j$;
    $(e_j^{\text{avg}}, \sigma_j) =$ `10foldCV`$(\mathbf{U}_0, \mathbf{Q}_0, \alpha, \lambda^*, \beta^*,)$;
    // e_j^{avg} the average classification error over ten folds.
    // σ_j the standard deviation of the ten folds.
end for
$j^* = \arg\max_{1 \le j \le J} j$, subject to $e_j^{\text{avg}} \le e_0 + \sigma_0$; // one-standard-error rule
$\alpha^* = \alpha_{j^*}$;
return $\alpha*$



(a) (b)

Figure 3.2 **Hyper-parameter optimization.** (a) The optimal value of λ is found to be 0.00056 by Algorithms 1 and 2. (a) shows the average classification error and the standard error over ten-CV folds. The optimal β value is marked with the dashed red vertical line. (b) Example graph for the hyper-parameter optimization when $k = 10$. (b) shows the optimal value for α (see Algorithm 3).

3.4 Additional Evaluation Metrics

In addition to the losses defined for each task, we rely on different evaluation metrics for assessing the performance of the different components. We use the concordance index (C-index) to evaluate the survival module (Harrell, Califf, Pryor, Lee & Rosati, 1982). C-index calculates the fraction of patient pairs that are predicted to have the correct partial rank among all acceptable pairs. An *acceptable* pair is the one for which we can conclusively decide which patient survived longer. These are the pairs for which the first patient is less than that of the second patient, and the first patient survival time is non-censored. The C-index is computed as follows:

$$\text{C-index} = \frac{1}{|\mathcal{A}|} \sum_{(i,j)\in A} \mathbb{1}\left(\hat{\mathbf{h}}^{(i)} < \hat{\mathbf{h}}^{(j)}\right) \tag{3.8}$$

Where $\mathcal{A}$ is defined as an *acceptable* pair when C-index ranges from 0-1, and the higher values are better. For a random guess, the C-index value will be around 0.50. $\mathbb{1}$ is the indicator function that evaluates to 1 if the condition inside holds.

To assess cluster quality, we also use the silhouette score (Rousseeuw, 1987) together with k-means loss. The silhouette score is the standard evaluation metric that measures an object's similarity to its cluster compared to other clusters. It is calculated

for each instance with the following mathematical formulation:

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}} \tag{3.9}$$

Where $a(i)$ is the average distance of data point i to the other data points in the same cluster and $b(i)$ is the average nearest cluster distance from data point i to any other cluster. The score results are between -1 and 1. A value close to 1 indicates that the instance is assigned to its own cluster, whereas -1 means that it is closed to another cluster than its own.

3.5 Step 2: Identifying Cross-Cancer Patients

By applying the k-means clustering with different numbers of clusters, we group the patients into clusters for a set of increasing values of k. $\mathbf{K} = [k_1, k_2, ..., k_{|K|}]$. We then compute a patient-by-patient similarity matrix, which holds the pairwise similarity scores computed per patient pair, $f_{i,j}$. $f_{i,j}$ is simply the frequency of co-clustering within the same cluster over the $|K|$ clustering. To find similar patients, we only consider those pairs where they always co-cluster, thus $f_{i,j} = 1$.

Once the similar pairs are identified, we find the cross-cancer patients by checking whether they are closer to their cancer type patients or the similar patients' cluster. To achieve this, we use the sign of the silhouette score. In the transformed space, Z, we calculated the silhouette score; in these calculations, the papers' clusters are the real cancer types of the patients. A positive silhouette score indicates that the patient is close to other patients diagnosed with the same cancer, and such a patient can well be a representative member of that cancer type. On the other hand, the negative silhouette score flags that this patient is closer to other cancer type patients than the patients with the same cancer.

In summary, a patient who has a similarity score of 1 with another patient from another cancer type and with a negative silhouette score in all clustering models are deemed as a cross cancer patient. For a cross-cancer patient i, we will denote the set of patients to whom this patient is similar to with the set S_i.

3.6 Deep SHAP for Detecting Patient Specific Important Genes

We would like to analyze whether the predictive genes are shared also across the cross-cancer patient and its similar patient set. As deep learning models are not readily interpretable, we use the Deep SHAP (SHapley Additive exPlanations) method, which assigns each feature a significant value for a given prediction by using Shapley values in cooperative game theory (Lundberg & Lee, 2017). Shapley values explain the contribution of an input feature to the difference between the predicted value and the average prediction value (see section 2.5).

We define Φ_{ij}^{m} as the SHAP value for each patient i, and input feature j in model m of the $|K|$ different models. Each model consists of different prediction tasks; in finding the SHAP values, we fit the DeepExplainer by specifying the clustering part. The quantity of a SHAP value gives the significance score of a similar feature. We consider the features in the top one percent of all features in all models when ranked based on the SHAP values. In doing so, we aim to find the genes that consistently emerge as the important ones. Once we obtain a list for the cross-cancer pair, we check how many shared features exist between the cross-cancer patient and the patient(s) similar to these patients. Thus, we can infer the similarity of the patients by looking at common genes. The pseudo-code of the proposed procedure is in Algorithm 4.

Algorithm 4 Getting Top Features with Deep SHAP

Input: List of number of clusters **K**, the number of samples n, the set of similar patients $S^{(i)} = \{S_1^{(i)}, ..., S_s^{(i)}\}$ to the cross-cancer patient i, and s is the number of similar patients.

Output: Common top feature list $P^{(i)}$ within the top 1% between the cross-cancer patient i and patients similar to the patient i.

for k in K **do**
  Load the trained model with the number of clusters k;
  Specify clustering part of the model m;
  Get SHAP values Φ^m with Deep SHAP;
  for $l = 0$ to n **do**
    Take absolute values of Φ_l^m;
    Get top features P_l^m whose SHAP values within the top 1%;
  end for
end for
for $l = 0$ to n **do**
  $P_l = \bigcap_{m=1}^{M} P_l^m$;
end for

Common top features between the cross-cancer patient i and similar patients $S^{(i)}$:

$P^{(i)} = \bigcap_{l=S_1^{(i)}}^{S_s^{(i)}} P_l$; // Repeat for each cross-cancer patient i.

3.7 Dataset and Dataset Processing

We use the TCGA (The Cancer Genome Atlas) patient data source (Network & others, 2008). We obtain the processed gene expression data and clinical information of ten different types of cancer from `http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html` (Rappoport & Shamir, 2018). The following cancer types are covered in this dataset: acute myeloid leukemia (LAML), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KIRC), liver hepatocellular carcinoma (LIHC), lung squamous cell carcinoma (LUSC), skin Cutaneous Melanoma (SKCM), ovarian serous cystadenocarcinoma (OV), sarcoma (SARC), and Glioblastoma multiforme (GBM). In the following analysis, we refer to these cancer types as AML, breast, colon, kidney, liver, lung, melanoma, ovarian, sarcoma, and GBM, respectively. The gene expression is quantified by the RNA-seq experiments and was processed by the RNA-Seq Analysis pipeline of TCGA. The gene expression values are normalized with RSEM count

estimates and cover 20,531 genes.

In our analysis, we use only primary solid tumor samples. Since most samples are metastatic samples for melanoma cancer, and there are only 103 primary solid tumor samples, we decided to exclude melanoma and were left with nine cancer types. The number of patients ranges from 161 AML to 1211 breast cancer samples (see Table 3.1 for more details). We obtain age, gender, and survival information from clinical annotation data provided by the TCGA. The survival time is the number of days to the last follow-up if the patient is alive and the number of days to death if the patient has passed away. We discretize age information. The discretized bins are $0-20$, $20-35$, $35-45$, $45-55$, $55-65$, $65-75$, $75-85$, and $85-....$

Sample Type	AML	Breast	Colon	Kidney	Liver	Lung	Melanoma	Ovarian	Sarcoma	GBM
Primary Solid Tumor	0	1077	278	537	367	489	103	294	258	151
Recurrent Solid Tumor	0	0	1	0	2	0	0	4	3	13
Primary Blood Derived	161	0	0	0	0	0	0	0	0	0
Additional-New Primary	0	0	0	1	0	0	0	0	0	0
Metastatic	0	7	1	0	0	0	358	0	1	0
Additional Metastatic	0	0	0	0	0	0	1	0	0	0
Solid Tissue Normal	0	111	40	72	48	51	1	0	2	0

Table 3.1 **The number of cancer patients with sample types as in the dataset obtained from (Rappoport & Shamir, 2018).**

Chapter 4

RESULTS

In this chapter, we will present the results of applying DeepCrossCancer to cancer patients. Next, we analyze the cross-cancer patients in light of the complementary molecular data on patient tumors.

4.1 Experimental Set-up

We use gene expression data of tumors biopsied from cancer patients and the clinical annotation data made available by the TCGA project. The dataset details are provided in Section 3.7. We split the dataset into a train and test set with a 0.2 ratio and normalized it into the range [0-1]. We design a five-layer neural network: an input layer, two hidden layers, a classification layer, and a survival prediction layer. The number of nodes for the input layer, hidden layers, classification layer, and survival layer is set to 20533, 32, 16, 9, and 1. Learning rate, batch size, and the number of epochs are set to 0.0001, 30, and 200, respectively, for pretraining, and 0.05, 24, and 150 for training. For the clustering part, we train seven clustering models with seven different numbers of cluster size k, but the number can be changed. By using the method proposed in Section 3.3, the regularization parameter λ, and the trade-off parameters α, and β are optimized in each iteration. We use Adam (Kingma & Ba, 2014) and SGD (Bottou, 2010) optimizers in training the model.

