Phenotype Relevant Network-based biomarker discovery
Integrating multiple Omics data
- EMT network-based lung cancer prognosis prediction

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

Borong Shao

Berlin, 2018

Erstgutachter: Prof. Dr. Tim Conrad

Zweitgutachter: Prof. Dr. Gunnar Klau

Tag der Disputation: 9 July, 2018

Acknowledgements

Foremost, I would like to thank my advisor Prof. Dr. Tim Conrad for his patience, advice, and support throughout my PhD project. I would also like to thank my group members Nada Cvetković, Fahrnaz Jayrannejad, Dr. Arthur You, Dr. Iliusi Vega and Victor Mireles for sharing their knowledge and experience. I also thank Prof. Dr. Christof Schütte for his helpful advice.

Meanwhile, I give thanks to Dr. Carlo Vittorio Cannistraci at Technische Universität Dresden for his advice in publications. I thank very much Prof. Dr. Åslang Helland and Maria Moksnes Bjaanæs at Oslo University Norway for providing me lung cancer patient data.

I thank my thesis reviewer Prof. Dr. Gunnar Klau at Heinrich-Heine-Universität Düsseldorf for his earnest effort and advice.

Last by not least, I would like to thank my family and friends.

Abstract

Biological networks have been employed as prior knowledge to identify robust molecular signatures. Many studies integrate protein-protein interaction(PPI) network with gene expression data, where the network is often used to guide the search in the feature space to select the important nodes or subnetworks as signatures. As multiple types of omics data become more and more available, it brings the questions of which data types contain better molecular signatures and whether it is beneficial to use multiple omics data simultaneously.

We investigated this topic with the task of cancer prognosis prediction. After conducting a comprehensive literature review, we selected 10 representative feature selection algorithms. Five of them integrate network information and the other five algorithms use only omics data. On single-omics levels, we performed feature selection alternatively on mRNA and miRNA expression data, DNA methylation data, and copy number alteration data. Then we evaluated the prediction performance of these features, as well as their stability, network properties, and biological interpretation. To obtain multi-omics signatures, we first combined the selected features from single-omics levels and evaluated the prediction performance of different combinations. Then we extended network-based feature selection algorithms to incorporate multi-omics data using a multiplex structure. For the feature selection algorithms that do not use network information, we employed them on concatenated data. Last but not least, we evaluated the predictive performance of both single-omics and multi-omics signatures on independent lung cancer multi-omics data.

One of the major challenges in biomarker discovery is the curse of dimensionality, which contributes to the low reproducibility of many proposed molecular signatures. Even with network-based feature selection algorithms, significant improvements are hard to achieve as the networks are often of large sizes. To avoid this issue and find biologically meaningful signatures, we propose to construct phenotype relevant gene regulatory networks based on Epithelial Mesenchymal Transition (EMT), which is demonstrated as highly relevant to the metastasis and prognosis of epithelial cancers. We integrated different types of omics data with EMT networks to select prognostic signatures. Although the dimensionality was reduced to less than 2.5% of the original, EMT-based feature selection gave even better prediction performance than selecting features from the original data. Frequently selected features achieved average AUC value of 0.83. The features were able to stratify patients into significantly different prognostic groups on both the training data and the independent testing data. Using combinations of single-omics signatures and using multiplex-based feature selection further improved the prediction performance.

Since biological data have large volume, high velocity, and wide varieties, it becomes necessary to employ database systems that meet the need of integrative omics data analysis. Therefore, we tested the performance of a few relational and non-relational databases for storing and retrieving omics data. Based on the results, we provided a few advices on building scalable omics data infrastructures.

Contents

Table of Contents v

1 Introduction **1**

1.1 Rich Omics Data and the Curse of Dimensionality 1

1.2 Methods for Identifying Molecular Signatures 7

1.2.1 Feature Selection on Single-omics Data 9

1.2.2 Network-based Feature Selection on Single-omics Data 11

1.2.3 Multi- and Integrative Omics Data Analysis 16

1.3 Thesis Overview 19

1.3.1 Motivations 19

1.3.2 Thesis Outline 21

2 Epithelial Mesenchymal Transition in Cancer Progression **23**

2.1 Biological Background 23

2.2 EMT Gene Regulations 25

2.3 A Novel EMT Network Model 27

3 Single-omics Prognostic Signatures for Lung Adenocarcinoma **33**

3.1 State-of-the-art Algorithms 33

3.2 Experiments 43

3.2.1 Data Pre-processing 43

3.2.2 Evaluation Metrics 44

3.2.3 EMT-based Feature Selection 47

3.3 Results 49

3.3.1 EMT Features vs. Random Features 50

3.3.2 AUC, AUPR and Accuracy of EMT Features 50

3.3.3 Frequently Selected EMT Features 56

3.3.4 The Effect of Adding Clinical Features 61

3.3.5 The Effect of Classification Thresholds 67

3.4 Analysis 68

3.4.1 Stability of Feature Selection Algorithms 68

3.4.2 Patterns in EMT Signatures 72

3.4.3 Survival Analysis Using EMT Signatures 80

3.5 Discussion 82

3.5.1 Advantages of EMT Features 82

3.5.2 Considerations among Multiple Evaluation Metrics 84

3.5.3 Comparisons to State-of-the-art Studies 84
3.5.4 Single Features vs. Composite Features - Our Insights 87

4 Multi-omics Prognostic Signatures for Lung Adenocarcinoma 89
4.1 Multi-omics Signatures from Single-omics Signatures 89
4.1.1 Survival Analysis on All Stage Samples 89
4.1.2 Survival Analysis on Early Stage Samples 91
4.1.3 The Application of Integrative Clustering Algorithms 91
4.2 A Novel Integrative Feature Selection Approach 99
4.2.1 The Construction of Multiplex Networks 99
4.2.2 Experiments 101
4.2.3 Results 101
4.2.4 Multi-omics Feature Compositions 107
4.3 EMT-based Prognosis Prediction on Real-world Clinical Datasets 110
4.3.1 Data Pre-processing 111
4.3.2 Results 111
4.4 Discussion 117
4.4.1 Advantages of Integrative EMT Signatures 117
4.4.2 The Importance of Relevant Network-based Feature Selection 121

5 Strategies for Handling Large Biological Data 123
5.1 Proteomics and Mass Spectrometry 123
5.1.1 The Complexity of Proteomics Data 123
5.1.2 MS and MS/MS Technologies for Protein Characterization 125
5.1.3 The Potential of Proteomics in Biomarker Discovery 126
5.2 The Big Data Challenge 127
5.3 Methods 128
5.3.1 Data for Testing the Databases 128
5.3.2 Databases and Data Models 129
5.3.3 Experiments 135
5.4 Results and Analysis 136
5.4.1 Storing and Querying MS Data 136
5.4.2 Storing and Querying MS/MS Data 138
5.5 Discussion 140
5.6 Conclusion 141

6 Summary 143
6.1 Contributions 143
6.2 Limitations 144
6.3 Outlook - an Intelligent Decision Support System 145

Bibliography 147

7 Appendix 173

Chapter 1

Introduction

1.1 Rich Omics Data and the Curse of Dimensionality

Cancer is a generic term for a large group of diseases characterized by uncontrollable cell growth and divisions, promotion of blood vessel construction, and the capability of invading other tissues, organs and forming metastases. Cancer accounts for more than 8 million deaths worldwide annually. According to the National Cancer Institute, there are more than 100 types of cancer, which are usually named after the organs or tissues where the cancer origins. The major cancer types are lung, liver, colorectal, breast cancer, etc. Cancers are also described by the type of cell that forms them, such as an epithelial cell or a squamous cell. Cancers that are formed by epithelial cells are called carcinomas. It is the most common type of cancer. Carcinomas that begin in different epithelial cell types have specific names. Adenocarcinoma is a cancer that forms in epithelial cells that produce fluids or mucus. Lung cancer is a leading cancer type, which causes more than 1.5 million deaths [222] per year. It is an aggressive, heterogeneous cancer type [60] and its long-term survival rate remains low despite the advances in surgery, radiotherapy, and chemotherapy [222]. As adenocarcinoma is the most common lung cancer histological subtype, accounting for almost half of all lung cancers, in this study we take Lung Adenocarcinoma as the example [50].

While it is difficult to cure a cancer patient at a late stage, it is very important to choose the right therapy according to the risk/prognosis of the patients. If we can predict the prognosis of cancer patients at diagnosis, e.g., classifying patients into high and low risk groups, it can help the clinicians and patients to make decisions between active versus supportive treatment in order to balance the benefits and toxicities. Traditional cancer prognosis prediction is based on clinical variables such as tumor stage, age, and disease history. The information of a patient is compared against population cancer registries [104]. However, these clinical parameters are insufficient to accurately predict the risk of patients [232]. Overwhelming evidence shows that histologically similar tumors can be of completely different diseases at the molecular level, each with different clinical behavior [143, 226, 255]. In addition, some medical tests such as computed tomography and biopsy can be harmful. The National Lung Screening Trial conducted a randomized trial using low-dose CT to screen lung cancer among high-risk persons. While the screening significantly reduced lung cancer mortality, the false positive rate is high (96%) [241]. Since cancer is a heterogeneous disease, more individualized prognosis prediction is necessary.

Fortunately, 'Omic' technologies provide detailed characterization of patients' molecular profiles to advance the clinical management of cancer.

Omics refers to the aggregate of studies in biology ending in omics, such as genomics, transcriptomics, epigenomics, proteomics, etc. The related suffix *-ome* is used to address the object under these studies - genome, transcriptome, epicurean, proteome, respectively. For example, genomics involves sequencing and the analysis of entire genome, which is the complete set of genetic material. It studies phenomenon such as heterosis, epistasis, and copy number variation (CNV) using technologies of DNA sequencing. Transcriptomics studies all RNA transcripts, including mRNAs and non-coding RNAs. The commonly used techniques are microarray and RNA sequencing (RNA-Seq). Epigenetics studies epigenetic modifications such as DNA methylation and histone modification on the genome. These modifications are heritable traits that affect gene activity and expression. These multiple levels of gene regulation interplay and lead to complex phenotypes, as shown in Figure 1.1, where we illustrated five gene regulation mechanisms (in reality there are much more) that happen within or between omics levels:

1. Single-nucleotide polymorphism (SNP). It is a variation in a single nucleotide at specific positions of a genome. It is present to a certain degree in a population and it underlies differences in our susceptibility to disease.

2. Epigenomic regulations such as DNA methylation and histone modification. The former adds methyl groups to the DNA molecule and typically acts to repress gene transcription. The latter contributes to the accessibility of genes for transcription.

3. Alternative splicing. On the transcriptome level a precursor mRNA can produce different mRNAs, which is regulated by trans-acting splicing activator and splicing repressor proteins. In this way, a single gene can code multiple proteins.

4. miRNA-related gene silencing. miRNA-related gene silencing: miRNAs bind to the target mRNA via base-paring to silence mRNA molecules by mechanisms such as cleavage and destabilization.

5. Transcription factor regulation. Transcription factors are proteins that bind to a specific DNA sequence (TFbs: transcription factor binding site) to control the transcription rate of genes.

6. The enzymes catalyze all catabolic pathways to break nutrients into cellular building blocks, produce energy and heat; and all anabolic pathways to synthesize complex molecules from simpler ones using energy.

It is not hard to acknowledge that biological systems are complex and dynamic, considering the gene regulations that happen at multiple Omics levels simultaneously. Therefore, heterogeneous molecular data, as mentioned above, can offer tremendous insights into the molecular changes in cancer pathogenesis. It is also necessary for building robust models to predict clinical outcomes. Thanks to technological advances, obtaining multiple levels of omics data becomes cheaper and more feasible. As one of the large-scale projects, the Cancer Genome Atlas (TCGA) project has generated genomic, transcriptomic, epigenomic, and proteomic data from more than 11,000 cases in 33 cancer types and subtypes [271].

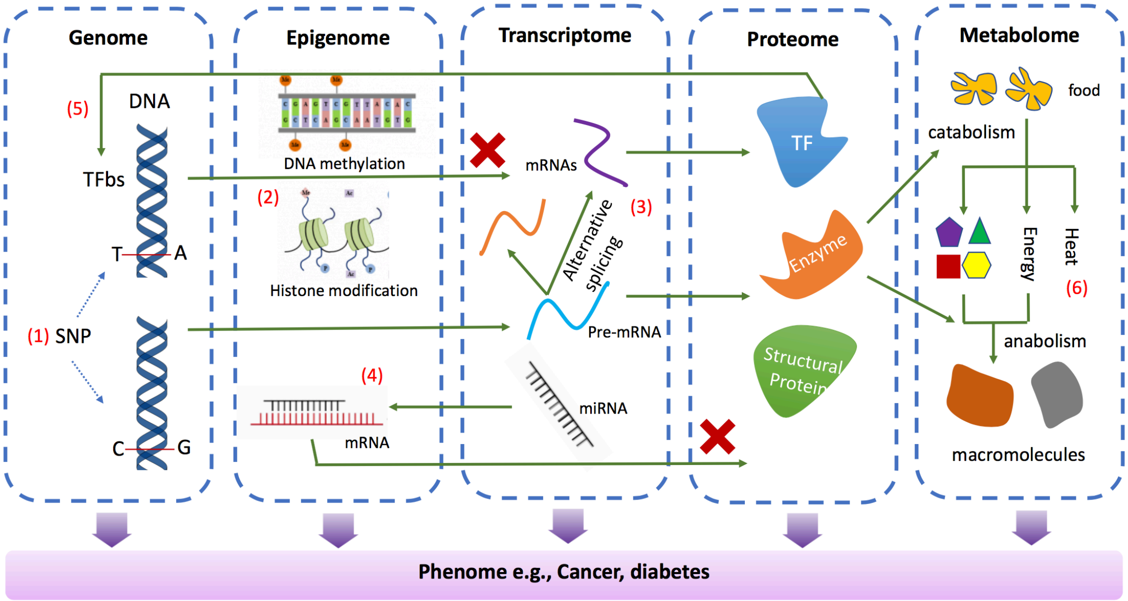


Figure 1.1: Multiple -omes interact with each other and jointly contribute to the phenome.

The data have been used for different applications such as prognosis prediction [303], cancer subtypes classification [30, 305] and cancer driver genes identification [305].

In prognosis prediction with omics data, machine learning is a widely applied technique to generate predictive models. Machine learning can be generally understood as learning from data using computer programs. It follows the procedure of induction and deduction. In the inductive step it learns the model from data (training set) and in the deductive step the model is employed to make predictions on new data (testing data). In data-driven applications, it is often a very good alternative to manual model construction, which is in some cases very time consuming or not feasible. The commonly accepted operational definition of a machine learning program is given by Tom M. Mitchell ”A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E” [172]. It shows that the computer program captures the patterns in the existing data so that it can generalize in unseen data. We would like to give an example of spam filter here to convey this idea. The task T is to classify emails into two classes: spam or not spam. This is a type of machine learning named *classification* and it is the most mature and most often used type. A simple classifier is illustrated in Figure 1.2. Given a good amount of spam and normal emails, there are mainly five steps:

1. Decide how to represent each email as a feature vector. We build a representative word list from a dictionary and use a Boolean vector to represent each email in terms of whether these words are present in the email.

2. Extract the Boolean features from all emails to generate the training set.

3. Train a decision tree classifier using these features where the leaf nodes give the most likely labels for the emails that reach the nodes.

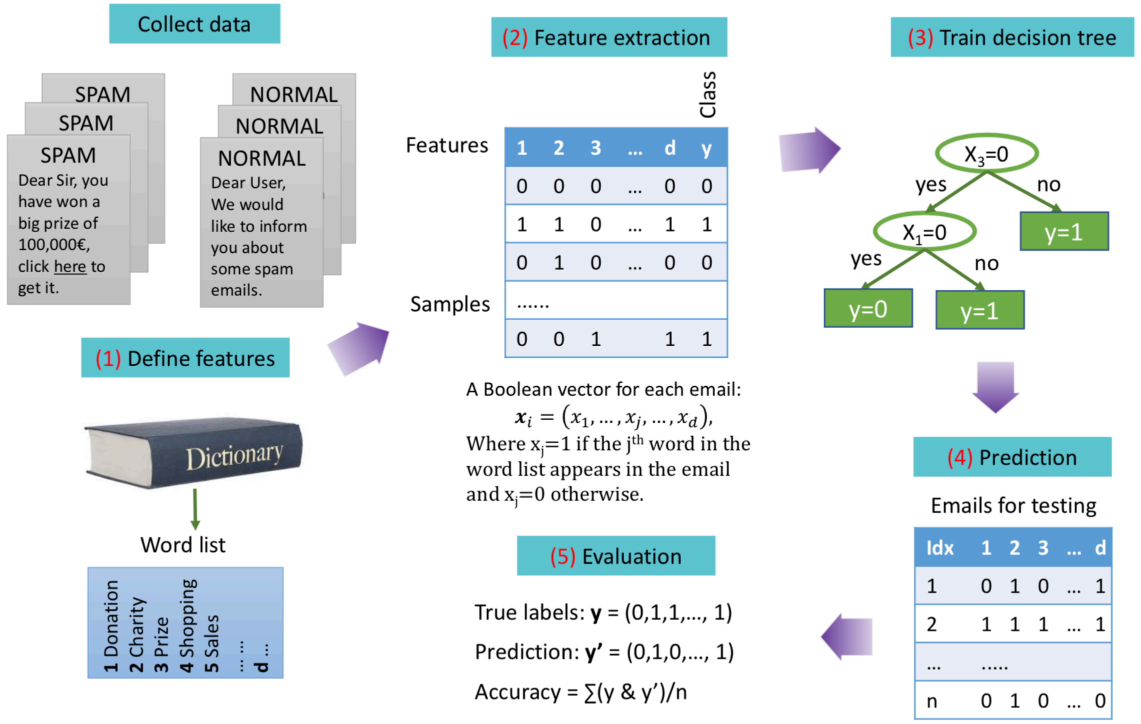


Figure 1.2: A simple machine learning application - spam classifier.

4. Test the classifier on new emails by extracting the same Boolean features and predict their labels using the decision tree.

5. Compare the prediction with the true labels to evaluate the classifier.

By its definition, machine learning applications depend on data, specifically, the amount of data and the quality of features. In extreme cases, when there are no data or no relevant data, it is infeasible to train a machine learning model. Compared with the application of classifying emails, classifying cancer patients into good and poor prognostic groups is more difficult due to the facts listed in Table 1.1. The class labels of patients are not so straightforward as emails. This is because usually one does not have the follow-up information for all samples. This happens when the patients withdraw the study or the event (death) has not occurred at the end of the study so their prognostic data are only partly known. In this case, right-censoring is present. If right-censoring happens in many samples, e.g., due to a short follow-up time, it can affect the reliability of analysis. In studies that try to identify prognostic biomarkers, censoring is usually handled in two ways, either by dichotomizing the samples into good and poor prognostic groups based on a threshold or by doing survival analysis that can take into account the censoring information. The feature noise is common in Omics data due to batch effect, which refers to technical sources of variation that have been added to the samples during handling. Factors include different experiment times, handlers, reagent lots, etc [81]. It can be detected and adjusted using batch effect correction algorithms [32, 144, 234]. The most

Table 1.1: Comparison of the applications of spam filter and cancer prognosis prediction.

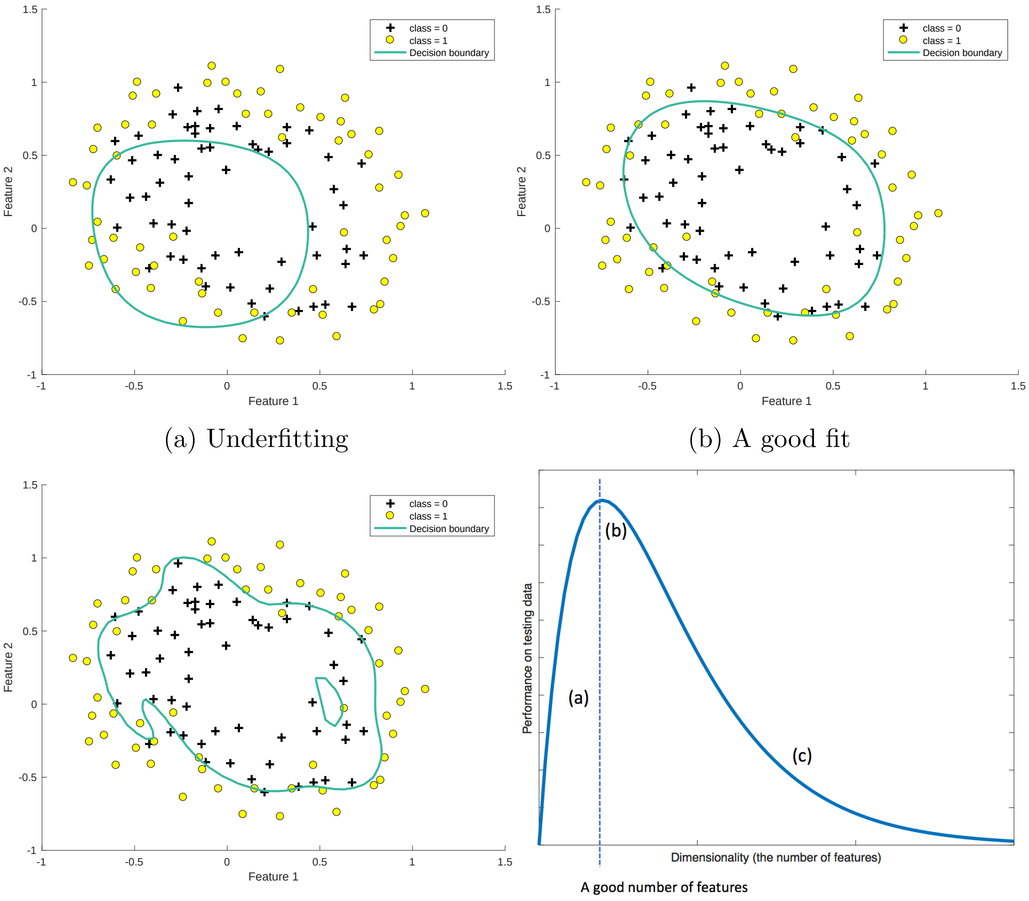
	Spam filter	Prognosis prediction
# instances (n)	millions	a few hundred
dimensionality (# features p)	Several thousand p <<n	Around 20,000 features (genes) p >>n
# class noise	avoidable	censoring of clinical data
# feature noise	hardly any	common

critical issue for building predictive prognostic model is, however, the high dimensionality of data, where the number of features overwhelms the number of samples (p » n). This brings lots of concerns in fitting machine learning models.

When the data lie in high-dimensional space, analyzing and organizing the data become challenging. It is because the sample density decreases exponentially with the increase of the dimensionality. For example, suppose we have 100 samples and the 1D space has a width of 50 unit intervals. The sample density is 100/50 = 2 samples/interval. The samples can completely cover the space. In the 2D space of 50*50 = 2500 unit squares, the sample density becomes 100/2500=0.04 samples/unit square. In the 3D space, spreading the samples onto the 50*50*50=1.25e5 unit cubes yields a sample density of only 100/125000 = 8e-4 samples/unit cube. If we keep increasing the dimensionality, the samples lie exponentially sparser in the space. We know that statistically sound analysis requires certain levels of sample density. When the data is too sparse in high-dimensional space, it is problematic for any method that requires statistical significance. It not only applies in machine learning, but also in numerical analysis, combinatorics, optimization, etc. These phenomena are in general called *the curse of dimensionality.*

In machine learning, each sample is represented by n features (variables). The word *feature space* refers to the n-dimensions where the features live. The high dimensionality of the feature space can lead to undesired consequences such as overfitting. Here we would like to give the intuition with the following example. Suppose we have 100 samples in the training set to train a classifier. When the dimensionality is low, with 2 features, the samples cannot be separated well and thus the patterns in the data are not recognized. This is called underfitting. As more features are added, the samples are described in higher dimensional space, one can find the hyperplane to separate the samples better. The underlying patterns in the data are captured. Note that when the dimensionality further increases, e.g., by constructing more features such as taking the polynomials, it becomes much easier to find a hyperplane that can separate the training samples perfectly into two classes (it is know that any dataset that lie in N dimensional space can be separated in N-1 dimensions). However, this may not bring the desired benefit. If we project the samples in low, moderate, and high dimensional spaces back to a two-dimensional space, the decision boundaries typically look like the ones shown Figure 1.3. It is clear that the trained model in high dimensional space overfits the data. In other words, the model is overly complex. It shows that the classifier can no longer differentiate between exceptions and

[0] Note that Figure 1.3 is only for giving an intuition of how the dimensionality may affect the performance, in reality machine learning has much trial and error. One can tune parameters and use techniques such as regularization to decrease overfitting to some extent.



(a) Underfitting

(b) A good fit

(c) Overfitting

(d) Performance

Figure 1.3: An illustration of underfitting, a good fit, and overfitting in terms of the dimensionality of the feature space.

the appearance of certain training samples and the real patterns. This needs to be avoided because it makes the model sensible to minor fluctuations in the training data and results in poor predictive performance on the testing data. Depending on the underlying model of the machine learning algorithm, e.g., linear or non-linear, the overfitting can happen earlier or later with increasing number of features. Note that the performance of the classifier is evaluated by how well it can generalize on the testing set and not how well the classifier can fit the training data.

Mathematically, theoretical analysis shows that distance metrics which are used to measure the similarity between samples show different behavior in high dimensional space than that in low dimensional space [3]. With L_k distance metric, [19] proves the following theorem:

$$\lim_{d\to\infty} \frac{{D_{max}}_d^k - {D_{min}}_d^k}{{D_{min}}_d^k} \to 0,$$

where d is the dimensionality of the feature space. It shows that the ratio of the difference between the maximum and minimum distances to the origin, and the minimum distance, tends to be zero. One can imagine that on a high dimensional hypercube, most of the data points lie on the corners of the cube and far from the origin. Therefore, distance measures lose their effectiveness. Since distance metrics are widely used in machine learning algorithms, high dimensionality makes it difficult to train machine learning models.

It is safe to say that the number of available samples limits the number of features to use without overfitting. With p » n, omics data are far from ideal for training machine learning models. If we use the data directly, the classifier will not able to find the true patterns, thus generalizing correctly becomes exponentially hard or even impossible, since the samples cover only a dwindling fraction of the feature space. The trained models will then give more random predictions. However, this does not put training machine learning models on omics data to an end because not all the features are necessary to use. In cancer biology it is acknowledged that only a small fraction of these features (the true signals) contribute to the phenotype of interest. It we can find these important features out of all features and train a predictive model with only these features, it can help prevent overfitting and lead to more interpretable models. This process is named *feature selection* and the identified feature subset is called molecular signatures. It is supposed to represent the footprint of the phenotype so that it can be used to predict the phenotype of new samples given their signatures. Selecting robust molecular signatures from Omics data, however, is challenging. Many feature selection algorithms have been proposed in the past decades. However, the robustness and the predictive capability of the biomarkers are not yet satisfactory [16, 171, 258]. In the next section we are going to have an overview of existing feature selection algorithms, many of which are proposed for identifying cancer prognostic signatures.

1.2 Methods for Identifying Molecular Signatures

Identifying disease signature (biomarkers) from omics data is one of the major endeavors in systems biology. Predictive, robust, and interpretable molecular signatures are generally considered as an important step towards personalized medicine. Since the emergence of gene expression microarray techniques, methods have been developed to find disease signatures - the set of genes whose expression values are indicative of a phenotype. The aim is to improve our understanding of disease by finding disease driver gene and help develop effective models to predict disease outcome [59,113,281]. From a machine learning perspective, identifying disease biomarkers is formulated as a feature selection task. The high-dimensional data is recorded as a matrix $X^{n \times p} = \{x_{i,j}\}$ containing the information of p genes in n samples, where $x_{i,j}$ is the expression level of gene j in sample i and $p \gg n$. In a classification problem, the n samples originate from different phenotype groups, which are denoted by a target variable $y = \{y_1, ..., y_n\}$. The entries of y are either 0 or 1. Feature selection is to find the subset of genes across all samples $S^{n \times k} \in X^{n \times p}, k \ll p$, which can best discriminate the samples into their groups. Unlike other dimensionality reduction techniques that are based on projection (e.g., principle component analysis) or compression, feature selection techniques do not alter the original features, but only select a subset of them. Thus, it offers the advantage of interpretability of the molecular signatures.

Finding the optimal feature subset to give the best prediction performance is NP hard

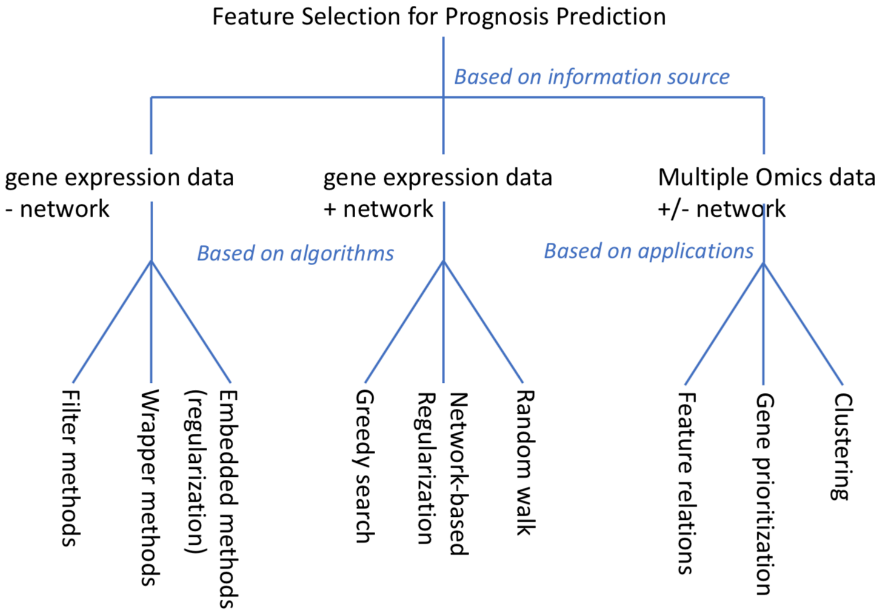

Feature Selection for Prognosis Prediction
Based on information source
gene expression data - network
gene expression data + network
Multiple Omics data +/- network
Based on algorithms
Based on applications
Filter methods
Wrapper methods
Embedded methods (regularization)
Greedy search
Network-based Regularization
Random walk
Feature relations
Gene prioritization
Clustering

Figure 1.4: An overview of feature selection algorithms for prognosis prediction. The methods are first divided into three broad categories based on their information sources. In each category, we give three representative branches based on the algorithm or application.

(nondeterministic polynomial-time hard). In bioinformatics, different algorithms are proposed to improve the prediction accuracy of the machine learning models trained using these features. Since the early development on mainly microarray data, the methods have been developed from using single type of omics data (typically gene expression data) to the integration of data with biological network, and to the integration of multiple types of omics data, thanks to the accumulation of genetic regulation knowledge and improvements in sequencing techniques.

Following this trend of increasing integration of biological knowledge and multiple data sources in biomarker discovery (and more or less in ascending chronological order), we would like to give the introduction in three parts according to whether and how data are integrated for feature selection. The first part involves methods that use only one type of omics data without network. The data is usually gene expression data. The second part introduces methods that integrate gene expression data with biological network. The third part introduces studies that employ multiple types of omics data (with or without biological network), with the goal to compare different omics data, cluster samples or perform predictive tasks. In each part we will give the motivations, some representative methods, and the potential advantages and drawbacks. For advanced readers, Figure 1.4 gives a high-level summary of the methods included. In the following text, we will use the word feature in the context of machine learning, and use the corresponding terms of biological entities, e.g., genes, molecules, when we talk about applications.

1.2.1 Feature Selection on Single-omics Data

This part concerns the earliest and the more frequently used feature selection algorithms. Features are selected only based on the intrinsic property of the data without considering their biological roles. Gene expression microarray (GEM) is the earliest high-dimensional omics data that captures expression level of many genes simultaneously. Most of the earlier feature selection studies used GEM data. As is generally accepted, these methods are divided into three categories: filter, wrapper, and embedded methods [101, 145, 209, 237]. They differ from each other in how they search in the space of feature subsets:

- Filter methods select features by their statistical and information theoretical measures without the use of machine learning model. Based on a predefined criterion, such as t-statistic, feature relevance scores are calculated for individual features with a higher score indicative of a more important feature. The features are then ranked by the scores and top ranked features are selected. A few other commonly used criterion include chi-squared statistic, the p-value of univariate cox proportional hazard model, information gain, and Fisher score [88]. Most of the filter methods are univariate. Several multivariate methods also exist, such as correlation-based feature selection [95, 295], Markov blanket filter [135], etc.

- Wrapper methods utilize a specific classifier to evaluate the quality of feature subsets, in order to guide feature selection. Wrapper methods generate various feature subsets and evaluate them by training a classifier and testing its performance. There are two types of wrappers: deterministic and randomized. The former uses search techniques such as sequential forward selection, sequential backward selection. The added or eliminated features in the previous step are not changed in later steps [200]. The latter often uses genetic algorithms and simulated annealing to obtain feature subsets [195, 208].

- Embedded methods embed feature selection with classifier construction. They can be divided into three types [237]: 1) pruning methods such as recursive feature elimination [90] 2) build-in mechanisms for feature selection, typically for tree-based classifiers and 3) regularization methods, which are shown to give good performance and are increasingly employed [159].

Filters are simple, fast, and independent of classifiers. They can easily handle high-dimensional data. Sometimes it is used as an initial gene selection step before more complicated feature selection algorithms are applied [77, 146, 255]. Despite its simplicity, it is shown to be able to achieve comparable or even better performance [97, 106, 215]. The disadvantage of filter methods is that feature dependencies are usually ignored, which may result in good features but worse classifier performance. For example, highly correlated features have been shown to undermine the stability of classifiers [251]. Another disadvantage is that it ignores the effect of selected features on the performance of the classifier, which is dependent on the biases and the heuristics of the classifier [97]. For example, the performance of Naive-Bayes classifier improves with the removal of irrelevant features, but not every filter is able to select the relevant features.

Wrappers can take into account feature dependencies and avoid the representational bias of the classifier. This renders the selected features to be specific for a classifier.

However, as the dimensionality of the data grows, the space of feature subsets grows exponentially. Wrapper methods become much more expensive. In this case, randomized search strategies are often applied instead of deterministic search. Compared with filter methods, wrapper methods are prone to over-fitting, especially if the classifier is complex. [157] shows that the more exhaustively the feature subset space is searched the greater the likelihood of finding a feature subset that has a high training accuracy while generalizing poorly. It is necessary to control the depth of search, and use strategies such as post-pruning of decision trees, adding noise to the training data [134], and early stopping [157] to decrease overfitting.

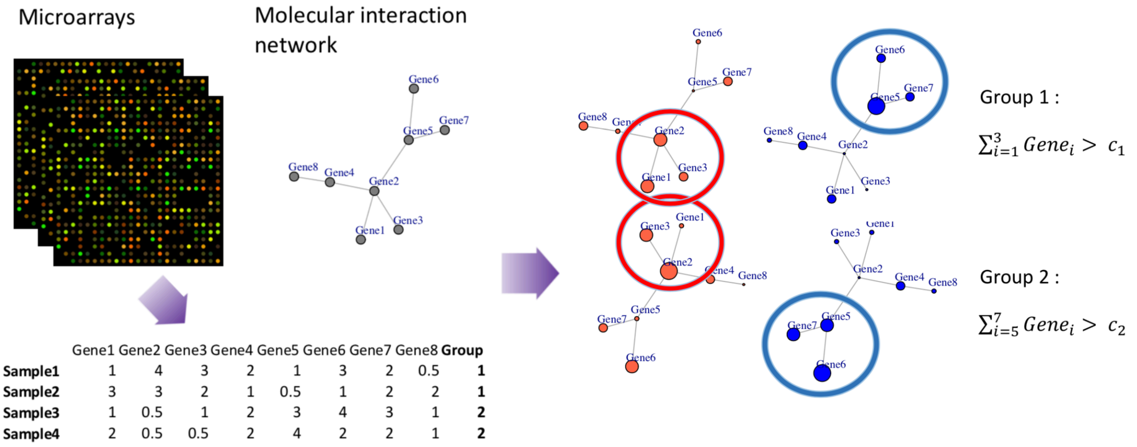
Embedded methods address the disadvantages of both filter and wrapper methods. It includes the interaction with the classification model and is far less computationally costly than wrapper methods. This is achieved by having an objective function that minimizes both the fitting error of the machine learning algorithm and a penalization term that is used to shrink the model coefficients. Taking the example of a regularization method with a linear classifier $\mathbf{w}$ and its objective function $c(.)$, an embedded method takes the form:

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} c(\mathbf{w}, \mathbf{X}) + \alpha\, penalty(\mathbf{w})$$

where $penalty(\mathbf{w})$ is the regularization term and α is the regularization parameter controlling the trade-off between $c(.)$ and the penalty [237]. For example, a well-known regularization algorithm Lasso [250] penalizes the l_1-norm of the coefficient $\mathbf{w}$: $penalty(\mathbf{w}) = \sum_{i=1}^{p} |\mathbf{w}_i|$. Lasso method is very widely applied [159] on high-dimensional omics data, although its consistency cannot be guaranteed when certain conditions are not met [302]. Regularization-based feature selection has attracted attention in recent years. The different combinations of model fitting and regularization terms can give rise to different embedded feature selection methods, such as group Lasso [114], bridge regularization [108], regularized Cox regression [224], regularized decision trees [56, 57], etc.

Although many algorithms are proposed and can be easily applied as off-the-shelf tools, it is still difficult to identify robust biomarkers. This is because while it is easy to identify a set of features that fits the training data very well, it is hard to find features that generalize well on independent data. Due to the properties of the data shown in Table 1.1, a good generalization is very hard to achieve. Indeed, as more molecular patterns are reported, the low reproducibility of many reported gene signatures has been criticized. It was noticed that there is hardly considerable overlap among biomarkers identified in different studies for the same disease, and biomarkers identified using one dataset may not work well on other datasets [66, 67, 171]. It is shown that there exist many feature subsets that perform similarly well, making it difficult to find the true signals. Even taking random sets of features from the entire feature space is shown to give comparable prediction performance [258]. To tell which set of features can generalize well seems to be very puzzling.

Another challenge is to produce feature sets that have good biological interpretation to assist biological research [62,66]. Given multiple sets of features that do not have significant difference in their prediction performance, it is impossible to tell from the data alone which set of features is more biologically meaningful, e.g., more relevant to cancer progression. In another scenario when a set of features does have good prediction performance, but the features have no relations with each other in gene regulations, it would also be reluctant


Microarrays
Molecular interaction network
Gene1
Gene2
Gene3
Gene4
Gene5
Gene6
Gene7
Gene8
Group 1 :
$\sum_{i=1}^{3} Gene_i > c_1$
Group 2 :
$\sum_{i=5}^{7} Gene_i > c_2$

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	Gene7	Gene8	**Group**
Sample1	1	4	3	2	1	3	2	0.5	**1**
Sample2	3	3	2	1	0.5	1	2	2	**1**
Sample3	1	0.5	1	2	3	4	3	1	**2**
Sample4	2	0.5	0.5	2	4	2	2	1	**2**

(a) Data-network mapping

(b) Subnetwork features

Figure 1.5: An illustration of the integration of gene expression data with molecular network. (a) shows the correspondence between features and network nodes (b) shows subnetwork features - the aggregation of features in red and blue circles can better differentiate between two sample groups.

to tell why this set of features is a set of good disease signatures. This is very likely to happen given the noise and the sample heterogeneity in omics datasets. The typical low reproducibility of molecular signatures and the difficulty to interpret them necessitate the integration of domain knowledge in biomarker discovery [169]. Various feature selection methods that integrate biological knowledge have been proposed and we are going to address it in the next part.

1.2.2 Network-based Feature Selection on Single-omics Data

Biological knowledge has been applied to assist feature selection in a few directions such as the integration of biological network or pathways, Gene-Set Enrichment Analysis (GSEA), and text mining. GSEA is a method to identify whether classes of genes (e.g., based on gene ontology terms) are statistically over-represented or deleted in a set of genes. This can be associated with disease phenotypes. It is usually used to help assess the potential biological indications of a set of features. Text mining employs natural language processing techniques to relate basic biomedical research to clinical practice, e.g., to find potential associations among genes from literature that are very time consuming for humans. Most of the effort so far has been made in network or pathway-based feature selection where biological knowledge is represented as a network of genes. The networks of interacting molecules have been placed between genotypes and phenotypes in systems biology and it is acknowledged that the properties of the network as a whole determine the phenotypes [17]. The underlying motivation is that the features (genes) are not isolated (independent) but regulate each other in a network (dependent) via various mechanisms. This property should be considered in a way so as to find features that act jointly to contribute to the phenotype.

Figure 1.5a shows intuitively how the omics data and network relate to each other.

It shows that molecular profiling of a group of patients can be put to a table, where the rows correspond to different patients and the columns correspond to molecules. If one only looks at the table and try to identify the signatures, it falls into the methods in section 1.2.1. However, the features and the nodes in a gene network have correspondence. It is very likely that the role that a gene plays in the network has something to do with the importance of the gene and its interactions with other features. By taking into account both the features and the underlying network structure, various feature selection methods have been proposed. Our scope is within the studies that try to identify molecular signatures for making predictions and does not include the studies that only identify active modules without building predictive models. While it is difficult to mention all important methods, we summarized three representative categories based on their methodologies.

The first category is network structure guided searching. This group of methods aims to identify subnetworks, instead of single nodes, that can best differentiate phenotype groups, as shown in Figure 1.5b. Each of the identified subnetworks is aggregated to produce one feature (the *metagene*) and these metagene features are used for training predictive models. The search is guided by a predefined scoring function which takes the input of a subnetwork (several connected feature vectors) and outputs a real value, which is used to rank the subnetwork. Because finding the maximal-scoring connected module is NP-hard, greedy search is frequently applied by starting at each network node and adding its neighbors incrementally until constraints are not satisfied. For example, the increase of the score does not meet a predefined threshold. Different methods can differ in their scoring functions, how metagene was calculated from individual genes, and the biological networks used. For example, [40] used mutual information as the scoring function and the addition operator to aggregate subnetworks. [152, 167] used the p-value of Cox PH model in defining the scoring functions. [190] dichotomized features and defined a scoring function based on information theory. [8] tested different aggregation operators on their effects on the prediction performance.

The second category is network-based regularization. Recall the regularization methods in section 1.2.1. where a penalty term is included to avoid overfitting, here the penalty term additionally takes into account the network structure. Adjacency matrix A and Laplacian matrix L are frequently used to represent a network G to be included in the penalty term. Since the majority of the methods in this category are based on linear classifiers, it can be written in the following form:

$$\hat{\mathbf{w}} = \min_{\mathbf{w}} c(\mathbf{w}^T\mathbf{X}, Y) + \alpha\, penalty(\mathbf{w}, G)$$

where G denote the network structure. For example, in graph Lasso $penalty(\mathbf{w}, G) = \lambda||\mathbf{w}||_1 + (1-\lambda)\sum_{i,j} A_{i,j}(\mathbf{w}_i - \mathbf{w}_j)$, which forces connected nodes to have similar weights [237, 287]. Using similar formulations, [149] proposed a network-constrained regularization and feature selection methods on genomic data. [299] added a network regularization term to the log-likelihood function of the Cox proportional hazard model. [35] developed a network-constrained support vector machine method, where the network-based regularization term is added to the objective function of SVM.

The third category of methods involves iteratively updating node importance scores on the network. Frequently used methods include network propagation and random walk. [175, 274] adapted Google's PageRank algorithm to rank genes in a network. PageRank is used to rank the importance of web pages based on the links among pages. Pages are assigned an initial rank $\mathbf{r^{[0]}} \in \mathbb{R}^N$. This rank is updated iteratively depending on the rank of pages that are linked to it. For page j, its rank from $r_j^{[n-1]}$ to $r_j^{[n]}$ is updated according to the formula:

$$r_j^{[n]} = 1 - d + d\sum_{i=1}^{N} \frac{A_{i,j} r_i^{[n-1]}}{deg_i}, 1 \leq j \leq N$$

where deg_i is the degree of the ith page and d is a fixed parameter. By iterating until convergence, a page will be highly ranked if it is linked to other highly ranked pages. This process has an alternative interpretation in terms of random walk theory [142]. [49] used random walk kernel to smooth gene-wise t-statistics over the network. This is achieved by assigning each node an initial score based on t-test and then multiply it with the random walk kernel. The p-step random walk kernel is used as a similarity measure to capture the relatedness of two nodes in the network. It is defined as:

$$K = (\alpha I - L^{norm})^p = ((\alpha - 1)I + D^{-1/2}AD^{-1/2})^p$$

where L^{norm} is the normalized graph Laplacian matrix, α is constant, and p is the number of random walk steps. The network-smoothed t-statistic $\tilde{t} = t^T K$ is used to measure node importance. Similarly, random walk-based scoring of network components is also applied in [136] to prioritize functional networks. [207] used network propagation method to score genes based on their proximity to mutated genes and differentially expressed genes in the network. The method is shown to be able to better prioritize known cancer driver genes and predict potential cancer genes. Both mutation and expression data are mapped on the network to jointed identify important genes.

