GRAPH EMBEDDINGS ON PROTEIN INTERACTION NETWORKS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Halil İbrahim Kuru
February 2019

Graph Embeddings on Protein Interaction Networks
By Halil İbrahim Kuru
February 2019

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

A. Ercüment Çiçek(Advisor)

Öznur Taştan Okan(Co-Advisor)

Can Alkan

R. Gökberk Cinbiş

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

GRAPH EMBEDDINGS ON PROTEIN INTERACTION NETWORKS

Halil İbrahim Kuru
M.S. in Computer Engineering
Advisor: A. Ercüment Çiçek
February 2019

Protein-protein interaction (PPI) networks represent the possible set of interactions among proteins and thereby the genes that code for them. By integrating isolated signals on single genes such as mutations or differential expression patterns, PPI networks have enabled various biological discoveries so far. Furthermore, even the connectivity patterns of proteins in such networks have been proven to be highly informative for various prediction tasks involving proteins or genes. These tasks; however, require task specific feature engineering. Graph embedding techniques that learn a deep representation of the nodes on the network, provides a powerful alternative and obviate the need for this extensive feature engineering on the network. In this study we use *graph embedding techniques* on PPI networks in two independent machine learning tasks. The first part of the present work focuses on predicting gene essentiality. Using two different node embedding techniques, node2vec and DeepWalk, we present a classifier which only uses node embeddings as input and show that it can achieve up to 88 % AUC score in predicting human gene essentiality.

The second part of the thesis proposes a novel representation of patients based on pairwise rank order of patient protein expression values and protein interactions, which we abbreviate as PRER. Specifically, we use the protein expression values of proteins, and generate a patient specific gene embedding to represent relative expression of a protein with other proteins in the neighborhood of that protein. The neighborhood is derived using a biased random-walk strategy. We first check whether a given protein is less or more expressed compared to the other proteins in their neighborhood for a specific tumor. Based on this we generate a

representation that not only captures the dysregulation patterns among the proteins but also accounts for the molecular interactions. To test the effectiveness of this representation, we use PRER for the problem of patient survival prediction. When compared against the representation of patients with their individual protein expression features, PRER representation demonstrates significantly superior predictive performance in 8 out of 10 cancer types. Proteins that emerge as important in the PRER as opposed to individual expression values provide a valuable set of biomarkers with high prognostic value. Additionally, they highlight other proteins that should be further investigated for the dysregulation patterns.

Keywords: graph representations, node embeddings, gene essentiality, network topological features, survival prediction, cancer, protein-protein interaction network.

ÖZET

PROTEİN ETKİLEŞİM AĞLARINDA ÇİZGE GÖMÜLÜMLERİ

Halil İbrahim Kuru
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: A. Ercüment Çiçek
Şubat 2019

Protein-protein etkileşimi (PPE) ağları, proteinleri ve dolayısı ile onları kodlayan genler arasındaki olası etkileşimler kümesini temsil eder. Mutasyonlar veya değişken ifade örüntüleri gibi tek tek genlerden gelen sinyalleri entegre edilmesini olanaklı kılarak PPE ağları günüze dek çeşitli biyolojik keşiflere vesile olmuştur. Ayrıca, bu tür ağlardaki proteinlerin bağlantı örüntülerinin, proteinleri veya genleri içeren çeşitli tahmin görevleri için oldukça bilgilendirici olduğu kanıtlanmıştır. Ancak, bu görevler göreve özel öznitelik mühendisliği gerektirmektedir. Ağdaki düğümlerin derin bir gösterimini öğrenen çizge gömülüm teknikleri, bu konuda güçlü bir alternatif sağlamakta ve söz konusu ağ için duyulan kapsamlı öznitelik mühendisliği ihtiyacını ortadan kaldırmaktadır. Bu çalışmada, biz *çizge gömülme tekniklerini* iki bağımsız makine öğrenmesi görevinde kullanıyoruz. Mevcut çalışmanın ilk kısmı, gen esaslılığını tahmin etmeye odaklanıyor. Bu bölümde, iki farklı düğüm gömülme tekniği, node2vec ve DeepWalk kullanarak, girdi olarak yalnızca düğüm gömülme kullanıldığında, insan genlerinin gerekliliğini tahmin etmede % 88'e varan AUC alabileceğini gösteriyoruz.

Tezin ikinci kısmı, protein ifade değerlerinin çiftli sıralamaları ve protein etkileşimlerine dayalı, açılımını PRER olarak kısalttığımız özgün bir hasta gösterimi önermektedir. Daha spesifik olarak, proteinlerin ifade değerlerini kullanıyor ve bir proteinin kendi komşuluk bölgesindeki diğer proteinlerle nispi ifadesini temsil eden hastaya özgü bir gen gömülmesi üretiyoruz. Komşuluk bölgesi PPE ağında yanlı rastgele yürüme stratejisi kullanılarak türetiliyor. öncelikle, belirli bir proteinin spesifik bir tümör için komşuluk bölgesindeki diğer proteinlere kıyasla daha az veya

daha fazla ifade edilip edilmediğini kontrol ediyoruz. Buna dayanarak, sadece proteinler arasındaki düzensizlik örüntülerini yakalayan değil, aynı zamanda moleküler etkileşimleri de hesaba katan bir gösterim üretiyoruz. Bu gösterimin etkinliğini test etmek için, PRER'i hasta sağkalım tahmin problemi için kullanıyoruz. Hastaların bireysel protein ifade özellikleriyle gösterimine kıyasla, PRER gösterimi 10 kanser türünden 8'inde istatiski olarak anlamlı bir şekilde üstün tahmin performansı gösteriyor. Bireysel ifade değerlerinin aksine PRER'de önemli olarak ortaya çıkan proteinler, yüksek prognostik değeri olan değerli bir biyobelirteç seti sağlıyor. Ek olarak, düzensizlik desenleri için daha fazla araştırılması gereken diğer proteinleri de vurguluyor.

Anahtar sözcükler: çizge gösterimler, düğüm gömülümleri, gen esaslılığı, topolojik özellikler, sağkalım tahmini, kanser, protein-protein etkileşim ağı.

Acknowledgement

Fist of all, I would like thank Öznur Taştan Okan for her full support, understanding, kindness and motivation on my entire graduate study despite of the distances. Her ambition to doing science and understanding for my problems motivated me throughout my study.

I am grateful and want to thank A. Ercüment Çiçek, Can Alkan and R. Gökberk Cinbiş for reading my thesis, giving feedbacks and directions to research, and also for accepting to be in my thesis committee.

Moreover, I would like to thank my friends who are always with me from the beginning, and the whole Bilkent BTO family. Their support and friendship made life easier.

Finally, I would like to express my gratitude to my parents, my brother and sister for their support in my entire life and study, and I know that it will be less how much I thank them. I cannot find a way of express my feelings, gratitude and love to my mother who passed away due to cancer. I will always try to be worthy of you, and I hope that the work I have done can advances the research and treatments in cancer.

Contents

1 Introduction **1**

2 Predicting Gene Essentiality with Graph Embeddings **4**

2.1 Introduction 4

2.2 Related Work 6

2.3 Methods 8

2.3.1 Problem Statement 8

2.3.2 Gene Essentiality Prediction with Gene Embeddings (GEGE) 9

2.3.3 Network Topology Measures 13

2.3.4 Datasets 15

2.4 Results and Discussion 16

2.4.1 Prediction Performance 16

2.4.2 Effect of the Embedding Sizes 18

2.4.3 Performance with Additional Homolog Genes 23

2.4.4 Exploring Consistently Misclassified as Essential 24

3 Pairwise Ranking Embeddings with Random Walks (PRER) 25

3.1 Introduction 25

3.2 Related Work 27

3.3 Methods 29

3.3.1 Pairwise Ranking Embeddings Based on Random Walks . . 29

3.3.2 Survival Prediction 32

3.3.3 Clinical and Molecular Patient Data 37

3.4 Results and Discussion 38

3.4.1 Prediction Performance 38

3.4.2 Predictive Feature Sets 40

3.4.3 Top Predictive PRER are Prognostic Biomarkers 44

3.4.4 Proteins that Emerge as Important only in the PRER Representation 47

4 Conclusion and Future Work 51

A List of Abbreviations 74

B Important PRER Features **76**

C PRER Networks **86**

List of Figures

2.1 Schematic describing how the feature vectors are created based on node embeddings. Input is the PPI network and the output is a latent low-dimensional representation of nodes in the network. The node2vec and DeepWalk algorithms are used to generate the embeddings based on the topological features of the graph. Additional features about the genes can be concatenated to the node embedding vectors. In this case, homology feature is added. 10

2.2 Accuracy obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed 19

2.3 Area Under Receiver Operating Characteristic curve obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed 19

2.4 F1 obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed 20

2.5 Average Precision obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed 20

2.6 Accuracy obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed 21

2.7 Area Under Receiver Operating Characteristic curve obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed 21

2.8 F1 obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed 22

2.9 Average Precision obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed 22

3.1 The pipeline of the pairwise rank representation which is based on random walks. By generating random walks on the graph, the neighborhood of node 2 is obtained. Then the pairwise comparison of the neighborhood proteins in terms of their protein expression quantities is used to form a representation of the protein. 31

3.2 Each bar represents a subject in the study and their duration in the study. The open circles indicate that the data is right censored; the subject left the study before the event had occurred or the event had occurred due to a completely different reason. The black circles denote patients for whom the event had occurred within the duration of the study. 33

3.3 General pipeline for survival prediction. The step that involves generating PRER is skipped when the experiment is run with the alternative method of individual expression values. 39

3.4 Comparison of RSF model performances that are trained with individual features and pairwise ranking embeddings(PRER) for different cancer types. 40

3.5 Variable importance of significant pairwise ranking embeddings for ovarian cancer. 42

3.6 Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for ovarian cancer; edges represent that two proteins participate in a pairwise rank order feature together. 43

3.7 Kaplan-Meier plot for each cancer type based on overall survival. Number at risk denotes the number of patients at risk at a given time, and p-value is calculated with the log-rank test. 46

B.1 Variable importance of significant pairwise ranking embeddings for kidney renal clear cell carcinoma 77

B.2 Variable importance of significant pairwise ranking embeddings for breast invasive carcinoma 78

B.3 Variable importance of significant pairwise ranking embeddings for bladder urothelial carcinoma 79

B.4 Variable importance of significant pairwise ranking embeddings for head and neck squamous cell carcinoma 80

B.5 Variable importance of significant pairwise ranking embeddings for colon adenocarcinoma 81

B.6 Variable importance of significant pairwise ranking embeddings for glioblastoma multiforme 82

B.7 Variable importance of significant pairwise ranking embeddings for lung adenocarcinoma 83

B.8 Variable importance of significant pairwise ranking embeddings for lung squamous cell carcinoma 84

B.9 Variable importance of significant pairwise ranking embeddings for uterine corpus endometrial carcinoma 85

C.1 Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for kidney renal clear cell carcinoma; edges represent that two proteins participate in a pairwise rank order feature together 86

C.2 Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for breast invasive carcinoma; edges represent that two proteins participate in a pairwise rank order feature together 87

C.3 Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for bladder urothelial carcinoma; edges represent that two proteins participate in a pairwise rank order feature together . 88

C.4 Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for head and neck squamous cell carcinoma; edges represent that two proteins participate in a pairwise rank order feature together 89

C.5 Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for colon adenocarcinoma; edges represent that two proteins participate in a pairwise rank order feature together 90

C.6 Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for glioblastoma multiforme; edges represent that two proteins participate in a pairwise rank order feature together 91

C.7 Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for lung adenocarcinoma; edges represent that two proteins participate in a pairwise rank order feature together 92

C.8 Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for lung squamous cell carcinoma; edges represent that two proteins participate in a pairwise rank order feature together . 93

C.9 Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for uterine corpus endometrial carcinoma; edges represent that two proteins participate in a pairwise rank order feature together 94

List of Tables

2.1 Gene essentiality prediction performance when different features are input to SVM classifier and the kernel of choice is varied. 17

3.1 Number of censored and deceased patients for each cancer type. . . 38

3.2 Comparison of RSF model performances that are trained with individual features and pairwise ranking embeddings(PRER) for different cancer types. The methods are compared at the mean with four different quantile values of the 100 bootstrapped runs. 41

3.3 Top-10 rank differentiated features in each cancer with PRER. . . . 48

Chapter 1

Introduction

There has been a dramatic increase in the availability of large and high dimensional datasets generated from biological experiments. These datasets present an opportunity to gain deeper insights into biological systems and complex diseases. Machine learning has been an instrumental tool in towards realizing this aim. By building a statistical association in diverse datasets, it has been used for a variety of prediction tasks in biomedical research. One rich source of information that proved instrumental in biological discoveries is the analysis of biological networks. By enabling the quantitative study of biologically relevant molecules through their interactions, networks fundamentally contribute to the analysis of biological structures and the functions of living cells. Particularly, the use of protein-protein interaction (PPI) networks has been instrumental in for various applications such as disease characterization, module discovery, protein function prediction and drug target prediction. A PPI network comprises nodes, which correspond to proteins or genes that code for them, and edges that are established between the nodes that are reported to be interacting.

There are two main ways by which PPIs can be used for downstream prediction tasks. In the first scenario, the structure of the network is used as the primary

source of information, and the connectivity patterns of the proteins (or genes) are used to make statistical associations, such as predicting gene essentiality, protein complexes, module discovery or protein functions [1, 2]. For the second route, a PPI network is used is to integrate signals that are on single genes or proteins, such as mutations, or differential expression, to make predictions for gene, patient or disease characterization [3]. By analyzing the missing and incomplete observations within the context of PPI interactions, the hope is to boost the signal-to-noise ratio [2].

In this thesis, we use graph embedding techniques to integrate PPIs in the downstream classification tasks. The first task involves gene essentiality prediction. A gene is considered essential for an organism if its function is indispensable for the viability or reproductive success[4]. Distinguishing essential genes from non-essential genes has long been a fundamental question [5, 6, 7]. Earlier methods have established that the connectivity patterns of the proteins in the protein-interaction networks are highly predictive of gene essentiality. Here, we offer an alternative; we hypothesize that the node embedding approaches can be used to capture the local and global connectivity patterns of the proteins. The embedding methods allow learning a latent representation of nodes in the graphs in lower dimensional space. In general, the graph embedding methods first create a neighborhood around the relevant node with multiple random walks starting from that node. Subsequently, by using an optimization model to generate vectors in $\mathbb{R}^d$ for each node in the graph. Optimization is applied to preserve the similarity of nodes and the graph structure in the new representation. The learned embeddings can be used to deal with different problems on graphs such as node classification or link prediction.

In the first part of the thesis, we use node embedding techniques on the PPI network for the prediction of genes essentiality. Our results in human dataset show that we can achieve prediction performance which is better than network topological features and the previously reported results in the literature on the same dataset. We show that graph embeddings lead to better representations of the global and local view of the graph.

In the second part of the thesis, we propose a new embedding technique inspired by the node2vec technique to integrate protein expression data with the PPI. This technique performs a pairwise ranking of expression of a node in the PPI network with its neighborhood. The neighborhood is formed by multiple biased random walks. Therefore, we call this new representation Pairwise Ranking Embeddings with Random Walks (*PRER*). Experiments in 10 different cancer types show that this new representation yields significant improvements and give promising results for future experiments.

The outline of the thesis is as follows: In Chapter 2, we propose a framework to assess whether a gene is essential by combining known graph embedding methods with an SVM classifier. This chapter provides the details of graph embedding methods and presents a literature review on graph features which are used in the essential gene prediction task. In Chapter 3, we propose a new embedding method that combines both protein interaction network and molecular data, and it is applied to a survival analysis problem. This new method is named Pairwise Ranking Embeddings with Random walks (PRER). A framework with Random Survival Forest classifier is applied both to individual molecular data and the proposed PRER features. PRER achieves significant improvements compared to expression features. Chapter 4 concludes the thesis and suggest possible future directions on graph usage in both prediction problems.

Chapter 2

Predicting Gene Essentiality with Graph Embeddings

2.1 Introduction

A gene is considered essential its function is indispensable for the viability or reproductive success of a cell or an organism [4]. Distinguishing essential genes from non-essential genes is a fundamental question, the answer of which is key to understand the minimal set of functional requirements of an organism or a cell [5, 6, 7]. For example, only about 300 genes for a bacteria are enough to maintain its life [8]. The identification of essential genes also has practical significance for drug target identification. Essential genes of a pathogen constitute potential drug targets for infectious diseases[9, 10] while development of antibacterial drugs. Similarly, a gene essential for a cancer cell but not for the normal cell reveal vulnerable points of the cancer cells, which can be targeted by drugs [11, 12].

Assessing the essentiality of a gene requires assessing the viability of the living system that either entirely lacks that gene or in which the expression or function of

that gene has been significantly compromised. There are small-scale experimental techniques for single gene-knockouts [13]. To find all essential genes in a cell requires disrupting the gene individually one-by-one and assessing its effect on the cell viability. Single gene knockout experiments [14], RNAi screens [15]and more recently CRISPR/Cas9 genome editing technologies [16] have been used for this purpose.

While experimental methods provide powerful results, they are laborsome, so thus it is also an important question of whether essential genes can be predicted computationally. Such tools not only allows making predictions for species not amenable to experimental studies but also provide insight into the question of what makes a gene essential. In a large body of work, many biological features compiled from experimental data or properties of the genes are used as features and used to predict the essentiality of the genes. There exist methods that make use of properties of the genes such as gene sequence [17, 18, 19], gene expressions [20], functional annotation such as gene ontology [21].

With the availability of protein interaction data at a large scale that become available due to high-throughput technologies, such as yeast two hybrid, tandem affinity purification and mass spectrometry, it become also possible to ask the question whether the positioning of a gene in the PPI have relation to essentiality of the gene coding that protein. Previous studies suggested that there is an association between the essentiality of a gene and network topological features such as degree centrality and betweenness. Highly connected nodes and the nodes that serves as a bridge between two connected subgroups are significant in the network theory. It is also suggested in the literature that these kinds of nodes in the PPI network tend to be essential [22, 23]. Although the network structure gives information about the essentiality, there is no consensus about the topological measures that are related to essentiality.

In this study, we applied a framework to predict gene essentiality. We utilized human protein-protein interaction network to create embeddings. Our framework

is flexible to use different kinds of biological networks, and this framework is also applicable to other organisms. We also assess whether the conservation of a gene across different organisms may give information about a gene being essential. Additionally we checked whether conservation of a gene across species would add value to the predictions. We added this feature to our embeddings in graph, and we found improvements in our results. Our approach finds promising results by only using graph structure.

In this chapter, we firstly cover the data descriptions, and where they are obtained to use in our framework. Second, we will give descriptions and formulations of topological features and explain the node embeddings methodology. Thirdly, we will review the related literature work which uses biological networks to predict gene essentiality. For literature review, we only focused on the works that are using some types of network features. Finally, we will show the results and discuss the outcomes.

2.2 Related Work

Accurate prediction of essential genes with computational methods is important because this predictions may lead the laboratory experiments so that researchers may only deal with small subset of genes. Therefore, many computational methods proposed in the literature to assess the essentiality of genes [17, 18, 19, 20, 21].

The current developments in the technology enable us to get information about the interactions of protein. Therefore, effects of physical interactions at the PPI network can be examined for the essentiality of a gene. Other than PPI networks, different biological networks might be used for prediction as well. For example metabolic networks [24], transcriptional networks [25] and gene coexpression networks [26, 20] were used in the literature. All different networks types can have

their advantages compared to others. Thus, several studies in the literature utilized the different network types to gain advantages of all. Two studies [25, 27] used and integration of PPI, metabolic and gene coexpression networks to capture whole cellular activity.

Several work in the literature report that network topological properties of a gene in the network are related to gene essentiality. In particular centrality measures are explored. Among them the local centrality measures local connectivity patterns of a node, while betweenness centrality measures whether a node is a key connector of different parts of the network. Early network analysis pointed that there is a correlation between lethality and the degree of a node, where highly connected proteins in PPI networks tend to be essential [22]. Although this idea has been challenged by others [28, 29], several other studies that examined datasets in different organisms supported the idea that proteins with high local centrality is positively correlated to gene essentiality[30, 31, 32, 33]. Others [34] reported correlation with clustering coefficient, which is defined as the ratio of the number of edges connecting the neighbors of a node to the maximum number of possible edges among them. Others reported nodes with high betweenness centrality are likely to be essential [23, 35, 20]. In the search for finding the right centrality measures that is related to gene essentiality, several other studies interrogated the relationship of gene essentiality with a series of centrality measures [36, 37, 38]. There are also works in the literature that integrated the topological information with other information to predict essential genes in a binary classification framework in a supervised learning setting [27, 39, 40] to predict the essentiality of the genes and showing its usefulness.

The methods discussed above aim at including information about the local or global positioning of a node in the network, and there is no straightforward way to encode this information into a feature vector. Although these previous works claimed to use these topological features to have accurate predictions, there is no consensus about what features are more correlated with the essentiality. Assuming that the position on the biological networks plays a key role for the

essentiality of a gene, directly learning the structure of the network may lead to better performances. In this work we ask the question whether node embeddings learned in a deep learning framework can represent the topological properties of a protein in the PPI and help discriminate essential genes from non-essential genes. For this purpose, we learned the node embeddings of the genes in a PPI network and use these low-dimensional representations of the genes as features to train a binary classifier. The main reason to use node embedding methods is that they are learning features from the network as an optimization problem so that the position of a gene and its relations to other genes are conserved in the embeddings space. Our results show that the deep node embedding methods help to find good representations of the features as opposed to preselected topological features. Additionally, node embedding methods provide flexibility to be applied in different networks, and these deep methods are able to learn its own features. We claim that directly using network via embeddings can capture the information about essentiality if the interactions of genes have a clue about essentiality. Embedding methods like DeepWalk [41] and node2vec [42] were able to used at a broad range of applications [42, 43] such as computational and systems biology applications. However, none of the previous works about the prediction of essential genes used node embedding methods. We found promising results on human genes with the protein-protein interaction network.

2.3 Methods

2.3.1 Problem Statement

In this study, we set out to predict the essentiality of genes by formulating this problem as a classification task over the nodes of a protein-protein interaction (PPI) network. We denote this dataset as $D = x_i, y_i,$ where x_i is a multi-dimensional numeric representation of the gene and $y_i \in -1, 1$ is the class label.

Here 1 indicates the *essential* class and -1 indicates the $non-essential$ class. The node classification task involves predicting the most probable label for the node.

2.3.2 Gene Essentiality Prediction with Gene Embeddings (GEGE)

We propose that the gene essentiality can be predicted with features that capture the structural roles of the genes in the PPI that are learned with graph embedding techniques. We are given a PPI network, which we denote with $G = (V, E)$ where V is the set of vertices representing the genes coding for the proteins, and E denotes an edge between two such genes. The feature vector for a gene, x_i, is created based on the node embeddings which we learn on G. This feature vector can be augmented with additional information on the genes. The GEGE framework comprises two main steps: representing each gene with feature vectors based on node embeddings and building a classifier based on the learned representations of the genes. Below, we provide a detailed account of these steps.

2.3.2.1 Representing Genes with Node Embeddings

A graph embedding of node v returns a feature representation of the node in d dimensional space such that the local structures and the similarities between the nodes are conserved in this new feature space. This representation is learned based on the relationship of the nodes with each other; thus, the topology of the graph. In this representation, nodes that are highly connected belong to the same communities due to the homophily principle, and they are expected to be embedded closely. Additionally, the nodes that are not necessarily close in the network but have similar structural nodes (e.g. hubs) shall have similar embeddings. In short, these methods operate with homophily [44] and structural equivalence [45]

principles.