The experiments are run with the following system configuration: CPU: Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz CPU. Memory: 256Gb. Operating system: Ubuntu 16.04.4 LTS.

k	**10**	**20**	**30**	**40**	**50**	**70**	**100**
Accuracy	0.97	0.98	0.97	0.98	0.98	0.97	0.98
C-index	0.69	0.70	0.73	0.69	0.72	0.71	0.72
Silhouette score	0.65	0.44	0.33	0.28	0.26	0.24	0.22

Table 4.1 **The performance measures are reported with different numbers of clusters (k).**

4.2 Cluster Evaluations

We evaluate the performance of clustering for different values of $k \in \mathbf{K} = [10, 20, 30, 40, 50, 70, 100]$. The evaluation metrics for k are reported in Table 4.1. The high silhouette score, accuracy, and the c-index show that the models are well trained. Further, we compare DeepCrossCancer with the K-means clustering algorithm based on silhouette scores for varying degrees of k. For the optimal number of clusters 10, the silhouette score of DeepCrossCancer is revealed to be 0.64 compared to a K-means silhouette score of 0.27. Overall, DeepCrossCancer shows better performance regardless of the number of clusters. While the number of clusters is increasing, the silhouette score stays at the same level. The comparison is shown in Figure 4.1 for all numbers of clusters.

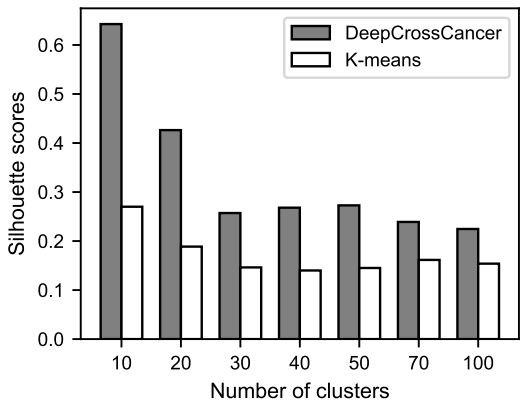


Figure 4.1 **Comparison of silhouette scores of DeepCrossCancer and K-means algorithm on different numbers of clusters.**

4.3 Cross-Cancer Patients Revealed

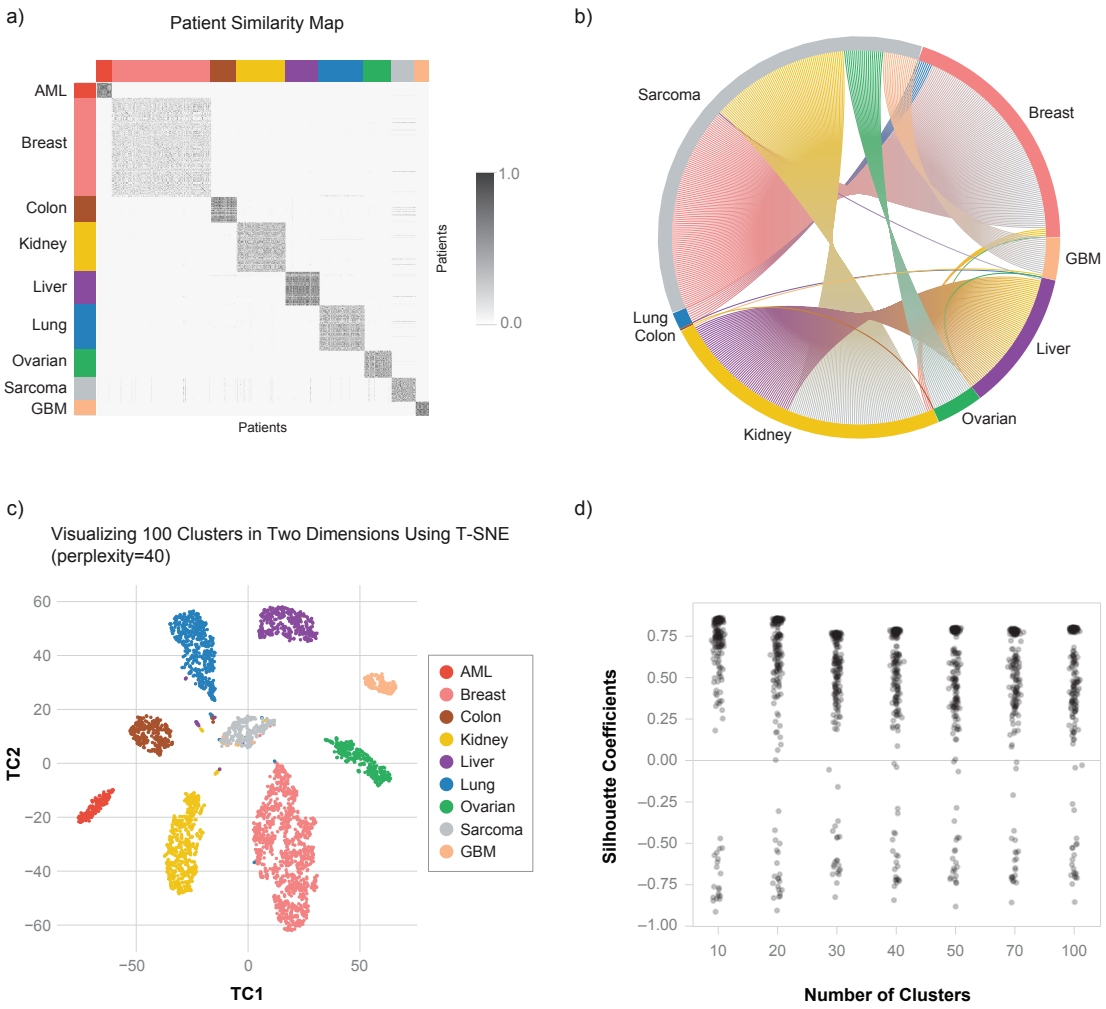


Figure 4.2 **Cross-cancer patients are revealed in the different types of cancer.** (a) The pairwise similarity of all patients is visualized by the heatmap. The similarity is based on how often the patients co-clusters. Off-diagonal black points represent similar patients across cancer. (b) Similar patients across cancers are shown by the chord diagram. (c) Example t-SNE plot for clustering with DeepCrossCancer with $k = 100$. The patients are colored by the actual cancer types. (d) The distribution of the silhouette coefficient of cross-cancer patients shown. Patients with a negative silhouette coefficient among similar patient pairs are the cross-cancer patients.

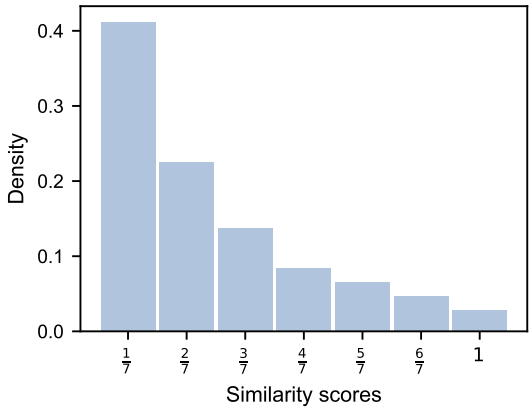


Figure 4.3 **Distribution of similarity scores of patients.** Similarity score is calculated for each patient pair as the fraction of frequency of co-clustering over multiple runs of clustering.

We apply t-SNE on the output of the encoder of the proposed model with $k = 10$ (Figure 4.2c). The figure is colored based on the actual cancer types. The plot shows that there are patients that are well-separated from their own cancer type members. For example, some breast, kidney, and liver cancer patients are closer to sarcoma patients. To find the patient pairs that are similar, we calculate a similarity score per patient pair based on how frequently the patient pair co-cluster (see Section 3.5). Figure 4.2a represents the heatmap of patients' similarity scores. The similarity scores are linearly scaled into [0,1] interval. Diagonal black points show pairwise patients that are in the same cancer type. Off-diagonal black points show similar patients across cancers. There are many off-diagonal black points that belong to sarcoma cancer. The distribution of similarity scores of patients is shown in Figure 4.3.

To find similar patients across cancers with a high similarity score, we only consider a small subset of patients from different cancer types with a similarity score of 1 because these patients are grouped into the same cluster across all models with the different numbers of clusters. This could be stringent but we aim for a high precision in finding cross-cancer patients. Figure 4.2b shows the similar patient pair distributions across different cancer types. Most similar patients are related to sarcoma. We also plot the distribution of the number of patients resembling a patient (Figure 4.4). For example, there is a kidney cancer patient that is found to be similar to 63 liver cancer patients.

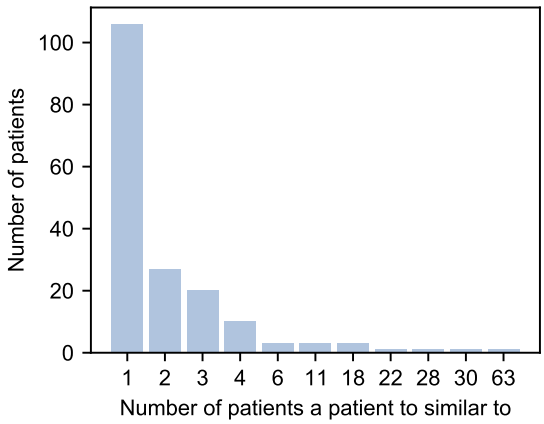


Figure 4.4 **The distribution of how many patients a patient is similar to.** There are 176 patients that show similarities across cancers.