Having discussed about the feature selection methods, we think it is also necessary to introduce the molecular interaction networks. The network mostly comes from biological knowledge on gene regulations, protein-protein interaction, etc. Different networks have been used, as listed in Table 1.2. Given the variability of options, it seems that before choosing which feature selection algorithm to use, one of the first step is to consider which network, or which combination of networks to use. Apparently, the results are network dependent. Intuitively, one may choose to use as much network information as possible because it enriches the information. However, it meanwhile increases the search space, and this makes it harder to differentiate signals and noises.

It is shown in many studies, e.g., from Table 1.2, that features selected based on a network outperform features that are selected based on the data alone, in terms of their prediction performance and biological interpretation. There was at one point a significant increase in network-based feature selection methods. However, in recent years, several studies show that network-based feature selection methods do not always show better prediction performance, but mainly improve the biological interpretation of the signatures [48, 228, 229]. [48] compared fourteen published gene selection methods, including 6 single gene methods and 8 network-based feature selection methods, on six breast cancer datasets with respect to prediction accuracy, signature stability and biological interpretability. The biological interpretability is measured by enrichment analysis of disease related genes, KEGG pathways and known drug targets. They found that network-based features

Table 1.2: Frequently used molecular/gene interaction networks in network-based feature selection studies. We list the networks, their sizes and information sources, as well as exemplary studies that use the networks. With STRING database, we only considered the edges with confidence score ≥ 0.9. Note that when the database includes many species, only Homo sapiens is considered. *Data* means the network size is dependent on the dimension of data. *App* means the network size is dependent on the application.

Idx	**interacting structures**	**Database**	**Version**	$\|\boldsymbol{E(G)}\|$	$\|\boldsymbol{V(G)}\|$
1	Protein-protein	STRING	v10.5	547621	19578
2	Protein-protein	HPRD	Release 9	41327	30047
3	Biological pathways	KEGG	Release 84.0	App	App
4	Biological pathways	Pathway Commons	v7	1912848	14863
5	miRNA-gene	miRTarBase	v7.0	502651	16822
6	transcription factor - target	TRANSFAC	v7.0 public	-	1648
7	Gene co-expression	None	None	Data	Data
8	Gene Functional linkage	Multiple	None	App	App
9	Gene ontology	Gene ontology	None	GO	GO

Idx	**Information**	**Studies**
1	binary PPIs, including predicted interactions	[51, 91, 118]
2	binary, and complex PPIs, PTMs, protein-DNA interactions	[35, 152, 167, 190]
3	Very broad, including systems information, genomic information, chemical information, and health information	[149, 215]
4	Biochemical reactions, gene regulations, interactions involving proteins, DNA, RNA, and small molecules.	[176]
5	microRNA-target interactions	[118]
6	Transcription factor binding sites	[111, 274]
7	Pearson correlation	[175, 299]
8	Multiple database resources	[40, 277, 299, 300]
9	Gene ontology	[175]

in most cases cannot improve prediction accuracy significantly but can improve the interpretability of gene signatures. [229] tested single gene and network-based algorithms on six breast cancer datasets in predicting breast cancer prognosis. They also found out that the composite feature classifiers do not outperform single gene classifiers. What's more, the randomization of the network structure, which destroys the biological information, does not result in a deterioration in the performance of composite feature classifiers. [228] extended the experiments in [229] by including more gene signatures and drew consistent conclusions. [228, 229] also show that when a proper correction of feature set size is performed, the stability of composite features is not higher than single gene features. Based on their experiments, the argument to prefer network-based features over single gene features does not hold true.

It now seems controversial whether network-based methods are superior to feature selection methods without network. In the first sight, this may seem puzzling - because by integrating biological network we give more relevant input to the algorithms and normally the outcome should be better as well. In our opinions, there could be several reasons that contribute to the puzzle.

1. There is no exact mapping from gene expression data to PPI network. Ideally, one would map protein expression value to the protein network, or the corresponding functional molecules to a molecular interaction network. For example, in the case of transcription factor - DNA interaction, the entities are protein and gene expression, but instead, only gene expression values are used. In fact, it is reluctant to represent protein levels by gene expression levels as there are several steps from gene expression to a functional protein, which are regulated by multiple mechanisms (as will be mentioned in Chapter 5.

2. There is not enough investigation in meta-gene operators - how individual gene expression levels are aggregated into composite features. A recent study [8] has addressed this issue by evaluating multiple operators to summarize genes into metagenes and choose the best operator. This significantly improves the stability of the signatures. Since there are still no available rules of how to combine features based on a biological network, this area potentially needs more efforts.

3. The curse of dimensionality is present in both no network methods and network-based methods. As there are many irrelevant signals in the omics data, the signals are also present in the network when the data are mapped to it. When the network is, e.g., searched from each node alternatively, many irrelevant nodes could be selected. Due to the high-dimensionality and heterogeneity of samples, which create great difficulty to differentiate signals and noise, the performance difference of single gene features and composite features cannot be determined.

While the first issue cannot be resolved at the moment (as will be mentioned in Chapter 5) and the second issue could be explored by trail-and-error, we think the last issue is by far the most critical one, which hinders the advantages of data integration to be seen in feature selection. As observed from both studies in section 1.2.1 and section 1.2.2, when different algorithms are compared on multiple datasets with the same experimental settings, the conclusions can be different than the studies where the algorithms were proposed. This shows that the algorithms are sensitive to changes in the data and networks. It resembles overfitting, where the information (both signals and noise) is overwhelming for the small set of samples at hand and the feature selection algorithms are not able to pick the signals. As discussed previously, to have good performance we need to reduce the dimensionality of data and select features with biological significance.

Some researchers have approached this question by integrating information from multiple omics data sources. The motivation is that these different sources of information depict the multiple dimensions of gene activity and thus can complement each other, although the dimensionality is not necessarily reduced. While selecting molecular signatures using only one type of data can have much bias, taking into account multiple omics data can better prioritize candidate genes [203]. In recent years, data integration-based feature selection has attracted more and more attention and it will be introduced in the following part. Due to the limited number of methods compared with the previous two method categories, we will not limit our introduction to feature selection in prognosis prediction, but also include methods that use multiple omics data in other relevant tasks such as patient clustering, disease gene prioritization, etc.

1.2.3 Multi- and Integrative Omics Data Analysis

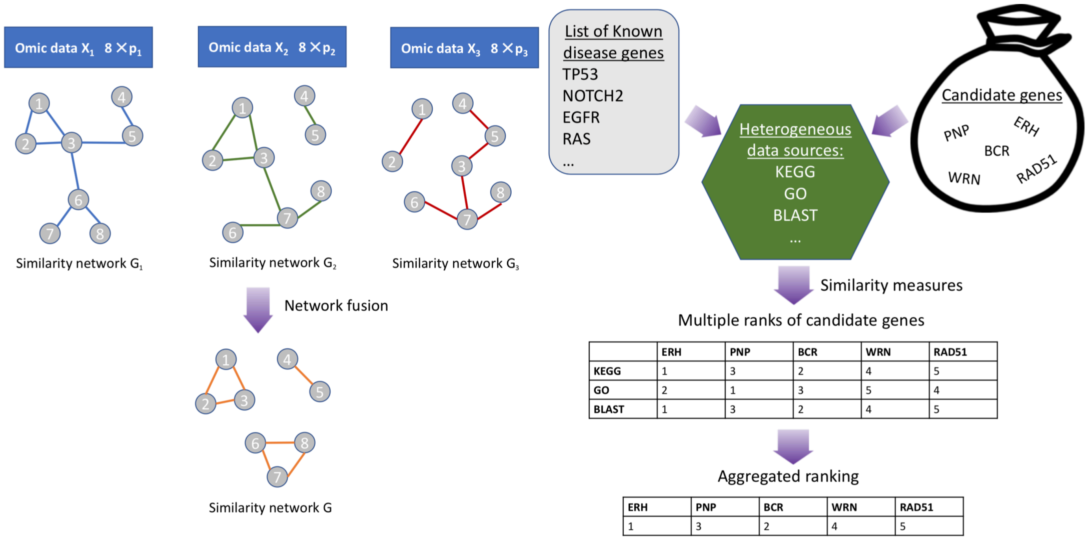
It is demonstrated that the molecular portrait of a tumor manifests at multiple omic levels [181]. The typical focus on one single omic level at a time thus can only explain a modest portion of complex disease. Thanks to the recent efforts in collecting multi-omics data in the same groups of individuals, integrative analysis has been developed to study complex diseases in a more comprehensive manner. For example, besides finding molecular patterns from one type of omics data, which can have certain bias, multiple types of omics data can be considered simultaneously to find more robust disease signatures. Later in Chapter 4 we proposed to use a multiplex structure to selected mixed features from multiple omic layers. In fact, for many questions in disease research, such as how genomic changes affect genetic pathways that drive cancer phenotypes, it is necessary to simultaneously look at multiple molecular data levels. In this part, we introduce some studies that employ multiple omics data. We decide not to limit our scope to cancer prognosis prediction studies (the number of studies is by far limited), but also include studies that use integrative approaches for other relevant tasks such as patient clustering, gene prioritization, etc.

A comprehensive study [296] compared the predictive capability of multiple levels of omics data: CNA, DNA methylation, mRNA expression, miRNA expression, and protein expression data, and their combination with clinical features, on four cancer types from TCGA. Using the same feature selection algorithms and classifiers, different omics data are used to train predictive models for prognosis prediction and their performance was evaluated by cross-validation. The results show that the prediction performance varies significantly across data types and cancer types while the effect of different machine learning algorithms is moderate. They show that clinical features in many cases outperform molecular features, while combining molecular features with clinical features can significantly improve the performance. [303] conducted a similar comparative study to compare the predictive capability of four types of omics data and their combinations. In addition, they evaluated the performance of the omics data combinations with clinical features. Their experimental results show that clinical features and mRNA expression features are the most informative. Adding other omics features does not give substantial improvement in predictions. [163] employed four omics data types (mRNA, DNA methylation, CNA, and miRNA) to identify prognostic signatures for serous ovarian cancer. Features are selected individually from these data types using regularized Cox model [188] and the selected features are combined to give integrative features. The integrated features were shown to outperform individual data in stratifying patients into significantly different prognostic groups. While these straightforward methods of combining multiple omics data are helpful and necessary, there are other studies that use more complex integration methods in different aspects of disease research.

One category of studies aims to cluster patients into different prognostic groups based on patient similarities defined on multiple omic levels. A simple illustration is given in Figure 1.6a. [128, 264] use similarity network fusion methods to cluster patients into different survival groups. [128] used the same four types of omics data and constructed a patients' k-nearest neighbor graph using each data type, where edges represent similarities and nodes represent patients. Then these four graphs are integrated using graph-based semi-supervised learning algorithms [252, 307]. They show that the integrative approach (using the best combination of model parameters) significantly outperforms individual data types. [264] proposed a network fusion algorithm on patient similarity networks constructed using mRNA expression, DNA methylation, and miRNA expression data. It iterative updates each of the networks with information from the other networks, making them more similar with each iteration and eventually obtaining the fused network. Based on the fused network, spectral clustering is applied to stratify patients into clusters. They tested this method on five cancer types and showed significant improvement in stratifying patients into different prognostic groups, compared with using individual data types. Similarity can also be defined on the pathway level. In [257], curated pathway interactions are incorporated to define the levels of pathway activities using different types of omics data. Grouping patients based on their pathway perturbations leads to subgroups that have significantly different survival outcomes.

While the similarity network clustering above is composed of separate clustering followed by network integration, there are methods proposed to incorporate all data types simultaneously and produce a single integrated cluster assignment. [227] applied unsupervised multiple kernel learning method on multiple omics data which reduces the dimensionality and meanwhile performs data integration. One or multiple kernels are generated from individual omics data and these kernels are linearly combined to give a unified kernel matrix. K-means clustering is applied on the unified kernel matrix to stratify patients into clusters. It is shown that these clusters have significantly different clinical outcome. [219] developed iCluster algorithm that used joint latent variable to associate different omic data types. The idea is to use the tumor subtypes vector as the latent variable that connects a set of models, which induces dependencies across different types of omics data. The algorithm is tested on breast and lung cancer samples to identify tumor subtypes. Compared with the patient stratification results in these studies, we will show in chapter 3 and 4 that using a much smaller but phenotype relevant biological network, one can achieve remarkably good patient stratification.

The second category of study aims to prioritize candidate genes using heterogeneous data sources. A simple illustration is given in Figure 1.6b. [2] developed Endeavour to



Multiple ranks of candidate genes

	ERH	PNP	BCR	WRN	RAD51
KEGG	1	3	2	4	5
GO	2	1	3	5	4
BLAST	1	3	2	4	5

Aggregated ranking

ERH	PNP	BCR	WRN	RAD51
1	3	2	4	5

(a) Similarity network fusion

(b) Gene prioritization

Figure 1.6: An illustration of integrating multiple omics data sources to improve predictions. (a) Multiple omics data are used to build patient similarity networks and these networks are fused into one similarity network. (b) Heterogeneous data sources are used to rank candidate genes based on their similarity with known disease genes. The individual ranks are aggregated into an overall rank.

prioritize genes that are involved in specific diseases or signaling pathways. The individual rankings using different data sources such as literature, functional annotation, and pathway membership are integrated into a single overall ranking using order statistics. This overall ranking is shown to significantly outperform individual rankings. [54] used kernel fusion technique to more accurately prioritize disease genes. They computed multiple kernel matrices based on different data sources to measure the similarities between candidate genes and diseases genes. These kernels are convexly combined to give an overall similarity matrix that can better capture gene similarities. [49] prioritized candidate genes for prognosis prediction using random walk kernel on a gene network. It first calculated the t-statistics of individual genes and miRNAs using prognostic information. Then a random walk kernel is used to smooth the t-statistics over the network structure. The details of the algorithm are given in Chapter 3. [36] proposed the method MAXDRIVER to identify potential cancer driver genes. It first constructed a fused gene functional similarity network where the edge weights are derived from protein-protein interaction, gene co-expression, gene sequence similarities, and pathway co-occurrence. Then this network was combined with a disease phenotypic similarity network and gene-disease associations to construct a heterogeneous network. Then an information flow method was applied on the heterogeneous network to find the relationships among cancers and candidate genes. This method was shown to be able to accurately rank known disease genes. [207] and [6] also used multiple data sources to discover potential cancer driver genes.

Methods have also been proposed for finding relationships among heterogeneous data sources. [158] proposed to find synergistic effect of DNA methylation and CNA on gene expression using statistical approach. They showed that for several oncogenes both hypomethylation and copy number amplification are present. [153] introduced a sparse multiblock partial least squares regression model to identify multi-dimensional regulatory modules from multiple omics data. It identified sets of features from gene expression data, DNA methylation data, and miRNA expression data that jointly contributed to the expression of a set of genes. They showed that the identified modules had significant functional enrichment and coupled impact on oncogenes. [155] predicted gene-phenotype relationships by using random walk method on a heterogeneous network. The network was constructed by connecting gene network (PPI) and phenotype network using a bipartite graph. The transition matrix was defined based on network connectivity. The steady state probability of finding the walker at each node was used to measure the proximity of network nodes to seed nodes. The results showed that compared with using random walk on gene network only, the accuracy of gene-phenotype link prediction was significantly improved with heterogeneous network. Besides the above discussed applications, heterogeneous data sources have also been applied in gene regulatory network (GRN) inference [308].

As we have seen, many studies integrate multiple omics data sources to better cluster samples, prioritize disease genes, and find the relationships between features of different omic levels. This is usually achieved by integrating similarity measures (kernels, random walk distance), using joint latent variables, and applying multiple statistical tests. However, so far, few methods have been developed to integrate multiple omics data sources for feature selection in cancer prognosis prediction. In addition, integrating multiple omics data sources and meanwhile utilizing biological network structure has not been well investigated. This is mainly because integrating several high-dimensional data will worsen the curse of dimensionality and make it more difficult to find the important features. Taking all features into account increases irrelevant information dramatically and thus diminishes the advantages of information enrichment. For network-based feature selection algorithms, finding composite features by mapping these data on a biological network would be hard to implement and evaluate.

1.3 Thesis Overview

1.3.1 Motivations

Biological systems are usually complex and high-dimensional by nature. It is recognized that the bottleneck for life science studies has shifted from generating data to the interpretation of data so as to derive insights into biological mechanisms [169]. When dealing with omics data, the problems of high-dimensionality and limited number of samples create great challenges [10]. How to select important features has been investigated since decades. Not only does it apply to the data, but also to the underlying gene interaction networks, where not all network components are relevant for the phenotype of interest. As introduced, feature selection algorithms have developed from using single omics data, typically microarray data, to the integration of omics data with biological network, and to the integration of multiple types of omics data and other biological information. In line with the innovation of network and multiple data resources integration, we identified some research gaps:

1. Biological networks are mostly mapped with only one type of omics data. It would be very interesting to know how the prediction performance differs when the networks are mapped with different omics data types alternatively. In addition, what are the relationships among the features that are selected using different omics data.

2. Data integration has not been realized directly on the networks. In our opinion, this is mainly due to the large network size, which makes this integration very costly. What's more, not all parts of the network are equivalent for the phenotype to predict. Usually a small fraction of the network plays a key role in a phenotype.

3. The integration of biological knowledge has not been detailed. The integration considers either the whole biological knowledge-base (e.g., a PPI network) or none. The specific information of which genes are important for a phenotype/disease that are discovered in biological research has not been utilized.

State-of-the-art studies have shown that feature selection, biological network integration, and multiple omics data integration are all important components for discovering robust cancer prognostic biomarkers. However, how to bring these elements under one framework for building predictive models remains to be investigated. Since the problem of high dimensionality hinders the discovery of robust molecular signatures, by far the potential clinical utility of the aggregate of these data remains largely unknown. In this thesis we propose a Phenotype Relevant Network-based Feature Selection (PRNFS) framework and demonstrates its superior performance with the use case of cancer prognosis prediction. Within this framework, we have successfully integrated four types of omics data with GRNs to identify robust molecular signatures. We show that significantly better prediction performance can be achieved using <2.5 % of the original dimensionality.

Concretely, we used biological knowledge to build a specific GRN for cancer prognosis prediction. We constructed an Epithelial Mesenchymal Transition (EMT) network by reviewing EMT literature and putting the genes and regulatory relationships to a graph. EMT process has long been demonstrated as closely related to cancer progression [245]. Many important EMT regulators such as SNAIL1 and SNAIL2 genes have been shown to correlate significantly with disease relapse and survival in patients with breast, colorectal, and ovarian carcinoma, where it is shown that EMT development can lead to poor clinical outcomes [96, 246]. In lung cancer, [214] demonstrated that mesenchymal gene expression signatures (genes that encode mesenchymal proteins) can divide early stage LUAD patients into significantly different survival groups. Since mesenchymal gene expression has a predictive value, their regulations are therefore important for prognosis. This is because the gene regulations in cancer progression have been widely acknowledged as a network consisting of multiple dysregulated pathways [14, 137, 240, 268]. The details of how the EMT networks are constructed are given in Chapter 2.

By using EMT networks and their composing molecules for feature selection, we aim to select highly relevant features for cancer prognosis prediction. As we narrow down the number of features dramatically, the dimensionality is reduced, and biologically important features are kept. On EMT networks, we mapped multiple types of omics data either alternatively or simultaneously to select single-omics features and multi-omics features. This is not only useful for building predictive models, but also for the analysis of molecular signatures. Since the process of EMT is complex and coordinately regulated by multiple pathways and regulatory levels, it is interesting to see which genes play important roles on different omics levels.

1.3.2 Thesis Outline

Finding important features from a reservoir of features is difficult. To ensure that the molecular signatures identified on one dataset will generalize well on other samples, it is necessary to guide the feature selection by using cancer domain knowledge. Currently, the domain knowledge - mainly protein-protein interaction networks, are either used as a whole, or not used at all. In depth domain knowledge has not been integrated with the identification of prognosis biomarkers. This lack of direction when integrating biological network thus could not be as fruitful as expected. This is supported by recent studies, which show that integrating biological network does not bring benefits in improving the prediction accuracy of molecular signatures. Based on these motivations, we aim to identify robust prognosis signatures using EMT networks. The thesis is organized into four parts:

Part I: Constructing EMT Gene Regulatory Networks

We introduce the biological background of EMT, e.g., what it is, where does it happen, and how it is related to cancer prognosis. Then we go further and introduce the key regulatory events during EMT and the genes and miRNAs that play important roles. Afterwards we show how EMT gene regulations are modeled using networks and provide network visualizations.

Part II: Identifying Single-omics Prognostic Signatures

We applied state-of-the-art feature selection algorithms - both network-based and none network-based algorithms, to select molecular signatures on individual omics data levels using EMT networks. We used three data levels alternatively - gene expression (including mRNA expression and miRNA expression), DNA methylation, and copy number alteration. We mapped each data level to the EMT networks, selected the features, and evaluated the prediction performance of the features using multiple evaluation metrics. To provide objective evaluations, we compared the performance of EMT features with that of several other groups such as random features. We also evaluated the prediction performance of frequently selected features, the combination of molecular features with clinical features, and the effect of classification thresholds. Additionally, the selected features were analyzed in terms of their stability, network properties and biological interpretations. At last, we related our results with state-of-the-art studies, where consistent findings were pointed out and new insights were addressed.

Part III: Identifying Multi-omics Prognostic Signatures

We identified multi-omics prognostic signatures and compared it with single-omics ones. We first obtained multi-omics signature by combining signatures from single data levels. The combined features showed significant improvement in patient stratification. Then we proposed an integrative feature selection algorithm based on multiplex networks, which is able to directly identify multi-omics signatures. Compared with feature selection on single-layer networks, we have obtained significantly better prediction performance with multiplex-based feature selection. We have also analyzed the feature compositions of the identified multi-omics signatures. Last but not least, we tested EMT signatures - both single- and multi-omics ones, on an independent dataset consisting of both gene expression and DNA methylation data for a cohort of patients. We showed that EMT signatures can stratify the independent samples into significantly different prognostic groups.

Part IV: Database Applications for Proteomics Data

Having focused on discovering and evaluating prognostic signatures, we are also concerned about how to effectively manage large biological data to support data integration and the development of individualized medicine. Here we take the example of proteomics data because of its complexity and large size. We benchmarked several relational and non-relational databases on their performance of storing and querying Mass Spectrometry (MS) and Tandem Mass Spectrometry (MS/MS) data. Based on the results we provided some advice on building omics data infrastructures.

Chapter 2

Epithelial Mesenchymal Transition in Cancer Progression

As introduced in Chapter 1, epithelial mesenchymal transition (EMT) gene regulations are highly relevant to cancer progression, which is what we aim to predict using feature selection and machine learning algorithms. We proposed to select features from EMT networks so that the molecular signatures can be more robust and biologically meaningful. However, we have not yet explained what is EMT and why it is relevant to cancer prognosis. This is the theme of this chapter. We will first introduce what is EMT. Then we describe the central EMT gene regulations that occur on multiple interconnected gene regulatory networks and pathways. Afterwards we explain how we represent EMT gene regulations in a network model. In the end we provide visualizations and basic information of the networks.

2.1 Biological Background

EMT is originally defined by a series of experiments where differentiated epithelial cells can convert into mesenchymal cells [84, 85]. It is a biological process that allows a polarized epithelial cell to obtain mesenchymal cell phenotypes including increased migratory capacity, invasiveness, and elevated resistance to apoptosis, etc [120]. Based on the biological context in which EMT occurs, it has been categorized into three subtypes [119, 121] 1) embryogenesis, implantation and organ development 2) tissue regeneration and organ fibrosis, and 3) cancer progression and metastasis. Here we put the emphasis on subtype 3.

In normal epithelial tissues, the cells are tightly connected laterally by cell junction structures, including adherence junctions, desmosomes, tight junctions, and gap junctions [285], as illustrated in Figure 2.1 from the book [26] (on page 1036 in chapter 19). The cells have an apical-basal polarity and anchor to the basement membrane. It ensures that the cells can only migrate laterally but not entering the underlying extracellular matrix (ECM). In contrast, mesenchymal cells are front-back polarized and rarely contact directly with neighboring cells [100]. They can invade as individual cells through ECM. In epithelial cancers, the activation of EMT program has been proposed as the critical mechanism for the acquisition of malignant phenotypes [245] because the epithelial cells undergoing EMT become invasive and can migrate to distant sites. At the distant site, the migrated cancer

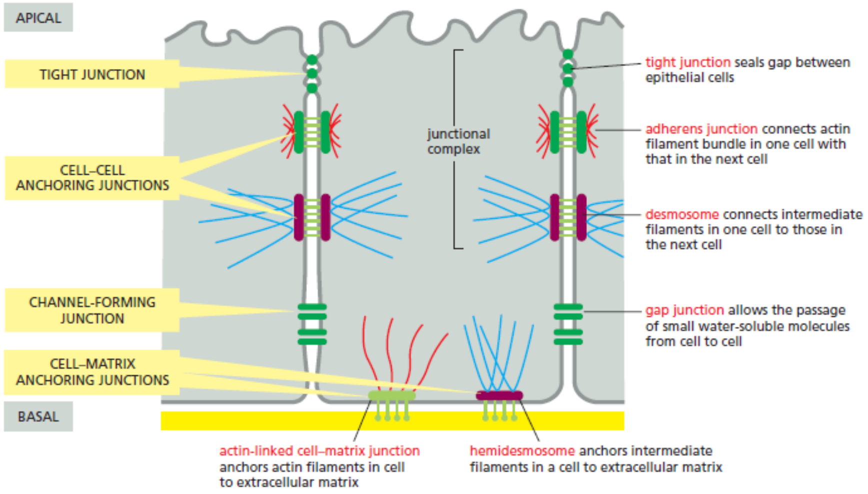

APICAL
TIGHT JUNCTION
CELL–CELL ANCHORING JUNCTIONS
CHANNEL-FORMING JUNCTION
CELL–MATRIX ANCHORING JUNCTIONS
BASAL
junctional complex
tight junction seals gap between epithelial cells
adherens junction connects actin filament bundle in one cell with that in the next cell
desmosome connects intermediate filaments in one cell to those in the next cell
gap junction allows the passage of small water-soluble molecules from cell to cell
actin-linked cell–matrix junction anchors actin filaments in cell to extracellular matrix
hemidesmosome anchors intermediate filaments in a cell to extracellular matrix

Figure 2.1: An illustration of cell junctions in epithelial cells. Figure source [26].

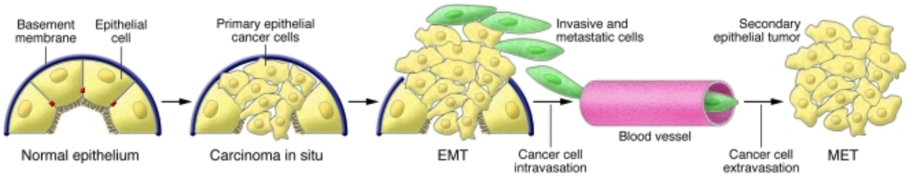

Basement membrane
Epithelial cell
Primary epithelial cancer cells
Invasive and metastatic cells
Secondary epithelial tumor
Normal epithelium
Carcinoma in situ
EMT
Cancer cell intravasation
Blood vessel
Cancer cell extravasation
MET

Figure 2.2: Contribution of EMT to cancer progression. It illustrates the process of normal epithelial cells losing their polarity and becoming invasive carcinoma, invading remote sites and establishing secondary tumors. Figure source [121].

cells can establish subsequent colonies via a MET (Mesenchymal Epithelial Transition, the reverse of EMT) process under the local micro-environments of distant organs [20, 116]. This is well illustrated by Figure 2.2 from Figure 5 in [121].

Many molecular processes are engaged in the EMT process such as the activation of transcription factors, expression of cytoskeletal proteins, expression changes of miRNAs, etc. [121] reviewed the potential pathways and transcription factors that induce EMT. In a later review study [24], the authors give an overview of the four different regulatory layers involved in EMT process including transcriptional control, Non-coding RNAs, EMT-associated differential splicing, translational and post-translational regulations. It reviewed the important molecules in each layer. In review paper [140], it is described the main changes that occur in cells that undergoes EMT, the roles of major EMT transcription factors, and signaling pathways involved in EMT. In review paper [285], the authors described the signaling pathways related to EMT and put these pathways in a molecular network to show their interactions. In the following we introduce in detail these molecular regulations, starting with hallmarks of EMT process and key transcriptional factors that mediate this process, then how these transcriptional factors are regulated at transcriptional, translational, and post-translational level, and ending up with a summary from the perspective of network and pathways.

A hallmark of EMT is the down-regulation of E-cadherin expression, which is essential for maintaining cell-cell adhesion [289]. Functional loss of E-cadherin has been associated with cancer progression and poor prognosis in human tumors [261]. In addition, genes encoding claudin and occludin are also repressed, which stabilizes the dissolution of apical tight junctions and desmosomes [109]. While the genes encoding epithelial cell junction proteins are repressed, the genes encoding proteins that promote mesenchymal adhesion are activated [74,291]. Specifically, the expression of N-cadherin is increased which provides a mechanism for trans-endothelial migration of cancer cells. Thus the 'cadherin switch' drastically changes the adhesive properties of cells and provokes cell migration and invasion [244, 272].

2.2 EMT Gene Regulations

As summarized in [24], EMT is regulated by four major interconnected regulatory networks - transcriptional control, non-coding RNA regulation, differential splicing and post-translational control. In the following paragraphs, we are going to follow this framework and describe the gene regulations in each layer. Although we could refer the readers to [24] without further explanations, we think it is necessary to provide the following gene regulations learned from this article. Because we are going to construct an EMT network, which is used in the following chapters as the basis for identifying molecular signatures for prognosis prediction, we think it is better to give the gene regulations in the following text. Different from the original article, we are going to introduce the gene regulations in a more concise way, only to make it sufficient for supporting this study.

At the transcription level, there are a few master transcription factors that repress the expression of cell-cell junction proteins [193]. SNAI1 and SNAI2 down-regulate CDH1 gene (encodes E-cadherin) by binding to CDH1 promoter [15, 29, 55, 93, 174]. ZEB1 [65], ZEB2 [41] and E47 [174, 196] also bind to CDH1 promoter and down-regulate CDH1 expression. SNAI1 and SNAI2 also down-regulate claudins and occludin [112], and other epithelial genes such as MUC1 and cytokeratin 18 [89]. ZEB1 also represses MUC1 expression [89]. ZEB2 down-regulates also P-Cadherin. Meanwhile SNAI1, SNAI2, [174] ZEB2 [254] and TWIST1 [182] up-regulate N-cadherin expression. SNAI1 is reported to induce the expression of mesenchymal markers fibronectin [29], LEF1 and the transcription repressor ZEB1 [89]. SNAI2 can induce the expression of mesenchymal marker vimentin [263]. LEF1 is shown to be an essential molecule for EMT [170]. It can facilitate the nuclear translocation of β-catenin, which drives the gene expression programme of cell cycle proteins and oncogenes that favor EMT [117].

The master transcription factors can also interact with epigenetic modifiers and cause genome-wide gene expression changes. DNA methyltransferase 1 (DNMT1) maintains the expression of E-cadherin in a methylation-independent way by interacting with SNAI1 and thereby preventing it from binding CDH1 promoter [68]. SNAI1 can also induce repressive histone modifications at CDH1 promoter by its interactions with HDAC1, HDAC2 and Sin3A [192]. ZEB1 recruits deacetylase sirtuin 1 (SIRT1) to the E-cadherin promoter region to deacetylate histone H3 and reduce the transcription of E-cadherin [27]. Similarly, ZEB1 interacts with the SWI/SNF chromatin-remodeling protein BRG1 to repress E-cadherin expression [212]. TWIST1 cooperates with BMI1 and EZH2 to down-regulated CDH1 [286].

In addition to transcriptional regulations, post-transcriptional gene regulations such

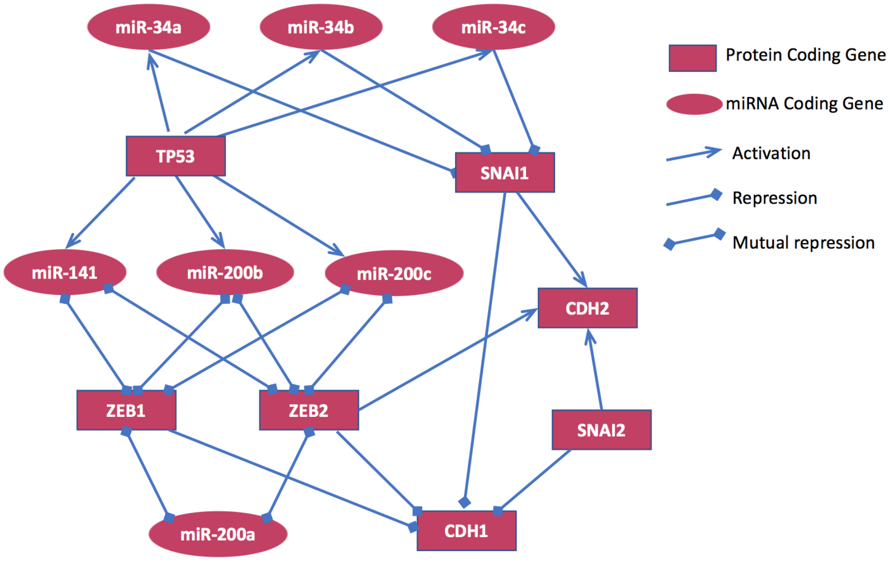


Figure 2.3: Construction of a small EMT network involving 14 genes and miRNAs. Two double-negative feedback loops - miR-200 family and ZEB family, miR-34 family and SNAIL1, are shown.

as pre-mRNA alternative splicing also contribute to EMT [24]. Many genes have different splicing isoforms and the balance between these isoforms is related to whether EMT will occur [78, 183]. Epithelial splicing regulatory protein 1 (ESRP1) and ESRP2 are involved in controlling the specific splicing of epithelial isoforms of many proteins such as CTNND1, CD44, etc [269, 284]. ESRP genes are found to be directly down-regulated by EMT transcription factors SNAI1, ZEB1 and ZEB2 [102, 202]. At the post-transcriptional level, miRNAs also regulate EMT gene expression. miRNAs are small non-coding RNA molecules that can bind to complementary sequences of mRNA molecules and thus silence the mRNAs. miRNA-200 family, including miR-200a, miR-200b, miR-200c, miR-141 and miR-429, and miR-205 repress the expression of ZEB family [87]. ZEB family also repress the expression of miR-200 family. The loop between these two families of transcription factors controls both EMT and MET [25, 87, 189]. miR-34 family members including miR-34a, miR-34b, and miR-34c also form a double-negative feedback loop with SNAI1 that regulates EMT [131, 223]. Tumor suppressor p53 inhibits EMT by activating miR-200 and miR-192 family members, which suppress ZEB1 and ZEB2. [31, 131]. EZH2 represses the expression of E-cadherin. miR-101 represses EZH2 and thus help maintaining E-cadherin expression [256]. To give a brief illustration of the regulatory network among these molecules, we have drawn a regulatory network including 14 central molecules, as shown in Figure 2.3.

The regulations of EMT also occur at translational and post-translational level. Increased expression of YB1 protein is shown to induce EMT [69]. It can stimulate the translation of SNAI1, ZEB2, LEF1, and TWIST1 [69]. Post-translational regulation of SNAI1 is also important for its protein level and sub-cellular localization. In the GSK3β dependent mechanism of SNAI1 regulation, SNAI1 is phosphorylated by CK1 and then phosphorylated by GSK3β for degradation [282, 293, 306]. TNFα stabilizes SNAIL by activating the NF-κB pathway [278]. GSK3β independent ubiquitin ligases MDM2 and FBXL14 can also target SNAI1 and SNAI2 for degradation. In contrast, LOXL2 interacts with SNAI1 and stabilizes it [194]. The subcellular location of SNAI1 also affects its activity as nuclear SNAI1 degrades slower than cytosolic SNAI1. PKD1 phosphorylates SNAI1 and facilitates its nuclear export, which decreases its effect on inducing EMT [63]. PAK1 and LATS2 phosphorylate SNAI1 and favor its nuclear retention, thus enhancing its activity [288, 298].

As is widely acknowledged, EMT involves the corporation of multiple pathways [140]. They often regulate these EMT master transcription factors such as SNAI1, ZEB1 [193]. As mentioned above, GSK3β phosphorylates SNAI1 for degradation. GSK3β also phosphorylates p53 and activates is transcriptional activity [253]. WNT pathway increases SNAI1 activity by inhibiting GSK3β. PI3K-AKT pathway also inhibits GSK3β [294]. When GSK3β is inhibited, β-catenin can accumulate in the cytoplasm and eventually translocate into the nucleus to activate TCF/LEF transcription factors and induce EMT [161, 276]. JAK/stat3 signaling pathway increases the expression of SNAI1 and TWIST1 [37, 283]. Growth factors that act through receptor tyrosine kinases such as EGF, FGF and VEGF can activate RAS-RAF-MEK-ERK MAPK signaling cascade. When activated, ERK2 can increase the expression of ZEB1, ZEB2, SNAI1, SNAI2 [39, 220]. Notch signaling pathway upregulates SNAI1 and SNAI2 [211]. In Hedgehog signaling pathways, GLI1 can induce SNAI1 expression [154]. Last but not least, TGFβ signaling pathway also contributes to EMT. In response to TGFβ1, SMAD2 and SMAD3 are phosphorylated and then combine with SMAD4 to form SMAD complexes. SMAD3 and SMAD4 activate HMGA2, which activates SNAI1 expression [249]. SNAI1 induces the nuclear translocation of ETS1, which is required for ZEB1 expression [52]. These pathways are not isolated but interconnected. TGF-*beta*/Smad signaling and Wnt signaling pathways are reported in developmental and pathological events. In EMT, Smad2 and Smad4 form a complex with LEF1 at the E-cadherin promoter, resulting in its transcriptional repression. In the TGFβ pathway, TGFβ receptor TGFBR1 can phosphorylates SHCA protein which activates SOS protein and then initiates the RAS-RAF-MEK-ERK MAPK pathway.

There are different perspectives with regard to EMT induction. Some propose that the convergence of signaling pathways is essential for EMT; Some say that the balance of different regulatory layers decides whether EMT will occur. Both are relevant to the network perspective of cancer. The EMT gene regulatory network is complex and how its components interact with each other to induce EMT is not yet well understood [247]. What is certain is that EMT is a crucial program for the invasion and metastasis of epithelial tumors which involves the loss of cell-cell adhesion and increase of cell mobility.

2.3 A Novel EMT Network Model

We constructed the EMT network based on the literature review above. It consists of 74 genes and miRNAs and their interactions. In our previous work [216], where EMT network was used to find molecular signatures for prognosis prediction, we showed that this EMT network did not give a good prediction performance as expected. One of the reason we consider is that the genes in the EMT network are mainly driver genes, which affect the expression of many downstream target genes that are usually more differentially expressed than the driver genes [66]. In this scenario, using only the driver genes to build predictive models is not a good option, especially when the data are noisy.

To compensate for it, we extended the EMT network by including the molecules that directly interact with or being regulated by the molecules in the network. In this way, the network not only contains driver genes but also some important downstream genes. We used *NetworkAnalyst* tool [279] to find these interactions. NetworkAnalyst is a comprehensive web-based tool for biological network analysis. One of its component is to take the input of gene or protein names and return their interacting partners in protein-protein interaction, miRNA-gene interaction, and TF-gene interactions. We thereby name the original network as *Core network* and the following extended versions of networks as *Extended network* and *Filtered network*.

- Extended network. We uploaded the official gene symbols of the core EMT network to the *NetworkAnalyst* web interface and obtained the first-order network of these core EMT genes and miRNAs using STRING interactome (confidence score cutoff: 900) [235], ENCODE transcription factor and gene target data [266], and miRNA-gene interaction data [105] respectively. Since we are only interested in aberrant pathways in cancer, for each obtained network we only kept the genes that belong to the *Pathways in cancer* term according to the KEGG pathway enrichment analysis. Note that we applied this selection criteria only to the newly added genes. None of the molecules in the core EMT network is removed. Accordingly, we obtained 3 new networks (one from each information source) and each network includes the core EMT network, the newly added genes and all interactions among them. Then we took the union of the three networks and generated the *Extended network*.
- Filtered network. Upon analysis of the extended EMT network, we notice that many genes have low variance and they are not useful for differentiating different sample groups. Thus, we removed the nodes (genes or miRNAs) whose variance are below a certain threshold, no matter whether the nodes belong to the core EMT network or not. We choose to remove nodes according to their variance because it is an unsupervised criterion. For example, we could also remove nodes based on univariate Cox proportional hazards model. However, as it is a supervised criterion, it would be hard to apply in reality because it requires the response information beforehand. Even though it is applicable, modifying the network in a supervised manner may lead to overfitting of subsequent analysis.

In biological literature the molecules are often referred to as, e.g., the corresponding kinases, transcription factors, or other functional molecules. However, for the convenience of mapping different types of omics data to the EMT network, we represent the molecules in the network using the names of their corresponding genes or miRNAs. For example, E-Cadherin, which is an epithelial cell marker, is represented using the name of its coding gene CDH1. The names of the molecules in the core EMT network are provided in Table 2.1. Their interactions are given in Table 7.1. The basic information of the three networks (as undirected) are given in Table 2.2. For the convenience of the readers, we visualized the three networks using R package *igraph* [47]. The networks are shown in Figure 2.4, Figure 2.5, and Figure 7.1.

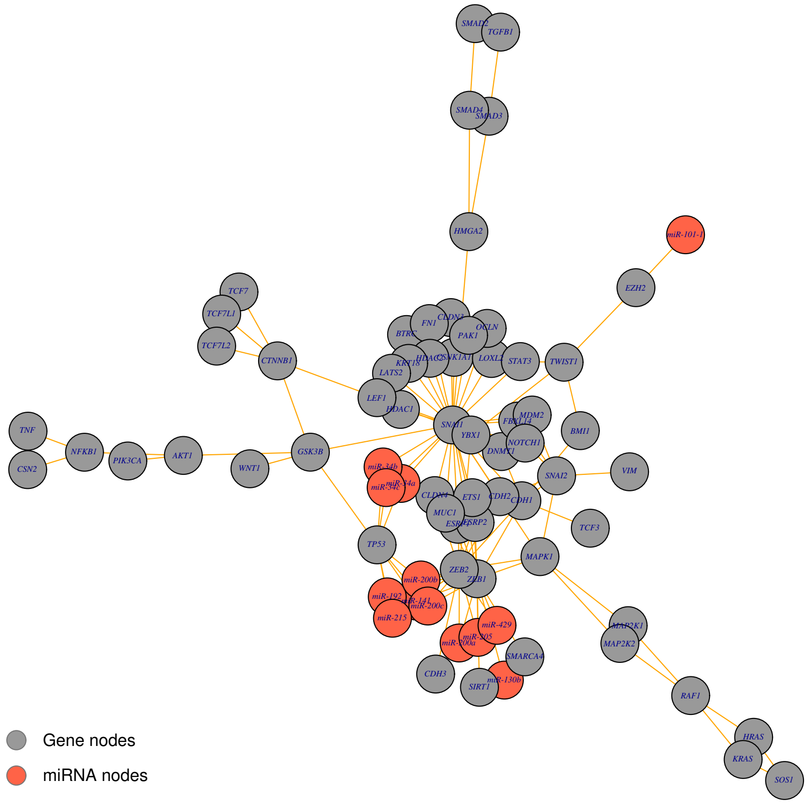


Figure 2.4: Core EMT Network. The names of genes and miRNAs are given on the nodes.

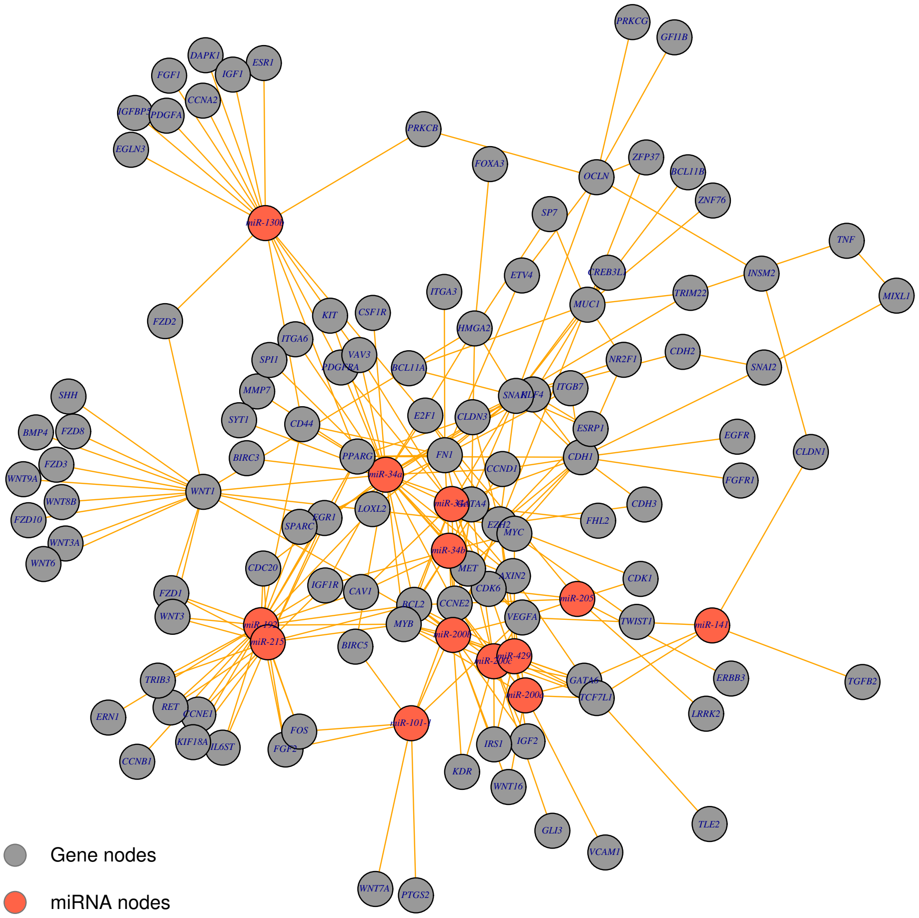


Figure 2.5: Filtered EMT Network. The names of genes and miRNAs are given on the nodes.