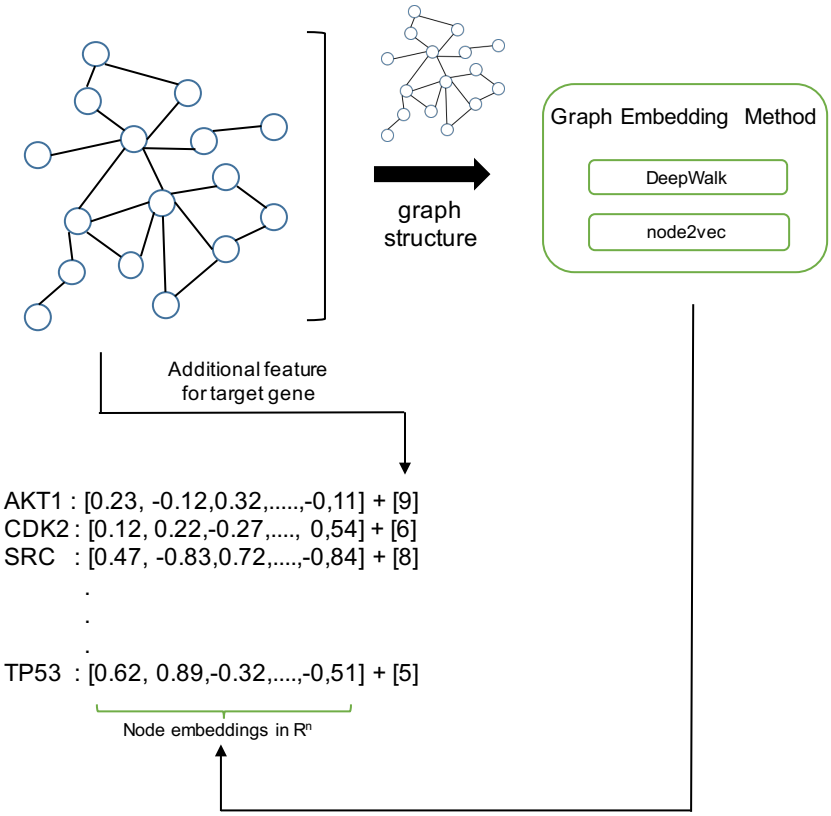


Figure 2.1: Schematic describing how the feature vectors are created based on node embeddings. Input is the PPI network and the output is a latent low-dimensional representation of nodes in the network. The node2vec and DeepWalk algorithms are used to generate the embeddings based on the topological features of the graph. Additional features about the genes can be concatenated to the node embedding vectors. In this case, homology feature is added.

In the current study, we experiment with two different node embedding methods: DeepWalk [41] and node2vec [42]. Both methods derive an embedding based on the neighborhood of a node, wherein the neighborhood is based on random walks on the graph. They both aim to minimize the differences between the graph

representation and the embedding representation. Random walks centered on a vertex v are used to derive a neighborhood of a given vertex v_i. In the literature, random walk approaches have been used for similarity measures and describing the local community information of a graph [46, 47]. On the other hand, using random walks to capture the local structures of a graph is a reasonable choice because it requires less computational power in comparison to the approaches that use the whole graph [41]. The details of the random walk and optimization strategies for each method are described in the following sections.

2.3.2.1.1 DeepWalk Embeddings DeepWalk firstly starts by taking a node v_i, and uniformly samples a node from the immediate neighbors. Then, it continues uniform sampling from the last visited node until it reaches to r_{th} node where r is the random walk length. For each node, DeepWalk applies this procedure λ times. Therefore, λ different random walks are examined for every node in the graph. At the end, it constructs $\lambda * \|V\|$ random walks. SkipGram algorithm [48] is applied to each random walk to update the current representation of the corresponding node v_i. In this context, SkipGram algorithm simply maximizes the probability of the neighbors of v_i in the given random walk [41]. In this setting, hierarchical softmax function is used to approximate the probability distribution and Stochastic Gradient Descent (SGD) [49] algorithm is applied to learn model parameters.

2.3.2.1.2 node2vec Embeddings Similar to DeepWalk, node2vec algorithm is based on random walks. Starting with node v_i, it samples nodes from the last visited node to create the neighborhood of node i. The key difference between DeepWalk and node2vec is that node2vec uses a biased search strategy to sample the neighborhood of a given node. node2vec chooses the next node from the last visited in accordance with the following probability distribution [42] :

$$P(n_i = v_j \| n_{i-1} = v_i) = \begin{cases} \frac{\Pi_{v_i v_j}}{Z} & if\ (v_i, v_j) \quad \epsilon \quad E \\ 0 & otherwise \end{cases} \tag{2.1}$$

, where $\frac{\Pi_{v_i v_j}}{Z}$ is the transition probability from node v_i to v_j. This transition probability is calculated by a bias parameter α that interpolates between Breadth-first Sampling (BFS) and Depth-first Sampling (DFS) [42] strategies while searching the next node from the last visited node in the graph. While BFS considers the immediate neighbors of the source (root) node of the random walk, DFS considers the nodes with sequentially increased distances from the source node. Therefore, BFS lead to capture structural similarity while DFS is inclined to capture the similarities in a macro-view [42].

2.3.2.2 Classification

After having represented each gene with a low-dimensional feature vector using node embedding methods we input them into a classifier. We use Support Vector Machine (SVM) due to its effectiveness in a variety of different domains [50]. Given a training set (x_i, y_i) where $x_i \in \Re^N$ and $y_i \in \{-1, +1\}, i = 1, 2, ..., k$ Support Vector Machines (SVMs) [51, 50] aims to find a linear seperating hyperplane which maximizes the margin. SVM formulates this tasks as the following optimization problem:

$$\begin{aligned} \underset{\mathbf{w},\xi}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(x_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{2.2}$$

where ξ is the slack variable, C is the penalty parameter for the error term and ϕ is a function that maps the feature vectors into the new space. This function is currently determined based on kernel function κ. We use the linear kernel, which is the dot products of the examples in the original space:

$$\kappa(x_i, x_j) = \mathbf{x_i}^T\mathbf{x_j} \tag{2.3}$$

We also experiment with the radial basis kernel (RBF) to find nonlinear decision boundaries. The RBF kernel function is given as follows:

$$\kappa(x_i, x_j) = \exp(-\gamma\|\mathbf{x_i} - \mathbf{x_j}\|^2) \quad where \quad \gamma > 0 \tag{2.4}$$

When the number of features is less than the number of instances, and there is a nonlinear relation between features and labels, RBF kernel becomes a reasonable choice. We use 10-fold cross-validation to test the performances of the parameters of our framework. For statistical measures, we use the area under curve (AUC) criterion as well as F1 and average precision (AP) scores.

2.3.3 Network Topology Measures

An alternative to the PPI network representation is to use a set of network topology measures to describe each node. Several centrality measures are known to be correlated with gene essentiality [22, 34, 32, 30, 35, 20]. Yu, et al. (2007) [23] argues that proteins with high betweenness are associated with important roles in the biological networks. In their study, they show that betweenness centrality is a robust predictive factor for the essentiality of a protein in the regulatory network. In the current study, we select four mostly used topological features which are closeness centrality, degree centrality, betweenness centrality, and clustering coefficient. We use SNAP library [52] to calculate each of these features. We briefly define them below:

Closeness Centrality: Closeness centrality calculates the average shortest paths from a given node i to all other nodes. It estimates how fast the flow of information would be from a given node i to all other nodes on average. It is calculated as follows:

$$\begin{aligned} ClosenessCentrality(i) = \frac{|V|-1}{\sum_{j=1}^{|V|} s(i,j)} \quad & where, \\ & i \neq j, \\ & s(i,j) \text{ is the shortest path} \\ & \text{length between node i and j,} \\ & |V| \text{ is the number of nodes in the graph} \end{aligned} \tag{2.5}$$

Degree Centrality: Degree centrality of a node is defined as its degree/(N-1),

where N is the number of nodes in the network:

$$DegreeCentrality(i) = \frac{d(i)}{N-1} \quad where, \\ d(i) \text{ is the degree of node i} \tag{2.6}$$

Betweenness Centrality: Betweenness centrality is a measure which calculates the frequency with which a node to appears in all the shortest paths in the network. It basically asses how often it forms a bridge between the nodes controlling the information flow of the graph. Betweenness centrality is calculated as follows:

$$BetweennessCentrality(i) = \sum_{j<k} \frac{s_{jk}(i)}{s_{jk}} \quad where, \\ s_{jk} \text{ is the number of shortest} \\ \text{paths between node j and k,} \\ s_{jk}(i) \text{ is the number of shortest paths} \\ \text{between node j and k that pass over i} \tag{2.7}$$

Clustering Coefficient: Clustering coefficient is a measure which calculates the average number of edges within the neighborhood of a node. It measures the tendency to conserve information flow within the densely connected subnetworks. For example, in PPI networks these groups of subnetworks may lead to functional modules of proteins which may work together in the biological processes. Clustering coefficient for undirected graphs is calculated as follows:

$$ClusteringCoeff(i) = 2 * \frac{e(i)}{k(i) * (k(i)-1)} \quad where, \\ e(i) \text{ is the number of edges} \\ \text{in the neighborhood of node i,} \\ k(i) \text{ is the number of nodes} \\ \text{in the neighborhood of node i} \tag{2.8}$$

2.3.4 Datasets

The information of whether a gene is essential or not is obtained from [19]. Guo et al. [19] gathered the raw data from DEG (http://tubic.tju.edu.cn/deg/) database. The original dataset of essentiality annotations is culled from three different works [16, 53, 54]. They obtain 11 different gene sets along with corresponding cell lines. They mark a gene as positive (essential) if it is selected as essential in more than half of the cell lines. After processing, they construct a benchmark dataset that consists of 12015 genes. Among these 12015 genes, 1516 of them are deemed essential. More details can be found at Guo et al. [19].

The protein-protein interaction network is obtained from InbioMap, which is publicly available at **https://www.intomics.com/inbio/map.html**. Inbiomap specifies a confidence score for each edge, which represents the support of the interaction in the literature. The interactions that have lower than 0.1 confidence cut-off are eliminated from the network. The network includes 17653 genes and 625.641 interactions between those genes. Among the 2015 genes that have information on their gene essentiality, 10579 are in the PPI network. The remaining 1436 are not incorporated in are the PPI network and only 2 of those are essential. Of the 10579 genes that are present in the PPI network genes, 1514 genes are essential, which constitute the positive class in our data, while remaining 9065 genes are not essential and they constitute the negative class.

Homology information of each human gene is obtained from HGNC Comparison of Orthology Predictions (HCOP) database (**https://www.genenames.org/cgi-bin/hcop**). The data contain homologous gene information of human genes with other 19 species. We consider the number of organisms in which a gene is conserved as an additional feature.

2.4 Results and Discussion

In this section, we present the results of different feature settings listed in section 2.3.3. In this experiment, we apply the node2vec and the DeepWalk algorithms to generate gene node embeddings on the PPI network. The model parameters which include the embedding size, the number of walks, walk length, p and q parameters are tuned via grid search strategy in a 10-fold cross-validation.

2.4.1 Prediction Performance

Gene essentiality prediction with graphs is well established in the literature [39, 38]. In these studies, ordinary pre-selected topological features such as degree centrality, clustering coefficient, closeness centrality, and betweenness centrality are used as relevant graph features. Therefore, we calculate these features from the PPI network for each target gene and compare them with the results of node embeddings.

Table 2.1 shows the best performances for each feature setting. With different feature sets, we identify best results with topological features and additional homology feature as 0.846 AUC, 0.609 F1 score and 0.426 AP with RBF SVM. For node2vec and DeepWalk embeddings, we outperform these results even when we use individual embedding features. When we further add the homology feature we find about 4 % improvement in terms of mean AUC score.

Embeddings	**Kernel**	**ACC**	**AUC**	**F1**	**AP**
DeepWalk	Linear	0.856	0.867	0.637	0.457
	RBF	0.871	0.874	0.661	0.483
DeepWalk + homology	Linear	0.875	0.883	0.672	0.497
	RBF	**0.885**	**0.884**	**0.687**	**0.514**
node2vec	Linear	0.856	0.84	0.588	0.405
	RBF	0.856	0.85	0.625	0.442
node2vec + homology	Linear	0.853	0.860	0.629	0.448
	RBF	0.880	0.868	0.669	0.491
Topological Features	Linear	0.584	0.736	0.395	0.244
	RBF	0.847	0.831	0.602	0.416
Topological Features + homology	Linear	0.802	0.824	0.553	0.371
	RBF	0.844	0.846	0.609	0.426

Table 2.1: Gene essentiality prediction performance when different features are input to SVM classifier and the kernel of choice is varied.

To assess the robustness of our approach, we evaluate the configuration of our best performance with 100 random bootstrap samples. We randomly split our dataset into test (20%) and train (80%) 100 times. Our best performance in a 10-fold cross-validation produces 0.884 mean AUC, 0.687 F1 and 0.514 average precision (AP), and we use the same parameters with the configuration of these results in 100 random bootstrapped samples. In this experiment, we find 0.881 mean AUC, 0.683 mean F1 and 0.508 AP. These results are close to the 10-fold cross-validation results. Therefore, we claim that our framework is robust against test dataset selection.

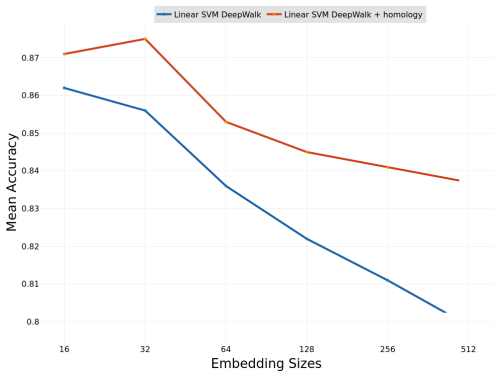
Results with additional homology feature reach to 0.884 mean AUC, 0.687 mean F1 and 0.514 mean AP scores for DeepWalk embeddings with RBF kernel. These results are better than the results of Guo et al. [19] who used the same gene essentiality dataset for their predictions and use nucleotide sequence features. The best performing model achieves 0.845 mean AUC in a 5-fold cross-validation. With a feature selection, their 5-fold cross-validation results achieved 0.885 mean AUC

score. They applied a similar strategy to our bootstrap experiment. They randomly split the data into test and train with %20 ratio, and found 0.854 mean AUC across 100 samples. We think that comparing our results with theirs may not be completely fair because due to different fold splits in cross-validation. However, to sum up, our results indicate that node embeddings are highly predictive of gene essentiality.

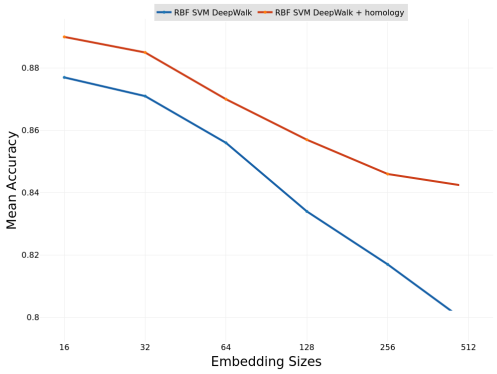
2.4.2 Effect of the Embedding Sizes

In this experiment, we explored the effect of the embedding size on the DeepWalk and node2vec performances. Embedding sizes are varied while the other parameters are fixed to their best values.

Figures 2.2, 2.3, 2.4 and 2.5 demonstrate the performance of DeepWalk in terms of different evaluation metrics.

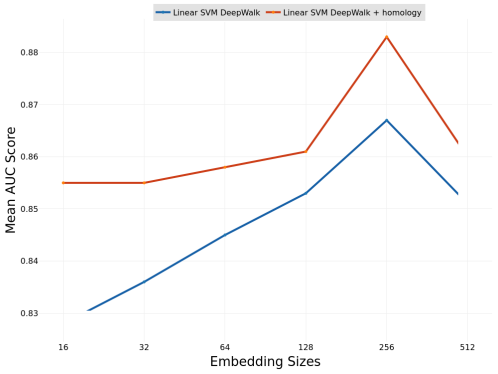


(a) Linear kernel

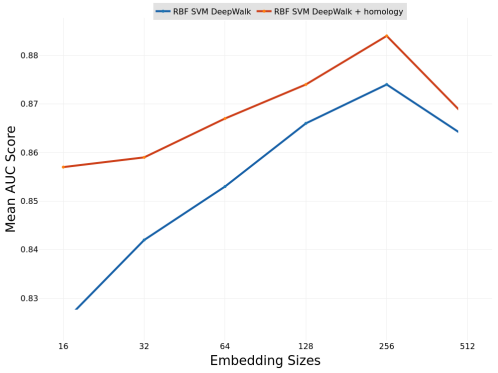


(b) RBF kernel

Figure 2.2: Accuracy obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed

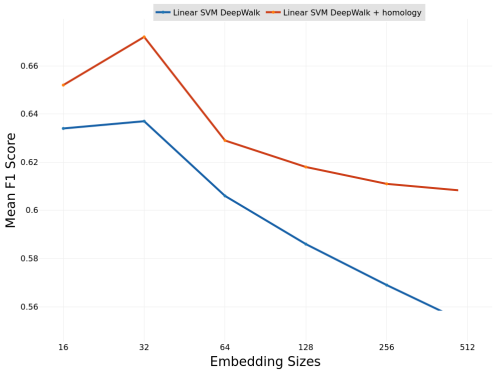


(a) Linear kernel

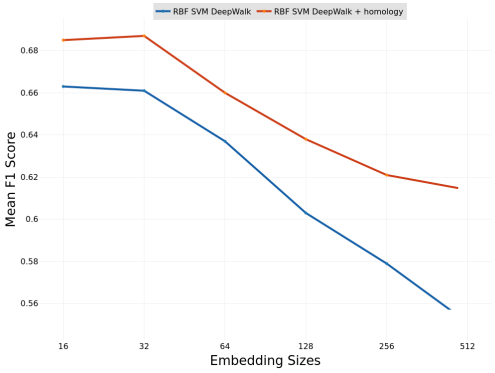


(b) RBF kernel

Figure 2.3: Area Under Receiver Operating Characteristic curve obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed

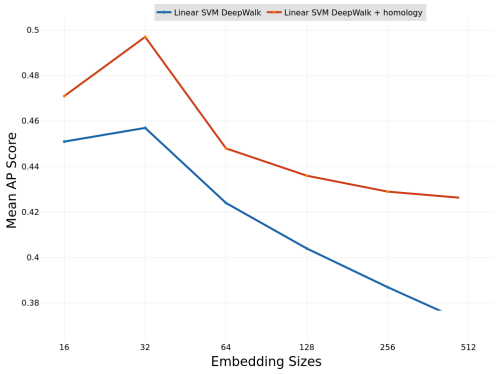


(a) Linear kernel

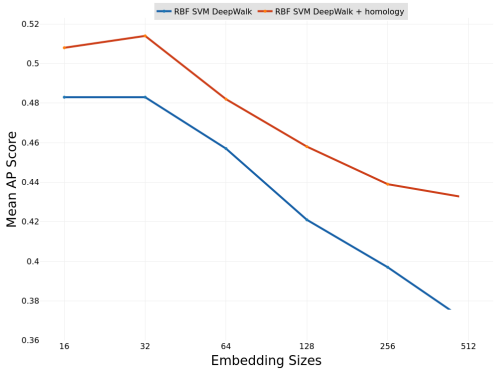


(b) RBF kernel

Figure 2.4: F1 obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed



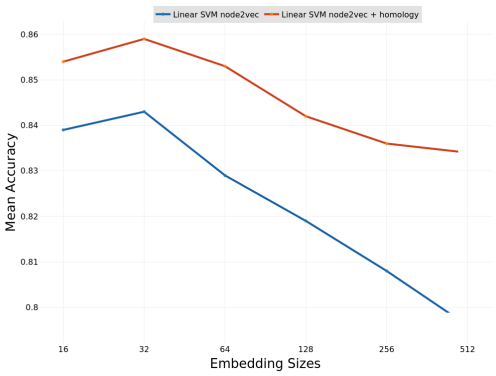
(a) Linear kernel



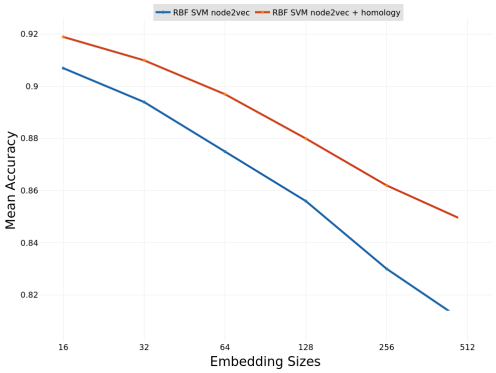
(b) RBF kernel

Figure 2.5: Average Precision obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed

Figures 2.6, 2.7, 2.8 and 2.9 demonstrate the performance of node2vec in terms of different evaluation metrics.

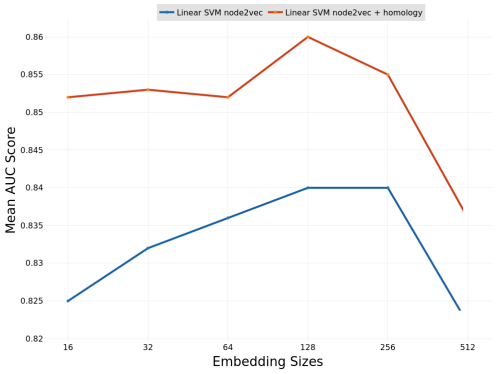


(a) Accuracy plot of linear kernel

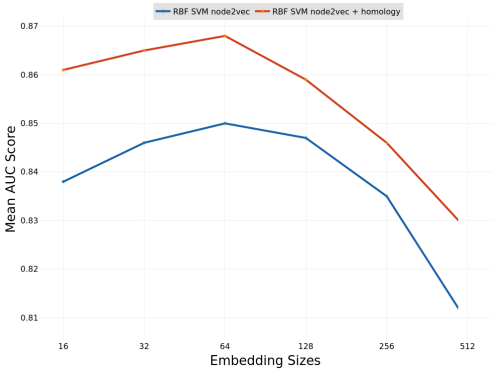


(b) Accuracy plot of RBF kernel

Figure 2.6: Accuracy obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed

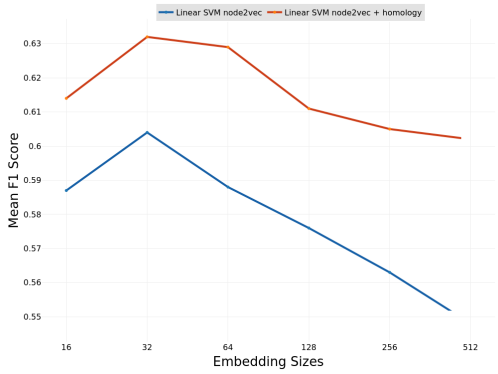


(a) AUC plot of linear kernel

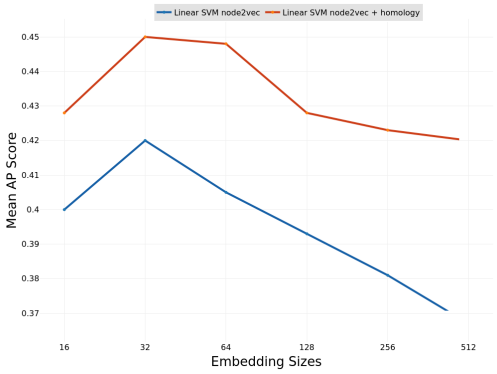


(b) AUC plot of RBF kernel

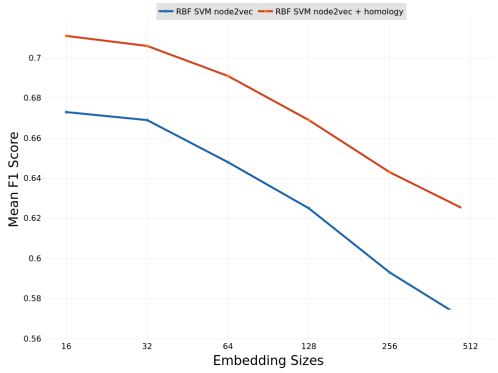
Figure 2.7: Area Under Receiver Operating Characteristic curve obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed



(a) F1 plot of linear kernel

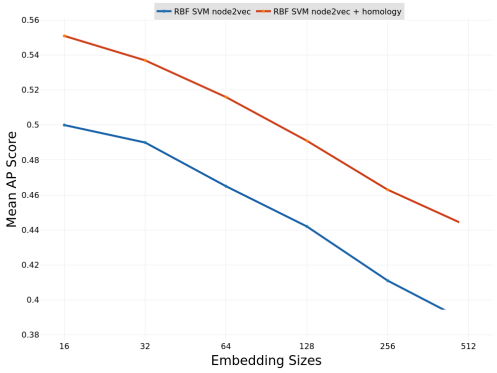


(a) Average precision plot of linear kernel



(b) F1 plot of RBF kernel

Figure 2.8: F1 obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed



(b) Average precision plot of RBF kernel

Figure 2.9: Average Precision obtained when DeepWalk is run with varying embedding sizes and when all the other parameters are fixed

Node embedding methods have a number of parameters to control the trade-off between over-fitting and over-generalization. The dimension of the embedding space is the most important parameter. Node embedding methods return a feature representation in R^d where d is the dimension of embedding. We experiment with different embedding sizes to construct the node embeddings and assess the performance of the classifier. Therefore, we show the effect of embedding size to

average performance under 10-fold cross-validation. Figures 2.3a and 2.3b, 2.7a and 2.7b show how the performance changes when different embedding sizes are used. We find the best result as 0.884 mean AUC score among the 10-folds. The patterns from the figures reveal that DeepWalk embeddings perform better than node2vec embeddings in the adopted settings. For the kernel parameter of SVM, RBF kernel gives about 1 % higher performance compared to the linear kernel. We find the best AUC score for the linear kernel with DeepWalk embeddings as 0.867, and 0.874 mean AUC score for RBF kernel when the embedding size d is set to 256. node2vec embeddings give their best performance when the embedding size d is set to 64 with RBF SVM, and it leads to 0.85 AUC score while the best performance with linear kernel achieves 0.84 AUC score. Nearly in all embedding sizes, DeepWalk embeddings consistently outperform node2vec embeddings.

