Efficient Machine Learning Models for Cancer Biology

by

Ayyüce Begüm Bektaş

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Industrial Engineering and Operations Management

KOÇ ÜNİVERSİTESİ

August 5, 2022

Efficient Machine Learning Models for Cancer Biology

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Ayyüce Begüm Bektaş

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Mehmet Gönen (Advisor)

_____

Prof. Ceyda Oğuz

_____

Prof. Füsun Can

_____

Prof. Mehmet Güray Güler

_____

Assoc. Prof. Arzucan Özgür

Date: _____

To my family.

ABSTRACT

Efficient Machine Learning Models for Cancer Biology
Ayyüce Begüm Bektaş
Doctor of Philosophy in Industrial Engineering and Operations Management
August 5, 2022

In the recent past, a variety of multiple kernel learning algorithms has been proposed in machine learning literature. A kernel corresponds to a measure of similarity between the data instances while multiple kernels correspond to multiple different measures of similarity. Learning with multiple kernels, in brief, serves to perform learning while integrating different inputs originated from different feature representations. This thesis contains three main extensions to original multiple kernel learning framework together with their implementations on cancer data sets.

Identification of molecular mechanisms that determine tumor progression in cancer patients is a prerequisite for developing new disease treatment guidelines. Even though the predictive performance of current machine learning models is promising, extracting significant and meaningful knowledge from the data simultaneously during the learning process is a difficult task considering the high-dimensional and highly correlated nature of genomic data sets. Thus, there is a need for models that not only predict tumor volume from gene expression data of patients but also use prior information coming from pathways/gene sets during the learning process to distinguish molecular mechanisms that play crucial role in tumor progression and disease prognosis.

In this thesis, we demonstrate a novel machine learning algorithm, PrognosiT, that combines optimization and kernel learning. Instead of initially choosing several pathways/gene sets from a candidate set and training a model on this previously chosen subset of features, our proposed algorithm accomplishes both tasks together. We tested our algorithm on thyroid carcinoma patients using gene expression profiles and cancer-specific pathways/gene sets. Predictive performance of our novel multiple kernel learning algorithm was comparable or even better than random forest (RF)

and support vector regression (SVR). It is also notable that, to predict tumor volume, PrognosiT used gene expression features less than one-tenth of what RF and SVR algorithms used. We demonstrated that during the learning process, our algorithm managed to extract relevant and meaningful pathway/gene set information related to the studied cancer type, which provides insights about its progression and aggressiveness. We also compared gene expressions of the selected genes by our algorithm in tumor and normal tissues, and we then discussed up- and down-regulated genes selected by our algorithm, which could be beneficial for determining new biomarkers.

The thesis also provides a novel multiple approximate kernel learning framework, namely, MAKL, that is fast, scalable and interpretable. Data set sizes in computational biology have been increased drastically with the help of improved data collection tools and increasing size of patient cohorts. Previous kernel-based machine learning algorithms proposed for increased interpretability started to fail with large sample sizes, owing to their lack of scalability. To overcome this problem, we proposed MAKL, a fast and efficient multiple kernel learning algorithm to be particularly used with large-scale data that integrates kernel approximation and group Lasso formulations into a conjoint model. Our method extracts significant and meaningful information from the genomic data while conjointly learning a model for out-of-sample prediction. It is scalable with increasing sample size by approximating instead of calculating distinct kernel matrices. To test MAKL, we demonstrated our experiments on three cancer data sets (i.e., created using multiple cancer cohort data sets from The Cancer Genome Atlas (TCGA) consortium and a melanoma single-cell data set) and showed that MAKL is capable to outperform the baseline algorithm, extreme gradient boosting, while using only a small fraction of the input features. We also reported selection frequencies of low-dimensional approximation matrices associated with feature subsets (i.e., pathways/gene sets), which helps seeing their relevance for the given classification task. Our fast and interpretable MKL algorithm producing sparse solutions is promising for computational biology applications considering its scalability and highly correlated structure of genomic data sets, and it can be used to discover new biomarkers and new therapeutic guidelines.

As another contribution, this thesis provides a novel multiple approximate kernel clustering framework. Kernel-based clustering algorithms are essential to genomic data analysis since they provide detection of nonlinear relationships within the

data while offering interpretability using one or more prior information sources. With this motivation, we designed a scalable multiple approximate kernel k-means clustering framework that is compatible with large-scale data sets and combines kernel approximation and k-means clustering approach into the same model. To test our algorithm, we combined information from multiple cancer cohorts provided by TCGA consortium. Our algorithm extracts relevant parts of the prior information to the clustering task while maximizing the silhouette score to improve the clustering results. To test the findings of our clustering framework, we performed k-means clustering on the full data set as a baseline method. The silhouette score resulted from the baseline experiment was 35.8% lower than the one resulted from our algorithm which uses information from only four gene sets instead of all 19 814 genes. Results of our proposed algorithm are supported by the existing literature and our approach is promising to be an easy and efficient way to provide sparse and interpretable results while integrating prior information as approximation matrices to k-means algorithm in a novel way, which may give hope to discover new cancer subtypes and insights related to novel cancer treatment options.

ÖZETÇE

Kanser Biyolojisi için Etkin Yapay Öğrenme Modelleri
Ayyüce Begüm Bektaş
Endüstri Mühendisliği ve İşletme Yönetimi, Doktora
5 Ağustos 2022

Yakın geçmişte yapay öğrenme için birçok çoklu çekirdek öğrenimi yöntemi önerilmiştir. Çekirdek genel anlamda bir benzerlik ölçütüdür; çoklu çekirdekler de bir veri kümesi için farklı kaynaklardan sağlanan ayrı benzerlik ölçütleri olarak düşünülebilir. Çoklu çekirdek öğrenimi ise özetle farklı kaynaklardan gelen çoklu benzerlik ölçütlerini öğrenme modeline dahil etmektedir. Bu tez, çoklu çekirdek öğrenimi için üç yeni yöntem önermekte ve bu yöntemlerin kanser veri kümeleri üzerinde gerçekleştirilmiş deneylerini raporlamaktadır.

Kanser çağımızın en yaygın ve ölümcül hastalıklarından biridir. Kanser hastalarında, yeni hastalık tedavi kılavuzlarının belirlenmesi için hastalığın ilerleyişinin altında yatan moleküler mekanizmaların tanımlanması ön şarttır. Mevcut yapay öğrenme modellerinin performans sonuçları ümit vadediyor olsa da, eş zamanlı olarak verinin anlamlı ve kayda değer kısımlarının tespiti ile öğrenme işinin yapılması, çözümü zor bir problemdir. Bu zorluk temel olarak genomik verilerin çok boyutlu, yüksek korelasyonlu ve doğrusal olmayan ilişkileri içeren yapısıyla ilişkilidir. O halde, kanser ilerlemesini tahmin eden bir model kurulurken bu modelin bir öncül bilgi kümesiyle (örneğin, yolak/gen kümesi bilgileri) birlikte çalışabilmesi, bu öncül bilgi kümesinin kanser ilerleyişi için önemli olan kısımlarının öğrenirken bulunmasına ihtiyaç vardır. Sonuç olarak altta yatan moleküler dinamiklerin anlaşılmasıyla, kanserin ilerlemesinin ve agresifliğinin daha iyi anlaşılması beklenmektedir.

Bu tezde ilk olarak eniyileme ve çoklu çekirdek öğrenimini birlikte kullanan yeni bir yapay öğrenme modeli önermekteyiz. Öncül bilgi olarak kanser yolak/gen kümesi bilgisi kullanıldığında, bu öncül bilginin önce önemli kısımlarının seçilip sonra öğrenme işinin yapılması yerine, eş zamanlı olarak öncül bilgiden önemli kısımları çıkararak öğrenme işini gerçekleştiren bir algoritma geliştirdik. Bu algoritmayı Kanser Genom Atlası tiroit kanseri hasta verisi üzerinde kanser için özel olarak

hazırlanmış yolak/gen kümesi öncül bilgilerini kullanarak test ettik. Yeni çoklu çekirdek öğrenimi algoritmamızın tahmin performansının temel referans algoritması olarak seçtiğimiz rassal orman (RO) ve regresyon destek vektör makinesi (RDVM) algoritmalarının tahmin performanslarıyla karşılaştırabilir ve hatta onların tahmin performanslarından daha iyi olduğunu gösterdik. Ayrıca, önerdiğimiz algoritmanın bu performansı RO ve RDVM algoritmalarının kullandıklarının onda birinden daha az öznitelik kullanarak gösterdiğini; çalışılan kanser türüne ilişkin önemli yolak/gen kümesi bilgilerini belirlediğini gösterdik. Öte yandan, algoritma tarafından seçilen genlerin ekspresyonlarını tümör ve normal dokularda karşılaştırdık; tümör dokusunda normale göre daha fazla veya daha az ifade edilen genlerin yeni biyolojik belirteç bulunması açısından nasıl faydalı olabileceği konusunu tartıştık.

Bu tezde ayrıca yeni bir hızlı, ölçeklenebilir ve yorumlanabilir çoklu yaklaşıklanmış çekirdek öğrenimi (ÇYÇÖ) yöntemi sunduk. Hesaplamalı biyoloji alanında gün geçtikçe gelişen veri toplama araçları ve büyüyen hasta kohortları ile veri boyutları hızla artmaktadır. Sağladığı yorumlanabilme ve verideki doğrusal olmayan ilişkileri yakalayabilme özellikleri sayesinde biyolojik veri analizinde özellikle tercih edilen çekirdek temelli öğrenme algoritmaları ölçeklenebilir olmadıklarından büyüyen veri kümesi boyutlarıyla birlikte çalışmamaya başlamışlardır. Bu problemi çözmek için hızlı ve etkin, büyük ölçekli veri ile kullanıma uygun, çekirdek yaklaşıklama ve grup Lasso tekniklerini birlikte yeni bir şekilde kullanan bir algoritma geliştirdik. Bu yöntem etkin bir şekilde sınıflandırma yapmayı öğrenirken genomik verinin anlamlı ve kayda değer kısımlarını belirlemektedir. Çekirdek matrislerini hesaplamak yerine yaklaşıkladığı için artan veri kümesi boyutlarıyla kullanıma uygun ve ölçeklenebilirdir. ÇYÇÖ yöntemini Kanser Genom Atlası'ndan çoklu kanser hasta kohort verilerini kullanarak oluşturduğumuz iki veri kümesi üzerinde ve melanoma tek hücre dizileme verisi üzerinde test ettik. ÇYÇÖ yönteminin temel referans algoritmanın performansından daha üstün bir performansa girdi özniteliklerin yalnızca çok küçük bir kısmını kullanarak ulaştığını gösterdik. Ayrıca, yolak/gen kümelerinin ilişkili olduğu yaklaşıklanmış çekirdeklerin seçilme sıklığını çalışılan sınıflandırma problemi için bu yolak/gen kümelerinin önemini işaret etmek amacıyla raporladık. Bu hızlı ve yorumlanabilir çoklu çekirdek öğrenimi yöntemi ve yolak/gen kümesi temelinde verdiği seyrek ve anlamlı bilgiler ile yüksek korelasyonlu genomik veri kümelerinin analizinin kolaylaşması sağlanmıştır. Bu yöntemin, daha önceden analizi yapılamayan büyük veri kümeleriyle çalışmayı mümkün kılarak yeni tedavi kılavuzlarının hazırlanması

ve yeni biyolojik belirteçlerin bulunması konusunda yardımcı olması beklenmektedir.

Literatüre bir diğer katkı olarak bu tez, yeni bir çoklu yaklaşıklanmış öbekleme yöntemi sağlamaktadır. Çekirdek temelli öbekleme yöntemleri genomik veri analizi için öncül bilgi ile bütünleşerek yorumlanabilirlik sunmaları ve verideki doğrusal olmayan ilişkileri bulmaya yardımcı olmaları açısından önemli gözetimsiz öğrenme yöntemleridir. Bu amaçla, ölçeklenebilir bir çoklu yaklaşıklanmış çekirdek k-ortalama öbekleme yöntemi geliştirdik. Bu yöntem çekirdek yaklaşıklama yapıp k-ortalama öbekleme algoritmasıyla bu yaklaşıklamayı birleştirmektedir. Algoritmamızı test etmek için Kanser Genom Atlası veri kümelerinden çoklu kohort genomik verisini birleştirdik. Algoritma, öbekleme sonuçlarını iyileştirmek için "silhouette" katsayısını eniyilerken öncül bilginin önemli kısımlarını bulmak üzerine tasarlanmıştır. Öbekleme için oluşturduğumuz bu tekniğin sonuçlarını test etmek üzere, tüm veri kümesine k-ortalama öbekleme yöntemini uyguladık. 19 814 geni kullanan referans deneyin sonucu yalnızca dört gen kümesinden gelen bilgiyi kullanarak öbekleme yapan önerdiğimiz öbekleme algoritmasından %35,8 daha az çıkmıştır. Önerilen öbekleme algoritmasının sonuçları literatür verileri ile desteklenmektedir. Kolay kullanılabilirliği ve ölçeklenebilir olması ile bu çalışma yeni kanser alt tiplerinin ve yeni tedavi yöntemlerinin keşfedilmesi açısından umut vericidir.

ACKNOWLEDGMENTS

I would first like to express my sincere gratitude to my advisor, Assoc. Prof. Mehmet Gönen for his scientific guidance and motivation throughout my PhD. He encouraged me to be persistent against the difficulties I encountered during my research. I will forever remember his support towards better research.

I am also very grateful to my examiners Prof. Ceyda Oğuz, Prof. Füsun Can, Prof. Mehmet Güray Güler and Assoc. Prof. Arzucan Özgür for their constructive comments and insights.

I wholeheartedly thank the members of Koç University - Department of Industrial Engineering where I have enhanced my research and teaching skills throughout my doctoral training.

My thanks extend to my friends I have met during my PhD and to the former and current members of the Machine Intelligence and Data Analysis in Science Laboratory (MIDASLAB).

My special thanks go to my lifelong friends who have always encouraged me and were there for me. Life would not be this beautiful and enjoyable without them.

I would also like to acknowledge the funding provided by the Scientific and Technological Research Council of Turkey (TÜBİTAK) and Koç University during my PhD. This thesis has been supported by TÜBİTAK under grant EEEAG 117E181.

Finally, I would like to thank my dearest family that I am deeply proud and feel very lucky to have. They have supported and encouraged me through the hardest times, accompanied me during the happiest moments of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

$\|\cdot\|_p$	ℓ_p-norm
$\top$	Transpose
$\langle\cdot,\cdot\rangle$	Dot product
C	Regularization parameter for support vector machines
$k(\cdot,\cdot)$	Kernel function
K	Number of clusters
$\mathbf{K}$	Kernel matrix
$\ell(\cdot,\cdot)$	Loss function
N	Number of training data points
$\mathbb{N}$	Natural numbers
P	Number of kernels
$\mathbb{R}$	Real numbers
$\mathbb{R}_{++}$	Positive real numbers
$\boldsymbol{x}$	Data point
y	Output value
$\mathbf{Z}$	Low-dimensional approximation matrix
$\boldsymbol{\alpha}_i^+, \boldsymbol{\alpha}_i^-$	Support vector coefficients
$\boldsymbol{\beta}$	Weights in group Lasso formula
ϵ	Tube width
$\boldsymbol{\eta}$	Kernel weights
λ	Regularization parameter for group Lasso formula
$\boldsymbol{\xi}^+$, $\boldsymbol{\xi}^+$	Sets of slack variables
$\Phi(\cdot)$	Mapping function

ABBREVIATIONS

AUROC	Area Under the Receiver Operating Characteristic Curve
GDC	Genomic Data Commons
Lasso	Least Angle Selection and Shrinkage Operator
MKL	Multiple Kernel Learning
MAKL	Multiple Approximate Kernel Learning
mRNA	Messenger RNA
MSigDB	Molecular Signatures Database
NRMSE	Normalized Root Mean Squared Error
PID	Pathway Interaction Database
RF	Random Forest
SVM	Support Vector Machine
SVR	Support Vector Regression
TCGA	The Cancer Genome Atlas
XGBoost	eXtreme Gradient Boosting

Chapter 1

INTRODUCTION

Machine learning serves to capture the patterns in the observed data to accurately classify the unseen data instances (i.e., for classification problems) or to make accurate predictions (i.e., for regression problems). While doing so, algorithms do not need to be explicitly programmed to capture these patterns, which refers to their capability "to learn". In supervised learning setting, we try to learn from the observed data instances with their corresponding outputs, $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where N denotes the size of the observed data set, $\boldsymbol{x}_i$ denotes the feature vector associated with data instance i, and y_i is the output associated with data instance i. Output $\{y_i\}_{i=1}^N$ correspond to class labels for classification tasks and to continuous values for prediction tasks. As to the unsupervised learning setting, we are given only the observed data instances without their corresponding output, $\{\boldsymbol{x}_i\}_{i=1}^N$. In this thesis, we have developed novel algorithms in both supervised and unsupervised learning settings and applied these algorithms to cancer data sets.

Cancer is one of the most common causes of mortality in our era, and its treatment may extremely be hard to patients, both from a psychological and economical perspective. Many genetic, epigenetic and environmental factors are effective in cancer pathogenesis and, for each type of cancer, these factors play different roles. Therefore, determining the molecular mechanisms that are related to driver genes and driver pathways is of great importance in terms of cancer diagnostics, prognostics and treatment. In recent years, projects about large-scale cancer genomics are giving researchers a chance to understand the genomic and epigenomic changes in patients. Thus, associated with the increasing opportunity of analyzing genomic characterizations of tumors biopsied from patients, standard machine learning algorithms like

random forest (RF) [Breiman, 2001] and support vector machines (SVM) [Cortes & Vapnik, 1995] have been used to make predictions related to cancer. Even though the predictive performance of these machine learning algorithms is usually good, these applications may not be successful in extracting significant and meaningful knowledge from the data since the genomic data sets are high-dimensional and highly correlated by their nature. For this reason, designing new machine learning algorithms that are capable of selecting meaningful parts of the genomic data sets and use these selected subsets for prediction is necessary.

In this thesis, we first consider the problem of predicting tumor volume using a novel multiple kernel learning algorithm. Tumor volume is considered to be one of the significant prognostic factors for oncological outcome after radiotherapy or chemotherapy [Timmermans et al., 2016]. Along with clinical T and N stages, tumor differentiation and circumferential tumor extent, tumor volume has been identified by numerous retrospective cohort studies as potential predictors of pathologic complete response [Tan et al., 2019]. Rather than TNM staging system, estimating tumor volume from the patient's genomic data while conjointly identifying the molecular mechanisms that affect tumor progression could be highly useful to foresee cancer aggressiveness. Although TNM staging system has been demonstrated to have prognostic information, different cure rates in the literature have raised concern about the efficiency of the T-classification [Issa et al., 2015].

Even though a reduction in tumor volume after therapy appears to be indicating a better prognosis than an unchanged or increasing tumor size, this assumption may not be correct in some cases. Tumor size is strongly related to cancer prognosis but dynamics of this relation have not yet been fully understood. Since the underlying biological mechanisms that affect tumor size have not yet been discovered completely, there arises a need to determine new ways to predict cancer prognosis using tumor volume. In other words, there is a need for models that learn to predict tumor volume while determining the important pathways/gene sets that affect the tumor progression and agressiveness at the genomic level [Weber, 2009].

Klement et al. (2014) studied SVM-based prediction of local tumor control, but

they trained their model on only seven potential input features. While there exists standard statistical tests and models applied to the clinical outcomes of patients; to the best of our knowledge, there is no study of predicting tumor volume using machine learning while simultaneously discovering the hidden molecular mechanisms towards tumor progression using genomic characterizations of the patients as input.

Integration of kernel matrices into a linear model to solve a nonlinear problem called the "kernel trick". A kernel can be described as a measure of similarity between data points, which corresponds to a dot product in some implicit feature space [Schölkopf & Smola, 2002]. The kernel trick basically enables us to map the input data to a higher dimensional feature space and then to solve the originally nonlinear problem in this high-dimensional and implicit feature space. Some of the widely used kernel functions used in the literature are the linear kernel ($k_{\mathcal{L}}$), the polynomial kernel ($k_{\mathcal{P}}$) and the Gaussian kernel ($k_{\mathcal{G}}$):

$$
\begin{aligned}
&k_{\mathcal{L}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle, \\
&k_{\mathcal{P}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + 1)^q \qquad q \in \mathbb{N}, \\
&k_{\mathcal{G}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{x}_i - \boldsymbol{x}_j \rangle}{2\sigma^2} \right) \qquad \sigma \in \mathbb{R}_{++}.
\end{aligned}
$$

Instead of learning with a single kernel matrix, one may choose to learn with multiple kernel matrices calculated using different kernel functions on the same data representation, using the same kernel function on different data partitions or using same/different kernel functions on different data representations, which is known as multiple kernel learning (MKL). Some of the existing MKL algorithms assign equal weights to the input kernels, which may not be the best way of discovering the underlying dynamics. For instance, in computational biology, to discover the underlying molecular mechanisms from genomic data, instead of initially setting the kernel weights, making the MKL algorithm choose the weights assigned to each kernel in a data-dependent manner is more convenient. Additionally, combining kernels in a nonlinear or data-dependent way is stated to seem more promising than linear combination in fusing information provided by simple linear kernels, whereas linear methods are stated as being more reasonable in case of combining complex

Gaussian kernels [Gönen & Alpaydın, 2011]. Amongst machine learning algorithms, kernel-based approaches have been shown to be successful in problems associated with cancer, such as gene essentiality prediction [Costello et al., 2014, Gönen et al., 2017], due to their capability of handling high-dimensional genomic input data. One of the most widely known kernel-based machine learning methods is the support vector machine (SVM) algorithm [Boser et al., 1992, Guyon et al., 1992, Cortes & Vapnik, 1995]. Using multiple kernel learning framework on multi-omics data, Li et al. (2020) built a linear mixed model with adaptive Lasso for phenotype prediction which is capable of selecting predictive regions and predictive layers of the data. As another recent work, Uzunangelov et al. (2021) designed a multiple kernel learning framework, a kernel-based stacked learner where kernels are integrated with random forests where each one is built from a specific pathway/gene set.

In Chapter 2 of this thesis, we give details of a novel multiple kernel learning (MKL) algorithm, namely, PrognosiT[1], that we developed to predict the tumor volume using genomic characterizations while integrating a prior information source to provide interpretability [Bektaş & Gönen, 2021]. In this algorithm, we have applied MKL using support vector regression (SVR) as base learner on gene sets to discover mechanisms at the molecular level related to tumor initiation and progression. In Chapter 2, we first give details of SVR, then we demonstrate integration of kernel matrices and update rules for our iterative approach. Then in Chapter 5, we give the experimental results of our novel algorithm PrognosiT, its ability on the task of predicting tumor volume from genomic data. We also provide the relevant pathway/gene set outputs from our algorithm with the existing literature about the studied cancer type (i.e., thyroid carcinoma). Lastly, we compare the tumor and normal tissue gene expressions for the genes resulted from our algorithm, which could be beneficial for determining genomic biomarkers for screening.

For genomic data sets that have relatively low number of training instances, the number of model parameters to be optimized using kernel-based approaches

[1]Our implementations of support vector regression (SVR) and PrognosiT for tumor volume prediction in R can be found at `https://github.com/begumbektas/prognosit`.

is proportional to the number of training instances (N, generally in the order of hundreds); not to the number of training features (D, generally in the order of thousands) [Schölkopf & Smola, 2002], which is a big computational advantage for genomic data analysis of small patient cohorts. Besides the advantages of integrating exact kernel matrices to the learning process, machine learning methods that need to calculate a kernel matrix over the data set scale poorly with increasing training sample size, since the size of the kernel matrix is quadratic in the number of training samples. Although integration of multiple kernels provides detection of nonlinear relationships in the data and interpretability, calculating exact kernel matrices is computationally highly expensive and solving multiple kernel learning problems while iteratively calculating kernel matrices may be computationally prohibitive in case of large-scale data. Finding large-scale solutions to apply to genomic data is essential since with the help of improved data collection tools and increased size of patient cohorts, the size of data sets increases (e.g., single-cell research).

To overcome this problem, we propose in Chapter 3 of this thesis, a fast and efficient MKL algorithm to be particularly used with large-scale data that integrates kernel approximation and group Lasso formulations into a conjoint model [Bektaş et al., 2022]. Our method extracts significant and meaningful information from the tested genomic data while conjointly learning a model for out-of-sample prediction. It is scalable with increasing sample size by approximating instead of calculating distinct kernel matrices[2]. Associated with our method, we built an R package named `MAKL` that is available at `https://cran.r-project.org/package=MAKL`.

This thesis also provides an application of the main idea of MAKL algorithm to unsupervised learning setting. In Chapter 4, we provide a multiple approximate kernel-based clustering framework that is scalable and interpretable.

Clustering algorithms are unsupervised machine learning methods that aim to detect inner grouping of data instances for exploratory data analysis. Clustering, in broad terms, is the classical problem of dividing a d-dimensional finite set of

[2]Our implementations of MAKL algorithm and the baseline XGBoost algorithm in R are available at `https://github.com/begumbektas/makl`.

N data instances $\mathbf{X} \in \mathbb{R}^{N \times d}$ into k disjoint groups. These disjoint groups are intuitively aimed to be optimized such that while the within-cluster dissimilarity is minimized and the between-cluster dissimilarity is maximized to achieve good and easy-to-differentiate clusters. A clustering problem may be formulated to minimize the pairwise dissimilarity within a cluster. One of the most commonly used clustering algorithms for this aim is k-means clustering.

The most prevalent heuristic algorithm for k-means is Lloyd's algorithm (1957, published in 1982) followed by other algorithms [Lloyd, 1982, Forgy, 1965, MacQueen, 1967]. Lloyd's algorithm starts with random initialization of centroids and then proceeds by repeatedly assigning each data point to the cluster represented by the nearest centroid then followed by updating all the cluster centroids. The convergence criterion is met for this approach when the assignments of data instances to clusters no longer change after a finite number of trials. However, its result highly depends on the random initialization of the centroids since it may be stuck in a local optimum that could result in a low-quality clustering result. For this purpose, Hartigan's method is reported to be less prone to the initial selection of centroids [Hartigan & Wong, 1979, Telgarsky & Vattani, 2010].