AKT1	ESRP2	LEF1	miR-205	SIRT1	TGFB1
BMI1	ETS1	LOXL2	miR-215	SMAD2	TNF
BTRC	EZH2	MAP2K1	miR-34a	SMAD3	TP53
CDH1	FBXL14	MAP2K2	miR-34b	SMAD4	TWIST1
CDH2	FN1	MAPK1	miR-34c	SMARCA4	VIM
CDH3	GSK3B	MDM2	miR-429	SNAI1	WNT1
CLDN3	HDAC1	miR-101-1	MUC1	SNAI2	YBX1
CLDN4	HDAC2	miR-130b	NFKB1	SOS1	ZEB1
CSN2	HMGA2	miR-141	NOTCH1	STAT3	ZEB2
CSNK1A1	HRAS	miR-192	OCLN	TCF3	
CTNNB1	KRAS	miR-200a	PAK1	TCF7	
DNMT1	KRT18	miR-200b	PIK3CA	TCF7L1	
ESRP1	LATS2	miR-200c	RAF1	TCF7L2	

Table 2.1: The list of gene names in the core EMT network.

Table 2.2: Basic information of the three EMT networks

Different EMT networks	$\lvert V(G)\rvert$	$\lvert E(G)\rvert$
Core EMT network	74	113
Filtered EMT network	123	253
Extended EMT network	455	2620

In the next two chapters, we are going to use these three EMT networks for extracting features for prognosis prediction in lung cancer. In fact, as shown by the experiments in the next chapters, both the extended network and the filtered networks can significantly outperform the core network. Many important features selected in the extended network and filtered network are not included in the core network.

Chapter 3

Single-omics Prognostic Signatures for Lung Adenocarcinoma

In this chapter, we apply state-of-the-art features selection algorithms on several omics data levels to identify molecular signatures for lung cancer prognosis prediction. The data levels include transcriptomics (including mRNA expression and miRNA expression), DNA methylation, and copy number alteration (CNA). We will employ 10 feature selection algorithms on each data level separately and evaluate their prediction performance using three evaluation metrics: ROC-AUC values, ROC-PR values and classification accuracy. All the evaluations are based on 30 times stratified 10-fold cross-validation. We will also analyze the selected features in terms of their stability, network properties, biological interpretations, and whether they can stratify the samples into significantly different prognostic groups. We begin with introducing the feature selection algorithms. Then we describe the experiments and analyze the results.

3.1 State-of-the-art Algorithms

There are mainly two types of studies for identifying prognostic signatures based on how follow-up data is used. In the original data, the survival times of individuals are either uncensored or right-censored. Right-censoring means that the individual was followed until time t, at which the individual was still alive, but then takes no further part in the study. Thus, we only know that the individual survived at least up to time t. The first type of algorithms is based on classification. The samples are firstly divided into two prognosis groups: a group that survived above a threshold and a group that did not. The goal of an algorithm is to find features that can discriminate the two groups. The second type of algorithms directly uses survival data and involved either Cox's proportional hazard model or survival analysis to select features and evaluate the predictions. The goal of an algorithm is to find features to better predict the risk of death. We included 10 algorithms for comparison, 6 from type one and 4 from type two. We also try to cover different underlying methodologies when we select algorithms. Since we will evaluate these algorithms using the same gene regulatory networks (GRN) - EMT networks, algorithms that are based on co-expression networks are excluded. We also excluded dimensionality reduction algorithms such as PCA because our goal is to select a subset of the original features. All the algorithms perform feature selection by selecting an optimal subset of

Table 3.1: Categorization of the 10 feature selection algorithms

Algorithm	Methodology	Network	Output	Ref
t-test	t-statistic	No	feature ranking	[97, 215]
Lasso	regularized regression	No	coefficients	[250]
NetLasso	network-based regularization	Yes	coefficients	[149]
AddDA2	subnetwork scoring and searching	Yes	subnetworks	[8, 40]
NetRank	feature importance on network	Yes	feature ranking	[274]
stSVM	random walk on network	Yes	feature ranking	[49]
Cox	Cox PH model	No	feature ranking	[43, 181]
RegCox	regularized Cox PH model	No	coefficients	[224]
MSS	random sampling	No	feature ranking	[150]
Survnet	subnetwork scoring and searching	Yes	subnetworks	[152]

features or by providing a rank of the features. Table 3.1 gives an overview of these algorithms.

Before describing these algorithms, we would like to give the necessary notations to be used in the following descriptions. Suppose that we have a training set $\boldsymbol{X}$ with m individuals and p features, $\boldsymbol{X} \in \mathbb{R}^{m \times p}$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, ..., \boldsymbol{X}_{ip})'$. For the purpose of exposition, we assume that X is gene expression data, although it can denote any type of omics measurement. The survival data for individual i is represented by (t_i, δ_i). t_i is the time at which death or censoring occurred to individual i. δ_i is the indicator of censoring. $\delta_i = 1$ if i is uncensored and $\delta_i = 0$ if censored. The response variable $y \in \{0, 1\}$. y_is are the class label for the corresponding $\boldsymbol{X}_i$s indicating whether individual i has survived up to a certain threshold. The right-censored samples with t_i below the threshold are thrown away because their class cannot be determined. The EMT network has p genes, which correspond to the p predictors. It is represented by a simple graph $G = (V, E)$[1], where V is the set of vertices that correspond to the p predictors and E is the set of edges. A is the adjacency matrix of the graph G. It is a square matrix such that its element A_{ij} is one when there is an edge from vertex i to vertex j, and zero when there is no edge. The diagonal elements of A are all zero, since there are no edges from a vertex to itself (loops) in the simple graph G. $\boldsymbol{D}$ is the degree matrix of graph G.

Feature selection based on classes

1. t-test. The intuition is that genes that are differentially expressed between the two sample groups are likely to be useful for sample classification. The more a gene is differentially expressed, the less likely that the difference of the mean expression values between the two groups is zero. For each feature vector x, let x_0 denotes the values of x where the samples belong to class 0 and x_1 denotes the values of x where the samples belong to class 1. Assuming unequal sample sizes and unequal

[1] Note that there is a directed edge from gene g_1k to gene g_2 if g_1 regulates g_2. However, in the case of gene product interactions, e.g., protein-protein interaction, it is hard to define a direction. Since the algorithms under comparison do not use edge direction information, we treated the graph as undirected. Likewise, since we could not find a good definition of the weight of each edge, we leave the graph edges as unweighted or with equal weight 1.

variances, we used Welch's t test to test whether the population means are different. The t statistic is calculated as:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

It is then used for significance testing under Student's t distribution with the degree of freedom calculated as

$$d.f. = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

The resulting p-values of the test are used to provide the feature ranking. Although this is a simple feature selection method, it has been shown to outperform many more sophisticated feature selection algorithms [97, 215].

2. Lasso (least absolute shrinkage and selection operator) [250]. It is a regression analysis method that performs both feature selection and regularization. It shrinks the coefficients of some features to zero, thus producing a sparse model. Let regression coefficient vector $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$, the Lasso estimate $(\hat{\beta}_0, \hat{\beta})$ is defined by [250] as:

$$(\hat{\beta}_0, \hat{\beta}) = arg\,min \left\{ \sum_{i=1}^{m} (y_i - \beta_0 - \sum_j \beta_j x_{ij})^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

   We used *glmnet* R package [73] to find the optimal coefficients via penalized maximum likelihood. *glmnet* solves the following problem

$$\min_{\beta_0, \beta} \frac{1}{m} \sum_{i=1}^{m} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \Big[(1-\alpha) \sum_{j=1}^{p} \beta_j^2 / 2 + \alpha \sum_{j=1}^{p} |\beta_j| \Big]$$

   over a range of λ values. The $l(y, \eta)$ is the negative log-likelihood contribution of observation i. α controls the elastic net penalty. $\alpha = 1$ is Lasso and $\alpha = 0$ is ridge regression. We set $\alpha = 0.95$. We used *glmnet* to perform 10-fold cross-validation on the training set to select the best λ value and the corresponding feature sets (features with nonzero coefficients).

3. Network-constrained regularization and variable selection [149]. It is a linear regression model that incorporates network information as a graph represented by its Laplacian matrix. The algorithm takes similar form as elastic net which fits a linear regression model that penalizes the L_1-norm and L_2-norm of the coefficients. Meanwhile, it adds a network-constrained penalty using the Laplacian matrix L of the graph. Such penalties can select subgroups of correlated features in the network and offer global smoothness of the coefficients over the network. The aim is to not only select a subset of important features but also potential subnetworks that are related to y_is. Given graph G, its Laplacian matrix $\boldsymbol{L}_{n \times n}$ is defined as:

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$$

The symmetric normalized Laplacian matrix $\boldsymbol{L}$ is defined as:

$$\boldsymbol{L}^{norm} := D^{-1/2}\boldsymbol{L}D^{-1/2}$$

The elements of L are given by

$$\boldsymbol{L}^{norm}_{i,j} := \begin{cases} 1, & \text{if } i = j \text{ and } deg(v_i) \neq 0 \\ -\dfrac{1}{\sqrt{deg(v_i)deg(v_j)}}, & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise.} \end{cases}$$

For any fixed non-negative λ_1 and λ_2, [149] defines the network-constrained regularization criterion as

$$L(\lambda_1, \lambda_2, \beta) = (\vec{y} - \boldsymbol{X}\beta)^T(\vec{y} - \boldsymbol{X}\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \beta^T \boldsymbol{L}\beta$$

$||\beta_1||$ is the L_1 norm, the same as used in Lasso. The third term $\lambda_2\beta^T\boldsymbol{L}\beta$ induces a smooth solution of β on the network. The goal is to find the estimator $\hat{\beta}$ to minimize $L(\lambda_1, \lambda_2, \beta)$. The author has rewritten the optimization problem as the following:

$$\hat{\beta} = arg\, min_\beta\{|\vec{y} - \boldsymbol{X}\beta|^2\},$$

$$\text{subject to } (1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha \sum_{\{i,j\}\in E}\left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_j}{\sqrt{d_j}}\right)^2 w(i,j) \leq t$$

From this network-constrained penalty one can see that if vertices v_i and v_j are neighbors in graph G, then β_i and β_j tend to be assigned similar values. We use the implementation of this algorithm in *Grace* R package. It selects λ_1 and λ_2 values by doing cross-validation on the training set.

4. Subnetwork aggregation algorithm [40]. This algorithm first overlaid the gene expression data on the corresponding proteins in the PPI (protein-protein interaction) network. Then it performs searching from each node and adds its neighbors until the mutual information between the aggregated gene expression of the subnetwork and the class vector does not increase above a rate r. When adding a neighbor, it selects the neighbor that maximizes the mutual information between the aggregation of the subnetwork and the class vector. After selecting all subnetworks, the algorithm picks out all significant subnetworks based on statistical tests using the null distributions of subnetwork scores of random networks and permuted data. This algorithm takes the averages of gene expressions within individual subnetworks as new features. However, this aggregation may not be appropriate when the genes exhibit opposite association

with the class label because then the predictive contributions can be canceled out. [8] investigated 11 operators to aggregate gene expressions. Many of them showed better prediction performance than the average operator. Here we adopted the Direction Aware Average (DA2) operator [8]. It is defined as:

$$DA2_g = \frac{1}{|\Psi_g|} \sum_{j \in \Psi_g} sgn(c_j) \times \boldsymbol{X}_{*j},$$

where Ψ_g is the gene set of seed gene g. $\boldsymbol{X}_{*j}$ is the expression values of gene j. c_j is the correlation value of gene j with the class vector. We therefore implemented the original algorithm but replaced the average operator with the DA2 operator. The pseudo-code is shown in Algorithm 1.

5. NetRank [274]. NetRank algorithm follows the idea of Google's PageRank algorithm [187], which decides the relevance of web documents based on the number of highly ranked documents that point to it. In the context of biological network, the rank of a gene is influenced by the ranks of genes that link to it. The rank can be computed iteratively:

$$r_j^n = (1-d)c_j + d\sum_{i=1}^{N} \frac{A_{ij} r_i^{n-1}}{deg_i}, \;\; 1 \leq j \leq |V(G)|,$$

where r_j^n denotes the ranking of gene j after n iterations, c_j is the absolute Pearson correlation of gene j expression values with the class vector, and $d \in (0,1)$ is a parameter describing the influence of the network on the ranking of genes. Iterating to convergence corresponds to solving the following equation [175, 274]:

$$(I - dA^T D^{-1})\vec{r} = (1-d)\vec{c}$$

The parameter d is set using Monte Carlo cross-validation on the training set. Values of d ranging from 0 to 1 with step size of 0.1 are tested. We followed the procedure described in [274].

6. Feature selection using network smoothed t-statistic (stSVM) [49]. This algorithm uses random walk kernel to characterize the degree of relatedness between network nodes. The p-step random walk kernel is defined as:

$$K = (\alpha I - L^{norm})^p$$

where α is a constant and p is the number of random walk steps. We used $\alpha = 1$ and $p = 2$, as proposed in this article. stSVM assesses the differential expression of each network gene by obtaining the t-statistic for each network node i. The t statistics $t_i, i = 1, ..., |V|$ are summarized into a vector $\vec{t}$. The final scores of nodes are calculated as: $\tilde{t} = t^T K$. The features are ranked based on the scores.

Algorithm 1 Subnetwork searching and aggregation algorithm

```
Input: $\boldsymbol{X} \in \mathbb{R}^{m \times p}$, $G = (V, E)$, $y = \{0, 1\}^m$, and $rate\, r$.
Output: $K_{subnet}, Scores$ //the selected subnetworks and their scores.

  Sign_of_node = $\{0\}^p$
  for each node $u \in V$ do
    $Sign\_of\_node_u = sgn(cor(\boldsymbol{X}_{*u}, y))$ //the sign of Spearman correlation
  end for
  $X' = \{1\}^m \times Sign\_of\_node^T \circ \boldsymbol{X}$
  for each node $u \in V$ do
    $Mutinfo\_seed = mutual\_information(\boldsymbol{X'}_{*u}, y)$
    $Mutinfo = \{0\}^{|N_G(u)|}$
    for each node $v \in N_G(u)$ //get neighbors do
      $aggrevec = \boldsymbol{X'}_{*v} + \boldsymbol{X'}_{*u}$
      $Mutinfo_v = mutual\_information(aggrevec, y)$
    end for
    $maxindex = arg\, max_i(Mutinfo_i)$
    if $Mutinfo_{maxindex} > Mutinfo\_seed \times (1 + r)$ then
      {extd_net, extd_score} = extend_subnet($X', y, \{u, maxindex\}, aggrevec$)
      $K_{subnet}.add(extd\_net)$
      $Scores.add(extd\_score)$
    end if
  end for

  Function extend_subnet $(X', y, snode, aggrevec)\{$
  $Mutinfo\_current = mutual\_information(aggrevec, y)$
  $Mutinfo = \{0\}^{|N_G(snode)|}$
  for each node $v \in N_G(snode)$ do
    if $v \notin snode$ then
      $aggrevec = \boldsymbol{X'}_{*v} + aggrevec$
      $Mutinfo_v = mutual\_information(aggrevec, y)$
    end if
  end for
  $maxindex = arg\, max_i(Mutinfo_i)$
  if $Mutinfo_{maxindex} > Mutinfo\_current \times (1 + r)$ then
    extend_subnet($X', y, \{snode, maxindex\}, aggrevec$)
  else
    $return(snode, Mutinfo\_current)$
  end if
  $\}$
  EndFunction
```

Feature selection based on survival data

In this part, the algorithms use censored survival data directly instead of dividing it into two classes. The goal of these algorithms is not to classify patients into two distinct groups but to assess how much more at risk one individual is than another. Since these algorithms share common components such as survival analysis and Cox proportional hazards model (Cox PH) [43], we would like to first describe these components before going into the details of each algorithm.

In survival analysis, usually one would estimate the survival function $S(t)$. If T is the time of death, then $S(t) = p(T > t)$. It is the probability of surviving at least to time t. It can be estimated non-parametrically in the presence of censoring using Kaplan-Meier estimate [124]. Assuming $d(t)$ is the number of deaths at time t, $q(t)$ is the number of right-censoring at time t, and $n(t^-)$ is the total number of individuals at risk an instant before time t, the Kaplan-Meier estimate of $S(t)$ is

$$\hat{S}(t) = \hat{S}(t^-)\hat{p}(T > t|T \geq t)$$

If no failures occur at time t, $\hat{p}(T > t|T \geq t) = 1$.

If one or more failures occur at time t,

$$\hat{p}(T > t|T \geq t) = \frac{n(t^-) - d(t)}{n(t^-)}$$

Note that the Kaplan-Meier curve only drops at the time when failures occur. If we write $t_{(i)}$ as the ith time point, and $d_{(i)}$, $q_{(i)}$, and $n_{(i-)}$ accordingly. Then the Kaplan-Meier estimate can be written as:

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} \frac{n_{(i-)} - d_{(i)}}{n_{(i-)}} \text{ , with } \hat{S}(t) = 1 \text{ for } t < t_{(1)}$$

The confidence interval of $\hat{S}(t)$ can be estimated, e.g., using Greenwood's estimator [86] or the Tsiatis/Aalen formula [1]. In prognostic biomarker evaluation, one typically wants to know whether a biomarker can separate the individuals to groups that have significantly different survival functions. In other words, we need to compare different Kaplan-Meier curves. It is illustrated in Figure 3.1, where the survival functions of two groups of individuals, e.g., case and control, are plotted with confidence intervals. This comparison can be accomplished using log-rank test. It is a nonparametric test and appropriate to use when the data are censored. The basic idea is that if the two groups have the same survival distributions, then the ratio between the number of events and the number of individuals at risk at the beginning of the time period for each observed time point should be the same between the two groups.

To explain this using notations, let $t = 1, ..., T$ be the distinct times of observed events in either group. For each time t, let N_{1j} and N_{2j} be the number of individuals at risk at the start of period j in the two groups. Let O_{1j} and O_{2j} be the number of deaths in the groups at time j. $N_j = N_{1j} + N_{2j}$, $O_j = O_{1j} + O_{2j}$. Under the null hypothesis that $S_1(t) = S_2(t)$, O_{1j} has a hypergeometric distribution with parameters N_j, N_{1j}, and O_j, and the expected

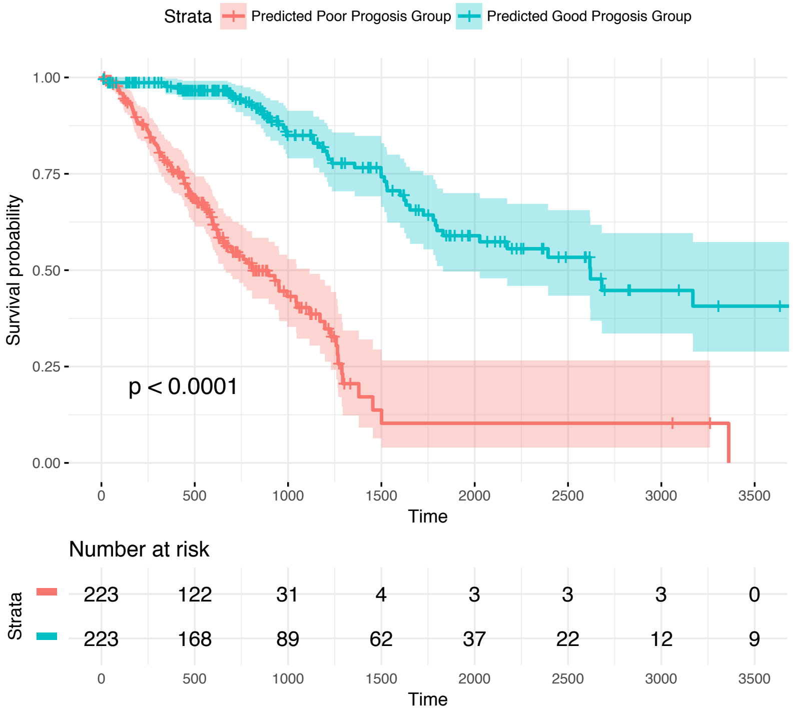


Figure 3.1: An example of estimated survival curves using Kaplan-Meier estimator and the result of log-rank test.

value $E_{1j} = \frac{O_j}{N_j} N_{1j}$, with variance $V_j = \frac{O_j(N_{1j}/N_j)(1 - N_{1j}/N_j)(N_j - O_j)}{N_j - 1}$. The log-rank test compares O_{1j} with E_{1j} under the null hypothesis. The test-statistic is:

$$q = \frac{\sum_{j=1}^{J} w_j (O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{J} w_j^2 V_j}}$$

for some weights w_j. If the null hypothesis is true, $q \sim \chi_1^2$ asymptotically. The p-value, $p = p(Q > q | H_0)$, follows the CDF of the χ_1^2 distribution. For log-rank test, $w_i = 1$. Thus, emphasis is put on larger values of time. For the generalized Wilcoxon test, $w_i = n_{(i^-)}$, where emphasis is put on smaller values of time.

From the test above, one can know whether there is a significant difference between the survival distributions of the two groups based on a categorical covariate. But it is not known how much more at risk one group is than the other. When dealing with continuous covariate, depending on how the covariate is binned into groups, the test results can be different. Cox's proportional hazards model (Cox PH) [43] can be used to incorporate continuous covariates into survival analysis and analyze their effect on survival.

Cox PH employs hazard function $h(t)$. Hazard function is defined as the instantaneous rate for the event to occur at time t conditional on survival until time t or later:

$$h(t) = \lim_{\triangle t \to 0} \frac{p[(t \leq T < t + \triangle t) | T \geq t]}{\triangle t}$$

The survival function can be written as a function of the hazard function:

$$S(t) = exp\Big\{ - \int_0^t h(\tau) d\tau \Big\}$$

The hazard function in Cox PH takes the form:

$$h(t, \boldsymbol{x}) = h_0(t, \boldsymbol{\alpha}) exp(\boldsymbol{\beta}^T \boldsymbol{x}) \ ,$$

where $h_0(t)$ is the baseline hazard function, α are parameters that influence the baseline hazard function. $\boldsymbol{\beta} = (\beta_1, ...\beta_p)'$ is a vector of regression coefficients. The baseline hazard function depends on time but not covariates. The second term depends on the covariates but not time.

Considering two individuals with covariate x_1 and x_2 (x_1 and x_2 are scalars). Then the ratio of their hazards at time t is:

$$\frac{h(t, x_1)}{h(t, x_2)} = exp\{\beta(x_1 - x_2)\}$$

As shown above, the hazard ratio does not depend on time. One can use the notion of hazard ratio to test whether a covariate influences survival. The coefficient $\boldsymbol{\beta}$ is estimated by maximizing the Cox's log-partial likelihood:

$$logL(\boldsymbol{\beta}) = \sum_{i=1}^{m} \delta_i \Big\{ \boldsymbol{X}_i' \boldsymbol{\beta} - log\Big[\sum_{j \in R(t_i)} exp(\boldsymbol{X}_j' \boldsymbol{\beta}) \Big] \Big\}$$

where t_i is the observed or censored survival time for ith individual. δ_i is the censoring indicator. $\delta_i = 1$ if i is observed and $\delta_i = 0$ if censored. $R(t_i)$ is the risk set at time t_i - the set of patients who survived prior to time t_i. β can be estimated without knowing the baseline hazard function. According to Cox, the partial likelihood is valid when there are no ties in the survival data [44]. In case there are ties, Efron approximation is used.

Below we introduce the details of each feature selection algorithm.

1. Univariate Cox PH model. We use each feature to fit a Cox PH model to obtain the effect of each feature on survival. Given covariate x and parameter β. The hazard rate at time t is given by:

   $$h(t, x) = h_0(t)exp(\beta x),$$

   where β is the regression coefficient and $h_0(t)$ is the baseline hazard function. Increasing x by one unit scales the hazard rate by e^{β}. Thus β can be interpreted as the increase in log hazard per unit of x. We use *survival* R package to fit the univariate Cox PH model for each feature and obtain the p-value. Then we rank the features by their p-values. Although straightforward, this approach has been widely applied in many studies to obtain a preliminary set of features for prognosis prediction [46, 181, 201, 267].

2. Regularized Cox PH model [224]. This algorithm employs cyclical coordinate descent to fit the Cox PH model with elastic net penalties. Assuming there are no ties in death/censoring time, the optimization problem is to find β that maximizes Cox's log-partial likelihood given above, subject to the constraint: $\alpha \sum_{j=1}^{p} |\beta_j| + (1 - \alpha) \sum_{j=1}^{p} \beta_j^2 \leq c$. The algorithm is implemented in R package *glmnet*.

3. Multiple Survival Screening (MSS) algorithm [150]. This algorithm mainly addresses the issue of high variability of tumor gene expression profiles. It means that in tumour cells, there are many 'passenger signals' and they differ among individual samples. These signals may 'bury' the real cancer gene expression signals. MSS algorithm addresses this issue by using random samplings of random gene set (RGS) and random data set (RDS). It aims to select more stable markers by using genes that are consistently good predictors in randomly sampled data sets. We adjusted the original algorithm to our experiment. The pseudocode is given in Algorithm 2.

4. Survnet [152]. This algorithm is similar to [40] where a scoring function is defined, and a searching algorithm is employed to find subnetworks as molecular signatures. Different from [40], Survnet evaluates each node using a Cox PH model and a searching criterion is defined accordingly. Specifically, each gene i is assigned a score s_i, which is transformed from the p-value of the univariate cox PH model p_i.

   $$s_i = \Phi^{-1}(1 - p_i)\ ,$$

   where Φ^{-1} is the inverse standard normal cumulative distribution function. The subnetwork scoring function F of a subnetwork G_s with n_s genes is defined as:

   $$F_{G_s} = \frac{1}{\sqrt{n}} \sum_{i \in G_s} s_i$$

Algorithm 2 Multiple survival screening algorithm

```
Input: $\boldsymbol{X} \in \mathbb{R}^{m \times p}$, survival data of $m$ individuals, the number of RDS $nD$, the number of RGS $nG$.
Output: The ranking of the p features.
  Initialize a matrix $S = \{0\}^{nG \times nD}$ to store the p-values of log-rank tests.
  Generate the indices $idx\_D$ and $idx\_G$ for $nD$ RDS and $nG$ RGS respectively.
  for i = 1 to $nD$ do
    for j = 1 to $nG$ do
      Extract the datasets $M$ consisting of $idx\_D[i]$ samples and $idx\_G[j]$ genes.
      Apply fuzzy clustering algorithm to divide samples in M into two clusters.
      Fit survival curves for the individuals in each cluster.
      Apply log-rank test to compare the two curves and assign the p-value to $S[i, j]$.
    end for
  end for
  Keep the gene sets whose p-values are less than 0.05 in at least 25% of all RDS.
  Derive the ranking of genes based on their frequencies in these gene sets.
```

. A greedy search algorithm [40] is employed to find the subnetworks.

3.2 Experiments

3.2.1 Data Pre-processing

We obtained the following types of data for LUAD from the FIREHOSE Broad GDAC website (Broad Genome Data Analysis Center [http://gdac.broadinstitute.org]). The data were from TCGA data version 28/01/2016.

1. Level 3 normalized mRNA-Seq data - the calculated expression signals of genes per sample. The data are generated with Illumina HiSeq 2000 RNA Sequencing Version2 analysis platform using RSEM (RNA-Seq by Expectation-Maximization) quantification [148].

2. Level 3 CNA data - copy number alterations for aggregated/segmented regions of chromosomes per sample. It is generated with Affymetrix Genome-Wide Human SNP Array 6.0 platform with Human Genome version 19 (hg19) reference.

3. Level 3 DNA methylation data - the calculated Beta-values mapped to the genome per sample. Beta-value is a methylation level measurement of an interrogated CpG site. It is calculated as the ratio of the methylated probe intensity and the sum of methylated and unmethylated probe intensities. The data is additionally preprocessed by calculating the mean Beta-values of the probes among each gene (GDAC Methylation Preprocessor Pipeline).

4. Level 3 miRNA-Seq data - the calculated expression for all reads aligning to miRNAs per sample. The data is generated with Illumina Genome Analyzer miRNA Sequencing platform using RPKM (Reads Per Kilobase exon Model per million mapped reads) quantification.

Table 3.2: The description of single-level datasets. This table shows (1) sample sizes for labeled data, (2)sample sizes for censored data and (3) the total number of features on each data level.

	labeled data			censored data	
	good prognosis	poor prognosis	total		Total # features
Level GE	84	99	183	497	19290
Level DM	74	93	167	447	20074
Level CNA	73	76	149	503	21456

5. Clinical and follow-up data of patients.

The CNA data we obtained contain the segment mean values of different regions on the chromosomes. To enable the mapping of CNA data to the network, we calculated the gene-level CNA values using the function *ProcessCNAData()* in the *TCGA-Assembler* R package [292]. In detail, for each gene on the reference genome hg19, the regions in the CNA data that overlap this gene were found. Then the function takes the sum of the element-wise products of the regions' length and the regions' segment mean values. This value is then divided by the total length of the overlapping regions to give the gene level CNA values. This calculation is performed for every sample to generate gene-level CNA values. In the end, we normalized each of the four data types feature-wise by subtracting the mean and dividing by the standard deviation. We removed the features with >20% missing values among the samples for each type. For the remaining features, missing values are imputed using the mean of the corresponding features.

We normalized each omics data feature-wise. We combined mRNA-Seq data and miRNA-Seq data as transcriptomics data. Since these data levels are multi-dimensional description of genes, for the convenience of the following text, we abbreviate transcriptomics data, DNA methylation data, and CNA data as GE, DM, and CNA respectively.

We mapped each data level to the EMT networks with corresponding features. For each data level we have 6 datasets: 3 for the data mapping with labeled samples, and 3 for the data mapping with censored samples. The datasets with labels have less samples than the dataset with censored samples because for many samples the labels are unknown and thus are removed. The information of the datasets is given in Table 3.2, together it is shown the total number of features from each data level. For CNA data, there are missing values for some nodes in the EMT network, thus we took subnetworks of the EMT networks such that each node in the network is mapped with real-valued features. The sizes of networks for CNA data are 70 for the core EMT network, 445 for the extended EMT network, and 117 for filtered EMT network.

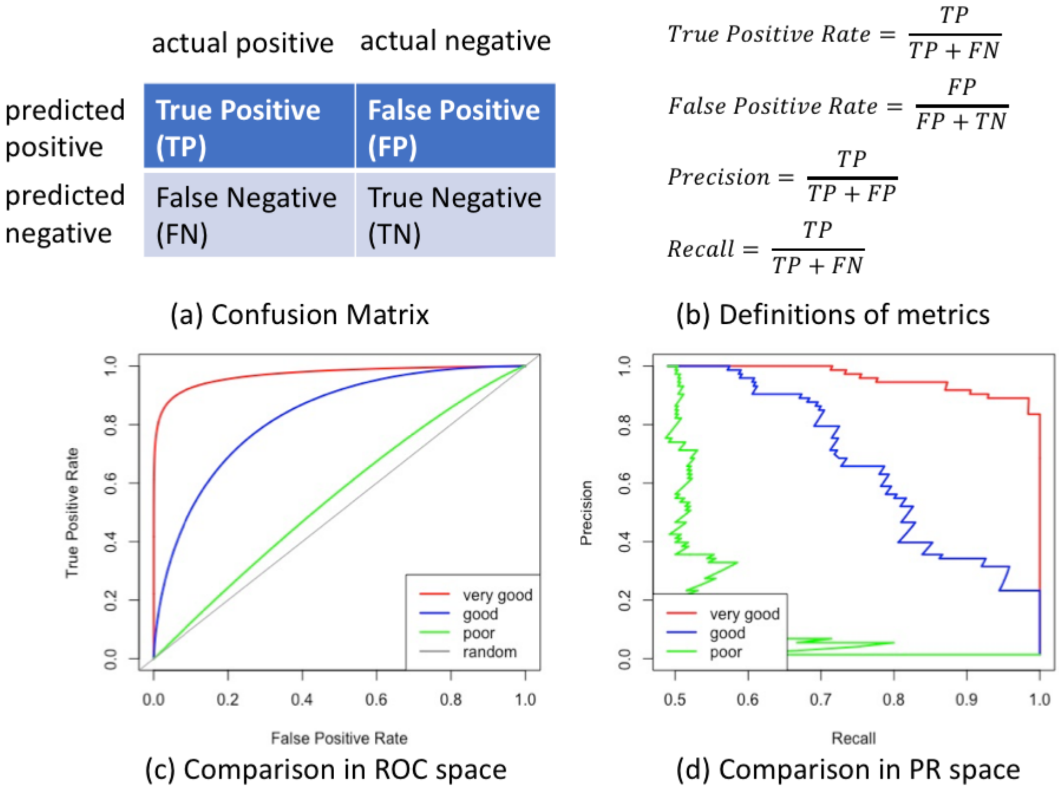
3.2.2 Evaluation Metrics

We have employed multiple evaluation metrics for evaluating the classifiers trained using the selected features. This is because the performance can be sensitive to the evaluation metrics [34]. We used accuracy, AUC-ROC, PR-ROC to evaluate classifiers as they address different aspects of the performance. The accuracy metric imposes a threshold for continuous classifiers and evaluates the resulting classifications. Classification accuracy is often a poor metric in the real world when the data is skewed [198, 199]. In our study, the datasets are balanced and stratified sampling techniques are applied in cross-validations. Therefore, the issue of imbalanced data is not of concern.

ROC curve measures the capability of a classifier to rank the positive samples relative to the negative samples. We use ROC (Receiver operating characteristic) curve to evaluate the prognostic potential of different feature sets. Let X denotes the predicted values, with higher value more prone to short survival, and C denotes the binary outcome of short or long survival, then the ROC curve for X is a plot of the True Positive Rate (TPR, also called sensitivity) associated with the dichotomized test $X > c$ versus the False Positive Rate (FPR) for all possible threshold values c. The ROC curve is a monotone increasing function in $[0, 1]$. It plots the value pairs of $\{P(X > c|C = 1), P(X > c|C = 0)\}$ for $c \in (-\infty, \infty)$. ROC curves are often used to measure the prognostic capability of biomarkers. It is shown that a perfect ROC curve with AUC equals 1 does not guarantee a perfect classification, which depends on the threshold [72]. Compared with accuracy measure, it has several advantages. First, a ROC curve captures the inherent discrimination capability of a test without relying on any specific threshold while using accuracy measure one often has to decide a threshold for the algorithms that output the probabilities of class values. Second, ROC curve can depict the relative trade-offs between TPR and FPR. Third, ROC curve provides a valid approach to compare different markers even when they are on different scales. Last, the area under a ROC curve (AUC-ROC) can be interpreted as the probability that the marker value from a randomly chosen positive individual exceeds the marker value of a randomly chosen negative individual. In this way, AUC-ROC values can be used to compare different ROC curves from different biomarkers.

Given the prediction vector X and a threshold c, one can derive from the prediction probabilities the confusion matrix, and calculate the TPR, FPR, prediction, and recall metrics, as shown in Figure 3.2. As proposed in [53], one can treat these metrics as functions that map the confusion matrix to a point in either ROC space or PR space. In the ROC space, the goal is to be closer to the upper-left corner. While in the PR space, the goal is to be closer to the upper-right corner. As shown in Figure 3.2, the red curve is the best one in both ROC and PR space. [53] has proven that for a given dataset of positive and negative samples, there exists a one-to-one correspondence between a curve in ROC space and a curve in PR space. Meanwhile, the authors also show that among different ROC curves, a higher AUC-ROC value does not always lead to a higher AUC-PR value. By the definition of PR curve, the prediction needs to achieve a high precision at low recall values to achieve high AUC-PR values. Since ROC and PR curves are both derived from the confusion matrix but cannot replace each other, we think it is necessary to evaluate the above algorithms using both ROC and PR curves.

The area under the curve is used as a metric to define how an algorithm performs over the whole space. The area under the ROC curve (AUC-ROC) is now commonly used to evaluate biomarker performance. It can be calculated using the trapezoidal areas created among ROC points. There are a few R packages for ROC analysis. Comparisons of these packages, with regard to their support of smoothing, partial AUC, confidence interval, and statistical tests are provided in [204, 230]. According to these studies, pROC R package is shown to have advantages over the other packages and we will employ it for the ROC analysis. In PR space, linear interpolation cannot be applied because it yields over-optimistic estimate of performance [53]. Therefore, we cannot directly calculate the area



	actual positive	actual negative
predicted positive	True Positive (TP)	False Positive (FP)
predicted negative	False Negative (FN)	True Negative (TN)

(a) Confusion Matrix

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

(b) Definitions of metrics

(c) Comparison in ROC space

(d) Comparison in PR space

Figure 3.2: Evaluation metrics. In (a) and (b) we give the definitions of a few common probabilities used in machine learning. In (c) and (d) we plotted curves in ROC and PR spaces. The curves illustrate very good, good, and poor predictions. The curves with the same color are generated using the same values.

under the PR curve using linear interpolation. [23, 125] proposed fine-grained, continuous interpolation between the PR points. The latter method has been implemented in the PRROC R package [83]. We will use this package for AUC-PR calculation.

3.2.3 EMT-based Feature Selection

As introduced in Chapter 2 we constructed three EMT networks: core EMT network, filtered EMT network, and extended EMT network. These networks and the features in the network will be used for feature selection. We would like to address three motivations of using EMT networks instead of the entire PPI network. The first motivation is from a biological perspective. Although numerous biological evidence has shown that EMT gene regulations play important roles in cancer prognosis, it has not been directly used for prognosis prediction. Therefore, it is necessary to test the predictive capabilities of EMT networks. The second motivation is that using EMT network can dramatically reduce the cost of the experiments. Compared with the original datasets which have >20,000 features, EMT networks have only 74, 123, and 455 features. High dimensionality is a major obstacle in selecting biomarkers. The PPI network has around >20,000 nodes, which cannot help with the dimensionality issue. The irrelevant parts of the network can cause high variance in selected features. Last but not least, we objectively compared different algorithms with the same network. As shown in the summary Table 1.2, the feature selection algorithms usually used different sources of underlying network. This makes it difficult to directly compare these algorithms. We tested all algorithms using the same data partitions, the same networks, and evaluated them using the same metrics.

In our previous work [216], where we decomposed the EMT network into network motifs and selected network motif features, the predictive performance was not as good as expected. Sometimes the EMT features were even performed worse than random features. This does not agree with the biological facts that EMT is highly relevant to cancer prognosis. We have found two potential reasons for it. One is the limited information in the core EMT network. We have resolved it by extending the EMT network to include downstream genes, as introduced in Chapter 2. The second and more critical reason is that we used Lasso feature selection on each of the feature sets, which may have overwhelmed the effects of different feature selection methods. Research also shows that Lasso could be biased when the features are highly correlated [251]. As we used network motif scores as new features, the overlap of many network motifs can cause high correlation among the features. In this case, Lasso cannot give stable predictions. To change it, we abandoned the idea of decomposing EMT network into network motifs and instead included many state-of-the-art feature selection algorithms that do not depend on network motifs. Lasso is only included as one of the algorithms.

Work Flow

The work flow of the experiments is shown in Figure 3.3. We take the processed data from the three data levels and map the data to the three EMT networks, which yields 9 combinations of data and networks. For each combination we performed 30 times 10-fold cross-validation to evaluate the 10 feature selection algorithms. Note that the indices for training and testing samples in each fold are kept identical across all algorithms. In the

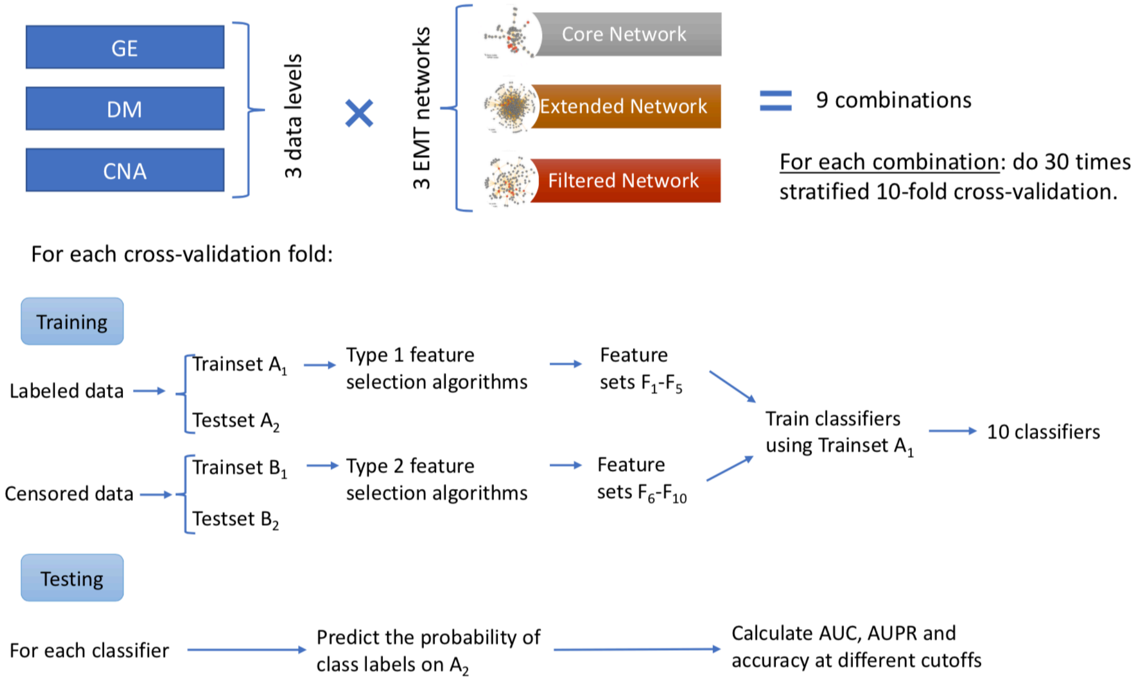

GE
DM
CNA
3 data levels
×
3 EMT networks
Core Network
Extended Network
Filtered Network
=
9 combinations
For each combination: do 30 times stratified 10-fold cross-validation.
For each cross-validation fold:
Training
Labeled data
Trainset A1
Testset A2
Type 1 feature selection algorithms
Feature sets F1-F5
Censored data
Trainset B1
Testset B2
Type 2 feature selection algorithms
Feature sets F6-F10
Train classifiers using Trainset A1
10 classifiers
Testing
For each classifier
Predict the probability of class labels on A2
Calculate AUC, AUPR and accuracy at different cutoffs

Figure 3.3: Experiment flow. This figure illustrates the combination of data levels with different networks and shows how the feature selection algorithms are evaluated in each cross-validation fold.

lower part of the figure, we show how the evaluation was performed in each cross-validation fold. As written in section 2.2.2, we have both labeled data and censored data for each data type. We divide both of them into 10 folds, use 9 folds for training and the fold left for testing. This gives training set A_1 and testing set A_2 for labeled data, and training set B_1 and B_2 for censored data. For the training phase, we apply type 1 algorithms on A_1 and type 2 algorithms on B_1 to select features. This gives feature sets F_1 to F_{10}. Each of these feature sets is used to train classifiers on A_1. For the testing phase, we use each trained classifier to make predictions on A_2. As A_2 is labeled data, we evaluate the prediction performance using AUC, AUPR, and classification accuracy at different cutoffs.

There are a few commonly used accuracy estimation methods such as holdout, k-fold cross-validation, stratified k-fold cross-validation, leave-one-out cross-validation, and bootstrap. We used stratified 10-fold cross-validation because it is demonstrated to give a more stable accuracy estimation [133]. We have noticed that in our experimental design some of the testing samples in A2 could appear in the training set of B1 and interfere with the evaluation. Therefore, we have tried another experimental setting (setting 2) to separate the data in a stricter way to make sure that A2 and B1 have no overlap in each cross-validation run. We have compared the results of setting 2 with the original setting and found that this did not cause significant effect on the evaluation. We think this is because both labeled data and censored data are divided into 10 folds independently and randomly. The fraction of overlapping samples between A2 and B1 is rather small and random. This could also be inferred from the data description in Table 3.2.