2.4.3 Performance with Additional Homolog Genes

We calculate the number of organisms which maintain genes homologous with the target gene, and we call this feature as homology. We add this feature to our graph embeddings for each node and apply the same procedure for assessing the essentiality. As shown in Table 2.1 and further evidenced in Figures 2.7a, 2.7b, 2.3a and 2.3b, where we vary the embedding sizes and summarize the best overall results , homology brings complementary information and improves the results about 2% in accuracy in all of the configurations. The DeepWalk algorithm's best performance with RBF SVM improves from 0.874 to 0.884, and the best performance with Linear SVM improves from 0.867 to 0.883. Similarly, the node2vec best performance with RBF kernel improves from 0.850 to 0.868 while the best performance with the linear kernel improves from 0.84 to 0.86.

2.4.4 Exploring Consistently Misclassified as Essential

We examine the genes that are labeled as non-essential in the dataset but are consistently predicted as essential genes in our repeated bootstrap experiments. These constitute the false positive predictions of the classifier. For each gene, we calculate the counts of false positive prediction in 100 bootstrap experiments. We refer to this fraction as the false positive rate. We examine the genes whose false positive rates are greater than 0.50. Among these genes, we find that some of the genes are actually reported as conditionally essential genes, that is in a given context the gene is important for the viability of the organism. For example, SERPINE1, AIMP1, FIGF, RPS6KA6 and PDK4 genes are listed as non-essential in our benchmark dataset. However, we find that Silva et al. (2008) [55] reports that these genes as essential. Study of Marcotte et al. (2012) [56] labels DAZ2 gene as essential while Luo et al. (2009) [57] labels KIAA0408 and ZCCHC13 genes as essential.

These outcomes show potentials for our framework. Relations among the protein-protein interaction network may give potential information about the essentiality of a gene. Current nonessential genes may be labeled as essential with the advances of new techniques in future experiments, and interactions between proteins may lead to discoveries of new essential genes.

Chapter 3

Pairwise Ranking Embeddings with Random Walks (PRER)

3.1 Introduction

Accurate prediction of clinical outcomes such as survival success remains to be a challenge for cancer patients. If achieved, it can guide the decision-making process for choosing optimal treatment and surveillance strategies among alternative options. Typically, clinical or pathological features such as the age of the patient, tumor stage or grade are employed to predict the clinical outcomes. With the advent of high-throughput technologies, molecular descriptions of the tumors for a large number of patients across many cancer types have become available. However, it remains to be a significant challenge to use this data due to the high level of genomic heterogeneity among patients.

Recent advances in the high-throughput technologies enable researchers to use transcriptomic, genomic and epigenomic data for the prediction of clinical outcomes [58, 59, 60]. One such strategy relies on molecular expression data and

represents each patient with numerical features that encode the individual expression levels of the proteins. For this group of survival models, individual expressions of the molecules in the tumor cells become central. A richer representation of the patients can be obtained if the molecular interactions are considered. These interactions give a prior knowledge about the relations between molecules which may give biological insight about characteristics of a cancer. Multiple studies in the literature examined this idea to represent patients by integrating expression profiles of genes and their interactions [61, 62, 63], and found promising results. With this motivation, Buyukozkan et al. proposed a representation of patients based on the partial ordering of the molecular feature values. However, the work did not consider known information of molecular interactions such as protein-protein interactions (PPI). Here, we extend this work by incorporating the PPI knowledge with a graph embedding approach.

In this study, we propose to represent patients with features that are based on the pairwise rank of protein expression values. The motivation stems from the fact that the molecular mechanism is affected by the level of expression and the pairwise relationships can hint to molecular dysregulation patterns in different cancer types. More specifically, we consider whether a protein is more or less expressed with respect to proteins within its neighborhood on the protein-protein interaction network. For a given gene, its neighborhood is defined based on a random walk search strategy on the PPI network. In the ensuing discussion, we shall call this representation as Pairwise Rank Embeddings with Random walks (PRER).

To test the effectiveness of PRER, we use them for the problem of survival prediction. We build random forest survival models to predict patient survival in different cancer data types. When compared to the representation of patients with their individual protein expression features, PRER representation demonstrates significantly superior predictive performance in multiple cancer types. Proteins that are found to possess interaction features with high predictive performance include those that are already known to have clinical significance in the literature.

Additionally, we identify several others with promising prognostic potential.

Below, we describe the PRER representation in detail, describe the survival prediction models built using this representation, elaborate on the data sources and finally discuss our findings in light of the input parameters and highlight the potential biological significance of the reported results.

3.2 Related Work

Survival prediction for cancer is a challenging task. Therefore, various studies have been proposed to find out a subset of molecular signatures that drive the cancer cells. Ritchie et al (2015) [64] proposed a method with genetic algorithm based on the expression quantities of a miRNA and corresponding mRNA pair. miRNAs were also used in [65], seven miRNA features were identified as biomarkers which are related to the survival of gastric cancer patients. Gene expression profiles from different studies are collected in [26] to reduce to effect of data collecting procedure differences among studies and found 64 genes as the gene signature for predicting survivals of Stage I non-small cell lung cancer patients.

To improve the survival models, finding expression signatures for cancer is a significant task, and many studies examined this problem as supervised and unsupervised manner. Grouping patients via unsupervised clustering models used to figure out the different characteristics of sub-group of patients. Although clustering techniques do not directly estimate the survival distributions, they can be powerful to apprehend biological insights and to find expression signatures for different subtypes of cancer [66, 67, 68]. For example, a clustering based approach proposed by Pagnotta and Ceccarelli (2011) [69], firstly clusters patients with respect to their survival rates and find biomarkers for each cluster with an iterative procedure.

To find out what molecular dysregulations can effect to survival rate (high or low) of patients, dichotomized survival models [70, 71, 72, 73, 74, 75] have been studied to reduce the continuous regression problem into a classification task. Survival times are discretized into classes based on predefined thresholds. Although these methods are not directly approximate the continuous survival time data, they may enable us to discover the genes that their up or down regulations can discriminate the patients with high survival from low survival. Therefore, these types of models could be used as the feature selection procedure.

On the other hand, many previous frameworks have been proposed to integrate different kinds of quantities in order to achieve better survival estimation. There are some frameworks that integrate different molecular data. Kim et al. (2015) [76] proposed a network-based method which integrate molecular data such as miRNA, copy number alteration, methylation and gene expression, with prior genomic information from signaling or regulatory networks. They showed that the predictive power of models for clinical outcomes of ovarian cancer patients improved with the integrated data.

Integration of molecular expressions with biological knowledge coming from pathways or biological networks was used in various tasks to find out features that are relevant to survival, and to improve the model performances. The study of Taylor et al. (2009) [61] integrated PPI network and co-expression profiles and claimed that genes with dysregulated neighbors in PPI network may be used as the indicator to predict the survival of breast cancer patients. Their experiments support the idea that good biomarkers might be found by focusing on pathways or physical interactions in the network rather than using genes individually. Another study [63] supports similar point and claimed that gene co-expression or functional relation networks can be used for improving the performance of predicting survival in ovarian cancer. Crijns et al. (2009) [62] analyzed the gene expression profiles of tumors to predict the survival of ovarian cancer patients, and further asses the relations of transcription factors and pathways to survival rates. A different approach called NetRank [77], a similar approach as Pagerank [78] uses both gene

expression and the network relation information to rank the genes with respect to their relevance to the outcome of pancreatic cancer.

While molecular data give biological insight about the cells and integration of different data sources lead to good results [79, 80, 81, 82, 64] to predict survival, the models can be improved by integrating clinical information to molecular data [59]. Shedden et al. (2008) [58] used gene expression profiling as well as clinical information such as age, gender, tumor stage and median follow-up time, with eight different classifiers to predict the hazard ratios of lung adenocarcinoma patients.

3.3 Methods

3.3.1 Pairwise Ranking Embeddings Based on Random Walks

We arrive at a patient feature representation using molecular expression data and the PPI network. This representation can be both built on mRNA expression or on protein expression data. In this work, we use the protein expression data as mRNA expression is a proxy for protein expression. Let $G = (V, E)$ be the given PPI network. Let $U \subset V$ be the proteins whose expression values are quantified; these constitute the source proteins. The expression values for node u is$X_u^{(k)} \in \mathbb{R}$ for patient k. For each source protein in U, we first define a neighborhood, N_u, which is the set of proteins that are proximal to the source protein u on G. Below we first describe how the neighborhood is defined and then detail how the features are derived within these neighborhoods.

3.3.1.1 Protein Neighborhoods on the Protein Interaction Network

To obtain the neighborhood for a node in the graph, a set of random walks are generated. For every source node $u \in U$, we sample neighbors of a source node. Towards this aim, we first describe the node2vec biased random walk strategy[42] algorithm.

A random walk with a fixed length of l starting at source node u is generated based on the following distribution:

$$P(c_i = x \mid c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & if\ if(v,x) \quad \epsilon \quad E \\ 0 & otherwise \end{cases} \tag{3.1}$$

Here, c_i denotes i_{th} node in the walk and $c_0 = u$. $\mathbf{Z}$ is the normalizing constant. $P(c_i = x \mid c_{i-1} = v)$ is the transition probability on edge (v, x), where the current node is v, and the next node to visit is x.

Transition probability is depends on the function π, and it is defined as:

$$\pi_{vx} = \alpha_{pq}(t, x) * w_{vx} \tag{3.2}$$

where w_{vx} is the edge weight between nodes v and x. However, in this work, we used an unweighted PPI network and thus set $w_{vx} = 1$. $\alpha_{pq}(t, x)$ is the random walk bias which is defined at equation 3.3 based on the parameters p and q.

$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases} \tag{3.3}$$

This bias controls the different search strategies to sample the next visited nodes. The random walk algorithms used in this study uses two different search methods: Depth-first Sampling (DFS) and Breadth-first Sampling (BFS). As explained in [42], BFS samples the nodes from the nearby nodes while DFS samples the nodes by sequentially increasing the distance from a source node. p and q parameters control the connection between BFS and DFS approaches. With a high q value,

sampled nodes in the random walk are aligned to BFS and get a local view over the source node. Small q value aligns random walk to DFS so that a global view of the network is explored. p controls the chance of revisiting the nodes. A high value of p decreases the probability of sampling of the already visited nodes while a small value of p aligns random walk to return the source node.

By using fixed length random walks, we sample a neighborhood for any given source node. To be consistent and to decrease the variance, multiple random walks per source node are applied so that different neighborhoods are sampled for each node. Frequencies of nodes in the multiple neighborhoods are calculated, and the nodes, that are involved in more than one random walk, are selected as the neighbors of the source node.

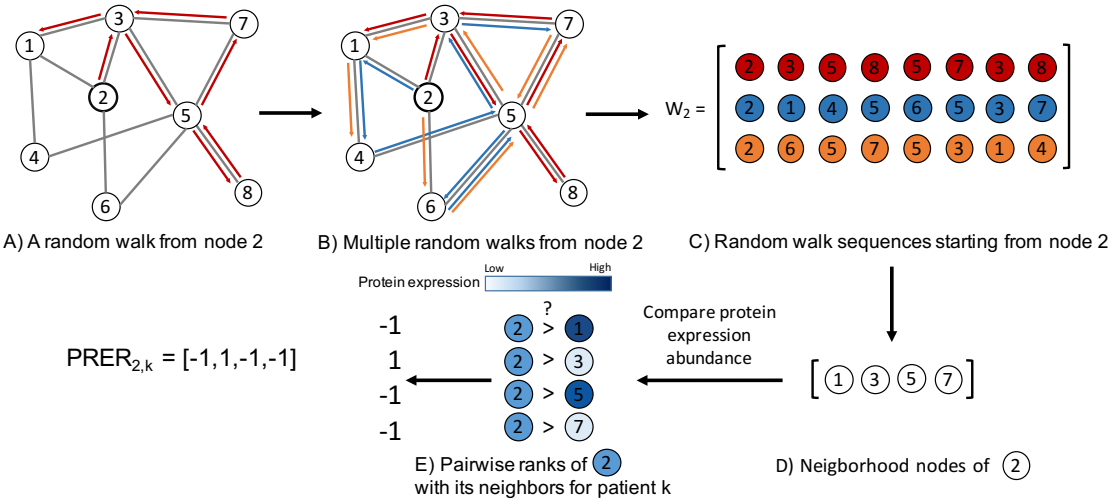


Figure 3.1: The pipeline of the pairwise rank representation which is based on random walks. By generating random walks on the graph, the neighborhood of node 2 is obtained. Then the pairwise comparison of the neighborhood proteins in terms of their protein expression quantities is used to form a representation of the protein.

3.3.1.2 Pairwise Rank Features

Using the above procedure, we define the neighborhood of a protein, N_i; however, some neighbors might be missing measurements. We define the subset of neighbor proteins with measurements $M_i = N_i \cap U$. Next, for a protein i, we generate pairwise rank features with every protein $i \in M_i$ as follows:

Let $X_i^{(k)}$ and $X_j^{(k)}$ denote the expression quantities for protein i and j for patient k. Protein i is the source protein, and protein j is a protein in the neighborhood of i. The pairwise rank expression embeddings (PRER) for this patient is defined as:

$$X_{i,j}^{(k)} = \begin{cases} 1 & if\ X_i^{(x)} > X_j^{(k)} \\ -1 & otherwise \end{cases} \tag{3.4}$$

$X_{i,j}^{(k)} = 1$ indicates that the molecule i is more upregulated with respect to molecule j for this patient, whereas $X_{i,j}^{(k)} = -1$ indicates otherwise. For every i in U and for every j in M_i, we define a pairwise rank protein. This representation constitutes a nonlinear interaction feature among original features which aims at capturing expression dysregulation among proteins that are potentially interacting. This serves as a node embedding that uses the node features, in this case, the protein expression patterns.

3.3.2 Survival Prediction

We apply the PRER representation for survival prediction. In this section, we will provide the background information on the survival models and the methods that we used. For each cancer type, the data is of the form, $D = \{\mathbf{X}^{(i)}, \mathbf{S}^{(i)}, \delta^{(i)}\}_{i=1}^{n}$; n is the number of patients. For each patient $\mathbf{X}$ is the derived features from protein expression data and $\mathbf{S}$ is the overall survival time and δ denotes censoring.

3.3.2.1 Survival Time and Censoring

The time interval between the beginning and the end of observation is called the survival time. The beginning can be the first visit or diagnosis of cancer. In our case, beginning refers to the patient's first visit to hospital whereas the end of observation is either date of death or the last follow-up which leads to censored survival time. Censoring means that the event does not happen within the corresponding time interval. In our case, this may include cases such as a patient gives up the treatment or death may not be related to cancer. It corresponds to cases with missing information about the real survival time of the patient. Censoring occurs in two ways: (1) right censored where the beginning is known but the observation of event does not take place within the corresponding time window, (2) left censored where the starting point is unknown. In this study, survival distributions of samples are right censored. For example, in Figure 3.2 s_2, s_3, s_6 are censored samples which means that the actual survival time is greater than censoring time because it is assumed that these patients left the treatment prematurely.

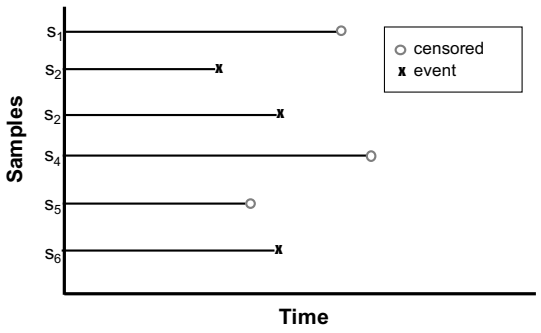


Figure 3.2: Each bar represents a subject in the study and their duration in the study. The open circles indicate that the data is right censored; the subject left the study before the event had occurred or the event had occurred due to a completely different reason. The black circles denote patients for whom the event had occurred within the duration of the study.

Censored survival data are not complete because the value of the time interval until the occurence of event (death) is partially known. It consists of two quantities:

$(1)\delta_i$ represents whether patient i is censored ($\delta_i = 1$) or not ($\delta_i = 0$). $(2)S_i$ represents the observed survival time. If we denote C_i^* as time until today since the beginning and T_i^* as the actual survival time, δ_i assigned as given at equation 3.5. On the other hand, S_i is defined as $min(T_i^*, C_i^*)$.

$$\delta_i = \begin{cases} 0 & if \quad \text{if} \quad T_i^* \leq C_i^* \\ 1 & otherwise \end{cases} \tag{3.5}$$

3.3.2.2 Hazard Function and Cumulative Hazard Function

Hazard function estimates the failure or death rate at time t conditional on survival until time t or later ($T \geq t$). Let $S(t)$ be a function that approximates the distribution of patients' survival times and p be the probability density function of survival time T. The probability that a patient is alive at time t ($t > 0$) is represented with survival function $S(t) = Pr[T > t]$. Let $\lambda(t) = p(t)/S(t)$ be the hazard function and it is found as $\lambda(t) = \lim_{\triangle t \to 0} Pr[t < T \leq t + \triangle t | T > t]/\triangle t$. Hazard function $\lambda(t)$ measures the instantaneous rate of failure [83]. The cumulative hazard function is the integral of the hazard function. It can be interpreted as the probability of an event at time t given that the patient has survived until time t

3.3.2.3 Random Survival Forest

Different methods have been proposed in the literature for survival prediction: (i)Non-parametric methods such as Random Survival Forests (RSF) [84], (ii) semi-parametric methods like Cox Proportional Hazard Rates (Cox-ph) [85] and (iii) parametric models with known distributions such as Weibull, Exponential or Gompertz [86].

In this study, we use Random Survival Forest(RSF)[84] for right-censored survival time predictions as it is a non-parametric method and has been shown to

perform well. RSF extends the Random Forest (RF) [87] method for survival analysis data. Random Forest is an ensemble method wherein the base learner is a tree and employs two randomization strategies. In RF, each tree is grown on a randomly drawn bootstrap sample. Furthermore, in growing a tree, at each node of the tree, a randomly selected subset of features is chosen as the candidate features for splitting. The randomization leads to good generalization errors. On the other hand, randomization in both feature space and sample space enable RF to catch the non-linear relations among the features. Apart from these advantages, RF is a non-parametric method which does not require any parameters to estimate and any prior assumption on the distribution of data.

In the RSF model, the node splitting approach is different. The node is split with the feature among the candidate features that maximizes survival difference between child nodes. Nodes are split until a specific number of death patients left in the given node, and this node becomes the leaf node. In each leaf node l of the tree, cumulative hazard function is estimated by Nelson-Aalen estimator 3.6 as follows:

$$\hat{\Lambda}_l(t) = \sum_{t_{j,l} \leq t} \frac{d_{j,l}}{R_{j,u}} \tag{3.6}$$

where $d_{j,u}$ is the number of deaths at time t and $R_{j,u}$ represents the number of patients at risk at time $t_{j,u}$ for leaf node u. This estimator is calculated for each leaf of the tree. The ensemble cumulative hazards for each patient is calculated by averaging the sum of all cumulative hazards at random trees.

3.3.2.4 Evaluation Criteria

To evaluate the survival models we used concordance index (c-index) [88]. C-index is the standard performance measure for model assessment in survival analysis and assesses the level of concordance between the pairwise ordering of patients based on their predicted survival times with the pairwise ordering based on their actual survivals. It can be interpreted as the fraction of all pairs of patients whose

predicted risks are correctly ranked among all permissible patient pairs. A pair is permissible if it is possible to tell which patient passed away first. Those are patient pairs, wherein patients are either both uncensored or within which one patient is censored but lived longer than a patient who has passed away. A value of 0.5 indicates no predictive discrimination and a value of 1.0 indicates the perfect ordering of the patients.

3.3.2.5 Kaplan-Meier Estimator

The Kaplan–Meier (KM) estimator is a non-parametric statistic to estimate the survival function. For unique each time points i in $\{t_1, \ldots, t_n\}$, where an event is observed, or a censoring take place, KM estimates the survival probability, that is the probability that the patient will live longer than t as follows:

$$for \quad t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_n, \quad \hat{S}(t) = \prod_{t_i < t} \frac{N_i - D_i}{N_i}, \tag{3.7}$$

Here, N_i is the number of individuals known to survive (have not yet had an event or been censored) at the time. D_i is the number of patients died at time point t_i.

We use KM plots to estimate the survival probabilities of the patients that take a specific feature value.

3.3.2.6 Log-rank Test

We also compare the patients with different feature values with respect to a protein pair using log-rank test [89]. For each group at each event time, log-rank test calculates the expected number of events occured since the previous event. These values are then summed over all event times. This summation gives the total expected number of events in each group (E_i for group i). The log-rank test compares the observed number of events(O_i for group i) to the expected number

of events (E_i) via the test statistic:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{3.8}$$

X^2 can be compared according to test statistic with a χ^2 distribution with $k-1$ degrees of freedom, where k is the number of groups.

3.3.3 Clinical and Molecular Patient Data

The Cancer Genome Atlas protein expression data and patient survival data are obtained from USCS Cancer Browser (`https://genome-cancer.ucsc.edu`) (April 11, 2017). The protein expression was quantified by reverse phase protein array (RPPA). The features in RPPA data is the expression values of multiple proteins as well as some phosphorylated versions of proteins. For example, RPPA data includes both STAT3 and STAT3PY705 where STAT3 is Signal Transducer And Activator Of Transcription 3 protein and STAT3PY705 is phosphorylation of STAT3 at tyrosine 705 residue. The data is collected for ten different cancer types: ovarian adenocarcinoma (OV), breast invasive carcinoma (BRCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), bladder urothelial carcinoma (BLCA), colon adenocarcinoma (COAD), uterine corpus endometrial carcinoma (UCEC). For each cancer type, the number of patients, who has deceased and who is censored are given at Table 3.1.

Number of Patients		
	censored	deceased
KIRC	297	156
OV	194	213
BRCA	628	99
LUAD	147	72
LUSC	106	78
COAD	260	60
HNSC	98	114
BLCA	70	52
UCEC	347	50
GBM	73	139

Table 3.1: Number of censored and deceased patients for each cancer type.

3.4 Results and Discussion

3.4.1 Prediction Performance

In order to understand the utility of PRER models, we build two types of RSF models for 10 different cancer types. In these two types of models, only the feature representations differ: in the first one, the individual expression values are input as features and in the second one the proposed PRER representation is used. For each cancer type, we randomly split the samples into two 100 times: 80% as the training set and 20% as the test set. In both cases, we perform a univariate feature selection step based on the hazard ratio of the univariate Cox model [90]. The likelihood ratio test p-values are used to assess the significance of hazard ratio; features with p-value ≤ 0.05 are retained for model training. For each cancer type and feature representations, 100 models are trained. Finally, the models are evaluated by the Concordance-Index (C-index) [88] on the test data.