In Hartigan's method, a data instance is analyzed and optimally re-assigned. The method considers the change of all centroids while re-assigning, and it is possible for a data instance not to be assigned to the nearest cluster. As a consequence of this, the set of local minima of Hartigan's algorithm is a strict subset of those of Lloyd's method, implying that the algorithm is less sensitive to the choice of the initial random partition of the data [Telgarsky & Vattani, 2010]. With these reasons, for our proposed integrative clustering framework, we used the heuristic approach for k-means designed by Hartigan and Wong (1979).

One of other problems related to standard k-means approaches is their lack of capturing nonlinear relationships between the data instances. Thus, k-means may not perform well if there are nonlinear relationships between the data instances such as in the case of genomic data sets. Therefore, one may think to integrate kernel methods to k-means clustering knowing that these methods are capable of detecting

such relationships. This integration is called kernel k-means clustering and proposed by Girolami (2002). Considering its aforementioned ability to capture nonlinear relationships of the data while clustering, kernel k-means clustering has been applied for handwritten digits recognition [Zhang & Rudnicky, 2002]. Moreover, a localized data fusion approach for kernel k-means has been applied to human colon and rectal cancer data in context of multiview learning [Gönen & Margolin, 2014].

In omics data clustering context, rather than clustering data coming from a single source, to examine the data in a deeper level, clustering of multiple omics data (i.e., multi-omics clustering) may be preferred since it may reduce the effect of experimental and biological noise and integrate different molecular aspects. Rappoport and Shamir (2018) prepared prepared a review for multi-omic and multi-view clustering algorithms and compared the performance of several existing methods using cancer data sets.

With the aforementioned motivations, we developed a multiple approximate kernel k-means clustering framework using approximation matrices and selecting the low-dimensional approximation matrices that are the most relevant to the clustering task[3]. Our proposed framework can be found in detail in Chapter 4.

In Chapter 5, we provide the experimental results associated with the algorithms proposed in Chapters 2–4. Finally, in Chapter 6, we conclude and discuss some possible future directions.

[3]Our proposed algorithm's implementation in R is available at `https://github.com/begumbektas/makkmeans`

Chapter 2

PATHWAY/GENE SET-BASED PREDICTION USING MULTIPLE KERNEL LEARNING

Most frequently used based learners for multiple kernel learning (MKL) algorithms are SVM and support vector regression (SVR) since they have been proven empirically successful and they are easily applicable as a building block [Gönen & Alpaydın, 2011]. In this chapter, we applied MKL algorithm using SVR as base learner on gene sets to discover mechanisms at the molecular level related to tumor initiation and progression (see Figure 2.1). The experimental results of our algorithm can be found in Chapter 5.

In this chapter, after describing the materials used for the defined problem, we discuss RF and SVR algorithms which we used as baseline algorithms. We then give the details of our novel algorithm, namely, PrognosiT.

2.1 Materials

We gathered genomic characterizations and clinical annotation files of over 10 000 cancer patients from 33 cancer cohorts in the Genomic Data Commons (GDC) data portal offered by TCGA consortium at `https://portal.gdc.cancer.gov`. TCGA provided the RNA-Seq measurements of the tumors from 33 cohorts and pre-processed them with a unified pipeline, which facilitated our analysis of tumor gene expression profiles. We downloaded `HTSeq-FPKM` files of all primary tumors from the most recent data freeze (i.e., `Data Release 33.1 - May 31, 2022`), which leads to 9 911 files in total. Please note that unless stated otherwise, this is the data freeze we used for the experiments throughout this thesis. Due to the strong elemental and hidden differences at the molecular level, we did not add metastatic tumors to our

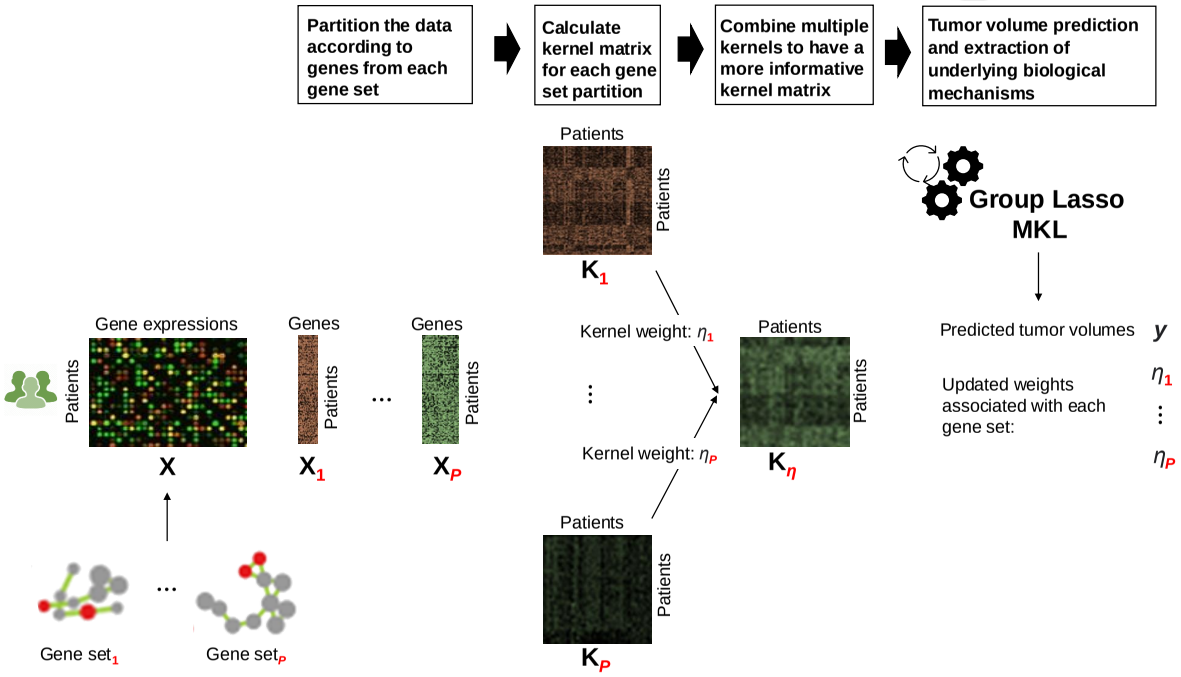


Figure 2.1: Our proposed MKL algorithm, named PrognosiT, which takes gene expression profiles of patients, denoted as $\mathbf{X}$, tumor volumes of patients, denoted as $\boldsymbol{y}$, and a pathway/gene set collection as its input. Then, it calculates distinct kernel matrices, denoted as $\mathbf{K}_1, \ldots, \mathbf{K}_P$, for input pathways/gene sets on gene expression matrices, denoted as $\mathbf{X}_1, \ldots, \mathbf{X}_P$, formed from the input matrix of gene expression profiles. Followingly, to have a kernel matrix between pairs of patients which is denoted as $\mathbf{K}_\eta$ and which carries more information, multiple kernel matrices are combined with a weighted sum. The resulting kernel matrix is later used to learn a function, denoted as f, to predict tumor volumes of out-of-sample cancer patients.

analysis. We used clinical annotation files of the patients to obtain the tumor volume information. We checked the clinical annotation files for tumor dimension information (i.e., tumor length, tumor depth and tumor width) and there were only two cohorts containing this information that are sarcoma (`SARC`) and thyroid carcinoma (`THCA`) cohorts. By nature, malignant soft tissue tumors (i.e., sarcomas) have numerous subtypes with different prognoses and therefore, with different molecular mechanisms related to cancer progression. Concordantly, when we checked the histological type information in clinical annotation files, we saw that histological types of cancer tissues were highly different within `SARC` cohort. Thus, we excluded `SARC` cohort from further analysis, and used `THCA` cohort that has 507 patients in total.

We first calculated tumor volume for each tumor by multiplying `neoplasm_length`, `neoplasm_width` and `neoplasm_depth` found in the clinical annotation files. Afterwards, we picked the patients that have both tumor dimension information and gene expression profile. We then discarded patients having their tumor volume `nonpositive` or `NA`, leading us to 402 primary tumors in `THCA` cohort. Lastly, to compare the gene expressions between tumor and normal tissues, we used the normal tissue gene expression profiles found in 58 patients.

In addition to having a predictive model for tumor volume, we aimed to discover the molecular processes that play key roles in this volume prediction task. Therefore, we used cancer-specific pathway/gene set collections previously depicted in the literature. Using these collections, we can determine the group of genes having similarities or dependencies in their functionalities.

Using the Molecular Signatures Database (MSigDB), we extracted `Hallmark` gene sets and pathways from Pathway Interaction Database (`PID`), which are specifically curated for cancer-related research tasks. `Hallmark` gene set collection contains computationally constructed list of genes that convey a particular biological state or process and shows coherent expression in cancers [Liberzon et al., 2015]. `PID` is a freely available collection of manually curated and peer-reviewed pathways that consists of human molecular signaling and regulatory events and major cellular processes [Schaefer et al., 2009]. The `Hallmark` gene set collection contains 50 gene

sets with sizes varying between 32 and 200, whereas the `PID` collection include 196 pathways with sizes varying between 10 and 137.

2.2 Random Forest

RF algorithm is a combination of weak decision trees. It uses an ensemble strategy to get more robust classification and regression trees than decision trees algorithm [Breiman, 2001]. Because of its data-adaptive structure, RF is appealing for high-dimensional genomic data analysis [Chen & Ishwaran, 2012]. We chose RF as a baseline algorithm due to the fact that it is highly used in the studies with genomic input data, and it is capable of handling the noise and the correlation among features [Shi et al., 2005, Díaz-Uriarte & Alvares de Andrés, 2006, Stephan et al., 2015]. Despite the fact that predictive performance of RF is claimed to be good in certain applications, their capability to extract meaningful knowledge from data is highly inadequate. Additionally, since RF models are generally built by randomly selecting bootstrap samples, their knowledge extraction process may vary remarkably.

2.3 Support Vector Regression

SVR is a modified version of SVM algorithm [Cortes & Vapnik, 1995] to be used for prediction tasks [Drucker et al., 1996]. Our proposed MKL algorithm uses SVR as the base learner. The optimization problem used for SVR is

$$\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}(\xi_i^+ + \xi_i^-) \\
\text{with respect to} \quad & \boldsymbol{w} \in \mathbb{R}^D, \;\; \boldsymbol{\xi}^+ \in \mathbb{R}^N, \;\; \boldsymbol{\xi}^- \in \mathbb{R}^N, \;\; b \in \mathbb{R} \\
\text{subject to} \quad & \epsilon + \xi_i^+ \geq y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b \qquad \forall i \\
& \epsilon + \xi_i^- \geq \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b - y_i \qquad \forall i \\
& \xi_i^+ \geq 0 \qquad \forall i \\
& \xi_i^- \geq 0 \qquad \forall i,
\end{aligned} \tag{2.1}$$

where $\boldsymbol{w}$ is the feature weight vector, C is the nonnegative regularization parameter, $\boldsymbol{\xi}^+$ and $\boldsymbol{\xi}^-$ are the sets of slack variables, D is the number of input features (number

of genes in gene expression profiles), b is the intercept parameter, and ϵ is the nonnegative tube width parameter.

Solving the dual of the above optimization problem would decrease the number of decision variables and thereby, we could integrate kernel functions in the problem formulation, to be able to model nonlinear problems. The corresponding dual optimization problem is

$$\begin{aligned}
\text{minimize} \quad & -\sum_{i=1}^{N} y_i(\alpha_i^+ - \alpha_i^-) + \epsilon \sum_{i=1}^{N}(\alpha_i^+ + \alpha_i^-) \\
& + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\
\text{with respect to} \quad & \boldsymbol{\alpha}^+ \in \mathbb{R}^N, \quad \boldsymbol{\alpha}^- \in \mathbb{R}^N \\
\text{subject to} \quad & \sum_{i=1}^{N}(\alpha_i^+ - \alpha_i^-) = 0 \\
& C \geq \alpha_i^+ \geq 0 \qquad \forall i \\
& C \geq \alpha_i^- \geq 0 \qquad \forall i.
\end{aligned} \tag{2.2}$$

In the dual optimization problem, there are $2N$ decision variables instead of $(D + 2N + 1)$, which is the number of decision variables in the primal problem. To build nonlinear models, we can add the kernel function to the dual formulation by replacing $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ term with $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where we encode the similarities between pairs of data points. This kernel function is usually selected with a model selection approach by trying several alternatives.

2.3.1 Derivation of dual optimization problem for support vector regression

The Langrangian function corresponding to the primal optimization problem shown in Equation 2.1 is calculated as

$$\begin{aligned}
\mathcal{L} = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{N}(\xi_i^+ + \xi_i^-) - \sum_{i=1}^{N}\alpha_i^+(\epsilon + \xi_i^+ - y_i + \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \\
- \sum_{i=1}^{N}\beta_i^+\xi_i^+ - \sum_{i=1}^{N}\alpha_i^-(\epsilon + \xi_i^- - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - b + y_i) - \sum_{i=1}^{N}\beta_i^-\xi_i^-.
\end{aligned} \tag{2.3}$$

We then take the derivative of the Lagrangian function with respect to all the decision variables of the primal optimization problem, we get

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{N} (\alpha_i^+ - \alpha_i^-)\boldsymbol{x}_i \tag{2.4}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_{i=1}^{N} (\alpha_i^+ - \alpha_i^-) = 0 \tag{2.5}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i^+} = 0 \Rightarrow C = \alpha_i^+ + \beta_i^+ \qquad \forall i \tag{2.6}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i^-} = 0 \Rightarrow C = \alpha_i^- + \beta_i^- \qquad \forall i. \tag{2.7}$$

Lastly, by plugging the derivation results back into the Lagrangian function we obtain the dual optimization problem shown in Equation 2.2.

2.4 Our Proposed Algorithm

We approached the problem of predicting the volume of primary tumors at the diagnosis while simultaneously determining the molecular mechanisms that affect tumor progression by applying machine learning algorithms on gene expression profiles extracted from the tumors. For a cohort consisting of N patients, the training data set can be represented as $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where N denotes the number of tumors, $\boldsymbol{x}_i$ denotes the gene expression profile related to tumor i, and y_i is the volume of tumor i.

Aforementioned problem may be formulated using a regression model and can be solved using algorithms for regression such as RF [Breiman, 2001] and SVR [Drucker et al., 1996]. With these algorithms, it may be possible to have good predictive performances. However, good prediction performance is not enough to extract insightful information about the mechanisms that play role in tumor progression. It is shown that, for predicting outcome in cancer, size of the training sample set should be at least in the order of thousands [Ein-Dor et al., 2006]. In other words, since gene expression data is highly correlated by its nature, if the training sample set is not in order of thousands, machine learning algorithms might use different subsets of a specific patient cohort to predict and might result with different biomarkers for

prediction. Therefore, it would be sensible to use our prior knowledge regarding genes, information from pathways/gene sets, and discover mechanisms at the molecular level using this prior information.

The predictive ability of machine learning algorithms that uses "kernel trick" to capture patterns between pairs of data points is quite dependent on the selected kernel function. The standard practice for selecting a kernel function consists of first comparing the performances of several candidate kernel functions with the aid of a cross-validation technique on the training data and then selecting the kernel function that performs best on the training set to make predictions on the test set. Nevertheless, usage of a single kernel function might not be sufficient for handling the complexity of the studied machine learning problem, but a combination of different kernels might give better predictive results than a single one. Moreover, there exist many algorithms that combine multiple kernels to capture the similarity between pairs of data points. Different kernels in MKL may correspond to using different measures of similarity or they may be using information coming from different sources (i.e., different feature representations or different feature subsets) [Gönen & Alpaydın, 2011]. For instance, MKL algorithms might combine kernel functions that have different complexities (e.g., linear, polynomial or Gaussian) defined on the same input representation or they might combine kernels prepared from different sources of input data (i.e., multiview learning; data fusion from multiple feature sets). To be used in cancer research, we can train algorithms using the same set of tumors in different representations such as gene expression, copy number or methylation profiles.

In this chapter, one of our purposes was to discover biological mechanisms that define tumor prognosis. To that end, we propose a modified version of SVR using multiple kernel learning on pathways/gene sets (PrognosiT). We first form a kernel matrix for each pathway/gene set and then we combine these kernel matrices using an MKL algorithm. We assume that we are given P kernel functions instead of a single one for PrognosiT algorithm, where we calculate a weighted sum of these kernel functions. In other words, we get a convex combination of these kernel functions

(i.e., sum of nonnegative kernel weights is set equal to one).

To integrate MKL into the SVR model, the dual of the support vector optimization model can be used as an inner problem within the following outer optimization model:

$$\begin{aligned} \text{minimize} \quad & J(\boldsymbol{\eta}) \\ \text{with respect to} \quad & \boldsymbol{\eta} \in \mathbb{R}^P \\ \text{subject to} \quad & \sum_{m=1}^{P} \eta_m = 1 \\ & \eta_m \geq 0 \qquad \forall m, \end{aligned} \tag{2.8}$$

where $\boldsymbol{\eta}$ represents the kernel weights, and $J(\boldsymbol{\eta})$ is the optimization problem depicted in Equation (2.2) with a modified objective function, that replaces $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ term with $\sum_{m=1}^{P} \eta_m k_m(\boldsymbol{x}_i, \boldsymbol{x}_j)$. It should be noted that the equality constraint in Equation (2.8), which is otherwise known as the unit simplex constraint, represents enforcing ℓ_1-norm on the kernel weights and leads us to have a sparse solution. The optimization problem in Equation (2.8) is convex with respect to $\boldsymbol{\eta}$, and the optimization problem in Equation (2.2) is convex with respect to $\{\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-\}$, but the nested optimization problem is not jointly convex with respect to $\boldsymbol{\eta}$ and $\{\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-\}$. Since we cannot solve this nested optimization problem globally, we inspire from group Lasso MKL algorithm that was initially constructed for binary classification tasks [Xu et al., 2010]. We initiate the algorithm by setting all kernel weights equal to each other. Since the summation of all kernel weights should be one, each kernel weight is $1/P$ at the initialization. We then solve the inner optimization problem (i.e., a standard SVR model) at each iteration t by utilizing the current kernel weights $\boldsymbol{\eta}^{(t)}$ to obtain the support vector coefficients $\{\boldsymbol{\alpha}^{+(t)}, \boldsymbol{\alpha}^{-(t)}\}$. We then calculate the kernel weights at the next iteration using the following update equation:

$$\eta_m^{(t+1)} = \frac{\eta_m^{(t)} \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i^{(t)} \alpha_j^{(t)} k_m(\boldsymbol{x}_i, \boldsymbol{x}_j)}}{\sum_{o=1}^{P} \eta_o^{(t)} \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i^{(t)} \alpha_j^{(t)} k_o(\boldsymbol{x}_i, \boldsymbol{x}_j)}} \qquad \forall m, \tag{2.9}$$

where $\alpha_i^{(t)} = (\alpha_i^{+(t)} - \alpha_i^{-(t)})$, and the superscripts (t) and $(t+1)$ depicts the current and next iterations, respectively. This is an iterative solution methodology and is

demonstrated to converge for binary classification problems [Xu et al., 2010]. In each iteration, we can solve the inner optimization problem to optimality since it is a standard SVR formulation when we fix the kernel weights $\boldsymbol{\eta}$. We can also solve the outer optimization problem to optimality when we fix the sample weights $\{\boldsymbol{\alpha}^+, \boldsymbol{\alpha}^-\}$. These two steps are monotonically decreasing the objective function, leading to convergence.

2.4.1 Kernel selection

When our algorithm converges, we can note the final $\boldsymbol{\eta}$ values to identify which kernels are included (i.e., nonzero η_m values) in the final model. We perform kernel selection in a supervised manner by conjointly learning regression coefficients and kernel weights. Thanks to the ℓ_1-norm on the kernel weights, we obtain sparse kernel weights, leading to eliminating irrelevant kernel from the combination. Additionally, we can compare the significance of pathways/gene sets by comparing their kernel weights, which could give us valuable information about biological processes towards tumor progression.

Chapter 3

MULTIPLE APPROXIMATE KERNEL LEARNING

With increasing size of patient cohorts and genomic characterizations collected from tumor biopsies, machine learning algorithms have been applied to different tasks in the hope of discovering molecular mechanisms related to the diagnosis, prognosis and treatment of cancer. Together with good predictive performance, building interpretable machine learning algorithms plays a crucial role in cancer research to discover new therapeutic biomarkers and to analyze the heterogeneity of tumors. Kernel-based machine learning algorithms have been widely used in computational biology, thanks to their capability to deal with highly correlated data and interpretability. However, the problem with these existing algorithms is their lack of scalability, which started to cause problems especially for single-cell cancer research, since traditional kernel-based methods are computationally prohibitive with large sample sizes.

The main contribution of this chapter is to extend well-known MKL idea to large-scale problems using an approximation procedure for kernel matrices. The proposed algorithm was shown to have four important characteristics: (i) running in a scalable manner on large-scale genomic data, (ii) getting sparse solutions, (iii) integrating prior information during the learning process to increase the interpretability and (iv) maintaining or even increasing the predictive performance.

3.1 Materials

In this chapter, we used three data sets formed using two data sources. First, we used data from TCGA consortium to discover the molecular mechanisms related to early- and late-stage cancers and two-year survival of patients. Second, we used single-cell RNA-Seq data provided by the Broad Institute Single Cell Portal to

understand molecular signatures related to cell malignancy, which could be beneficial for understanding tumor plasticity and cancer progression mechanisms.

3.1.1 TCGA data set

We gathered genomic data and clinical annotation files of over 10 000 cancer patients from 33 cancer cohorts in the GDC data portal offered by TCGA consortium at `https://portal.gdc.cancer.gov`. TCGA shared the RNA-Seq measurements of the tumors from 33 cohorts and pre-processed them with a unified pipeline. We downloaded `HTSeq-FPKM` profiles of all primary tumors from the most recent data freeze and finally got 9 911 files in total. To obtain tumor grade levels and survival-related information, we used clinical annotation files of the patients.

For binary classification task of cancer stages, we considered patients clinically annotated as `Stage I` as having early-stage cancer; patients clinically annotated with `Stage II`, `III` or `IV` as having late-stage cancer. We included cohorts having at least 20 tumors from both categories. We combined data coming from all cohorts to form one big data set of 19 814 features. As a result, the final data set contains 5 547 patients coming from 15 distinct cohorts.

For two-year survival classification, we considered patients that have `vital_status`, `days_to_death`, `days_to_last_followup`, `days_to_last_known_alive` information. We combined data from 33 cohorts by including patients with the required survival information. We included cohorts having at least 20 patients whose `vital_status` is dead and at least 100 patients in total. The resulting data set contains 5 168 patients coming from 20 distinct cohorts with 19 814 expression features. We labeled the patient as having greater than or equal to two-year survival (i) if `vital_status` is alive and `days_to_last_followup` or `days_to_last_known_alive` is greater than or equal to two years; (ii) if `vital_status` is dead and `days_to_death` is greater than or equal to two years. On the other hand, we labeled the patient as having less than two-year survival if `vital_status` is dead and `days_to_death` is less than two years.

3.1.2 Single-cell RNA-Seq data set

Analysis of single-cell RNA-Seq data holds promise to find new ways to treat cancer. To reveal the mechanisms related to tumor plasticity for targeted cancer therapies and to find new biomarkers, single-cell analysis is in the spotlight. Having this motivation, we searched for the data sets in the Broad Institute Single Cell Portal at `https://singlecell.broadinstitute.org` and found the melanoma immunotherapy data set, which is appropriate for our binary classification task, provided by Jerby-Arnon et al. (2018) as a part of their research.

We intersected the gene expression profiles of 7 186 cells with malignancy information coming from 6 738 clinical annotation files. For each cell, there were 23686 gene expression measurements in the data set. If a cell is annotated as `B.cell`, `CAF`, `Endothelial.cell`, `Macrophage`, `NK`, `T.CD4`, `T.CD8` or `T.cell`, we labeled this cell as `nonmalignant`. All other cells were labeled as `malignant`.

3.1.3 Pathway/gene set collections

In this chapter, we used the two pathway/gene set collections that we have described in Chapter 2, that are extracted from the MSigDB and specifically curated for cancer research: `Hallmark` gene set collection [Liberzon et al., 2015] and Pathway Interaction Database (`PID`) pathway collection [Schaefer et al., 2009], as prior information sources to be used for calculation of multiple approximate kernel matrices.

3.2 MAKL: Fast Multiple Approximate Kernel Learning

In MKL context, storage requirements associated with exact kernel matrices and learning with them are the reason of high computational cost. To bypass this computationally expensive step of MKL, we used low-dimensional approximation matrices instead of exact kernel matrices. After accelerating the kernel matrix computation step, we combined approximation matrices calculated for each pathway/gene set. We then fed the resulting combined matrix into group Lasso formulation, which led to sparsity in pathway/gene set level and gave interpretable results.

We would like to begin by noting that in kernel-based methods, the idea of integrating kernel matrices into a linear model to solve a nonlinear problem is achieved by applying the "kernel trick". The kernel trick is a result from the fact that any positive definite function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^d$ defines an inner product and a mapping function $\Phi(\cdot)$ so that the inner product between mapped data points can be quickly computed as $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle$ [Schölkopf & Smola, 2002].

To accelerate the kernel matrix computation step of MKL, we used a modified version of random Fourier features [Rahimi & Recht, 2007]. Instead of using the implicit mapping by the aforementioned kernel trick, Rahimi and Recht (2007) proposed explicitly mapping the data to a low-dimensional Euclidean inner product space using a randomized feature map $z \colon \mathbb{R}^d \to \mathbb{R}^D$, so that the inner product between a pair of mapped points $z(\boldsymbol{x}_i)$ and $z(\boldsymbol{x}_j)$ approximates their kernel evaluation: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle \approx \langle z(\boldsymbol{x}_i), z(\boldsymbol{x}_j) \rangle$. This calculation of random features is notable since this new feature map $z(\cdot)$ is low-dimensional. In fact, the dimensionality of randomized feature space D is given as an input to algorithm, and it can cope with large-scale data well, since instead of calculating kernel matrices which are of size $N \times N$, we now need to calculate matrices of size $N \times D$, where $D \ll N$.

Once the appropriate low-dimensional approximation method for the kernel matrices is set, another problem arises at how to integrate this approximation matrices into the learning process.