Comparative Groups

We have proposed to employ feature selection algorithms on the EMT networks, which contain at most 2.5% features from the original features. This is an original and very bold assumption we have made based on biological domain knowledge. Consequently, we are urged to compare the performance of EMT features with that of considering all features from the corresponding data levels to intuitively assess how much ”performance loss” is caused due to the assumption. This can tell us how useful EMT networks are in selecting prognostic signatures. Therefore, we included the following four comparative groups:

1. Random networks. A random network has the same structure as an EMT network but with randomly selected features on each node. We have used the same samples and the same cross-validation folds to select features on the random networks. This is to enable paired statistical tests between EMT networks and random networks. In total, 150 random networks were generated, each was tested using 10-fold cross-validation.
2. All features that are in the EMT networks. We include all features in the EMT networks as features.
3. All features from the corresponding data levels (>19,000 features).
4. Take all features from the corresponding data levels and select a subset of features using Lasso.

Additionally, we used clinical features for prediction. We used 9 clinical features from patient data. The features are: age at initial diagnosis, gender, overall pathologic stage, the variables T and N using the TNM cancer staging system [1], smoking history (discrete variables denoting the heaviness of smoking), radiation therapy indicator and molecular therapy indicator. We have also combined clinical features with frequently selected molecular features for prediction on the samples that have both clinical features and the corresponding molecular features. What’s more, we tested the effect of thresholds in prediction performance, which is usually omitted in most of the studies where only one arbitrary threshold was chosen.

3.3 Results

In this section we present our results into five parts:

1. The comparison of EMT features with random features.
2. The AUC, AUPR, and accuracy of EMT features.
3. Frequently selected EMT features.
4. The effect of adding clinical features.
5. The effect of classification thresholds.

[1] TNM Classification of Malignant Tumor (TNM) is the most widely used cancer staging system. T refers to the size and extend of the main tumor. N refers to the number of nearby lymph nodes that have cancer. M refers to whether the cancer has spread from the primary tumor to other parts of the body.

3.3.1 EMT Features vs. Random Features

We compared the prediction performance of EMT features with that of randomly chosen features using AUC, AUPR and prediction accuracy metrics. We chose two combinations of data levels and network sizes for the comparison. For each combination, 150 sets of random features were sampled to estimate the overall prediction performance, each with stratified 10-fold cross-validation. Figure 3.4 shows the distributions of AUC, AUPR, and prediction accuracies for DNA methylation data with core EMT network. Figure 3.5 shows the distributions of the metrics for gene expression data with filtered EMT network. In each figure, the probability density functions of the evaluation metrics are estimated using kernel density estimation method with Gaussian kernel.

Since our goal here is to compare random features with EMT features, we only included 5 feature selection algorithms and one comparative group (use all EMT features). The data folds used in 30 times 10-fold cross-validation are kept the same for both EMT features and random features. This enabled us to perform paired t-tests between EMT features and random features for each feature selection algorithm. Since we have sampled 150 random networks, each tested with 1 time 10-fold cross-validation, we randomly chose 30 random networks for the statistical tests. Note that the probability density functions of random networks are estimated using all 150 random networks. In each sub-figure we give the p-value of the paired t-test. We observe that EMT features obtained significantly better prediction performance for every feature selection algorithm and evaluation metric.

3.3.2 AUC, AUPR and Accuracy of EMT Features

In Figure 3.6 we show the boxplot of AUC values of the 10 feature selection algorithms, with 3 different data levels and 3 different EMT networks using SVM classifier. The comparative groups 2, 3, and 4 are included. In Figure 3.7 we give the boxplot for AUPR values. Table 3.3 shows the average AUC and AUPR values. All results are averaged from 30 times stratified 10-fold cross-validation. We observe multiple interesting comparisons that are consistent among the majority of the algorithms.

1. GE and DM data gave more accurate predictions than CNA data.

2. Larger network does not necessarily give better predictions. For example, with GE data, filtered network in general gives better predictions than the core network and the extended network. With DM data, the core network is in general the one with higher AUC values. With CNA level, the advantage of larger network size is almost negligible.

3. Compared with the first column, where all features within the EMT network were used without feature selection, many feature selection algorithms do not show improved performance. Some combinations of algorithms and networks significantly outperform EMT features while some others gave worse performance.

4. Compared with using all features from GE data (19290 features), EMT network-based features achieved significantly higher AUC values. Compared with using all features of DM data (20074 features) and CNA data (21456 features), EMT network-based features achieved significantly better or at least equal prediction performance.

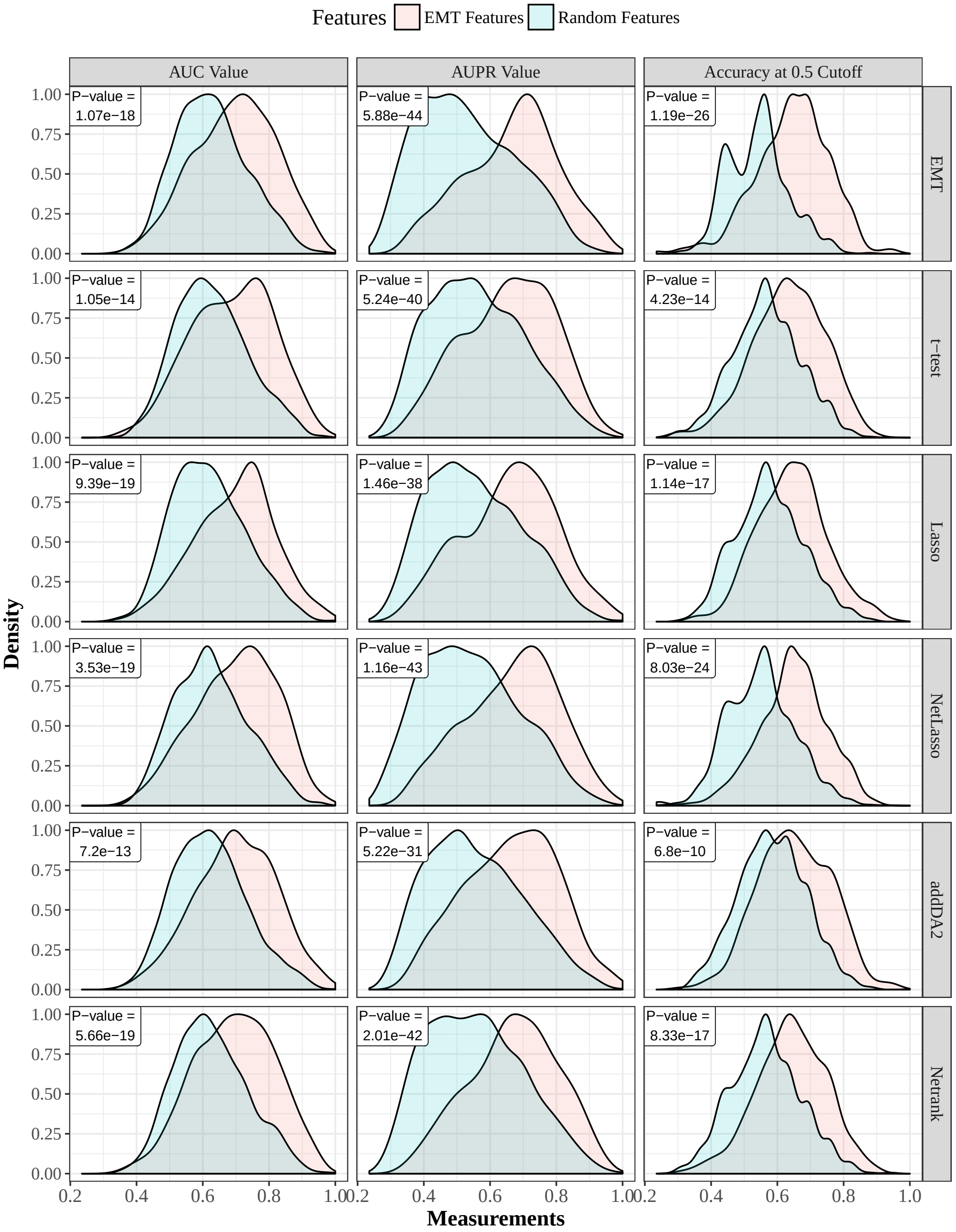


Figure 3.4: The AUC, AUPR, and accuracies of EMT features versus random features using DNA methylation data with the core EMT network. Gaussian kernel is used to estimate the density functions based on results from 30 times 10-fold cross-validation. For each cross-validation fold, EMT features and random features are tested on the same training and testing samples. The comparisons on five feature selection algorithms together with the comparative group of using all EMT features are shown. The p-values of paired t-tests are provided.

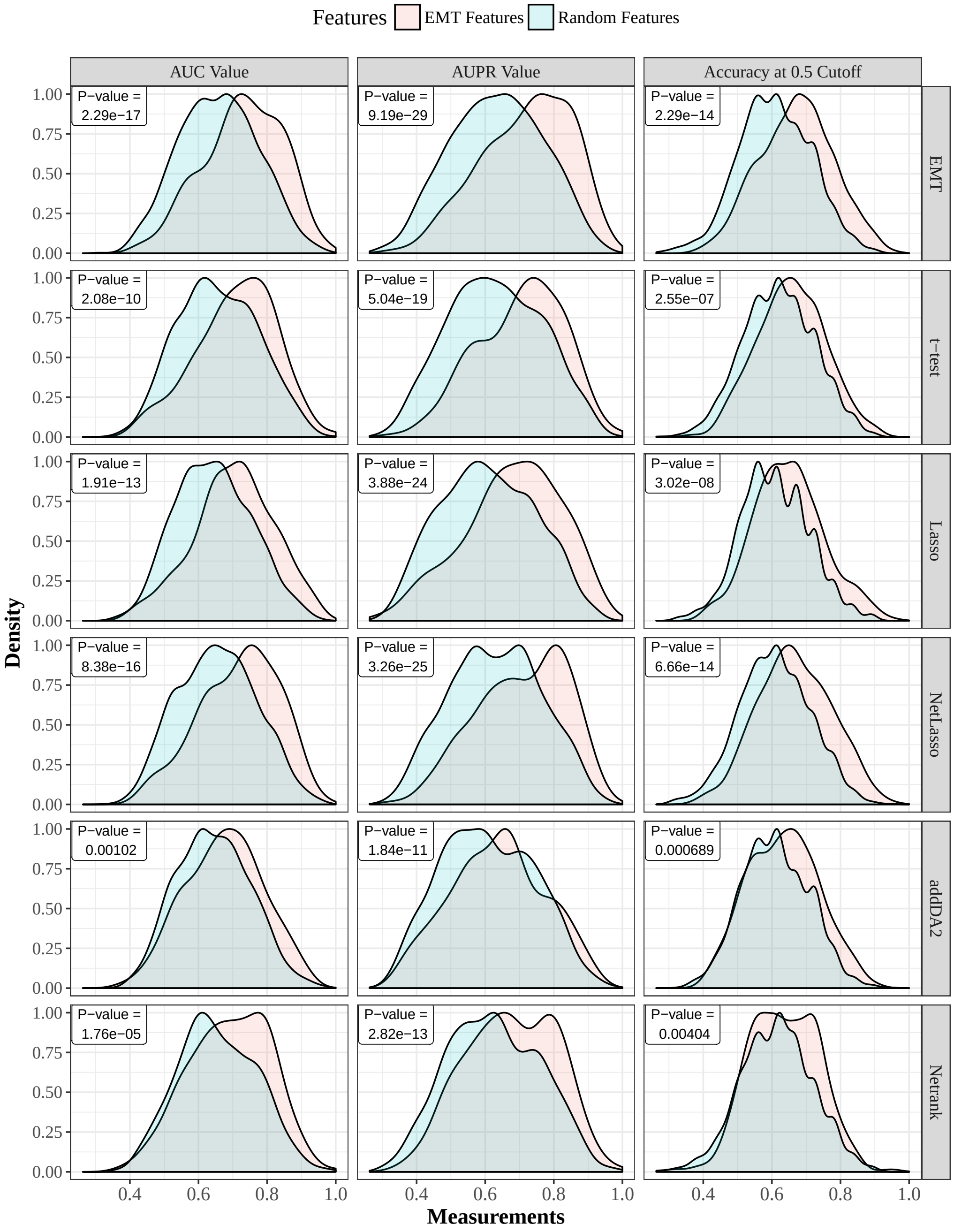


Figure 3.5: The AUC, AUPR, and accuracies of EMT features versus random features using gene expression data with filtered EMT network. Gaussian kernel is used to estimate the density functions based on results from 30 times 10-fold cross-validation. For each cross-validation fold, EMT features and random features are tested on the same training and testing samples. The comparisons on five feature selection algorithms together with the comparative group of using all EMT features are shown. The p-values of paired t-tests are provided.

Table 3.3: The average AUC and AUPR values of the 10 feature selection algorithms using SVM classifier. For each algorithm, we evaluated its prediction performance using 3 data levels. Within each data level, we used 3 different sizes of EMT networks.

Data Level	Gene expression			DNA Methylation			CNA			Metric
$\lvert V(G) \rvert$	74	123	455	74	123	455	70	117	445	
EMT	0.662	0.728	0.691	0.698	0.679	0.671	0.616	0.645	0.608	AUC
	0.627	0.719	0.651	0.666	0.648	0.652	0.521	0.608	0.506	AUPR
t-test	0.658	0.709	0.677	0.688	0.675	0.669	0.616	0.626	0.621	AUC
	0.612	0.702	0.650	0.654	0.645	0.647	0.570	0.603	0.578	AUPR
Lasso	0.616	0.703	0.620	0.697	0.666	0.667	0.615	0.619	0.617	AUC
	0.565	0.680	0.577	0.658	0.624	0.646	0.565	0.592	0.578	AUPR
NetLasso	0.659	0.718	0.686	0.700	0.678	0.677	0.619	0.635	0.621	AUC
	0.623	0.706	0.649	0.666	0.647	0.672	0.553	0.615	0.580	AUPR
addDA2	0.650	0.675	0.651	0.699	0.661	0.702	0.597	0.626	0.616	AUC
	0.611	0.636	0.607	0.662	0.610	0.678	0.530	0.616	0.579	AUPR
Netrank	0.656	0.691	0.668	0.695	0.685	0.693	0.615	0.619	0.610	AUC
	0.610	0.675	0.637	0.670	0.652	0.674	0.569	0.590	0.564	AUPR
stSVM	0.651	0.693	0.639	0.669	0.668	0.687	0.608	0.617	0.616	AUC
	0.617	0.669	0.591	0.621	0.631	0.664	0.531	0.554	0.552	AUPR
Cox	0.673	0.705	0.712	0.703	0.707	0.696	0.620	0.664	0.675	AUC
	0.623	0.686	0.673	0.667	0.673	0.671	0.541	0.642	0.659	AUPR
RegCox	0.648	0.698	0.729	0.696	0.717	0.666	0.645	0.669	0.653	AUC
	0.614	0.676	0.689	0.652	0.674	0.607	0.613	0.642	0.621	AUPR
MSS	0.662	0.694	0.659	0.674	0.654	0.640	0.608	0.627	0.625	AUC
	0.625	0.679	0.626	0.636	0.601	0.580	0.509	0.584	0.555	AUPR
Survnet	0.646	0.661	0.679	0.702	0.688	0.680	0.626	0.693	0.682	AUC
	0.611	0.650	0.654	0.664	0.637	0.628	0.574	0.673	0.645	AUPR

It is shown that applying Lasso feature selection on all the dimensions improved the predictions for DM data, but not for GE and CNA data.

5. Core EMT network gives more accurate predictions than the other two networks using DM data while filtered or extended EMT networks give more accurate predictions using GE data. Since the latter networks include the downstream genes of the core network, it is interesting to investigate the relationships between network scope and the data level that is more predictive.

To test the effect of classifiers on the prediction performance, we also applied random forest classifier using the selected features of each algorithm. The boxplot is shown in Figure 3.8. The average AUC and AUPR values are given in Table 7.2. We observe that with GE data both classifiers give similar performance, while random forest classifier performs better with using all 19290 features, which is still equivalent or lower than the prediction performance of EMT network-based features. With DM data, we observed less difference among the three EMT networks than using SVM classifier. The AUC values of all algorithms, especially when using the core EMT network, are lower than that of SVM classifier. With CNA data, we again observed similar performance as SVM classifier.

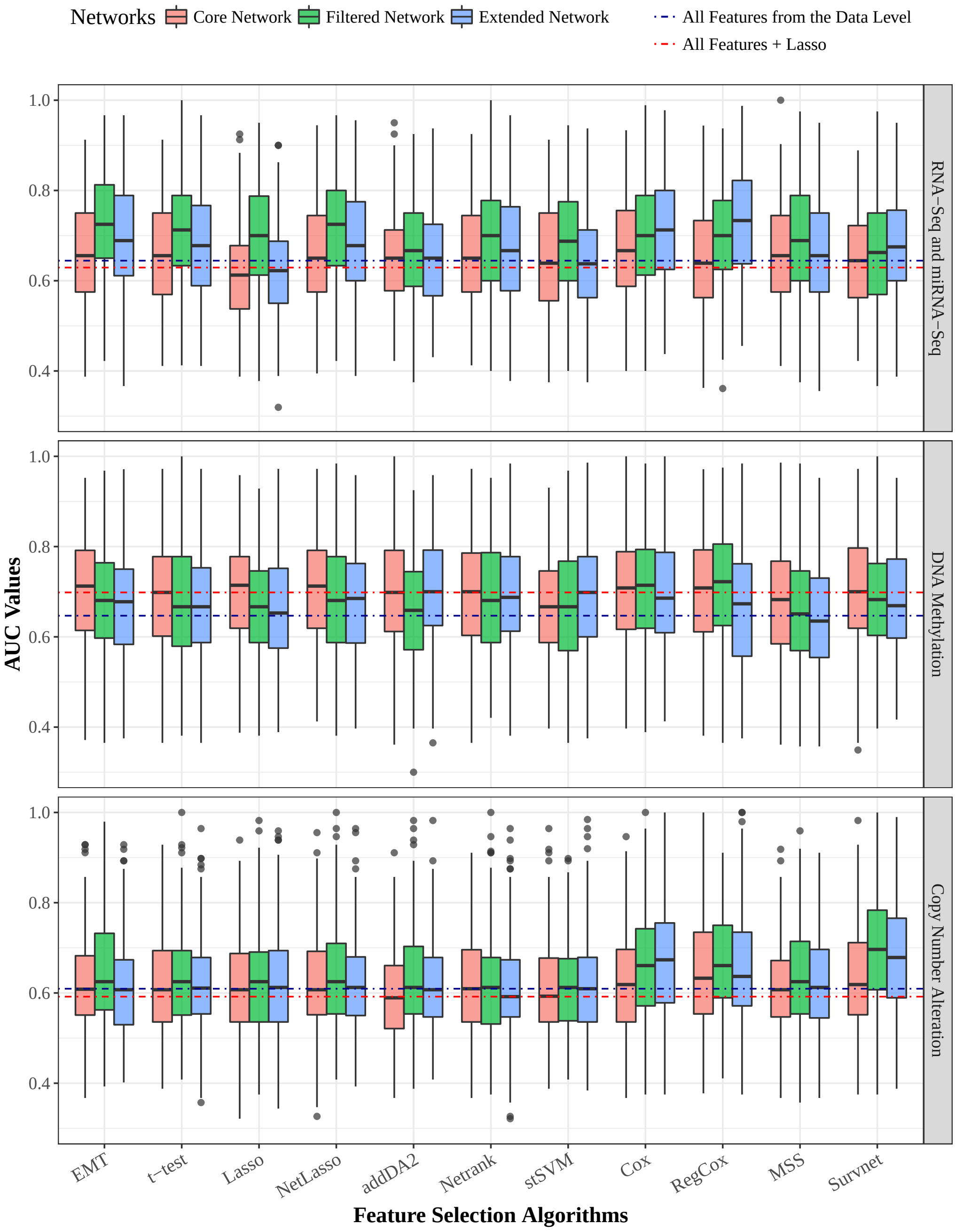


Figure 3.6: The AUC values of 10 feature selection algorithms using SVM classifier. The three panels correspond to three data levels. Within each panel, the AUC values of the 10 algorithms are plotted. Each algorithm has three boxes of different colors denoting the 3 EMT networks. The blue and red dotted lines within each panel are the median AUC values of two comparative groups: 1) using all data level features and 2) Lasso feature selection on all data level features.

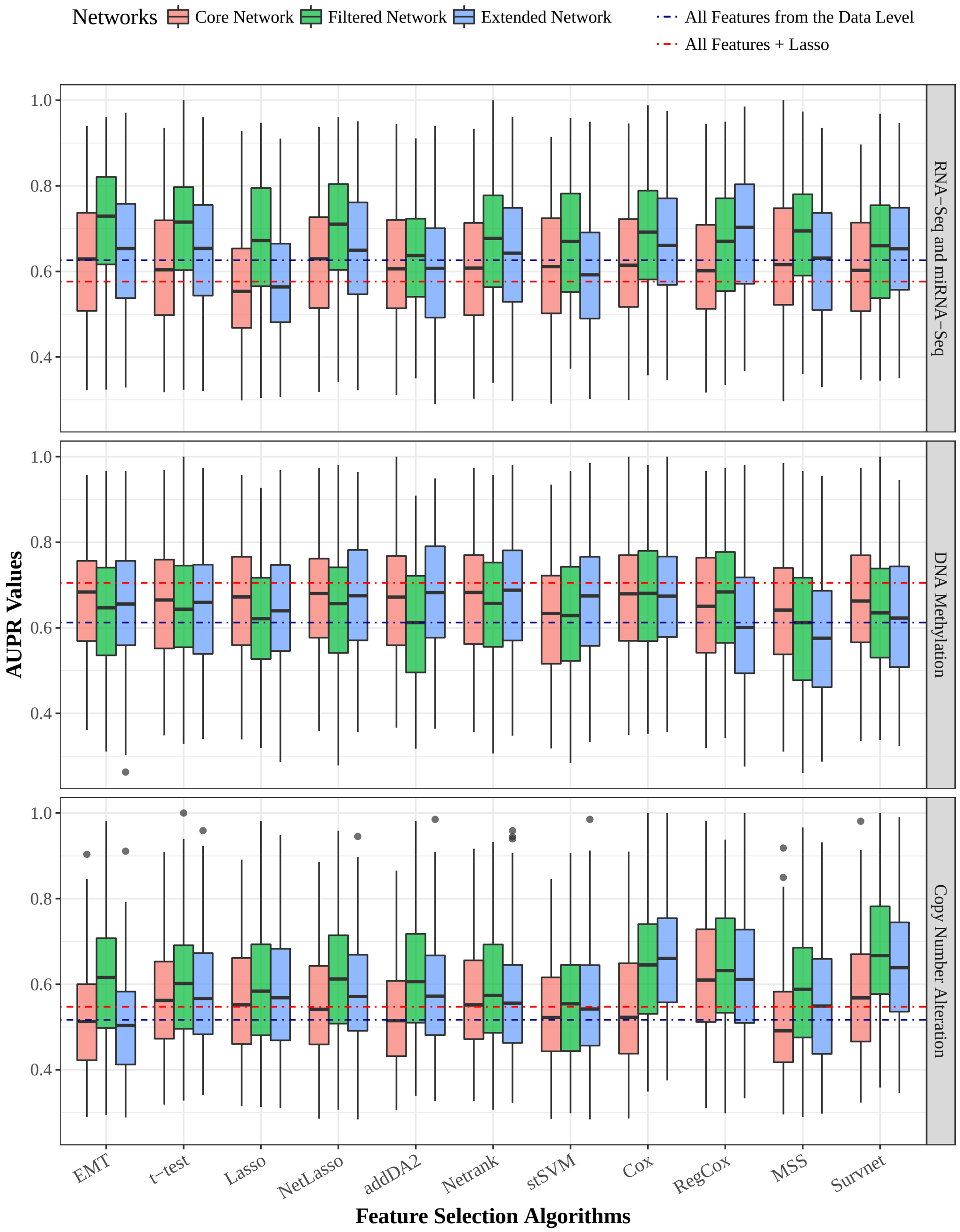


Figure 3.7: The AUPR values of 10 feature selection algorithms using SVM classifier. The three panels correspond to three data levels. Within each panel, the AUPR values of the 10 algorithms are plotted. Each algorithm has three boxes of different colors denoting the 3 EMT networks. The blue and red dotted lines within each panel are the median AUPR values of two comparative groups: 1) using all data level features and 2) Lasso feature selection on all data level features.

As also shown in our other experiments, random forest classifier gives comparable predictions as SVM classifier, with lower or higher prediction performance than SVM in some minority cases (certain combinations of networks and feature selection algorithms). Since random forest classifier gives different results upon repetition because of its randomness, our following discussions will be mainly based on the results from SVM classifier by default. When the performance of these two classifiers are significantly different for the majority of algorithms, we will discuss separately.

As shown in section 2.2.3 about evaluation metrics, accuracy measure is necessary because a higher AUC value does not necessarily guarantee accurate predictions at a certain cutoff. We have obtained the classification accuracy of each algorithm at different cutoffs. Since we observed that most of the algorithms achieved highest prediction accuracy at the cutoff of around 0.5 (as shown in Figure 7.2), we compared the algorithms' average prediction accuracy at 0.5 cutoff and presented it in Figure 3.9. We have observed the following:

1. EMT network-based features achieved equivalent or higher prediction accuracy than using all features from the corresponding data levels, with or without Lasso features selection.

2. CNA data give lower prediction accuracy than the other two data levels for most of the algorithms.

3. Larger network size does not always lead to better predictions. For the majority combinations of algorithms and data levels, the prediction accuracy does not correlate with network sizes.

3.3.3 Frequently Selected EMT Features

While it is difficult to select a single best feature selection algorithm, all evaluation metrics show high variance from the results of 30 times 10-fold cross-validation. It can be observed from both the density plots and box plots. This phenomenon has raised our concerns. It could be attributed to the high heterogeneity of the samples. Depending on the division of samples into training and testing sets, even very small variations can cause large difference in the predictions on the testing set. We show an example below in Figure 3.10 and Figure 3.11. These two different cross-validation runs are from the same set of samples. The two training sets have 89% of the samples in common. However, we observe the large difference in prediction performance on the testing set. We have also performed clustering analysis by firstly using t-SNE (t-distributed stochastic neighbor embedding) algorithm [1] to project the samples into two dimensions and then applying hierarchical clustering on 2D. We have performed this analysis for a set of ”good/bad pairs” of cross-validation runs but we have not found a common explanation for this phenomenon. What we have observed is that it occurs very often. This means that selecting biomarkers based on single cross-validation runs or a certain division of samples into training and testing set is highly unreliable.

To improve the stability of EMT signatures, we proposed to use the frequently selected features (FSFs) from all 30 times 10-fold cross-validation. We selected 20 FSFs and

[1] t-SNE is a nonlinear dimensionality reduction technique that is particularly suitable for embedding high-dimensional data into a space of two or three dimensions [160].

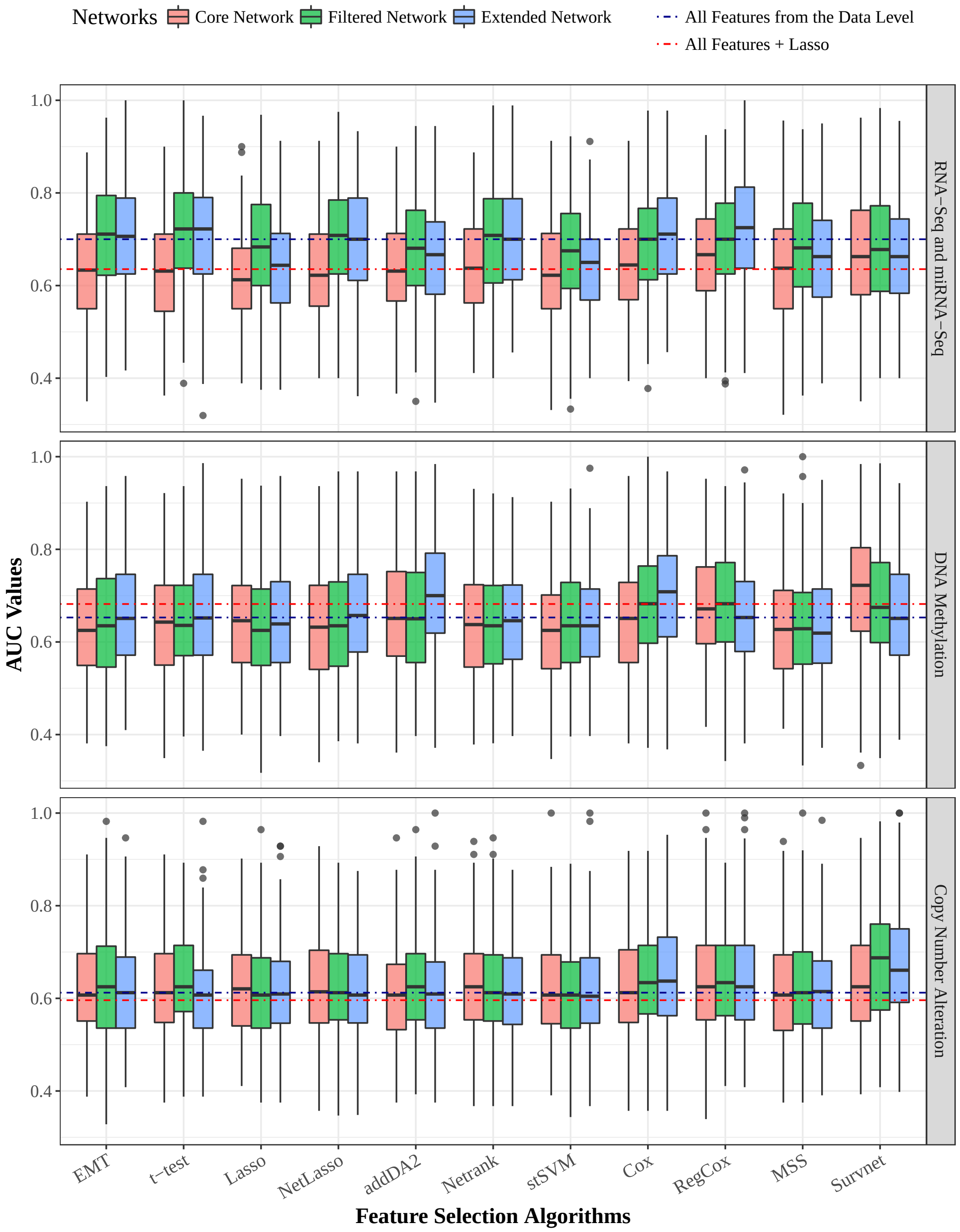


Figure 3.8: The AUC values of 10 feature selection algorithms using random forest classifier. The three panels correspond to three data levels. Within each panel, the AUC values of the 10 algorithms are plotted. Each algorithm has three boxes of different colors denoting the 3 EMT networks. The blue and red dotted lines within each panel are the median AUC values of two comparative groups: 1) using all data level features and 2) Lasso feature selection on all data level features.

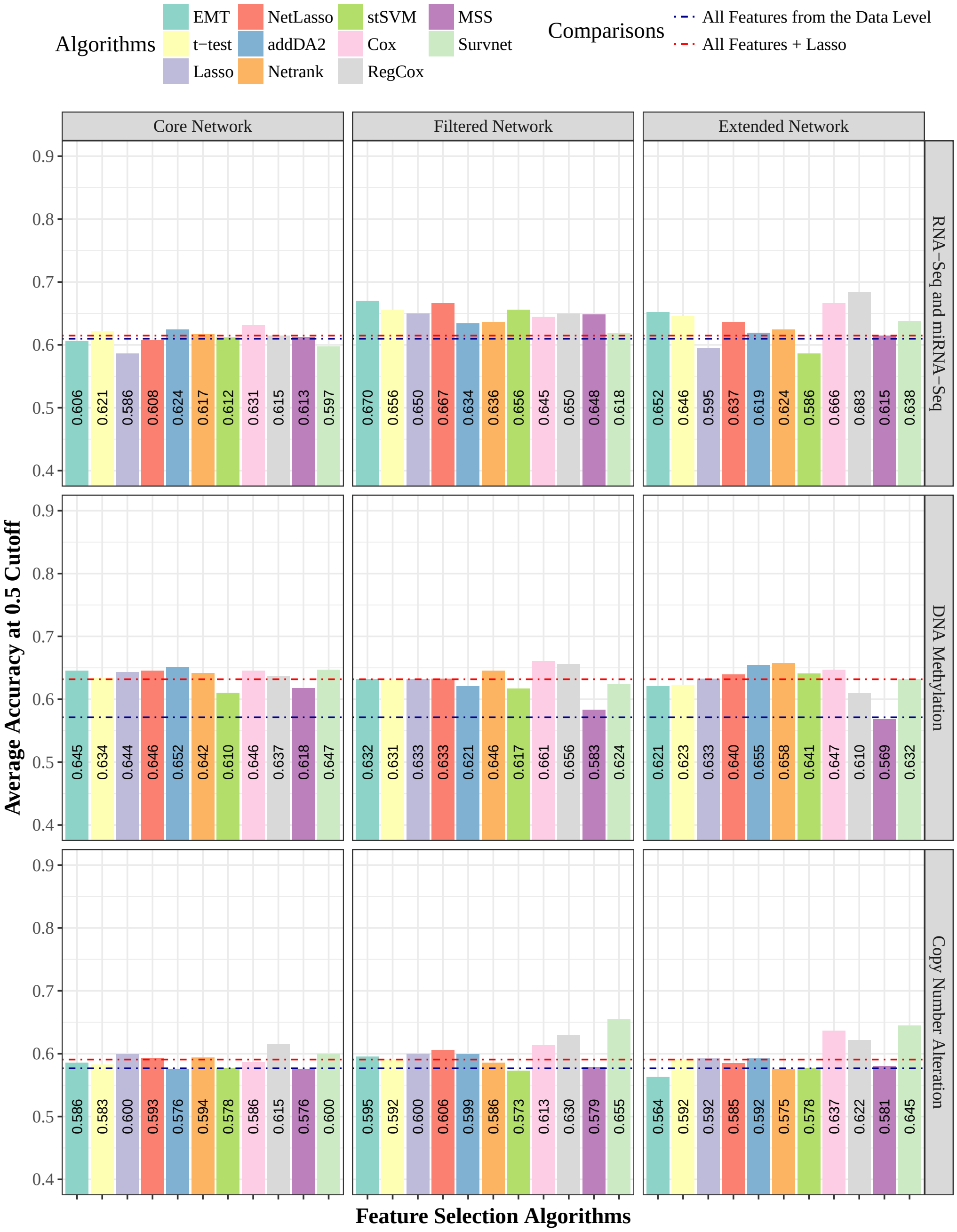


Figure 3.9: The prediction accuracy of the 10 feature selection algorithms at 0.5 cutoff using SVM classifier. The three panels correspond to three data levels. Within each panel, the average prediction accuracy of the 10 algorithms are plotted, which is grouped by EMT networks. The blue and red dotted lines within each panel are the average accuracies of two comparative groups: 1) using all data level features and 2) Lasso feature selection on all data level features.

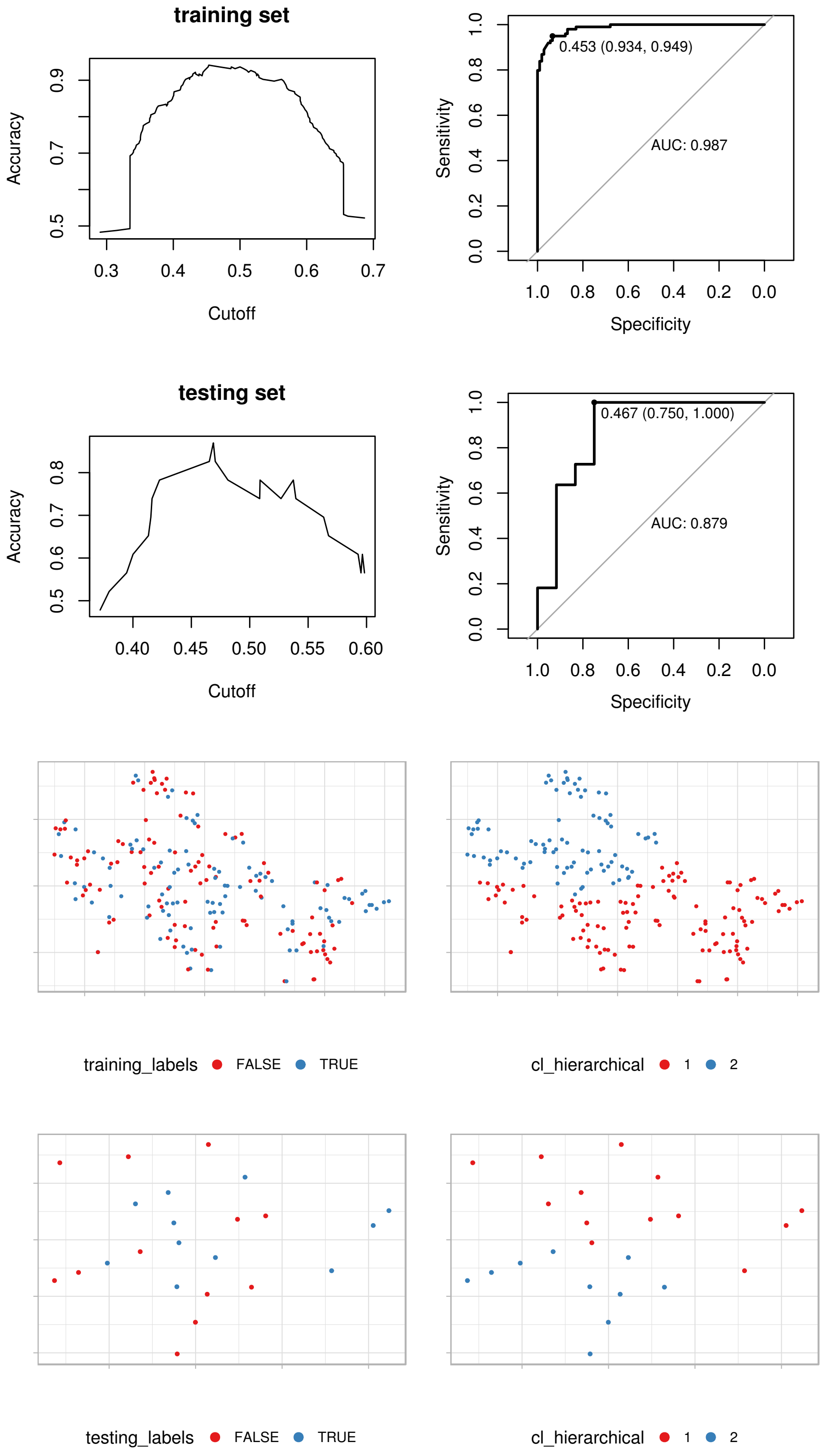


Figure 3.10: An example of cross-validation test that gives good prediction performance. The upper figure shows ROC curves on training and testing data. The lower figure shows clustering analysis on training and testing data.

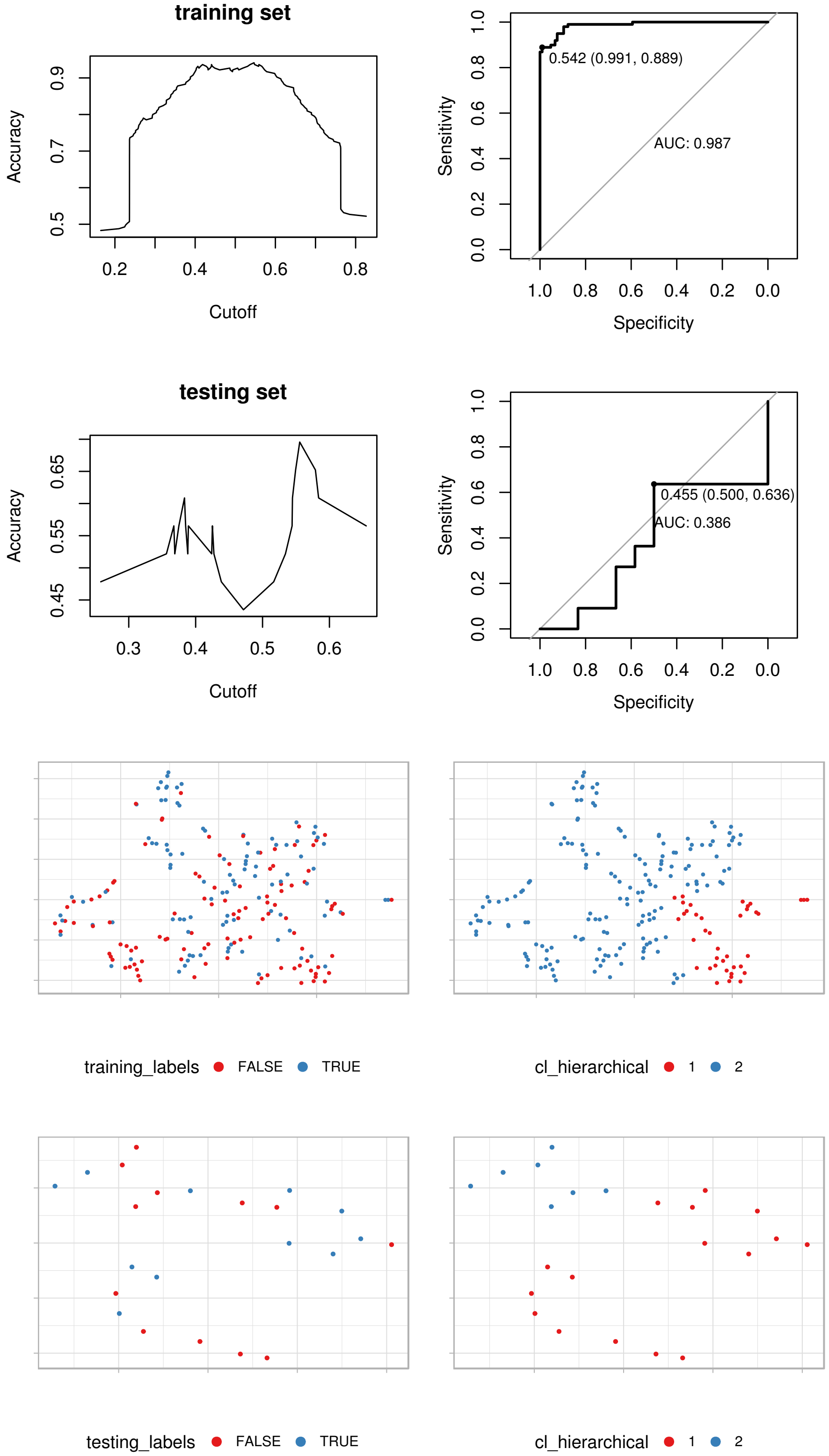


Figure 3.11: An example of cross-validation test that gives unsatisfactory prediction performance. The upper figure shows ROC curves on training and testing data. The lower figure shows clustering analysis on training and testing data.

performed 30 times 10-fold cross-validation to test their performance. The density plot of AUC, AUPR, and accuracy values are shown Figure 3.12, with the comparison of individually selected features. Since the sample folds are kept the same for both groups, we used paired t-test to compare their performance. The p-values are shown in each sub-figure. The corresponding boxplot of AUC and AUPR values is given in Figure 3.13.

The results show that FSFs outperform individually selected features very significantly (except Netrank algorithm). FSFs show much higher AUC, AUPR, and accuracy values and lower variance. The average AUC values for t-test, Lasso, NetLasso, and addDA2 have reached 0.773, 0.825, 0.796, and 0.833. This predictive performance is rather remarkable. Recall that we only used less than 2.5% of the original dimensionality, namely EMT features, for feature selection and prognosis prediction. This shows the predictive capability of EMT features, which is consistent with biological findings that EMT process is highly relevant to cancer prognosis. The results also suggest that using FSFs can effectively mitigate the effect of sample heterogeneity. We acknowledge that testing FSFs on the same set of samples causes dependencies. But considering the significant performance gain, we attribute most of the contributions to the stability of FSFs. In the next chapter we will show that on independent samples the FSFs can also give significant sample stratifications.

3.3.4 The Effect of Adding Clinical Features

Next, we combined the FSFs with clinical features to make predictions. We obtained the AUC values and average prediction accuracy and compared it with using only clinical features. The boxplot of AUC values and the barplot of the average prediction accuracy are shown in Figure 3.14 and Figure 3.15. Table 3.4 shows the average AUC and AUPR values. Random forest classifier shows different AUC values than SVM classifier. The boxplot of AUC values and the barplot of the average prediction accuracy are shown in Figure 7.3 and Figure 7.4. The average AUC and AUPR values are shown in the Table 7.3. Note that the sample sizes for each data level are different from that of section 2.3.1 because we can only use the samples who have both the molecular data and the clinical data. Due to the same reason, the performance of clinical features in the three panels is also different as the sample sizes differ. In this way, the performance of clinical features within each panel is comparable with other feature groups. From these results we have the following observations:

1. Combining frequently selected EMT features with clinical features can significantly outperform clinical features. Some combinations of data levels, networks and algorithms achieved average AUC values of above 0.8. However, in some cases, the combination of features leads to worse prediction performance than using only the FSFs (as shown in Figure 3.13).

2. Lasso, NetLasso, and addDA2 algorithms give better performance than the other feature selection algorithms using GE and DM data. Using CNA data, combining features mostly does not significantly improve prediction performance.