The general pipeline of the model training and evaluation is detailed in Figure 3.3.

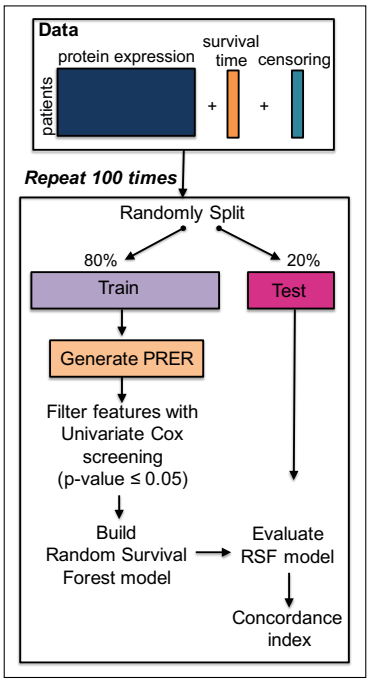


Figure 3.3: General pipeline for survival prediction. The step that involves generating PRER is skipped when the experiment is run with the alternative method of individual expression values.

Figure 3.4 compares the distribution of C-indices for 100 models trained with two feature representations for 10 different cancer types. In 8 of 10 cancer types, PRER representation yields significant improvements ($(p-value < 0.05)$ Wilcoxon signed-rank test). The C-index quantiles of 100 bootstrap results and corresponding *p-values* are shown in Table 3.2. The best improvements are found in *UCEC*, *BRCA* and *KIRC*. *GBM* and *LUAD* are the two cancer types with the least significant improvement. This could be due to the fact that these cancer types are associated with other genomic alterations.

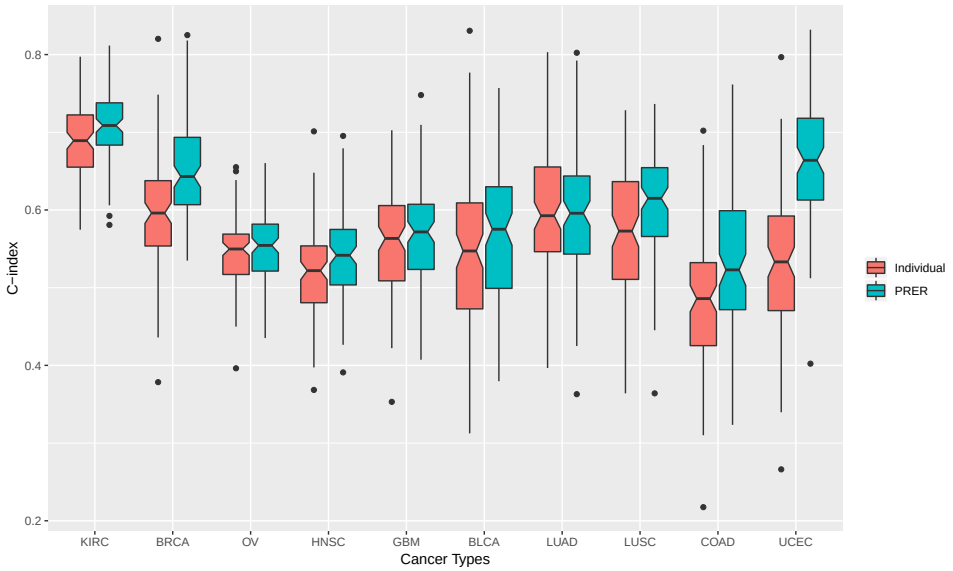


Figure 3.4: Comparison of RSF model performances that are trained with individual features and pairwise ranking embeddings(PRER) for different cancer types.

3.4.2 Predictive Feature Sets

We analyzed the features that emerge as important in the Random Survival Forest Models. The feature importances are assessed by the performance difference between the original features and the case where the feature vector is randomized. A large difference indicates a salient feature, whose absence leads to performance degradation [91]. The overall importance scores for each feature are calculated from the sum of scores over 100 bootstrap models. The top importance scores for each cancer type are shown in Figure 3.5, Appendix B.1, B.2, B.3, B.4, B.5, B.6, B.7, B.8, B.9.

As shown in Figure 3.5, some proteins repeatedly show up as partners in the informative feature list. To analyze this relationship, we form a network where the nodes are the participants of top 50 PRER proteins and edges are formed if a given protein pair are partners of a PRER representation. As shown in Figure 3.6 there are genes whose rank order with many different proteins appear as important.

Random Survival Forest C-indices						
	p-value	C-index				
		1st	2nd	median	3rd	4th
KIRC	**1.53x10^{-11}**	0.575	0.655	0.689	0.722	0.797
KIRC-prer		0.581	0.683	**0.709**	0.738	0.811
OV	**0.041**	0.396	0.517	0.549	0.568	0.655
OV-prer		0.435	0.521	**0.554**	0.582	0.661
BRCA	**9.41x10^{-11}**	0.378	0.554	0.596	0.637	0.820
BRCA-prer		0.535	0.607	**0.643**	0.693	0.825
LUAD	0.490	0.397	0.546	0.593	0.655	0.803
LUAD-prer		0.363	0.543	**0.596**	0.644	0.802
LUSC	**3.36x10^{-4}**	0.363	0.511	0.573	0.636	0.728
LUSC-prer		0.364	0.566	**0.615**	0.654	0.736
COAD	**4.67x10^{-5}**	0.218	0.423	0.486	0.532	0.702
COAD-prer		0.323	0.472	**0.522**	0.599	0.761
HNSC	**2.42x10^{-4}**	0.368	0.481	0.522	0.554	0.701
HNSC-prer		0.391	0.504	**0.542**	0.575	0.695
BLCA	**0.003**	0.312	0.473	0.547	0.609	0.831
BLCA-prer		0.380	0.499	**0.575**	0.630	0.757
UCEC	**3.38x10^{-17}**	0.266	0.470	0.533	0.592	0.797
UCEC-prer		0.402	0.613	**0.664**	0.718	0.832
GBM	0.136	0.353	0.509	0.563	0.606	0.703
GBM-prer		0.407	0.523	**0.572**	0.607	0.748

Table 3.2: Comparison of RSF model performances that are trained with individual features and pairwise ranking embeddings(PRER) for different cancer types. The methods are compared at the mean with four different quantile values of the 100 bootstrapped runs.

These are the hub nodes in the network. There are several studies that support that these genes have relevance to ovarian cancer. Below, we discuss this literature evidence.

Among the top features, Caveolin-1 (also known as CAV1) takes part in many biological processes inside the cell including survival, cell proliferation, cell migration and programmed cell death [92] and are reported to be dysregulated in different cancer types, Zhang and Luo (2016) [93] studied the gene expressions of ovarian cancer to find out prognostic genes and pathways related to chemotherapy resistance, and CAV1 is mentioned as one of the key genes for ovarian cancer in this study. An early study by Wiechen et al. (2001) [94] reports that CAV1 is dysregulated among the ovarian cancer patients based on microarray expression data.

The authors suggest that down-regulation of CAV1 leads to ovarian carcinoma since CAV1 is a tumor suppressor.

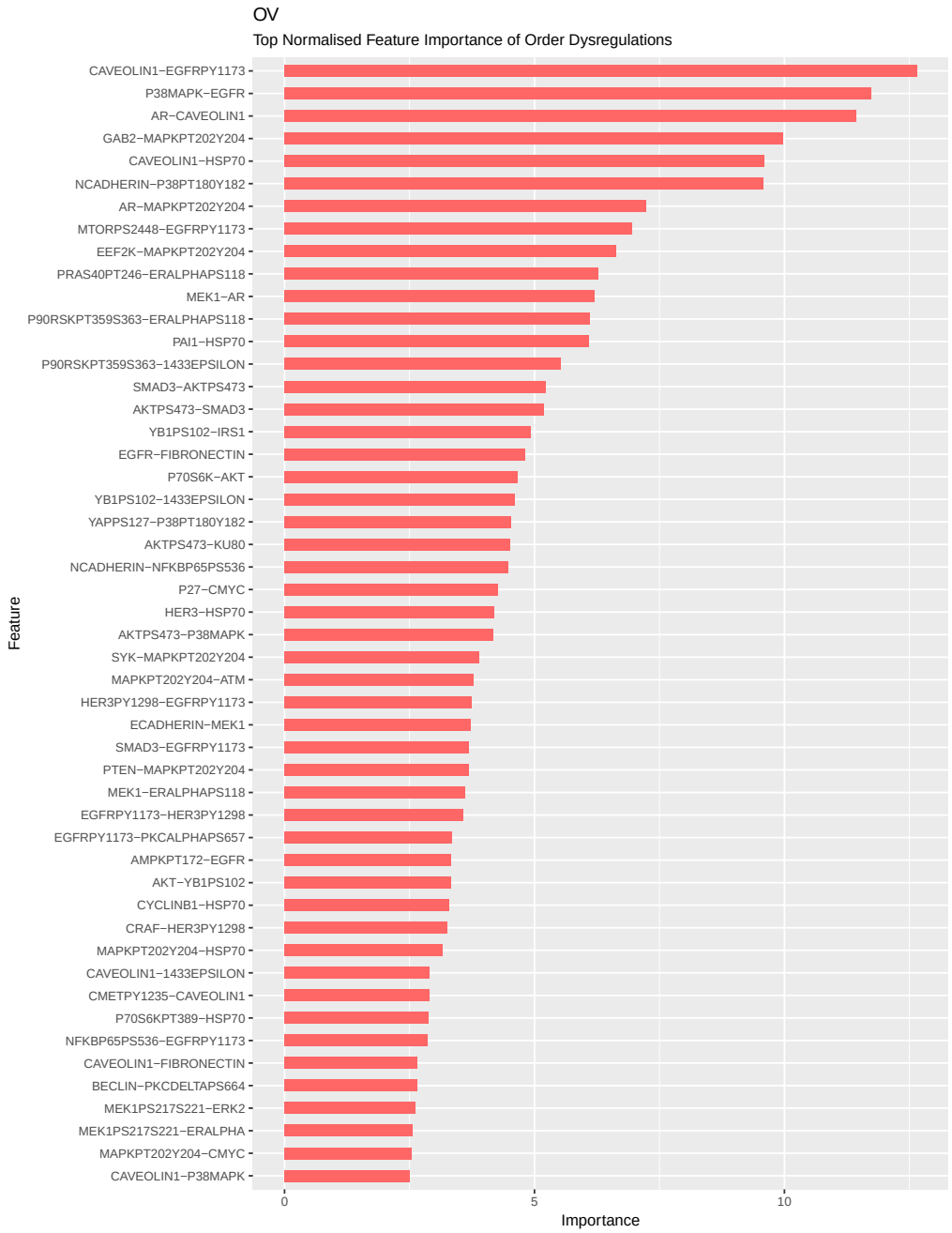


Figure 3.5: Variable importance of significant pairwise ranking embeddings for ovarian cancer.

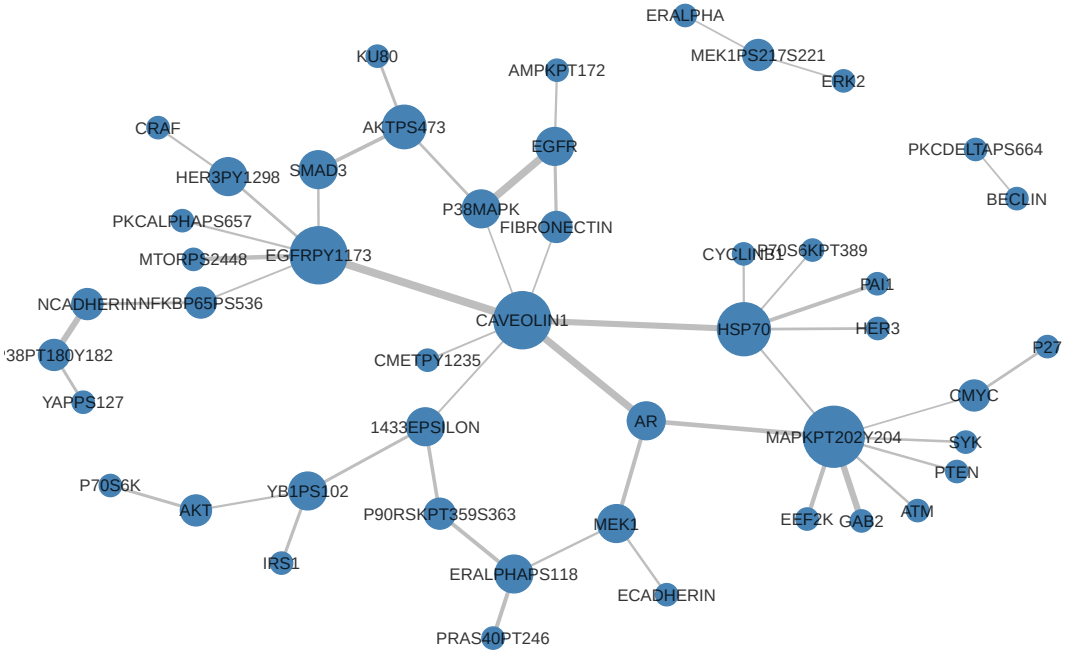


Figure 3.6: Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for ovarian cancer; edges represent that two proteins participate in a pairwise rank order feature together.

Epidermal growth factor receptor protein (EGFR) and its phosphorylation EGFRPY1173 are among the top features in PRER representation. EGFR is a receptor protein that receives and transmits signals from the environment to the cell. Marozkina et al. (2010) provides results that changes in expression of EGFR [95] may lead to ovarian carcinoma. Others [96, 97, 98] also claim that up-regulation of EGFR expression promotes ovarian cancer. Supporting this claim, EFGR is used as the target of drugs in therapies for many cancer types including ovarian cancer [99, 100].

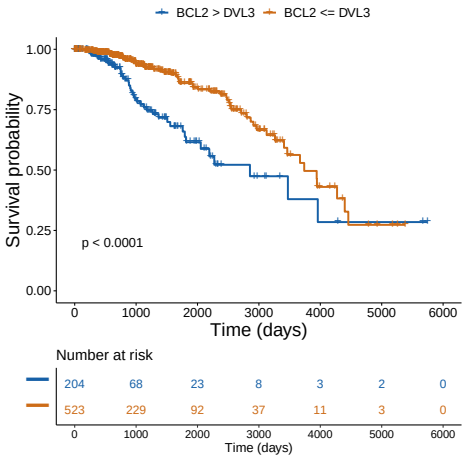
Dysregulations of mitogen-activated protein kinase 1 (MAPK1) and its phosphorylation MAPKPT202Y204 show significant impact in the survival prediction. MAPK1 takes part in migration, apoptosis, reproduction of cell and angiogenesis [101]. Zou et al. (2016) [102] mutation in MAPK1 may promote tumorigenesis

in ovarian epithelium. MAPK1 amplification is related to the growth of ovarian carcinoma and has a negative effect on survival of patients [103]. Similar to MAPK1, MEK1 is involved in MAP kinase signal transduction pathways, and it also involves similar cellular processes. It is considered as a prognostic biomarker for ovarian cancer. [104].

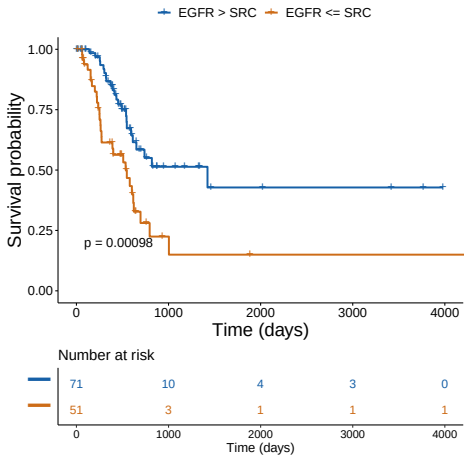
HSP70, heat shock protein 70 family, takes part in tumor development [105], and it is targeted to improve the treatment of ovarian cancer patients [106]. Androgen receptor (AR) activity has been shown related to the risk of being ovarian cancer because its activity increases the migration and growth of ovarian cancer cells [107]. Up-regulation of GRB2-associated binding protein 2 (GAB2) supports the development of tumors in ovarian cancer [108] and normal cells turn up to cancerous cells by amplification of GAB2 may result normal cells to turn [109]. However, we should note that many of the genes that goes into the RPPA assay in the TCGA study are selected due to their relevance to cancer. Thus, these important genes are likely to exhibit the individual importance of the PRER partners. Thus, we suggest an alternative way to analyze those features that emerge as important exclusively in the PRER in section 3.3.1.

3.4.3 Top Predictive PRER are Prognostic Biomarkers

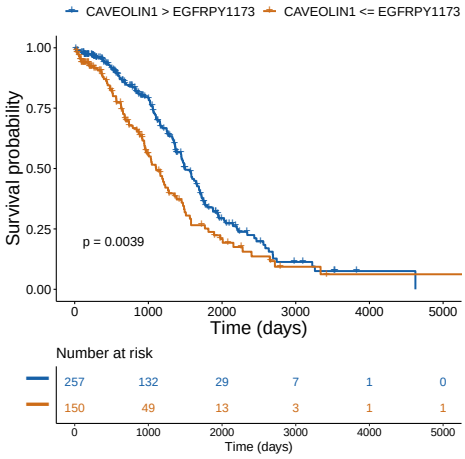
To check the relevance of the PRER genes, we check whether the patients with different feature values are different in terms of their survival distribution. For a PRER protein pair, A-B, we group patients such that those patients with protein expression values A¿B are in one group and the rest are in the other group. We then check if the survival distribution of these groups is different using the log-rank test. The Figure 3.7 shows the KM plots for each cancer type and the top-ranked feature.



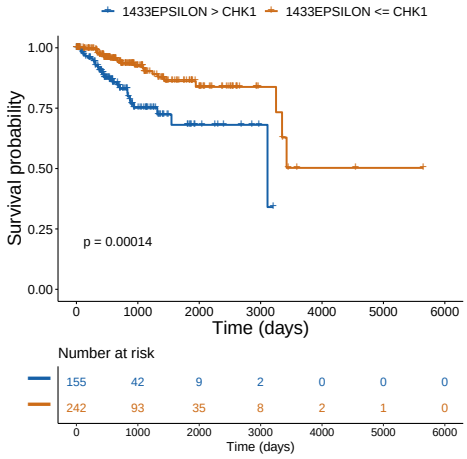
(a) BRCA



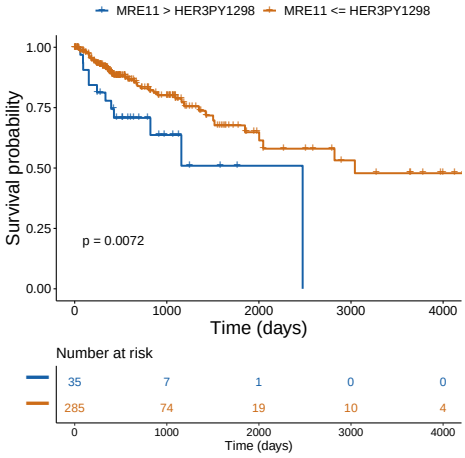
(b) BLCA



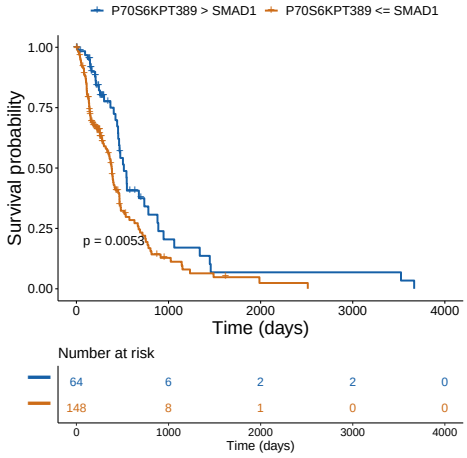
(c) OV



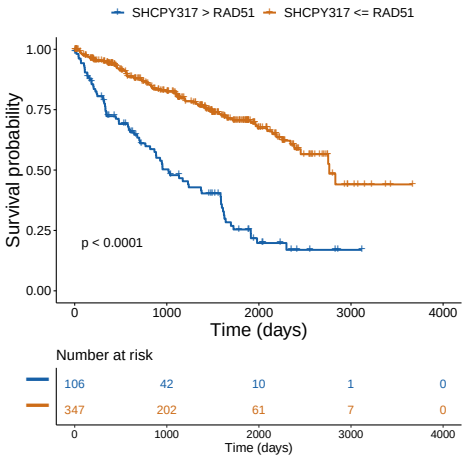
(d) UCEC



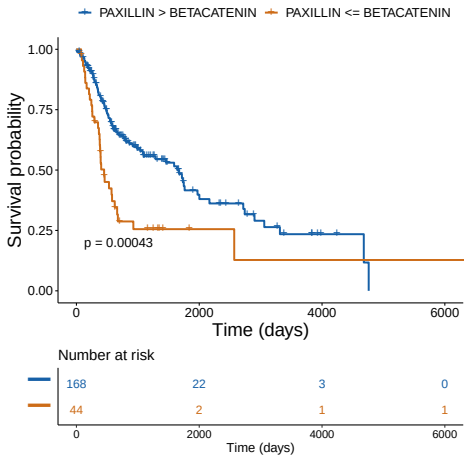
(e) COAD



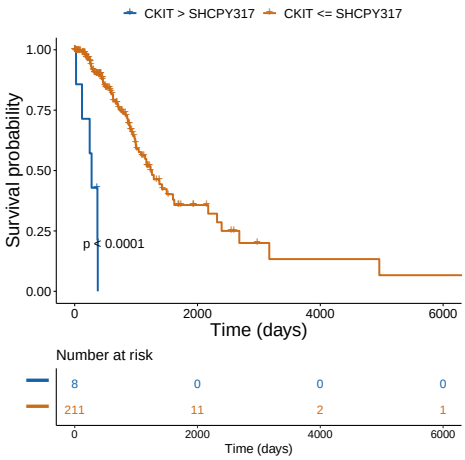
(f) GBM



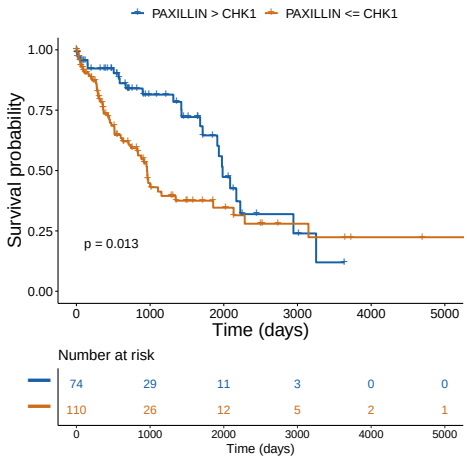
(g) KIRC



(h) HNSC



(i) LUAD



(j) LUSC

Figure 3.7: Kaplan-Meier plot for each cancer type based on overall survival. Number at risk denotes the number of patients at risk at a given time, and p-value is calculated with the log-rank test.

3.4.4 Proteins that Emerge as Important only in the PRER Representation

An important problem in the survival analysis tasks is to find a set of molecules whose up or down-regulation are associated with the survival likelihood of the patients. Since many of the proteins are cancer related, the list of features that are detected as important are already relevant to cancer. However, we are interested in finding proteins that emerge as important in the PRER representation and unimportant if their individual expression values are used exclusively. To achieve this, we first assign a feature importance to each protein in the PRER representation. As the feature importance is calculated for a pair of proteins, we calculate the feature importance of a protein by summing the importance of the PRER feature which it contributes. Let $F_{i,j}$ denotes the feature importance score of pairwise ranking between molecules i and j in the RSF model. Individual feature importance score for molecule i is calculated as follows:

$$S(i) = \frac{1}{\|N_i\|} \sum_{j \epsilon N_i} F_{i,j} \tag{3.9}$$

where N is the set of all pairwise ranking embeddings that include molecule i. $S(i)$ can be seen as the average of importances in the molecule i's neighborhood.

To find the features that are important, we look at the differences of feature ranks. Those features that gain importance in the PRER representation hint at proteins whose relations to other proteins carry prognostic value. Table 3.3 list the top-ranked features in term of the $S(i)$ score.