As a popular model selection and shrinkage estimation method, the Lasso estimator had been proposed [Tibshirani, 1996]. Then, the group Lasso has been proposed as an extension to the Lasso algorithm [Bakin, 1999, Antoniadis & Fan, 2001, Cai, 2001, Yuan & Lin, 2006]. The group Lasso is the least-square regression problem with regularization by a block ℓ_1-norm. This estimator works well in case of dealing with grouped data which in fact leads to the idea of MKL algorithm. Bach (2008) discussed theoretical aspects of the consistency of the group Lasso and using multiple kernel matrices.

Despite giving interpretable results, the problem with all ordinary MKL methods is their scalability issue due to their need to calculate distinct kernel matrices, which

scale quadratically with the sample size. This issue could be problematic once they are applied to large-scale data sets in areas such as computational biology, computer vision, etc.

3.2.1 Random features for kernel approximation

Owing to high computational complexity of using kernel trick in large-scale data sets, the question of finding faster methods arises. Instead of computing the kernel matrices, to accelerate the learning process, there is a need to find a low-dimensional approximation for kernel matrix calculation. For this purpose, we approximate the kernel matrices using a modified version of random feature mapping originally has been introduced by Rahimi and Recht (2007). This approximation method is a natural result from a classical harmonic analysis as presented below.

Theorem 1 *A continuous kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \kappa(\boldsymbol{x}_i - \boldsymbol{x}_j)$ on $\mathbb{R}^d$ is positive definite if and only if $\kappa(\theta)$ is the Fourier transform of a nonnegative measure [Rudin, 1994].*

Briefly, as a result of Bochner's theorem, we are ensured that for any properly scaled shift-invariant kernel $k(\cdot, \cdot)$, its Fourier transform is a proper probability distribution and $\langle z(\boldsymbol{x}_i), z(\boldsymbol{x}_j)\rangle$ is an unbiased estimate of $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in case the random samples are drawn from the Fourier transform $p(\cdot)$. The Fourier transform $p(\cdot)$ for a given kernel $\kappa(\cdot)$ is

$$p(\boldsymbol{\omega}) = \frac{1}{2\pi} \int e^{-j\boldsymbol{\omega}^\top \theta} \kappa(\theta)\, d\theta. \tag{3.1}$$

To calculate similarity between pairs of gene expression profiles, we used the Gaussian kernel on feature sets in Algorithm 1, which is a modified version of the random feature mapping algorithm proposed by Rahimi and Recht (2007). We approximated a given Gaussian kernel matrix by getting a low-dimensional approximation matrix $\mathbf{Z}$ as the output from this algorithm, where we used the $\sin(\cdot)$ function in addition to the $\cos(\cdot)$ function to guarantee that $\mathbf{Z}\mathbf{Z}^\top$ produces a kernel matrix with ones on the diagonal. Due to this modification, we also changed the normalizing constant from $\sqrt{2/D}$ to $\sqrt{1/D}$.

The Gaussian kernel was previously reported in the literature as being reliable to be used with high-dimensional genomic data [Costello et al., 2014, Gönen et al., 2017]. We used the same approach to discover highly nonlinear dependency between the RNA-Seq data and the corresponding clinical annotations. The Gaussian kernel is

$$k_{\mathcal{G}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{x}_i - \boldsymbol{x}_j \rangle}{2\sigma^2}\right), \tag{3.2}$$

where σ is the kernel width parameter. The Fourier transform of the Gaussian kernel is

$$p(\boldsymbol{\omega}) = \sqrt{2\pi}\sigma e^{-\|\boldsymbol{\omega}\|_2^2 \sigma^2/2}. \tag{3.3}$$

It is worth noting here that the Fourier transform of a Gaussian function has also Gaussian distribution. Additionally, the kernel width parameter σ in time space corresponds to σ^{-1} in Fourier space.

3.2.2 Fast integration with group Lasso

To be used in an MKL setting, we integrated the kernel approximations into the group Lasso formulation. After partitioned the input data according to feature sets, we computed distinct kernel matrix approximations for each feature set. Then, we concatenated the approximation matrices and used the concatenated matrix as an input to the group Lasso formulation.

The group Lasso minimizes the following objective function with respect to model parameters $\beta_0 \in \mathbb{R}, \boldsymbol{\beta}_1 \in \mathbb{R}^{d_1}, \boldsymbol{\beta}_2 \in \mathbb{R}^{d_2}, \ldots, \boldsymbol{\beta}_P \in \mathbb{R}^{d_P}$:

$$\sum_{i=1}^{N} \ell\left(y_i, \beta_0 + \sum_{p=1}^{P} \langle \boldsymbol{\beta}_p, \boldsymbol{x}_{ip} \rangle\right) + \lambda \sum_{p=1}^{P} \sqrt{d_p} \|\boldsymbol{\beta}_p\|_2, \tag{3.4}$$

where $\ell(\cdot, \cdot)$ denotes the loss function according to problem type (e.g., square loss for regression problems, logistic loss for binary classification problems), y_i is the output for i^{th} data point (i.e., response value for regression problems, class label for binary classification problems), λ is a positive regularization coefficient, d_p refers to the size of feature set p, and $\|\cdot\|_2$ is the Euclidean norm. This optimization problem is

convex with respect to the weights $\boldsymbol{\beta}_p$ associated with feature sets. The fact that the Euclidean norm of a vector is zero only if all vector entries are zero, the group Lasso formulation leads to sparse solutions at group level, which facilitates the feature set selection process.

Overall, our algorithm for MKL using any kernel approximation method is described in Algorithm 2 with details. Figure 3.1 displays an overview of our proposed algorithm that combines Algorithm 1 and Algorithm 2.

Algorithm 1 Fast computation of low-dimensional approximation matrix using random Fourier features

---

Input: A matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, a positive definite shift-invariant kernel, dimension D ($D \ll N$) to calculate random Fourier features, subset size S ($S \ll N$) to calculate the distance matrix.

Output: Low-dimensional approximation matrix $\mathbf{Z} \in \mathbb{R}^{N \times 2D}$.

Draw S random rows from $\mathbf{X}$.

Compute the Euclidean distance matrix $\mathbf{S} \in \mathbb{R}^{S \times S}$.

Calculate the kernel width parameter σ as the mean of the distance matrix $\mathbf{S}$.

Draw D i.i.d. random samples $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_D \in \mathbb{R}^d$ from $p(\cdot)$ using (3.1).

Draw D i.i.d. random samples $b_1, b_2, \ldots, b_D \in \mathbb{R}$ from the uniform distribution on $[0, 2\pi]$.

Compute $\mathbf{Z}$ as

$$\mathbf{Z} = \sqrt{\frac{1}{D}}[\cos(\mathbf{X}\boldsymbol{w}_1 + b_1) \ \ldots \ \cos(\mathbf{X}\boldsymbol{w}_D + b_D) \ \sin(\mathbf{X}\boldsymbol{w}_1 + b_1) \ \ldots \ \sin(\mathbf{X}\boldsymbol{w}_D + b_D)].$$

Algorithm 2 Fast integration of low-dimensional approximation matrices into group Lasso

Input: Training matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, P feature sets $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_P$, binary output vector $\boldsymbol{y}$ of length N.

Output: Classification AUROC value, ℓ_2-norm of the weights assigned to $\mathcal{F}_p$ denoted as η_p for all $p \in P$.

From the training matrix $\mathbf{X}$, generate P submatrices, using $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_P$.

Compute low-dimensional approximation matrices $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_P$.

Concatenate distinct approximation matrices $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_P$ to obtain one single matrix $\mathbf{Z} \in \mathbb{R}^{N \times 2DP}$.

Solve the group Lasso formulation using $\mathbf{Z}$ as the input matrix.

Calculate ℓ_2-norm using the weights of each $\mathcal{F}_p$ and denote the corresponding norm calculated for $\mathcal{F}_p$ as η_p for all $p \in P$.

If η_p is nonzero, count $\mathcal{F}_p$ selected for the classification task.

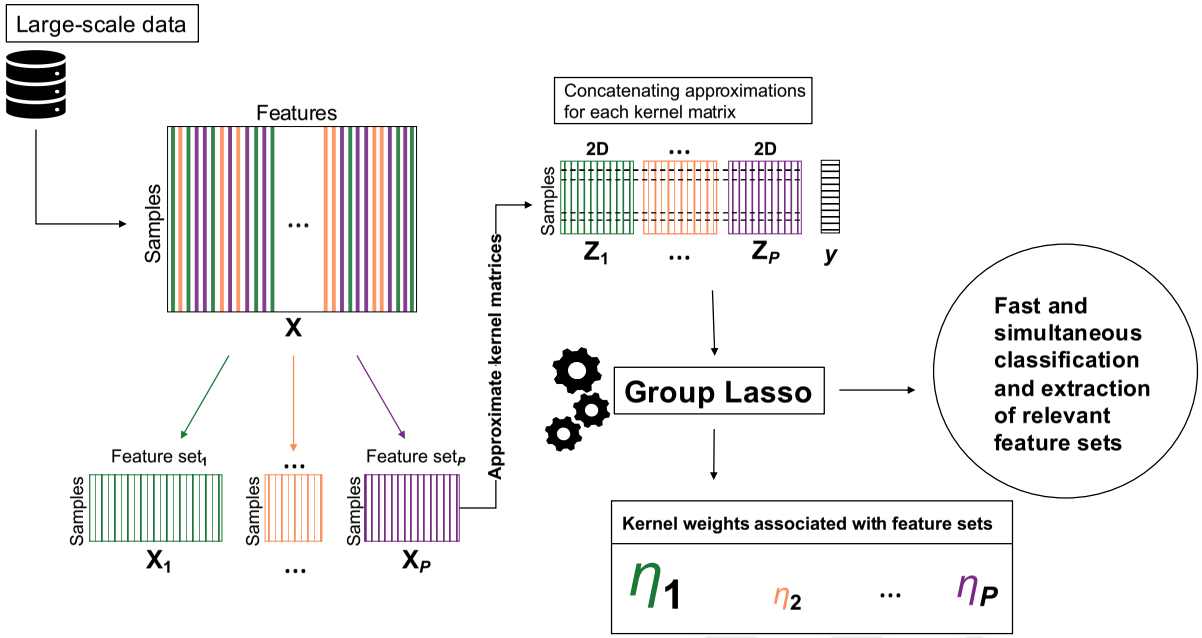


Figure 3.1: Our proposed multiple approximate kernel learning algorithm, namely MAKL, that integrates kernel approximation and group Lasso formulations into a conjoint model. It takes the input matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, a binary outcome vector $\boldsymbol{y}$ of length N, and a pathway/gene set collection consisting of P pathways/gene sets, as its inputs. It then computes distinct kernel approximation matrices for each pathway/gene set, denoted as $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_P$; and finally combine all kernel approximation matrices to get $\mathbf{Z} \in \mathbb{R}^{N \times 2DP}$, to be used in the group Lasso formulation. During learning process, it determines the weights associated with each pathway/gene set, denoted as $\eta_1, \ldots, \eta_P$. It is scalable since it integrates a low-dimensional kernel matrix approximation instead of usual kernel matrix computation; it is interpretable since it performs feature selection and sorts pathways/gene sets according to their relevance by comparing their ℓ_2-norms.

Chapter 4

MULTIPLE APPROXIMATE KERNEL K-MEANS CLUSTERING

The main contribution of this chapter is to integrate a kernel approximation method to well-known k-means clustering algorithm to be easily used with large-scale data sets while giving interpretable clustering results by extracting relevant parts from prior multiple gene set information. It is expected that, in this chapter, using data from multiple cancer patient cohorts provides a general idea about the separation characteristics across all used cancer types. The proposed design has three notable characteristics: (i) being scalable in terms of the data set size, (ii) capturing nonlinear relationships of the data while clustering, (iii) integrating prior multiple gene set information to the clustering process for interpretability.

4.1 Materials

For this chapter, we used the data provided by TCGA consortium that includes RNA-Seq measurements, mutation information and clinical annotations associated with cancer patients from 33 patient cohorts. The data we used in our analyses are publicly available at the GDC data portal at `https://portal.gdc.cancer.gov`.

4.1.1 RNA-Seq measurements

We downloaded `HTSeq-FPKM` profiles of all primary tumors from the most recent data freeze which provided us 9 911 patient files from 33 patient cohorts. We selected the patient cohorts that have at least 300 patients within, so that the resulting data set to be clustered would be balanced in terms of cancer tissue type. There were 15 such cohorts that are bladder urothelial carcinoma (`BLCA`), breast invasive carcinoma

(BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), brain lower grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA) and uterine corpus endometrial carcinoma (UCEC). The number of patients included in the final RNA-Seq data set was 7 474.

4.1.2 Mutation information

We extracted the mutation information associated with each patient cohort in TCGA and combined them to build a mutation data set. We first counted the total number of Missense_Mutation for each patient on each gene. We then formed a dichotomised version of this data set to analyze the existence of a Missense_Mutation on a specific gene of a patient, rather than the total number of missense mutations. If a patient for a specific gene, had one or more missense mutations, we overwrote the corresponding number of missense mutations as 1. If a patient for a specific gene, had no Missense_Mutation, we left the corresponding number of missense mutations as 0.

4.1.3 Histological grade information

TCGA consortium provides clinical information associated with RNA-Seq profiles and mutation information for not all but many patients. We extracted histological grade information from the clinical annotation files where the patients were labeled into 8 categories which are G1, G2, G3, G4, GB, GX, High Grade, Low Grade. According to the information provided by National Cancer Institute on tumor grades at https://www.cancer.gov/about-cancer/diagnosis-staging/prognosis/tumor-grade-fact-sheet, grading system is defined such that GX refers to Grade cannot be assessed (undetermined grade); G1 refers to Well differentiated, G2 refers to Moderately differentiated; G3 refers to Poorly differentiated and

`G4` refers to `Undifferentiated`. We overwrote the histological grade information for patients that were originally categorised into `G1` and `Low Grade` as `Low Grade`; `G2` as `Intermediate Grade`; `G3`, `G4` and `High Grade` as `High Grade`.

4.1.4 Gene set collection

In this chapter, to provide interpretability while clustering in the context of cancer research, we integrated information coming from `Hallmark` gene set collection from the MSigDB [Liberzon et al., 2015] into the clustering algorithm.

4.2 Scalable and Integrative Multiple Approximate Kernel k-Means Clustering

As mentioned in earlier sections, learning with exact kernel matrices for algorithms that use multiple kernels is computationally expensive and challenging to use particularly in large-scale data analysis. With this motivation, we used low-dimensional approximation matrices instead of calculating exact kernel mappings. We then integrated prior information coming from gene set collection by designing a forward selection k-means framework to get interpretable clustering results. This framework provides a subset of the prior information fed to clustering algorithm, that is relevant to the clustering task.

In the previous chapter, we have developed a fast and scalable binary classification algorithm, namely, MAKL, that integrates randomized kernel approximation to group Lasso algorithm in a novel way, and we tested our algorithm using data from TCGA and a single-cell immunotherapy data [Bektaş et al., 2022]. The promising results of our previous work led us to design a clustering approach that integrated pathway/gene set information, that is scalable and capable of capturing the nonlinear patterns of the data.

In this chapter, after obtaining the low-dimensional approximation matrices instead of exact kernel matrices, we designed a forward set selection algorithm to integrate these multiple approximation matrices into the clustering algorithm and to make it interpretable. We calculated silhouette score for each selection combination

until a pre-defined stopping criterion is met and selected the best gene set combination as the one that gives the highest silhouette score. At the end, along with clustering results, we had the relevant gene set information for the clustering task.

4.2.1 Randomized approximation of kernel matrices

To accelerate the kernel integrated clustering for large-scale data sets, due to the computationally expensive nature of learning with multiple kernel matrices, we approximated the kernel matrices using a modified version of random feature mapping that had been contributed to literature by Rahimi and Recht (2007). This modified version of random feature mapping has been experimented for a variety of classification tasks and demonstrated to give promising results in the genomic data analysis context [Bektaş et al., 2022]. As has been mentioned in Chapter 3, random feature mapping is a natural result from the classical harmonic analysis theorem called Bochner's theorem [Rudin, 1994]. As a result of this theorem, it is guaranteed that for any properly scaled shift-invariant kernel $k(\cdot, \cdot)$, its Fourier transform is a proper probability distribution and $\langle z(\boldsymbol{x}_i), z(\boldsymbol{x}_j)\rangle$ is an unbiased estimate of $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in case the random samples are drawn from the Fourier transform $p(\cdot)$.

The steps that are associated to randomized kernel approximation are given in detail in Algorithm 1. Inputs of this algorithm are a matrix $\mathbf{X} \in \mathbb{R}^{N\times d}$, a positive definite shift-invariant kernel function, dimension D ($D \ll N$) to calculate random Fourier features, subset size S ($S \ll N$) to calculate the distance matrix. It is noteworthy to state that, in addition to the steps in Algorithm 1, we also calculate the Euclidean distance matrix $\mathbf{E} \in \mathbb{R}^{S\times S}$ using S ($S \ll N$) rows of the input data to calculate the silhouette scores in Algorithm 3. Overall, the outputs of the algorithm to be used in our clustering approach are low-dimensional approximation matrix $\mathbf{Z} \in \mathbb{R}^{N\times 2D}$, Euclidean distance matrix $\mathbf{E} \in \mathbb{R}^{S\times S}$.

4.2.2 Selection of low-dimensional approximation matrices

To integrate prior information to clustering and to get sparsity in gene set level, we developed a greedy heuristic framework, which selects the combination of low-

dimensional approximation matrices giving the best silhouette score. Silhouette score is a measure to detect how similar a data instance is to the cluster it belongs, compared to other clusters (i.e., measure of tightness and separation), which is originally contributed to literature by Rousseeuw (1987). It is used to identify the quality of clusters in terms of how-well the clusters are separated. By its nature, it takes values on the range $[-1, +1]$, where higher values indicate better separability. Silhouette score for a data instance is calculated as $(n - m)/\max(n, m)$, where m refers to the mean distance to the most similar cluster, and n refers to the mean intra-cluster dissimilarity.

Silhouette score is calculated using a similarity matrix for all the data instances, where we used again a novel approach to provide scalability in our framework. We used dissimilarity matrices (i.e., Euclidean distance matrices) that had already been calculated to approximate the kernel matrices using a subset of the data set (i.e., S rows where $S \ll N$) to calculate the silhouette score. After having calculated the Euclidean distance matrices using only a small number of rows from the raw data, we calculated their corresponding kernel matrix $\mathbf{K}$ and used $2 - 2\mathbf{K}$ as the distance matrix in our silhouette score calculations.

The details of the algorithm that integrates prior information to the clustering framework and iteratively calculates silhouette score in a scalable way can be found in Algorithm 3. An overall demonstration of our proposed approach that uses randomized feature mapping and Algorithm 3 is shown in Figure 4.1.

Algorithm 3 Forward selection of low-dimensional approximation matrices

Input: A matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, P feature sets $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_P$, Euclidean distance matrices resulted from randomized kernel approximation algorithm, a list of gene set candidates $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_P\}$, a number of gene sets to select G.

Output: An array $\boldsymbol{c}$ of length N of cluster memberships, an array $\boldsymbol{s}$ of length G of silhouette scores, a list $\mathcal{L}$ of length G of relevant gene set information.

From $\mathbf{X}$, using $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_P$, generate $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_P$.

Compute low-dimensional approximation matrices $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_P$.

Initialization: Create a list $\mathcal{L}$ that will correspond to selected gene sets.

$t \leftarrow 0$

while $t \leq G$

Apply k-means to $\mathbf{Z}_{\mathcal{L}_t \cup m}$ $\forall m$ where $m \in \mathcal{C}$.

Calculate the corresponding silhouette scores.

Pick $\mathbf{Z}_{\mathcal{L}_t \cup m}$ giving the maximum silhouette score to determine the selected gene set.

Store this silhouette score as s_t.

Add the selected gene set to $\mathcal{L}_t$.

Delete the selected gene set from $\mathcal{C}$.

$t \leftarrow t + 1$

end while

Pick the maximum silhouette score among elements of $\boldsymbol{s}$.

Store the corresponding cluster memberships as $\boldsymbol{c}$.

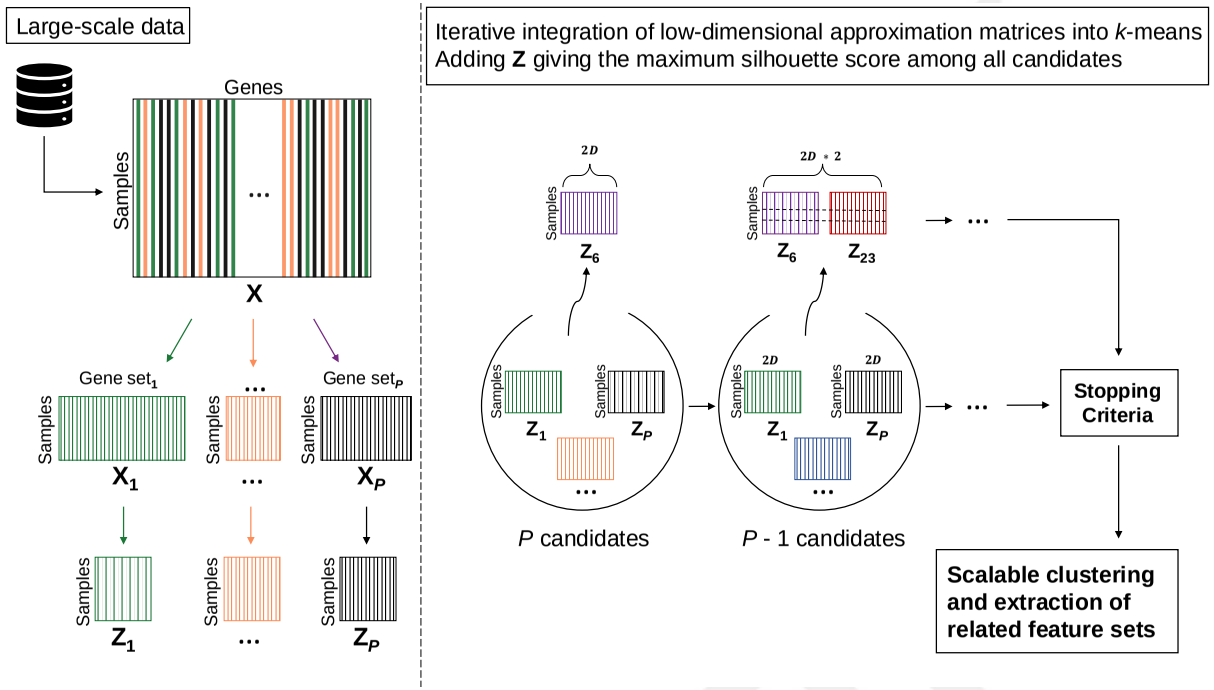


Figure 4.1: Our proposed forward multiple approximate kernel k-means clustering framework that integrates kernel approximation and k-means formulations into a conjoint model. It takes an input matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$, a gene set collection consisting of P gene sets, and a number of gene sets to stop p. During each iteration, a silhouette score is calculated from k-means clustering results, to select the next low-dimensional approximation matrix to combine with the one from the previous iteration. It is scalable since it integrates a low-dimensional kernel matrix approximation instead of usual kernel matrix computation; it is interpretable since it provides relevant gene set information to the clustering task by tracking the maximum silhouette score.

Chapter 5

EXPERIMENTAL RESULTS AND DISCUSSION

5.1 Experimental Results of Pathway/Gene Set-Based Prediction Using Multiple Kernel Learning

To test our proposed algorithm, namely, PrognosiT, we performed computational experiments on TCGA thyroid carcinoma data set, which contains the volume information of tumors. We compared PrognosiT against two baseline algorithms that are widely used for genomic data analysis, namely, RF and SVR.

5.1.1 Experimental settings

We split the data set into two parts, 80% of the tumors formed the training set and the remaining 20% of the tumors formed the test set. We normalized each feature in the training set to have zero mean and unit standard deviation. For the test set, we normalized each feature with the mean and standard deviation calculated on the original training set. We applied cube root transformation to the tumor volume values since it is the multiplication of three dimensions. We performed 100 replications of our analysis to have more robust results, and we reported the results of these 100 replications. The hyper-parameters for RF (i.e., number of trees to grow), SVR (i.e., regularization parameter C and tube width multiplier) and MKL (i.e., regularization parameter C and tube width multiplier) were selected by using a four-fold cross-validation strategy on the training set.

We used `randomForestSRC` R package version 2.9.3 [Ishwaran & Kogalur, 2020] for RF experiments. We chose the number of trees to grow, ntree, from the set $\{500, 1000, \ldots, 2500\}$ by the aforementioned cross-validation approach.

For SVR and our proposed MKL algorithm PrognosiT, we built our own imple-

mentations in R, which use CPLEX version 12.6.3 for solving quadratic optimization problems [1]. We used Gaussian kernel in our algorithm to form a similarity measure between gene expression profiles of primary thyroid tumors. As a recall, the Gaussian kernel is

$$k_{\mathcal{G}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\langle \boldsymbol{x}_i - \boldsymbol{x}_j, \boldsymbol{x}_i - \boldsymbol{x}_j \rangle}{2\sigma^2}\right),$$

where σ is the kernel width parameter, and we set it to the mean pairwise Euclidean distances between training samples. For both algorithms, the tube width multiplier parameter, which is to be multiplied by the standard deviation of the current training samples to form the tube width, is chosen from the set $\{0, 0.25, 0.50, \ldots, 2\}$, whereas the regularization parameter C is chosen from the set $\{10^{-3}, 10^{-2}, \ldots, 10^{+3}\}$ using the previously described four-fold inner cross-validation strategy.

In our implementation, we chose Gaussian kernel to discover the highly nonlinear dependency between the tumor progression and gene expression profiles. We calculated the Gaussian kernel matrices on subsets of tumor gene expression profiles by examining the pathway and gene set content and selecting the corresponding kernel widths. For PrognosiT algorithm, knowing that the algorithm converges in the order of tens of iterations, we performed 200 iterations to guarantee the convergence.