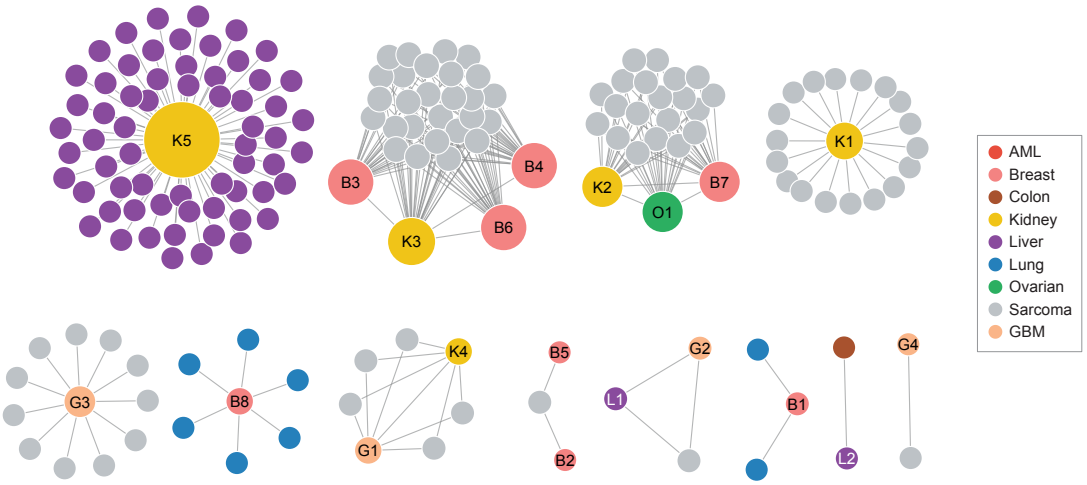


Figure 4.5 **The network of cross-cancer patients.** The relationship of patients across cancers is shown in the network. Cross-cancer patients are assigned an ID and are shown in the center of the network. The TCGA study abbreviations for cancer types in the legend are as follows: LAML, BRCA, COAD, KIRC, LIHC, LUSC, OV, SARC, and GBM. TCGA patient IDs of the patients are listed in Table 4.2.

Patient ID in the Network Figure	TCGA Patient ID	Cross-cancer Type	Cancer Type of Patients Similar to	Number of Patients Similar to
B8	TCGA.E9.A5FL.01	Breast	Lung	6
B1	TCGA.AR.A5QQ.01	Breast	Lung	2
B3	TCGA.A2.A4S1.01	Breast	Sarcoma	27
B4	TCGA.AO.A1KO.01	Breast	Sarcoma	27
B6	TCGA.AC.A2QJ.01	Breast	Sarcoma	27
B7	TCGA.AC.A2QH.01	Breast	Sarcoma	20
B5	TCGA.AC.A7VC.01	Breast	Sarcoma	1
B2	TCGA.BH.A6R9.01	Breast	Sarcoma	1
L2	TCGA.CC.5260.01	Liver	Colon	1
K1	TCGA.B0.4698.01	Kidney	Sarcoma	18
K2	TCGA.BP.4770.01	Kidney	Sarcoma	20
K3	TCGA.B0.4697.01	Kidney	Sarcoma	27
K4	TCGA.B0.4696.01	Kidney	Sarcoma	5
K5	TCGA.AS.3777.01	Kidney	Liver	63
L1	TCGA.CC.A7IJ.01	Liver	Sarcoma	1
O1	TCGA.61.1721.01	Ovarian	Sarcoma	20
G3	TCGA.06.2569.01	GBM	Sarcoma	11
G2	TCGA.02.0055.01	GBM	Sarcoma	1
G4	TCGA.06.0130.01	GBM	Sarcoma	1
G1	TCGA.28.5218.01	GBM	Sarcoma	5

Table 4.2 **TCGA patient IDs and cancer types of cross-cancer patients.**

Of similar patient pairs, not all members are cross-cancer patients (Section 3.1). For example, as seen in the t-SNE graph (Figure 4.2c), one patient could be in the center of its cluster member; thus, the other patient, though, could be the outsider to its cancer type cluster and then that patient will be the cross-cancer patient. To find such cross-cancer patients, we use silhouette coefficients. In these calculations, clusters are based on the actual cancer type classes, but the representation is obtained from the deep learning model's encoding layer. Patients that always have a negative score over 7 runs of clustering are deemed cross cancer patients. 20 out of 176 patients possess a negative silhouette coefficient consistently across all 7 cluster runs (Figure 4.2d). We will refer to these 20 patients as cross-cancer patients.

The 20 cross-cancer patients are shown in Figure 4.5 as the center node. Each node represents a patient who shows similarity to patients carrying other cancer types.

The nodes are colored based on the cancer type. The number of such patients is provided in Table 4.2. We present an analysis of cross cancer patients, along with those who exhibit similarities in the upcoming sections.

4.4 Detailed Analysis of Cross-cancer Patients Discovered by DeepCrossCancer

We analyze the genomics profile of cross-cancer patients in different aspects. We first looked at common predictive genes in the models. Next, we conducted an analysis of the gene expression, mutation, and copy number variations.

4.4.1 Significance of Common Genes Found with Deep SHAP

For each cross-cancer patient identified (the center node in Figure 4.5), we have a set of patients that this patient is similar to (the connected nodes to the center node in Figure 4.5). Let i denote the cross-cancer patient and S_i be the number of patients that this patient is similar to (the last column in Table 4.2). First, we identify the predictive genes for each of the patients and check if these predictive genes are shared across the cross-cancer patient and its similar patients.

We find the set of genes that are consistently emerging as important for replacing the cross-cancer patient in a cluster. Then, we check which of these genes are also important for the set of patients; this cross-cancer patient is similar to (Figure 4.5). The importance of a gene in the model is quantified by the SHAP values, as described in Section 3.6). The top 1% for the important list is taken, and the number of common genes is found. We next statistically test which of the number of shared genes is surprisingly large.

We use a non-parametric permutation test for testing the null hypothesis that the number is significantly large. Let i denote a cross-cancer patient, and S_i is the set of patients that this patient is similar to. B is the number of permutations, a test statistic t is calculated over patient i, and the patients in $S_{i'}$ drawn from the population in the cancer type of i. $S_{i'}$ patients are drawn B times and alternative test statistics (t') is calculated over i and $S_{i'}$. Based on the number of times $t \leq t'$ in B samplings (let it be c times), the empirical test statistic is calculated (p-value) by the ratio c/B. Similar permutation tests are also used in the following parts.

In the analysis of SHAP genes, the test statistic is the number of common genes

TCGA Patient ID	Cancer Type	Patient ID in the Network Figure	Adjusted P-value	Number of Common Genes Found by Deep SHAP
TCGA.A2.A4S1.01	Breast	B3	<0.0001	18
TCGA.AO.A1KO.01	Breast	B4	<0.0001	18
TCGA.AS.3777.01	Kidney	K5	<0.0001	13
TCGA.B0.4697.01	Kidney	K3	<0.0001	14
TCGA.AC.A2QJ.01	Breast	B6	0.0012	12
TCGA.06.2569.01	GBM	G3	0.0450	12
TCGA.06.0130.01	GBM	G4	0.0757	37
TCGA.BH.A6R9.01	Breast	B2	0.0825	42

Table 4.3 **Significance Results of Common Genes Found with Deep SHAP.**

between the patient i and its similar patient set S_i. The alternative test statistic is calculated similarly over i and $S_{i'}$ patients. We set the significance level to 0.05 and use Benjamini and Hochberg (B&H) correction (Benjamini & Hochberg, 1995). The number of samplings B is set to $10,000$. We repeat the method for each cross-cancer patient. For example, the kidney patient (K5) shares 13 common genes with 63 liver patients in the first cross-cancer subnetwork. The number of common genes is always bigger than the number of common genes for the randomly selected 63 kidney patients (p-value ≤ 0.0001). The result shows that there is a significant cross-cancer similarity between the kidney patient and liver patients in the subnet. The adjusted p-values are listed in Table 4.3 for each cross-cancer patient. The results show that there are 8 cross-cancer patients, out of 20, sharing a significantly large number of genes with their similar patient set.

4.4.2 Gene Expression Analysis of the Cross-cancer Patient K5

To further understand the nature of the similarity between these cross-cancer patients, we conduct a gene expression analysis. In this analysis, for every 20 cross-cancer patients', we identify the genes for which the expression level is more typical for the cross-cancer type rather than the patients' diagnosed cancer type. For example, for K5, we check genes, where the gene expression level of K5 is not typical in the kidney cancer gene expression distribution but is more like the liver patients gene expression distribution.

Since gene expression values are skewed, we use log-transformation ($\log_2(x+1)$, where x is the gene expression value). Let x_i be expression value of gene i for a cross-cancer patient. To quantify these genes that support the cross-cancer relation,

we calculate two z-scores. The mean and standard distribution are calculated over patients in different cancer types, one for the patient's actual cancer type and one for the cross-cancer patients exhibiting similarities. $Z_i^{(1)}$ is the z-score of gene i for a particular cross-cancer patient, calculated by the mean and standard deviation of gene expression calculated over the cancer type of the cross-cancer patient, whereas $Z^{(2)}$ is calculated with the mean and standard deviation of the 63 liver patients to which K5 is similar. We define ΔZ as the difference between two scores: $\Delta Z = |z_2| - |z_1|$. Since Z_1 and Z_2 scores follow a normal standard distribution, ΔZ follows a normal distribution with a mean of 0 and a standard deviation of 2. We standardize ΔZ dividing by 2. We test whether $\Delta Z < 0$. We adjust the p-values using the B&H correction (Benjamini & Hochberg, 1995), where the number of tests is the number of genes. We use the False Discovery Rate (FDR) threshold of 0.1 to subset the lists.