KIRC	BRCA	BLCA	GBM	OV
STAT3PY705	VEGFR2	PAI1	YAPPS127	PDK1PS241
PR	SRCPY527	XRCC1	YAP	SRCPY416
PKCALPHA	PDK1PS241	YAPPS127	RBPS807S811	RAD50
HSP70	YB1	YAP	INPP4B	STAT3PY705
CLAUDIN7	S6	VEGFR2	LKB1	INPP4B
BAK	ASNS	P27	P53	MTOR
VEGFR2	MTOR	P27PT157	PKCALPHAPS657	NF2
EIF4E	GATA3	PKCDELTAPS664	CLAUDIN7	YB1
CHK2PT68	P38PT180Y182	S6	ERALPHA	S6
PI3KP110ALPHA	SMAD1	P90RSKPT359S363	PCNA	SRCPY527

HNSC	LUAD	LUSC	COAD	UCEC
MTORPS2448	STAT5ALPHA	SMAD1	YB1	STAT5ALPHA
ASNS	S6PS240S244	SRCPY416	ASNS	S6PS240S244
YAPPS127	MTORPS2448	P70S6K	SRCPY416	PKCALPHAPS657
SRCPY527	RAD50	HER3	SMAD4	ASNS
S6PS240S244	TUBERIN	EEF2K	SMAD1	SRCPY416
YB1PS102	YB1PS102	RBPS807S811	STAT3PY705	PKCALPHA
PI3KP110ALPHA	S6PS235S236	CMETPY1235	RAD50	SRCPY527
PR	SRCPY416	STATHMIN	JNKPT183Y185	MAPKPT202Y204
PEA15	NOTCH1	YAP	P38PT180Y182	P38PT180Y182
HER3	YAPPS127	AKTPS473	YAPPS127	NCADHERIN

Table 3.3: Top-10 rank differentiated features in each cancer with PRER.

YAPPS127 and YAP proteins, which are encoded with YAP1 (Yes-associated protein 1) gene, found important in 7 distinct cancer types in Table 3.3. YAP1 is involved in the Hippo signaling pathway that is associated with growth, development and repair of the cells, and influences the survival of multiple cancers [110]. A recent study by Poma et al. (2018) [111] reports, in their extensive literature review, that 17 genes (out of 32) in the Hippo pathway have effects on survival in more than 20 different cancer types and conclude that YAP1 relevance on the survival of head and neck carcinoma, hepatocellular, lung adenocarcinoma, gastric, pancreatic and colorectal cancers. Further, other studies also suggest that survival for different cancer types is associated with the expression level of YAP1 and its differential expression is considered a biomarker for bladder urothelial carcinoma (BLCA) [112], breast invasive carcinoma (BRCA) [113, 114, 115, 116], ovarian

serous cystadenocarcinoma (OV) [117, 118].

ASNS (Asparagine Synthetase) is involved in the synthesis of asparagine [119], which is an amino acid used for controlling the functions of brain and nerve cells. ASNS protein expression [120, 121]. Sircar et al. (2012) [122] indicate that the upregulation of ASNS promotes the prostate cancer and using inhibitors against to asparagine may be a new choice in the treatment of prostate tumor cells. Similarly, Krall et al. (2016) [123] show that ASNS's product asparagine regulates the cells growth and balances the amino acids in cancer cells. ASNS activation improves the resistance of cells to drugs so that it might have prognostic value for radiochemotherapy for head and neck squamous cell carcinoma (HNSC) [124]. ASNS are also suggested as biomarker and target at treatments for breast cancer [125], and The Human Protein Atlas found ASNS as prognostic factor for renal (KIRC), liver, head and neck (HNSC) and endometrial (UCEC) cancers [126].

VEGFR2 (Vascular endothelial growth factor receptor 2), also known as kinase insert domain containing receptor (KDR), takes roles in vascular permeability regulation, vascular development and it is a biomarker for tumor angiogenesis so that it is a significant drug target for anticancer researches [127]. Neuchrist et al. (2001) [128] showed the expression of VEGFR2 is associated with head and neck cancer (HNSC) tumors. There are different studies in the literature that VEGFR2 plays a crucial role in the prognosis of breast cancer, and overexpression of VEGFR2 leads to lower survival in breast cancer patients [129, 130, 131]. The potential detrimental effect of VEGFR2 is reported different cancer types as well. For example, studies show that recurrence-free survivals of bladder cancer patients with overexpressed VEGFR2 are influenced negatively [132, 133, 134]. It is associated with bad prognosis in kidney cancer types, renal clear cell and renal papillary cell [135, 136, 137, 138].

The other two genes are the SMAD1 and SMAD4. Their activities are is important in the regulation of different signaling pathways, such as TGF-beta signaling

pathway which acts as tumor suppressor and this pathway was found to be associated with breast, colon, lung, prostate and pancreas cancers [139]. It is also found that there is a negative relation between the overall survival of colorectal cancer patients and loss of SMAD4 is associated with metastases [140, 141].

STAT3PY705 (STAT3 phosphorylation at tyrosine 705), phosphorylated state of STAT3 (Signal Transducer and Activator Of Transcription 3) protein, and STAT5ALPHA (Signal Transducer And Activator Of Transcription 5A) are both in STAT protein family. While STAT3PY705 is observed as significant in COAD, KIRC and OV, STAT5ALPHA shows up in LUAD and UCEC in Table 3.3. Experiments on mice show that STAT3 involved in cellular processes related to cellular proliferation, programmed cell death, movement and division and cell [142]. Activation in the STAT family, especially for STAT3 and STAT5, consistently occurs in several cancer cell lines including head and neck, breast, kidney, ovarian and colorectal[143, 144, 145, 146] because its activation contributes to the growth and survival of tumor cells.

YB1 its phosphorylation YB1PS102 shows correlation with many genes that have functions such as resistance to drugs, transcription and translation of cancerous cells [147]. Despite the downregulation of YB1 is found to be correlated with the reduction in progression, development of cell and programmed cell death at various cancer cell such as breast, colon, lung, prostate and pediatric glioblastoma by some studies [148, 149], there are studies [150, 151, 152, 153, 154] showing the association between overexpression of YB1 and different cancer types such as breast, colorectal, glioblastoma, lung, liver, ovarian cancers.

Chapter 4

Conclusion and Future Work

In this thesis, we use graph embedding methods on PPI to solve two different prediction tasks. The first task involves predicting gene essentiality based on node embeddings of the genes in the PPI. We propose the method GEGE. We learn a latent lower dimensional representation of the nodes in the PPI network with two different graph embedding methods, DeepWalk [41] and node2vec [42]. By applying machine learning algorithms to this new representation of the genes, we show that the gene essentiality can be predicted with high success. We compare these results, with a previously reported work that report results on the same dataset, GEGE, overperforms this method by 4%. We also compare our results to the alternative of representing each node with topological node features. Graph embeddings achieve significant improvements in all settings.

In our experiments, when compared to node2vec, DeepWalk embeddings achieve the best performance but their results are very close. The framework also allows for the addition of other gene features. When we augment the node embeddings with homology information, we observe performance improvements in all settings. We perform a robustness analysis with 100 random bootstrap samples which show that that the results are not affected by the selection of random test genes. We

also investigate the genes whose true labels are in the benchmark dataset yet are predicted as essential genes repeatedly in the 100 bootstrap samples. Some of these genes are reported to be conditionally essential genes; depending on the context they can be conditional. This work can be extended in different directions:

- The gene essentiality predictions can be tested for other organisms.
- In addition to homology information, other relevant features can be incorporated into this framework.

In the second part of the thesis, we propose a novel embedding method to represent proteins and molecular alterations on PPI. This method is based on pairwise comparison of the feature values of a node with the other proteins in its neighborhood. In this way, an embedding of the genes that reproduce the pairwise rank order of the molecular expression patterns with its potential interacting proteins are captured. The neighborhood of the nodes is generated with a biased random walk procedure. We call this feature representation Pairwise Ranking Embeddings with Random Walks (*PRER*). We generated this presentation with protein expression data as measured by the RPPA experiments in TCGA and input them to random forest survival models to predict survival in cancer patients for 10 different cancer types. When compared to the feature representation with individual protein expression values, same experiments are performed for both *PRER* with Random Survival Forest (RSF) model. Models are applied to 10 different cancer types, and the models with PRER representation outperform the models with individual molecules in each cancer type.

There may be multiple directions to follow as a future work that builds on the present studies in this thesis. These are listed as follows:

- PRER representation can be tested with other data types such as mRNA expression.

- The survival model can be improved with the addition of clinical features such as age, duration of the follow-up, cancer stage.
- Finally, we have not conducted a pan-cancer analysis and this is a noteworthy follow-up with potentially wider biological implications.

Bibliography

[1] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, *et al.*, "A large-scale evaluation of computational protein function prediction," *Nature methods*, vol. 10, no. 3, p. 221, 2013.

[2] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: a universal amplifier of genetic associations," *Nature Reviews Genetics*, vol. 18, no. 9, p. 551, 2017.

[3] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular systems biology*, vol. 3, no. 1, p. 140, 2007.

[4] G. Rancati, J. Moffat, A. Typas, and N. Pavelka, "Emerging and evolving concepts in gene essentiality," *Nature Reviews Genetics*, vol. 19, no. 1, p. 34, 2018.

[5] M. Itaya, "An estimation of minimal genome size required for life," *FEBS letters*, vol. 362, no. 3, pp. 257–260, 1995.

[6] A. R. Mushegian and E. V. Koonin, "A minimal gene set for cellular life derived by comparison of complete bacterial genomes," *Proceedings of the National Academy of Sciences*, vol. 93, no. 19, pp. 10268–10273, 1996.

[7] E. V. Koonin, "How many genes can make a cell: the minimal-gene-set concept," *Annual review of genomics and human genetics*, vol. 1, no. 1, pp. 99–116, 2000.

[8] X. Zhang, M. L. Acencio, and N. Lemke, "Predicting essential genes and proteins based on machine learning and network topological features: a comprehensive review," *Frontiers in physiology*, vol. 7, p. 75, 2016.

[9] M. Y. Galperin and E. V. Koonin, "Searching for drug targets in microbial genomes," *Current opinion in biotechnology*, vol. 10, no. 6, pp. 571–578, 1999.

[10] A. F. Chalker and R. D. Lunsford, "Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach," *Pharmacology & therapeutics*, vol. 95, no. 1, pp. 1–20, 2002.

[11] H. Farmer, N. McCabe, C. J. Lord, A. N. Tutt, D. A. Johnson, T. B. Richardson, M. Santarosa, K. J. Dillon, I. Hickson, C. Knights, *et al.*, "Targeting the dna repair defect in brca mutant cells as a therapeutic strategy," *Nature*, vol. 434, no. 7035, p. 917, 2005.

[12] N. J. O'Neil, M. L. Bailey, and P. Hieter, "Synthetic lethality and cancer," *Nature Reviews Genetics*, vol. 18, no. 10, p. 613, 2017.

[13] A. Cho, N. Haruyama, and A. B. Kulkarni, "Generation of transgenic mice," *Current protocols in cell biology*, vol. 42, no. 1, pp. 19–11, 2009.

[14] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, *et al.*, "Functional profiling of the saccharomyces cerevisiae genome," *nature*, vol. 418, no. 6896, p. 387, 2002.

[15] J. M. Silva, K. Marran, J. S. Parker, J. Silva, M. Golding, M. R. Schlabach, S. J. Elledge, G. J. Hannon, and K. Chang, "Profiling essential genes in human mammary cells by multiplex rnai screening," *Science*, vol. 319, no. 5863, pp. 617–620, 2008.

[16] T. Wang, K. Birsoy, N. W. Hughes, K. M. Krupczak, Y. Post, J. J. Wei, E. S. Lander, and D. M. Sabatini, "Identification and characterization of essential genes in the human genome," *Science*, vol. 350, no. 6264, pp. 1096–1101, 2015.

[17] L. W. Ning, H. Lin, H. Ding, J. Huang, N. N. M. Rao, and F.-B. Guo, "Predicting bacterial essential genes using only sequence composition information.," *Genetics and molecular research : GMR*, vol. 13 2, pp. 4564–72, 2014.

[18] W. C. Wei, L.-W. Ning, Y.-N. Ye, and F.-B. Guo, "Geptop: A gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny," in *PloS one*, 2013.

[19] F.-B. Guo, C. Dong, H.-L. Hua, S. Liu, H. Luo, H.-W. Zhang, Y.-T. Jin, and K.-Y. Zhang, "Accurate prediction of human essential genes using only nucleotide composition and association information," *Bioinformatics*, vol. 33 12, pp. 1758–1764, 2017.

[20] J. Deng, L. Deng, S. Su, M. Zhang, X. Lin, L. Wei, A. A. Minai, D. J. Hassett, and L. J. Lu, "Investigating the predictability of essential genes across distantly related organisms using an integrative approach," in *Nucleic acids research*, 2011.

[21] L. Chen, Y.-H. Zhang, S. Wang, Y. Zhang, T. Huang, and Y. D. Cai, "Prediction and analysis of essential genes using the enrichments of gene ontology and kegg pathways," in *PloS one*, 2017.

[22] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, p. 41, 2001.

[23] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, "The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics," *PLoS computational biology*, vol. 3, no. 4, p. e59, 2007.

[24] K. Plaimas, R. Eils, and R. König, "Identifying essential genes in bacterial metabolic networks with machine learning methods," *BMC systems biology*, vol. 4, no. 1, p. 56, 2010.

[25] J. P. M. da Silva, M. L. Acencio, J. C. M. Mombach, R. Vieira, J. C. da Silva, N. Lemke, and M. Sinigaglia, "In silico network topology-based prediction of gene essentiality," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 4, pp. 1049–1055, 2008.

[26] Y. Lu, J. Deng, J. C. Rhodes, H. Lu, and L. J. Lu, "Predicting essential genes for identifying potential drug targets in aspergillus fumigatus," *Computational biology and chemistry*, vol. 50, pp. 29–40, 2014.

[27] M. L. Acencio and N. Lemke, "Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information," *BMC Bioinformatics*, vol. 10, p. 290, Sep 2009.

[28] S. Coulomb, M. Bauer, D. Bernard, and M.-C. Marsolier-Kergoat, "Gene essentiality and the topology of protein interaction networks," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 272, no. 1573, pp. 1721–1725, 2005.

[29] X. He and J. Zhang, "Why do hubs tend to be essential in protein networks?," *PLoS genetics*, vol. 2, no. 6, p. e88, 2006.

[30] M. W. Hahn and A. D. Kern, "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks," *Molecular biology and evolution*, vol. 22, no. 4, pp. 803–806, 2004.

[31] N. N. Batada, L. D. Hurst, and M. Tyers, "Evolutionary and physiological importance of hub proteins," *PLoS computational biology*, vol. 2, no. 7, p. e88, 2006.

[32] Y. Chen and D. Xu, "Understanding protein dispensability through machine-learning analysis of high-throughput data," *Bioinformatics*, vol. 21, no. 5, pp. 575–581, 2004.

[33] E. Zotenko, J. Mestre, D. P. O'Leary, and T. M. Przytycka, "Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality," *PLoS computational biology*, vol. 4, no. 8, p. e1000140, 2008.

[34] Y.-C. Hwang, C.-C. Lin, J.-Y. Chang, H. Mori, H.-F. Juan, and H.-C. Huang, "Predicting essential genes based on network and sequence analysis," *Molecular BioSystems*, vol. 5, no. 12, pp. 1672–1678, 2009.

[35] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, "High-betweenness proteins in the yeast protein interaction network," *BioMed Research International*, vol. 2005, no. 2, pp. 96–103, 2005.

[36] J. Wang, M. Li, H. Wang, and Y. Pan, "Identification of essential proteins based on edge clustering coefficient," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 9, no. 4, pp. 1070–1080, 2012.

[37] J. Zhong, J. Wang, W. Peng, Z. Zhang, and Y. Pan, "Prediction of essential proteins based on gene expression programming," *BMC genomics*, vol. 14, no. 4, p. S7, 2013.

[38] J. Cheng, Z. Xu, W. Wu, L. Zhao, X. Li, Y. Liu, and S. Tao, "Training set selection for the prediction of essential genes," *PLoS ONE*, vol. 9, no. 1, 2014.

[39] J. Cheng, W. Wu, Y. Zhang, X. Li, X. Jiang, G. Wei, and S. Tao, "A new computational strategy for predicting essential genes," in *BMC Genomics*, 2013.

[40] M. C. Palumbo, A. Colosimo, A. Giuliani, and L. Farina, "Functional essentiality from topology features in metabolic networks: a case study in yeast.," *FEBS letters*, vol. 579 21, pp. 4642–6, 2005.

[41] B. Perozzi, R. Al-Rfou', and S. Skiena, "Deepwalk: Online learning of social representations," in *KDD*, 2014.

[42] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining*, vol. 2016, pp. 855–864, 2016.

[43] E. Palumbo, G. Rizzo, R. Troncy, E. Baralis, M. Osella, and E. Ferro, "Knowledge graph embeddings with node2vec for item recommendation," in *ESWC*, 2018.

[44] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[45] F. Lorrain and H. C. White, "Structural equivalence of individuals in social networks," *The Journal of mathematical sociology*, vol. 1, no. 1, pp. 49–80, 1971.

[46] R. Andersen, F. Chung, and K. Lang, "Local graph partitioning using pagerank vectors," in *null*, pp. 475–486, IEEE, 2006.

[47] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation," *IEEE Transactions on knowledge and data engineering*, vol. 19, no. 3, pp. 355–369, 2007.

[48] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[49] L. Bottou, "Stochastic gradient learning in neural networks," *Proceedings of Neuro-Nımes*, vol. 91, no. 8, p. 12, 1991.

[50] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[51] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, ACM, 1992.

[52] J. Leskovec and R. Sosič, "Snap: A general-purpose network analysis and graph-mining library," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 1, p. 1, 2016.

[53] T. Hart, M. Chandrashekhar, M. Aregger, Z. Steinhart, K. R. Brown, G. MacLeod, M. Mis, M. Zimmermann, A. Fradet-Turcotte, S. Sun, *et al.*, "High-resolution crispr screens reveal fitness genes and genotype-specific cancer liabilities," *Cell*, vol. 163, no. 6, pp. 1515–1526, 2015.

[54] V. A. Blomen, P. Májek, L. T. Jae, J. W. Bigenzahn, J. Nieuwenhuis, J. Staring, R. Sacco, F. R. van Diemen, N. Olk, A. Stukalov, *et al.*, "Gene essentiality and synthetic lethality in haploid human cells," *Science*, vol. 350, no. 6264, pp. 1092–1096, 2015.

[55] J. M. Silva, K. Marran, J. S. Parker, J. Silva, M. Golding, M. R. Schlabach, S. J. Elledge, G. J. Hannon, and K. Chang, "Profiling essential genes in human mammary cells by multiplex rnai screening," *Science*, vol. 319, no. 5863, pp. 617–620, 2008.

[56] R. Marcotte, K. R. Brown, F. Suarez, A. Sayad, K. Karamboulas, P. M. Krzyzanowski, F. Sircoulomb, M. Medrano, Y. Y. Fedyshyn, J. L. Y. Koh, D. van Dyk, B. Fedyshyn, M. Luhova, G. C. Brito, F. J. Vizeacoumar, F. S. Vizeacoumar, A. Datti, D. Kasimer, A. Buzina, P. Mero, C. Misquitta, J. Normand, M. Haider, T. Ketela, J. L. Wrana, R. Rottapel, B. G. Neel, and J. Moffat, "Essential gene profiles in breast, pancreatic, and ovarian cancer cells.," *Cancer discovery*, vol. 2 2, pp. 172–189, 2012.

[57] J. Luo, M. J. Emanuele, D. Li, C. J. Creighton, M. R. Schlabach, T. Westbrook, K. kin Wong, and S. J. Elledge, "A genome-wide rnai screen identifies multiple synthetic lethal interactions with the ras oncogene," *Cell*, vol. 137, pp. 835–848, 2009.

[58] K. Shedden, J. M. Taylor, S. A. Enkemann, M.-S. Tsao, T. J. Yeatman, W. L. Gerald, S. Eschrich, I. Jurisica, T. J. Giordano, D. E. Misek, *et al.*, "Gene

expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study," *Nature medicine*, vol. 14, no. 8, p. 822, 2008.

[59] Y. Yuan, E. M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, L. A. Byers, Y. Xu, K. R. Hess, L. Diao, *et al.*, "Assessing the clinical utility of cancer genomic and proteomic data across tumor types," *Nature biotechnology*, vol. 32, no. 7, p. 644, 2014.

[60] A. Fernandez-Teijeiro, R. A. Betensky, L. M. Sturla, J. Y. Kim, P. Tamayo, and S. L. Pomeroy, "Combining gene expression profiles and clinical parameters for risk stratification in medulloblastomas," *Journal of clinical oncology*, vol. 22, no. 6, pp. 994–998, 2004.

[61] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nature biotechnology*, vol. 27, no. 2, p. 199, 2009.

[62] A. P. Crijns, R. S. Fehrmann, S. de Jong, F. Gerbens, G. J. Meersma, H. G. Klip, H. Hollema, R. M. Hofstra, G. J. te Meerman, E. G. de Vries, *et al.*, "Survival-related profile, pathways, and transcription factors in ovarian cancer," *PLoS medicine*, vol. 6, no. 2, p. e1000024, 2009.

[63] W. Zhang, T. Ota, V. Shridhar, J. Chien, B. Wu, and R. Kuang, "Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment," *PLoS computational biology*, vol. 9, no. 3, p. e1002975, 2013.

[64] D. Kim, R. Li, S. M. Dudek, and M. D. Ritchie, "Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer," *Journal of biomedical informatics*, vol. 56, pp. 220–228, 2015.

[65] X. Li, Y. Zhang, Y. Zhang, J. Ding, K. Wu, and D. Fan, "Survival prediction of gastric cancer by a seven-microrna signature," *Gut*, vol. 59, no. 5, pp. 579–585, 2010.

[66] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, *et al.*, "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.

[67] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[68] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14863–14868, 1998.

[69] S. M. Pagnotta and M. Ceccarelli, "An algorithm for finding gene signatures supervised by survival time data," in *Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 568–578, Springer, 2011.

[70] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488–492, 2005.

[71] M. Zou, Z. Liu, X.-S. Zhang, and Y. Wang, "Ncc-auc: an auc optimization method to identify multi-biomarker panel for cancer prognosis from genomic and clinical data," *Bioinformatics*, vol. 31, no. 20, pp. 3330–3338, 2015.

[72] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.

[73] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Breast cancer survivability via adaboost algorithms," in *Proceedings of the second Australasian workshop on Health data and knowledge management-Volume 80*, pp. 55–64, Australian Computer Society, Inc., 2008.

[74] Y.-J. Lee, O. L. Mangasarian, and W. H. Wolberg, "Survival-time classification of breast cancer patients," *Computational Optimization and Applications*, vol. 25, no. 1-3, pp. 151–166, 2003.

[75] J.-S. Lee, I.-S. Chu, J. Heo, D. F. Calvisi, Z. Sun, T. Roskams, A. Durnez, A. J. Demetris, and S. S. Thorgeirsson, "Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling," *Hepatology*, vol. 40, no. 3, pp. 667–676, 2004.

[76] D. Kim, J.-G. Joung, K.-A. Sohn, H. Shin, Y. R. Park, M. D. Ritchie, and J. H. Kim, "Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction," *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. 109–120, 2015.

[77] C. Winter, G. Kristiansen, S. Kersting, J. Roy, D. Aust, T. Knösel, P. Rümmele, B. Jahnke, V. Hentrich, F. Rückert, *et al.*, "Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes," *PLoS computational biology*, vol. 8, no. 5, p. e1002511, 2012.

[78] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[79] H. Gómez-Rueda, E. Martínez-Ledesma, A. Martínez-Torteya, R. Palacios-Corona, and V. Trevino, "Integration and comparison of different genomic data for outcome prediction in cancer," *BioData mining*, vol. 8, no. 1, p. 1, 2015.

[80] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype-phenotype interactions," *Nature Reviews Genetics*, vol. 16, no. 2, pp. 85–97, 2015.

[81] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine, and C. Sander, "Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles," *PLoS One*, vol. 6, no. 11, p. e24709, 2011.