5.1.2 Performance metric

We used a form of normalized root mean squared error (i.e., NRMSE) for comparison of prediction performances of the three algorithms, namely, RF, SVR and our proposed MKL algorithm PrognosiT. NRMSE can be calculated as

$$\text{NRMSE} = \sqrt{\frac{(\boldsymbol{y} - \widehat{\boldsymbol{y}})^\top (\boldsymbol{y} - \widehat{\boldsymbol{y}})}{(\boldsymbol{y} - \mathbf{1}\bar{y})^\top (\boldsymbol{y} - \mathbf{1}\bar{y})}}, \tag{5.1}$$

where $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}$ stand for the vectors of observed and predicted tumor volumes, respectively, and $\bar{y}$ denotes the mean of $\boldsymbol{y}$. Note that smaller NRMSE values correspond to better predictive performance, and if the value is less than one, it means that the model is capable of learning from the data set.

5.1.3 Predictive performance of the proposed algorithm

We compared three machine learning algorithms in our experiments: Random forest (denoted as RF), support vector regression (denoted as SVR) and our proposed algorithm PrognosiT that integrates multiple kernel learning (denoted as MKL). For RF and SVR algorithms, we provided all the available gene expression features (i.e., 19 814 features in total) as the input data. For PrognosiT algorithm, MKL[H] and MKL[P] used the `Hallmark` and `PID` pathway/gene set collections, respectively.

Figure 5.1 displays the predictive performances of RF, SVR, MKL[H] and MKL[P] algorithms on `THCA` data set for tumor volume prediction problem by using gene expression profiles as the input. The box-and-whisker plots compares the NRMSE values of the four algorithms resulted from 100 random training/test splits. RF algorithm is compared against SVR, MKL[H] and MKL[P] algorithms using a two-tailed paired t-test to check whether there is a significant difference between their performances whilst SVR is compared against MKL[H] and MKL[P] algorithms. The NRMSE values of these algorithms resulted from each replication for `THCA` data set is available in Supplementary Table 1 of Bektaş and Gönen (2021)[1].

We observed that SVR algorithm outperformed RF in thyroid carcinoma data set. Due to having a highly nonlinear kernel (i.e., the Gaussian kernel) integrated in its implementation, SVR performed better than RF in this prediction task. Our algorithm, PrognosiT, which is an extension of SVR algorithm and integrates our prior knowledge about pathways/gene sets into the machine learning model with a multiple kernel learning formulation, outperformed both RF and SVR.

MKL[H] and MKL[P] algorithms picked significantly fewer gene expression features than RF and SVR algorithms and eliminated uninformative pathways/gene sets from the machine learning model thus, identified relevant pathways/gene sets for tumor volume prediction. RF and SVR algorithms benefited from all available gene expression features (i.e., 19 814 features in total) whereas the average numbers of

[1]Supplementary Table 1 can be found in `Supplementary Information` section at `https://doi.org/10.1186/s12859-021-04460-6`.

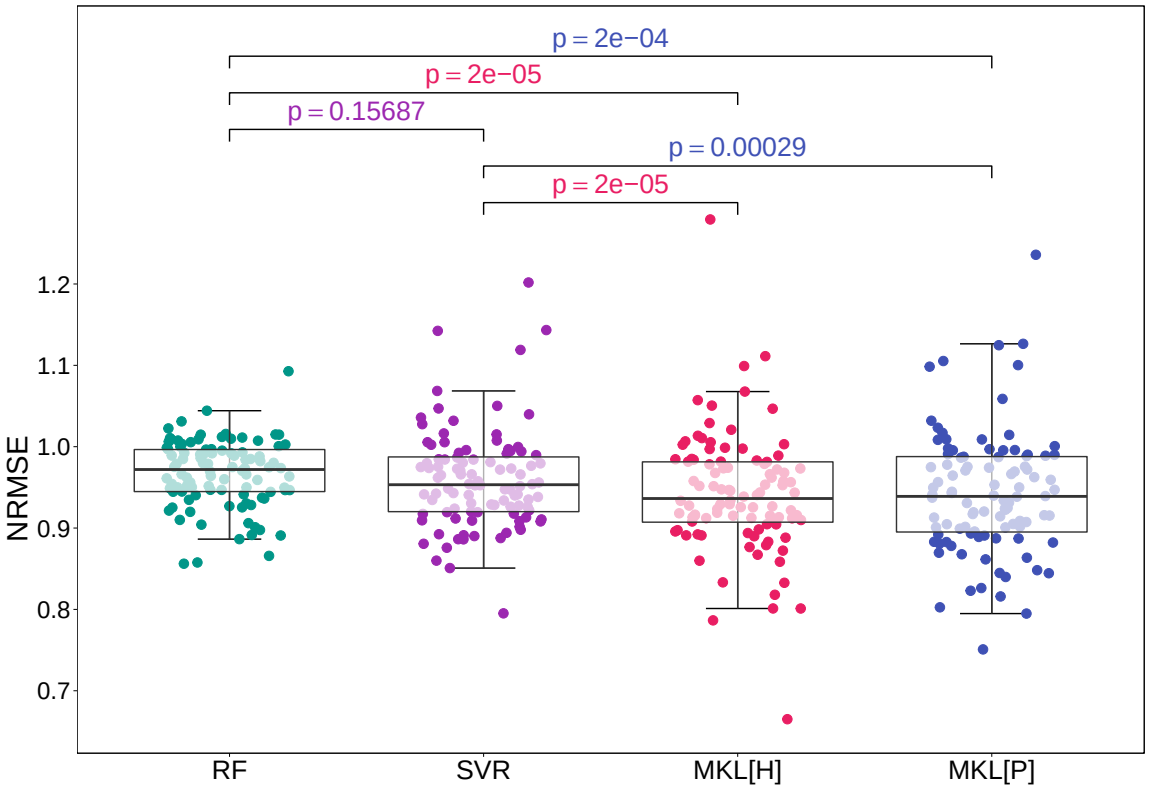


Figure 5.1: The predictive performances of RF algorithm, SVR algorithm, PrognosiT algorithm with the `Hallmark` gene set collection (MKL[H]) and PrognosiT algorithm with the Pathway Interaction Database pathway collection (MKL[P]) on thyroid carcinoma (`THCA`) data set. Comparison of the NRMSE values resulted from 100 replications for each of the four algorithms is made using the box-and-whisker plots. RF is compared against SVR, MKL[H] and MKL[P] algorithms using the two-tailed paired t-test to check whether there is a significant difference between their performances. In addition, SVR is compared against MKL[H] and MKL[P]. The resulting p-values are reported. The color of p-values matches with the color of the winning algorithm, for each pairwise comparison.

used gene expression features by MKL[H] and MKL[P] algorithms were respectively 1782 and 797. The exact numbers of gene expression features that are used by these algorithms in each replication are available in Supplementary Table 2 of Bektaş and Gönen (2021)[2].

5.1.4 Informative pathways/gene sets for tumor progression

In addition to comparing the predictive performances of MKL[H] and MKL[P] algorithms to RF and SVR algorithms, we investigated the pathways/gene sets chosen by our proposed PrognosiT algorithm on thyroid carcinoma data set. Supplementary Table 3 of Bektaş and Gönen (2021)[3] displays the selection frequencies of 50 gene sets in the `Hallmark` collection for 100 replications. We assumed that a pathway/gene set was added in the final model if the corresponding kernel weight was greater than 0.01. The exact kernel weights assigned to 50 gene sets in 100 replications for thyroid carcinoma cohort is displayed in Supplementary Table 4 of Bektaş and Gönen (2021)[4]. Moreover, we also reported the selection frequencies of 196 pathways in the `PID` collection for 100 replications in Supplementary Table 5 of Bektaş and Gönen (2021)[5] and the exact kernel weights assigned to these pathways in 100 replications in Supplementary Table 6 of Bektaş and Gönen (2021)[6]. The selection frequencies in MKL[H] algorithm averaged to 28.5 gene sets, whereas those of MKL[P] algorithm averaged to 13.4 pathways.

We checked the column sums of the selection frequencies of the pathways/gene

---

[2]Supplementary Table 2 can be found in `Supplementary Information` section at `https://doi.org/10.1186/s12859-021-04460-6`.

[3]Supplementary Table 3 can be found in `Supplementary Information` section at `https://doi.org/10.1186/s12859-021-04460-6`.

[4]Supplementary Table 4 can be found in `Supplementary Information` section at `https://doi.org/10.1186/s12859-021-04460-6`.

[5]Supplementary Table 5 can be found in `Supplementary Information` section at `https://doi.org/10.1186/s12859-021-04460-6`.

[6]Supplementary Table 6 can be found in `Supplementary Information` section at `https://doi.org/10.1186/s12859-021-04460-6`.

sets for the Hallmark and PID collections shown in Supplementary Tables 3 and 5 of Bektaş and Gönen (2021) to discover informative and uninformative pathways/gene sets for tumor volume prediction and showed them in Table A.1 and A.2. In the final model, MKL[H] algorithm chose HYPOXIA gene set in 99 replications over 100. Hypoxia is known as one of the most important signatures of solid tumors and is related to radiotherapy and chemotherapy resistance, which leads to poor clinical prognosis [Ruan et al., 2009]. We know that thyroid cancer is mostly an ERK-driven malignity and mutations that activate the RAS/ERK mitogenic signaling pathway are responsible for up to 70% of thyroid carcinomas [Zaballos et al., 2019]. Thus, having high selection frequencies for KRAS signaling gene sets in the final model shows that our predictive model is in agreement with the literature. Another gene set that our model selected frequently over 100 replications is GLYCOLYSIS. It is known that a near-universal property of primary and metastatic cancers is up-regulation of glycolysis, leading to increased glucose consumption [Gatenby & Gillies, 2004].

One other highly selected pathway that attracted our attention in our final model is P53_PATHWAY. It is well known that there exists a complex network among p53 family members and interactions of these members with other elements accelerates thyroid cancer progression [Manzella et al., 2017]. High selection frequency over 100 replications by our final model of ANGIOGENESIS, INFLAMMATORY_RESPONSE and EPITHELIAL_MESENCHYMAL_TRANSITION gene sets, which are highly associated with tumor progression and rapid changes in cellular phenotype, and frequent selection of the metabolism-related gene sets such as PANCREAS_BETA_CELLS, XENOBIOTIC_METABOLISM, FATTY_ACID_METABOLISM, BILE_ACID_METABOLISM show that the results of our proposed algorithm are consistent with the existing knowledge and may contribute to the discovery of unknown mechanisms towards tumor progression.

5.1.5 Significantly up- and down-regulated genes between tumor and normal tissues

After comparing predictive performances and analyzing highly selected pathways/gene sets to check the biological mechanisms that lead to tumor progression, we also analyzed the genes that have been selected frequently by our final model over 100

replications. We noted the genes that are selected in every replication separately by our MKL[H] and MKL[P] algorithms. There were 137 genes resulted from MKL[H] algorithm and 37 genes resulted from MKL[P] algorithm. We then checked whether these genes are significantly up- or down-regulated during the tumor progression by comparing the gene expressions of tumor tissues to the gene expressions of normal tissues collected from the same patients. There were 58 cancer patients that had both tumor and normal tissue information. We performed paired Wilcoxon test, which is a nonparametric statistical test that compares two paired groups, to determine whether there is a significant difference (i.e., p-value < 0.05) between the gene expressions coming from tumor and normal tissues. As a result, there were 113 genes that are significantly up- or down-regulated resulted by MKL[H] and 32 genes resulted by MKL[P].

We showed the scatter plot visualizations of some of the significantly differed expressions of the genes resulted from our MKL[H] and MKL[P] algorithms (see Figure 5.2). We checked whether these genes are used as prognostic factor for thyroid cancer evaluation from The Human Protein Atlas website at `https://www.proteinatlas.org`. The displayed genes in Figure 5.2 are already in use as prognostic factor for thyroid cancer. This situation contributes to the possibility that the remaining genes resulted from our algorithm to predict tumor volume might be considered as prognostic factors in the future as well.

5.2 Experimental Results of Multiple Approximate Kernel Learning Algorithm

To test our algorithm, we performed computational experiments for three binary classification tasks. To better compare the outcomes of the experiments, we benchmarked all algorithms over 100 independent replications. As the predictive performance metric for these binary classification tasks, we used area under the receiver operating characteristics curve (AUROC), whose larger values correspond to a better classification performance.

Owing to the serious scalability issues presented at earlier sections, we did not

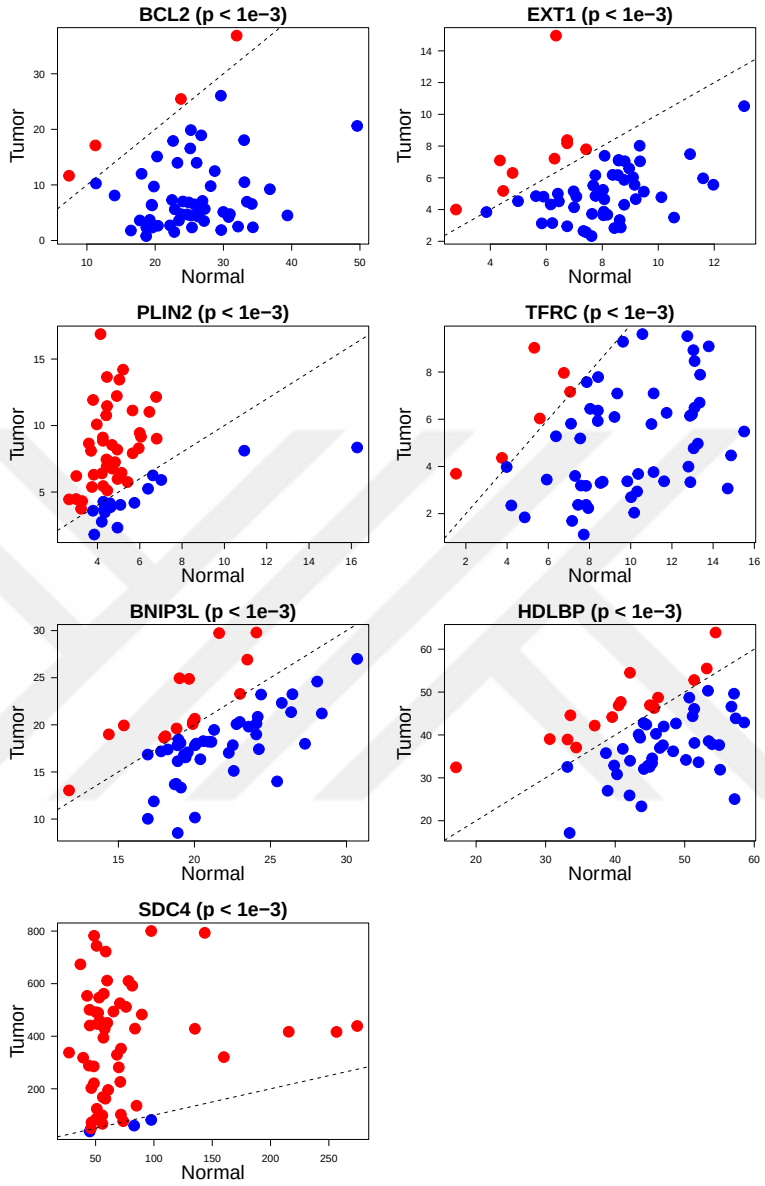


Figure 5.2: Comparison of mRNA gene expression levels of several genes resulted by our algorithms MKL[H] and MKL[P] on thyroid carcinoma data set. Corresponding gene up- and down-regulations can be understood by scatter plot visualizations of `Tumor` and `Normal` tissues. Red points refers to overexpression of the specified gene in tumor whereas blue points refer to underexpression in tumor. The p-values are resulted from paired Wilcoxon test that we used to determine whether there is a significant difference between the gene expressions of `Tumor` and `Normal` tissues.

choose standard MKL algorithms for baseline comparison. We chose extreme gradient boosting (XGBoost) [Chen & Guestrin, 2016] as the baseline algorithm to compare against our MAKL algorithm. Extreme gradient boosting is an ensemble of weak prediction models and, by its nature, its predictive performance is generally better compared to the random forest algorithm. It is also known as being scalable.

For all experiments, we split the corresponding data set into two parts, in a way that 80% of the samples were included into the training set and the remaining 20% of the samples were included into the test set. We normalized each feature in the training set to have zero mean and unit standard deviation. For the test set, we normalized each feature with the mean and standard deviation calculated on the original training set. We performed four-fold inner cross validation on the training set for hyper-parameter tuning. We used two aforementioned pathway/gene set collections.

For XGBoost algorithm that we used as the baseline algorithm, the hyper-parameter, maximum depth of a tree, was chosen using four-fold inner cross validation on the training set with a maximum number of iterations of 1 000 and learning rate of 0.2. Since, in this paper, we cover binary classification tasks, we used "binary:logistic" option as the objective function argument to the algorithm. For XGBoost experiments, we used `xgboost` R package [Chen & He, 2019].

The hyper-parameter related to the regularization, λ, for MAKL was chosen using four-fold cross validation on the training set. The λ parameter set used for cross validation was $\{0.9, 0.8, 0.7, 0.6\}$ to get fast and sparse solutions, since parameter values closer to 1 leads to sparse solutions.

The dimensionality parameter D (i.e., the number of random features used to approximate distinct kernel matrices) was given as an input to the MAKL algorithm. We ran experiments using different D values $\{100, 150, 200, 250\}$ to perform sensitivity analysis on MAKL. For MAKL experiments, we used `grplasso` R package with default parameters [Meier, 2020].

5.2.1 Early- and late-stage cancer classification

For this binary classification task, we performed sensitivity analysis for the AUROC values using different D values for MAKL with `Hallmark` gene sets (i.e., different numbers of randomly chosen Fourier samples). We compared MAKL algorithm against the baseline algorithms, XGBoost (Subset) and XGBoost (Full). XGBoost (Subset) denotes the algorithm trained using the genes from the `Hallmark` gene set collection, whereas XGBoost (Full) denotes the algorithm trained using all available 19 814 genes.

Figure 5.3 shows how the classification performance changes with respect to the number of feature sets (i.e., pathways/gene sets) used. It also shows that MAKL with `Hallmark` gene set collection outperforms XGBoost (Subset) using only a small fraction of the available features. Considering the mean number of pathways used for classification, MAKL outperforms the baseline algorithm XGBoost (Subset) using only 18.54% (i.e., 9.27 out of 50) of the total feature sets (i.e., gene sets) with $D = 150$. It also outperformed XGBoost (Full) while using only 12.47% of the all available 19 814 genes on the average. Additionally, Figure 5.4 displays the early- and late stage cancer classification sensitivity analysis for the AUROC values using different D values for MAKL with `PID` pathway collection and compares MAKL against the baseline algorithms, namely, XGBoost (Subset) and XGBoost (Full).

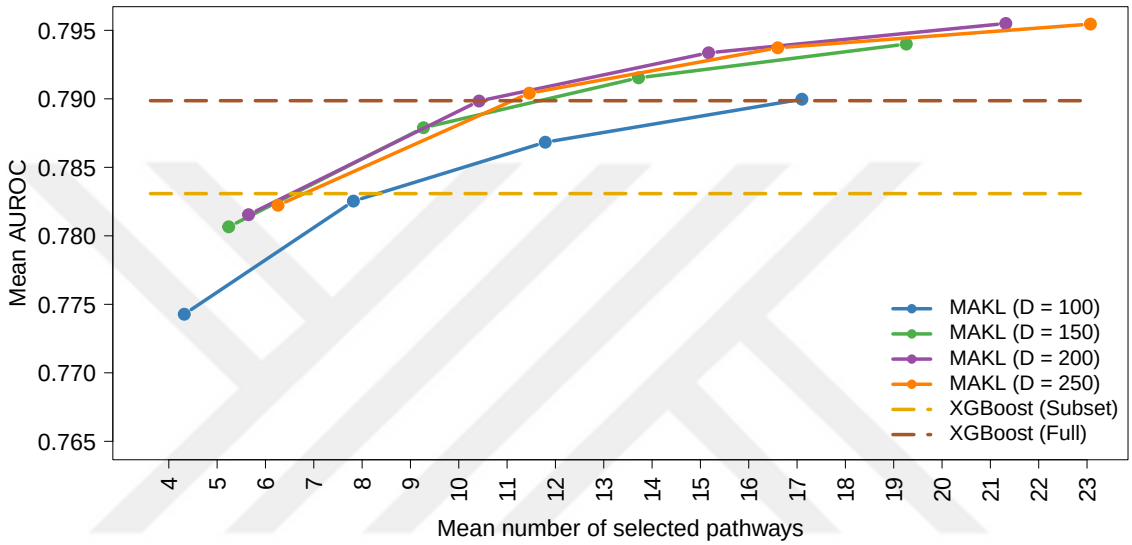


Figure 5.3: Sensitivity analysis on the early- and late-stage cancer classification performance of MAKL using four different D values (i.e., different numbers of randomly chosen Fourier samples), and their comparison to the baseline algorithms, XGBoost (Subset) and XGBoost (Full). The figure depicts the AUROC values resulted from 100 replications. As feature sets, the `Hallmark` gene set collection is used. XGBoost (Subset) denotes that the algorithm trained using the genes from the `Hallmark` gene set collection, and XGBoost (Full) denotes that the algorithm trained using all available 19 814 genes. Note that the results reported for XGBoost (Subset) and XGBoost (Full) are not affected from the values in the x-axis, and they are reported as dashed lines for easy comparison.

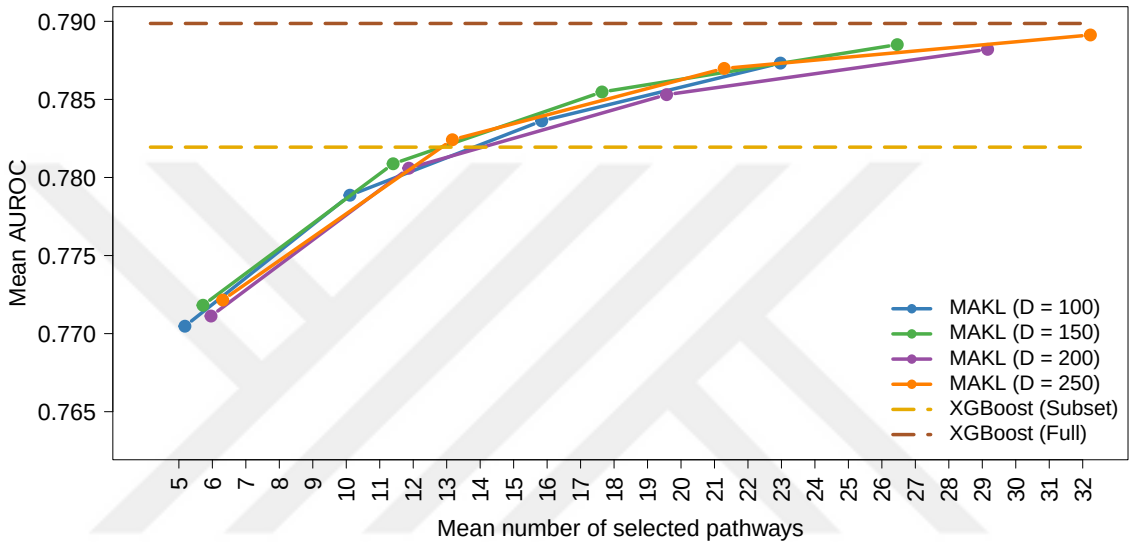


Figure 5.4: Sensitivity analysis on the early- and late-stage cancer classification performance of MAKL using four different D values (i.e., different numbers of randomly chosen Fourier samples), and their comparison to the baseline algorithms, XGBoost (Subset) and XGBoost (Full). The figure depicts the AUROC values resulted from 100 replications. As feature sets, the Pathway Interaction Database (`PID`) pathway collection is used. XGBoost (Subset) denotes that the algorithm trained using the genes from the `PID` pathway collection, and XGBoost (Full) denotes that the algorithm trained using all available 19 814 genes. Note that the results reported for XGBoost (Subset) and XGBoost (Full) are not affected from the values in the x-axis, and they are reported as dashed lines for easy comparison.

5.2.2 Two-year survival classification

For two-year survival classification task, MAKL with `PID` pathway collection was able to outperform the baseline algorithms XGBoost (Subset) using only 5.96% (i.e., 11.69 out of 196) of the total feature sets (i.e., pathways) with $D = 150$. It also outperformed XGBoost (Full) while using only 4.15% of all available 19 814 genes on the average.

Figure 5.5 shows sensitivity analysis for the AUROC values using four different D values for MAKL with `PID` pathway collection (i.e., different numbers of randomly chosen Fourier samples), and compares them against the baseline algorithms, XGBoost (Subset) and XGBoost (Full). XGBoost (Subset) denotes the algorithm trained using the genes from the Pathway Interaction Database (`PID`) pathway collection, and XGBoost (Full) denotes the algorithm trained using all available 19 814 genes. Additionally, Figure 5.6 shows the sensitivity analysis for the AUROC values using four different D values of MAKL with `Hallmark` gene sets and compares the results with the baseline algorithms, XGBoost (Subset) and XGBoost (Full). XGBoost (Subset) denotes the algorithm trained using the genes from the `Hallmark` gene set collection, and XGBoost (Full) denotes the algorithm trained using all available 19 814 genes.