Significant genes for the cross-cancer patient (K5) are listed in Table A.2. We plot the top significant 15 genes for the cross-cancer patient. In Figure 4.6, the expression values of kidney and liver patients are shown for the cross-cancer patient K5. The yellow-point represents K5, and purple-points represent liver patients that are similar to K5. While the gene is over-expressed in K5, this expression level is typical for the 63 liver cancer patients to which K5 is similar. We also repeat the same analysis by controlling age and gender. To this end, we only take into account the patients with the same gender and the same age range as the cross-cancer patient. Gene expression profiles with the controlled gender and together with age for the cross-cancer patient are given in Figure 4.7. When we control the age and gender, the number of significant genes is decreased; however, the top most significant genes are almost the same as the previous analysis results.

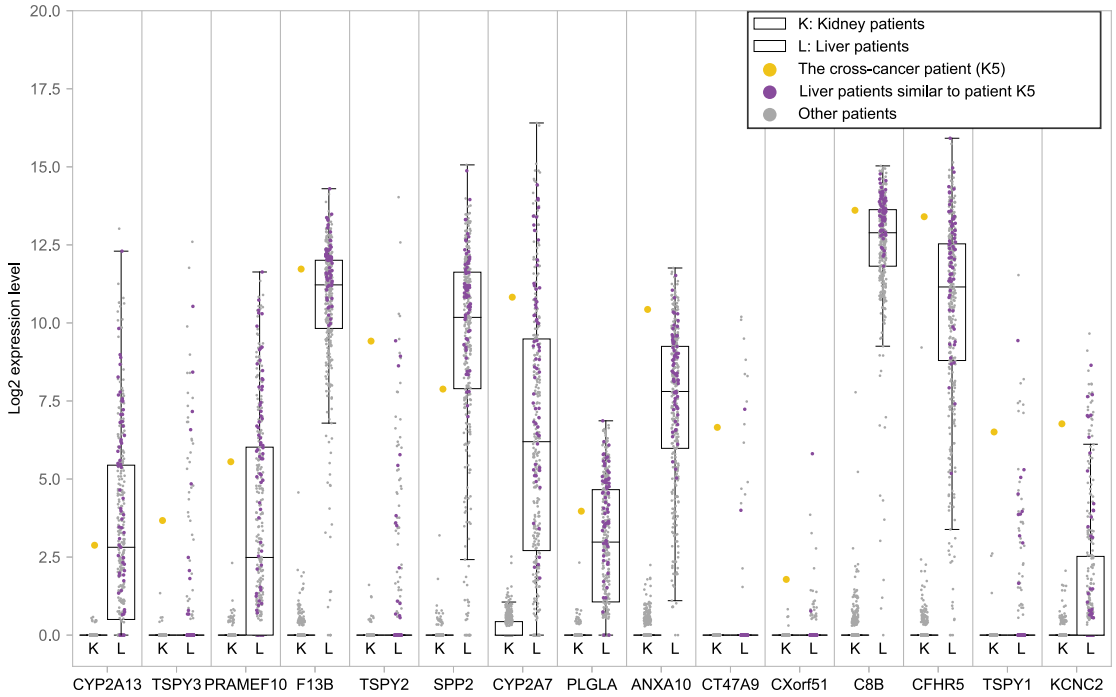
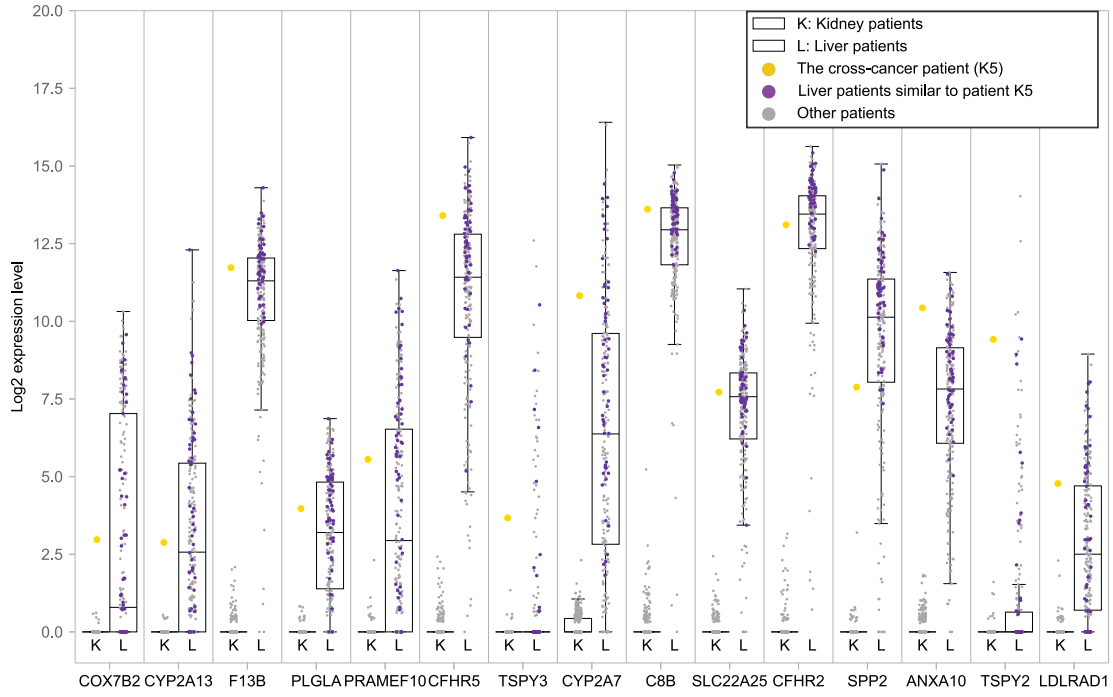
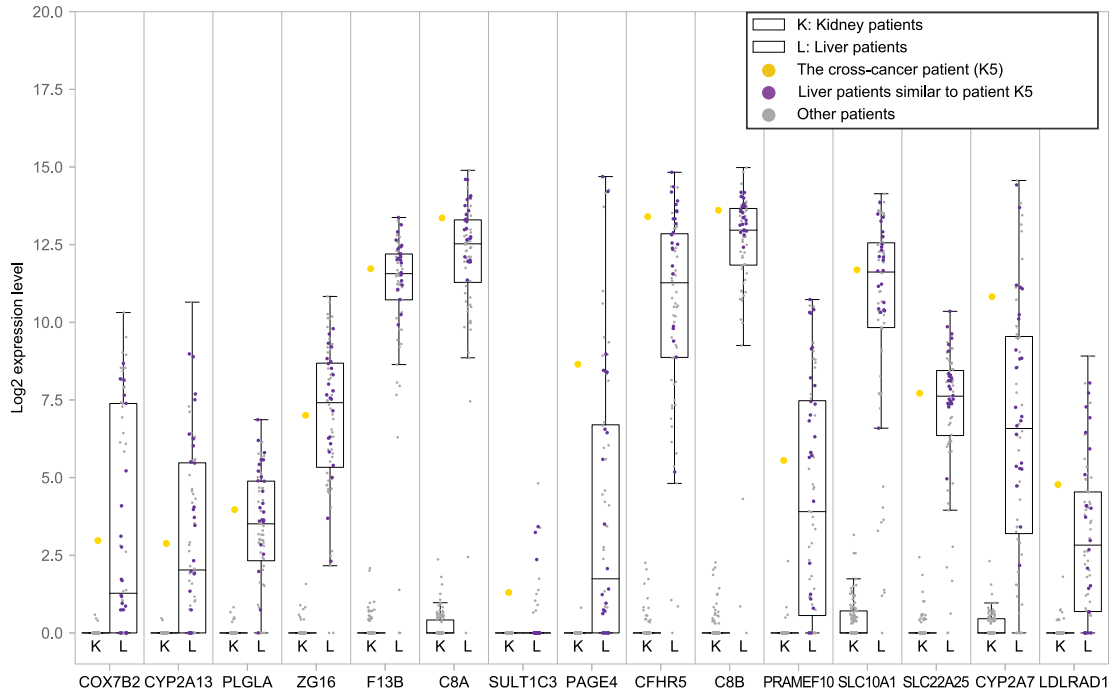


Figure 4.6 **Gene expression profiles of kidney (KIRC) and liver (LIHC) patients.** The cross-cancer patient K5 is represented with yellow-point and liver patients that are similar to K5 are shown with purple points. The most 15 significant genes (q-value $\leq$ 6.83e-14) are listed in the figure. Others are in Table A.2.



(a)



(b)

Figure 4.7 **Gene expression profiles of subset of kidney (KIRC) and liver (LIHC) patients based on gender and age.** (a) As a result of testing with gender subset on liver patients that are similar to K5 in Section 4.4.2, the most 15 significant genes (q-value $\leq$ 3.59e-10) were represented. (b) The test was done with liver patients who are similar to K5 in the same age and gender subset and the most 15 significant genes (q-value $\leq$ 5.04e-3) were represented.

4.4.3 Significance of Commonly Mutated Genes

Next, we analyze the mutated genes of a cross-cancer patient and the patients that they are similar to. The mutation annotation files (MAFS) were obtained from the Broad Institute TCGA GDAC Firehose repository (Deng, Brägelmann, Kryukov, Saraiva-Agostinho & Perner, 2017). For each gene that is mutated in the cross-cancer patient and in at least one of the similar patients set, we conduct a permutation test analysis to test if there are significantly many people in the similar patient group that bears the same mutated gene. When conducting the test for gene g and for patient i, the test statistic t is the number of patients in the similar patients set that have at least one mutation in g. The same statistic is calculated over random samplings of the cancer type of the cross-cancer patient. For example, when testing the significance of K5, 63 random patients are selected from the kidney; note that this is a more difficult test than selecting the random samples from liver patients.