[82] L. Xu, L. Fengji, L. Changning, Z. Liangcai, L. Yinghui, L. Yu, C. Shanguang, and X. Jianghui, "Comparison of the prognostic utility of the diverse molecular data among lncrna, dna methylation, microrna, and mrna across five human cancers," *PloS one*, vol. 10, no. 11, p. e0142433, 2015.

[83] H. Steck, B. Krishnapuram, C. Dehing-oberije, P. Lambin, and V. C. Raykar, "On ranking in survival analysis: Bounds on the concordance index," in *Advances in neural information processing systems*, pp. 1209–1216, 2008.

[84] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *The Annals of Applied Statistics*, pp. 841–860, 2008.

[85] D. R. Cox, "Partial likelihood," *Biometrika*, vol. 62, no. 2, pp. 269–276, 1975.

[86] M. J. Bradburn, T. G. Clark, S. Love, and D. Altman, "Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods," *British journal of cancer*, vol. 89, no. 3, pp. 431–436, 2003.

[87] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[88] F. E. Harrell Jr, R. M. Califf, D. B. Pryor, K. L. Lee, R. A. Rosati, *et al.*, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[89] R. Peto and J. Peto, "Asymptotically efficient rank invariant test procedures," *Journal of the Royal Statistical Society. Series A (General)*, pp. 185–207, 1972.

[90] T. Therneau, "A package for survival analysis in s. r package version 2.37-4. 2013," 2013.

[91] H. Ishwaran *et al.*, "Variable importance in binary regression trees and forests," *Electronic Journal of Statistics*, vol. 1, pp. 519–537, 2007.

[92] C. Boscher and I. R. Nabi, "Caveolin-1: role in cell signaling," in *Caveolins and Caveolae*, pp. 29–50, Springer, 2012.

[93] M. Zhang and S. Luo, "Gene expression profiling of epithelial ovarian cancer reveals key genes and pathways associated with chemotherapy resistance," *Genet Mol Res*, vol. 15, no. 1, p. 11, 2016.

[94] K. Wiechen, L. Diatchenko, A. Agoulnik, K. M. Scharff, H. Schober, K. Arlt, B. Zhumabayeva, P. D. Siebert, M. Dietel, R. Schäfer, *et al.*, "Caveolin-1 is down-regulated in human ovarian carcinoma and acts as a candidate tumor suppressor gene," *The American journal of pathology*, vol. 159, no. 5, pp. 1635–1643, 2001.

[95] N. V. Marozkina, S. M. Stiefel, H. F. Frierson Jr, and S. J. Parsons, "Mmtv-egf receptor transgene promotes preneoplastic conversion of multiple steroid hormone-responsive tissues," *Journal of cellular biochemistry*, vol. 103, no. 6, pp. 2010–2018, 2008.

[96] I. Dimova, B. Zaharieva, S. Raitcheva, R. Dimitrov, N. Doganov, and D. Toncheva, "Tissue microarray analysis of egfr and erbb2 copy number changes in ovarian tumors," *International Journal of Gynecological Cancer*, vol. 16, no. 1, pp. 145–151, 2006.

[97] J. V. Ilekis, J. P. Connor, G. S. Prins, K. Ferrer, C. Niederberger, and B. Scoccia, "Expression of epidermal growth factor and androgen receptors in ovarian cancer," *Gynecologic oncology*, vol. 66, no. 2, pp. 250–254, 1997.

[98] I. Skirnisdóttir, B. Sorbe, and T. Seidal, "The growth factor receptors her-2/neu and egfr, their relationship, and their effects on the prognosis in early stage (figo i-ii) epithelial ovarian carcinoma," *International Journal of Gynecological Cancer*, vol. 11, no. 2, pp. 119–129, 2001.

[99] L. G. Hudson, R. Zeineldin, M. Silberberg, and M. S. Stack, "Activated epidermal growth factor receptor in ovarian cancer," in *Ovarian Cancer*, pp. 203–226, Springer, 2009.

[100] J. A. Wilken, T. Badri, S. Cross, R. Raji, A. D. Santin, P. Schwartz, A. J. Branscum, A. T. Baron, A. I. Sakhitab, and N. J. Maihle, "Egfr/her-targeted therapeutics in ovarian cancer," *Future medicinal chemistry*, vol. 4, no. 4, pp. 447–469, 2012.

[101] A. Handra-Luca, H. Bilal, J.-C. Bertrand, and P. Fouret, "Extra-cellular signal-regulated erk-1/erk-2 pathway activation in human salivary gland mucoepidermoid carcinoma: association to aggressive tumor behavior and tumor cell proliferation," *The American journal of pathology*, vol. 163, no. 3, pp. 957–967, 2003.

[102] Y. Zou, W. Deng, F. Wang, X.-H. Yu, F.-Y. Liu, B.-C. Yang, M.-Z. Huang, J.-B. Guo, Q.-H. Xie, M. He, *et al.*, "A novel somatic mapk1 mutation in primary ovarian mixed germ cell tumors," *Oncology reports*, vol. 35, no. 2, pp. 725–730, 2016.

[103] M. T. Rahman, K. Nakayama, M. Rahman, H. Katagiri, A. Katagiri, T. Ishibashi, M. Ishikawa, E. Sato, K. Iida, N. Nakayama, *et al.*, "Kras and mapk1 gene amplification in type ii ovarian carcinomas," *International journal of molecular sciences*, vol. 14, no. 7, pp. 13748–13762, 2013.

[104] Z. Pénzváltó, A. Lánczky, J. Lénárt, N. Meggyesházi, T. Krenács, N. Szoboszlai, C. Denkert, I. Pete, and B. Győrffy, "Mek1 is associated with carboplatin resistance and is a prognostic biomarker in epithelial ovarian cancer," *BMC cancer*, vol. 14, no. 1, p. 837, 2014.

[105] N. Gupta, N. Jagadish, A. Surolia, and A. Suri, "Heat shock protein 70-2 (hsp70-2) a novel cancer testis antigen that promotes growth of ovarian cancer," *American journal of cancer research*, vol. 7, no. 6, p. 1252, 2017.

[106] H. Hatakeyama, S. Y. Wu, M. S. Lingegowda, C. Rodriguez-Aguayo, G. Lopez-Berestein, L. Ju-Seog, C. Rinaldi, E. J. Juan, A. K. Sood, M. Torres-Lugo, *et al.*, "Hsp70 inhibition synergistically enhances the effects of magnetic fluid hyperthermia in ovarian cancer," *Molecular cancer therapeutics*, pp. molcanther–0519, 2017.

[107] H. Zhu, X. Zhu, L. Zheng, X. Hu, L. Sun, and X. Zhu, "The role of the androgen receptor in ovarian cancer carcinogenesis and its clinical implications," *Oncotarget*, vol. 8, no. 17, p. 29395, 2017.

[108] C. Duckworth, L. Zhang, S. Carroll, S. Ethier, and H. Cheung, "Overexpression of gab2 in ovarian cancer cells promotes tumor growth and angiogenesis by upregulating chemokine expression," *Oncogene*, vol. 35, no. 31, p. 4036, 2016.

[109] G. P. Dunn, H. W. Cheung, P. K. Agarwalla, S. Thomas, Y. Zektser, A. M. Karst, J. S. Boehm, B. A. Weir, A. M. Berlin, L. Zou, *et al.*, "In vivo multiplexed interrogation of amplified genes identifies gab2 as an ovarian cancer oncogene," *Proceedings of the National Academy of Sciences*, vol. 111, no. 3, pp. 1102–1107, 2014.

[110] E. Lorenzetto, M. Brenca, M. Boeri, C. Verri, E. Piccinin, P. Gasparini, F. Facchinetti, S. Rossi, G. Salvatore, M. Massimino, *et al.*, "Yap1 acts as oncogenic target of 11q22 amplification in multiple cancer subtypes," *Oncotarget*, vol. 5, no. 9, p. 2608, 2014.

[111] A. M. Poma, L. Torregrossa, R. Bruno, F. Basolo, and G. Fontanini, "Hippo pathway affects survival of cancer patients: extensive analysis of tcga data and review of literature," *Scientific reports*, vol. 8, no. 1, p. 10623, 2018.

[112] J.-Y. Liu, Y.-H. Li, H.-X. Lin, Y.-J. Liao, S.-J. Mai, Z.-W. Liu, Z.-L. Zhang, L.-J. Jiang, J.-X. Zhang, H.-F. Kung, *et al.*, "Overexpression of yap 1 contributes to progressive features and poor prognosis of human urothelial carcinoma of the bladder," *BMC cancer*, vol. 13, no. 1, p. 349, 2013.

[113] F. Cheng, J. Zhao, A. B. Hanker, M. R. Brewer, C. L. Arteaga, and Z. Zhao, "Transcriptome-and proteome-oriented identification of dysregulated eif4g, stat3, and hippo pathways altered by pik3ca h1047r in her2/er-positive breast cancer," *Breast cancer research and treatment*, vol. 160, no. 3, pp. 457–474, 2016.

[114] L. Cao, P.-L. Sun, M. Yao, M. Jia, and H. Gao, "Expression of yes-associated protein (yap) and its clinical significance in breast cancer tissues," *Human pathology*, vol. 68, pp. 166–174, 2017.

[115] S. K. Kim, W. H. Jung, and J. S. Koo, "Yes-associated protein (yap) is differentially expressed in tumor and stroma according to the molecular subtype of breast cancer," *International journal of clinical and experimental pathology*, vol. 7, no. 6, p. 3224, 2014.

[116] H. M. Kim, W. H. Jung, and J. S. Koo, "Expression of yes-associated protein (yap) in metastatic breast cancer," *International journal of clinical and experimental pathology*, vol. 8, no. 9, p. 11248, 2015.

[117] C. He, X. Lv, G. Hua, S. M. Lele, S. Remmenga, J. Dong, J. S. Davis, and C. Wang, "Yap forms autocrine loops with the erbb pathway to regulate ovarian cancer initiation and progression," *Oncogene*, vol. 34, no. 50, p. 6040, 2015.

[118] Y. Xia, T. Chang, Y. Wang, Y. Liu, W. Li, M. Li, and H.-Y. Fan, "Yap promotes ovarian cancer cell tumorigenesis and is indicative of a poor prognosis for ovarian cancer patients," *PloS one*, vol. 9, no. 3, p. e91770, 2014.

[119] C. L. Lomelino, J. T. Andring, R. McKenna, and M. S. Kilberg, "Asparagine synthetase: Function, structure, and role in disease," *Journal of Biological Chemistry*, pp. jbc–R117, 2017.

[120] P. L. Lorenzi, W. C. Reinhold, M. Rudelius, M. Gunsior, U. Shankavaram, K. J. Bussey, U. Scherf, G. S. Eichler, S. E. Martin, K. Chin, *et al.*, "Asparagine synthetase as a causal, predictive biomarker for l-asparaginase activity in ovarian cancer cells," *Molecular cancer therapeutics*, vol. 5, no. 11, pp. 2613–2623, 2006.

[121] P. L. Lorenzi, J. Llamas, M. Gunsior, L. Ozbun, W. C. Reinhold, S. Varma, H. Ji, H. Kim, A. A. Hutchinson, E. C. Kohn, *et al.*, "Asparagine synthetase is a predictive biomarker of l-asparaginase activity in ovarian cancer cell lines," *Molecular cancer therapeutics*, vol. 7, no. 10, pp. 3123–3128, 2008.

[122] K. Sircar, H. Huang, L. Hu, D. Cogdell, J. Dhillon, V. Tzelepi, E. Efstathiou, I. H. Koumakpayi, F. Saad, D. Luo, *et al.*, "Integrative molecular profiling reveals asparagine synthetase is a target in castration-resistant prostate cancer," *The American journal of pathology*, vol. 180, no. 3, pp. 895–903, 2012.

[123] A. S. Krall, S. Xu, T. G. Graeber, D. Braas, and H. R. Christofk, "Asparagine promotes cancer cell proliferation through use as an amino acid exchange factor," *Nature communications*, vol. 7, p. 11457, 2016.

[124] I. Summerer, J. Hess, A. Pitea, K. Unger, L. Hieber, M. Selmansberger, K. Lauber, and H. Zitzelsberger, "Integrative analysis of the microrna-mrna response to radiochemotherapy in primary head and neck squamous cell carcinoma cells," *BMC genomics*, vol. 16, no. 1, p. 654, 2015.

[125] H. Yang, X. He, Y. Zheng, W. Feng, X. Xia, X. Yu, and Z. Lin, "Down-regulation of asparagine synthetase induces cell cycle arrest and inhibits cell proliferation of breast cancer," *Chemical biology & drug design*, vol. 84, no. 5, pp. 578–584, 2014.

[126] "The human protein atlas: Asns."

[127] N. R. Smith, D. Baker, N. H. James, K. Ratcliffe, M. Jenkins, S. E. Ashton, G. Sproat, R. Swann, N. Gray, A. Ryan, *et al.*, "Vascular endothelial growth factor receptors vegfr-2 and vegfr-3 are localized primarily to the vasculature in human primary solid cancers," *Clinical Cancer Research*, pp. 1078–0432, 2010.

[128] C. Neuchrist, B. M. Erovic, A. Handisurya, G. E. Steiner, P. Rockwell, C. Gedlicka, and M. Burian, "Vascular endothelial growth factor receptor 2 (vegfr2) expression in squamous cell carcinomas of the head and neck," *The Laryngoscope*, vol. 111, no. 10, pp. 1834–1841, 2001.

[129] J.-D. Yan, Y. Liu, Z.-Y. Zhang, G.-Y. Liu, J.-H. Xu, L.-Y. Liu, and Y.-M. Hu, "Expression and prognostic significance of vegfr-2 in breast cancer," *Pathology-Research and Practice*, vol. 211, no. 7, pp. 539–543, 2015.

[130] S. Guo, L. S. Colbert, M. Fuller, Y. Zhang, and R. R. Gonzalez-Perez, "Vascular endothelial growth factor receptor-2 in breast cancer," *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, vol. 1806, no. 1, pp. 108–121, 2010.

[131] S. Ghosh, C. A. Sullivan, M. P. Zerkowski, A. M. Molinaro, D. L. Rimm, R. L. Camp, and G. G. Chung, "High levels of vascular endothelial growth factor and its receptors (vegfr-1, vegfr-2, neuropilin-1) are associated with worse outcome in breast cancer," *Human pathology*, vol. 39, no. 12, pp. 1835–1843, 2008.

[132] P. C. Black and C. P. Dinney, "Bladder cancer angiogenesis and metastasis—translation from murine model to clinical trial," *Cancer and Metastasis Reviews*, vol. 26, no. 3-4, p. 623, 2007.

[133] A. Verma, J. DeGrado, A. B. Hittelman, M. A. Wheeler, H. Z. Kaimakliotis, and R. M. Weiss, "Effect of mitomycin c on concentrations of vascular endothelial growth factor and its receptors in bladder cancer cells and in bladders of rats intravesically instilled with mitomycin c," *BJU international*, vol. 107, no. 7, pp. 1154–1161, 2011.

[134] P. K. Kopparapu, S. A. Boorjian, B. D. Robinson, M. Downes, L. J. Gudas, N. P. Mongan, and J. L. Persson, "Expression of vegf and its receptors vegfr1/vegfr2 is associated with invasiveness of bladder cancer," *Anticancer research*, vol. 33, no. 6, pp. 2381–2390, 2013.

[135] A. Takahashi, H. Sasaki, S. J. Kim, K.-i. Tobisu, T. Kakizoe, T. Tsukamoto, Y. Kumamoto, T. Sugimura, and M. Terada, "Markedly increased amounts of messenger rnas for vascular endothelial growth factor and placenta growth factor in renal cell carcinoma associated with angiogenesis," *Cancer Research*, vol. 54, no. 15, pp. 4233–4237, 1994.

[136] D. Nicol, S.-I. Hii, M. Walsh, B. Teh, L. Thompson, C. Kennett, and D. Gotley, "Vascular endothelial growth factor expression is increased in renal cell carcinoma," *The Journal of urology*, vol. 157, no. 4, pp. 1482–1486, 1997.

[137] M. Tomisawa, T. Tokunaga, Y. Oshika, T. Tsuchida, Y. Fukushima, H. Sato, H. Kijima, H. Yamazaki, Y. Ueyama, N. Tamaoki, *et al.*, "Expression pattern of vascular endothelial growth factor isoform is closely correlated with tumour stage and vascularisation in renal cell carcinoma," *European Journal of Cancer*, vol. 35, no. 1, pp. 133–137, 1999.

[138] B. J. Ljungberg, J. Jacobsen, S. H. Rudolfsson, G. Lindh, K. Grankvist, and T. Rasmuson, "Different vascular endothelial growth factor (vegf), vegf-receptor 1 and-2 mrna expression profiles between clear cell and papillary renal cell carcinoma," *BJU international*, vol. 98, no. 3, pp. 661–667, 2006.

[139] D. Samanta and P. K. Datta, "Alterations in the smad pathway in human cancers," *Frontiers in bioscience (Landmark edition)*, vol. 17, p. 1281, 2012.

[140] M. Miyaki, T. Iijima, M. Konishi, K. Sakai, A. Ishii, M. Yasuno, T. Hishima, M. Koike, N. Shitara, T. Iwama, *et al.*, "Higher frequency of smad4 gene mutation in human colorectal cancer with distant metastasis," *Oncogene*, vol. 18, no. 20, p. 3098, 1999.

[141] L. Losi, H. Bouzourene, and J. Benhattar, "Loss of smad4 expression predicts liver metastasis in human colorectal cancer," *Oncology reports*, vol. 17, no. 5, pp. 1095–1099, 2007.

[142] S. Akira, "Roles of stat3 defined by tissue-specific gene targeting," *Oncogene*, vol. 19, no. 21, p. 2607, 2000.

[143] R. Buettner, L. B. Mora, and R. Jove, "Activated stat signaling in human tumors provides novel molecular targets for therapeutic intervention," *Clinical cancer research*, vol. 8, no. 4, pp. 945–954, 2002.

[144] H. Yu and R. Jove, "The stats of cancer—new molecular targets come of age," *Nature Reviews Cancer*, vol. 4, no. 2, p. 97, 2004.

[145] A. Lavecchia, C. Di Giovanni, and E. Novellino, "Stat-3 inhibitors: state of the art and new horizons for cancer treatment," *Current medicinal chemistry*, vol. 18, no. 16, pp. 2359–2375, 2011.

[146] I. Souissi, I. Najjar, L. Ah-Koon, P. O. Schischmanoff, D. Lesage, S. Le Coquil, C. Roger, I. Dusanter-Fourt, N. Varin-Blank, A. Cao, *et al.*, "A stat3-decoy oligonucleotide induces cell death in a human colorectal carcinoma cell line by blocking nuclear transfer of stat3 and stat3-bound nf-κb," *BMC cell biology*, vol. 12, no. 1, p. 14, 2011.

[147] Y. Basaki, K.-i. Taguchi, H. Izumi, Y. Murakami, T. Kubo, F. Hosoi, K. Watari, K. Nakano, H. Kawaguchi, S. Ohno, *et al.*, "Y-box binding protein-1 (yb-1) promotes cell cycle progression through cdc6-dependent pathway in human cancer cells," *European journal of cancer*, vol. 46, no. 5, pp. 954–965, 2010.

[148] Y. Basaki, F. Hosoi, Y. Oda, A. Fotovati, Y. Maruyama, S. Oie, M. Ono, H. Izumi, K. Kohno, K. Sakai, *et al.*, "Akt-dependent nuclear localization of y-box-binding protein 1 in acquisition of malignant characteristics by human ovarian cancer cells," *Oncogene*, vol. 26, no. 19, p. 2736, 2007.

[149] A. Lasham, W. Samuel, H. Cao, R. Patel, R. Mehta, J. L. Stern, G. Reid, A. G. Woolley, L. D. Miller, M. A. Black, *et al.*, "Yb-1, the e2f pathway, and regulation of tumor cell growth," *Journal of the National Cancer Institute*, vol. 104, no. 2, pp. 133–146, 2011.

[150] R. C. Bargou, K. Jürchott, C. Wagener, S. Bergmann, S. Metzner, K. Bommert, M. Y. Mapara, K.-J. Winzer, M. Dietel, B. Dörken, *et al.*, "Nuclear localization and increased levels of transcription factor yb-1 in primary human breast cancers are associated with intrinsic mdr1 gene expression," *Nature medicine*, vol. 3, no. 4, p. 447, 1997.

[151] T. Kamura, H. Yahata, S. Amada, S. Ogawa, T. Sonoda, H. Kobayashi, M. Mitsumoto, K. Kohno, M. Kuwano, and H. Nakano, "Is nuclear expression of y box-binding protein-1 a new prognostic factor in ovarian serous adenocarcinoma?," *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 85, no. 11, pp. 2450–2454, 1999.

[152] K. Shibao, H. Takano, Y. Nakayama, K. Okazaki, N. Nagata, H. Izumi, T. Uchiumi, M. Kuwano, K. Kohno, and H. Itoh, "Enhanced coexpression of yb-1 and dna topoisomerase ii α genes in human colorectal carcinomas," *International journal of cancer*, vol. 83, no. 6, pp. 732–737, 1999.

[153] K. Shibahara, K. Sugio, T. Osaki, T. Uchiumi, Y. Maehara, K. Kohno, K. Yasumoto, K. Sugimachi, and M. Kuwano, "Nuclear expression of the y-box binding protein, yb-1, as a novel marker of disease progression in non-small cell lung cancer," *Clinical cancer research*, vol. 7, no. 10, pp. 3151–3155, 2001.

[154] M. Yasen, K. Kajino, S. Kano, H. Tobita, J. Yamamoto, T. Uchiumi, S. Kon, M. Maeda, G. Obulhasim, S. Arii, *et al.*, "The up-regulation of y-box binding proteins (dna binding protein a and y-box binding protein-1) as prognostic markers of hepatocellular carcinoma," *Clinical cancer research*, vol. 11, no. 20, pp. 7354–7361, 2005.