3. It can be seen from both Figure 3.13 and Figure 3.14 that most of the contributions come from FSFs, and not the clinical features for the top performing feature selection algorithms.

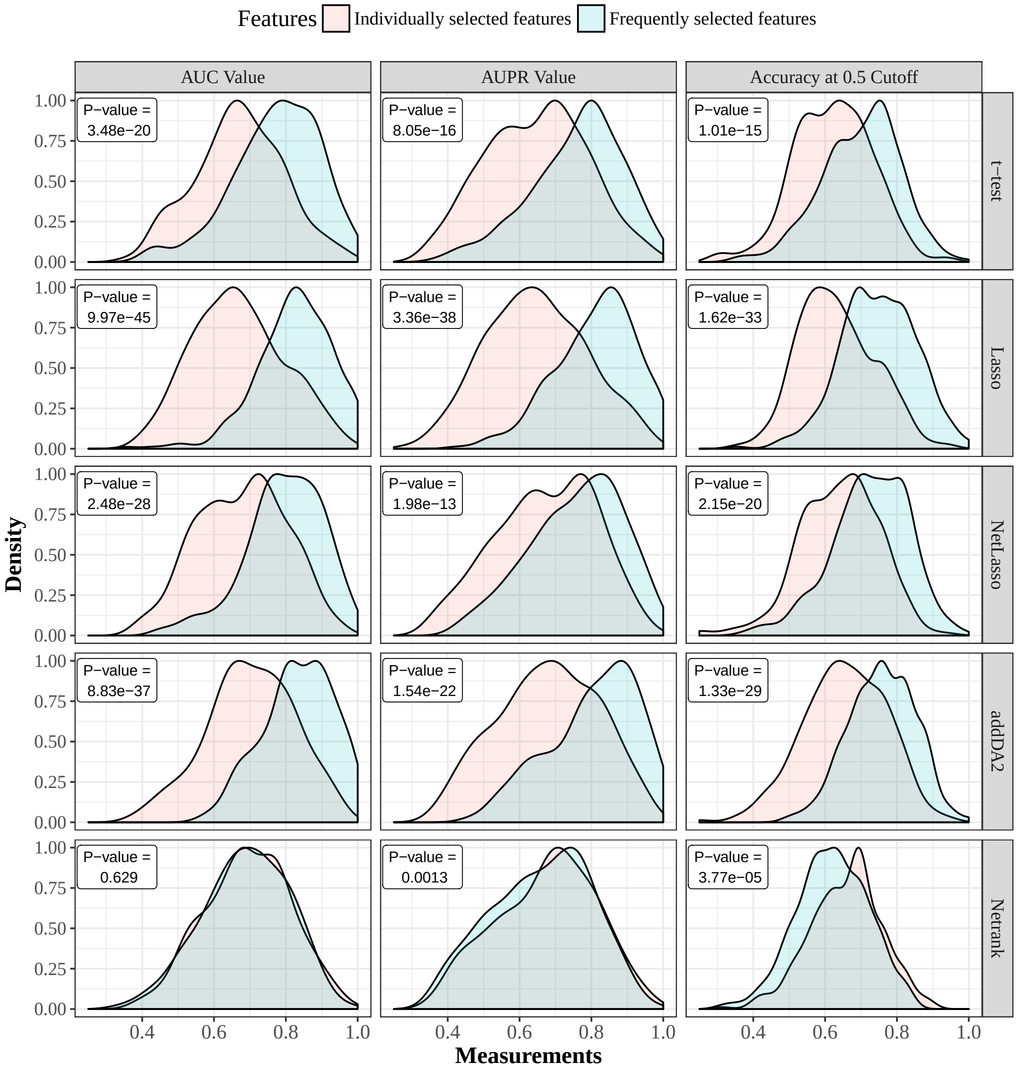


Figure 3.12: The comparison of FSFs with individually selected features in terms of AUC, AUPR, and accuracy values. We used DNA methylation data and extended EMT network for feature selection and SVM classifier for classification. Gaussian kernel is used to estimate the density functions based on results from 30 times stratified 10-fold cross-validation. For each cross-validation iteration, individually selected features and FSFs are tested on the same training and testing samples. The comparison between the two feature groups is shown on five feature selection algorithms together with the p-values of paired t-tests.

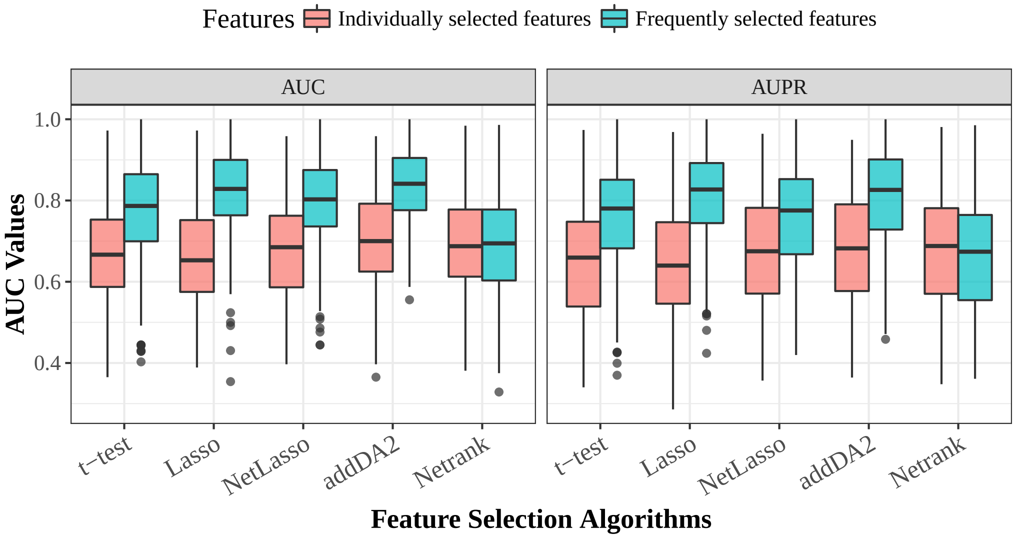


Figure 3.13: The comparison of FSFs with individually selected features in terms of AUC and AUPR values. We used DNA methylation data and extended EMT network for feature selection and SVM classifier for classification. The boxplot is based on the results from 30 times stratified 10-fold cross-validation. The comparison between the two feature groups is shown on five feature selection algorithms.

4. Comparing Figure 3.14 with Figure 7.3, random forest classifier significantly improved the AUC values of some algorithms using GE data and DM data, especially for addDA2 algorithm. Comparing Figure 3.15 with Figure 7.4, random forest algorithm gave much better prediction accuracy for most of the algorithms using GE data and DM data. It suggests that random forest classifier can do better with mixed features - both numerical features and categorical features.

We further analyzed which variables in clinical features are relatively important. We used a few methods to measure the feature importance, as given below:

1. The variable importance measure using random forest. It is based on the mean decrease in Gini impurity after the splitting of a node. The Gini impurity index is defined as $G = \sum_{i=1}^{n_c} p_i(1 - p_i)$, where n_c is the number of classes in the target variable and p_i is the ratio of this class. The Gini decrease is calculated as $I = G_{parent} - G_{split1} - G_{split2}$. The Gini decrease of a feature is averaged over all splits that are based on this feature in the forest. Since clinical variables are categorical and with different number of levels, in which case random forests are biased to favor features with more levels [231], we transformed the 8 clinical variables into 25 dummy variables where each dummy variable corresponds to one level of the original variables. For age variable, we binned it into three levels: young (30 to 50), middle (50 to 70), and old (70 to 90). The dot chart of feature importance is given in Figure 3.16a.

2. Statistical measures including AUC values, Spearman correlation, and p-values of univariate Cox model for each clinical variable (shown in Figure 3.16b)

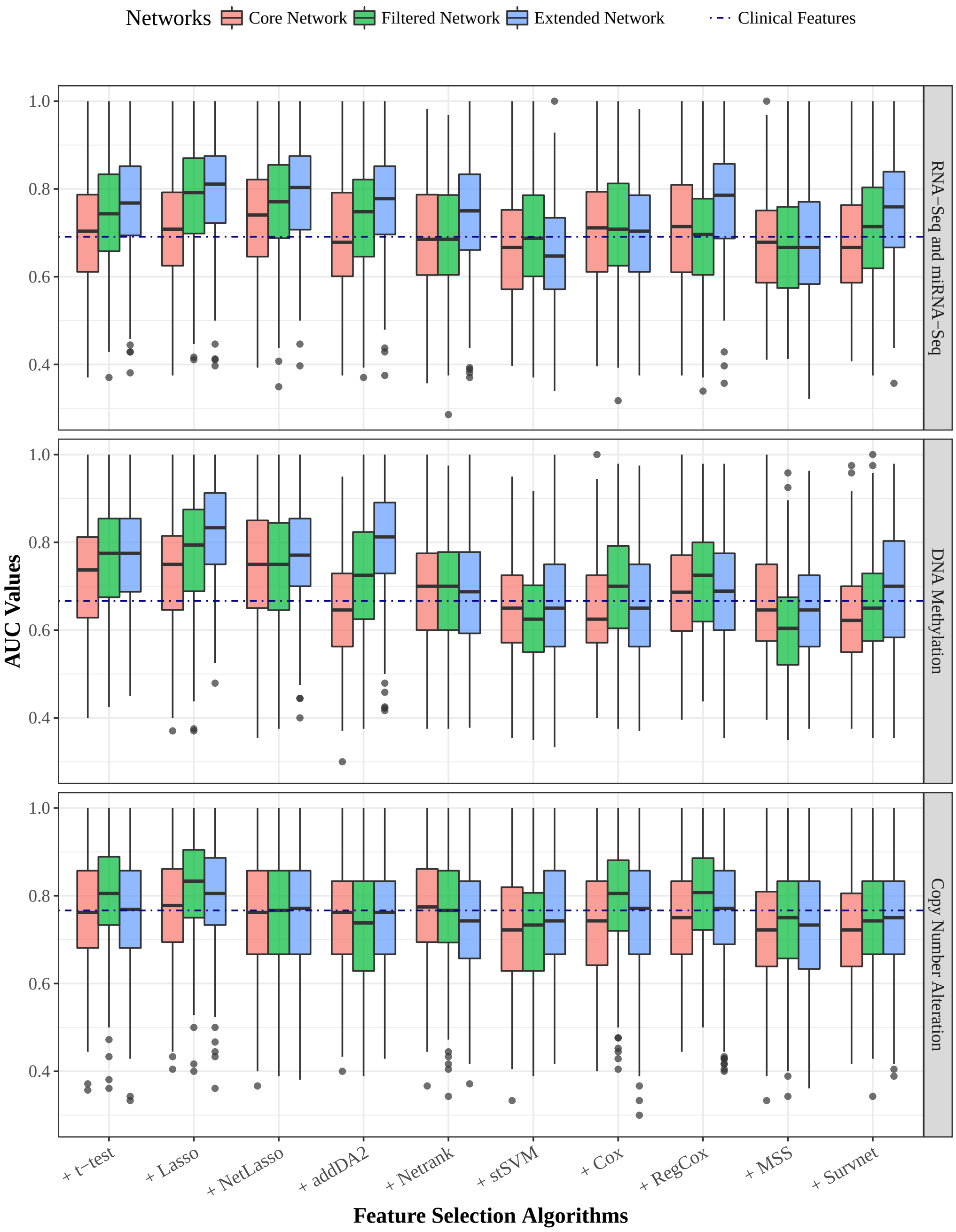


Figure 3.14: The average AUC values of FSFs combined with clinical features for three data levels and three EMT networks using SVM classifier. The blue dotted lines show the median AUC values of using only clinical features.

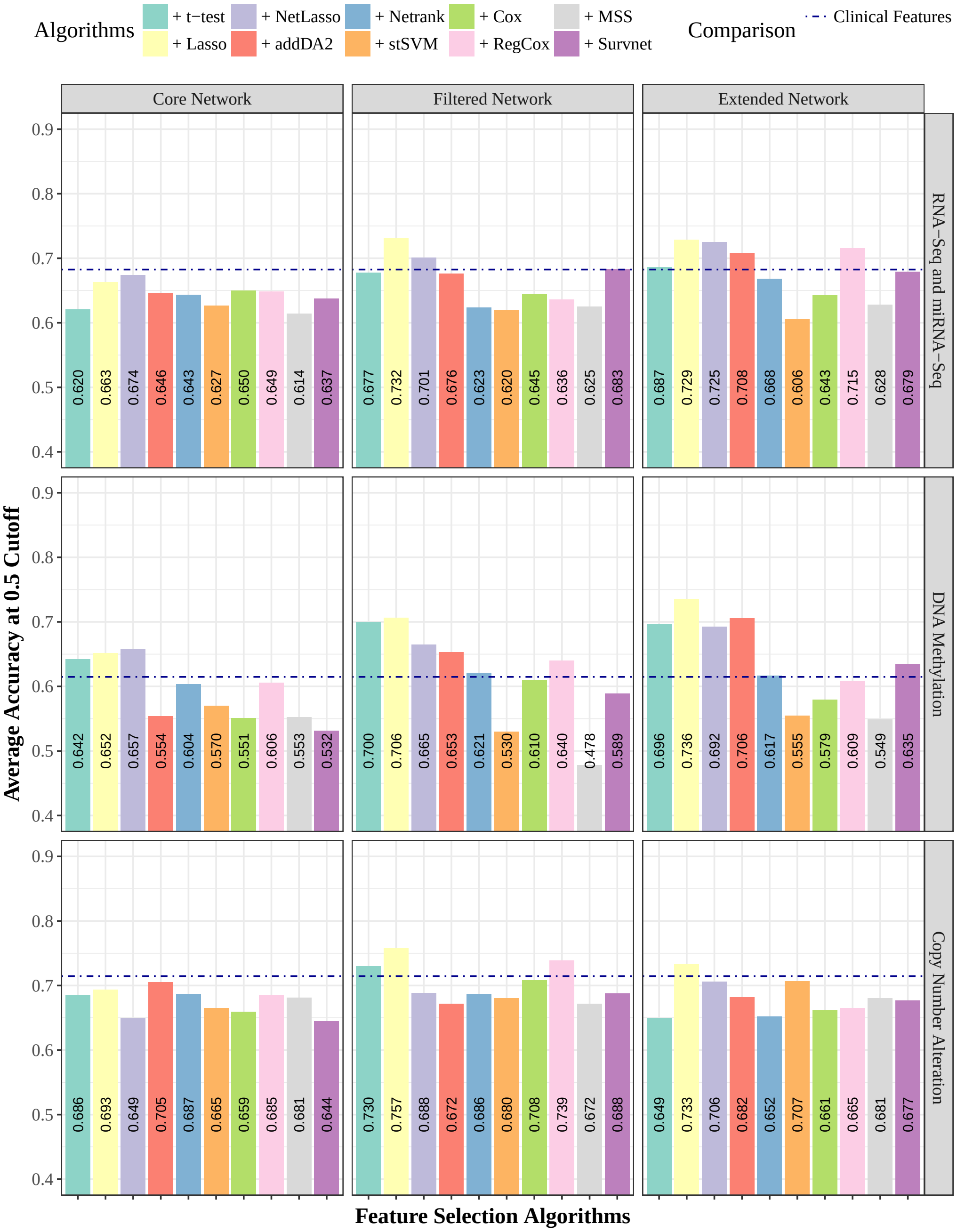
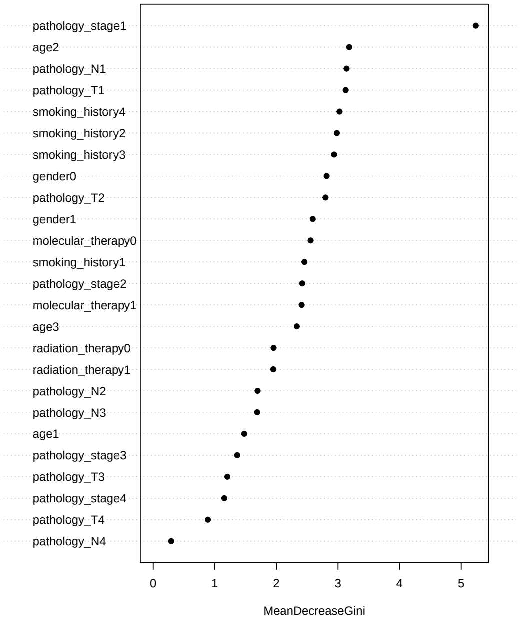


Figure 3.15: The average prediction accuracy at 0.5 cutoff of FSFs combined with clinical features for three data levels and three EMT networks using SVM classifier. The blue dotted lines are the median accuracy values of using only clinical features.

Table 3.4: The average AUC and AUPR values of FSFs together with clinical features using SVM classifier. For each algorithm, we evaluated its prediction performance using 3 data levels. On each data level, we used 3 different sizes of EMT networks.

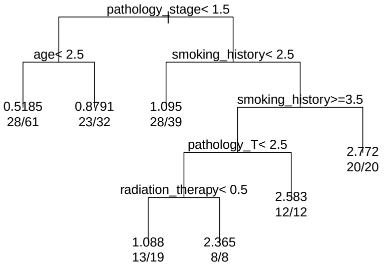
Data Level	Gene expression			DNA Methylation			CNA			Metric
$\|V(G)\|$	74	123	455	74	123	455	70	117	445	
clinical	0.695			0.682			0.756			AUC
	0.651			0.596			0.751			AUPR
t-test	0.701	0.735	0.762	0.724	0.766	0.770	0.755	0.806	0.756	AUC
	0.648	0.697	0.736	0.626	0.692	0.690	0.697	0.767	0.733	AUPR
Lasso	0.711	0.777	0.793	0.729	0.773	0.824	0.776	0.824	0.801	AUC
	0.648	0.761	0.777	0.633	0.700	0.756	0.704	0.780	0.753	AUPR
NetLasso	0.731	0.762	0.787	0.743	0.739	0.768	0.752	0.756	0.762	AUC
	0.690	0.738	0.761	0.644	0.672	0.689	0.684	0.703	0.730	AUPR
addDA2	0.684	0.733	0.772	0.648	0.717	0.794	0.752	0.724	0.749	AUC
	0.629	0.671	0.745	0.495	0.646	0.714	0.694	0.678	0.697	AUPR
Netrank	0.693	0.688	0.741	0.692	0.685	0.693	0.773	0.766	0.738	AUC
	0.638	0.648	0.712	0.563	0.604	0.581	0.708	0.714	0.690	AUPR
stSVM	0.669	0.689	0.660	0.648	0.630	0.656	0.722	0.721	0.755	AUC
	0.628	0.639	0.592	0.528	0.450	0.510	0.659	0.667	0.705	AUPR
Cox	0.705	0.706	0.701	0.647	0.694	0.662	0.734	0.791	0.756	AUC
	0.649	0.683	0.643	0.500	0.595	0.568	0.676	0.738	0.739	AUPR
RegCox	0.706	0.690	0.770	0.684	0.712	0.687	0.758	0.799	0.760	AUC
	0.642	0.655	0.717	0.589	0.608	0.574	0.687	0.752	0.714	AUPR
MSS	0.674	0.671	0.671	0.657	0.610	0.645	0.728	0.741	0.734	AUC
	0.621	0.631	0.655	0.501	0.370	0.496	0.683	0.707	0.699	AUPR
Survnet	0.678	0.706	0.750	0.625	0.658	0.693	0.718	0.744	0.747	AUC
	0.603	0.666	0.714	0.455	0.575	0.634	0.673	0.702	0.719	AUPR



(a) Mean Gini decrease

	Spearman's rho	AUC	Cox p-value
pathology_T	-0.15	0.43	4.39e-3
pathology_N	-0.28	0.64	8.62e-5
pathology_stage	-0.33	0.68	3.57e-7
smoking_history	-0.058	0.47	7.41e-1
age	-0.02	0.51	3.81e-1
gender	-0.068	0.47	4.87e-1
radiation_therapy	-0.16	0.44	8.15e-4
molecular_therapy	-0.12	0.45	1.28e-1

(b) Statistical measures



(c) Recursive partitioning

Figure 3.16: Assessment of the importance of individual clinical features. We calculated the relative importance of features using different methods: (a) the mean Gini decrease over all trees in the random forest (b) the Spearman correlation, AUC value, and the p-value of univariate Cox PH model, and (c) the recursive partitioning of samples based on event rate.

3. Recursive partitioning for survival trees [242, 243]. The method can build either classification or regression trees using a two-stage procedure. In the first stage the tree is built based on a splitting criterion and in the second stage the tree is pruned using cross-validation. Here we used event rate data to build a regression tree to divide the samples into groups of different event rates. The result is shown in Figure 3.16c. At each leaf node, one can observe the number of events (death) and the relative death rate compared to the overall rate. For example, for cancer in stage 1 with age <70, the relative death rate is 0.519. In contrast, for cancer in later stages and with the smoking history (ranges from 1 to 5) ≥ 4, the relative death rate increases to 2.772.

3.3.5 The Effect of Classification Thresholds

As mentioned in the Experiments section, we performed the same experimental procedure using four different thresholds that are in the order of increasing discrepancy: 3 years, <900 or >1200 days, <700 or >1400 days, <500 or >1500 days. DM data with core EMT network were used for the testing. Figure 3.17 shows the AUC values of the 10 feature selection algorithms with four different thresholds using both SVM and random

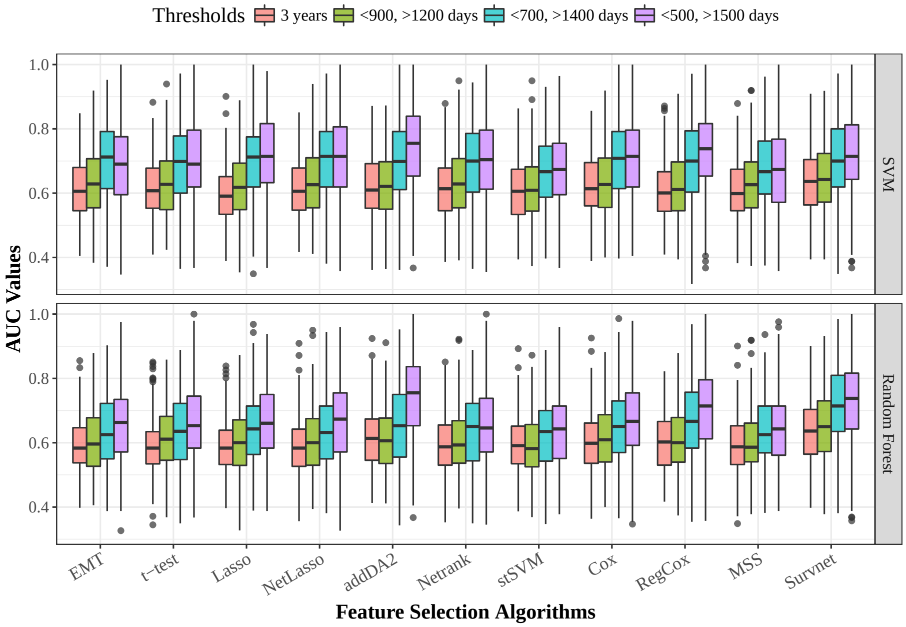


Figure 3.17: The AUC values of 10 feature selection algorithms using different thresholds for good prognosis class and poor prognosis class. The data level is DNA methylation data. The network is EMT core network.

forest classifiers. The corresponding AUPR values show similar trend and can be found in the Figure 7.5. The average prediction accuracies with these four thresholds are shown in Figure 3.18. We have observed the following:

1. Classification thresholds have an obvious effect on the AUC values and accuracy. The more discrepant the threshold is, the higher is the AUC value and the prediction accuracy. This apply to all feature selection algorithms. For all algorithms, the latter two thresholds gave significantly higher AUC values and prediction accuracy than the former two less discrepant thresholds.

2. The algorithms differ from each other in how sensitive they are to the effect of thresholds. For example, addDA2, RegCox, and Survnet algorithms are more sensitive to the effect of thresholds, compared with t-test and MSS algorithms.

3.4 Analysis

3.4.1 Stability of Feature Selection Algorithms

It is shown that the results of 30 times 10-fold cross-validation have high variability. Although the overlap of training instances among different training folds is around 78%, the prediction performance on the testing set can vary greatly from one to another. Therefore, we would like to look closer into the selected features and quantify how sensitive these

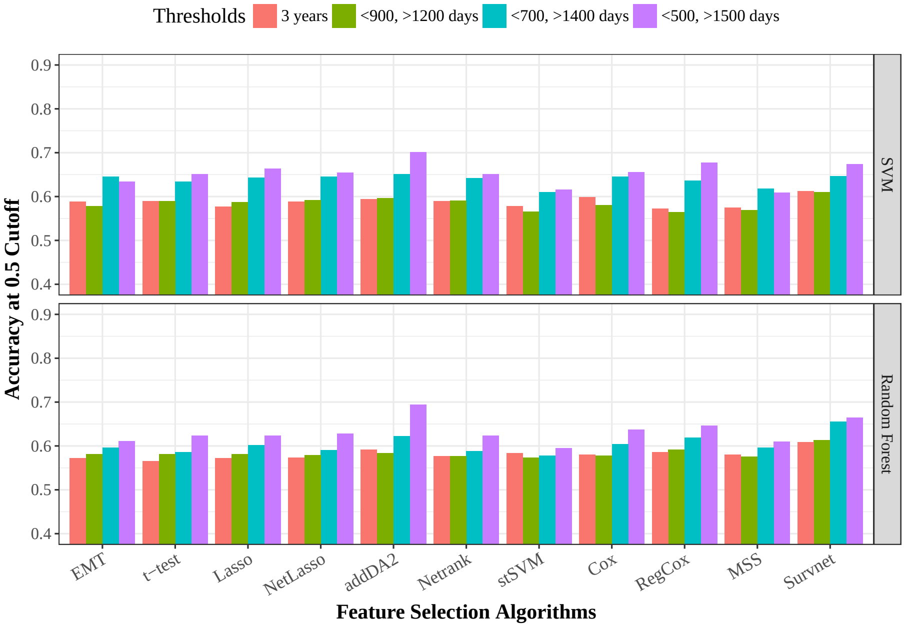


Figure 3.18: The average prediction accuracy of 10 feature selection algorithms using different thresholds for good prognosis class and poor prognosis class. The data level is DNA methylation data. The network is EMT core network.

feature selection algorithms are to the variations in the training set. [122, 123] proposed the notion of feature preference stability to quantify the sensitivity of feature selection algorithms to differences in training sets drawn from the same distribution. The feature preference can take different forms, e.g., a subset of selected features, a weighting-scoring, or the ranking of features. Especially, the motivation of investigating the stability of feature selection algorithms is particularly strong in biomarker discovery because the data are usually high-dimensional, and the selected molecular signatures are required to be stable - high overlap of selected features given variations on the training set. The reason is that if the selected features vary too much upon slight variations in the training data, then it is difficult for domain experts to have confidence in the molecular signatures.

The measurements of the similarity for feature preferences are defined for feature selection algorithms that provide weighting-scoring, ranking, or feature subsets [122, 123]. Let the feature vector $f = (f_1, f_2, ..., f_m)$, the similarity of feature preferences is calculated accordingly for the three cases:

- A weighting vector: $w = (w_1, w_2, ..., w_m), w \in W \subseteq \mathbb{R}^m$. In this case, the similarity between two weightings w, w' is calculated using the Pearson's correlation coefficient.

$$S_W(w, w') = \frac{\sum_i (w_i - \mu_w)(w'_i - \mu_{w'})}{\sqrt{\sum_i (w_i - \mu_w)^2 \sum_i (w'_i - \mu_{w'})^2}}$$

- A ranking: $r = (r_1, r_2, ..., r_m), 1 \leq r_i \leq m$. In this case, the similarity between two rankings r, r' is calculated using Spearman's rank correlation coefficient.

$$S_R(r, r') = 1 - 6 \sum_i \frac{(r_i - r'_i)^2}{m(m^2 - 1)}$$

- A subset of features: $s = (s_1, s_2, ..., s_m), s_i \in \{0, 1\}$, with 0 indicating absence of a feature and 1 presence. The similarity between two subsets s, s' (bitmaps) is calculated using Tanimoto similarity [64, 205].

$$S_S(s, s') = \frac{\sum_i (X_i \wedge Y_i)}{\sum_i^{(} X_i \vee Y_i)}$$

, where $\wedge, \vee$ are bitwise and, or operators.

As seen from the definitions, S_S focuses on a given number of top ranked features while S_W and S_R consider all features. In biomarker identification, S_S is of greater interest since we need to have small set of features as markers. Therefore, we use S_S to measure the feature preference similarity for each feature selection algorithm on each pair of cross-validation tests. We set the subset size of 10 for the core network, 15 for the filtered network, and 20 for the extended network . Since we applied 30 times 10-fold cross-validation, there are $300(300 - 1)/2$ pairs. Then we take the average similarity over all pairs as the final stability score. We picked the network that gave the highest prediction performance to evaluate the stability. We have also calculated the score for each comparative group. The results are shown in Table 3.5.

In biomarker discovery it is desired to find a feature selection algorithm that has both high stability and good prediction performance. Thus, we copied the average AUC values of algorithms from Table 3.3 into Table 3.5 for a convenient comparison. We observed the following:

1. EMT network-based feature selection achieved significantly higher stability than using all features plus Lasso from the corresponding data level, while having better prediction performance. Especially when using all EMT features on DNA methylation data using core EMT network, the stability score is 1, while having AUC value of 0.698. Table 3.5 shows that in many cases EMT feature selection achieved much higher prediction performance than using all data level features while having higher stability scores. Therefore, from the perspectives of both feature stability and prediction performance, EMT network shows its value in biomarker selection.

2. The goodness of the 10 feature selection algorithms can be hard to distinguish, which depends on the trade-off between AUC and stability score. stSVM and t-test feature selection show higher stability than other algorithms. Note that both algorithms are based on t-test while the former smoothed the t-statistic on a network. t-test feature selection achieved higher AUC values than stSVM and around half of other feature selection algorithms in each column of the table. Thus, if one takes into account both stability and prediction performance, t-test is a good choice, although it usually does not lead to the best AUC scores.

Because of the high-dimensionality of data, it is widely acknowledged that molecular signatures produced by different studies usually differ widely and have few genes in common

Table 3.5: The stability scores of feature selection algorithms. For each pair of cross-validation tests we calculated the Tanimoto similarity and averaged the scores from all pairs.

Data	Gene expression			DNA Methylation			CNA			Metric
$\lvert V(G)\rvert$	74	123	455	74	123	455	70	117	445	
EMT	1			1			1			stability
	0.662	0.728	0.691	0.698	0.679	0.671	0.616	0.645	0.608	AUC
t-test	0.721	0.762	0.609	0.747	0.505	0.496	0.656	0.550	0.564	stability
	0.658	0.709	0.677	0.688	0.675	0.669	0.616	0.626	0.621	AUC
Lasso	0.299	0.443	0.238	0.539	0.359	0.347	0.315	0.336	0.211	stability
	0.616	0.703	0.620	0.697	0.666	0.667	0.615	0.619	0.617	AUC
NetLasso	0.618	0.562	0.324	0.692	0.408	0.198	0.546	0.380	0.396	stability
	0.659	0.718	0.686	0.700	0.678	0.677	0.619	0.635	0.621	AUC
addDA2	0.479	0.538	0.245	0.559	0.314	0.258	0.454	0.314	0.223	stability
	0.650	0.675	0.651	0.699	0.661	0.702	0.597	0.626	0.616	AUC
Netrank	0.577	0.652	0.402	0.728	0.582	0.600	0.626	0.474	0.565	stability
	0.656	0.691	0.668	0.695	0.685	0.693	0.615	0.619	0.610	AUC
stSVM	0.639	0.838	0.858	0.780	0.695	0.869	0.833	0.646	0.775	stability
	0.651	0.693	0.639	0.669	0.668	0.687	0.608	0.617	0.616	AUC
Cox	0.597	0.690	0.613	0.682	0.580	0.483	0.641	0.558	0.492	stability
	0.673	0.705	0.712	0.703	0.707	0.696	0.620	0.664	0.675	AUC
RegCox	0.473	0.505	0.424	0.518	0.413	0.281	0.455	0.425	0.316	stability
	0.648	0.698	0.729	0.696	0.717	0.666	0.645	0.669	0.653	AUC
MSS	0.179	0.304	0.062	0.404	0.095	0.055	0.276	0.162	0.075	stability
	0.662	0.694	0.659	0.674	0.654	0.640	0.608	0.627	0.625	AUC
Survnet	0.418	0.220	0.164	0.176	0.453	0.318	0.298	0.371	0.287	stability
	0.646	0.661	0.679	0.702	0.688	0.680	0.626	0.693	0.682	AUC
All features + Lasso			0.228			0.261			0.136	stability
			0.648			0.652			0.612	AUC

[66, 67, 150, 171, 258]. This can already be observed using the same dataset with 10-fold cross-validation. This lack of agreement raised doubts about the reliability of the identified biomarkers. Using probably approximately correct (PAC) sorting algorithm, [67] shows that to achieve a typical overlap of 50% between two sets of prognostic markers, one would need the molecular profiles of several thousand patients, which is very challenging in clinical settings. Here we propose that by using a prognosis relevant network (EMT), one can increase the feature overlap significantly and meanwhile obtaining similar or even better prediction performance.

3.4.2 Patterns in EMT Signatures

One thing that is of top interest is to analyze which features are selected in different data levels and what do they indicate. We will first look at the network properties of the FSFs. Then we investigate how these features relate to each other by using association rule mining method.

Network Properties of FSFs

Cancer is nowadays acknowledged as a disease that involve dysregulation of multiple pathways that function in a complex GRN [80, 137]. Graph theory has been increasingly employed to better understand the network properties [168,191]. Here we use a few commonly used centrality measures to characterize the FSFs at each data level. These measures are introduced below:

- Degree centrality. It shows how many interactions is a node involved in. For a node i, the degree centrality $C_d(i) = deg(i)$. Nodes with high centrality are called *hubs* since they are connected to many neighbors.

- Betweenness centrality. It gives higher rank to nodes that lie on a high proportion of paths between other nodes in the network. These nodes are important for other nodes to communicate with each other. For distinct nodes $i, j, w \in V(G)$, let σ_{ij} be the total number of shortest paths between i and j and $\sigma_{ij}(w)$ be the number of shortest paths from i to j that pass through w. The betweenness centrality is calculated as $C_b(w) = \sum_{i \neq j \neq w} \frac{\alpha_{ij}(w)}{\alpha_{ij}}$.

In Table 3.6 and Table 3.7, we show the average C_d and C_b of the FSFs of feature selection algorithms, grouped by data levels and network sizes. We used Wilcoxon signed-rank test to test the differences between each pair of different data levels with the same network size. The results of the test are given in Table 3.8. We observe that when the network is small, there is hardly significant differences among the three data levels. When the size of network increases, features selected from DM data have significantly higher C_d and C_b than the other two data levels, while the difference between GE and CNA data may or may not be significant.

We would like to further investigate why this pattern emerges. As the features are selected according to their predictive capability of cancer prognosis, this suggests that the important features on different data levels have different properties. To give an intuition

Table 3.6: Average degree centrality of the FSFs of 10 feature selection algorithms.

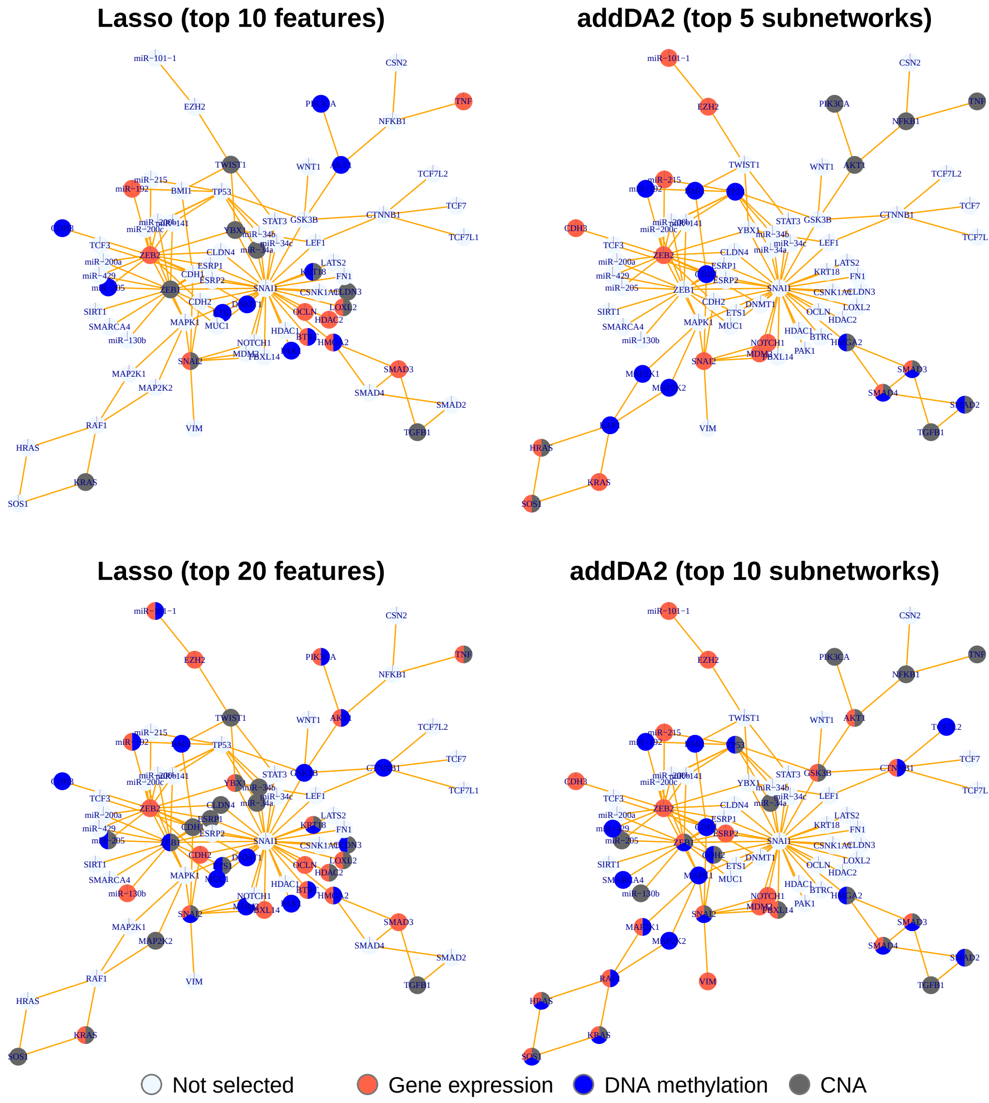
	GE74	DM74	CA70	GE123	DM123	CA117	GE455	DM455	CA445
t-test	2.10	2.15	2.35	3.30	6.15	4.90	9.35	15.15	9.95
Lasso	2.80	3.00	2.55	2.80	5.45	4.40	5.25	21.85	10.95
NetLasso	3.05	2.55	2.30	4.60	3.80	4.50	7.40	12.65	10.05
addDA2	3.91	3.91	2.38	5.52	7.00	5.46	24.50	18.26	14.79
Netrank	5.15	4.50	2.45	6.10	9.60	8.30	15.10	67.50	20.70
stSVM	2.40	4.35	2.80	6.30	9.55	7.70	45.10	52.90	15.95
Cox	3.50	2.00	3.65	2.55	5.75	3.45	6.70	19.00	13.75
RegCox	4.45	2.85	4.20	2.70	6.10	4.30	11.00	11.35	17.10
MSS	2.20	2.20	2.00	2.05	2.65	3.20	4.50	11.45	12.85
Survnet	3.42	3.38	2.29	4.53	7.71	5.00	17.58	23.79	11.72
ensemble	3.35	3.35	2.45	3.20	6.55	6.10	12.35	26.50	19.55

of the selected features, we visualized the FSFs of two algorithms - Lasso and addDA2 on the EMT network, as shown in Figure 3.19. We observe that some nodes are selected for more than one data levels and some are not. Important genes in one data level may not be the important genes in other data levels. These two algorithms show clearly different feature preference. addDA2 algorithm identified some important modules in which features are selected on multiple data levels while features selected by Lasso are more separately distributed. One can also observe that addDA2 algorithm selected less features that are in the center of the network. We have performed feature visualization for all algorithms on all three networks. However, when the network size grows, one can hardly identify the patterns by eyes. It is hard even for small networks, to identify how these features interplay in the context of cancer prognosis.

Although it is well-known that gene regulations take place in different levels and work closely with each other in molecular biology, this interplay has not been revealed in cancer prognosis prediction. Therefore, we propose the hypothesis that features from different data levels play different and supplementary roles in cancer prognosis prediction. To test this hypothesis, we need an approach that can make predictions while offering a model with clear biological interpretations. We think association rule mining algorithm is suitable for this purpose. It can give association rules for prognosis prediction and the rules can be derived simultaneously from different data levels. In this way, one can check the biological interpretation of the rules, e.g., how features from different data levels interact and how this is associated with cancer prognosis. Additionally, one can use the derived rules to make predictions to see how well these rules can stratify patients of different prognostic outcome.

Association Rule Mining

Association rule mining is a machine learning method for discovering interesting relations between features in a dataset. It consists of two steps: finding frequent itemsets and representing them in the form of rules. The problem of association rule mining is originally defined as the following [4]: Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n binary features called items. Let $D = \{t_1, t_2, ..., t_m\}$ be a set of transactions called the database. A rule is defined in


Lasso (top 10 features)

addDA2 (top 5 subnetworks)

Lasso (top 20 features)

addDA2 (top 10 subnetworks)

Figure 3.19: Visualization of the FSFs by Lasso and addDA2. Features selected from the three data levels are visualized with different colors.

Table 3.7: Average betweenness centrality of the FSFs of 10 feature selection algorithms.

	GE74	DM74	CA70	GE123	DM123	CA117	GE455	DM455	CA445
t-test	76.94	41.09	48.31	70.26	311.22	200.32	228.04	504.66	334.72
Lasso	67.37	96.56	40.67	101.29	219.24	138.85	38.34	1180.07	243.55
NetLasso	59.53	87.78	20.78	202.22	160.59	142.33	140.74	424.31	504.14
addDA2	118.93	117.42	89.39	241.17	291.52	198.98	1152.11	783.38	743.91
Netrank	197.06	188.94	55.75	241.97	476.54	371.82	391.56	4816.10	1398.33
stSVM	78.33	170.34	82.97	231.83	448.63	282.70	2737.36	3273.28	960.79
Cox	125.55	52.20	140.78	61.37	219.61	108.57	103.73	930.45	393.60
RegCox	150.91	82.61	153.80	53.57	238.37	150.54	287.48	177.61	618.29
MSS	54.66	71.42	70.44	25.09	99.06	88.50	24.65	629.72	690.07
Survnet	153.94	113.28	95.79	195.83	321.77	186.74	782.45	1106.00	353.09
ensemble	107.74	116.46	60.90	92.99	249.79	245.22	305.96	1678.27	1119.17

Table 3.8: Results of Wilcoxon signed-rank test on the centrality measures of the FSFs from different data levels. In the parenthesis after each p-value we give the comparative relationship of the two groups.

Network size	Data levels	Degree Centrality	Betweenness Centrality
74	GE vs. DM	p=0.5286	p=0.7646
74	DM vs. CNA	p=0.213	p=0.2783
70	GE vs. CNA	p=0.05545	p=0.04199 (>)
123	GE vs. DM	p=0.00293 (<)	p=0.001953 (<)
123	DM vs. CNA	p=0.009766 (>)	p=0.0009766 (>)
117	GE vs. CNA	p=0.008686 (<)	p=0.04199 (<)
455	GE vs. DM	p=0.009766 (<)	p=0.01367 (<)
455	DM vs. CNA	p=0.01855 (>)	p=0.04199 (>)
445	GE vs. CNA	p=0.5771	p=0.4648

the form: $X \Rightarrow Y$, where $X, Y \subseteq I$. The itemsets X and Y are called left-hand-side (LHS) and right-hand-side (RHS). In order to select interesting rules from the set of all possible rules, constraints on various measures of significance and interest are applied. Let a rule $X \Rightarrow Y$ be identified on a set of transactions T. Commonly used constraints are given below:

- Support. It indicates how frequently the itemset appears in T.

$$\mathit{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

- Confidence. It indicates how often a rule has been found to be true.

$$conf(X \Rightarrow Y) = \frac{\mathit{supp}(X \cup Y)}{\mathit{supp}(X)}$$

- Lift. It indicates the degree to which X and Y depend on each other.

$$lift(X \Rightarrow Y) = \frac{\mathit{supp}(X \cup Y)}{\mathit{supp}(X) \times \mathit{supp}(Y)}$$

To apply association rule learning, we first discretized the features using the mean. Since we are trying to find molecular patterns for predicting prognosis, we set the RHS of the rules to be the class labels of prognosis. Then we applied Apriori algorithm [5] implemented in the *arules* R package [92] to discover rules, with the constraints of $confidence \geq 0.8$ and $\mathit{support} \geq 0.1$. In principle, we can find association rules for each combination of data levels, algorithms, and the size of networks. However, based on our motivation, which is to investigate the association patterns among different data levels in prognosis prediction, we will focus on the following questions:

1. Whether association rules agree with the underlying EMT gene regulations. For example, whether the LHS of a rule gives a reasonable explanation of the prognosis outcome.