We use B&H for multiple hypothesis correction (Benjamini & Hochberg, 1995), 16 genes are under the significance level with an FDR threshold of 0.1. These genes are namely *TP53, RB1, HOXA10, CHD6, NKAPL, PCLO, PREX2, LDHAL6A, SLC22A14, ADCY2, AHNAK, FHOD3, SDK1, FAT1, LRP1B*, and *LARP1B*. These genes are listed in detail with the associated cross-cancer patients in Figure 4.8. We also calculate the overall mutational frequency of these genes in the cross-cancer patient's cancer type and the cross-cancer type. We observe interesting findings. For example, while *TP53* is the most frequently mutated gene in most human cancers (Olivier, Hollstein & Hainaut, 2010), for example, in sarcoma patients 34.41% of them carry a *TP53* mutation. However, in kidney *TP53* mutation is observed in only 3.33% of all kidney cancer patients. One of the cross-cancer patients (K1) is found to be similar to two sarcoma patients. And *TP53* is found to be one of the significantly shared genes for this relationship. A similar case is observed *RB1* gene, while it is not frequently mutated in the kidney cohort, it is frequently mutated in the sarcoma and these kidney patients are found to be similar to sarcoma patients.

For example, *PCLO* is the commonly mutated gene between the kidney patient (K5) and 11 liver patients in the subnetwork. We observe that 11 out of 63 similar patients also have at least one mutation of *PCLO* (q-value ≤ 0.00010). This gene has been reported to be important for liver cancer in whole-genome analysis study (Fujimoto, Furuta, Shiraishi, Gotoh, Kawakami, Arihiro, Nakamura, Ueno, Ariizumi, Nguyen & others, 2015).

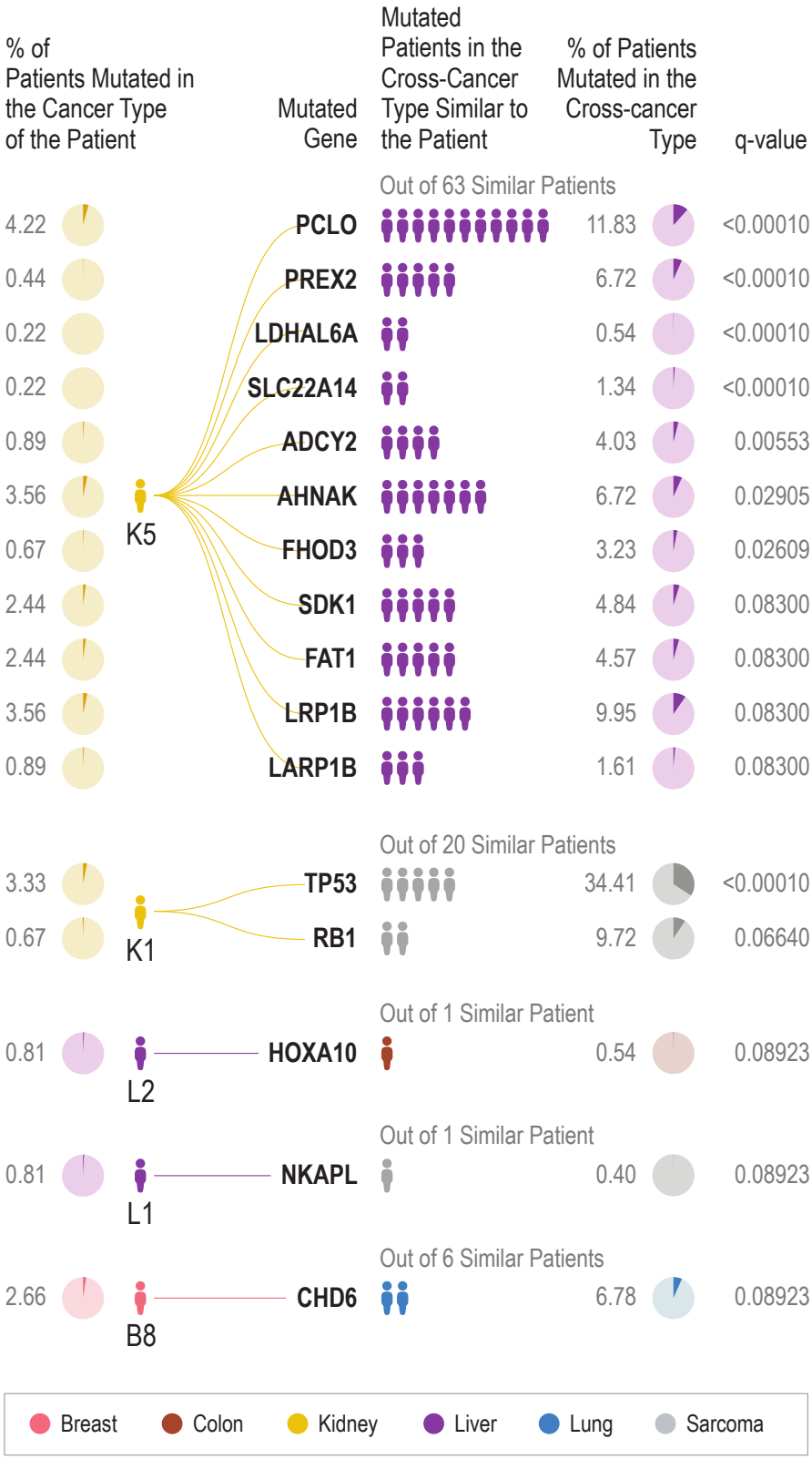


Figure 4.8 **Mutated gene profiles of cross-cancer patients.** Five cross-cancer patients share a significant number of commonly mutated genes with similar patients. These mutated genes appeared with a 0.1 FDR threshold. Details of the figure are shown in Table 4.4.

TCGA Patient ID	Network Patient ID	Cross-cancer Type (Number of Patients Similar to the Patient)	Mutated Gene	% of Patients Mutated in the Cancer Type of the Patient	% of Patients Mutated in the Cross-cancer Type	Number of Mutated Patients in the Cross-cancer Type Similar to the Patient	B&H Adjusted P-value	Mutation Type in the Patient	Mutation Type in Patients of the Cross-cancer Type (Number of Patients in the Mutation Type)
TCGA-BP-4770	K1	Sarcoma (20)	TP53	3.33	34.41	5	<0.00010	Splice Site	Missense Mutation (4) Frame Shift Del (1)
			RB1	0.67	9.72	2	0.066	Splice Site	Splice Site (1) Frame Shift Del (1)
TCGA-CC-5260	L2	Colon (1)	HOXA10	0.81	0.54	1	0.089	Frame Shift Ins	Missense Mutation
TCGA-E9-A5FL	B8	Lung (6)	CHD6	2.66	6.78	2	0.089	Missense Mutation	Silent
TCGA-CC-A7IJ	L1	Sarcoma (1)	NKAPL	0.81	0.4	1	0.089	Silent	Missense Mutation
TCGA-AS-3777	K5	Liver (63)	PCLO	4.22	11.83	11	<0.00010	Missense Mutation	Missense Mutation (9) Silent (2)
			PREX2	0.44	6.72	5	<0.00010	Splice Site	Missense Mutation (4) Silent (1)
			LDHAL6A	0.22	0.54	2	<0.00010	Silent	Missense Mutation (1) Splice Site (1)
			SLC22A14	0.22	1.34	2	<0.00010	Missense Mutation	Missense Mutation (1) Silent (1)
			ADCY2	0.89	4.03	4	0.006	Missense Mutation	Missense Mutation (3) In Frame Ins (1)
			AHNAK	3.56	6.72	7	0.029	Missense Mutation	Missense Mutation (3) Silent (3) Frame Shift Del (1)
			FHOD3	0.67	3.23	3	0.026	Missense Mutation	Missense Mutation (1) Silent (1) Nonsense Mutation (1)
			SDK1	2.44	4.84	5	0.083	Missense Mutation	Missense Mutation (4) Splice Site (1)
			FAT1	2.44	4.57	5	0.083	Missense Mutation	Missense Mutation (3) Silent (1) Frame Shift Del (1)
			LRP1B	3.56	9.95	6	0.083	Missense Mutation	Missense Mutation (4) Splice Site (1) Silent (1)
			LARP1B	0.89	1.61	3	0.083	Silent	Missense Mutation (2) Nonsense Mutation (1)

Table 4.4 **The significance of commonly mutated genes was tested by a permutation test with B&H correction.** Four cross-cancer patients show common genes that have been mutated significantly with patients similar to them.

4.4.4 Significance of Copy Number Variation (CNV) Overlapped Genes

We also analyze the common number of variations between the cross-cancer and its similar patient set. Copy number thresholded gene-level data from GISTIC2.0 (last analyze date 20160128)(Mermel, Schumacher, Hill, Meyerson, Beroukhim & Getz, 2011) were obtained from the Broad Institute TCGA GDAC Firehose repository by using the RTCGA-Toolbox R/BioConductor package, version 2.16.2 (Samur, 2014). Since the number of copy number alterations is too many in CNV data, we limit the analysis to cytobands with an alteration that is observed both in the cross-cancer patient and in at least 70% of her/his similar patients in each subnetwork. We analyzed the amplification and deletion events separately.