Appendix A

List of Abbreviations

List of abbreviations used in this study is given as follows:

- **AP** : Average precision
- **AUC** : Area under curve
- **BFS** : Breadth-first Sampling
- **BLCA** : Bladder urothelial carcinoma
- **BRCA** : Breast invasive carcinoma
- **C-index** : Concordance Index
- **COAD** : Colon adenocarcinoma
- **Cox-ph** : Cox proportional hazard
- **CRISPR** : Clustered regularly interspaced short palindromic repeats
- **DFS** : Depth-first Sampling
- **HNSC** : Head and neck squamous cell carcinoma

- **GBM** : Glioblastoma multiforme
- **KIRC** : Kidney renal clear cell carcinoma
- **LUAD** : Lung adenocarcinoma
- **LUSC** : Lung squamous cell carcinoma
- **OV** : Ovarian adenocarcinoma
- **PRER** : Pairwise Rank Expression embeddings with Random walks
- **PPI** : Protein-protein interaction
- **RPPA** : Reverse phase protein array
- **RSF** : Random survival forest
- **SGD** : Stochastic gradient descent
- **SVM** : Support vector machine
- **UCEC** : Uterine corpus endometrial carcinoma

Appendix B

Important PRER Features

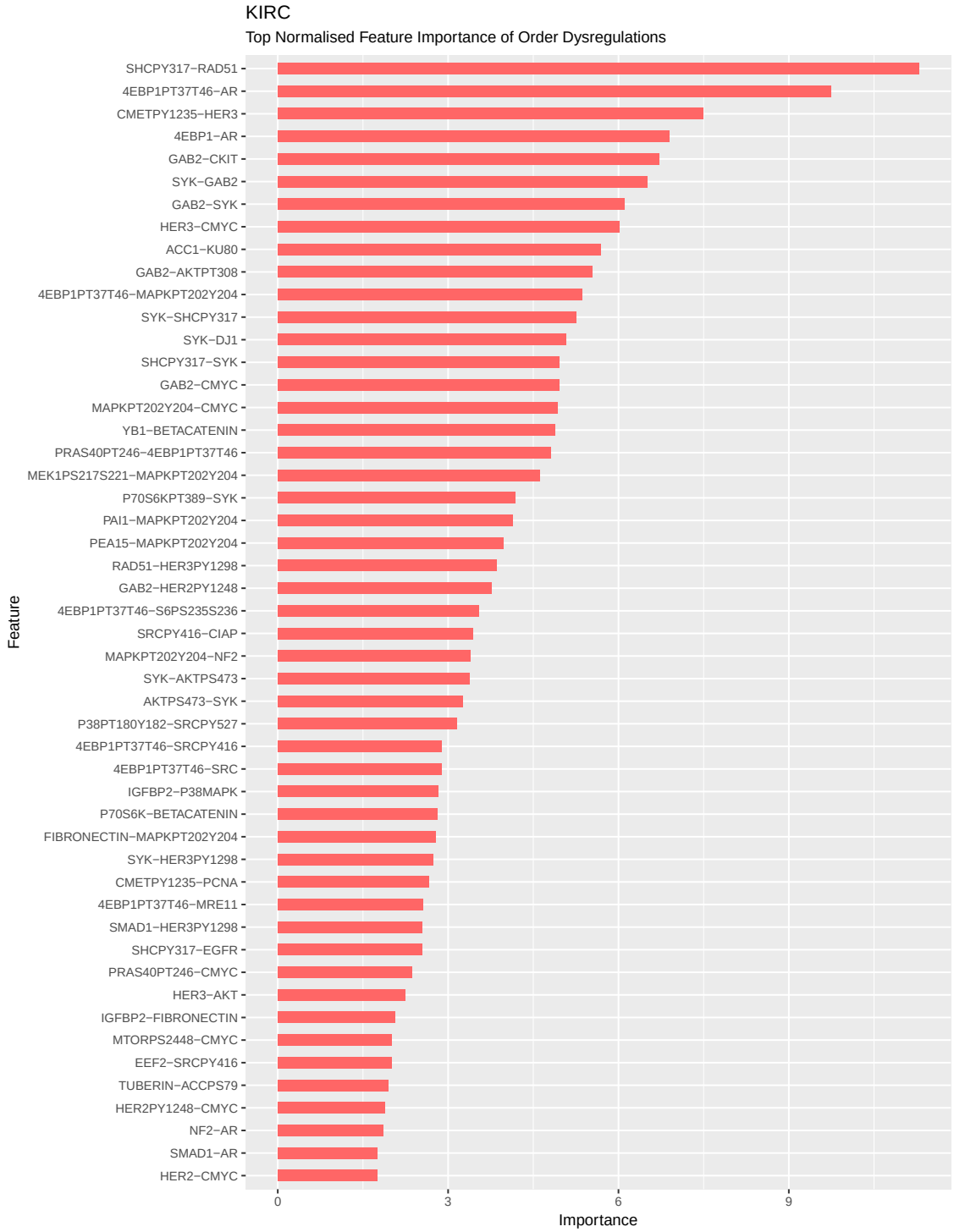


Figure B.1: Variable importance of significant pairwise ranking embeddings for kidney renal clear cell carcinoma

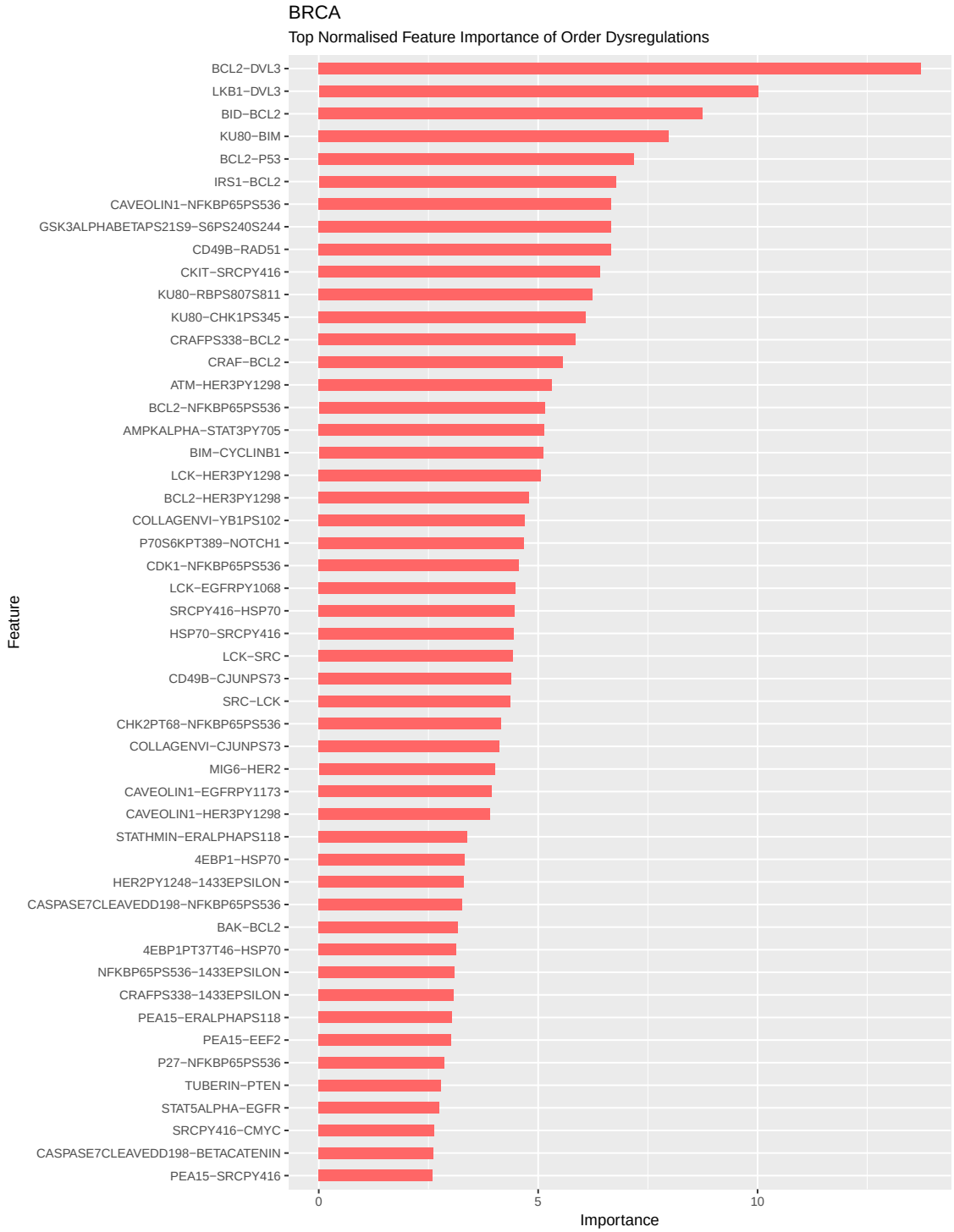


Figure B.2: Variable importance of significant pairwise ranking embeddings for breast invasive carcinoma

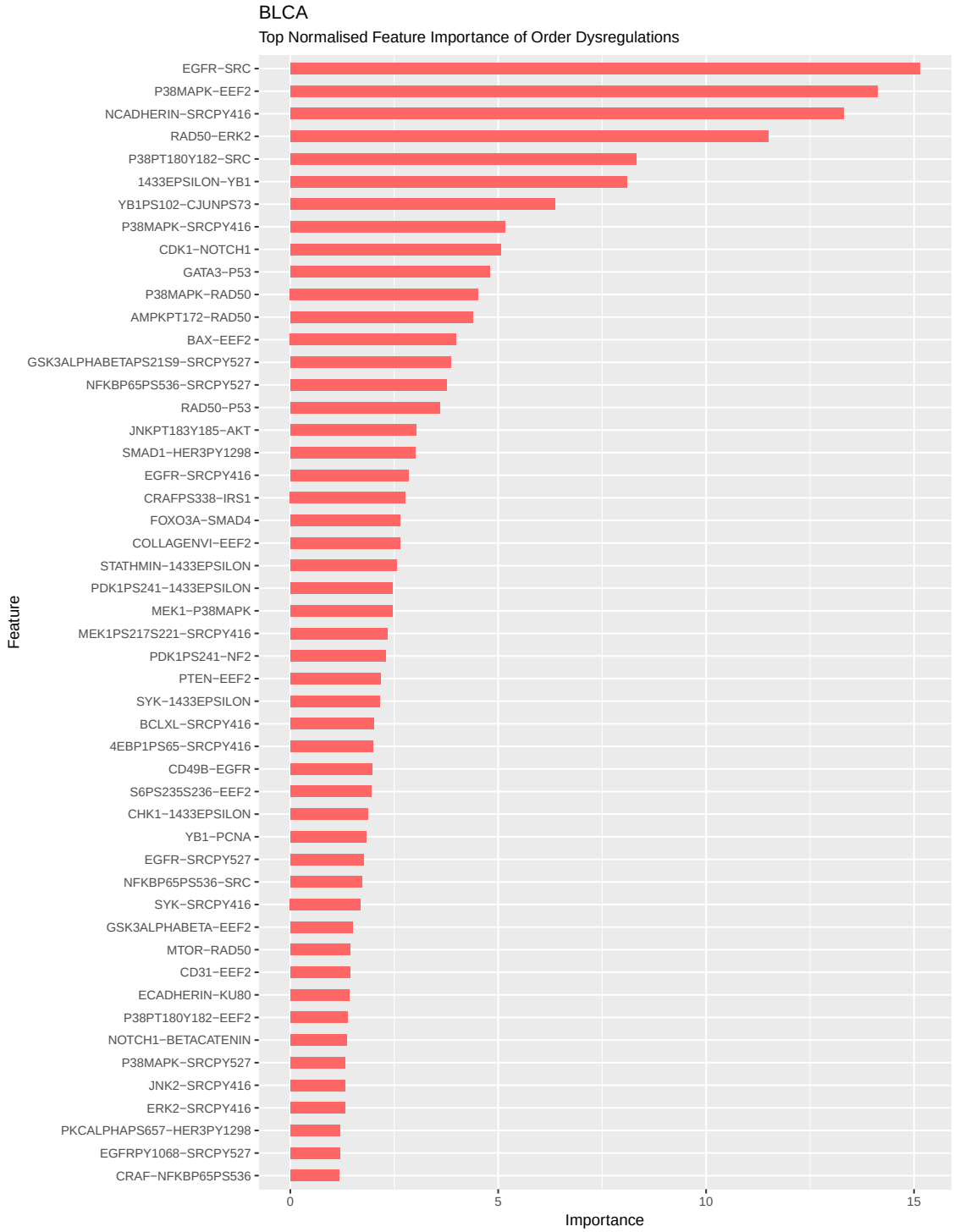


Figure B.3: Variable importance of significant pairwise ranking embeddings for bladder urothelial carcinoma

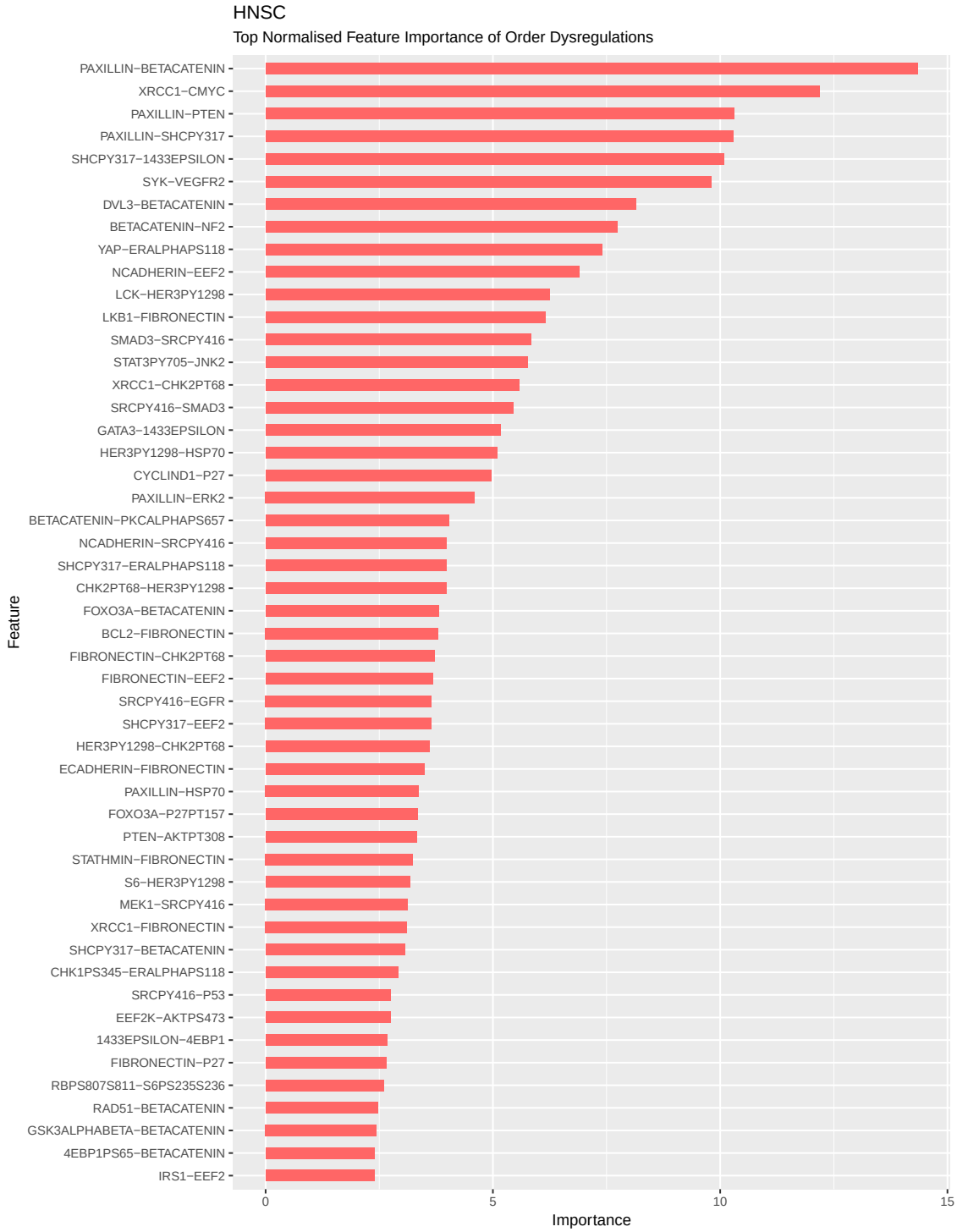


Figure B.4: Variable importance of significant pairwise ranking embeddings for head and neck squamous cell carcinoma

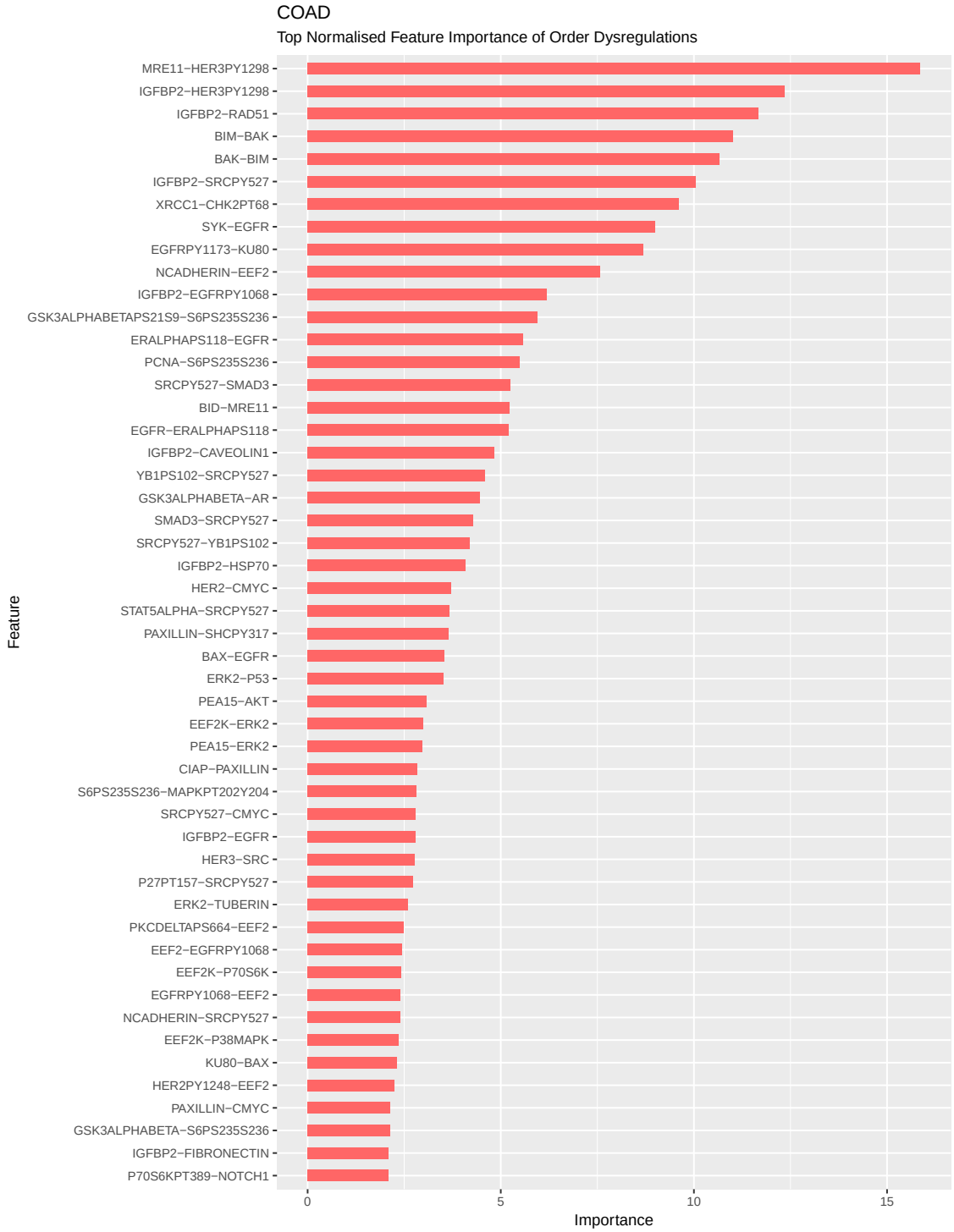


Figure B.5: Variable importance of significant pairwise ranking embeddings for colon adenocarcinoma

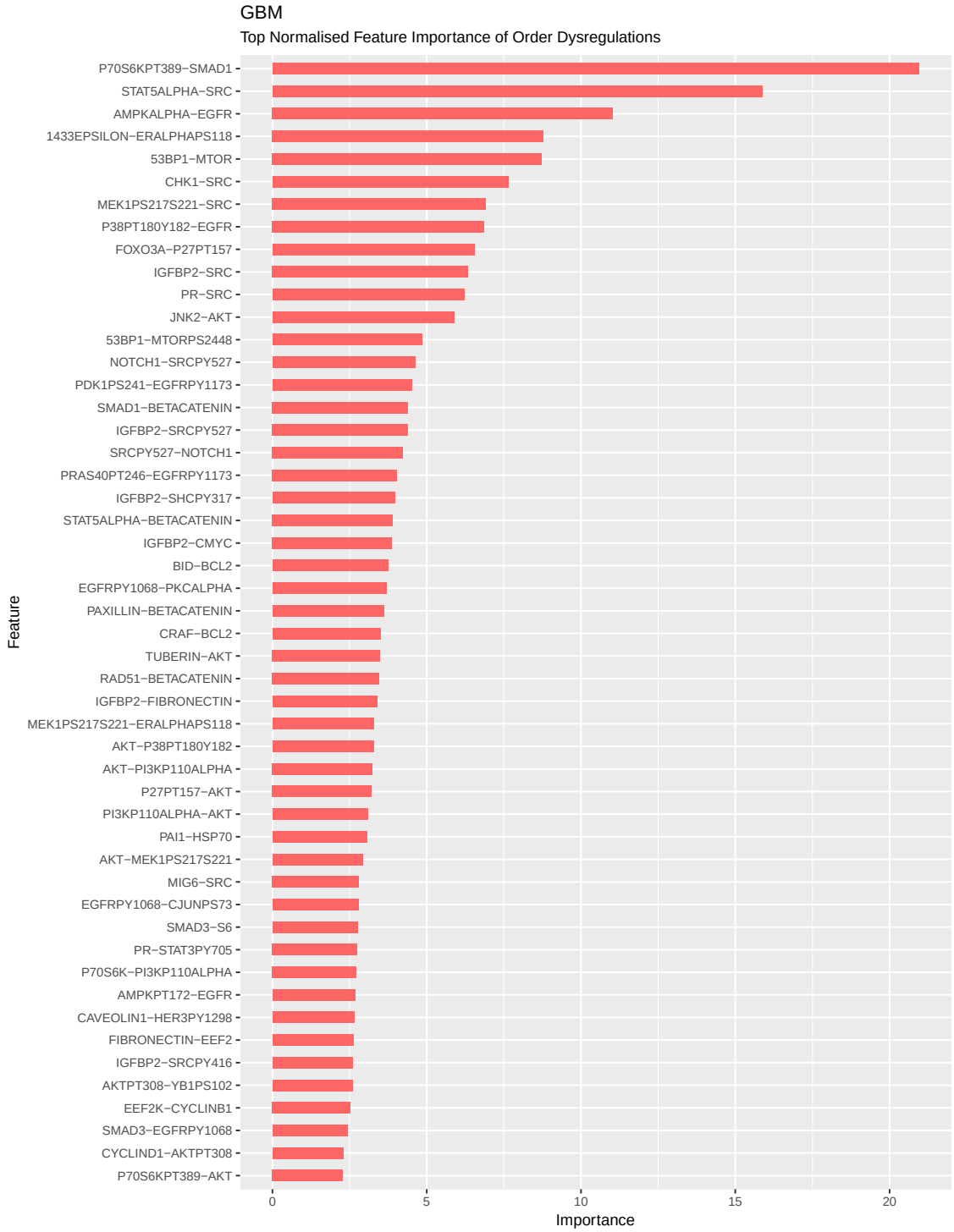


Figure B.6: Variable importance of significant pairwise ranking embeddings for glioblastoma multiforme

Figure B.7: Variable importance of significant pairwise ranking embeddings for lung adenocarcinoma

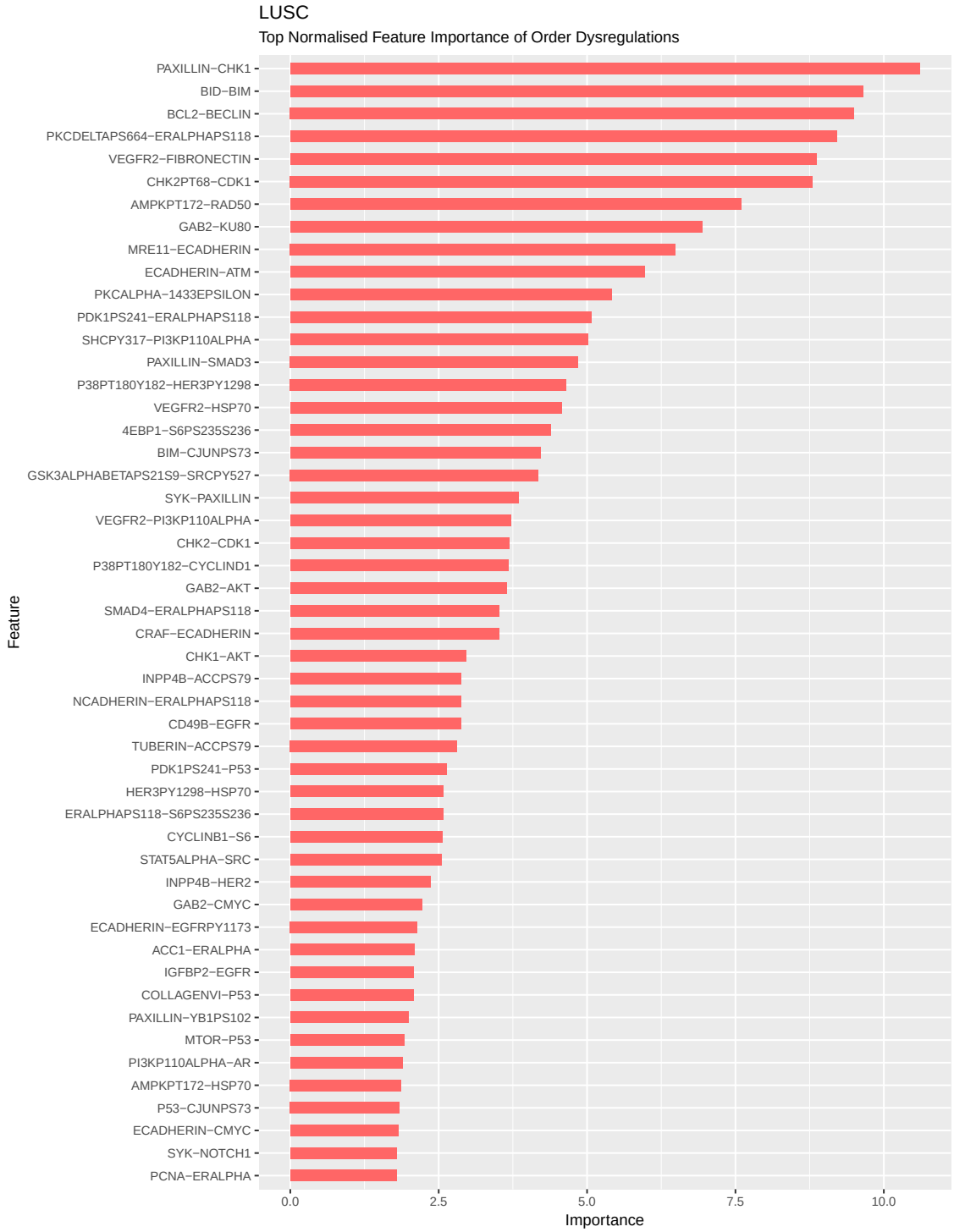


Figure B.8: Variable importance of significant pairwise ranking embeddings for lung squamous cell carcinoma