2. Whether combining different data levels gives significantly more rules than using the same number of features within a single data level. If this is true, it indicates that combining data levels can potentially reveal more prognosis relevant molecular interactions.

3. Whether the quality of rules, as measured by the metrics given above, is significantly different among different combinations of data levels and feature selection algorithms.

We designed experiments fur the purpose of answering these questions. We identified rules from the following combinations of data levels using the FSFs selected by each algorithm. Table 3.9 shows the data combinations and the corresponding sample sizes. We used the top 10 FSFs from each feature selection algorithm to identify rules. For a fair comparison, we used in total 20 FSFs for each data combination. For single data levels, we used top 20 features. For a combination of two data levels, we used 10 features from

Table 3.9: Sample distribution of datasets containing combinations of data levels.

Data combinations	Sample size	Good prognosis	Poor prognosis
Gene expression (GE)	183	84	99
DNA methylation (DM)	167	74	93
Copy number alteration (CNA)	149	73	76
GE + DM	165	73	92
GE + CNA	146	72	74
DM + CNA	135	66	69
GE + DM + CNA	133	65	68

each data level. For the combination of 3 data levels, we used 7 features from each data level. Below we provide results attempting to answer the above questions.

Association rules provide interesting biological insights.

We have identified many rules for good and poor prognosis groups. Many of them agree with the established findings in EMT literature, but some do not. Below we would like to illustrate with three sets of examples. In the first two sets of examples the rules have good biological interpretation and in the third set of examples the rules seem to be counter-intuitive.

1. Rules from the core EMT network agree with the underlying gene regulations. $\{LOXL2_{GE} = high, TGFB1_{GE} = high, miR.34a_{GE} = low\} \Rightarrow \{prognosis = poor\}$, with $support = 0.135, confidence = 1, lift = 2.046$. This rule has a perfect confidence for all samples that have these 3 items (features of gene expression). It has been shown that LOXL2 can stabilize SNAI1. TGFB1 can phosphorylate SMAD2 and SMAD3, which interact with SMAD4 and activates HMGA2, which then activates SNAI1. When LOXL2 and TGFB1 gene expression are high, it not only induces SNAI1 gene expression but also stabilizes SNAI1 protein. miR.34a has the role of repressing SNAI1. Thus, when miR.34a has low gene expression, as indicated by the rule, SNAI1 is less repressed. Therefore, these three conditions all lead to the direction of the high expression of SNAI1, which is a key transcription factor to induce EMT, thus leading to poor prognosis. In contrast, another rule which has the opposite state of LOXL2 indicates good prognosis: $\{LOXL2_{GE} = low, ETS1_{GE} = low, LOXL2_{DM} = high\} \Rightarrow prognosis = good$, with $support = 0.105, confidence = 1, lift = 1.956$. In this condition, LOXL2 has high methylation status and low expression. It is highly liked to remain low gene expression and not able to stabilize SNAI1. ETS1 gene can increase the expression of ZEB1 which induces EMT. In this rule ETS1 has low expression so it does not contribute to inducing EMT. These two opposite rules have both perfect confidence for a small group of samples and sound biological interpretations.

2. Rules from the filtered EMT network give good biological interpretations across multiple data levels. Here we pick the top 20 rules, which all have confidence score of 1, as shown in Table 3.10. We observed that these rules consist of only 16 genes and the associations of gene status with the prognosis outcome are consistent. For example, the high expression of GATA6 gene always associates with poor prognosis

and its low expression always associates with good prognosis. Another interesting example is miR-34a gene. Its high methylation or low gene expression corresponds to good prognosis, while its low methylation or high expression corresponds to poor prognosis, as shown in rule 1, 7, 8, and 18. These consistent patterns suggest that the rules are not random but are related to the underlying gene regulations in the context of cancer prognosis. Although many molecules in the rules are not in the core EMT network, upon literature review, we found out evidence for most of these molecules. We list three examples below:

- High GATA6 gene expression is always associated with poor prognosis. GATA6 transcription factor can activate SNAI2 and down-regulates E-cadherin to induce EMT. [28,139,225]. Thus, it is reasonable that a high expression of GATA6 contributes to poor prognosis.
- High expression of BIRC3 gene is always associated with poor prognosis. BIRC3 is shown to be a biomarker of mesenchymal phenotype in glioblastoma [265]. It is also related to cell motility and invasion in breast cancer [166].
- High expression of BIRC5 gene is always associated with poor prognosis. BIRC5 gene encodes Survivin protein, which is expressed highly in most human tumors [210] and it is required for tumor maintenance [9].

3. There are some items or rules that may be less apparent to explain or counter-intuitive. We give three examples below.

   - The low gene expression of CDH3 is always associated with good prognosis while its high expression is associated with poor prognosis. CDH3 encodes P-cadherin. It is a cell-cell adhesion molecule and plays an important role in conserving the structural integrity of epithelial tissues. Meanwhile, it is known as both a tumor suppressor and a tumor promoting molecule, depending on the molecular context [260]. Therefore, the high expression of CDH3 is supposed to be associated with good prognosis.
   - The low DNA methylation of HMGA2 is always associated with good prognosis. This is counter-intuitive as HMGA2 gene is known to promote cancer metastasis by activating SNAI1 gene to repress E-cadherin [270, 304]. Thus, a low DNA methylation level of HMGA2 is supposed to lead to poor prognosis.
   - There are items in high confidence rules but their roles in cancer are not yet well studied. For example, Table 3.10 shows that the high expression of FOXA3 is always associated with poor prognosis and its low expression is always associated with good prognosis. However, we did not find information about its role in cancer.

One can see from the rules that there are multiple alternative molecular mechanisms that are associated with prognosis. Since there are rules with high confidence and good biological interpretations, one can suggest that instead of finding a model for all samples, it could be more appropriate to use multiple rules for prognosis prediction. On the other hand, there are potentially ”false positive” rules which also have a high confidence but do not match the state-of-the-art biological insight. It would be interesting to differentiate the

Table 3.10: Top 20 prognostic association rules extracted from the FSFs using filtered EMT network. All the following rules have confidence scores of 1.

	LHS	prognosis	supp	lift
1	$CDH3_{GE} = low, miR.34a_{DM} = high, HMGA2_{DM} = low$	good	0.113	1.956
2	$EGLN3_{GE} = low, CDH3_{GE} = low, HMGA2_{DM} = low$	good	0.113	1.956
3	$GFI1B_{GE} = high, CDH3_{GE} = low, HMGA2_{DM} = low$	good	0.135	1.956
4	$GATA6_{GE} = low, E2F1_{GE} = low, HMGA2_{DM} = low$	good	0.120	1.956
5	$GATA6_{GE} = high, CDC20_{GE} = low, FOXA3_{GE} = high$	poor	0.150	2.046
6	$GATA6_{GE} = high, CDH3_{GE} = high, FOXA3_{GE} = high$	poor	0.135	2.046
7	$GATA6_{GE} = high, miR.34a_{DM} = low, FOXA3_{GE} = high$	poor	0.143	2.046
8	$miR.34a_{GE} = low, CCND1_{GE} = high, FOXA3_{GE}3 = low$	good	0.113	1.956
9	$GATA6_{GE} = low, miR.34a_{GE} = low, CCND1_{GE} = high$	good	0.105	1.956
10	$BIRC3_{GE} = low, CCND1_{GE} = high, FOXA3_{GE} = low$	good	0.113	1.956
11	$LOXL2_{GE} = high, miR.34a_{DM} = low, FOXA3_{GE} = high$	poor	0.158	2.046
12	$BIRC3_{GE} = low, miR.34a_{GE} = low, CDH3_{GE} = low$ $HMGA2_{DM} = low$	good	0.105	1.956
13	$GFI1B_{GE} = high, miR.34a_{GE} = low, HMGA2_{DM} = low$ $FOXA3_{GE} = low$	good	0.105	1.956
14	$GFI1B_{GE} = high, BIRC3_{GE} = low, miR.34a_{GE} = low$ $HMGA2_{DM} = low$	good	0.105	1.956
15	$ITGA6_{GE} = high, BIRC3_{GE} = high, BIRC5_{GE} = high$ $GATA6_{GE} = high$	poor	0.113	2.046
16	$BIRC3_{GE} = high, GATA6_{GE} = high, E2F1_{GE} = low$ $miR.34a_{DM} = low$	poor	0.105	2.046
17	$BIRC3_{GE} = high, GATA6_{GE} = high, E2F1_{GE} = low$ $GATA4_{GE} = low$	poor	0.135	2.046
18	$BIRC5_{GE} = high, GATA6_{GE} = high, miR.34a_{GE} = high$ $FOXA3_{GE} = high$	poor	0.128	2.046
19	$EGLN3_{GE} = high, BIRC5_{GE} = high, GATA6_{GE} = high$ $FOXA3_{GE} = high$	poor	0.113	2.046
20	$BIRC5_{GE} = high, GATA6_{GE} = high, miR.192_{GE} = low$ $FOXA3_{GE} = high$	poor	0.120	2.046

true rules and false rules by the properties of items, either in the context of gene regulations or network structures.

Combining multiple data levels improves the qualities of rules

We firstly measured the total number of rules inferred from the FSFs of different feature selection algorithms. Especially, we have tested all different combinations of data levels. As shown in Table 3.9, combining the data levels results in a smaller sample size, which tends to have more rules. Therefore, we used only the samples that have all three data levels for comparison. Here we have included the results of 8 algorithms, excluding algorithms addDA2, stSVM, and Survnet. The reason is that these algorithms have an aberrant lower number of rules compared with other algorithms, as shown in Table 7.4. In our opinion, this is due to the network constraints of these algorithms. In addition, the quality of the resulting rules is also lower and features from different data levels do not correlate well with each other. Thus, to investigate the effect of combining different data levels, we did not include these three algorithms.

We calculated the average length of rules, the average support, confidence, and lift of rules using different combinations of data levels. The results are shown in Table 7.4. The summary of rules based on 8 feature selection algorithms is shown in Figure 3.20. We observe that combining different data levels gives significantly more rules. The rules have significantly higher confidence and lift, while being shorter in length.

3.4.3 Survival Analysis Using EMT Signatures

In many state-of-the-art studies, selected signatures are evaluated using Kaplan-Meier survival curves and log-rank test to see whether these features can stratify patients into significantly different survival groups. Having selected features using different EMT networks and data levels, we would like to investigate whether these features can cluster patients into groups with significantly different survival distributions. Therefore, we clustered samples into 3 clusters using k-means algorithm with the FSFs of each feature selection algorithm. We used top 20 FSFs for single features and top 10 for subnetwork features. We plotted Kaplan-Meier survival curves for the 3 clusters and performed log-rank tests. To test whether feature selection can improve sample stratification, we also performed clustering using all features in the EMT networks and using all features in the data levels. To test whether an ensemble feature selection can improve the clustering, we also used top 20 most FSFs across all 10 feature selection algorithms for comparison. Table 3.11 shows the p-values of log-rank tests for each algorithm and comparative group, including the combinations of data levels and networks.

We observe that compared with using all EMT features, feature selection in most of the cases increased the quality of clustering for GE and DM data. EMT features can significantly stratify the patients into different prognostic groups ($p < 0.05$). Figure 3.21 shows two examples where good patient stratifications are obtained. The first (Figure 3.21a) is to use the FSFs (subnetworks) of addDA2 algorithm with extended EMT network and DM data. Log-rank test gives a p-value of 4.06e-9. The 10 subnetworks are [1] HSP90AA1, JAK1, MAP2K2, PIAS4, PIK3CA, STAT3; [2] AKT1, BCL2L2, CCNB1, PRKDC, TRAF6; [3] GATA6, GRB2, LCK, miR-200a, MTOR, PIK3CA; [4] FLT4, JAK1, LCK, PAK1, PIK3CA, SRC; [5] HSP90AA1, PAK1, PDGFRB, PIK3CA, SP1, STAT3; [6] AKT1, DNMT1, GLI2, KLF4, RAC3; [7] ESR1, JAK1, KRT18, PIK3CA, SP1; [8] CDC42,

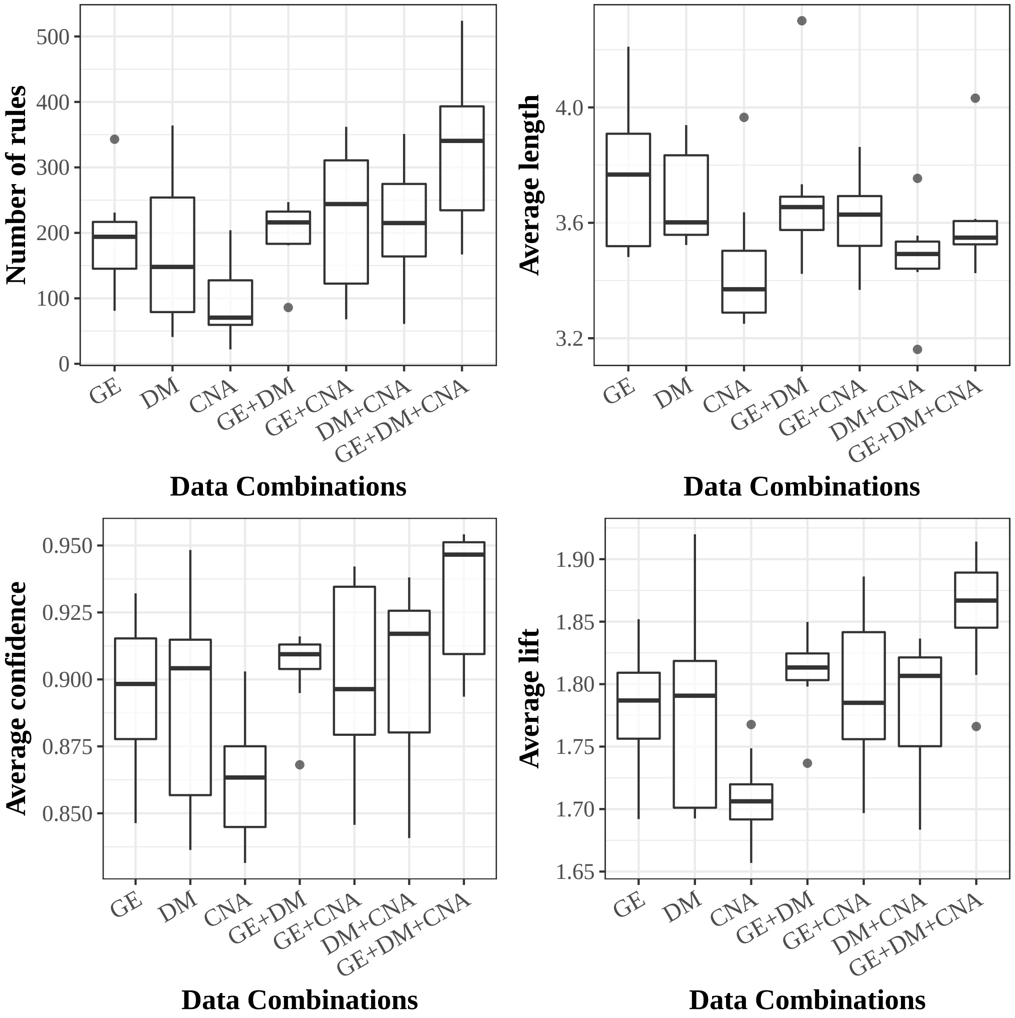


Figure 3.20: The statistics of prognostic association rules using different combinations of data levels. The top-left figure shows the number of rules given by each data combination. The top right figure shows the average rule length. The bottom two figures show the average confidence and lift of rules.

Table 3.11: The p-values of log-rank tests based on k-means clustering for individual data levels.

Data	Gene Expression			DNA Methylation			Copy Number Alteration		
$\|V(G)\|$	74	123	455	74	123	455	70	117	445
EMT	1.59e-1	3.88e-3	3.50e-2	9.76e-1	9.60e-1	3.45e-1	5.13e-1	9.75e-1	2.24e-1
t-test	2.97e-3	2.39e-5	5.56e-6	9.00e-2	4.19e-1	3.53e-2	4.80e-3	6.01e-4	1.43e-2
Lasso	1.60e-3	6.27e-4	4.88e-10	2.74e-1	8.40e-1	3.73e-2	4.46e-1	4.29e-1	8.06e-1
NetLasso	6.04e-2	2.41e-1	3.81e-1	3.58e-1	6.27e-1	4.24e-1	7.81e-1	9.27e-1	9.65e-1
addDA2	6.25e-5	5.63e-2	1.56e-6	9.11e-1	1.08e-3	4.06e-9	8.48e-3	8.69e-4	4.46e-5
Netrank	2.79e-4	1.14e-3	5.18e-8	7.08e-1	9.11e-1	5.59e-2	6.58e-3	5.85e-3	3.64e-3
stSVM	8.84e-2	3.86e-2	3.81e-2	3.79e-1	6.79e-1	3.48e-1	6.89e-2	5.30e-1	9.25e-1
Cox	9.19e-7	8.65e-8	8.22e-4	7.24e-2	1.77e-1	9.85e-4	4.24e-2	9.42e-2	1.36e-5
RegCox	4.19e-4	1.29e-4	2.68e-9	4.20e-1	1.29e-1	1.86e-2	1.40e-2	8.80e-2	3.36e-3
MSS	2.45e-3	3.80e-5	3.15e-4	2.45e-2	9.66e-2	2.52e-1	2.48e-1	1.86e-1	2.10e-1
Survnet	6.93e-4	2.89e-2	9.20e-6	3.19e-1	9.17e-1	5.50e-3	2.29e-1	5.52e-3	1.52e-2
ensemble	1.65e-4	2.00e-5	1.50e-10	5.62e-1	6.11e-1	7.63e-2	1.91e-1	7.66e-2	1.77e-2
All data level features			3.46e-2	3.91e-1			3.83e-1		

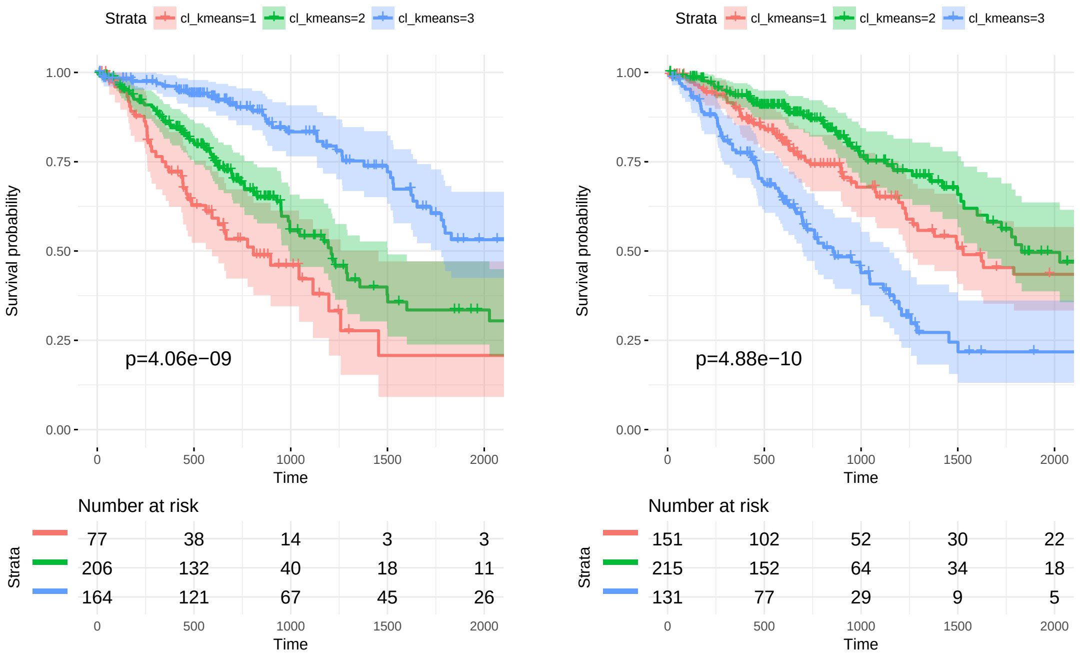
FGFR1, GRB2, PIK3CA, RAC2, SP1; [9] AKT1, BCL2L2, GLI2, RAC3; and [10] AKT1, BCL2L2, CCNB1, HIC1, RAC2, RAC3. The second example (Figure 3.21b) is to use the FSFs of Lasso algorithm with extended EMT network and GE data. Log-rank test gives a p-value of 4.88e-10. These 20 genes are ITGA6, EGLN3, ZNF2, YES1, CDC42, ABL1, ZNF146, PIK3CG, PPP1R13B, PTGS2, GLI2, WNT3A, SHC1, RELA, EPAS1, HIST1H1A, IGF1R, KAT2B, BIRC3, and LOXL2. Note that both feature selection algorithms selected features based on only labeled data, which contain 183 (GE) and 167 (DM) samples. Here we use these features to cluster all 497 (GE) and 447 (DM) samples. Although many samples were not exposed to the algorithms during feature selection, the FSFs can very well cluster these unseen samples. Last but not least, we observe that using all features from the data levels does not give satisfactory clustering results, which highlights the importance of feature selection.

3.5 Discussion

3.5.1 Advantages of EMT Features

We constructed EMT GRN from the review of EMT research articles. In this way we captured the central genes in EMT gene regulations. It is shown in Figure 3.4 and 3.5 that the features selected from EMT networks always obtained significantly better prediction performance than the features selected from random networks, no matter which feature selection algorithm was employed. This shows that feature selection based on biological knowledge makes a real difference. Given that the major challenge in biomarker discovery is to differentiate between the true signatures and statistically equivalent random signatures, our proposed approach of using phenotype relevant network-based feature selection can increase the robustness of molecular signatures.

One would ask that lots of information are lost when we select features only from EMT



(a) DM data using FSFs of addDA2 algorithm (b) GE data using FSFs of Lasso algorithm

Figure 3.21: Kaplan-Meier survival curves of 3 patient clusters and log-rank tests using individual data levels.

GRN, which has at most 2.5 % of the original dimensionality. However, our experiments show that selecting features based on EMT networks can significantly outperform the features selected from the original data (Figure 3.6, 3.7, 3.8, and 3.9), which corresponds to 19,290 GE features, 20,074 DM features, and 21,456 CNA features. Consistently, EMT-based features can stratify all stage samples and early stage samples into significantly different prognostic groups while using all data level features cannot (Table 3.11). Additionally, EMT-based features have much better biological interpretations, as shown from the prognostic association rules derived from the FSFs. These performance gains suggest that using biological knowledge to shrink the feature space to phenotype relevant features is an effective method to mitigate the curse of dimensionality. The Wolpert's famous ”no free lunch” machine learning theorems [275], which state that no learner can beat random guessing over all possible functions to be learned, also support our conclusions. Without certain level of domain knowledge, data alone are not enough for a machine learning model to have good generalization [61].

We meanwhile compared the performance of three EMT networks: core network, filtered network, and extended network. In most of the cases, the extended EMT network gives the best performance. This observation is consistent with state-of-the-art studies, where it is shown that the central transcription factors may not change their expressions as much as the genes they regulate. Thus, the central network may not show obvious difference among patients that belong to different prognostic groups. Attention needs to be paid on this point because it can let us utilize biological knowledge to a better extent. For example, one can first identify the central GRN and extend it to involve down-stream genes using biological databases that describe PPIs, TF-gene interactions, etc.

3.5.2 Considerations among Multiple Evaluation Metrics

Having evaluated the feature selection algorithms using different metrics, we observe that these metrics can give different conclusions if we base our conclusions only on statistical tests. For example, in Figure 3.6 one can perform statistical tests between each pair of boxes. If the same statistical tests are applied again on the results in Figure 3.9, one cannot hold the same conclusions, let alone when different classifiers are employed with different parameters. The same phenomenon is observed when we use other evaluation metrics. [34] performed prognosis prediction in breast cancer employing five predictive models. They found that different metrics of evaluating survival prediction models gave different conclusions. They also showed that the performance of binary classifiers is highly dependent on how the two risk groups are defined. A slight change of the threshold can lead to very different prediction results. This agrees with our results of experimenting with different prognostic thresholds. Taking these factors into account, we see that it is not reliable to evaluate prognostic feature selection algorithms using a single metric, as the conclusions can change upon the variations of evaluation metrics, classifiers and parameters. Most of the state-of-the-art studies use ≤ 2 evaluation metrics. As the metrics can be different among studies and these metrics do not guarantee to agree with each other, the best biomarker found in one study may not be good for the other studies. In our opinion, this contributes to the very low overlap among the biomarkers reported in different studies. Therefore, we advocate to employ a few different metrics when evaluating feature selection algorithms.

When we draw conclusions based on several evaluation metrics, addDA2 algorithm is one of the best performing algorithms. Especially, the FSFs of addDA2 perform very well in stratifying samples into distinct survival groups. Lasso algorithm has also achieved consistent good performance. t-test feature selection has a very good stability while performing better than several other algorithms. It is hard to rank the other feature selection algorithms as their performance measured by different evaluation metrics is not consistent. What we find more interesting is the comparison of different data levels. The comparisons among 3 data levels: GE, DM and CNA are consistent regardless of evaluation metrics. EMT GE and DM features are more predictive than CNA features.

3.5.3 Comparisons to State-of-the-art Studies

As shown in Chapter 1, we have reviewed studies that aim to identify molecular signatures from omics data for cancer prognosis prediction. In the following we relate our results to theirs to either support our findings or to share some new insights.

Random Features Can Significantly Predict Prognosis

[258] conducted a study to identify signatures for breast cancer outcome prediction. They found out that most random gene expression signatures are significantly associated with breast cancer outcome. Upon comparison of 47 published breast cancer outcome signatures to random gene signatures, 28 of them (60%) were not significantly better outcome predictors than random signatures of identical size and 11 (23%) were worse. They showed

that in breast cancer any set of 100 or more genes selected at random has a 90% chance to be significantly associated with clinical outcome. Our experiments show similar results. One can observe in Figure 3.4 and Figure 3.5 that at least half of the feature sets selected from random networks (consisting of random chosen features) are significantly predictive, in terms of AUC value, AUPR value, and accuracy. What's more, this holds true for both GE data and DM data, and for different network sizes. This demonstrates that statistical significance is not enough for the conclusion that a good molecular signature has been discovered. One needs to deal with the underlying irrelevant features that could be hidden in the signature.

Another study [8] also agrees with our conclusions on using random networks. They evaluated the prognosis prediction performance of features selected from 3 different networks - PPI network in [228,229], co-expression network, and a random network by shuffling the nodes in a PPI network. Their results show that features selected from the random network give similar prediction AUC values as features selected from the PPI network. Several other studies [66, 97, 229] on breast cancer prognosis prediction using gene expression data have drawn consistent conclusions. Briefly, [97] did not detect any feature selection method to be significantly better than random features using paired ANOVA test. [66] showed that random signatures can reach a baseline AUC value comparable to that of feature selection algorithms.

EMT Features Achieved Competitive Prediction Performance

We related the prediction performance of EMT features to that achieved in state-of-the-art studies of prognosis prediction. [297] carried out a comprehensive study to evaluate the features selected from different omics data types in their prognosis prediction performance. Four cancer types from TCGA are included: kidney renal clear cell carcinoma (KIRC), glioblastoma multiforme (GBM), ovarian serous cystadenocarcinoma (OV), and lung squamous cell carcinoma (LUSC). The omics data types under comparison are: (a) mRNA expression, (b) miRNA expression, (c) DNA methylation, (d) CNA, and (e) protein expression. What's more, they investigated the effect of classifiers (8 classifiers including SVM), feature selection algorithms, and the number of features on the classification performance by calculating the average AUC values of 10-fold cross-validation. They tuned the parameters of classifiers and picked the one that gave the highest AUC values. Using SVM classifier, they achieved average AUC values of 0.658, 0.531, 0.628, and 0.696 with data types a, b, d, and e for LUSC, and the values of 0.626, 0.625, 0.619, 0.615, and 0.625 with data types a, b, c, d, and e for OV. In our experiment, we used the default parameters of RBF kernel. Our results in Table 3.3 show higher average AUC values. Although the AUC values are not directly comparable because the cancer type differs, we can at least conclude that EMT-based feature selection is competitive. Since [297] investigated different factors that can potentially influence the prediction performance, they concluded that the effect of machine learning algorithms is moderate compared with the effect of data types. This is consistent with our findings, where the performance difference between SVM and random forest classifiers is much smaller than the difference among data types.

[163] selected features using four data types (mRNA, miRNA, DM, and CNA), as well as integrated data (a concatenation of individual data types with features that meet certain criterion) to predict prognosis in OV with L1 regularized Cox PH model [188] .

With the selected features from integrated data, they clustered the holdout test data into three clusters. The results of log-rank tests using either individual data types or integrated data did not show satisfactory separation of patients in terms of overall survival (p >0.05). The integrated data did not lead to a significantly better results. As the concatenated data have even higher dimensionality, the potential to find more important signatures may be out-weighted by the increased noise. The CoxReg algorithm [224] we employed has the same optimization function, except that a different method - cyclical coordinate descent is used to find the optimal model parameters. Using EMT networks, CoxReg algorithm can stratify patients into significantly different survival groups.

FSFs Boosted the Prediction Performance of EMT Features

This has been shown in Figure 3.12 and Figure 3.13. It suggests that using FSFs can better cope with the heterogeneity and noise of individual cross-validation tests and obtain more robust molecular signatures. We have observed from Figure 3.10 and Figure 3.11 that feature sets can fit the training sets very well but have very different performance on the testing set. In other words, it is hard to know whether a set of features will perform well on the testing set based on its performance on the training set. By using FSFs, we have a much better chance to select the true signals and avoid overfitting. This idea has been used in algorithm MSS [150], where the author also found out that the features selected from individual cross-validation folds are highly unstable. However, probably due to the high dimensionality, the performance gain of using FSFs in [150] was not as satisfactory as shown in our study, where only phenotype relevant features are employed for feature selection.

Clinical Features Alone Can Only Give Moderate Prediction Performance

As shown in Figure 3.14 and Figure 3.15, clinical features alone without molecular features give inferior performance than the combination of clinical features with molecular features. This agrees with several other studies [34, 218, 297, 303]. These studies, however, have not tried to understand how much clinical features and molecular features contribute to the prediction improvement. Our results in Figure 3.14 (second panel on DM data with extended EMT network) and Figure 3.13 can be compared in parallel to shed light on this question. We observe that the performance gain is mainly from the FSFs. For t-test and NetLasso algorithms adding clinical features to FSFs has even decreased the performance. It meanwhile suggests that combining clinical features directly with molecular features may not be a good implementation, as the variables come in different scales and types.

While state-of-the-art studies usually assess clinical features as a whole, we have instead analyzed the variable importance of clinical features using different methods - mean Gini decrease, Spearman correlation coefficient, AUC value, univariate Cox PH model, and survival trees, as shown in Figure 3.16. In univariate analysis, we found out that pathological stage and pathology_N are the two most important variables. What we found more useful is to build a survival tree like Figure 3.16c that can better utilize all clinical variables to assess the risks of patients and assist therapy choice. For example, on the node split of radiation therapy, it is shown that for the patients whose size and/or extension of the primary tumor $\leq$ T2, radiation therapy leads to increased death rate.

3.5.4 Single Features vs. Composite Features - Our Insights

Multiple studies have shown that composite features (e.g., subnetworks) are superior to single genes as biomarkers. However, it is still under debate whether the integration of network connectivity information can improve the prediction performance of the selected features [48, 228, 229]. We have shown in our experiments that composite features do not necessarily outperform single gene features. t-test feature selection can outperform more complicated network-based feature selection methods, as shown in Figure 3.6, Figure 3.7, Table 3.3, and Figure 3.9. Our results are consistent with the results in a few recent studies [97, 215]. [97] compared 32 feature selection methods on 4 gene expression datasets for breast cancer prognosis prediction. They found out that feature selection algorithms significantly influence the prediction accuracy. Overall, t-test feature selection gives the highest average AUC values.

A few studies including [8] conclude that network-based prognosis prediction methods mainly contribute to the robustness of features, rather than the prediction performance. [8] found out that the frequently used average operator is a poor choice to summarize genes into meta-genes. They proposed the Direction Aware Average (DA2) operator, which takes into account the directions of genes before taking the average. We have adopted this operator to improve the original algorithm [40] and yields the addDA2 algorithm. In our experiments, this algorithm often gave superior prediction performance than other feature selection algorithms. It shows that the operators used for generating meta-genes are important. [8] employed both PPI network and random network to test the performance of network-based feature selection algorithms. Their results show that the contribution of biological networks to the prediction performance is negligible. They accredit this to the existence of a large number of genes that are correlated with the target labels. As we have discussed before, it is very hard to differentiate the marker genes and irrelevant genes when they have similar statistical relevance with the target labels. We agree that using a global biological network does not help much on mitigating this issue. However, we show that using EMT networks, which cover <2.5% of the original dimensionality, addDA2 algorithm can select robust features that give superior prediction performance according to several evaluation metrics.

The potential of EMT molecules in prognosis prediction has also been studied before. [38] studied whether individual EMT molecules in primary tumor can be used as biomarkers for prognosis prediction in LUAD. They analyzed the correlation of the expression values of a few proteins, e.g., E-cadherin, vimentin, and fibronectin, with the survival data. They found some associations between these variables and pathological stages. However, survival analysis showed that none of these molecules are significantly associated with prognosis. [301] assessed whether the protein expression of EMT markers E-cadherin, Twist, and Vimentin could be predictive of patient survival in bladder cancer. Their analysis showed that none of these molecules were significantly predictive of the overall survival. We think that there could be two potential reasons why these biological meaningful markers did not do well in predictions. The first one is that these studies did not consider different variables simultaneously. The conclusions were made based on univariate analysis. Since the molecules usually regulate each other, it is reasonable to expect that a set of variables can jointly lead to a more predictive model. The second reason we think is that there could exist patient subgroups. This is supported by our observations on prognostic association rules, as shown in Table 3.10. We found many rules of high confidence with moderate support, which suggests that there may exist patient subgroups that have different disease mechanisms.

We would also like to address the wide application of Lasso algorithm. It is often employed to select features before building predictive models. What's more, the selected features were sometimes considered as the representatives of the corresponding omics data type for comparing the predictive capability among multiple data types [297, 303]. [303] applied Lasso feature selection and used the selected features to make prognosis predictions on four cancer types using four omics data types. It is shown that for all combinations of data types and cancer types, the average AUC values are between 0.5 and 0.7. 80% of the AUC values are below 0.6. This very much agrees with our results, as shown in Figure 3.6 and Figure 3.8, where the red dashed lines correspond to the average AUC values of Lasso algorithm applied on all data level features. These results illustrate that it is hard to select the true signals by applying Lasso directly on high-dimensional omics datasets. Therefore, by performing Lasso feature selection we cannot obtain a reliable estimation of the predictive potential of certain types of omics data. In contrast, applying Lasso feature selection on all data level features rarely outperforms EMT features. [8] also showed that using features selected by Lasso does not give a good estimate of the prediction performance that could be achieved if features were selected with other principles.

Last but not least, we have analyzed the patterns of the FSFs on different data levels. We found out that they have significantly different network properties. Besides, features from different data levels often appear in the same prognostic association rules and together they can give sound biological interpretations. Therefore, we want to further investigate whether features from different data levels can complement each other and further improve prognosis prediction performance. This will be introduced in Chapter 4.

Chapter 4

Multi-omics Prognostic Signatures for Lung Adenocarcinoma

The molecular portrait of a tumor manifests at multiple omic levels. We have learned from the last chapter that EMT features are very useful in prognosis prediction. We have evaluated EMT-based features on three data levels - gene expression (GE), DNA methylation (DM), and Copy Number Alteration (CNA), alternatively. In addition to observing their difference in prediction performance, we discovered that the network properties of the frequently selected features (FSFs) from these data levels are significantly different. This inspires us to explore whether the features from different data levels are complementary to each other. We therefore investigate whether more accurate predictions can be obtained if features from different data levels are used simultaneously.

In this chapter we will identify and test multi-omics EMT signatures. The most straightforward approach is to combine the FSFs from individual omics levels. Besides, we propose a multiplex-based integrative feature selection approach to directly select features spanning multiple omics data levels. We show that both approaches have achieved significantly better prognosis predictions than using single-omics features. Last but not least, we have tested the EMT signatures, both single-omics and multi-omics ones, on real-world clinical datasets. Using EMT features we are able to separate the patients into significantly different prognostic groups. Further, multi-omics signatures can often achieve superior performance than single-omics signatures.

4.1 Multi-omics Signatures from Single-omics Signatures

4.1.1 Survival Analysis on All Stage Samples

Having stratified the samples successfully using the FSFs from single data levels, we would like to know whether combining features from different data levels can better stratify the samples than using single data levels. We employed both k-means and spectral clustering algorithms to test this hypothesis. Note that for different data level combinations, we try to use the same number of features for clustering. For example, with GE data level, we used top 20 FSFs from each feature selection algorithm. When we use the combination of GE and DM data levels, we picked top 10 FSFs from each data level and combined them. When three data levels are combined, we picked top 7 FSFs from each data level. For

Table 4.1: The p-values of log-rank tests based on the clustering of k-means algorithm for different data level combinations using extended EMT network. We highlighted all p-values that are lower than 10e-5.

	GE	DM	CNA	GE+DM	GE+CNA	DM+CNA	GE+DM +CNA
t-test	**7.87e-06**	1.12e-01	3.62e-03	**7.55e-06**	8.75e-04	1.90e-03	**8.25e-06**
Lasso	**4.58e-06**	5.56e-02	5.54e-01	1.66e-04	**2.28e-07**	8.71e-01	1.18e-04
NetLasso	2.71e-01	6.45e-01	7.58e-02	2.96e-02	1.53e-02	4.83e-01	4.34e-01
addDA2	**5.20e-10**	**5.17e-09**	1.24e-04	**8.99e-18**	3.75e-05	**1.11e-09**	**1.35e-07**
Netrank	**1.13e-07**	1.79e-01	2.19e-02	**2.50e-07**	**6.97e-06**	7.31e-02	**3.28e-06**
stSVM	4.14e-02	3.39e-01	8.91e-01	8.86e-01	6.37e-02	5.85e-01	6.83e-01
Cox	**2.55e-09**	2.91e-03	**5.30e-07**	3.12e-04	**1.70e-06**	1.13e-04	**6.11e-06**
RegCox	**1.78e-07**	8.52e-03	2.67e-01	**2.36e-09**	**1.52e-10**	**1.81e-07**	**2.52e-07**
MSS	1.48e-03	5.95e-01	2.78e-01	6.29e-05	2.28e-04	2.59e-01	1.63e-03
Survnet	7.36e-05	6.25e-03	5.19e-03	**2.59e-06**	1.54e-03	3.77e-05	2.19e-05
Ensemble	**2.32e-09**	4.72e-02	6.05e-03	1.20e-04	1.62e-05	1.01e-01	1.18e-04
allemt	1.39e-02	4.30e-01	1.07e-01	7.64e-01	1.45e-02	2.07e-01	5.34e-01

subnetwork-based algorithms addDA2 and Survnet, we used top 10 subnetworks for single data levels, top 7 subnetworks when combining two data levels, and top 5 subnetworks when combining 3 data levels. We experimented with 439 samples which have all three data levels available. We applied k-means algorithm to divide the samples into 3 clusters and performed survival analysis on the clusters. We used extended EMT network because it is shown in the previous section to give better sample stratifications. The results of log-rank tests are shown in Table 4.1, where the columns show different data level combinations and the rows correspond to feature selection algorithms. We have also included the comparative group *allemt*. It stands for using all EMT features without feature selection.

We observe that the p-values of log-rank test are the same or remarkably lower after combining features from different data levels. This is also observed when we applied spectral clustering algorithm. The results are given in Table 7.5. Note that no matter whether we used features from single data level or combinations, feature selection shows its importance in clustering performance. When no feature selection is performed, as shown in the row named *allemt*, the clustering performance is very poor compared with a clustering with selected features.

addDA2 algorithm gives very good clustering performance when combining features from GE and DM data levels. Using top 10 features from GE and DM data separately, the p-values of log-rank tests are 5.20e-10 and 5.19e-7 respectively. When using 7 top subnetwork features from both data levels altogether, it achieved a much lower p-value of 8.99e-18. The survival curves of using individual data levels, compared with those of using combined data levels are shown in Figure 4.1. The top 10 subnetwork features from DM data can be found in Chapter 3. The top 10 subnetwork features from GE data are [1] DNMT1, miR-215, RAF1, RB1, TP53; [2] GFI1B, GLI2, MAP2K1, MAPK1, SMAD3; [3] BCL2L11, BIRC5, MED1, miR-101-1, PRKDC, PTGS2, TP53; [4] FOS, miR-192, RXRB, SMAD3, TGFB2; [5] CEBPB, KLF4, LCK, NFKB1, PRKCA, RXRB; [6] CDKN1B, IGF1R, KAT2B, LOXL2, miR-192, STK3; [7] KLF4, KRT18, TGIF2; [8]

APPL1, CSNK2A1, HDAC1, MAP3K7, SMAD3; [9] BTRC, GFI1B, GLI2, MAP2K1, ZNF2; and [10] IGF1R, LOXL2, miR-192, NFYA, ZEB2.

We are also interested to know whether the patient clusters in the three sub-figures of Figure 4.1 are consistent. Thus, we counted the number of common samples in these three clusters. This is done in two steps. First, we found the matching patient clusters in Figure 4.1a, 4.1b, and 4.1c by picking the matchings that have the highest sample overlap. Since there are 3 clusters in each sub-figure, we found 3 groups, each containing the matched 3 clusters. Then we plotted a Venn diagram for each group. R package *VennDiagram* [33] was used for the visualization. In Figure 4.2 we show the cluster matchings across these three sub-figures and the Venn diagrams. We observe that within each group, the three clusters have large fraction of common samples, especially in Group 2 and Group 3.

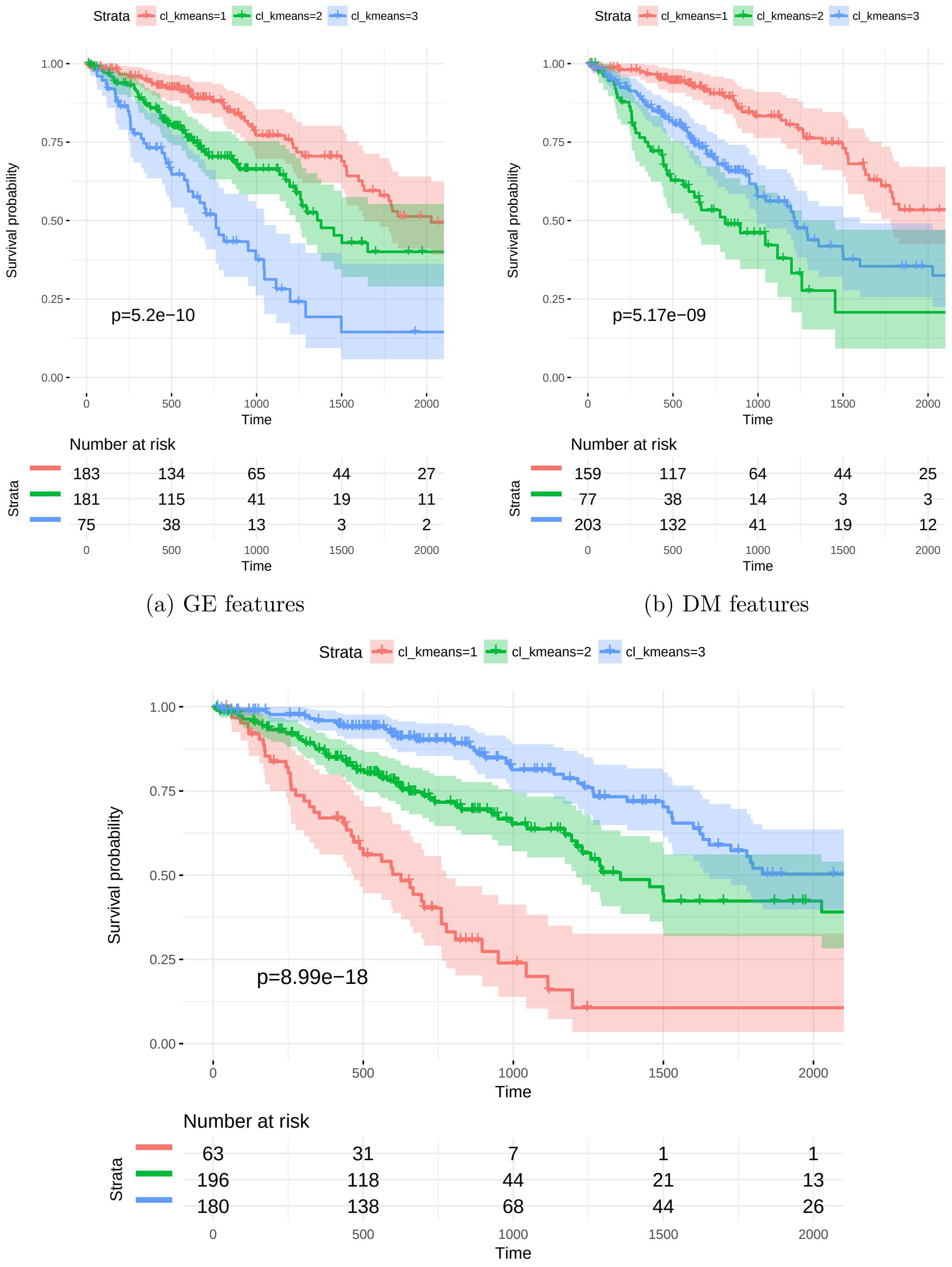
4.1.2 Survival Analysis on Early Stage Samples

We also tested the EMT signatures on early stage patients by selecting the samples with pathologic stages either I or II. We obtained 125 patients in early stage which have all three data levels. 80 patients belong to stage I and 45 patients belong to stage II. We performed the same clustering approach as above using different combinations of FSFs from individual data levels. The results of log-rank tests are given in Table 4.2. We noticed that the p-values are less significant than the values in Table 4.1. This is as expected because it is much harder to differentiate the prognosis of early stage patients. Even though it is a hard task, EMT-based features can separate the patients into significantly different prognosis groups. Consistent with our previous results, Lasso and addDA2 algorithms obtained lower p-values than the other feature selection algorithms. Combining features from multiple data levels achieved equivalent or significantly better sample stratifications than using features from single data levels. In Figure 4.3 we visualized the Kaplan-Meier survival curves of the patient clusters obtained using the FSFs of addDA2 algorithm. Figure 4.3a and Figure 4.3b show the results of using GE features and DM features separately. Figure 4.3c shows the results of using both GE and DM features.