We tested the number of common copy number alterations that are significantly large between a cross-cancer patient and similar patients set in a subnetwork using a permutation test with 1000 samplings. Test statistic t is the number of patients with deletion (or amplification) in cytoband c and alternative test statistic t' is calculated over the randomly selected patients from the cross-cancer patient type. For $B = 1000$, we repeated the method for each common thresholded cytoband.

After correcting with B&H, 55 cytobands in amplification events rejects the null hypothesis test with FDR threshold of 0.1 (q-value ≤ 0.066). Four cross-cancer patients including G3, G1, K4, and B8 have shown significant amplification events shared with their similar patients. The most frequent chromosomal arm alterations included copy number gains in *1q* and *1p* of the cross-cancer patient G3, in *17p* and *Xp* of G1, in *4p, 17p* and *Xp* of K4, and in *3q* of B8. Since G1 and K4 are connected in the same subnetwork, they show the common frequent copy number gains in *17p* and *Xp*. The significantly amplified cytobands are listed in Table A.3 together with the number of similar patients, raw p-values, and q-values.

119 cytobands pass the significance test with an FDR threshold of 0.1 (q-value $\leq$ 0.094) for the deletion event. Six cross-cancer patients including G1, K4, K1, K2, B7, and B8 have shown significant deletion events shared with their cross-cancer similar patient set. The significant chromosomal losses are revealed in *16q* and *17p* of the cross-cancer patient G1, in *1q, 2q, 4q, 10p, 10q, 13q, 17p, 18q,* and *Xq* of the patient K4, in *13q* of K1, in *13q* and *16q* of K2, in *13q* of B7, in *4p, 4q,* and *9p* of B8. G1 and K4 have the common copy number losses on the chromosomal arm *17p*, and K2 and B7 have the common losses on the chromosomal arm *13q*. Because they are also similar to each other. The significantly deleted cytobands are listed in Table A.4.

Chapter 5

CONCLUSION AND FUTURE WORK

In this work, we present a framework, DeepCrossCancer, for discovering cross-cancer patients using patient molecular profiles and clinical information. The proposed methods focus on patients individually rather than cancer subtypes and employ a semi-supervised deep learning framework. In this framework's clustering step, we use a model that uses survival prediction, classification prediction, and clustering. These auxiliary tasks help to learn a representation that is useful for the clustering step.

Applying our method to patients that cover nine cancer types, we find 20 cross-cancer patients. These patients appear in eight cancer types, except AML. None of the AML patients had a cross-cancer relationship. The reason might be that AML is a type of blood cancer, whereas the sample type of other cancers is a primary solid tumor.

We analyze the 20 cross-cancer patients, and the patients that they have this relationship with have standard predictive features. Using deep learning interpretability tools, we find the genes that were predictive per patient. Next, we analyzed whether cross-cancer patients and their similar patients share common genes as their important genes. For example, for a kidney diagnosed patient, whom we find similar to liver cancer patients, the shared common genes include genes reported to be prognostic markers for liver cancer. As a secondary analysis, we examined whether genes across these cross-cancer patients have common copy number variations or mutations. These results identified a set of loci and genes that are shared across these cancers. The reason why these patients similar is a mystery. The underlying cause could be the exposure to common carcinogens, similar lifestyles, or typical genomic

architecture that predispose them to cancer in the same way. There can also be shared unknown factors; for example, the same specific drug intake or radiation therapy may affect similarities in their transcriptomes. Since we do not have access to medication history, we cannot correct for such hidden factors. However, the fact that these patients also share other genomic alterations supports the case that the reason is not such a common therapy history.

The method also has its limitations. Since deep learning models require large training set sizes, DeepCrossCancer is not applicable to small patient cohorts. Since DeepCrossCancer consists of multiple loss functions, the parameter tuning step takes time.

There are several routes for further investigation as future work. Here, we worked on patients from nine cancer types. In future work, we plan to extend this work to find similar cancers in patients from 33 cancer types made available in the TCGA project. The information used in the different modules can be expanded. For example, as input to clustering, we used the transcriptome data and used the mutation and copy number to analyze the discovered cross-cancer patients. The framework can be extended to a multi-modal framework with other types of omic characterization of the patients, such as methylation, mutation, and copy number variation. Another research direction could be the integration of the pathways, gene sets, and protein interaction networks to learn a better representation of the patients. In the classification module, the class labels could be enriched with the known subtype information of the cancer types. In the survival module, other clinical information such as the stage can be incorporated.

The study presents new opportunities to treat different cancer patients who share transcriptomic similarities but respond poorly to tissue-specific treatments. Transferring clinical treatment strategies from one patient to another patient could expedite the clinical management of the cancer patients. The identified common alterations could be further investigated experimentally to decipher the common molecular mechanisms shared across these patients. Finally, although DeepCrossCancer is designed for identifying cross-cancer patients, the framework can be extended to other diseases such as neurological diseases where commonalities are reported.

BIBLIOGRAPHY

Abeshouse, A., Ahn, J., Akbani, R., Ally, A., Amin, S., Andry, C. D., Annala, M., Aprikian, A., Armenia, J., Arora, A., et al. (2015). The molecular taxonomy of primary prostate cancer. *Cell*, *163*(4), 1011–1025.

Ali, H. R., Rueda, O. M., Chin, S.-F., Curtis, C., Dunning, M. J., Aparicio, S. A., & Caldas, C. (2014). Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome biology*, *15*(8), 431.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, *403*(6769), 503–511.

Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine*, *8*(8), 816–824.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, (pp. 153–160).

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., Liu, Y., Fan, H., Shen, H., Ravikumar, V., et al. (2018). A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer cell*, *33*(4), 690–705.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, *98*(24), 13790–13795.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010* (pp. 177–186). Springer.

Brannon, A. R., Reddy, A., Seiler, M., Arreola, A., Moore, D. T., Pruthi, R. S., Wallen, E. M., Nielsen, M. E., Liu, H., Nathanson, K. L., et al. (2010). Molecular stratification of clear cell renal cell carcinoma by consensus clustering reveals distinct subtypes and survival patterns. *Genes & cancer*, *1*(2), 152–163.

Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, *101*(12), 4164–4169.

Campbell, J. D., Yau, C., Bowlby, R., Liu, Y., Brennan, K., Fan, H., Taylor, A. M., Wang, C., Walter, V., Akbani, R., et al. (2018). Genomic, pathway network, and immunologic features distinguishing squamous carcinomas. *Cell reports*, *23*(1), 194–212.

Cavalli, F. M., Remke, M., Rampasek, L., Peacock, J., Shih, D. J., Luu, B., Garzia,

L., Torchia, J., Nor, C., Morrissy, A. S., et al. (2017). Intertumoral heterogeneity within medulloblastoma subgroups. *Cancer cell*, *31*(6), 737–754.

Chaudhary, K., Poirion, O. B., Lu, L., & Garmire, L. X. (2018). Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, *24*(6), 1248–1259.

Chen, R., Yang, L., Goodison, S., & Sun, Y. (2020). Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics*, *36*(5), 1476–1483.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187–202.

Cross-Disorder Group of the Psychiatric Genomics Consortium and others (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet*, *381*(9875), 1371–1379.

Damrauer, J. S., Hoadley, K. A., Chism, D. D., Fan, C., Tiganelli, C. J., Wobker, S. E., Yeh, J. J., Milowsky, M. I., Iyer, G., Parker, J. S., et al. (2014). Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. *Proceedings of the National Academy of Sciences*, *111*(8), 3110–3115.

de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., & Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. *BMC bioinformatics*, *9*(1), 497.

Deng, M., Brägelmann, J., Kryukov, I., Saraiva-Agostinho, N., & Perner, S. (2017). Firebrowser: an r client to the broad institute's firehose pipeline. *Database*, *2017*.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), 14863–14868.

Fujimoto, A., Furuta, M., Shiraishi, Y., Gotoh, K., Kawakami, Y., Arihiro, K., Nakamura, T., Ueno, M., Ariizumi, S.-i., Nguyen, H. H., et al. (2015). Whole-genome mutational landscape of liver cancers displaying biliary phenotype reveals hepatitis impact and molecular diversity. *Nature communications*, *6*(1), 1–8.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, (pp. 2672–2680).

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543–2546.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: prediction, inference and data mining. *Springer-Verlag, New York*, *2*, 241–244.

Hinton, G. E. (2012). A practical guide to training restricted boltzmann machines. In *Neural networks: Tricks of the trade* (pp. 599–619). Springer.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527–1554.

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, *173*(2), 291–304.

Hofree, M., Shen, J. P., Carter, H., Gross, A., & Ideker, T. (2013). Network-based

stratification of tumor mutations. *Nature methods*, *10*(11), 1108–1115.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241–254.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, *18*(1), 24.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint*, *arXiv:1412.6980*.

Kotila, M. (2018). Talos documentation.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, (pp. 1097–1105).

Lee, D. D. & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, (pp. 556–562).

Lei, Z., Tan, I. B., Das, K., Deng, N., Zouridis, H., Pattison, S., Chua, C., Feng, Z., Guan, Y. K., Ooi, C. H., et al. (2013). Identification of molecular subtypes of gastric cancer with different responses to pi3-kinase inhibitors and 5-fluorouracil. *Gastroenterology*, *145*(3), 554–565.

Levine, D. A., Network, C. G. A. R., et al. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, *497*(7447), 67–73.

Li, L., Cheng, W.-Y., Glicksberg, B. S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E. P., & Dudley, J. T. (2015). Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science translational medicine*, *7*(311), 311ra174–311ra174.