5.2.3 Single-cell melanoma immunotherapy data set

As mentioned earlier, our algorithm is capable of extracting meaningful information while learning a predictive model. The weights associated with each pathway/gene set resulting from MAKL shows the relevance of each pathway in the given classification task. Owing to the computational preprocessing and labeling techniques used in single-cell data set that we used (i.e., the labeling is not done manually and by observation like in patient labeling), we concentrated on the pathways/gene sets that play crucial role in cell malignancy classification task, rather than the classification performance. We demonstrated that the pathways/gene sets that our algorithm MAKL found important for this classification task are relevant in terms of finding new immunotherapy techniques, new ways to discover tumor heterogeneity since the pathways/gene sets extracted from our algorithm are in line with the most

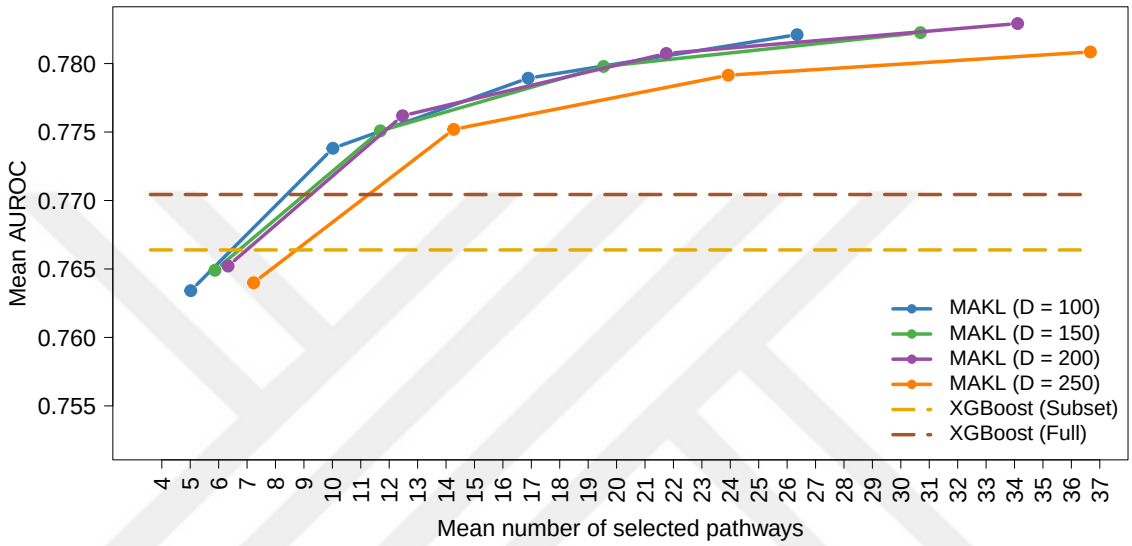


Figure 5.5: Sensitivity analysis on the two-year survival classification performance of MAKL using four different D values (i.e., different numbers of randomly chosen Fourier samples), and their comparison to the baseline algorithms, XGBoost (Subset) and XGBoost (Full). The figure depicts the AUROC values resulted from 100 replications. As feature sets, the Pathway Interaction Database (`PID`) pathway collection is used. XGBoost (Subset) denotes that the algorithm trained using the genes from the Pathway Interaction Database (`PID`) pathway collection, and XGBoost (Full) denotes that the algorithm trained using all available 19 814 genes. Note that the results reported for XGBoost (Subset) and XGBoost (Full) are not affected from the values in the x-axis, and they are reported as dashed lines for easy comparison.

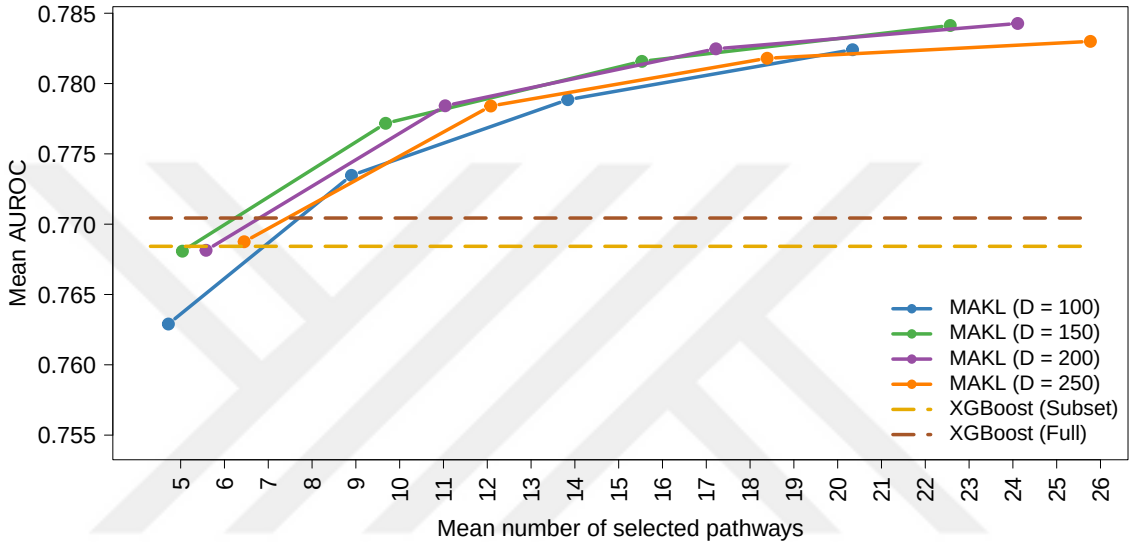


Figure 5.6: Sensitivity analysis on the two-year survival classification performance of MAKL using four different D values (i.e., different numbers of randomly chosen Fourier samples), and their comparison to the baseline algorithms, XGBoost (Subset) and XGBoost (Full). The figure depicts the AUROC values resulted from 100 replications. As feature sets, `Hallmark` gene set collection is used. XGBoost (Subset) denotes that the algorithm trained using the genes from the `Hallmark` gene set collection, and XGBoost (Full) denotes that the algorithm trained using all available 19 814 genes. Note that the results reported for XGBoost (Subset) and XGBoost (Full) are not affected from the values in the x-axis, and they are reported as dashed lines for easy comparison.

recent melanoma treatment trials, which are supported with the literature findings as detailed below.

Table C.1 displays the selection frequencies of 50 gene sets in the `Hallmark` gene set collection for over 100 replications, using four different regularization parameter (i.e., λ multiplier) values. If norm of weights associated with a feature set, resulting from MAKL is positive, we considered the corresponding feature set (i.e., pathway/gene set) as selected. Also, Table C.2 displays the selection frequencies of 196 pathways in the `PID` pathway collection for 100 replications, using four different regularization parameter (i.e., λ multiplier) values. According to the average selection frequencies over 100 replications, the most selected pathway by MAKL with `PID` pathway collection was `CXCR4` pathway. This result is relevant since targeting of `CXCR4` pathway is under consideration in melanoma treatment trials. Moreover, Table C.2 shows that, even using four pathways, MAKL is capable of finding this relevant pathway for cell malignancy classification. In other words, our algorithm can find the relevant feature sets towards a classification task using a small fraction of all the feature sets (e.g., 2.26 out of 196).

It is stated in literature that targeting melanoma by `CXCR4` inhibition may be a good way to destroy the tumor cells and that `CXCR4` regulates tumor immunity [Yang et al., 2018]. Additionally, the chemokine receptor `CXCR4` is stated to be associated with cancer growth, invasion and metastasis, and identified as an independent predictor of poor prognosis in primary melanoma [Scala et al., 2006]. The three most frequently chosen pathways resulted from MAKL with `PID` pathway collection over 100 independent replications are `CXCR4`, `SYNDECAN_4` and `CASPASE`, respectively. Likewise, the three most frequently chosen pathways resulted from MAKL with `Hallmark` gene set collection over 100 independent replications are `ALLOGRAFT_REJECTION`, `INTERFERON_ALPHA_RESPONSE` and `KRAS_SIGNALING_UP`, respectively. Further related details are given in Tables C.1–C.2.

5.3 Experimental Results of the Proposed Clustering Algorithm

To test our scalable multiple approximate kernel clustering framework, we clustered the RNA-Seq data from TCGA consortium. We combined data coming from 15 cohorts that has at least 300 patients within. To provide sparsity on gene set level and to find the gene sets most relevant to the clustering task, we chose the maximum number of gene sets that the algorithm is allowed to choose as 5 and the number of random features D to approximate the exact kernel matrices as 100. It is previously shown that increased number of random features for a machine learning task may give better approximations of kernel matrices while a better learning performance is not guaranteed [Bektaş et al., 2022].

As mentioned in earlier sections, after having calculated the Euclidean distance matrices using only S rows of the raw data where $S \ll N$, we calculated the corresponding kernel matrix $\mathbf{K}$ and used $2 - 2\mathbf{K}$ as the distance matrix in our silhouette coefficient calculations. We then calculated the mean of S silhouette coefficients and reported this mean as the mean silhouette score. The mean silhouette score of our approach is 0.279 while using information from only four relevant gene sets. We compared this result to a baseline algorithm where we applied k-means to raw data that contains RNA-Seq information on 19 814 genes of 7 474 patients. The mean silhouette score resulting from this baseline experiment is 0.179 which is approximately 64.2% of the result from our proposed algorithm.

5.3.1 Clustering in terms of cancer type

To interpret results from the clustering experiment, we analyzed the clustered RNA-Seq data together with the corresponding cancer tissue types. For the sake of visualization, we concentrated specifically in two cases. The first case is the breast cancer case where majority of triple negative breast cancer (`TNBC`) patients are separated clearly from other breast cancer patients. `TNBC` is a subtype of breast cancer that does not express estrogen receptor (`ER`), progesterone receptor (`PR`) and human epidermal growth factor receptor 2 (`HER2`). It is usually characterised by poor prognosis. These

tumors are insensitive to hormonal therapy and/or HER2 treatment, and there is a need for standardised treatment regimens [Yin et al., 2020]. The second case is the cluster of ovarian cancer (OV) and uterine corpus endometrial carcinoma (UCEC), which are gynaecological cancers. For both cases, we prepared heat maps showing the RNA-Seq data of the gene sets that are resulted as "relevant" from our algorithm. The four relevant gene sets that are resulted from our algorithm are ESTROGEN_RESPONSE_EARLY, ESTROGEN_RESPONSE_LATE, HEDGEHOG_SIGNALING and XENOBIOTIC_METABOLISM. It is not surprising that the first two selected gene sets to play a crucial role in the clustering are estrogen response related gene sets since 27% of the data that is clustered belongs to estrogen-dependent cancers which are BRCA, OV and UCEC. As to other selected gene sets, HEDGEHOG_SIGNALING plays a crucial role in embryonic development and tissue homeostasis. Its disregulation is demonstrated to be associated with neoplastic transformations, malignant tumors, and drug resistance of a multitude of cancers [Sari et al., 2018]. The fourth relevant gene set, XENOBIOTIC_METABOLISM is associated with the elimination of outside agents to the body. It is demonstrated to be linked to cancer. For instance, cytochrome P450 enzymes (P450s), that are noted as bridges between the body and the environment are noted as important targets in cancer due to their role in xenobiotic metabolism [Tamási et al., 2011].

For the heat maps of RNA-Seq data, we ordered the genes using Spearman's rank-order correlation between the gene expression values and the clustering results for better visualization. We annotated the heat maps with the existence of several genes that involve in cell growth and proliferation, namely KRAS, HRAS, NRAS, BRAF, PTEN, PIK3CA, EGFR and the existence of TP53 gene that is related to stress response and apoptosis. By combining information of missense mutation existence on these genes with the clustering results on the heat maps; we aimed to see whether between the clusters, there exist a significant difference in terms of the existence of missense mutations on a gene or not. We used Fisher's exact test for this aim. We reported the genes that have a p-value smaller than 0.001 on the heat maps.

Figure 5.7 shows the heat map using the RNA-Seq data corresponding to genes from ESTROGEN_RESPONSE_EARLY gene set which is the gene set selected in the first

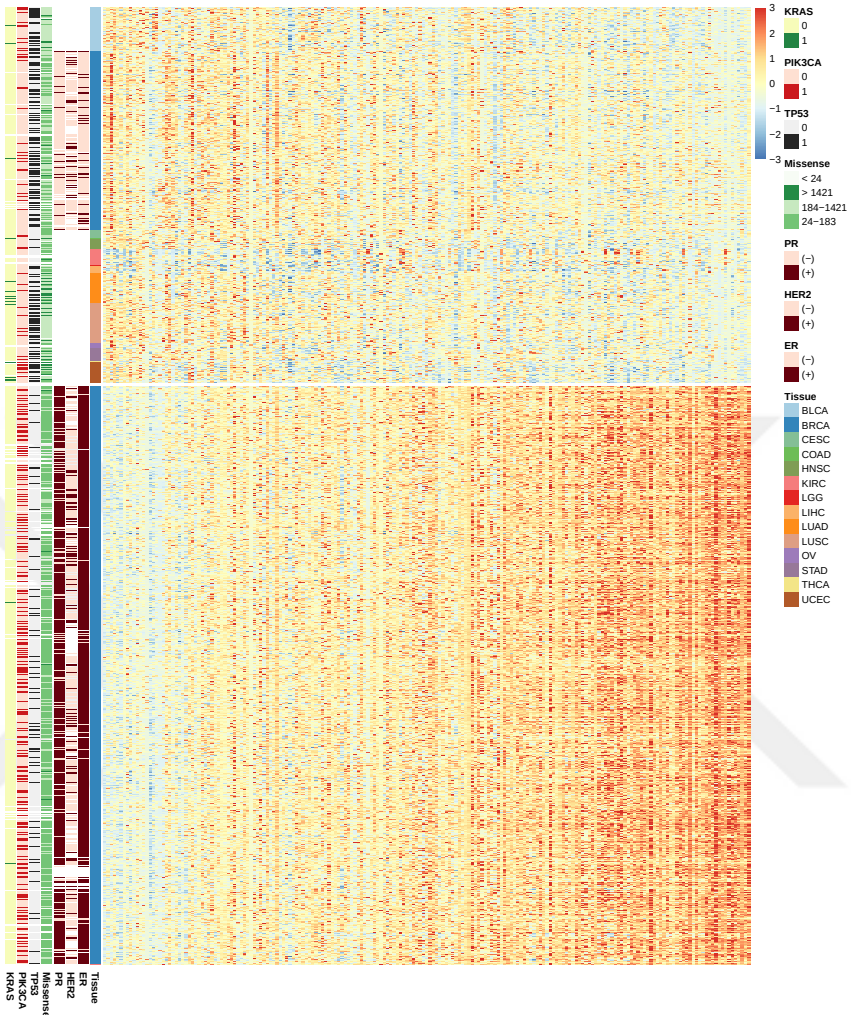


Figure 5.7: Heat map showing patient-based RNA-Seq data corresponding to genes from `Hallmark ESTROGEN_RESPONSE_EARLY` gene set, zoomed to clusters that separate the `TNBC` patients from other breast cancer patients; column ordered by Spearman's correlation between the clustering results and the gene expression values, for better visualization. The heat map is annotated with cancer tissue type; `ER` positivity, `PR` positivity and `HER2` positivity; scaled total number of missense mutations; existence of missense mutations on `TP53`, `PIK3CA`, `KRAS` genes sorted by their ascending p-values from right to left, indicating the significance of the difference between the clusters in terms of the existence of the mutations on the annotated genes on 0.001 significance level.

iteration of our algorithm as relevant. This is a meaningful result since breast cancer is an estrogen-dependent cancer. The heat map is annotated with clustering results, total number of missense mutations and the existence of missense mutations on several cancer-related genes. After an initial investigation of cluster memberships resulting from our approach, we noticed that there was a well separation between two subtypes of breast cancer. We then focused on these clusters to be able to get a more detailed view. While analyzing the heat map of `ESTROGEN_RESPONSE_EARLY`; we checked `ER` positivity, `PR` positivity and `HER2` positivity on these clusters and saw that the two clusters were separating `TNBC` patients from other breast cancer patients.

Figure 5.8 shows the heat map using the RNA-Seq data corresponding to genes from `ESTROGEN_RESPONSE_LATE` gene set which is selected as relevant by our algorithm. This relevance is supported by the existing literature since estrogen signalling is known to represent a potential target for therapy in ovarian cancer [Langdon et al., 2020]. The heat map is annotated with clustering results, total number of missense mutations, the existence of missense mutation on several cancer-related genes and the histological grade associated with each patient. It is worth noting that in primary endometrial cancer lesions, `PIK3CA` is reported to be the second most frequently significant mutated gene after `PTEN` with a frequency of 53% and noted as an indicator of poor prognosis. It is shown in the figure that the existence of mutations on both genes are significantly higher on the above cluster, where the corresponding histological grades indicate poor prognosis. Thus, it may be commented that the above cluster separates from the below cluster in terms of cancer aggressivenes.

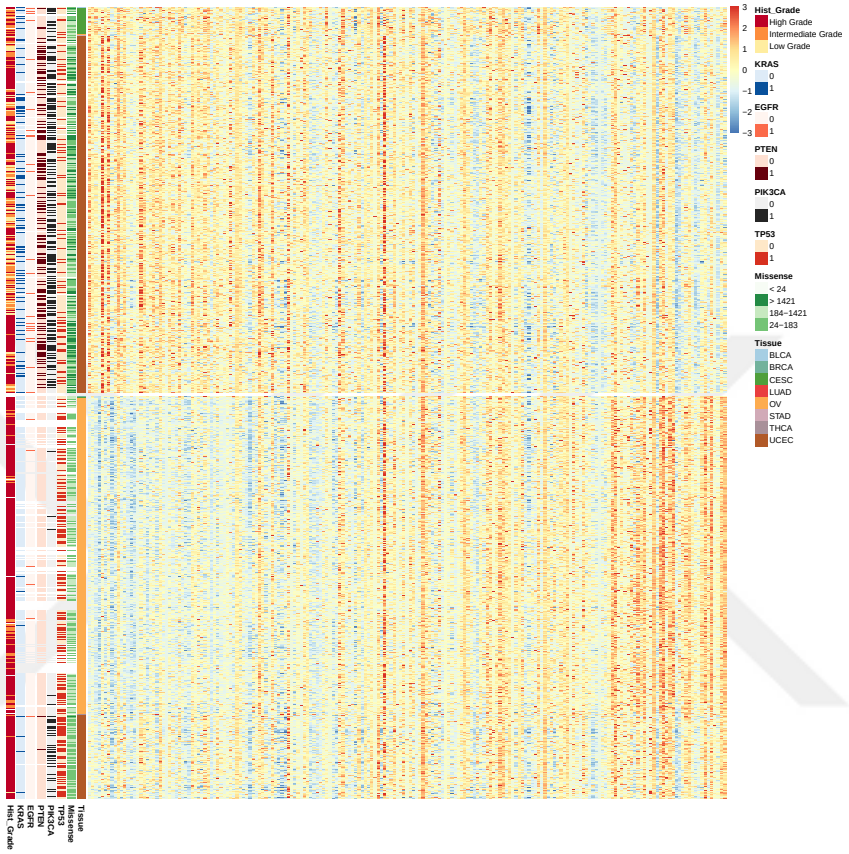


Figure 5.8: Heat map showing the patient-based RNA-Seq data corresponding to genes from `Hallmark ESTROGEN_RESPONSE_LATE` gene set, zoomed to clusters that separates ovarian cancer (`OV`) and uterine corpus endometrial carcinoma (`UCEC`) patients; column ordered by Spearman's correlation between clustering results and the gene expression values, for better visualization. The heat map is annotated with cancer tissue type; scaled total number of missense mutations; existence of missense mutations on `TP53`, `PIK3CA`, `PTEN`, `EGFR`, `KRAS` genes sorted by their ascending p-values from right to left, indicating the significance of the difference between the clusters in terms of the existence of the mutations on the annotated genes on significance 0.001 level; histological grade of tumors.

5.3.2 Visualization and analysis of clustering results using t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a technique visualizing high-dimensional data by locating each data instance to a two or three-dimensional plane [van der Maaten & Hinton, 2008]. We used t-SNE to visualize the clusters identified from the low-dimensional approximation matrix selected by our approach in the first iteration (see Figure 5.9). We cross-checked the results of our clustering framework with t-SNE and saw that, as had been provided by our proposed approach, `TNBC` patients are separated from other breast cancer patients; `OV` and `UCEC` patients are clustered together. It is intuitive that `OV` and `UCEC` cancers that are both gynaecologic cancers share similarities in their genomic profiling. It has also been reported that both ovarian and endometrial cancer may be seen together as part of the same autosomal dominantly inherited syndrome, namely, Lynch syndrome [Gayther & Pharoah, 2010]. Another notable result revealed by our clustering approach and t-SNE is that the patients with colon adenocarcinoma (`COAD`) and stomach adenocarcinoma (`STAD`) are clustered together. This is also meaningful result since these two cancer types are both gastrointestinal cancers. It has previously been reported that gastric cancer patients are reported to be at high risk of developing synchronous cancer; the most common one being colorectal cancer [Lee et al., 2006]. Additionally, as a common finding from both our approach and t-SNE, patients with head and neck cancer (`HNSC`) and lung squamous cell carcinoma (`LUSC`) are clustered together which suggest they may share similar genomic alterations. It is reported by Polo et al. (2016) that squamous cell carcinomas of lung and head and neck region share strong association with smoking habits and are characterised by smoke-related genetic changes.

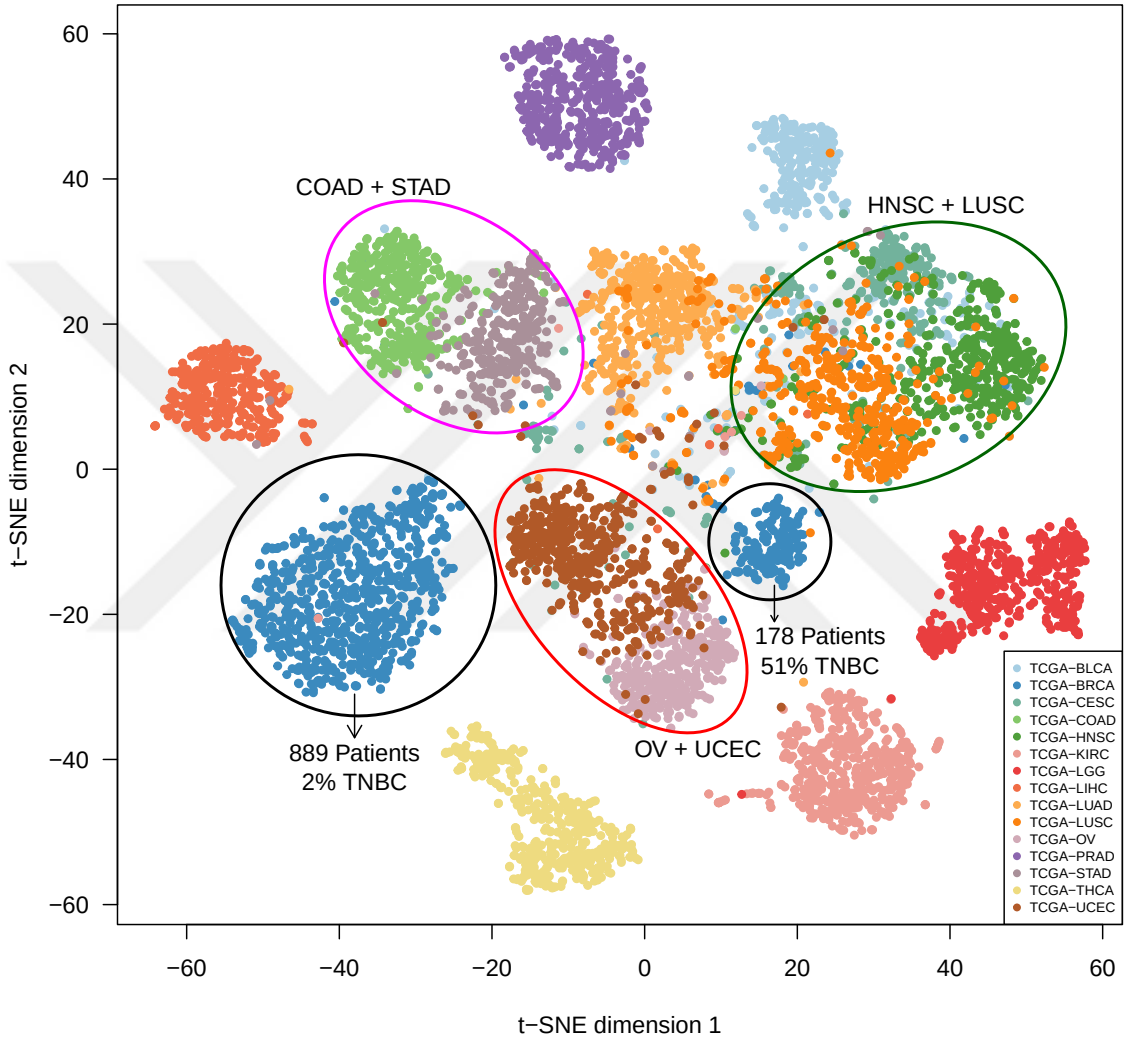


Figure 5.9: *t*-SNE plot of the low-dimensional approximation matrix, associated to the gene set selected as the most informative in the first iteration, namely `ESTROGEN_RESPONSE_EARLY`. The data instances are colored according to their corresponding cancer tissue types. For breast cancer patients, number of patients in both clusters and the percentage of the number of `TNBC` patients are reported using the TCGA clinical receptor data.

Chapter 6

CONCLUSIONS

Predicting tumor volume while discovering the underlying molecular mechanisms towards tumor progression using genomic characterizations of cancer patients is critical to foresee the disease prognosis and to be able to develop new therapeutic strategies. The first contribution of this thesis addresses the problem of scarcity of integrated computational methods that perform tumor volume prediction and knowledge extraction simultaneously on genomic data, to get insightful information related to cancer progression. Instead of solving these problems separately (i.e., tumor volume prediction and knowledge extraction at separate times), using our integrated approach, we are able to gather robust knowledge about the molecular mechanisms that are related to tumor progression. We tested our proposed algorithm PrognosiT on `THCA` cohort from TCGA using two pathway/gene set collections, which are curated specifically for cancer, namely, `Hallmark` gene set collection [Liberzon et al., 2015] and `PID` pathway collection [Schaefer et al., 2009], as prior information sources. The predictive performance results we obtained showed that PrognosiT performed comparable or even statistically significantly better than RF [Breiman, 2001] and SVR [Drucker et al., 1996] (see Figure 5.1), which are two standard baseline machine learning algorithms used for prediction from genomic data. The power of our method comes from the fact that the number of used gene expression features was significantly fewer (i.e., less than one-tenth) in PrognosiT while having comparable or even better predictive performance results while conjointly extracting the relevant pathways to predict the tumor volume.

To show the biological relevance of the results of our algorithm, we provided the selection frequencies of pathways/gene sets for `THCA` data set (see Tables A.1–A.2). We also showed the unique genes that are selected in every replication of our algorithm

and checked their gene expression levels between tumor and normal tissues. Among these genes, we displayed several statistically significantly up- or down-regulated ones (see Figures 5.2–B.1). We saw that frequently selected pathways/gene sets and unique genes in `THCA` cohort are supported by the existing literature, and some of the genes that are resulted from our algorithm are already in use as prognostic factors for thyroid carcinoma.