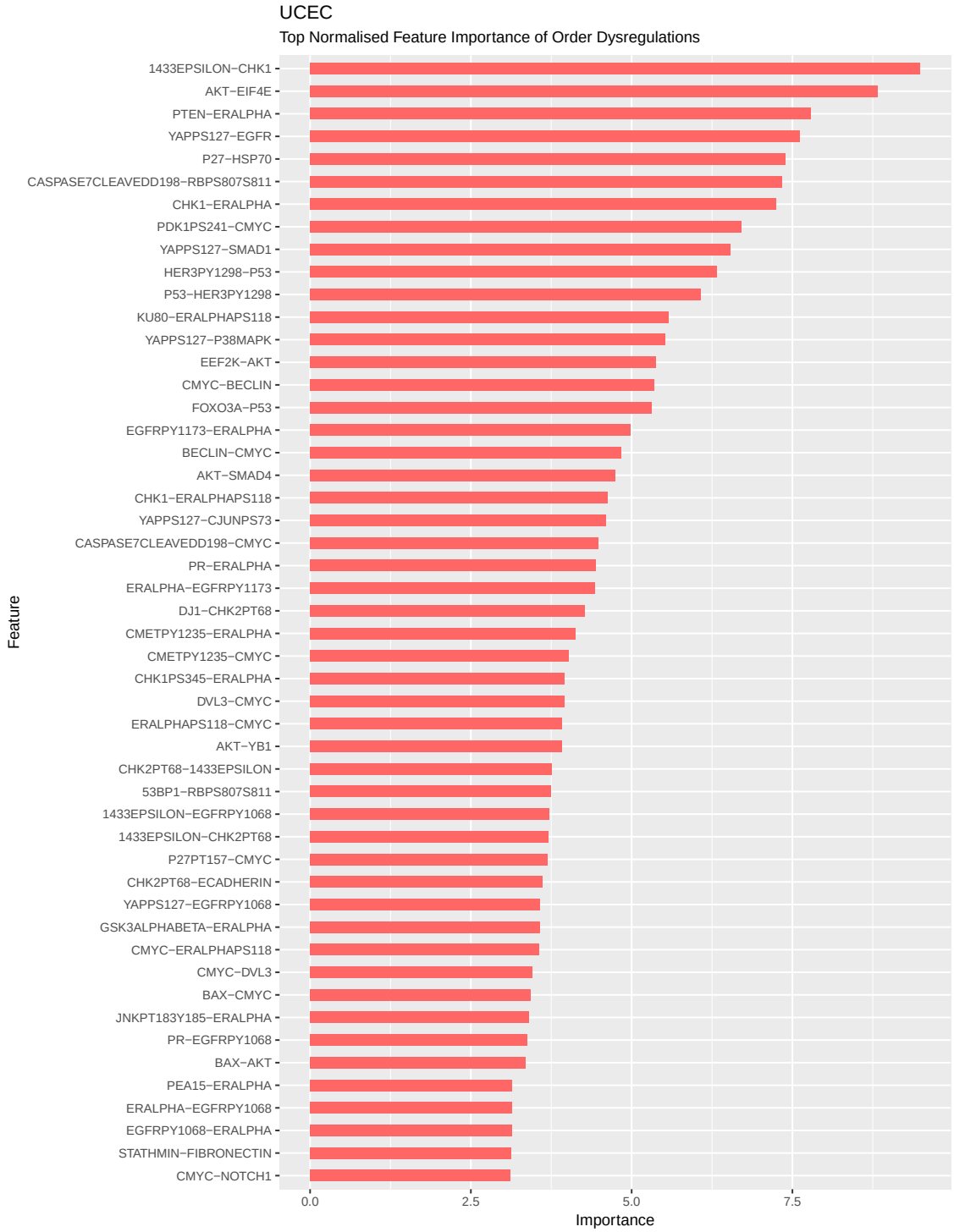


Figure B.9: Variable importance of significant pairwise ranking embeddings for uterine corpus endometrial carcinoma

Appendix C

PRER Networks

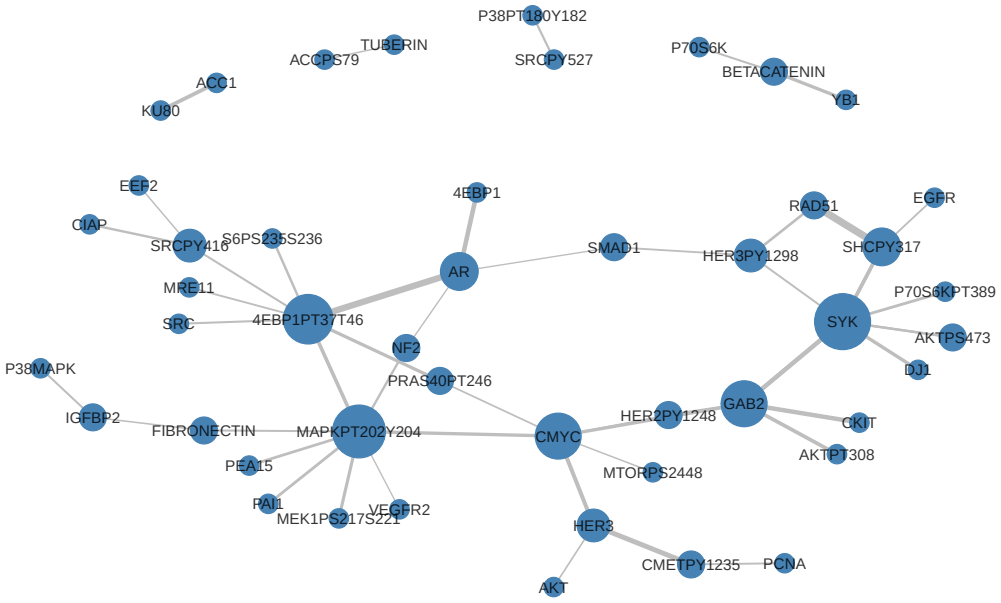


Figure C.1: Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for kidney renal clear cell carcinoma; edges represent that two proteins participate in a pairwise rank order feature together

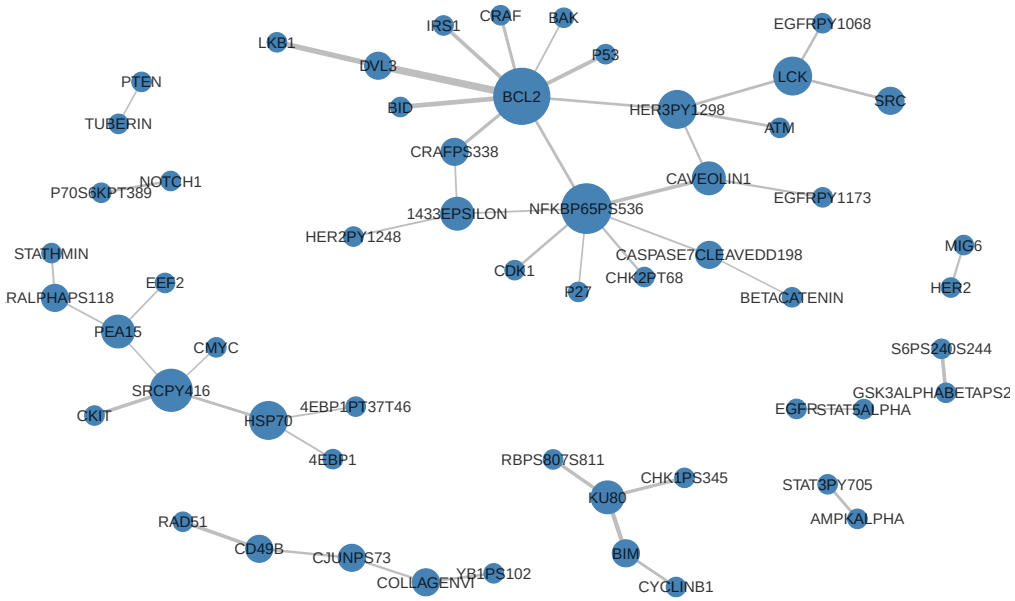


Figure C.2: Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for breast invasive carcinoma; edges represent that two proteins participate in a pairwise rank order feature together

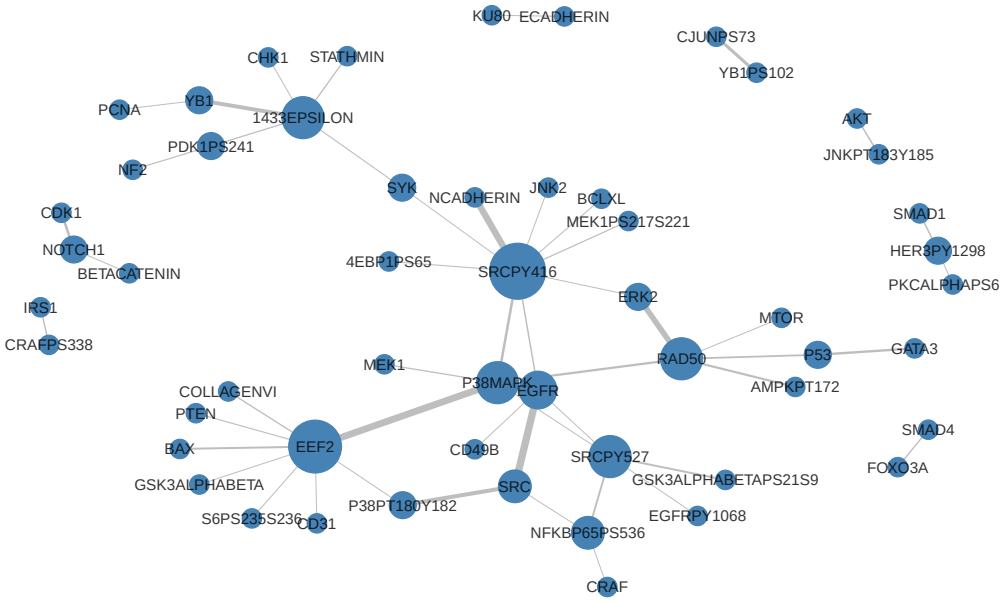


Figure C.3: Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for bladder urothelial carcinoma; edges represent that two proteins participate in a pairwise rank order feature together

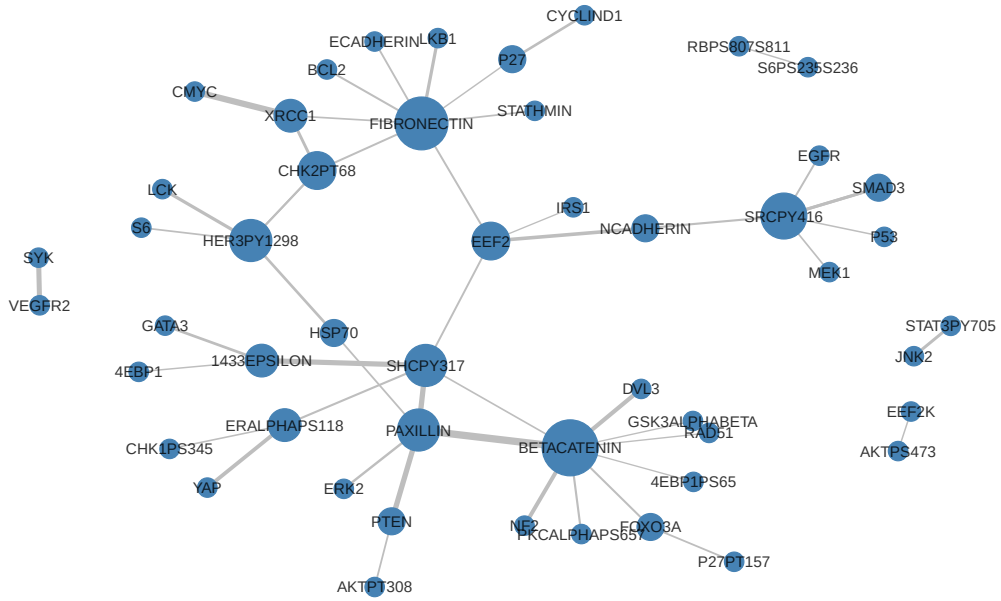


Figure C.4: Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for head and neck squamous cell carcinoma; edges represent that two proteins participate in a pairwise rank order feature together

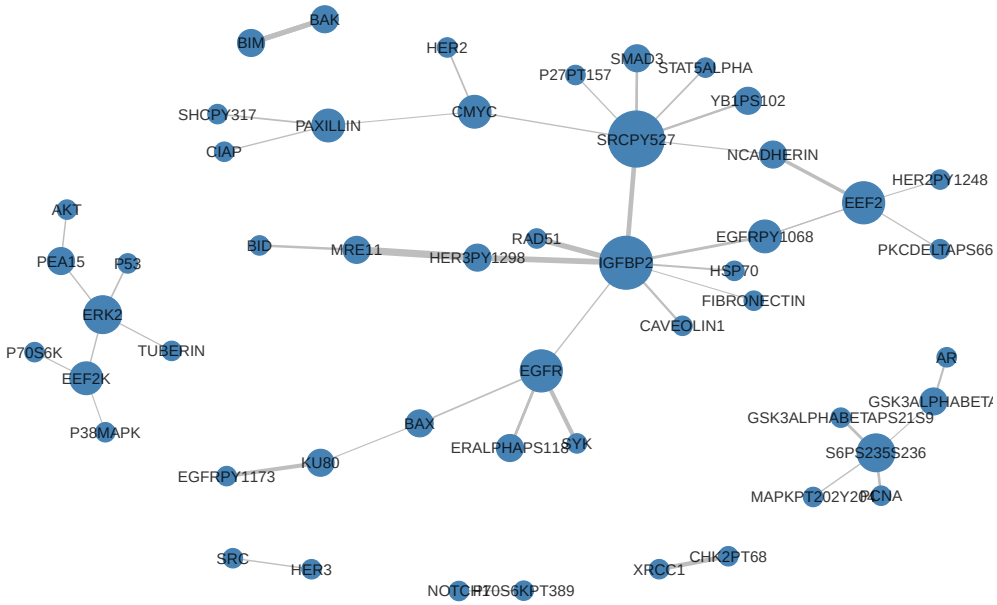


Figure C.5: Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for colon adenocarcinoma; edges represent that two proteins participate in a pairwise rank order feature together

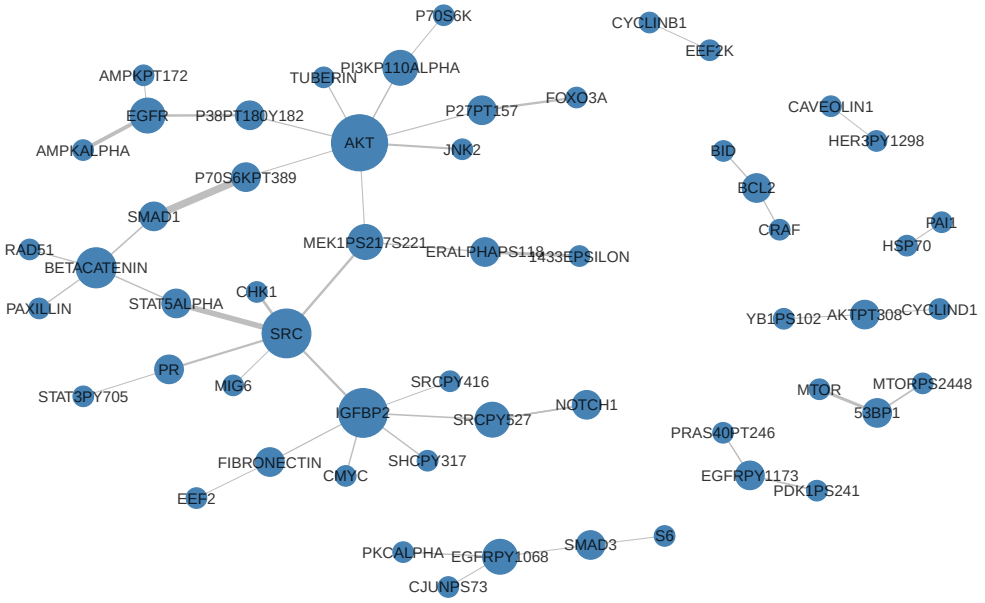


Figure C.6: Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for glioblastoma multiforme; edges represent that two proteins participate in a pairwise rank order feature together

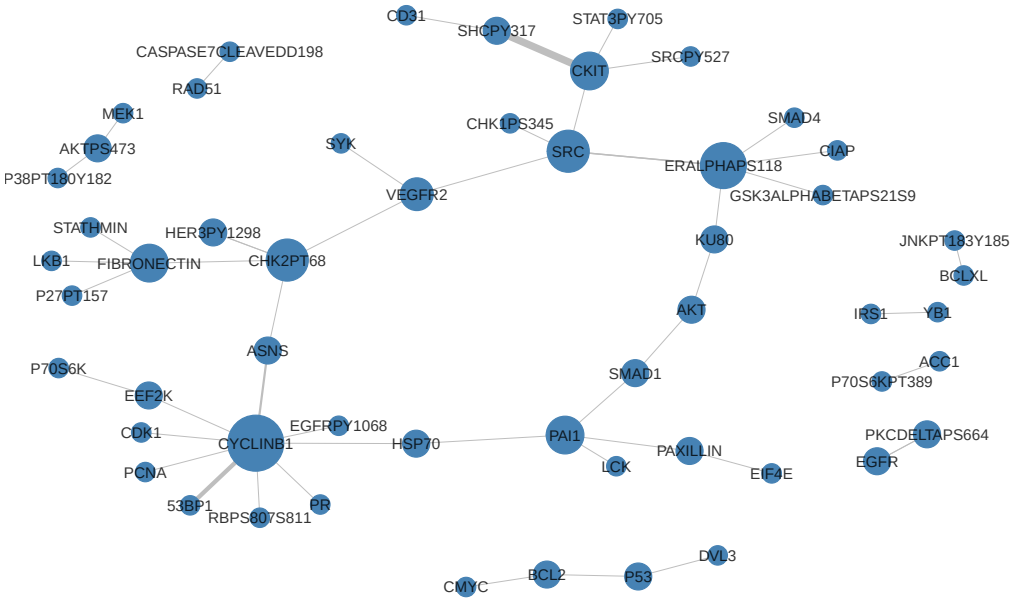


Figure C.7: Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for lung adenocarcinoma; edges represent that two proteins participate in a pairwise rank order feature together

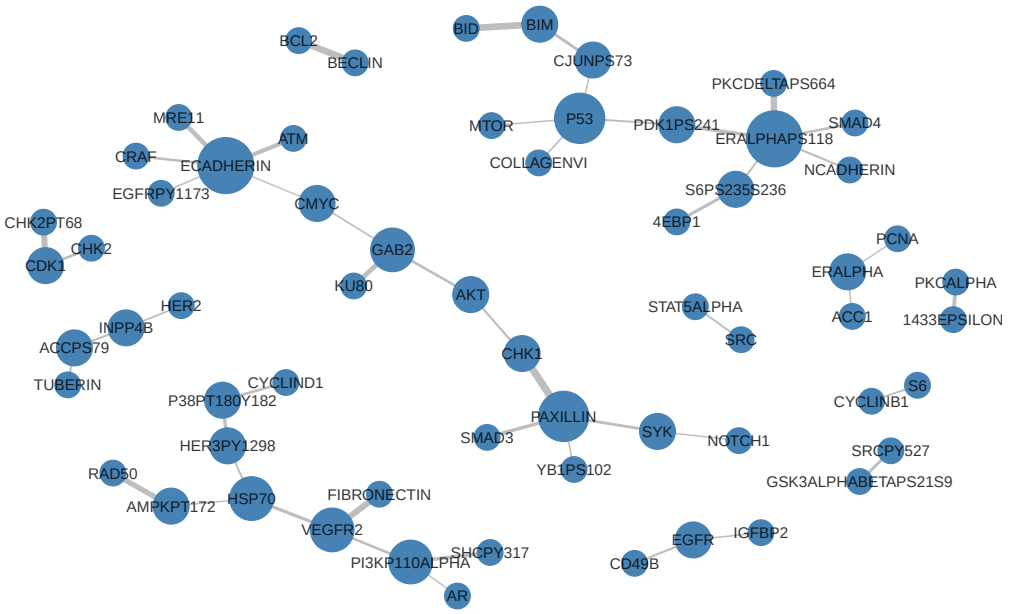


Figure C.8: Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for lung squamous cell carcinoma; edges represent that two proteins participate in a pairwise rank order feature together

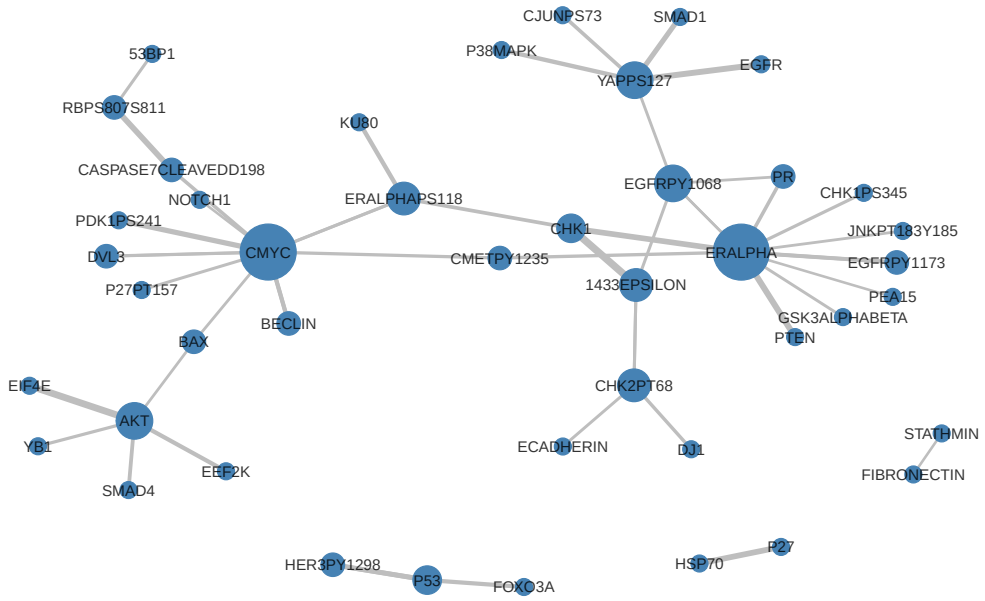


Figure C.9: Nodes represent proteins that appear in the top 50 pairwise ranking embeddings for uterine corpus endometrial carcinoma; edges represent that two proteins participate in a pairwise rank order feature together