Liang, M., Li, Z., Chen, T., & Zeng, J. (2014). Integrative data analysis of multiplatform cancer data with a multimodal deep learning approach. *IEEE/ACM transactions on computational biology and bioinformatics*, *12*(4), 928–937.

Liu, Y., Sethi, N. S., Hinoue, T., Schneider, B. G., Cherniack, A. D., Sanchez-Vega, F., Seoane, J. A., Farshidfar, F., Bowlby, R., Islam, M., et al. (2018). Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer cell*, *33*(4), 721–735.

Liu, Z. & Zhang, S. (2015). Tumor characterization and stratification by integrated molecular profiles reveals essential pan-cancer features. *BMC genomics*, *16*(1), 503.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, *28*(2), 129–137.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, (pp. 4765–4774).

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology*, *12*(4), R41.

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, *52*(1-2), 91–118.

Nair, V. & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann

machines. In *ICML*.

Network, C. G. A. et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61.

Network, C. G. A. R. et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061.

Network, C. G. A. R. et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, *474*(7353), 609.

Network, C. G. A. R. et al. (2014a). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, *513*(7517), 202–209.

Network, C. G. A. R. et al. (2014b). Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, *507*(7492), 315–322.

Nie, F., Huang, H., Cai, X., & Ding, C. H. (2010). Efficient and robust feature selection via joint $\ell 2$, 1-norms minimization. In *Advances in neural information processing systems*, (pp. 1813–1821).

Olivier, M., Hollstein, M., & Hainaut, P. (2010). Tp53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology*, *2*(1), a001008.

Pai, S. & Bader, G. D. (2018). Patient similarity networks for precision medicine. *Journal of molecular biology*, *430*(18), 2924–2938.

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *nature*, *406*(6797), 747.

Quackenbush, J. (2001). Computational analysis of cdna microarray data. *Nature Reviews*, *2*(6), 418–428.

Raphael, B. J., Hruban, R. H., Aguirre, A. J., Moffitt, R. A., Yeh, J. J., Stewart, C., Robertson, A. G., Cherniack, A. D., Gupta, M., Getz, G., et al. (2017). Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer cell*, *32*(2), 185–203.

Rappoport, N. & Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids research*, *46*(20), 10546–10562.

Ricketts, C. J., De Cubas, A. A., Fan, H., Smith, C. C., Lang, M., Reznik, E., Bowlby, R., Gibb, E. A., Akbani, R., Beroukhim, R., et al. (2018). The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell reports*, *23*(1), 313–326.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Samur, M. K. (2014). Rtcgatoolbox: a new tool for exporting tcga firehose data. *PloS one*, *9*(9).

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., Ladanyi, M., & Sander, C. (2012). Integrative subtype discovery in glioblastoma using icluster. *PloS one*, *7*(4).

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, *25*(22), 2906–2912.

Shrikumar, A., Greenside, P., & Kundaje, A. (JMLR, 2017). Learning important features through propagating activation differences. *In Proc. of the 34th In-*

ternational Conference on Machine Learning, *70*, 3145–3153.

Slonim, D. K. (2002). From patterns to pathways: gene expression data analysis comes of age. *Nature genetics*, *32*(4), 502–508.

Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., & Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences*, *100*(18), 10393–10398.

Tepeli, Y. I., Ünal, A. B., Akdemir, F. M., & Tastan, O. (2020). Pamogk: A pathway graph kernel based multi-omics approach for patient clustering. *Bioinformatics*.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., & Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol*, *6*(1), e1000641.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, *17*(1), 98–110.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, (pp. 1096–1103).

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, *11*(3), 333.

Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C., Clarke, D., Gu, M., Emani, P., Yang, Y. T., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. *Science*, *362*(6420), eaat8464.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, *45*(10), 1113.

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer cell*, *1*(2), 133–143.

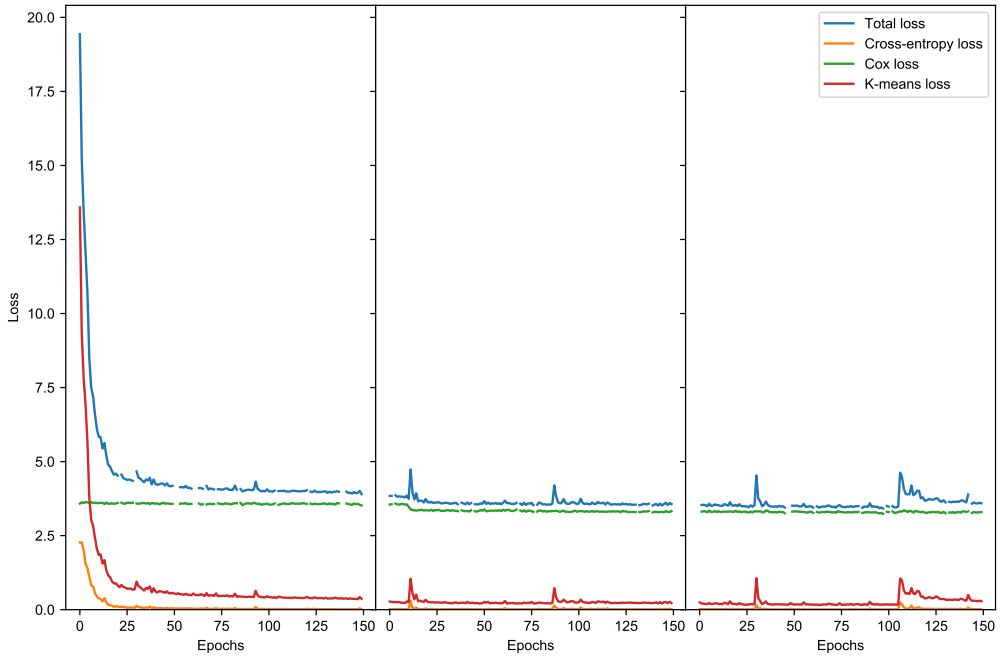
Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., Tao, Y., Guo, Y., Ni, X., & Shi, T. (2018). Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Frontiers in genetics*, *9*, 477.

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., & Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic acids research*, *40*(19), 9379–9391.

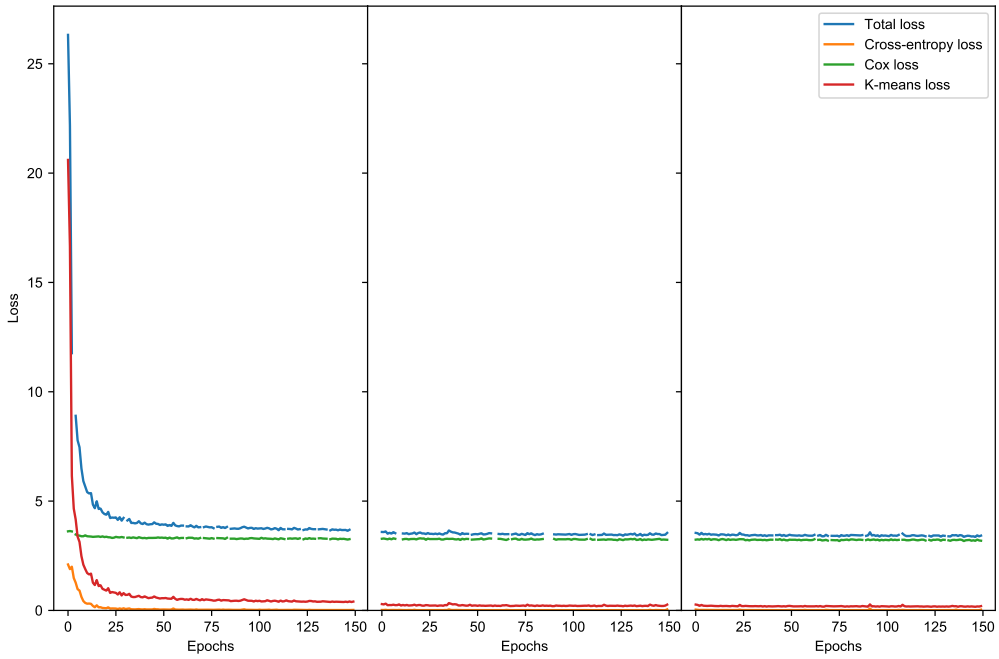
APPENDIX A

Table A.1 **The notation used throughout the study.**

Symbol	Explanation
X	input space $\in \mathbb{R}^d$
Z	latent space derived from the last hidden layer of the network input space $\in \mathbb{R}^p$, where $p << d$.
$\mathbf{x}$	input feature vector for patient $\in X$
$\mathbf{z}$	transformed feature vector for patient $\in Z$
n	number of patients
y	class label representing the patient's diagnosed cancer type
$\hat{y}$	predicted class label
m	number of cancer types, number of classes for the classification task
t	survival time
$\hat{h}$	predicted survival time
c	censorship status of the survival time
$\mathbf{W}$	network weight matrix
$\mathbf{b}$	network bias term
$\mathbf{o}$	network output layer
M	number of layers in the network
$\mathbf{\Theta}$	network trainable parameters $(\mathbf{W}, \mathbf{b})$
$\mathbf{U}$	cluster center matrix, where each column is a cluster center
$\mathbf{u}$	cluster center for a cluster $\in Z$
$\mathbf{Q}$	cluster assignment matrix, where each row is a 0-1 assignment vector for a patient
$\mathbf{q}$	cluster assignment vector. For example, for patient i, $q_{ji} = 1$ if the i-th patient belongs to the j-th cluster and 0 otherwise.
α	trade-off parameter for clustering loss
β	trade-off parameter for survival prediction
λ	regularization parameter
k	number of clusters
$\mathbf{K}$	the set of k values used for clustering
f	similarity score
Φ	Shap value



(a)



(b)

Figure A.1 **Training losses vs epochs for 3 iterations with updated Q and U values. Since convergence is met after the first iteration, we left with one iteration. Overfitting and underfitting was not observed.** (a) Training losses vs epochs for the number of clusters 20. (b) Training losses vs epochs for the number of clusters 50.