Even though the regression problem in this chapter is about to predict tumor volumes of cancer patients utilizing their gene expression profiles, it is possible to easily adapt the used computational framework to other disease types, other phenotypes and other prior knowledge sources with slight modifications. However, since there exist different underlying mechanisms related to different diseases, the prior knowledge source should give us insightful information about the studied prediction task. Thus, the compatibility of pathway/gene set collection with the studied prediction problem should be the priority to get good predictive performance using PrognosiT.

As the second main contribution of this thesis, we introduced a scalable multiple approximate kernel learning algorithm, named MAKL, which is designed specifically for large-scale genomic data sets. Our approach can combine any low-dimensional kernel approximation with a group Lasso formulation. We used a modified version of the random feature mapping as kernel approximation algorithm in our experiments (see Algorithm 1 and 2 for details). MAKL is fast and appropriate for large-scale genomic data such as single-cell sequencing data having large sample-size with many features. Our method can integrate prior information in the form of pathways/gene sets into the learning process to increase the interpretability without sacrificing the predictive accuracy.

To benchmark MAKL algorithm, we used three data sets constructed from two data sources. As prior information sources, we used `Hallmark` and `PID` pathway/gene set collections particularly curated for cancer research. We processed genomic data from TCGA consortium to form the early- and late-stage cancer data set together with two-year cancer survival data set. We also used single-cell RNA-Seq data provided

by the Broad Institute Single Cell Portal to understand the molecular underpinnings of cell malignancy. Our algorithm MAKL was capable of outperforming the baseline algorithms based on XGBoost using only a small fraction of the pathways/gene sets available (see Figures 5.3–5.6); thus, it provided sparse solutions. It also provided selection frequencies associated with the pathways/gene sets used as prior information sources for the corresponding classification task (see Tables C.1–C.2), which can be used to understand the biological mechanisms towards the applied classification problem. Regarding scalability, MAKL was trained (while simultaneously extracting the significant information from it, unlike the baseline algorithms) less than 30 seconds on the average for the three problems we discussed while the baseline algorithms XGBoost (Subset) and XGBoost (Full) perform the same classification task between 1 minute and 4 minutes on the average, respectively. This scalability property, unlike the standard MKL algorithms, makes it possible for MAKL to be used with large-scale data, such as single-cell genomic data sets. As a result, this characteristic could lead to discovery of new biomarkers for tumor plasticity and of new techniques for immunotherapy and targeted cancer therapies.

An interesting direction for future studies may be building a multi-class extension of MAKL to be used with multi-class problems. Also, integration of low-dimensional kernel approximation into other machine learning models is exciting considering the increasing need for fast data processing and interpretability. Our MAKL approach is promising in several domains including but not limited to computational biology thanks to its scalability, flexibility and also its ability to deal with highly-correlated features. It may also be used for analyzing time series data associated with forecasting problems such as financial forecasting problems and electric load forecasting problems. As to computer vision domain, using multiple kernels for object detection and categorization has been widely studied [Yang et al., 2009, Vedaldi et al., 2009, Bucak et al., 2014]. In a neuroimaging study, multiple kernels have been calculated to determine the brain regions that are informative [Schrouff et al., 2018]. To benefit from its scalability, our proposed MAKL approach can be used to determine the regions that are likely to contain significant information in visual inputs.

In this thesis, we also introduced a scalable multiple kernel k-means clustering framework that integrates prior information and is particularly designed for large-scale data clustering. It combines any low-dimensional approximation of kernel matrices with k-means clustering and performs a forward set selection by maximizing the silhouette score associated with clustering result of each iteration. In this framework, we used random features to approximate the kernel matrices. Thanks to integration of approximation matrices, this clustering method is capable of capturing nonlinear relationships within the data, which is particularly important in genomic data analysis context considering the highly complex nature of genomic profiles. Our proposed algorithm is scalable, can be used in large-scale genomic data clustering problems such as high-dimensional single-cell sequencing data. As prior information source, we used `Hallmark` gene set collection; we processed and combined genomic data from TCGA consortium to cluster. We also benefited from the mutation information, `ER` positivity, `PR` positivity, `HER2` positivity and histological grade information provided from TCGA consortium to better analyze the results. To compare the findings of our algorithm, we performed a baseline k-means clustering experiment on the full data set which resulted 35.8% decrease in silhouette score. We also cross-checked the results of our approach by applying two-dimensional t-SNE to the approximation matrix selected in the first iteration of our algorithm (see Figure 5.9). Our approach uses only a small fraction of the full data set while performing clustering and providing interpretability on gene set level. Interpretation of clustering results while analyzing genomic data is promising for future studies since discovering hidden molecular dynamics of the data set may give a hope to find new cancer subtypes that are sensitive to new targeted therapies and/or to set new treatment regimens.

Appendix A

SELECTION FREQUENCIES OF GENE SETS WITH PROGNOSIT ALGORITHM

Table A.1: The selection frequencies of gene sets in the `Hallmark` collection over 100 replications resulted from PrognosiT algorithm.

Gene set name	**Selection frequency**
HYPOXIA	99
KRAS_SIGNALING_DN	99
SPERMATOGENESIS	96
PANCREAS_BETA_CELLS	95
INFLAMMATORY_RESPONSE	86
XENOBIOTIC_METABOLISM	85
FATTY_ACID_METABOLISM	75
NOTCH_SIGNALING	73
KRAS_SIGNALING_UP	62
ANGIOGENESIS	54
P53_PATHWAY	41
MYOGENESIS	39
GLYCOLYSIS	37
BILE_ACID_METABOLISM	34
HEDGEHOG_SIGNALING	33
EPITHELIAL_MESENCHYMAL_TRANSITION	33
COAGULATION	32
CHOLESTEROL_HOMEOSTASIS	30

G2M_CHECKPOINT	30
E2F_TARGETS	30
APICAL_JUNCTION	26
APOPTOSIS	23
ESTROGEN_RESPONSE_LATE	22
MYC_TARGETS_V2	19
TNFA_SIGNALING_VIA_NFKB	18
ADIPOGENESIS	18
PEROXISOME	18
WNT_BETA_CATENIN_SIGNALING	13
DNA_REPAIR	12
COMPLEMENT	12
MYC_TARGETS_V1	12
INTERFERON_GAMMA_RESPONSE	10
IL2_STAT5_SIGNALING	10
ANDROGEN_RESPONSE	9
HEME_METABOLISM	7
TGF_BETA_SIGNALING	6
INTERFERON_ALPHA_RESPONSE	6
REACTIVE_OXIGEN_SPECIES_PATHWAY	6
ESTROGEN_RESPONSE_EARLY	5
APICAL_SURFACE	3
UV_RESPONSE_UP	3
IL6_JAK_STAT3_SIGNALING	2
PI3K_AKT_MTOR_SIGNALING	2
UV_RESPONSE_DN	2
MITOTIC_SPINDLE	0
PROTEIN_SECRETION	0
UNFOLDED_PROTEIN_RESPONSE	0
MTORC1_SIGNALING	0

OXIDATIVE_PHOSPHORYLATION	0
ALLOGRAFT_REJECTION	0

Table A.2: The selection frequencies of pathways in the PID collection over 100 replications resulted from PrognosiT algorithm.

Gene set name	**Selection frequency**
CONE_PATHWAY	98
HIF1_TFPATHWAY	98
NFAT_TFPATHWAY	88
ARF6_PATHWAY	85
HNF3A_PATHWAY	85
FRA_PATHWAY	84
ERBB_NETWORK_PATHWAY	82
HNF3B_PATHWAY	81
WNT_SIGNALING_PATHWAY	77
RHODOPSIN_PATHWAY	77
UPA_UPAR_PATHWAY	74
TCR_CALCIUM_PATHWAY	66
FANCONI_PATHWAY	63
LIS1_PATHWAY	63
HES_HEY_PATHWAY	60
IL23_PATHWAY	57
ERB_GENOMIC_PATHWAY	53
VEGF_VEGFR_PATHWAY	49
ATR_PATHWAY	46
ALPHA_SYNUCLEIN_PATHWAY	46
SYNDECAN_4_PATHWAY	44
FOXM1_PATHWAY	43

BETA_CATENIN_NUC_PATHWAY	43
EPHB_FWD_PATHWAY	39
CMYB_PATHWAY	39
REG_GR_PATHWAY	37
PLK1_PATHWAY	37
CD8_TCR_DOWNSTREAM_PATHWAY	37
TRAIL_PATHWAY	34
CDC42_PATHWAY	33
LKB1_PATHWAY	32
P53_DOWNSTREAM_PATHWAY	30
GLYPICAN_1PATHWAY	29
SYNDECAN_3_PATHWAY	28
FGF_PATHWAY	28
SYNDECAN_1_PATHWAY	26
IL3_PATHWAY	25
NEPHRIN_NEPH1_PATHWAY	24
DELTA_NP63_PATHWAY	24
RAS_PATHWAY	24
AP1_PATHWAY	22
MYC_REPRESS_PATHWAY	22
REELIN_PATHWAY	21
PRL_SIGNALING_EVENTS_PATHWAY	20
AURORA_B_PATHWAY	20
ARF_3PATHWAY	20
INSULIN_GLUCOSE_PATHWAY	20
PS1_PATHWAY	19
P38_MK2_PATHWAY	17
AVB3_OPN_PATHWAY	16
S1P_META_PATHWAY	15
HEDGEHOG_2PATHWAY	15

BARD1_PATHWAY	15
AMB2_NEUTROPHILS_PATHWAY	13
P38_ALPHA_BETA_PATHWAY	12
CDC42_REG_PATHWAY	12
CIRCADIAN_PATHWAY	12
IL12_STAT4_PATHWAY	12
ANGIOPOIETIN_RECEPTOR_PATHWAY	11
IL1_PATHWAY	11
IL5_PATHWAY	11
ATF2_PATHWAY	11
TOLL_ENDOGENOUS_PATHWAY	11
MAPK_TRK_PATHWAY	11
AR_TF_PATHWAY	10
INTEGRIN5_PATHWAY	9
ATM_PATHWAY	8
HDAC_CLASSI_PATHWAY	8
E2F_PATHWAY	7
DNA_PK_PATHWAY	7
FAS_PATHWAY	7
INTEGRIN2_PATHWAY	6
TAP63_PATHWAY	6
MYC_ACTIV_PATHWAY	5
RHOA_REG_PATHWAY	5
RAC1_REG_PATHWAY	5
IL8_CXCR1_PATHWAY	5
RB_1PATHWAY	5
LYSOPHOSPHOLIPID_PATHWAY	4
HDAC_CLASSIII_PATHWAY	4
S1P_S1P4_PATHWAY	4
ERBB2_ERBB3_PATHWAY	4

HIF1A_PATHWAY	4
BMP_PATHWAY	4
ENDOTHELIN_PATHWAY	3
RHOA_PATHWAY	3
TXA2PATHWAY	3
CASPASE_PATHWAY	3
VEGFR1_PATHWAY	3
NOTCH_PATHWAY	2
NFKAPPAB_ATYPICAL_PATHWAY	2
IL4_2PATHWAY	2
ER_NONGENOMIC_PATHWAY	2
CD40_PATHWAY	2
INTEGRIN3_PATHWAY	2
S1P_S1P1_PATHWAY	2
TNF_PATHWAY	2
MYC_PATHWAY	2
P75_NTR_PATHWAY	2
PI3K_PLC_TRK_PATHWAY	2
EPHA2_FWD_PATHWAY	2
WNT_NONCANONICAL_PATHWAY	1
IL27_PATHWAY	1
NFKAPPAB_CANONICAL_PATHWAY	1
PTP1B_PATHWAY	1
NETRIN_PATHWAY	1
ARF6_DOWNSTREAM_PATHWAY	1
THROMBIN_PAR4_PATHWAY	1
AJDISS_2PATHWAY	1
ECADHERIN_NASCENT_AJ_PATHWAY	1
ALK2_PATHWAY	1
THROMBIN_PAR1_PATHWAY	1

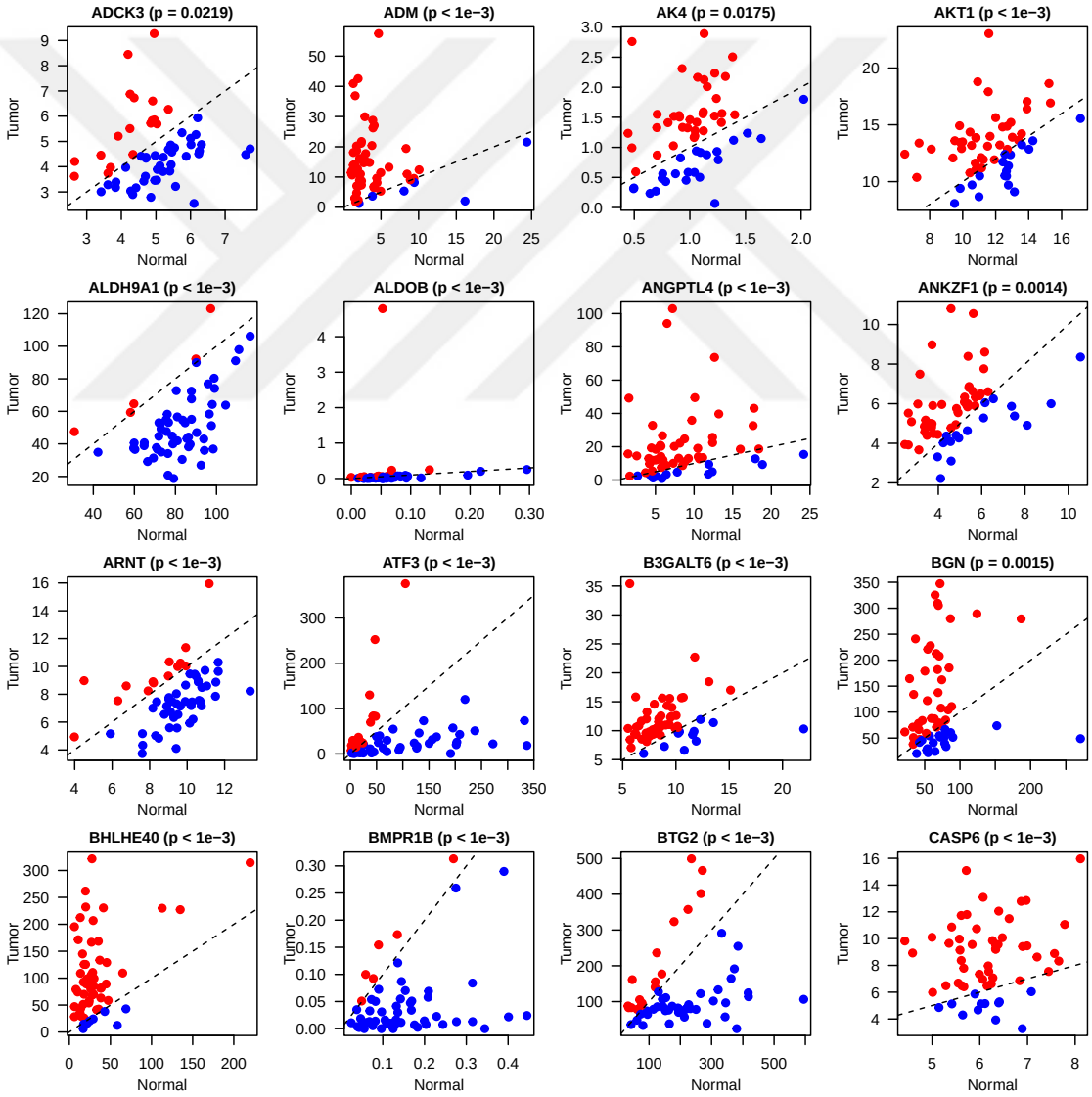
SYNDECAN_2_PATHWAY	1
EPHRINB_REV_PATHWAY	1
NCADHERIN_PATHWAY	1
INTEGRIN_A4B1_PATHWAY	1
SMAD2_3NUCLEAR_PATHWAY	0
FCER1_PATHWAY	0
BCR_5PATHWAY	0
ERBB4_PATHWAY	0
INSULIN_PATHWAY	0
INTEGRIN1_PATHWAY	0
P73PATHWAY	0
P38_MKK3_6PATHWAY	0
GMCSF_PATHWAY	0
HDAC_CLASSII_PATHWAY	0
BETA_CATENIN_DEG_PATHWAY	0
TCR_PATHWAY	0
HIF2PATHWAY	0
INTEGRIN_CS_PATHWAY	0
MET_PATHWAY	0
IL12_2PATHWAY	0
S1P_S1P3_PATHWAY	0
LPA4_PATHWAY	0
AR_PATHWAY	0
ARF6_TRAFFICKING_PATHWAY	0
ILK_PATHWAY	0
NECTIN_PATHWAY	0
RET_PATHWAY	0
CD8_TCR_PATHWAY	0
WNT_CANONICAL_PATHWAY	0
TCPTP_PATHWAY	0

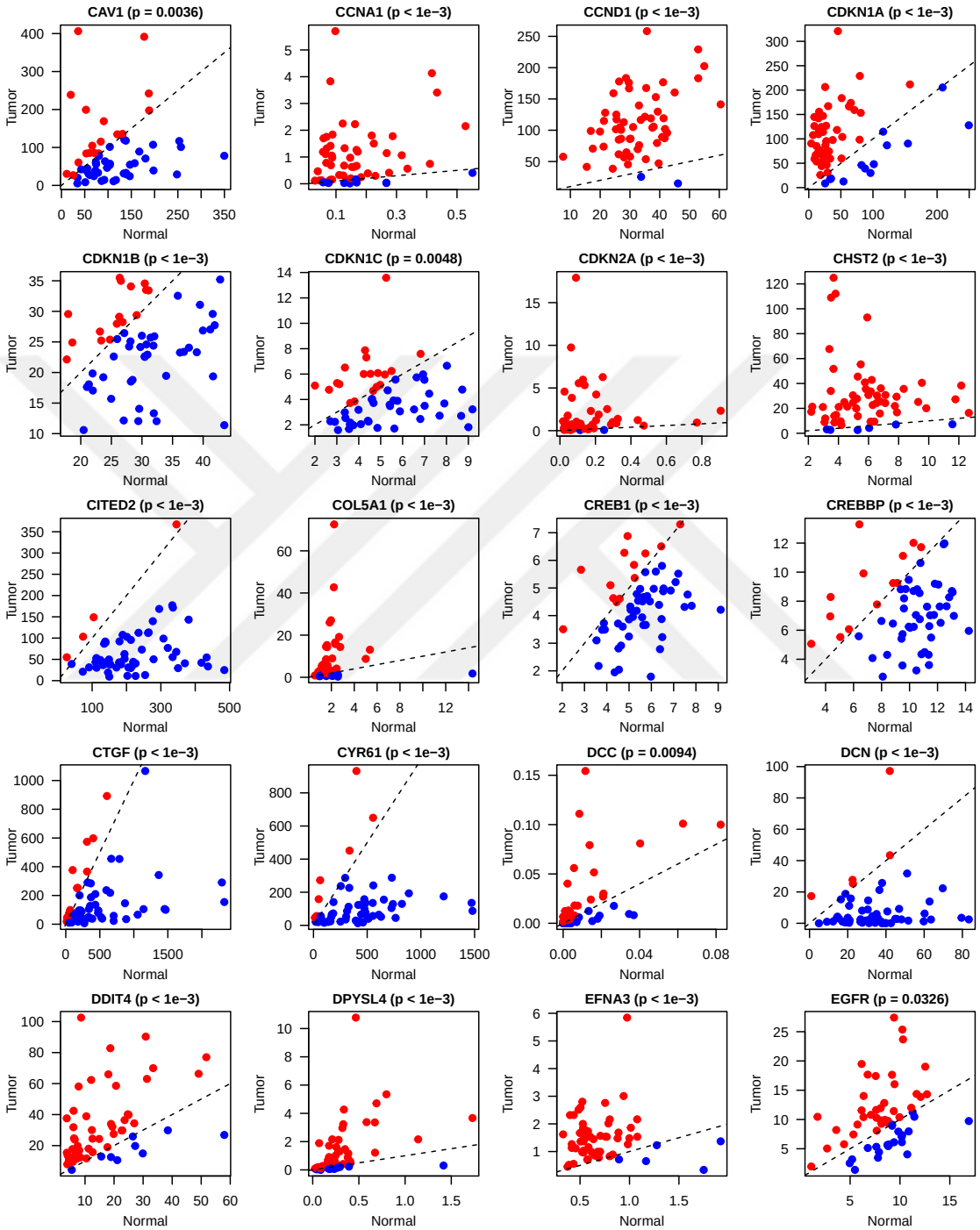
SHP2_PATHWAY	0
TELOMERASE_PATHWAY	0
NFAT_3PATHWAY	0
INTEGRIN_A9B1_PATHWAY	0
MTOR_4PATHWAY	0
IL2_1PATHWAY	0
CXCR4_PATHWAY	0
IGF1_PATHWAY	0
ERBB1_RECEPTOR_PROXIMAL_PATHWAY	0
TCR_RAS_PATHWAY	0
FOXO_PATHWAY	0
RANBP2_PATHWAY	0
PI3KCI_PATHWAY	0
IL2_PI3K_PATHWAY	0
CERAMIDE_PATHWAY	0
INTEGRIN4_PATHWAY	0
AVB3_INTEGRIN_PATHWAY	0
IFNG_PATHWAY	0
RXR_VDR_PATHWAY	0
ERBB1_DOWNSTREAM_PATHWAY	0
EPHA_FWDPATHWAY	0
IL6_7_PATHWAY	0
ECADHERIN_KERATINOCYTE_PATHWAY	0
ALK1_PATHWAY	0
PDGFRB_PATHWAY	0
TRKR_PATHWAY	0
TCR_JNK_PATHWAY	0
HIV_NEF_PATHWAY	0
ERA_GENOMIC_PATHWAY	0
PDGFRA_PATHWAY	0

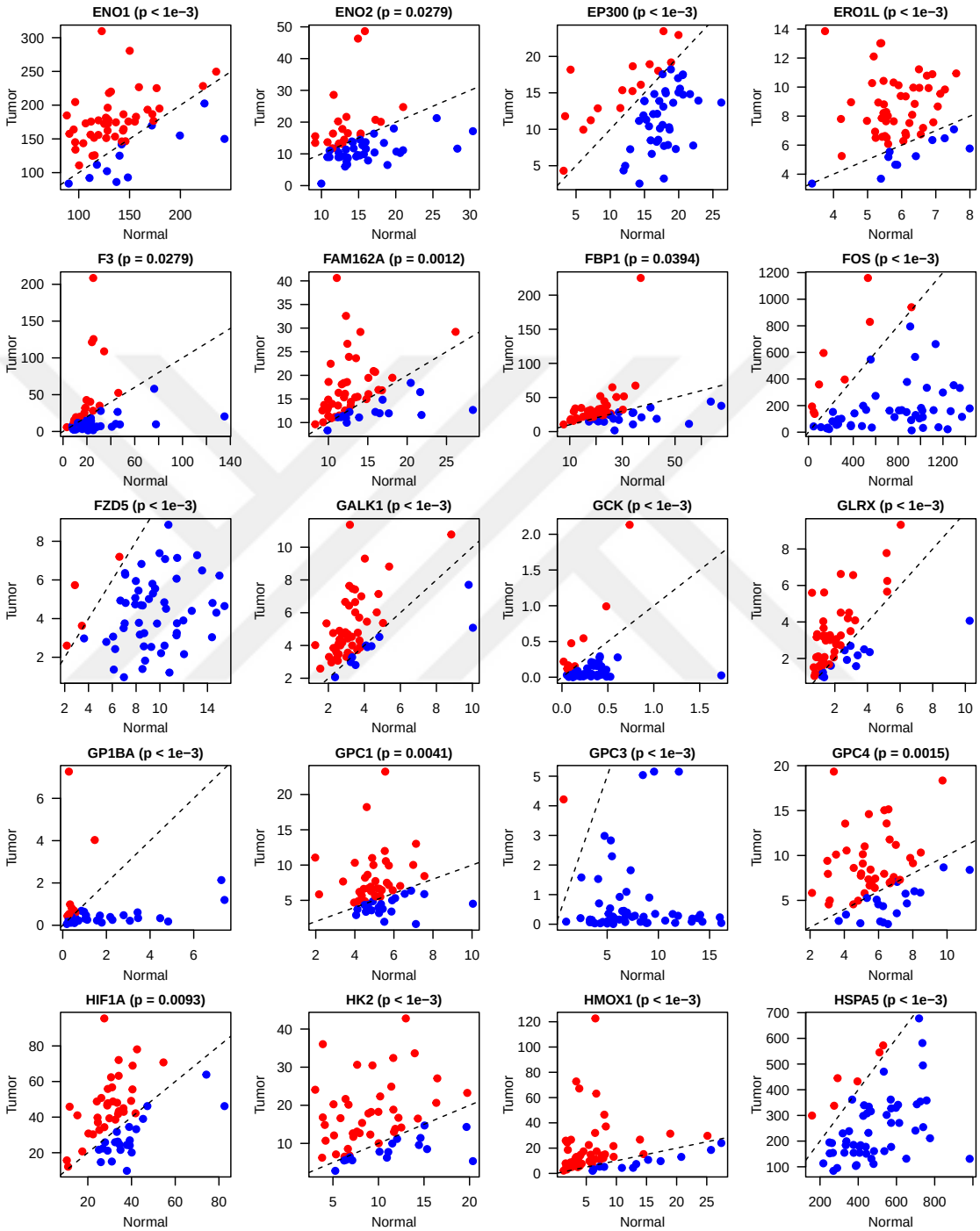
RETINOIC_ACID_PATHWAY	0
P38_GAMMA_DELTA_PATHWAY	0
IL8_CXCR2_PATHWAY	0
AR_NONGENOMIC_PATHWAY	0
ERBB1_INTERNALIZATION_PATHWAY	0
HEDGEHOG_GLI_PATHWAY	0
CXCR3_PATHWAY	0
SMAD2_3PATHWAY	0
P38_ALPHA_BETA_DOWNSTREAM_PATHWAY	0
KIT_PATHWAY	0
ECADHERIN_STABILIZATION_PATHWAY	0
EPO_PATHWAY	0
IL2_STAT5_PATHWAY	0
VEGFR1_2_PATHWAY	0
A6B1_A6B4_INTEGRIN_PATHWAY	0
AURORA_A_PATHWAY	0
PI3KCI_AKT_PATHWAY	0
P53_REGULATION_PATHWAY	0
ANTHRAX_PATHWAY	0
S1P_S1P2_PATHWAY	0
LYMPH_ANGIOGENESIS_PATHWAY	0
RAC1_PATHWAY	0
FAK_PATHWAY	0
TGFBR_PATHWAY	0