Next, we would like to visualize how these features are related to each other. We extracted these 14 subnetworks from the extended EMT network. The top 7 subnetworks from GE data have 32 unique genes. The top 7 subnetworks from DM data have 27 unique genes. In total, there are 53 unique genes. They are connected within the extended EMT network. The visualization is given in Figure 4.4. The nodes in gray correspond to GE data and nodes in tomato correspond to DM data. Figure 4.4a highlights all nodes and edges. Figure 4.4b , 4.4c , and 4.4d highlight the edges and the associated nodes that belong to different gene regulatory networks. We observe that with PPIs, features from GE data level were more frequently selected, while with TF-gene and gene-miRNA regulations features from DM data level were more frequently selected.

4.1.3 The Application of Integrative Clustering Algorithms

As introduced in Chapter 1, clustering algorithms that can integrate multiple omics data have been proposed. This is achieved by, for example, integrating multiple similarity networks or using joint latent variables. Representative algorithms are SNF [264] and iCluster [219]. Here we would like to employ both algorithms on EMT FSFs from multiple



(a) GE features

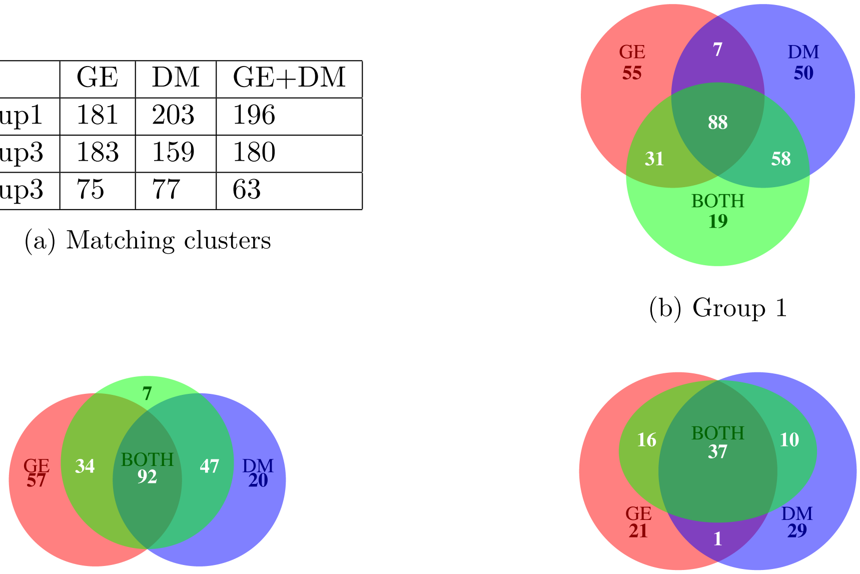
(b) DM features

(c) GE and DM features

Figure 4.1: Kaplan-Meier survival curves of 3 patient clusters using k-means algorithm based on the subnetwork features selected by addDA2 algorithm. Clustering using (a) top 10 subnetwork features selected from GE data, (b) top 10 subnetwork features selected from DM data, and (c) top 7 subnetwork features from GE data + top 7 subnetwork features from DM data.

	GE	DM	GE+DM
Group1	181	203	196
Group3	183	159	180
Group3	75	77	63

(a) Matching clusters



(b) Group 1

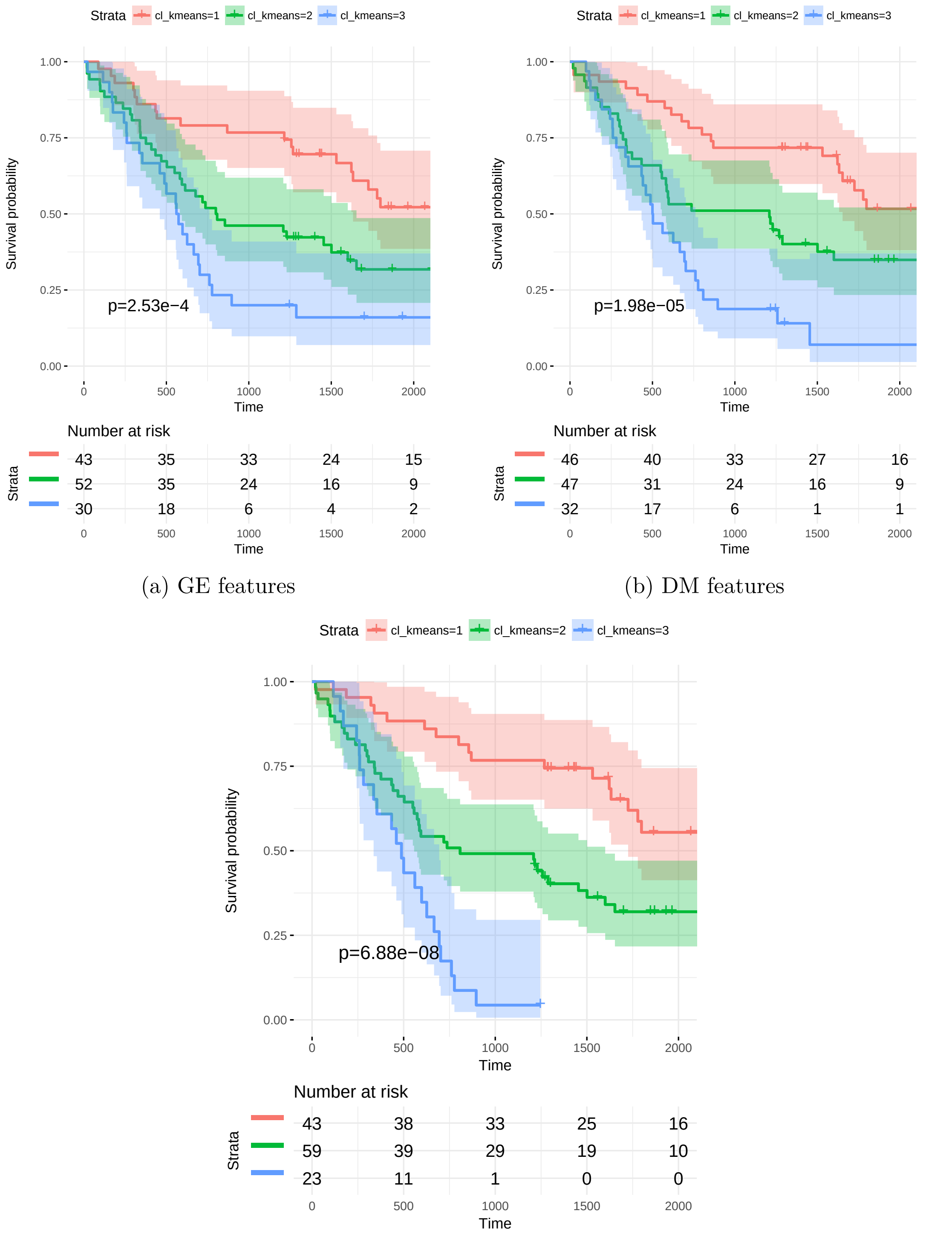
(c) Group 2

(d) Group 3

Figure 4.2: The Venn diagrams of the matching patient clusters in Figure 4.1a, 4.1b, and 4.1c. (a) The matching clusters in three patient prognostic groups. (b) Venn diagram of prognostic Group 1. (c) Venn diagram of prognostic Group 2. (d) Venn diagram of prognostic Group 3.

Table 4.2: The p-values of log-rank tests on early stage patients for different data level combinations using extended EMT network. We highlighted all p-values that are lower than 10e-3.

	GE	DM	CNA	GE+DM	GE+CNA	DM+CNA	GE+DM +CNA
t-test	6.28e-3	9.38e-2	1.93e-1	**4.78e-4**	8.40e-3	1.55e-1	2.47e-1
Lasso	**1.82e-04**	1.20e-03	1.01e-01	2.35e-01	**1.67e-06**	2.51e-03	4.94e-03
NetLasso	7.95e-3	8.56e-1	2.46e-1	2.29e-1	1.01e-1	9.69e-1	9.31e-1
addDA2	**2.53e-04**	**1.98e-05**	1.63e-03	**6.88e-08**	3.52e-02	**1.03e-05**	**8.51e-04**
Netrank	**9.31e-06**	5.39e-01	4.08e-03	3.52e-03	**5.35e-04**	7.54e-03	**8.57e-04**
stSVM	3.17e-2	2.99e-1	2.86e-1	4.00e-1	2.16e-2	1.32e-1	8.57e-1
Cox	**4.40e-4**	2.15e-1	2.42e-2	1.85e-2	3.30e-2	1.10e-2	**6.43e-4**
RegCox	**8.52e-04**	3.36e-01	2.33e-02	8.58e-03	**2.18e-05**	2.03e-03	3.90e-02
MSS	6.51e-2	6.51e-1	2.43e-1	2.34e-2	3.10e-2	9.91e-1	5.91e-2
Survnet	4.16e-3	3.05e-1	6.24e-2	6.78e-2	8.66e-2	2.27e-2	8.08e-3
Ensemble	**4.03e-4**	1.54e-1	4.54e-2	5.06e-3	**4.01e-4**	**5.79e-4**	**7.95e-4**
allemt	2.59e-1	9.45e-1	7.04e-3	6.31e-1	1.38e-2	4.16e-1	9.73e-1

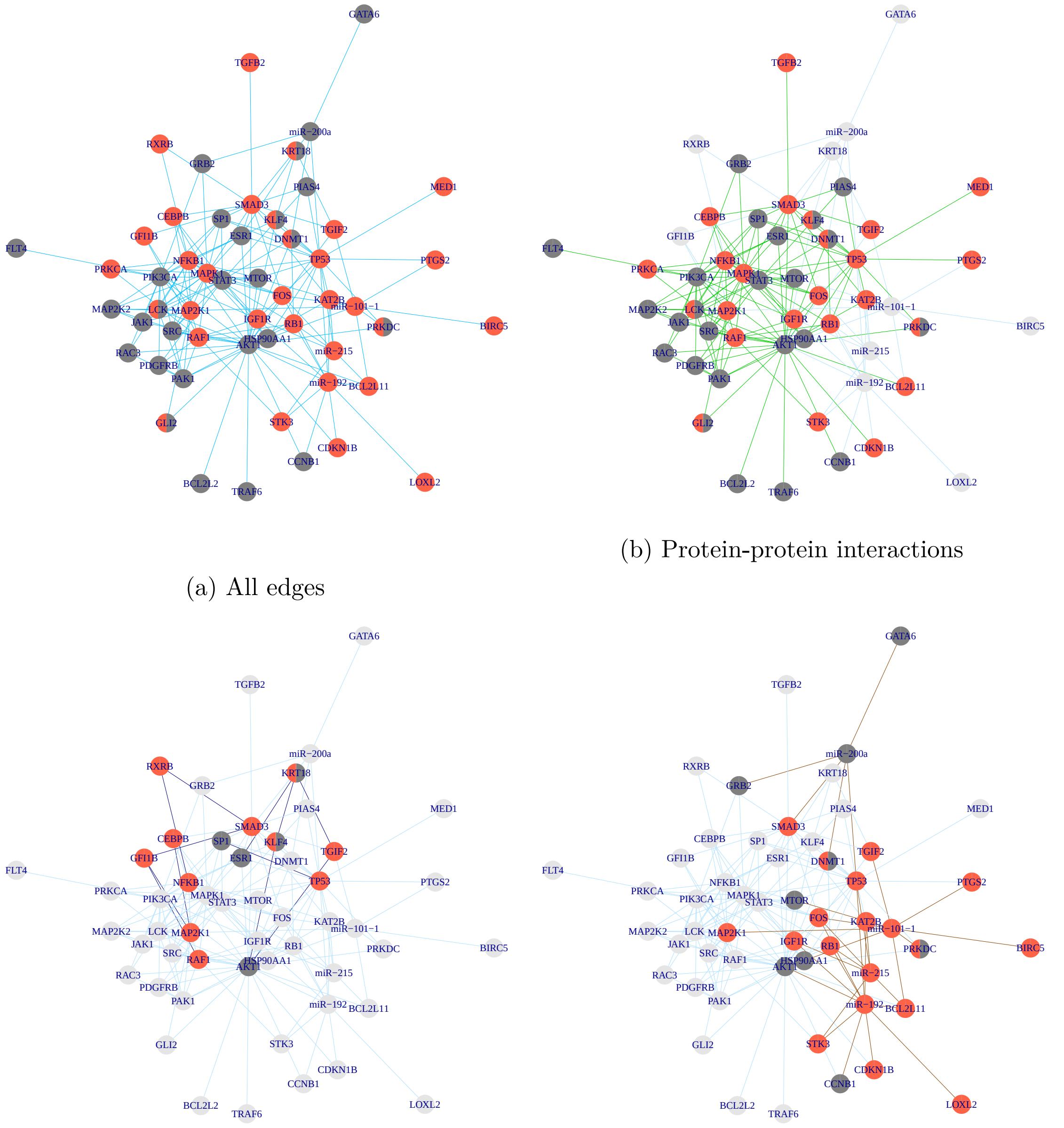


(a) GE features

(b) DM features

(c) GE and DM features

Figure 4.3: Kaplan-Meier survival curves of 3 early stage patient clusters using k-means algorithm based on the subnetwork features selected by addDA2 algorithm. Clustering using (a) top 10 subnetwork features selected from GE data, (b) top 10 subnetwork features selected from DM data, and (c) top 7 subnetwork features from GE data + top 7 subnetwork features from DM data.



(a) All edges

(b) Protein-protein interactions

(c) TF-gene regulations

(d) gene-miRNA regulations

Figure 4.4: Visualization of the top 14 selected subnetworks using addDA2 algorithm. These 14 subnetworks consist of 7 subnetworks from GE data (nodes in gray) and 7 subnetworks from DM data (nodes in tomato). If a node is involved in both data levels, we show both colors on the node. In total there are 53 nodes. A subgraph was extracted from the extended EMT network that included these nodes and the edges among them. The subgraph is connected. Subgraph (a) shows all nodes and highlighted all edges. Subgraph (b), (c), and (d) highlighted the edges that belong to protein-protein interactions, TF-gene regulations, and gene-miRNA regulations.

data levels to cluster samples. The aim is to investigate whether these algorithms can further improve patient stratification into different prognostic groups. In the following we briefly introduce the two algorithms.

Introduction of Algorithms

The main idea of [264] is to construct a sample similarity network for each omic data type (mRNA expression, DNA methylation, and miRNA expression) and then fuse these networks into a single similarity network. Suppose we have n samples and m features. Let $G = (V, E)$ represent a sample similarity network. The vertices V correspond to the samples. $\boldsymbol{W}$ is an $n \times n$ matrix of edge weights. $\boldsymbol{W}(i, j)$ indicates the similarity between sample x_i and x_j. The weight of an edge is calculated as:

$$\boldsymbol{W}(i, j) = exp\Big(- \frac{\rho^2(x_i, x_j)}{\mu \varepsilon_{i,j}} \Big)$$

where μ is a hyperparameter and $\varepsilon_{i,j}$ is used for scaling. A full and sparse kernel $\boldsymbol{P}$ is defined on the vertex set V. It is a normalized weight matrix as follows:

$$P(i, j) = \begin{cases} \dfrac{\boldsymbol{W}(i, j)}{2 \sum_{k \neq i} \boldsymbol{W}(i, k)}, j \neq i \\ \\ 1/2, j = i \end{cases}$$

Let $\boldsymbol{N}_i$ represent a set of x_i's neighbors including x_i in G. The local similarity matrix $\boldsymbol{S}$ is defined as:

$$\boldsymbol{S}(i, j) = \begin{cases} \dfrac{\boldsymbol{W}(i, j)}{\sum_{k \in \boldsymbol{N}_i} \boldsymbol{W}(i, k)}, j \in \boldsymbol{N}_i \\ \\ 0, otherwise \end{cases}$$

It is assumed that local similarities are more reliable than remote ones. This approach assigns similarities to non-neighbors through graph diffusion - start from $\boldsymbol{P}$ as the initial state and use $\boldsymbol{S}$ as kernel matrix. Suppose there are two omics data types and corresponding kernel matrices and similarity matrices $\boldsymbol{P}^{(1)}$, $\boldsymbol{P}^{(2)}$, $\boldsymbol{S}^{(1)}$, $\boldsymbol{S}^{(1)}$. $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$ are updated in each iteration t:

$$\begin{cases} \boldsymbol{P}^{(1)}_{t+1} = \boldsymbol{S}^{(1)} \times \boldsymbol{P}^{(2)}_t \times (\boldsymbol{S}^{(1)})^T \\ \\ \boldsymbol{P}^{(2)}_{t+1} = \boldsymbol{S}^{(2)} \times \boldsymbol{P}^{(1)}_t \times (\boldsymbol{S}^{(2)})^T \\ \end{cases}$$

After t steps, the final kernel matrix $\boldsymbol{P}^{(c)} = \dfrac{\boldsymbol{P}^{(1)}_t + \boldsymbol{P}^{(2)}_t}{2}$. When there are more than two data types ($m > 2$), the kernel matrices are updated as the following:

$$\boldsymbol{P}^{(v)} = \boldsymbol{S}^{(v)} \times \Big(\frac{\sum_{k \neq v} \boldsymbol{P}^{(k)}}{m - 1} \Big) \times (\boldsymbol{S}^{(v)})^T, v = 1, 2, ..., m$$

Spectral clustering is applied on the final kernel matrix $\boldsymbol{P}^{(c)}$ to determine patient clusters. In [264] the method has been applied on four cancer types to divide the samples into 3 to 5 clusters. Log-rank tests are performed to evaluate the clusters. It shows that individual data types are mostly not able to stratify samples into significantly different prognostic groups while the proposed method can lead to significant stratifications.

The second algorithm we employed is iCluster [219]. It integrates multiple omics data simultaneously to cluster samples by employing the joint latent variable model. Let $\boldsymbol{X}$ denote the data matrix of dimension $p \times n$ with rows being genes and columns being samples. $\boldsymbol{Z} = (\boldsymbol{z_1}, \boldsymbol{z_2}, ..., \boldsymbol{z_{K-1}})'$ is the cluster indicator matrix of dimension $(K-1) \times n$. K is the number of clusters. $\boldsymbol{W}$ is the coefficient matrix of dimension $p \times (K-1)$. $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_p)'$ is a set of independent error terms. A Gaussian latent variable model of the eigengene k-means clustering is written as:

$$\boldsymbol{X} = \boldsymbol{W}\boldsymbol{Z} + \boldsymbol{\varepsilon}$$

The basic concept of iCluster is to jointly estimate $\boldsymbol{Z}$ from multiple omics data: $\boldsymbol{X}_1$ of dimension $p_1 \times n$ for CNA data, $\boldsymbol{X}_2$ of dimension $p_2 \times n$ for DM data, $\boldsymbol{X}_3$ of dimension $p_3 \times n$ for GE data. The form of the integrative model is:

$$\begin{aligned} \boldsymbol{X}_1 &= \boldsymbol{W}_1\boldsymbol{Z} + \boldsymbol{\varepsilon_1} \\ \boldsymbol{X}_2 &= \boldsymbol{W}_2\boldsymbol{Z} + \boldsymbol{\varepsilon_2} \\ &\vdots \\ \boldsymbol{X}_m &= \boldsymbol{W}_m\boldsymbol{Z} + \boldsymbol{\varepsilon_m} \end{aligned}$$

$\boldsymbol{Z}$ is the latent component that connects the m omics data types measured on the same set of samples. It is estimated using expectation maximization algorithm to decide the cluster memberships of samples.

Results and Analysis

We used EMT FSFs as the input to SNF and iCluster algorithms. Each algorithm takes as input the combinations of the FSFs from individual data levels and divides the samples into 3 clusters. Log-rank tests are performed on the resulting clusters. Table 4.3 shows the results of using SNF algorithm. Table 4.4 shows the results of using iCluster algorithm.

Compared with the results of using k-means algorithms in Table 4.1 for all stage samples and in Table 4.2 for early stage samples, we have observed the following:

- Although more sophisticated, these two algorithms give inferior performance than k-means algorithm, while SNF performed better than iCluster. In the original article [264] both SNF and iCluster algorithms were applied on multiple cancer types. Although the results of integrative clustering were shown to be better than using single data levels, the p-values they obtained from log-rank tests (the lowest p-value was 2.0e-4) are much less significant than our results, not to mention that much more features were used in these studies. Even for clustering early stage samples, we have obtained good sample stratifications. This further demonstrates the advantages of using phenotype relevant biological networks for feature selection.

Table 4.3: The p-values of log-rank tests based on SNF clustering using different data level combinations with extended EMT network. We highlighted all p-values that are lower than 10e-5.

	GE	DM	CNA	GE+DM	GE+CNA	DM+CNA	GE+DM +CNA
t-test	2.71e-05	6.33e-02	3.32e-01	6.81e-03	**5.61e-06**	2.22e-03	1.72e-03
Lasso	7.95e-04	9.79e-02	5.05e-01	4.39e-03	**8.30e-08**	1.87e-03	8.73e-03
NetLasso	7.09e-02	5.91e-01	2.51e-01	2.70e-01	2.38e-02	9.23e-02	7.39e-02
addDA2	**7.19e-07**	**8.27e-10**	8.50e-04	**9.60e-12**	1.36e-03	7.41e-02	**5.25e-09**
Netrank	2.05e-03	3.45e-01	3.17e-01	**1.61e-06**	1.09e-03	1.76e-03	2.07e-05
stSVM	9.83e-02	4.11e-01	5.76e-01	5.49e-01	5.58e-01	7.25e-01	6.61e-01
Cox	7.50e-05	9.18e-04	1.50e-01	3.33e-03	3.53e-03	4.47e-03	8.89e-04
RegCox	2.01e-03	6.60e-01	1.76e-01	**2.39e-08**	4.34e-02	**6.70e-06**	**1.28e-08**
MSS	9.45e-04	6.22e-02	3.47e-01	1.04e-02	1.63e-02	2.48e-01	6.00e-04
Survnet	9.44e-05	5.75e-02	3.14e-02	2.71e-03	4.95e-05	6.96e-05	9.74e-05
Ensemble	1.14e-03	4.10e-03	6.14e-02	1.20e-03	1.05e-02	5.73e-04	2.51e-04
allemt	1.94e-02	7.26e-01	2.70e-01	1.91e-01	3.41e-01	7.15e-01	5.92e-01

Table 4.4: The p-values of log-rank tests based on iCluster clustering using different data level combinations with extended EMT network. We highlighted all p-values that are lower than 10e-5.

	GE	DM	CNA	GE+DM	GE+CNA	DM+CNA	GE+DM +CNA
t-test	1.32e-02	3.16e-01	7.70e-02	1.24e-01	4.92e-04	6.65e-04	2.79e-01
Lasso	2.16e-01	8.66e-01	7.33e-01	1.26e-03	4.74e-05	1.64e-01	2.00e-04
NetLasso	4.23e-01	8.34e-01	7.63e-01	6.22e-01	3.75e-02	3.49e-01	4.62e-03
addDA2	1.69e-02	1.57e-02	4.23e-03	2.36e-01	3.48e-03	2.75e-02	2.74e-01
Netrank	2.41e-02	2.66e-01	4.33e-02	2.47e-02	5.04e-03	7.67e-02	7.71e-04
stSVM	**2.77e-08**	7.28e-01	1.33e-01	2.12e-01	7.46e-02	6.51e-01	9.37e-01
Cox	9.48e-04	6.85e-04	6.13e-01	4.01e-03	5.17e-04	3.07e-04	2.57e-02
RegCox	9.51e-01	2.61e-01	2.72e-02	2.22e-03	**6.12e-06**	3.57e-03	**2.96e-08**
MSS	3.76e-02	2.30e-01	1.33e-01	4.24e-03	5.75e-01	9.48e-01	2.06e-01
Survnet	1.25e-03	2.60e-01	6.94e-02	7.15e-01	2.71e-02	5.93e-02	4.51e-02
Ensemble	9.75e-03	8.77e-01	7.73e-02	2.50e-05	2.10e-03	1.10e-03	1.09e-03
allemt	7.15e-01	9.65e-01	2.68e-01	1.29e-01	7.58e-02	5.54e-01	1.70e-01

- In terms of feature selection algorithms, we have observed consistent conclusions that addDA2 algorithm, Lasso, t-test, Netrank algorithms give better clustering results than the other algorithms. It indicates that feature selection algorithms are more stable than the choice of clustering algorithms.

- We have observed consistent conclusions with SNF algorithm that integrated molecular signatures from more than one data levels can better divide the samples into different prognostic groups.

4.2 A Novel Integrative Feature Selection Approach

Although individual omic data levels have been investigated for biomarker identification, intrinsically integrative data analysis methods are still lacking. The benefit of identifying molecular signatures across multiple data levels simultaneously remains to be revealed. In this section we will explore this direction. Since we have three omics data levels, one can imagine having three identical EMT networks mapped with values. These three layers of networks can be laid on top of each other to form a vertical network - a multiplex, with the correspondence of genes on each layer. One can also imagine that each network node is not only mapped with one value, but a vector of values from different data levels. Then the network becomes a network of vectors. Following different network constructions, one can formulate the task differently, e.g., identifying molecular signatures on a multiplex, or identifying signatures by combining vectors of values.

4.2.1 The Construction of Multiplex Networks

We will focus on the formulation of multiplex for two main reasons. The first is that it offers a natural extension of the feature selection methods used in the last chapter, which take the input of a network and a data matrix to be mapped on the network. The second reason is that multiplex is a suitable structure for representing hierarchies of networks. This is desired because biological systems are often composed of multiple layers of regulations [173, 280]. The regulations happen both within layers and across layers [75, 130]. Multiplex network has been recently applied on multiple omics data of yeast (transcriptomic and fluxomic layers) to better infer similarity of growth conditions.

Recent reviews [18, 110] included studies that use multiple omic data sequentially or simultaneously for different analysis in bioinformatics. However, most of the studies focus on gene privatization, studying the intrinsic relations between different data sources, and sample clustering. Finding features for prognostic prediction using multiple data sources simultaneously has not been investigated. Based on the EMT networks, which are shown to be predictive and of reasonable sizes, we aim to identify molecular signatures using multiplex-based feature selection.

We extend the algorithms in the previous chapter to incorporate multiple data levels. For algorithms that do not use network information, we apply them on concatenated features of multiple levels. For methods that use network information, we form a multiplex and apply those algorithms. While concatenation is rather straightforward, we would like to elaborate on how to build the multiplex. An interesting question is how to decide edges

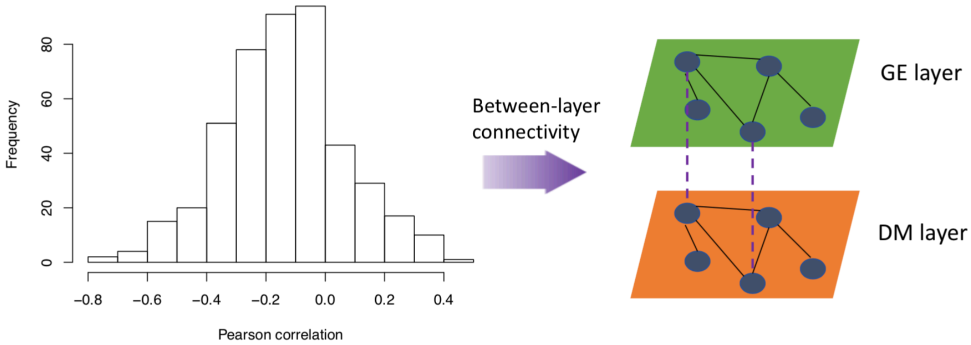


Figure 4.5: An illustration of multiplex construction based on Pearson correlation coefficients using GE and DM data.

that connect vertical layers. Although it has been long acknowledged that DNA methylation patterns and copy number alterations affect gene expression, we have not found previous work that try to identify molecular signatures using a multi-layered molecular network structure. Existing studies have, as introduced previously, simultaneously considered different layers of patient similarity networks to more accurately cluster samples into subtypes [129, 264]. Here we use two unsupervised methods for constructing the inter-layer edges.

- Pearson correlation coefficient (PCC). We calculated the correlations between GE and DM, and between GE and CNA for each gene. Edges are added when the absolute value of correlation exceeds a threshold. Assuming that certain levels of correlation indicate the effects of DM and CNA on gene expression, adding edges between them reflects this influence. An illustration is provided in Figure 4.5.

- CNAmet scores proposed by [158]. They employed signal-to-noise ratio [98] to relate CNA and DM data to GE data for the same gene. The method first dichotomizes DM and CNA data $M_{cn}, M_{me} \in \{0,1\}^{m \times p}$, where m is the number of samples and p is the number of features. For each DM feature i, a weight score is calculated as:

$$W^i_{dm} = \frac{\mu^i_{dm,1} - \mu^i_{dm,0}}{\sigma^i_{dm,1} + \sigma^i_{dm,0}},\ \sigma^i_{dm,1} > 0,\ \sigma^i_{dm,0} > 0,$$

  where $\mu^i_{dm,1}$ is the average expression of gene i when its DM level is high and $\sigma^i_{dm,1}$ is the standard deviation of gene i expression when its DM level is high. Likewise, the scores for CNA W^i_{cna} can be calculated. Then for each gene i permutation tests are performed on W^i_{dm} and W^i_{cna} by randomly permuting the labeling vectors and recalculating the scores. The p-values of the permutation tests are used to assess the influence of DM and CNA status on gene expression.

Table 4.5: The numbers of between-layer and within-layer edges using multiplex network for all three EMT networks using GE and DM data.

Edge type	Between-layer # edges					Within-layer # edges
Scoring method	PCC		Permutation test		all edges	
Threshold	*abs* > 0.2	*abs* > 0.1	p<0.01	p<0.05	none	
$\|V(G)\| = 74$ (core network)	34	51	24	34	74	113
$\|V(G)\| = 123$ (filtered network)	65	95	46	62	123	253
$\|V(G)\| = 455$ (extended network)	198	318	141	215	455	2620

Table 4.6: The number of samples after combining different data levels. For each data level combination, we kept the samples that have data for all involved data levels. Within one data level combination, the number of samples is the same for different network sizes.

Data level combinations	Classification		Survival analysis
	good prognosis	poor prognosis	
GE+DM	73	92	442
GE+CNA	72	74	494
DM+CNA	66	69	444
GE+DM+CNA	65	68	439

4.2.2 Experiments

We would like to evaluate the features selected from a multiplex that is mapped with different levels of omics data. For network-based feature selection algorithms, we used the above two methods to construct a multiplex. Certainly, the number of inter-layer edges depends on the threshold on Pearson correlation and CNAmet scores. We calculated both scores and used different thresholds for determining the inter-layer edges. We first constructed the multiplex using GE and DM data. The thresholds for determining the edges, and the resulting numbers of between-layer edges are given in Table 4.5. Besides, we would like to compare the prediction performance of multi-level molecular signatures with that of single-level signatures. Note that the results obtained from the last chapter are not directly comparable because here we can only use the samples that have the information of all the involved data levels. Therefore, we experimented on single data levels in parallel to enable paired comparisons. The sample sizes for classification and survival analysis are given in Table 4.6.

4.2.3 Results

We have obtained the AUC, AUPR, accuracy for both concatenation-based algorithms and multiplex-based algorithms. The test was performed using 30 times 10-fold cross validation. For network-based feature selection algorithms, this procedure was performed on each different settings of between-layer connectivity. Eventually, for each combination of network and feature selection algorithm we picked the network connectivity that gave then highest average AUC value. To enable an objective comparison, the data folds were kept the same for comparative groups. For example, concatenation-based feature selection using GE and DM data was tested using the same cross-validation folds as individual GE and DM data levels. Multiplex-based feature selection was tested using the same cross-validation folds as network-based feature selection algorithms on individual data levels. Since we have observed from Chapter 3 the importance of using FSFs and clinical features, here we also tested the combination of FSFs with clinical features.

Concatenation-based algorithms

Figure 4.6 shows the average AUC values with GE data, DM data, and concatenated data. Figure 4.7 shows the average AUC values of all three individual data levels and concatenated data. Both figures have two sides. On the left side we used only molecular features and on the right side we combined clinical features with FSFs. In each side there are three panels, corresponding to the three EMT network sizes. The results for the concatenation of GE and CNA features are given in Figure 7.6.

We have observed the following from Figure 4.6 and Figure 4.7:

- Without using FSFs, the concatenation of data levels mostly did not give better performance. Comparatively, DM data have better performance with the core EMT network. FSFs are shown to be much more robust and have achieved much better prediction performance, as also observed in Chapter 3.

- When combining FSFs with clinical features, Lasso feature selection algorithm has achieved significantly better prediction performance with concatenated data in both figures. t-test feature selection has achieved better performance when concatenating GE data with DM data using filtered and extended EMT features.

Multiplex-based algorithms

Figure 4.8 shows the average AUC values with GE network, DM network, and the multiplex. Figure 4.9 shows the average AUC values with all three individual networks and the multiplex. Both figures have two sides. On the left side we used only molecular features and on the right side we combined clinical features with FSFs. In each side there are three panels, corresponding to the three EMT networks. The results for the multiplex combining GE and CNA features are given in Figure 7.7.

We have observed from Figure 4.8 and Figure 4.9 that in most of the cases, selecting molecular features from multiplex did not give significant performance gain. However, when combining FSFs with clinical features, multiplex-based feature selection has shown significant performance gain. For example, with addDA2 algorithm, using FSFs from the multiplex combining GE and DM data increased the average AUC value from 0.656 to 0.777. With FSFs from the multiplex of GE, DM and CNA data, the prediction performance is remarkable. The highest average AUC value (0.881) is achieved by addDA2 algorithm with extended EMT network.

Last, we compared concatenation-based feature selection and multiplex-based feature selection. We would like to address the following two observations:

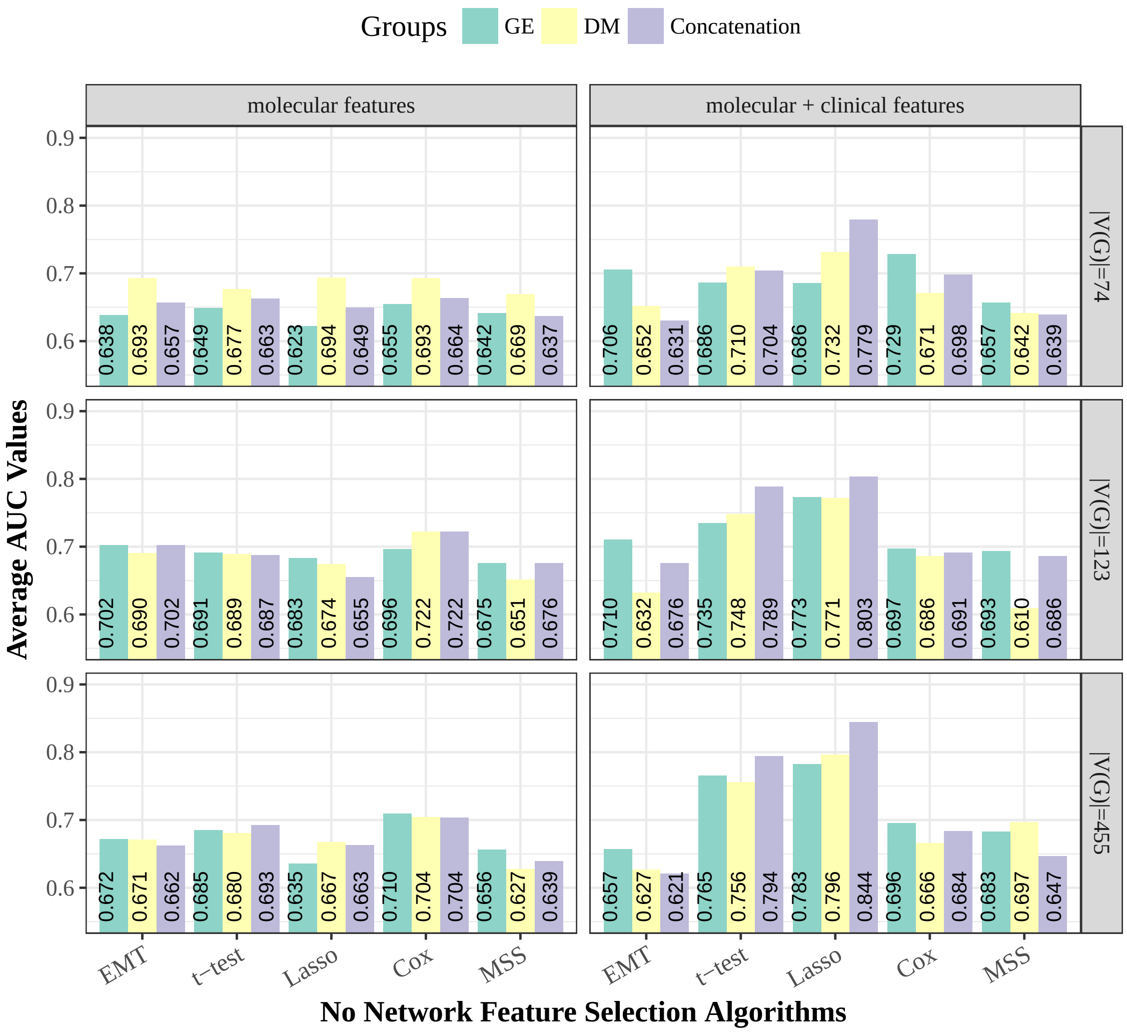


Figure 4.6: The average AUC values of concatenation-based feature selection using GE and DM data. The left side shows the performance of molecular features. The right side shows the performance of FSFs combined with clinical features. On each side we show the results of using 3 EMT networks.

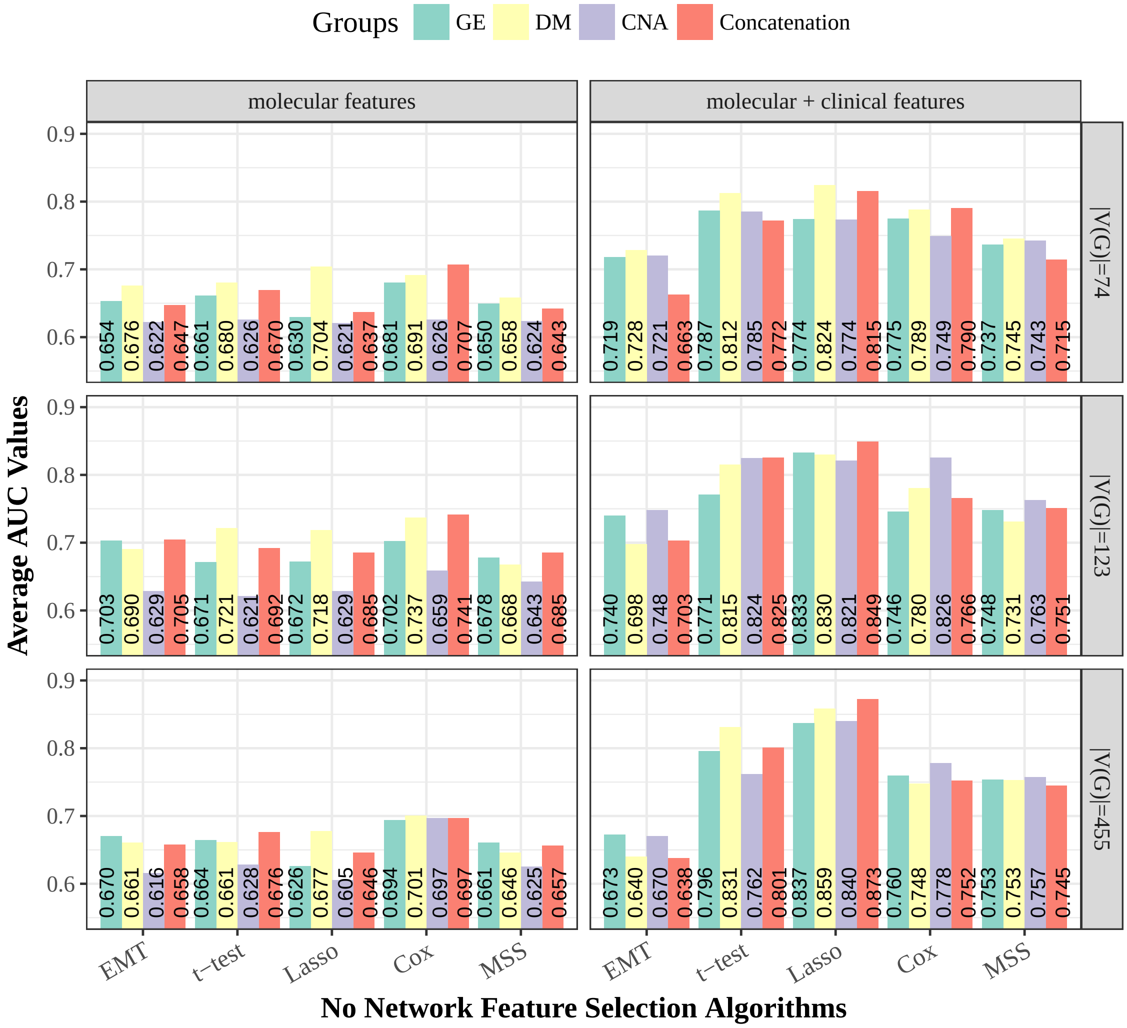


Figure 4.7: The average AUC values of concatenation-based feature selection using GE, DM, and CNA data. The left side shows the performance of molecular features. The right side shows the performance of FSFs combined with clinical features. On each side we show the results of using 3 EMT networks.

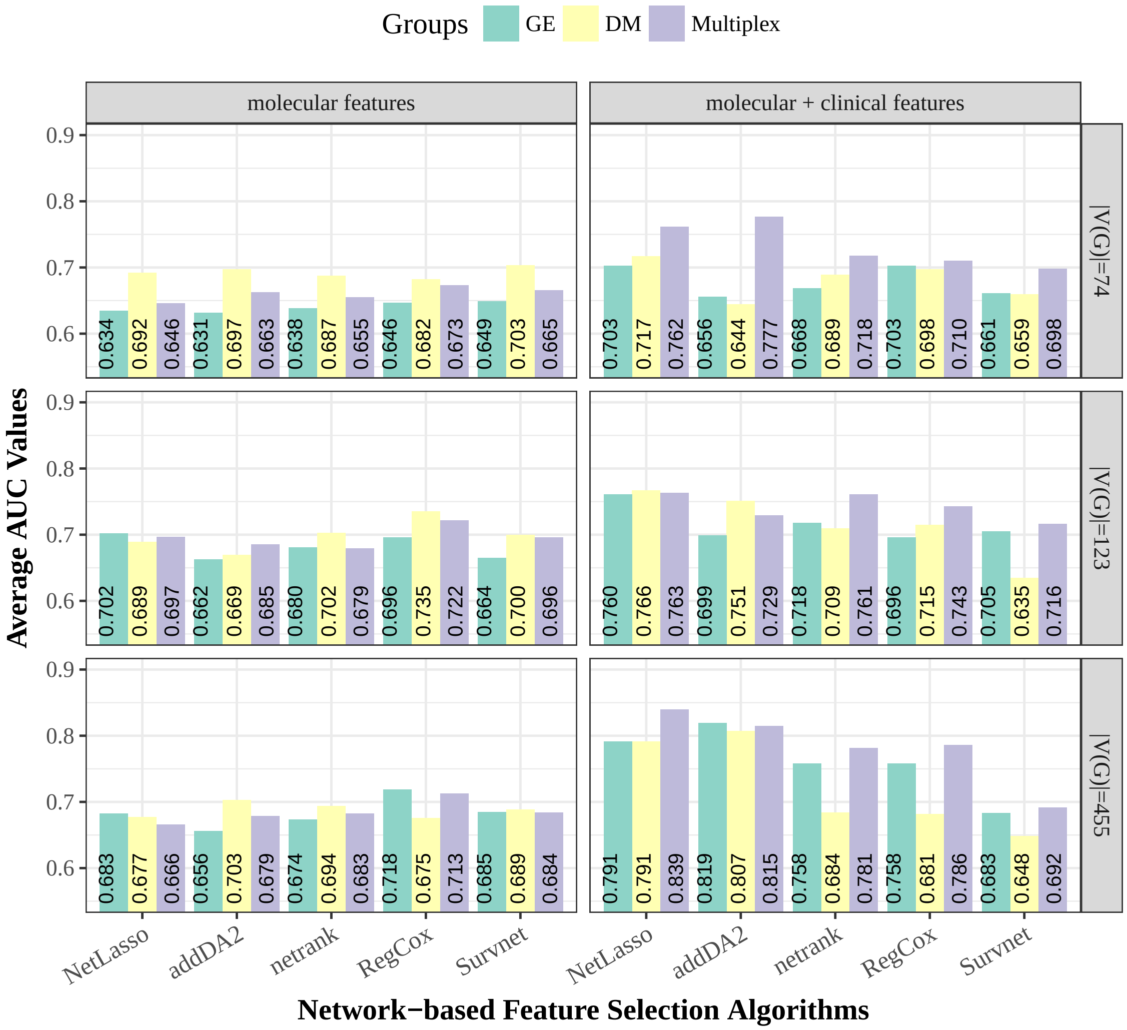


Figure 4.8: The average AUC values of multiplex-based feature selection using GE and DM data. The left side shows the performance of molecular features. The right side shows the performance of FSFs combined with clinical features. On each side we show the results of using 3 EMT networks.