Table A.2 **Top 30 significant genes of the kidney patient K5 from the gene expression values.**

Genes	Z2	Z1	ΔZ	Z.score	P-value	BH Adjusted P-value
CYP2A13	−0.53	20.67	−20.15	−10.07	3.59E-24	7.13E-20
TSPY3	1.38	21.02	−19.64	−9.82	4.55E-23	4.51E-19
PRAMEF10	0.03	19.37	−19.34	−9.67	2.00E-22	1.32E-18
F13B	−0.13	19.24	−19.10	−9.55	6.35E-22	3.15E-18
TSPY2	3.85	22.19	−18.34	−9.17	2.37E-20	9.40E-17
SPP2	−1.83	20.11	−18.28	−9.14	3.15E-20	1.04E-16
CYP2A7	0.70	18.41	−17.72	−8.86	4.04E-19	1.14E-15
PLGLA	0.03	17.51	−17.48	−8.74	1.16E-18	2.88E-15
ANXA10	1.45	18.66	−17.21	−8.60	3.84E-18	8.46E-15
CT47A9	6.26	23.13	−16.87	−8.43	1.67E-17	3.30E-14
CXorf51	2.28	19.11	−16.83	−8.41	1.98E-17	3.57E-14
C8B	0.19	16.97	−16.78	−8.39	2.43E-17	3.94E-14
CFHR5	0.74	17.51	−16.77	−8.38	2.58E-17	3.94E-14
TSPY1	3.24	19.86	−16.62	−8.31	4.82E-17	6.83E-14
KCNC2	2.06	18.38	−16.32	−8.16	1.67E-16	2.20E-13
ZG16	−0.25	16.47	−16.21	−8.11	2.59E-16	3.22E-13
CT47B1	3.81	19.91	−16.11	−8.05	4.04E-16	4.70E-13
SLC22A25	−0.22	16.31	−16.09	−8.05	4.27E-16	4.70E-13
PRAMEF5	0.39	16.44	−16.04	−8.02	5.24E-16	5.47E-13
CYP2A6	0.46	16.18	−15.72	−7.86	1.89E-15	1.88E-12
RBMY2FP	1.71	17.30	−15.59	−7.80	3.21E-15	3.03E-12
LDLRAD1	0.63	16.18	−15.54	−7.77	3.86E-15	3.48E-12
SLCO1B1	−1.16	16.62	−15.46	−7.73	5.30E-15	4.57E-12
APOF	0.35	15.73	−15.38	−7.69	7.43E-15	6.14E-12
CFHR2	−1.08	16.34	−15.26	−7.63	1.17E-14	9.25E-12
FAM99A	−0.70	15.93	−15.23	−7.61	1.32E-14	1.01E-11
CT47A7	8.21	23.13	−14.92	−7.46	4.38E-14	3.22E-11
SLC10A1	−0.33	15.14	−14.81	−7.40	6.62E-14	4.69E-11
ZNF679	3.75	18.33	−14.58	−7.29	1.53E-13	1.03E-10
C8A	0.44	15.02	−14.58	−7.29	1.55E-13	1.03E-10

Table A.3 **The significantly amplified cytobands on chromosomes of cross-cancer patients. (q-value** ≤ 0.0010**)**

Cross-cancer Patient TCGA ID	**Cytoband**	**Number of Similar Patients that Show an Amplification Event on the Cytoband**
TCGA.06.2569.01	1q21.2	9
TCGA.06.2569.01	1p12	8
TCGA.06.2569.01	1q21.3	9
TCGA.06.2569.01	1p11.2	8
TCGA.06.2569.01	1q21.1	8
TCGA.28.5218.01	Xp11.22	4
TCGA.28.5218.01	17p12	4
TCGA.28.5218.01	Xp11.21	4
TCGA.28.5218.01	Xp11.23	4
TCGA.28.5218.01	Xp21.2	4
TCGA.28.5218.01	Xp11.3	4
TCGA.28.5218.01	Xp22.13	4
TCGA.28.5218.01	Xp21.1	4
TCGA.28.5218.01	Xp22.12	4
TCGA.28.5218.01	Xp11.4	4
TCGA.28.5218.01	Xp21.3	4
TCGA.B0.4696.01	Xp11.22	4
TCGA.B0.4696.01	17p12	4
TCGA.B0.4696.01	Xp11.21	4
TCGA.B0.4696.01	4p16.3	5
TCGA.B0.4696.01	Xp11.23	4
TCGA.B0.4696.01	17p11.2	4
TCGA.B0.4696.01	Xp21.2	4
TCGA.B0.4696.01	Xp11.3	4
TCGA.B0.4696.01	Xp21.1	4
TCGA.B0.4696.01	4p16.1	4
TCGA.B0.4696.01	Xp11.4	4
TCGA.E9.A5FL.01	3q25.2	6
TCGA.E9.A5FL.01	3q26.33	6
TCGA.E9.A5FL.01	3q29	6
TCGA.E9.A5FL.01	3q25.1	6

TCGA.E9.A5FL.01	3q26.32	6
TCGA.E9.A5FL.01	3q26.31	6
TCGA.E9.A5FL.01	3q23	6
TCGA.E9.A5FL.01	3q25.33	6
TCGA.E9.A5FL.01	3q27.1	6
TCGA.E9.A5FL.01	3q24	6
TCGA.E9.A5FL.01	3q27.3	6
TCGA.E9.A5FL.01	3q25.31	6
TCGA.E9.A5FL.01	3q25.32	6
TCGA.E9.A5FL.01	3q28	6
TCGA.E9.A5FL.01	3q26.1	6
TCGA.E9.A5FL.01	3q22.3	6
TCGA.E9.A5FL.01	3q27.2	6
TCGA.E9.A5FL.01	3q26.2	6

Table A.4 **The significantly deleted cytobands on chromosomes of cross-cancer patients. (q-value** ≤ 0.0010**)**

Cross-cancer Patient TCGA ID	**Cytoband**	**Number of Similar Patients that Show an Deletion Event on the Cytoband**
TCGA.06.2569.01	16q11.2	8
TCGA.06.2569.01	16q12.2	9
TCGA.06.2569.01	16q12.1	9
TCGA.06.2569.01	17p13.1	9
TCGA.B0.4696.01	13q14.13	5
TCGA.B0.4696.01	13q14.2	5
TCGA.B0.4696.01	1q43	4
TCGA.B0.4696.01	Xq27.2	5
TCGA.B0.4696.01	17p13.2	5
TCGA.B0.4696.01	13q14.3	5
TCGA.B0.4696.01	18q22.3	5
TCGA.B0.4696.01	1q41	5
TCGA.B0.4696.01	1q31.3	4
TCGA.B0.4696.01	1q32.2	4
TCGA.B0.4696.01	Xq24	5

TCGA.B0.4696.01	Xq26.1	5
TCGA.B0.4696.01	2q37.3	4
TCGA.B0.4696.01	10p12.1	4
TCGA.B0.4696.01	13q21.1	5
TCGA.B0.4696.01	1q44	4
TCGA.B0.4696.01	1q42.2	4
TCGA.B0.4696.01	10p15.3	4
TCGA.B0.4696.01	Xq27.3	5
TCGA.B0.4696.01	Xq22.2	5
TCGA.B0.4696.01	Xq25	5
TCGA.B0.4696.01	Xq26.2	5
TCGA.B0.4696.01	2q37.2	4
TCGA.B0.4696.01	Xq27.1	5
TCGA.B0.4696.01	1q42.3	4
TCGA.B0.4696.01	1q42.11	5
TCGA.B0.4696.01	Xq26.3	5
TCGA.B0.4696.01	1q32.1	4
TCGA.B0.4696.01	13q14.12	5
TCGA.B0.4696.01	1q32.3	4
TCGA.B0.4696.01	1q42.13	5
TCGA.B0.4696.01	Xq22.3	5
TCGA.B0.4696.01	Xq28	4
TCGA.B0.4696.01	18q22.1	5
TCGA.B0.4696.01	Xq22.1	5
TCGA.B0.4696.01	13q14.11	5
TCGA.B0.4696.01	1q42.12	5
TCGA.B0.4696.01	Xq23	5
TCGA.B0.4698.01	13q21.33	12
TCGA.B0.4698.01	13q31.1	12
TCGA.B0.4698.01	13q21.32	12
TCGA.B0.4698.01	13q22.3	12
TCGA.B0.4698.01	13q22.1	12
TCGA.B0.4698.01	13q21.31	12
TCGA.B0.4698.01	13q22.2	12
TCGA.BP.4770.01	13q14.13	14
TCGA.BP.4770.01	13q13.3	14
TCGA.BP.4770.01	13q14.3	15

TCGA.BP.4770.01	13q34	14
TCGA.BP.4770.01	13q14.11	14
TCGA.BP.4770.01	13q21.2	14
TCGA.BP.4770.01	13q21.31	14
TCGA.BP.4770.01	13q14.2	15
TCGA.BP.4770.01	13q21.1	15
TCGA.BP.4770.01	13q14.12	14
TCGA.BP.4770.01	16q21	14