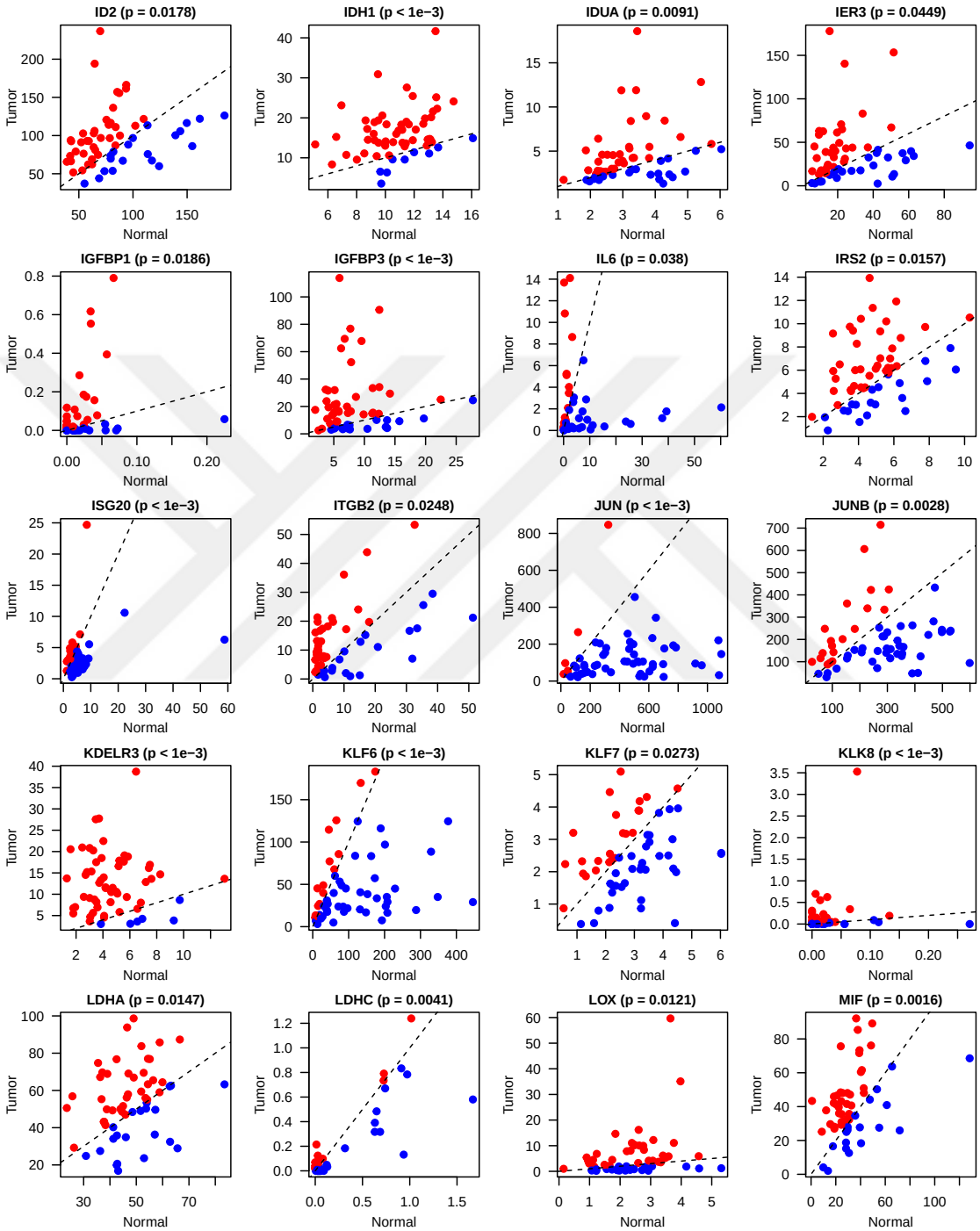
Appendix B

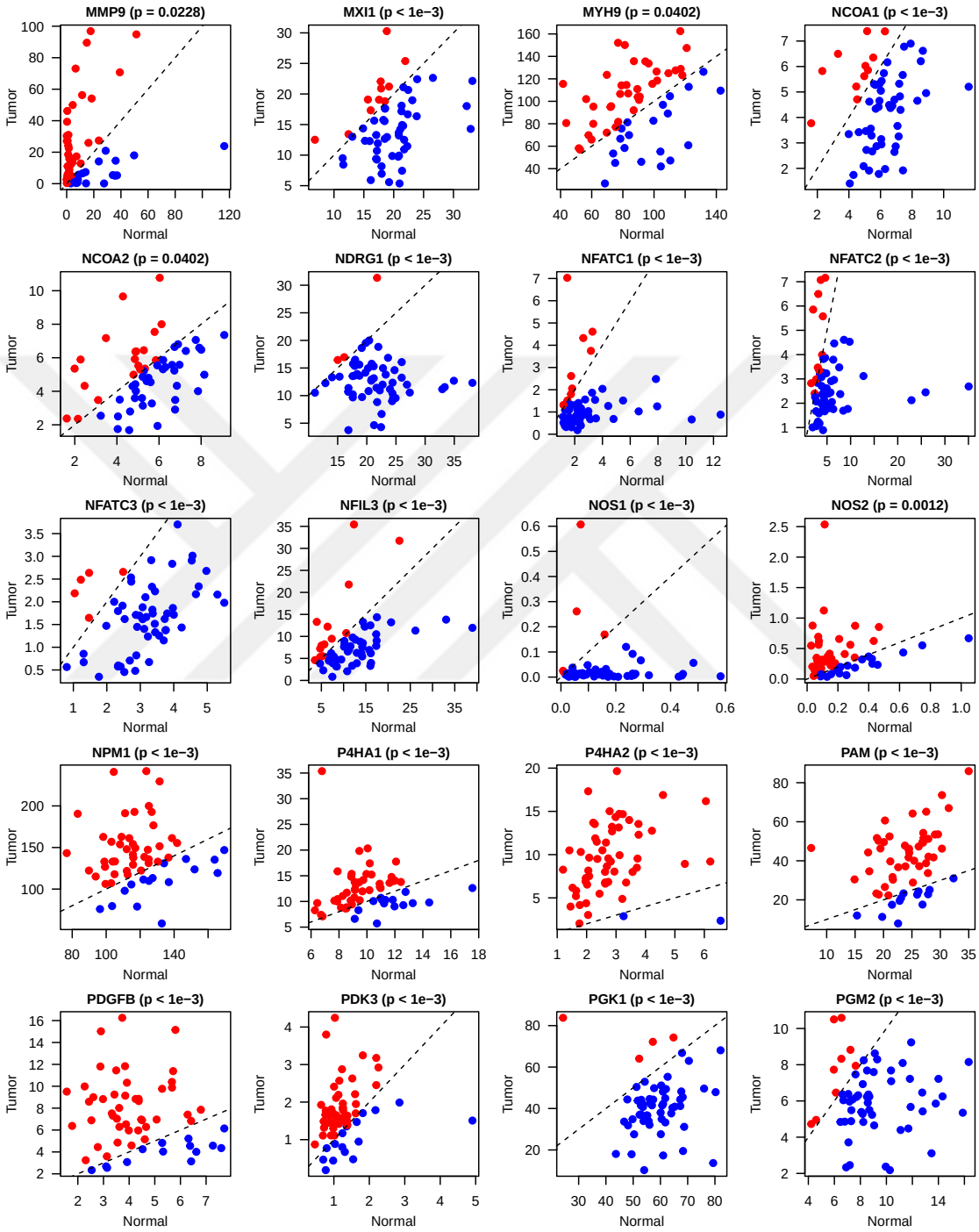
STATISTICALLY SIGNIFICANT GENES IDENTIFIED BY PROGNOSIT ALGORITHM

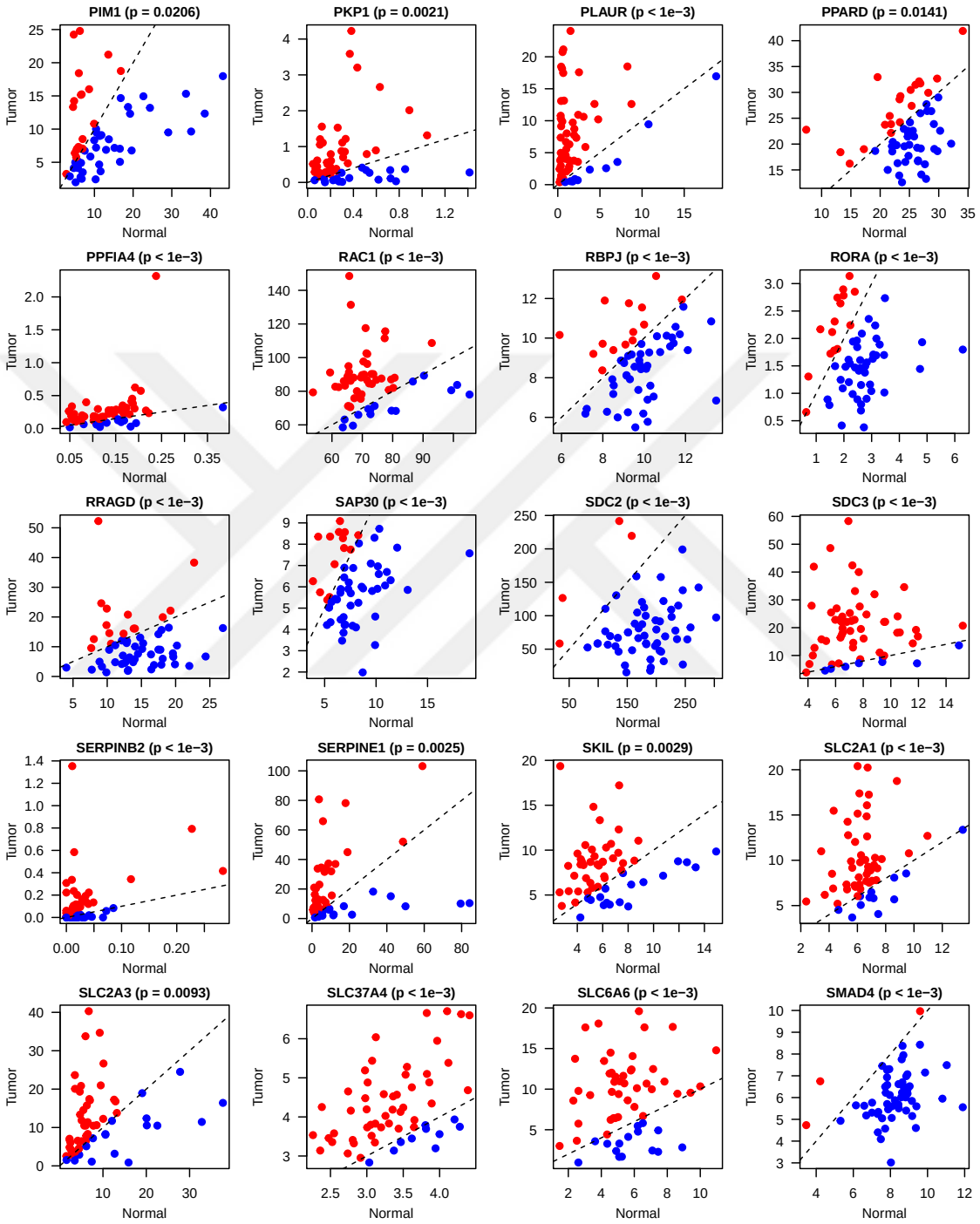



CAV1 (p = 0.0036)
CCNA1 (p < 1e−3)
CCND1 (p < 1e−3)
CDKN1A (p < 1e−3)
CDKN1B (p < 1e−3)
CDKN1C (p = 0.0048)
CDKN2A (p < 1e−3)
CHST2 (p < 1e−3)
CITED2 (p < 1e−3)
COL5A1 (p < 1e−3)
CREB1 (p < 1e−3)
CREBBP (p < 1e−3)
CTGF (p < 1e−3)
CYR61 (p < 1e−3)
DCC (p = 0.0094)
DCN (p < 1e−3)
DDIT4 (p < 1e−3)
DPYSL4 (p < 1e−3)
EFNA3 (p < 1e−3)
EGFR (p = 0.0326)
Tumor
Normal


ENO1 (p < 1e−3)
ENO2 (p = 0.0279)
EP300 (p < 1e−3)
ERO1L (p < 1e−3)
F3 (p = 0.0279)
FAM162A (p = 0.0012)
FBP1 (p = 0.0394)
FOS (p < 1e−3)
FZD5 (p < 1e−3)
GALK1 (p < 1e−3)
GCK (p < 1e−3)
GLRX (p < 1e−3)
GP1BA (p < 1e−3)
GPC1 (p = 0.0041)
GPC3 (p < 1e−3)
GPC4 (p = 0.0015)
HIF1A (p = 0.0093)
HK2 (p < 1e−3)
HMOX1 (p < 1e−3)
HSPA5 (p < 1e−3)
Tumor
Normal


ID2 (p = 0.0178)
IDH1 (p < 1e−3)
IDUA (p = 0.0091)
IER3 (p = 0.0449)
IGFBP1 (p = 0.0186)
IGFBP3 (p < 1e−3)
IL6 (p = 0.038)
IRS2 (p = 0.0157)
ISG20 (p < 1e−3)
ITGB2 (p = 0.0248)
JUN (p < 1e−3)
JUNB (p = 0.0028)
KDELR3 (p < 1e−3)
KLF6 (p < 1e−3)
KLF7 (p = 0.0273)
KLK8 (p < 1e−3)
LDHA (p = 0.0147)
LDHC (p = 0.0041)
LOX (p = 0.0121)
MIF (p = 0.0016)
Tumor
Normal


MMP9 (p = 0.0228)
MXI1 (p < 1e−3)
MYH9 (p = 0.0402)
NCOA1 (p < 1e−3)
NCOA2 (p = 0.0402)
NDRG1 (p < 1e−3)
NFATC1 (p < 1e−3)
NFATC2 (p < 1e−3)
NFATC3 (p < 1e−3)
NFIL3 (p < 1e−3)
NOS1 (p < 1e−3)
NOS2 (p = 0.0012)
NPM1 (p < 1e−3)
P4HA1 (p < 1e−3)
P4HA2 (p < 1e−3)
PAM (p < 1e−3)
PDGFB (p < 1e−3)
PDK3 (p < 1e−3)
PGK1 (p < 1e−3)
PGM2 (p < 1e−3)
Tumor
Normal


PIM1 (p = 0.0206)
PKP1 (p = 0.0021)
PLAUR (p < 1e−3)
PPARD (p = 0.0141)
PPFIA4 (p < 1e−3)
RAC1 (p < 1e−3)
RBPJ (p < 1e−3)
RORA (p < 1e−3)
RRAGD (p < 1e−3)
SAP30 (p < 1e−3)
SDC2 (p < 1e−3)
SDC3 (p < 1e−3)
SERPINB2 (p < 1e−3)
SERPINE1 (p = 0.0025)
SKIL (p = 0.0029)
SLC2A1 (p < 1e−3)
SLC2A3 (p = 0.0093)
SLC37A4 (p < 1e−3)
SLC6A6 (p < 1e−3)
SMAD4 (p < 1e−3)
Tumor
Normal

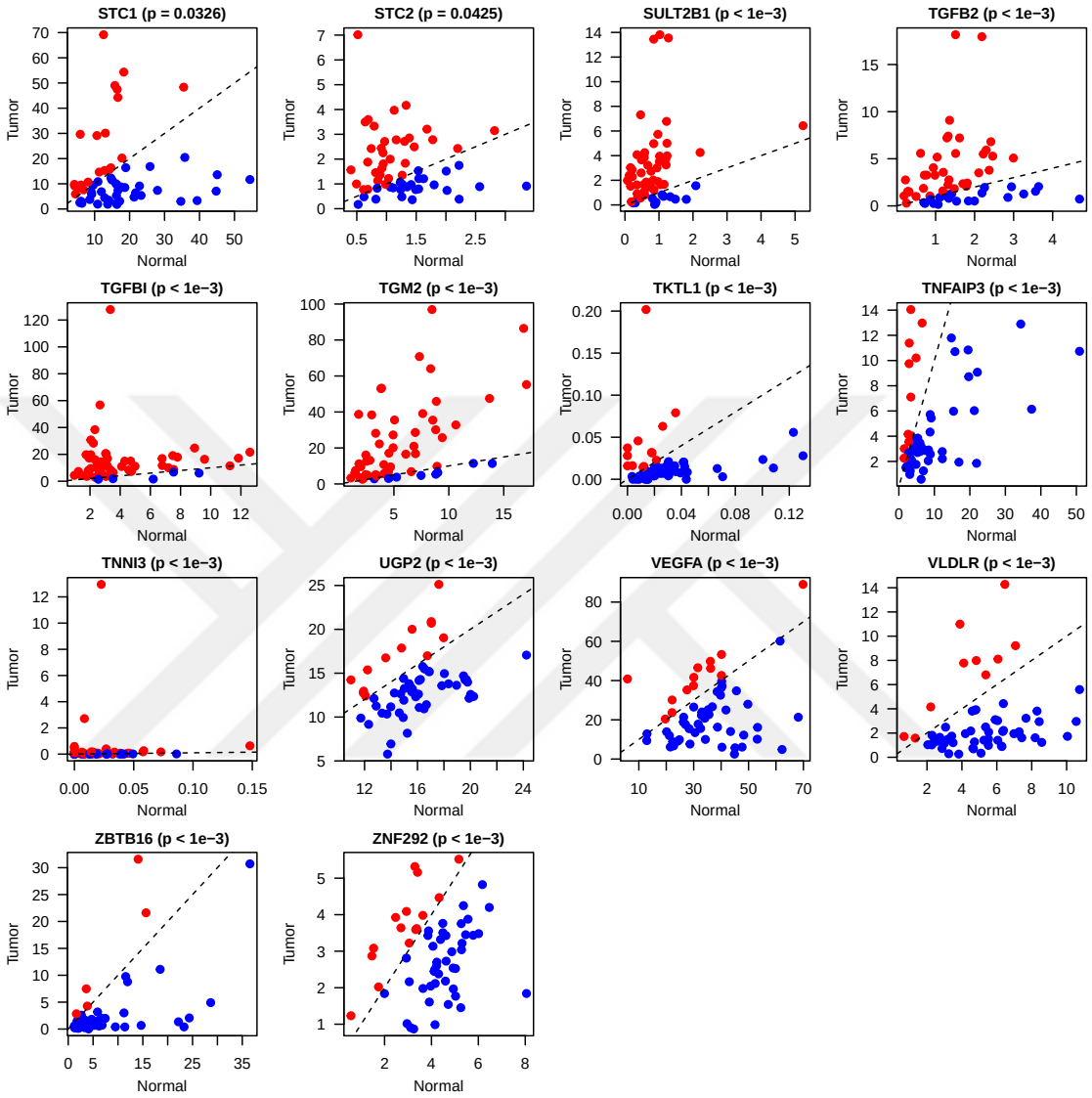


Figure B.1: Statistically significantly up- or down-regulated genes during tumor progression selected by PrognosiT algorithm on thyroid carcinoma (i.e., `THCA`) data set over 100 replications. We performed the Wilcoxon signed-rank test to determine whether there is a significant difference (i.e., p-value < 0.05) between the gene expressions coming from tumor and normal tissues.

Appendix C

SELECTION FREQUENCIES OF GENE SETS WITH MAKL ALGORITHM

Table C.1: Selection frequencies of 50 gene sets in the `Hallmark` collection by MAKL algorithm over 100 replications using single-cell melanoma immunotherapy data set, for detailed pathway and gene analysis regarding the cell malignancy classification task. Selection frequencies using 4 different regularization parameter (i.e., λ multiplier) values with $D = 100$ are reported. At the last column, the mean selection frequency for each gene set is reported and the gene sets are sorted from the most selected one to the least selected one.

	λ				**Mean selection**
Gene set name	**0.9**	**0.8**	**0.7**	**0.6**	**frequency**
ALLOGRAFT_REJECTION	95	98	100	100	98.25
INTERFERON_ALPHA_RESPONSE	67	86	95	96	86.00
KRAS_SIGNALING_UP	52	78	95	99	81.00
EPITHELIAL_MESENCHYMAL_TRANSITION	24	52	76	91	60.75
APOPTOSIS	25	52	75	87	59.75
ANDROGEN_RESPONSE	17	46	76	94	58.25
INTERFERON_GAMMA_RESPONSE	13	30	36	52	32.75
APICAL_JUNCTION	5	19	36	51	27.75
COAGULATION	1	2	8	22	8.25
IL2_STAT5_SIGNALING	0	2	8	15	6.25
MYOGENESIS	2	2	5	8	4.25
HYPOXIA	0	0	3	9	3.00
COMPLEMENT	0	1	1	4	1.50

INFLAMMATORY_RESPONSE	0	0	0	0	0
ESTROGEN_RESPONSE_LATE	0	0	0	0	0
ESTROGEN_RESPONSE_EARLY	0	0	0	0	0
MTORC1_SIGNALING	0	0	0	0	0
GLYCOLYSIS	0	0	0	0	0
UV_RESPONSE_DN	0	0	0	0	0
APICAL_SURFACE	0	0	0	0	0
TNFA_SIGNALING_VIA_NFKB	0	0	0	0	0
REACTIVE_OXIGEN_SPECIES_PATHWAY	0	0	0	0	0
HEME_METABOLISM	0	0	0	0	0
UV_RESPONSE_UP	0	0	0	0	0
P53_PATHWAY	0	0	0	0	0
ADIPOGENESIS	0	0	0	0	0
XENOBIOTIC_METABOLISM	0	0	0	0	0
PI3K_AKT_MTOR_SIGNALING	0	0	0	0	0
IL6_JAK_STAT3_SIGNALING	0	0	0	0	0
OXIDATIVE_PHOSPHORYLATION	0	0	0	0	0
PANCREAS_BETA_CELLS	0	0	0	0	0
CHOLESTEROL_HOMEOSTASIS	0	0	0	0	0
MITOTIC_SPINDLE	0	0	0	0	0
WNT_BETA_CATENIN_SIGNALING	0	0	0	0	0
TGF_BETA_SIGNALING	0	0	0	0	0
DNA_REPAIR	0	0	0	0	0
G2M_CHECKPOINT	0	0	0	0	0
NOTCH_SIGNALING	0	0	0	0	0
PROTEIN_SECRETION	0	0	0	0	0
HEDGEHOG_SIGNALING	0	0	0	0	0
UNFOLDED_PROTEIN_RESPONSE	0	0	0	0	0
E2F_TARGETS	0	0	0	0	0
MYC_TARGETS_V1	0	0	0	0	0

MYC_TARGETS_V2	0	0	0	0	0
FATTY_ACID_METABOLISM	0	0	0	0	0
ANGIOGENESIS	0	0	0	0	0
BILE_ACID_METABOLISM	0	0	0	0	0
PEROXISOME	0	0	0	0	0
SPERMATOGENESIS	0	0	0	0	0
KRAS_SIGNALING_DN	0	0	0	0	0

Table C.2: Selection frequencies of 196 pathways in the PID collection by MAKL algorithm over 100 replications using single-cell melanoma immunotherapy data set, for detailed pathway and gene analysis regarding the cell malignancy classification task. Selection frequencies using 4 different regularization parameter (i.e., λ multiplier) values with $D = 100$ are reported. At the last column, the mean selection frequency for each gene set is reported and the gene sets are sorted from the most selected one to the least selected one.

	λ				**Mean selection**
Pathway name	**0.9**	**0.8**	**0.7**	**0.6**	**frequency**
CXCR4_PATHWAY	100	100	100	100	100.00
SYNDECAN_4_PATHWAY	70	96	100	100	91.50
CASPASE_PATHWAY	55	100	100	100	88.75
IL12_2PATHWAY	1	23	77	98	49.75
BETA_CATENIN_NUC_PATHWAY	0	1	25	74	25.00
PDGFRB_PATHWAY	0	8	30	51	22.25
IL12_STAT4_PATHWAY	0	5	18	48	17.75
AP1_PATHWAY	0	1	4	44	12.25
TCR_PATHWAY	0	2	6	12	5.00
IL4_2PATHWAY	0	1	1	13	3.75
ECADHERIN_NASCENT_AJ_PATHWAY	0	0	1	11	3.00
ARF6_TRAFFICKING_PATHWAY	0	0	3	8	2.75

CD8_TCR_PATHWAY	0	0	0	5	1.25
P53_DOWNSTREAM_PATHWAY	0	0	0	1	0.25
INTEGRIN1_PATHWAY	0	0	0	0	0
AVB3_INTEGRIN_PATHWAY	0	0	0	0	0
SYNDECAN_1_PATHWAY	0	0	0	0	0
CMYB_PATHWAY	0	0	0	0	0
HIF1_TFPATHWAY	0	0	0	0	0
MYC_ACTIV_PATHWAY	0	0	0	0	0
NCADHERIN_PATHWAY	0	0	0	0	0
INTEGRIN2_PATHWAY	0	0	0	0	0
ILK_PATHWAY	0	0	0	0	0
MYC_REPRESS_PATHWAY	0	0	0	0	0
RHOA_REG_PATHWAY	0	0	0	0	0
INTEGRIN3_PATHWAY	0	0	0	0	0
RB_1PATHWAY	0	0	0	0	0
INTEGRIN_A4B1_PATHWAY	0	0	0	0	0
FGF_PATHWAY	0	0	0	0	0
AJDISS_2PATHWAY	0	0	0	0	0
VEGFR1_2_PATHWAY	0	0	0	0	0
S1P_S1P3_PATHWAY	0	0	0	0	0
PI3KCI_PATHWAY	0	0	0	0	0
ENDOTHELIN_PATHWAY	0	0	0	0	0
FOXM1_PATHWAY	0	0	0	0	0
FCER1_PATHWAY	0	0	0	0	0
SYNDECAN_3_PATHWAY	0	0	0	0	0
ERBB1_DOWNSTREAM_PATHWAY	0	0	0	0	0
EPHRINB_REV_PATHWAY	0	0	0	0	0
ALPHA_SYNUCLEIN_PATHWAY	0	0	0	0	0
PTP1B_PATHWAY	0	0	0	0	0
P38_MK2_PATHWAY	0	0	0	0	0

A6B1_A6B4_INTEGRIN_PATHWAY	0	0	0	0	0
IL2_1PATHWAY	0	0	0	0	0
DELTA_NP63_PATHWAY	0	0	0	0	0
AURORA_B_PATHWAY	0	0	0	0	0
CD8_TCR_DOWNSTREAM_PATHWAY	0	0	0	0	0
P73PATHWAY	0	0	0	0	0
SHP2_PATHWAY	0	0	0	0	0
P38_ALPHA_BETA_DOWNSTREAM_PATHWAY	0	0	0	0	0
ECADHERIN_STABILIZATION_PATHWAY	0	0	0	0	0
FANCONI_PATHWAY	0	0	0	0	0
SMAD2_3NUCLEAR_PATHWAY	0	0	0	0	0
BCR_5PATHWAY	0	0	0	0	0
PRL_SIGNALING_EVENTS_PATHWAY	0	0	0	0	0
RHOA_PATHWAY	0	0	0	0	0
ERBB4_PATHWAY	0	0	0	0	0
LYSOPHOSPHOLIPATHWAY	0	0	0	0	0
INSULIN_PATHWAY	0	0	0	0	0
NOTCH_PATHWAY	0	0	0	0	0
P38_MKK3_6PATHWAY	0	0	0	0	0
GMCSF_PATHWAY	0	0	0	0	0
WNT_NONCANONICAL_PATHWAY	0	0	0	0	0
NFKAPPAB_ATYPICAL_PATHWAY	0	0	0	0	0
HDAC_CLASSII_PATHWAY	0	0	0	0	0
BETA_CATENIN_DEG_PATHWAY	0	0	0	0	0
HDAC_CLASSIII_PATHWAY	0	0	0	0	0
GLYPICAN_1PATHWAY	0	0	0	0	0
IL27_PATHWAY	0	0	0	0	0
NFKAPPAB_CANONICAL_PATHWAY	0	0	0	0	0
E2F_PATHWAY	0	0	0	0	0
ER_NONGENOMIC_PATHWAY	0	0	0	0	0

DNA_PK_PATHWAY	0	0	0	0	0
HIF2PATHWAY	0	0	0	0	0
CD40_PATHWAY	0	0	0	0	0
ATR_PATHWAY	0	0	0	0	0
INTEGRIN_CS_PATHWAY	0	0	0	0	0
MET_PATHWAY	0	0	0	0	0
LPA4_PATHWAY	0	0	0	0	0
AR_PATHWAY	0	0	0	0	0
NFAT_TFPATHWAY	0	0	0	0	0
EPHB_FWD_PATHWAY	0	0	0	0	0
AVB3_OPN_PATHWAY	0	0	0	0	0
S1P_S1P4_PATHWAY	0	0	0	0	0
FRA_PATHWAY	0	0	0	0	0
REELIN_PATHWAY	0	0	0	0	0
PS1_PATHWAY	0	0	0	0	0
NECTIN_PATHWAY	0	0	0	0	0
P38_ALPHA_BETA_PATHWAY	0	0	0	0	0
WNT_SIGNALING_PATHWAY	0	0	0	0	0
TRAIL_PATHWAY	0	0	0	0	0
CDC42_PATHWAY	0	0	0	0	0
RET_PATHWAY	0	0	0	0	0
CDC42_REG_PATHWAY	0	0	0	0	0
ATM_PATHWAY	0	0	0	0	0
ARF6_PATHWAY	0	0	0	0	0
LKB1_PATHWAY	0	0	0	0	0
WNT_CANONICAL_PATHWAY	0	0	0	0	0
TCPTP_PATHWAY	0	0	0	0	0
ANGIOPOIETIN_RECEPTOR_PATHWAY	0	0	0	0	0
FAS_PATHWAY	0	0	0	0	0
CIRCADIAN_PATHWAY	0	0	0	0	0

TXA2PATHWAY	0	0	0	0	0
HDAC_CLASSI_PATHWAY	0	0	0	0	0
S1P_S1P1_PATHWAY	0	0	0	0	0
TELOMERASE_PATHWAY	0	0	0	0	0
HNF3B_PATHWAY	0	0	0	0	0
NETRIN_PATHWAY	0	0	0	0	0
IL1_PATHWAY	0	0	0	0	0
NFAT_3PATHWAY	0	0	0	0	0
REG_GR_PATHWAY	0	0	0	0	0
CONE_PATHWAY	0	0	0	0	0
INTEGRIN_A9B1_PATHWAY	0	0	0	0	0
ERB_GENOMIC_PATHWAY	0	0	0	0	0
ARF6_DOWNSTREAM_PATHWAY	0	0	0	0	0
MTOR_4PATHWAY	0	0	0	0	0
IGF1_PATHWAY	0	0	0	0	0
ERBB1_RECEPTOR_PROXIMAL_PATHWAY	0	0	0	0	0
TNF_PATHWAY	0	0	0	0	0
PLK1_PATHWAY	0	0	0	0	0
TCR_RAS_PATHWAY	0	0	0	0	0
IL5_PATHWAY	0	0	0	0	0
FOXO_PATHWAY	0	0	0	0	0
VEGF_VEGFR_PATHWAY	0	0	0	0	0
THROMBIN_PAR4_PATHWAY	0	0	0	0	0
MYC_PATHWAY	0	0	0	0	0
RANBP2_PATHWAY	0	0	0	0	0
IL2_PI3K_PATHWAY	0	0	0	0	0
CERAMIDE_PATHWAY	0	0	0	0	0
AR_TF_PATHWAY	0	0	0	0	0
P75_NTR_PATHWAY	0	0	0	0	0
S1P_META_PATHWAY	0	0	0	0	0

INTEGRIN4_PATHWAY	0	0	0	0	0
AMB2_NEUTROPHILS_PATHWAY	0	0	0	0	0
IFNG_PATHWAY	0	0	0	0	0
RXR_VDR_PATHWAY	0	0	0	0	0
LIS1_PATHWAY	0	0	0	0	0
ATF2_PATHWAY	0	0	0	0	0
UPA_UPAR_PATHWAY	0	0	0	0	0
ERBB2_ERBB3_PATHWAY	0	0	0	0	0
EPHA_FWDPATHWAY	0	0	0	0	0
HIF1A_PATHWAY	0	0	0	0	0
BMP_PATHWAY	0	0	0	0	0
IL3_PATHWAY	0	0	0	0	0
IL6_7_PATHWAY	0	0	0	0	0
ECADHERIN_KERATINOCYTE_PATHWAY	0	0	0	0	0
ALK1_PATHWAY	0	0	0	0	0
TRKR_PATHWAY	0	0	0	0	0
TCR_JNK_PATHWAY	0	0	0	0	0
NEPHRIN_NEPH1_PATHWAY	0	0	0	0	0
IL23_PATHWAY	0	0	0	0	0
HIV_NEF_PATHWAY	0	0	0	0	0
ERA_GENOMIC_PATHWAY	0	0	0	0	0
ERBB_NETWORK_PATHWAY	0	0	0	0	0
ALK2_PATHWAY	0	0	0	0	0
RHODOPSIN_PATHWAY	0	0	0	0	0
PDGFRA_PATHWAY	0	0	0	0	0
RETINOIC_ACID_PATHWAY	0	0	0	0	0
P38_GAMMA_DELTA_PATHWAY	0	0	0	0	0
IL8_CXCR2_PATHWAY	0	0	0	0	0
HEDGEHOG_2PATHWAY	0	0	0	0	0
INTEGRIN5_PATHWAY	0	0	0	0	0

AR_NONGENOMIC_PATHWAY	0	0	0	0	0
ERBB1_INTERNALIZATION_PATHWAY	0	0	0	0	0
HEDGEHOG_GLI_PATHWAY	0	0	0	0	0
CXCR3_PATHWAY	0	0	0	0	0
VEGFR1_PATHWAY	0	0	0	0	0
SMAD2_3PATHWAY	0	0	0	0	0
KIT_PATHWAY	0	0	0	0	0
EPO_PATHWAY	0	0	0	0	0
IL2_STAT5_PATHWAY	0	0	0	0	0
TCR_CALCIUM_PATHWAY	0	0	0	0	0
THROMBIN_PAR1_PATHWAY	0	0	0	0	0
SYNDECAN_2_PATHWAY	0	0	0	0	0
RAC1_REG_PATHWAY	0	0	0	0	0
AURORA_A_PATHWAY	0	0	0	0	0
ARF_3PATHWAY	0	0	0	0	0
INSULIN_GLUCOSE_PATHWAY	0	0	0	0	0
PI3KCI_AKT_PATHWAY	0	0	0	0	0
IL8_CXCR1_PATHWAY	0	0	0	0	0
TAP63_PATHWAY	0	0	0	0	0
BARD1_PATHWAY	0	0	0	0	0
P53_REGULATION_PATHWAY	0	0	0	0	0
TOLL_ENDOGENOUS_PATHWAY	0	0	0	0	0
ANTHRAX_PATHWAY	0	0	0	0	0
S1P_S1P2_PATHWAY	0	0	0	0	0
RAS_PATHWAY	0	0	0	0	0
MAPK_TRK_PATHWAY	0	0	0	0	0
PI3K_PLC_TRK_PATHWAY	0	0	0	0	0
EPHA2_FWD_PATHWAY	0	0	0	0	0
LYMPH_ANGIOGENESIS_PATHWAY	0	0	0	0	0
RAC1_PATHWAY	0	0	0	0	0

FAK_PATHWAY	0	0	0	0	0
HNF3A_PATHWAY	0	0	0	0	0
TGFBR_PATHWAY	0	0	0	0	0
HES_HEY_PATHWAY	0	0	0	0	0

Appendix D

STATISTICAL TESTS USED

In this chapter, we review the statistical tests that we used to analyze the experimental results of three algorithms demonstrated in this thesis.

D.1 Paired t-Test

We used paired t-test to compare the NRMSE values obtained as a result of 100 independent replications (see Figure 5.1) to check whether there is a significant difference between the mean NRMSE values of studied algorithms. The two-tailed paired t-test is a statistical technique used to determine whether the mean difference between the two sets of paired observations is zero (i.e., the null hypothesis H_0 is $\mu_d = 0$.). The test statistic is calculated as

$$t = \frac{\bar{d} - \mu_d}{\hat{\sigma}/\sqrt{n}},$$

where $\bar{d}$ is the mean of the differences between pairs of observed data points, $\hat{\sigma}$ is the standard deviation of the differences between pairs of observed data points and n is the sample size.

To check whether the null hypothesis is true, let us define the significance level as α. To calculate the probability of observing the test statistic under the null hypothesis, we check the t-distribution with $(n-1)$ degrees of freedom. If the corresponding p-value is less than α, we reject the null hypothesis H_0.

D.2 Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test [Wilcoxon, 1945] is a nonparametric test used to determine whether medians of two samples are equal or not. It is a test based on difference scores. It is assumed that, under the null hypothesis, paired differences

come from a continuous symmetric distribution hence the median and mean are identical. The test statistic is W, defined as the $\min(W^+, W^-)$, where W^+ is the sum of the positive ranks and W^- is the sum of the negative ranks.

Let d_i be the difference between the values of observed data pair $i = 1, 2, \ldots, N$. The differences between the observed data pairs are ranked regarding their absolute values. In case $d_i = 0$, ranks are split evenly between W^+ and W^- and if the number of ties is odd, one is not taken into account.

$$W^+ = \sum_{d_i>0} \text{rank}(d_i) + \frac{1}{2} \sum_{d_i=0} \text{rank}(d_i)$$
$$W^- = \sum_{d_i<0} \text{rank}(d_i) + \frac{1}{2} \sum_{d_i=0} \text{rank}(d_i)$$

The exact critical value of W can be found when N is small. If N is large, we can approximate the null distribution

$$z = \frac{W - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}} \sim Z,$$

where z is normally distributed. The null hypothesis that there is no difference between the medians of two samples is rejected at significance level, α, if this calculated value is less than $Z_{\alpha/2}$.

D.3 Fisher's Exact Test

Fisher's exact test [Fisher, 1922] is a statistical test to determine the nonrandom relations between two categorical variables. It is used to analyze contingency tables.

Let us consider the categorical variables X and Y with s and t observed outputs. Then, let us form a matrix of dimensionality $s \times t$, where each of its entries e_{ij} represent the number of observations when $x = i$ and $y = j$. Let us now compute the row and column sums as R_i and C_j. Then the total sum T of all entries in the matrix is

$$T = \sum_i R_i = \sum_j C_j.$$

Now let us calculate the conditional probability P_C of getting the matrix we already have, given the specific row and column sums:

$$P_c = \frac{(R_1!R_2!\dots R_s!)(C_1!C_2!\dots C_t!)}{T!\prod_i \prod_j e_{ij}!} \tag{D.1}$$

Here, it is noteworthy that Equation D.1 corresponds to multivariate version of hypergeometric distribution. Given row and column sums, we should now calculate all possible matrices containing nonnegative entries and calculate their corresponding probability using D.1.

In this thesis, we applied this test to contingency tables which are matrices of dimensionality 2×2. In this case, the p-value of the test can be computed as the sum of all probabilities for all possible such matrices, that are less than P_c.

BIBLIOGRAPHY

[Antoniadis & Fan, 2001] Antoniadis, A., & Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, *96*, 939–955.

[Bach, 2008] Bach, F. R. (2008). Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, *9*, 1179–1225.

[Bakin, 1999] Bakin, S. (1999). Adaptive regression and model selection in data mining problems [Doctoral thesis, Australian National University]. Australian National University Research Repository. https://openresearch-repository.anu.edu.au/handle/1885/9449

[Bektaş & Gönen, 2021] Bektaş, A. B., & Gönen, M. (2021). PrognosiT: Pathway/gene set-based tumour volume prediction using multiple kernel learning. *BMC Bioinformatics*, *22*, 537.

[Bektaş et al., 2022] Bektaş, A. B., Ak, Ç., & Gönen, M. (2022). Fast and interpretable genomic data analysis using multiple approximate kernel learning. *Bioinformatics*, *38*(Suppl 1), i77–i83.

[Boser et al., 1992] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop of Computational Learning Theory*, *5*, 144–152.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

[Bucak et al., 2014] Bucak, S. S., Jin, R., & Jain, A. K. (2014). Multiple kernel learning for visual object recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(7), 1354–1369.

[Cai, 2001] Cai, T. T. (2001). Regularization of wavelet approximations: Discussion. *Journal of the American Statistical Association*, *96*, 960–962.

[Chen & He, 2019] Chen, T., & He, T. (2019). *xgboost: Extreme Gradient Boosting*. R package version 0.90.0.2.

[Chen & Guestrin, 2016] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[Chen & Ishwaran, 2012] Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, *6*, 323–329.

[Cortes & Vapnik, 1995] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297.

[Costello et al., 2014] Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S. A., Mpindi, J. P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., NCI DREAM Community, Collins, J. J., Gallahan, D., Singer, D., Saez-Rodriguez, J., ... Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology. 32*, 1202–1212.

[Díaz-Uriarte & Alvares de Andrés, 2006] Díaz-Uriarte, R., & Alvares de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, *7*, 3.

[Drucker et al., 1996] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*, 155–161.

[Ein-Dor et al., 2006] Ein-Dor, L., Zuk, O., & Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in

cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(15), 5923–5928.

[Fisher, 1922] Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, *85*(1), 87–94.

[Forgy, 1965] Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, *21*, 768–769.

[Gatenby & Gillies, 2004] Gatenby, R., & Gillies, R. (2004). Why do cancers have high aerobic glycolysis? *Nature Reviews Cancer*, *4*, 891–899.

[Gayther & Pharoah, 2010] Gayther, S. A., & Pharoah, P. D. (2010). The inherited genetics of ovarian and endometrial cancer. *Current Opinion in Genetics & Development*, *20*(3), 231–238.

[Girolami, 2002] Girolami, M. (2002). Mercer kernel-based clustering in feature space. *IEEE Transactions on Neural Networks*, *13*(3), 780–784.

[Gönen & Alpaydın, 2011] Gönen, M., & Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, *12*, 2211–2268.

[Gönen & Margolin, 2014] Gönen, M. & Margolin, A. A. (2014). Localized data fusion for kernel k-means clustering with application to cancer biology. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, *1*.

[Gönen et al., 2017] Gönen, M., Weir, B. A., Cowley, G. S., Vazquez, F., Guan, Y., Jaiswal, A., Karasuyama, M., Uzunangelov, V., Wang, T., Tsherniak, A., Howell, S., Marbach, D., Hoff, B., Norman, T. C., Airola, A., Bivol, A., Bunte, K., Carlin, D., Chopra, S., Deran, A., ... Margolin, A. A. (2017). A Community

challenge for inferring genetic predictors of gene essentialities through analysis of a functional screen of cancer cell lines. *Cell Systems*, *5*, 485–497.

[Guyon et al., 1992] Guyon, I. M., Boser, B. E., & Vapnik, V. N. (1992). Automatic capacity tuning of very large VC-dimension classifiers. *Advances in Neural Information Processing Systems*, *5*.

[1] IBM. (2016). *ILOG CPLEX Interactive Optimizer*. Version 12.6.3.

[Hartigan & Wong, 1979] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, *28*(1), 100–108.

[Ishwaran & Kogalur, 2020] Ishwaran, H., & Kogalur, U. B. (2020). *randomForestSRC: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*. R package version 2.9.3.

[Issa et al., 2015] Issa, M. R., Samuels, S. E., Bellile, E., Shalabi, F. L., Eisbruch, A., & Wolf, G. (2015). Tumor volumes and prognosis in laryngeal cancer. *Cancers*, *7*(4), 2236–2261.

[Jerby-Arnon et al., 2018] Jerby-Arnon, L., Shah, P., Cuoco, M. S., Rodman, C., Su, M. J., Melms, J. C., Leeson, R., Kanodia, A., Mei, S., Lin, J. R., Wang, S., Rabasha, B., Liu, D., Zhang, G., Margolais, C., Ashenberg, O., Ott, P. A., Buchbinder, E. I., Haq, R., Hodi, F. S., ... Regev, A. (2018). A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell*, *175*, 984–997.

[Klement et al., 2014] Klement, R. J., Allgäuer, M., Appold, S., Dieckmann, K., Ernst, I., Ganswindt, U., Holy, R., Nestle, U., Nevinny-Stickel, M., Semrau, S., Sterzing, F., Wittig, A., Andratschke, N., & Guckenberger, M. (2014). Support vector machine-based prediction of local tumor control after stereotactic body radiation therapy for early-stage non-small cell lung cancer. *International Journal of Radiation Oncology, Biology, Physics*, *88*(3), 732–738.

[Langdon et al., 2020] Langdon, S. P., Herrington, C. S., Hollis, R. L., & Gourley, C. (2020). Estrogen signaling and its potential as a target for therapy in ovarian cancer. *Cancers (Basel)*, *12*(6), 1647.

[Lee et al., 2006] Lee, J. H., Bae, J. S., Ryu, K. W., Lee, J. S., Park, S. R., Kim, C. G., Kook, M. C., Choi, I. J., Kim, Y. W., Park, J. G., & Bae, J. M. (2006). Gastric cancer patients at high-risk of having synchronous cancer. *World Journal of Gastroenterology*, *12*(16), 2588–2592.

[Li et al., 2020] Li, J., Lu, Q., & Wen, Y. (2020). Multi-kernel linear mixed model with adaptive Lasso for prediction analysis on high-dimensional multi-omics data. *Bioinformatics*, *36*(6), 1785–1794.

[Liberzon et al., 2015] Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems*, *1*(6), 417–425.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, *28*(2), 129–137.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–297.

[Manzella et al., 2017] Manzella, L., Stella, S., Pennisi, M. S., Tirrò, E., Massimino, M., Romano, C., Puma, A., Tavarelli, M., & Vigneri, P. (2017). *New Insights in thyroid Cancer and p53 family proteins. International Journal of Molecular Sciences*, *18*(6), 1325.

[Meier, 2020] Meier, L. (2020). *grplasso: Fitting User-Specified Models with Group Lasso Penalty.* R package version 0.4-7.

[Polo et al., 2016] Polo, V., Pasello, G., Frega, S., Favaretto, A., Koussis, H., Conte, P., & Bonanno, L. (2016). Squamous cell carcinomas of the lung and of the head and neck: New insights on molecular characterization. *Oncotarget*, *7*(18), 25050–25063.

[Rahimi & Recht, 2007] Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, *20*, 1177–1184.

[Rappoport & Shamir, 2018] Rappoport, N., & Shamir, R. (2018) Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. *Nucleic Acids Research*, *46*, 10546–10562.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, *20*, 53–65.

[Rudin, 1994] Rudin, W. (1994). *Fourier analysis on groups*. Wiley Classics Library. Wiley-Interscience.

[Ruan et al., 2009] Ruan, K., Song, G., & Ouyang, G. (2009). Role of hypoxia in the hallmarks of human cancer. *Journal of Cellular Biochemistry*, *107*(6), 1053–1062.

[Sari et al., 2018] Sari, I. N., Phi, L., Jun, N., Wijaya, Y. T., Lee, S., & Kwon, H. Y. (2018). Hedgehog Signaling in Cancer: A Prospective Therapeutic Target for Eradicating Cancer Stem Cells. *Cells*, *7*(11), 208.

[Scala et al., 2006] Scala, S., Giuliano, P., Ascierto, P. A., Ieranò, C., Franco, R., Napolitano, M., Ottaiano, A., Lombardi, M. L., Luongo, M., Simeone, E., Castiglia, D., Mauro, F., De Michele, I., Calemma, R., Botti, G., Caracò, C., Nicoletti, G., Satriano, R. A., & Castello, G. (2006). Human melanoma metastases express functional CXCR4. *Clinical Cancer Research*, *12*, 2427–2433.

[Schaefer et al., 2009] Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., & Buetow, K. H. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Research*, *37*, 674–679.

[Schölkopf & Smola, 2002] Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

[Schrouff et al., 2018] Schrouff, J., Monteiro, J. M., Portugal, L., Rosa, M. J., Phillips, C., & Mourão-Miranda, J. Embedding anatomical or functional knowledge in whole-brain multiple kernel learning models. *Neuroinformatics*, *16*, 117–143.

[Shi et al., 2005] Shi, T., Seligson, D., Belldegrun, A. S., Palotie, A., & Horvath, S. (2005). Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. *Modern Pathology*, *18*(4), 547–557.

[Stephan et al., 2015] Stephan, J., Stegle, O. & Beyer, A. (2015). A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, *6*, 7432.

[Tamási et al., 2011] Tamási, V., Monostory, K., Prough, R. A., & Falus, A. (2011). Role of xenobiotic metabolism in cancer: involvement of transcriptional and miRNA regulation of P450s. *Cellular and Molecular Life Sciences*. *68*, 1131–1146.

[Tan et al., 2019] Tan, Y., Fu, D., Li, D., Kong, X., Jiang, K., Chen, L., Yuan, Y., & Ding, K. (2019). Predictors and risk factors of pathologic complete response following neoadjuvant chemoradiotherapy for rectal cancer: A population-based analysis. *Frontiers in Oncology*, *9*, 497.

[Telgarsky & Vattani, 2010] Telgarsky, M., & Vattani, A. (2010). Hartigan's method: k-means clustering without Voronoi. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, *9*, 820–827.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society.*, *58*, 267–288.

[Timmermans et al., 2016] Timmermans, A. J., Lange, C. A., de Bois, J. A., van Werkhoven, E., Hamming-Vrieze, O., Hilgers, F. J., & van den Brekel, M. W. (2016). Tumor volume as a prognostic factor for local control and overall survival in advanced larynx cancer. *The Laryngoscope*, *126*(2), E60–E67.

[Uzunangelov et al., 2021] Uzunangelov, V., Wong, C. K., & Stuart, J. M. (2021). Accurate cancer phenotype prediction with AKLIMATE, a stacked kernel learner integrating multimodal genomic data and pathway knowledge. *PLoS computational biology*, *17*(4), e1008878.

[van der Maaten & Hinton, 2008] van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605.

[Vedaldi et al., 2009] Vedaldi, A., Gulshan, V., Varma, M., & Zisserman, A. (2009). Multiple kernels for object detection. *IEEE 12th International Conference on Computer Vision*, 606–613.

[Weber, 2009] Weber, W. A. (2009). Assessing tumor response to therapy. *Journal of Nuclear Medicine*, *50 Suppl 1*, 1S–10S.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80–83.

[Xu et al., 2010] Xu, Z., Jin, R., Yang, H., King, I., & Lyu, M. (2010). Simple and efficient multiple kernel learning by group Lasso. *Proceedings of the 27th International Conference on Machine Learning*.

[Yang et al., 2009] Yang, J., Li, Y., Tian, Y., Duan, L., & Gao, W. (2009). Group-sensitive multiple kernel learning for object categorization. *Proceedings of IEEE 12th International Conference on Computer Vision*, 436–443.

[Yang et al., 2018] Yang, J., Kumar, A., Vilgelm, A. E., Chen, S. C., Ayers, G. D., Novitskiy, S. V., Joyce, S., & Richmond, A. (2018). Loss of CXCR4 in myeloid cells enhances antitumor immunity and reduces melanoma growth through NK cell and FASL mechanisms. *Cancer Immunology Research*, *6*(10).

[Yin et al., 2020] Yin, L., Duan, J. J., Bian, X. W., & Yu, S. C. (2020). Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Research*, *22*(61).

[Yuan & Lin, 2006] Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, *68*, 49–67.

[Zaballos et al., 2019] Zaballos, M. A., Acuña-Ruiz, A., Morante, M., Crespo, P., & Santisteban, P. (2019). Regulators of the RAS-ERK pathway as therapeutic targets in thyroid cancer. *Endocrine-related Cancer*, *26*(6), R319–R344.

[Zhang & Rudnicky, 2002] Zhang, R., & Rudnicky, A. I. (2002). A large scale clustering scheme for kernel k-means. *International Conference on Pattern Recognition*, *4*, 289–292.