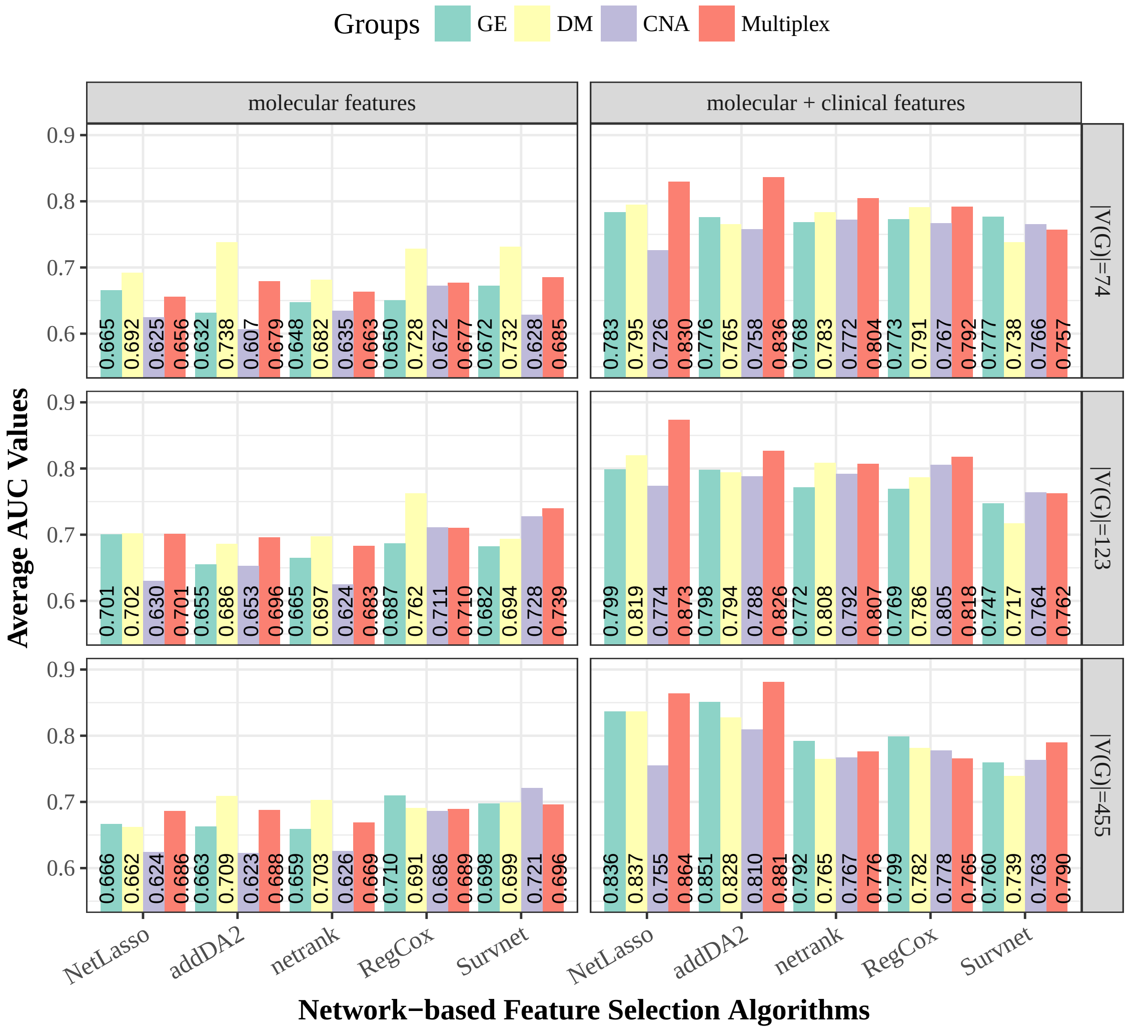


Figure 4.9: The average AUC values of multiplex-based feature selection using GE, DM, and CNA data. The left side shows the performance of molecular features. The right side shows the performance of FSFs combined with clinical features. On each side we show the results of using 3 EMT networks.

- While data concatenation did not show significant performance gain in most of the cases, using FSFs of multiplex-based feature selection has significant improvements in prediction performance.

- With network-based feature selection algorithms, the performance of FSFs together with clinical features increases as the size of network increases. With none network feature selection algorithms this is not apparent.

The second observation attracts our attention. It seems that with the increase of network size, the FSFs become more and more robust. This agrees with our experimental results where FSF boosted the prediction performance, especially when we use the extended EMT network. Therefore, we are curious to look closer at the FSFs with respect to the network sizes and the usage of multiplex network structures.

4.2.4 Multi-omics Feature Compositions

Since we mostly observed performance gain with FSFs, in this part we would like to understand two questions. The first one is that why the FSFs from concatenation improved the performance with some algorithms but not some other algorithms. The second one is why the FSFs from multiplex give better performance than the FSFs from single layer networks. Since the samples and cross-validation folds are kept the same in each experiment, we attribute the reasons to the differences in the features. Specifically, we will compare the features selected on individual data levels with that of joint selections to find out the potential reasons.

On FSFs from concatenated data

First, we analyzed FSFs selected on concatenated data. While individual sets of molecular features are not beneficial for the prediction, FSFs can give significantly better predictions. Even though clinical features are included to add more information, we know that the main contribution comes from the FSFs. This is supported by Figure 3.12 in chapter 3, where FSFs remarkably increased the AUC values. We think that whether concatenation-based feature selection improves the performance may be associated with the feature composition - the ratio of GE and DM features. It may be also relevant to the overlap between features selected on concatenated data and features selected on individual data levels.

To test our hypotheses, we calculated the feature ratios of the FSFs of the four none network-based feature selection algorithms on three EMT network sizes as well as the feature overlap. The results are given Table 4.7 for the concatenation of GE and DM data and Table 4.8 for the concatenation of all three data levels. We also list the average AUC values in the tables for a convenient comparison. We would like to address the following findings:

- Except for Cox and MSS algorithms with filtered EMT network, a higher ratio between non-GE features and GE features always associates with a better prediction performance in both tables. This suggests that algorithms that can select more balanced features across omics data levels can benefit more from data concatenation.

Table 4.7: The ratios between DM and GE features and the average AUC values for concatenation-based feature selection.

Algorithms	$\lvert V(G)\rvert = 74$			$\lvert V(G)\rvert = 123$			$\lvert V(G)\rvert = 455$		
	Ratio	AUC	Overlap	Ratio	AUC	Overlap	Ratio	AUC	Overlap
t-test	0.45	0.704	0.95	0.15	0.789	0.95	0.3	0.794	1
Lasso	0.70	0.779	0.85	0.50	0.803	0.9	0.4	0.844	1
Cox	0.4	0.698	1	0.05	0.691	1	0.15	0.684	1
MSS	0.05	0.639	0.55	0.20	0.686	0.7	0.00	0.647	0.45

Table 4.8: The ratios between non-GE and GE features and the average AUC values for concatenation-based feature selection.

Algorithms	$\lvert V(G)\rvert = 74$			$\lvert V(G)\rvert = 123$			$\lvert V(G)\rvert = 455$		
	Ratio	AUC	Overlap	Ratio	AUC	Overlap	Ratio	AUC	Overlap
t-test	0.5	0.772	1	0.35	0.825	0.9	0.3	0.801	1
Lasso	0.75	0.815	1	0.5	0.849	0.8	0.45	0.873	0.95
Cox	0.5	0.790	1	0.05	0.766	0.95	0.15	0.752	1
MSS	0.05	0.715	0.5	0.05	0.751	0.75	0.05	0.745	0.5

- We have not observed associations between feature overlap and AUC values. However, we find that when the size of network increases the feature overlap either becomes higher (Table 4.7) or remains on the same level (Table 4.8, except for MSS algorithm where the feature overlap is much worse than the other algorithms). This is very interesting because with a larger network size, the signatures from concatenated feature selection become more consistent with individual data levels. It indicates that the important features become more apparent to identify. But we think if the scope further increases, e.g., to the whole PPI network, this may not hold true. Since we used a phenotype relevant network, the extension from the core EMT network to the extended network added more useful information than noise.

On FSFs from multiplex

In this part we look at the FSFs of multiplex. There are two types of feature selection algorithms in Figure 4.8 and Figure 4.9. One type selects single features based on network connectivity - NetLasso, netrank, and RegCox. The other type selects subnetworks as features - addDA2 and Survnet. We have shown that the FSFs of these algorithms in most cases gave significantly better predictions, especially for NetLasso and addDA2 algorithms. We would like to find clues in the features.

Recall that in Chapter 3, we have found that the network properties of features selected at different data levels are significantly different. We have also used association rule approach to show that combining features from multiple data levels can provide interesting biological insights and improve the quality of rules. What is different now is that the features are not combined from different omics data levels, but directly selected from a multiplex. Therefore, first it is interesting to see how much the features selected from the two scenarios overlap. It is straightforward for single features. For subnetwork

Table 4.9: The overlap of FSFs on multiplex with FSFs on individual networks. The multiplex is composed of GE and DM layers.

Algorithms	$\vert V(G)\vert = 74$		$\vert V(G)\vert = 123$		$\vert V(G)\vert = 455$	
	GE	DM	GE	DM	GE	DM
NetLasso	0.3	0.45	0.35	0.25	0.25	0.25
addDA2	0.3	0.33	0.36	0.25	0.24	0.49
Netrank	0.5	0.4	0.65	0.2	0.55	0.25
RegCox	0.6	0.4	0.65	0.35	0.75	0.20
Survnet	0.63	0.16	0.42	0.0	0.53	0.00
addDA2(net)	3	3	4	2	2	4
Survnet(net)	9	2	3	0	1	0

Table 4.10: The overlap of FSFs on multiplex with FSFs on individual networks. The multiplex is composed of GE, DM, and CNA layers.

Algorithms	$\vert V(G)\vert = 74$			$\vert V(G)\vert = 123$			$\vert V(G)\vert = 455$		
	GE	DM	CNA	GE	DM	CNA	GE	DM	CNA
NetLasso	0.15	0.25	0.15	0.25	0.25	0.05	0.20	0.25	0
addDA2	0.14	0.45	0	0.27	0.18	0	0.11	0.38	0
Netrank	0.60	0.15	0.10	0.50	0.20	0.05	0.50	0.25	0
RegCox	0.60	0.2	0.2	0.55	0.25	0.20	0.70	0.20	0
Survnet	0.44	0.17	0.11	0.57	0	0.04	0.47	0	0
addDA2(net)	1	4	0	3	2	0	0	0	0
Survnet(net)	5	0	0	6	0	0	1	0	0

features, we calculated both the molecular level overlap and the proportion of multiplex subnetworks that contain FSFs (subnetworks) of individual data levels. Table 4.9 shows the feature overlap using the multiplex of GE and DM networks. Table 4.10 shows the feature overlap using the multiplex of GE, DM and CNA networks. We observe that the feature overlap in general is smaller than using concatenation-based feature selection. The total feature overlap tends to decrease with the increase of network size. We have not observed associations between feature ratios of non-GE features and GE features and the prediction performance (results not shown).

Then we want to know whether multiplex-based features are superior, e.g., can provide better association rules for prognosis prediction. Following the experimental setting in Chapter 3 we derived association rules and evaluated the confidence of top ranked rules. The results are given in Figure 4.10. We observed that multiplex-based features give equivalent or higher rule confidence in most of the cases. With the increase of network size, we observed that multiplex-based rules have always increasing average confidence. The same does not hold true for the features from single-layered networks. This shows the advantage of integrating information on a multiplex, where the FSFs show higher importance. We would like to draw the attention that due to the high sensitivity of feature selection algorithms to the sample folds, individually selected features with multiplex did not show clear advantages, which can be seen on the left sides of Figure 4.8 and 4.9.

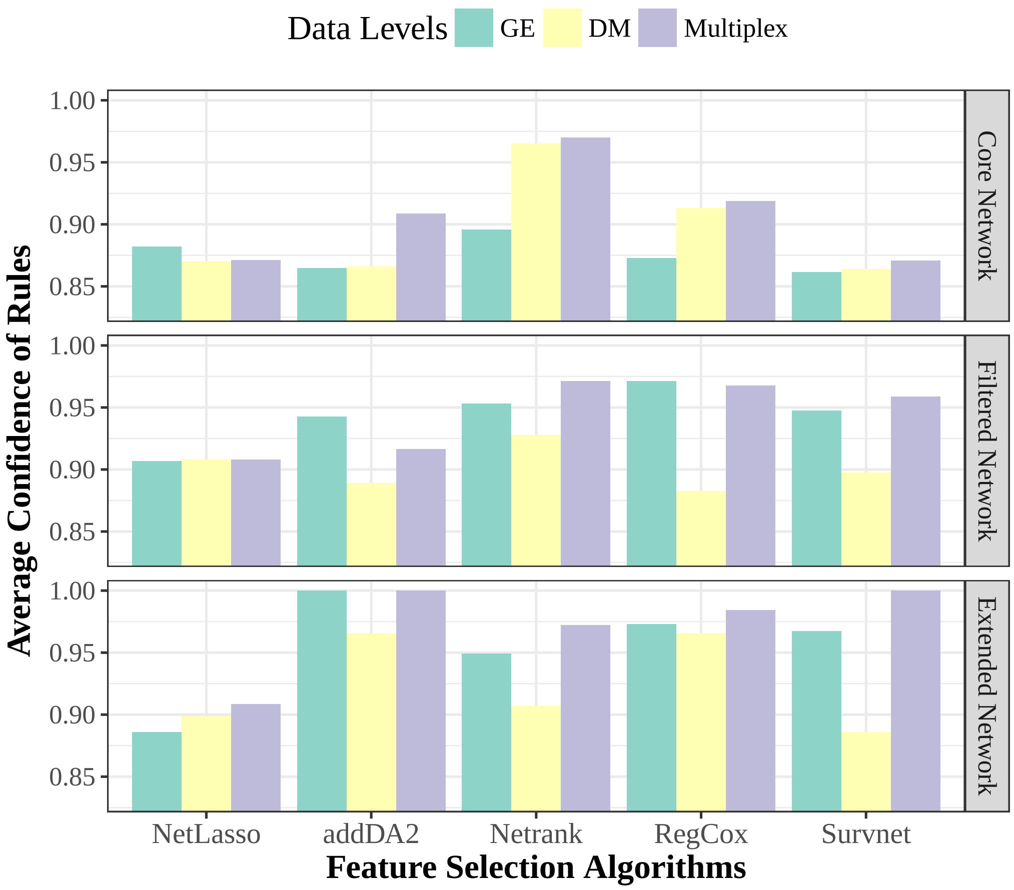


Figure 4.10: The average confidence of prognostic association rules derived from the FSFs on single-layered networks and multiplex. The top 30 rules are taken for evaluation.

However, with FSFs, multiplex always shows superiority over single-layered networks.

4.3 EMT-based Prognosis Prediction on Real-world Clinical Datasets

In this part we tested EMT-based molecular signatures identified in Chapter 3 and Chapter 4 on an independent cohort of LUAD samples. Out testing is novel because the independent data have both GE and DM profiles for the same group of samples. The data were provided by our collaborating partner in Oslo University Hospital (OUH) from their study [21]. The data include 164 samples for DNA methylation data. A subset of these samples (n=121) has mRNA expression data available. The mRNA expression analysis was assessed using gene expression microarrays from Agilent technologies (SurePrint G3 human GE, 8 x 60 K). The mRNA expression data were log2 transformed and normalized between arrays by using the 75th percentile method in Genespring GX analysis Software v.12.1 (Agilent technology). The DNA methylation data were generated using Illumina Infinium

Table 4.11: Fitting univariate Cox PH models with probes and gene aggregates for microarray data. We listed below the number of probes/genes with p <0.01 or p <0.05 out of all probes or genes that belong to the EMT network.

Microarray data	p <0.05	p<0.01
Probe level	72/671	20/671
Aggregate by mean	53/455	14/455
Aggregate by median	51/455	16/455

HumanMethylation450BeadChips. Detailed experimental procedures and the processing of raw data to level 3 data are provided in [21].

The patient follow-up time ranges between 2 and 99 months with the median of 44 months. The outcome (event) is defined as the occurrence of relapse, distant metastasis or death. The time to progression was calculated from the date of surgery to the date of event. We evaluated individual level EMT features, combined GE and DM features, and multiplex features on this dataset using hierarchical clustering and survival analysis. The reason why we use hierarchical clustering was that the original article [21] used the same approach and thus we can relate our analysis directly to theirs. In the following we will introduce data pre-processing steps and then give the test results and analysis.

4.3.1 Data Pre-processing

We need to summarize the probe level microarray and DNA methylation data to gene level data. Since there can be multiple ways to do so, we used the p-value of univariate Cox PH model (using the follow up data) to decide which processing method to adopt. For microarray data there are in some cases more than one probes for one gene, either referring to transcript variants or being repetitions. We tried taking the mean or the median of duplicated probes. The results are shown in Table 4.11. As both options give similar outcome and median is more often used than mean aggregation, we adopted the median aggregation.

With DNA methylation data we tried three ways to take the aggregates: averaging all probes per gene, averaging the probes at the promoter region (TSS200 and TSS1500) per gene, and taking the probe that has the lowest Pearson correlation coefficient with the corresponding gene (using the median processing above). The results are shown in Table 4.12. Based on the results, we choose to average all probes per gene. This is also the data pre-processing method that is adopted for TCGA DNA methylation data in our previous experiments.

4.3.2 Results

We extract EMT signatures from the test data and conducted survival analysis. Note that we took the features directly as selected using TCGA data and applied on the test data without any training or using additional information. We would like to present the results in two parts: using single-omics features and using multi-omics features.

Table 4.12: Fitting univariate Cox PH models with probes and gene aggregates for DNA methylation data. We listed below the number of probes/genes with p <0.01 or p <0.05 out of all probes or genes that belong to the EMT network.

DNA methylation data	p <0.05	p<0.01
Probe level	342/11434	171/11434
Average all probes	89/455	33/455
Average promoter region probes	43/455	8/455
Least correlated probe	34/455	6/455

Using single-omics features

We tested the EMT features selected from TCGA RNA-Seq and miRNA data on the test (microarray) data. Since miRNA measurements are not available in the test data, we removed miRNA features from the signatures before the testing. If a subnetwork feature contains miRNA features, we removed the miRNAs from the subnetwork. Using EMT features in the testing set, we applied hierarchical clustering to divide the samples into 2 or 3 clusters and performed survival analysis. Table 4.13 shows the results of log-rank tests, on both 2 sample clusters and 3 sample clusters. With the same approach we used EMT DNA methylation features to stratify the samples. The testing results are given in Table 4.14.

Table 4.13: The p-values of log-rank tests using EMT gene expression features on independent data. The samples are divided into either 2 clusters (upper table) or 3 clusters (lower table). We highlighted all p-values that are lower than 0.01.

Network	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
74 nodes	4.55e-01	4.53e-01	2.53e-01	1.11e-01	8.11e-01	9.28e-02	1.83e-01	8.20e-01	7.00e-01	7.83e-02	5.50e-01	3.85e-01
123 nodes	3.36e-01	2.60e-01	9.54e-02	3.47e-01	8.17e-01	9.45e-01	1.33e-01	7.36e-02	1.85e-01	2.58e-01	2.55e-01	1.79e-02
455 nodes	2.04e-01	9.98e-01	1.21e-01	1.05e-01	3.25e-01	5.81e-01	4.68e-02	5.55e-01	2.04e-01	4.31e-02	**2.01e-03**	3.39e-01

Network	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
74 nodes	6.41e-01	4.59e-01	4.34e-01	**4.93e-03**	9.07e-01	2.39e-01	1.95e-01	9.37e-01	9.19e-01	1.50e-01	6.50e-01	2.35e-01
123 nodes	7.23e-02	2.48e-01	2.35e-01	6.43e-01	1.48e-02	2.42e-01	1.10e-01	1.85e-01	3.02e-02	1.84e-01	1.63e-01	5.81e-02
455 nodes	4.23e-01	8.33e-01	8.39e-02	9.00e-02	5.88e-01	6.53e-01	2.88e-02	6.87e-01	3.82e-01	6.41e-02	**6.14e-03**	5.19e-01

Table 4.14: The p-values of log-rank tests using EMT DNA methylation features on independent data. The samples are divided into either 2 clusters (upper table) or 3 clusters (lower table). We highlighted all p-values that are lower than 0.01.

Network	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
74 nodes	4.34e-01	8.00e-02	8.05e-01	3.38e-01	2.48e-01	2.21e-01	6.09e-01	2.33e-01	4.69e-01	7.23e-01	**2.20e-04**	5.17e-02
123 nodes	3.38e-01	8.92e-01	5.05e-01	**4.51e-03**	2.28e-01	1.01e-01	2.54e-02	2.54e-02	6.23e-01	1.11e-02	1.01e-01	2.29e-01
455 nodes	**5.81e-03**	1.05e-01	**8.84e-03**	**5.24e-03**	4.38e-01	**6.29e-03**	**1.60e-03**	2.39e-02	1.10e-01	7.28e-02	1.22e-01	**7.11e-03**

Network	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
74 nodes	6.76e-01	2.16e-01	4.92e-01	5.21e-01	3.52e-01	1.67e-01	1.02e-01	1.45e-01	7.59e-01	9.10e-01	**1.08e-03**	8.42e-02
123 nodes	5.89e-01	1.93e-01	6.48e-01	1.63e-02	1.28e-01	**7.07e-03**	8.21e-02	6.59e-02	7.07e-01	**3.78e-04**	2.60e-01	4.83e-01
455 nodes	1.58e-02	1.64e-01	2.85e-02	**8.44e-03**	3.62e-01	2.26e-02	**6.73e-03**	4.48e-02	8.82e-02	4.01e-02	2.75e-01	2.55e-02

Using multi-omics features

Based on the good sample stratification results of using combined features on TCGA data, we have performed the same testing on OUH data. We combined EMT signatures from GE data and DM data for each feature selection algorithm as the new feature set. For single gene features, we combine the top 10 features from GE data with the top 10 features from DM data, and compare it with using 20 features from GE data or using 20 features from DM data. With subnetwork features, we combine the top 7 subnetwork features from GE data with the top 7 subnetwork features from DM data, and compare it with using 10 subnetwork features from GE data or using 10 subnetwork features from DM data. The reason why we do not use 5 but 7 subnetwork features is to account for the overlap of subnetwork features. Using EMT features in the testing set, we applied hierarchical clustering to divide the samples into 2 or 3 clusters and performed survival analysis. Table 4.15 shows the results of clustering samples into 2 clusters. Table 4.17 shows the results of clustering samples into 3 clusters. We have also tested multiplex features on the test data. The results are given in Table 4.16 for separating samples into 2 clusters and Table 4.18 for separating samples into 3 clusters.

Table 4.15: The p-values of log-rank tests using combined EMT gene expression and DNA methylation features on independent data. The samples are divided into 2 clusters. We highlighted all p-values that are lower than 0.01.

Core EMT network

Data	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
GE	4.55e-01	4.53e-01	2.53e-01	1.11e-01	8.11e-01	9.28e-02	1.83e-01	8.20e-01	7.00e-01	7.83e-02	5.03e-01	1.23e-01
DM	1.60e-01	6.79e-01	9.25e-01	3.05e-01	3.37e-01	1.71e-01	9.20e-01	8.07e-02	4.43e-01	4.26e-01	1.96e-02	7.09e-01
GE+DM	3.37e-01	7.08e-01	1.79e-01	**4.90e-03**	5.67e-02	5.07e-01	1.49e-02	1.74e-01	3.94e-01	9.33e-01	2.71e-02	7.65e-01

Filtered EMT network

Data	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
GE	3.36e-01	2.60e-01	9.54e-02	3.47e-01	8.17e-01	9.45e-01	1.33e-01	7.36e-02	1.85e-01	2.58e-01	**4.19e-03**	4.88e-01
DM	3.37e-01	3.37e-01	5.03e-01	6.61e-01	2.23e-01	6.06e-02	8.96e-01	2.56e-02	4.41e-01	3.77e-02	3.37e-01	2.78e-01
GE+DM	4.30e-01	4.47e-02	5.03e-01	2.23e-01	2.23e-01	**6.77e-05**	5.87e-01	5.87e-01	2.77e-01	1.52e-01	**2.00e-03**	5.36e-01

Extended EMT network

Data	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
GE	2.04e-01	9.98e-01	1.21e-01	1.05e-01	3.25e-01	5.81e-01	4.68e-02	5.55e-01	2.04e-01	4.31e-02	3.73e-02	1.07e-01
DM	4.62e-01	2.48e-01	5.25e-01	3.58e-01	4.83e-01	1.99e-01	1.55e-01	3.21e-01	1.49e-01	7.23e-02	1.91e-02	4.66e-01
GE+DM	**2.78e-03**	6.58e-01	1.87e-01	3.49e-02	9.64e-02	6.22e-01	2.46e-01	1.50e-02	2.34e-02	7.58e-02	5.46e-02	7.87e-01

Table 4.16: The p-values of log-rank tests using concatenation/multiplex EMT features to separate samples into 2 clusters. We highlighted all p-values that are lower than 0.01.

	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
74 nodes	1.79e-02	6.55e-01	2.75e-01	4.76e-02	1.52e-02	1.52e-02	2.26e-01	1.33e-01	9.54e-02	6.81e-02	**3.82e-04**	6.45e-01
123 nodes	1.23e-01	2.49e-01	1.28e-01	9.54e-02	**1.30e-04**	4.28e-02	2.25e-02	8.96e-01	2.28e-02	6.04e-02	3.81e-01	1.44e-01
455 nodes	4.43e-01	1.17e-01	1.58e-02	1.87e-02	**1.08e-04**	**6.10e-03**	2.43e-01	9.54e-02	2.73e-02	3.22e-02	9.54e-02	7.87e-01

Table 4.17: The p-values of log-rank tests using combined EMT gene expression and DNA methylation features on independent data. The samples are divided into 3 clusters. We highlighted all p-values that are lower than 0.01.

Core EMT network

	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
GE	6.41e-01	4.59e-01	4.34e-01	**4.93e-03**	9.07e-01	2.39e-01	1.95e-01	9.37e-01	9.19e-01	1.50e-01	6.05e-01	2.80e-01
DM	2.84e-01	8.62e-01	5.07e-01	5.08e-01	1.56e-01	3.31e-01	8.27e-01	2.01e-01	3.80e-01	6.36e-01	6.18e-02	4.35e-01
GE+DM	9.45e-02	7.20e-01	2.64e-01	1.78e-02	1.53e-01	6.65e-01	**4.32e-03**	3.29e-01	1.23e-01	9.90e-01	6.21e-02	3.61e-01

Filtered EMT network

	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
GE	7.23e-02	2.48e-01	2.35e-01	6.43e-01	1.48e-02	2.42e-01	1.10e-01	1.85e-01	3.02e-02	1.84e-01	1.51e-02	2.19e-01
DM	6.27e-01	2.66e-01	7.95e-01	6.14e-01	1.87e-01	1.21e-01	1.97e-01	2.67e-02	5.96e-01	1.15e-01	4.95e-01	4.32e-01
GE+DM	6.14e-01	7.98e-02	6.08e-01	2.61e-01	1.84e-01	**2.83e-04**	4.19e-01	6.76e-01	5.27e-01	2.70e-01	**7.72e-03**	3.28e-01

Extended EMT network

	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
GE	4.23e-01	8.33e-01	8.39e-02	9.00e-02	5.88e-01	6.53e-01	2.88e-02	6.87e-01	3.82e-01	6.41e-02	9.99e-02	1.78e-01
DM	5.93e-01	3.88e-01	6.74e-01	2.36e-01	7.77e-01	3.54e-01	3.63e-01	3.96e-01	2.99e-01	1.94e-01	6.36e-02	7.67e-01
GE+DM	**5.84e-03**	5.80e-01	1.13e-01	1.08e-01	**6.24e-03**	**3.38e-03**	1.14e-01	1.73e-02	7.49e-02	1.18e-01	1.53e-01	4.26e-01

Table 4.18: The p-values of log-rank tests using concatenation/multiplex EMT features to separate samples into 3 clusters. We highlighted all p-values that are lower than 0.01.

	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	ensemble	all
74 nodes	3.99e-02	7.18e-01	4.63e-02	9.86e-02	1.74e-02	3.01e-02	2.40e-01	2.35e-01	1.34e-01	6.19e-02	**1.04e-03**	3.51e-01
123 nodes	2.02e-02	**1.85e-04**	2.60e-01	5.20e-02	**6.16e-04**	1.04e-01	3.54e-02	6.11e-01	4.35e-02	**2.46e-03**	3.24e-02	1.98e-01
455 nodes	2.42e-01	1.80e-01	5.10e-02	**7.40e-03**	**2.00e-04**	1.88e-02	1.52e-01	9.94e-02	**1.22e-03**	3.07e-02	9.37e-02	4.26e-01

From the test results above we have a few interesting findings:

- Some sets of EMT features selected from TCGA data are able to significantly stratify samples in OUH data. Examples are shown in Figure 4.11. EMT DM features in general show better performance than EMT GE features. This may be due to the platform difference in transcriptomics profiling between TCGA data and the testing data. It could also because of the absence of miRNA information in the testing data.

- With some feature selection algorithms, combining features from the two data levels can lead to significantly improved clustering results. Examples are shown in Figure 4.12 and Figure 4.13. Since the sample size of the test data is much smaller than TCGA data, it is harder to obtain a significant sample stratification. Nevertheless, the p-values we obtained in several cases are even lower than the results obtained in the original study. This shows that EMT features have biological significance and can give robust prognosis predictions.

- Multiplex-based feature selection shows slightly better sample stratifications than combining features from individual data levels.

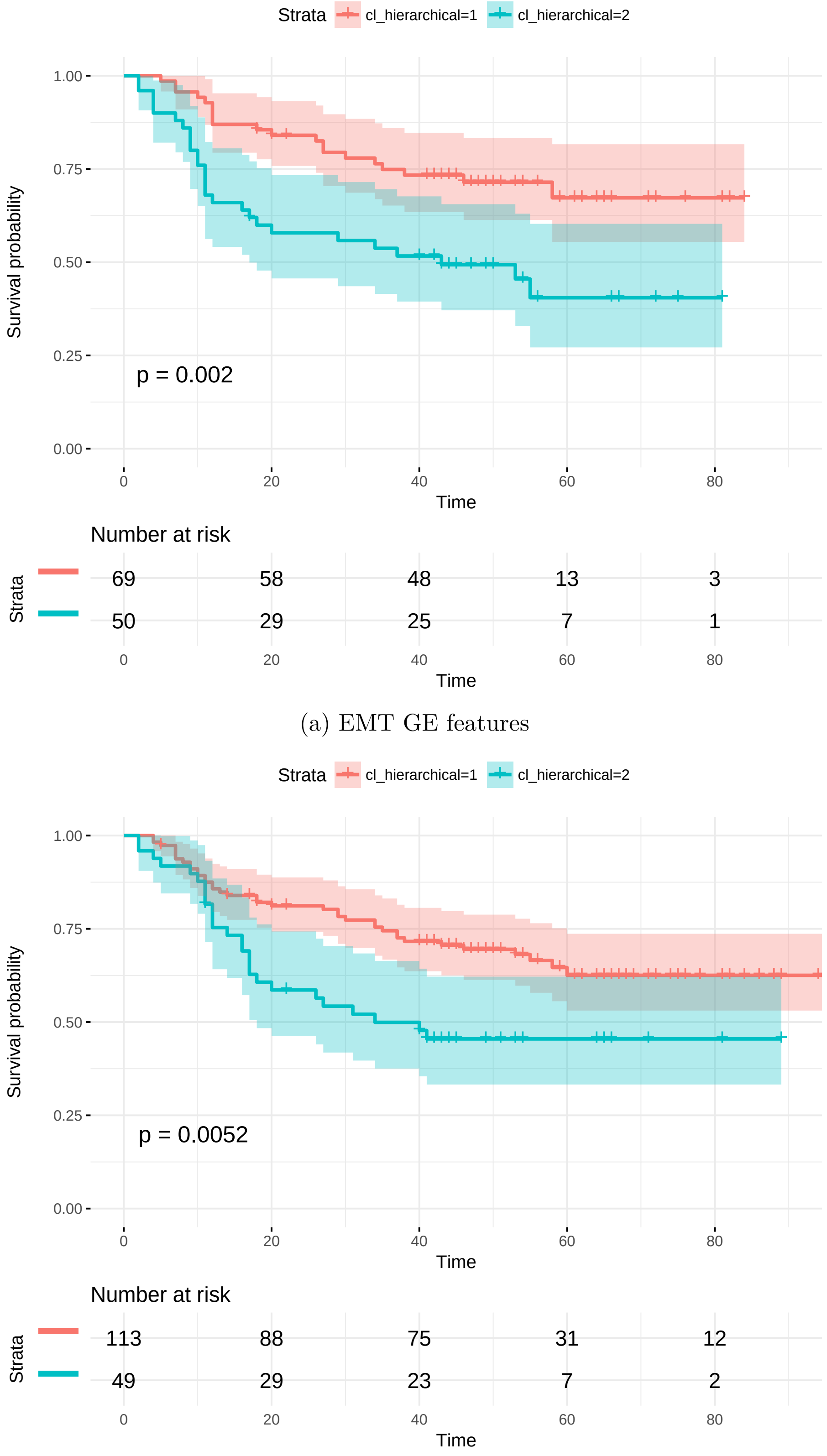
4.4 Discussion

4.4.1 Advantages of Integrative EMT Signatures

In this chapter we fist improved the prediction performance remarkably by combining EMT features selected from individual data levels. This has been shown on all stage samples and also on early stage samples. Although it is generally hard to predict the prognosis of early stage patients, EMT features achieved significant sample stratifications. We have also tested two integrative clustering algorithms - SNF [264] and iCluster [219], to see whether they can further improve sample clustering. However, although the resulting sample clusters show significantly different clinical outcome, the stratification is no better than using k-means clustering. As we already show that EMT features are capable of clustering patients into different outcome groups, the inferior results could be due to the algorithms themselves. [264] assumes that local similarities on the patient similarity network are more important than remote similarities. [219] models the relationships among multiple omics data using latent variable model assuming that there is a single cluster assignment for the samples across all omics levels. These assumptions may not agree with the data. The exact causes of this inferior performance need to be further investigated.

Note that in the original articles, these two algorithms use all the dimensions of several omics data types without feature selection. This causes much irrelevant information to influence the sample clustering. Nevertheless, using multiple omics data types improved sample stratification. However, in both articles it is shown by log-rank tests that the significance level was rather low (on the scale of 10e-4 for 3 sample clusters) compared with our results, although the data they employed contain all the EMT features that we used. This highlights the importance of extracting phenotype relevant features for subsequent analysis.

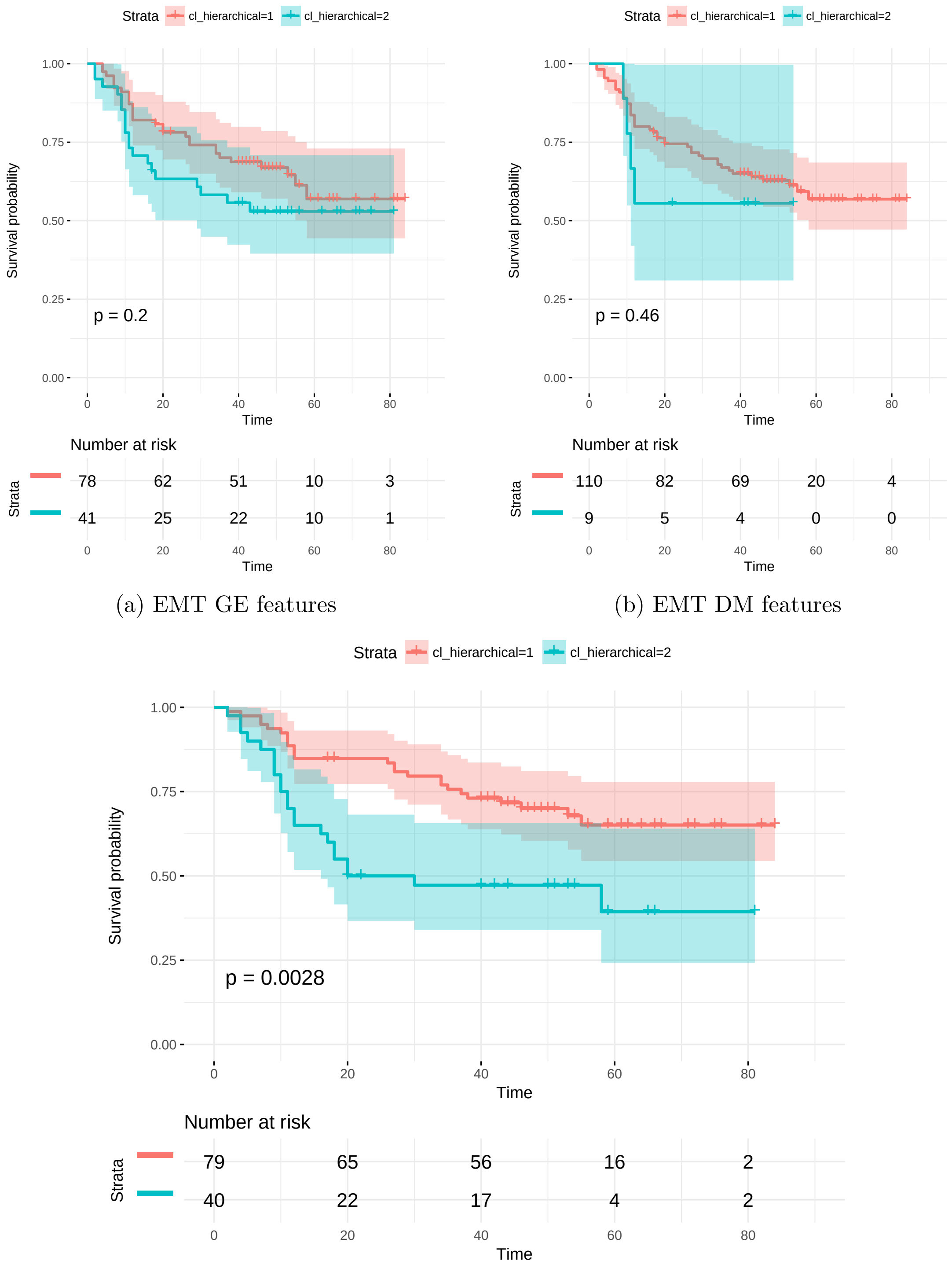
We have also proposed a multiplex-based feature selection approach that extends network-based feature selection algorithms to incorporate multiple network layers. To



(a) EMT GE features

(b) EMT DM features

Figure 4.11: Survival analysis on independent data using EMT single-level signatures. The samples are divided into two clusters with hierarchical clustering using (a) top 20 FSFs by ensemble algorithm with TCGA GE data, (b) top 10 FSFs (subnetworks) by addDA2 algorithm with TCGA DM data.

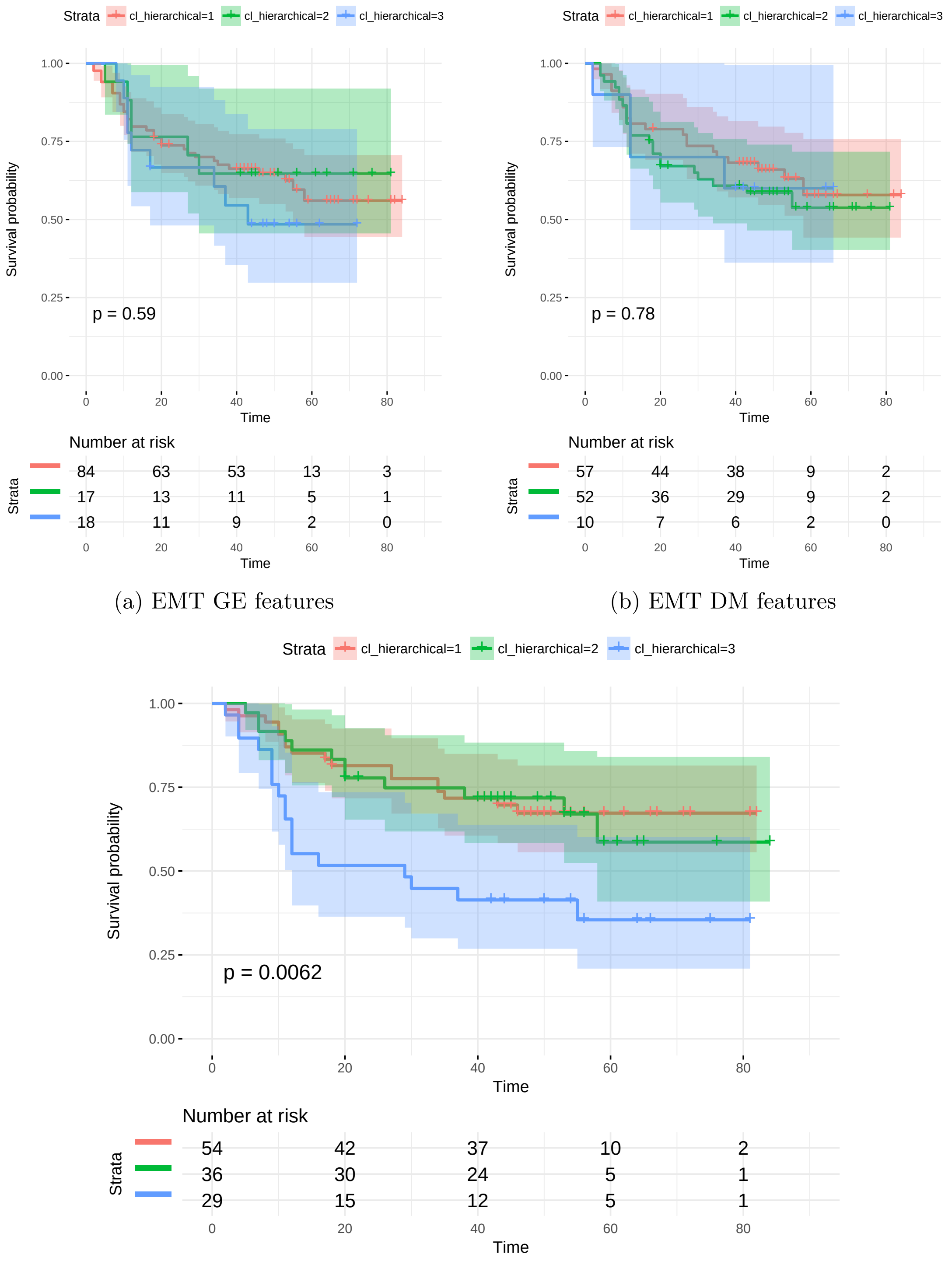


(a) EMT GE features

(b) EMT DM features

(c) Combined EMT GE and DM features

Figure 4.12: Survival analysis on independent data using EMT single- and multi-level signatures. The samples are divided into two clusters with hierarchical clustering using (a) top 20 FSFs by t-test with TCGA GE data, (b) top 20 FSFs by t-test with TCGA DM data, and (c) top 10 FSFs by t-test on TCGA GE data + top 10 FSFs by t-test on TCGA DM data.



(a) EMT GE features

(b) EMT DM features

(c) Combined EMT GE and DM features

Figure 4.13: Survival analysis on independent data using EMT single- and multi-level signatures. The samples are divided into three clusters with hierarchical clustering using (a) top 20 FSFs by Netrank algorithm with TCGA GE data, (b) top 20 FSFs by Netrank algorithm with TCGA DM data, and (c) top 10 FSFs by Netrank algorithm on TCGA GE data + top 10 FSFs by Netrank algorithm on TCGA DM data.

our knowledge, this is the first time when multiplex is employed in feature selection. We have tested two methods - Pearson correlation and CNAmet [158] for determining between-layer connections. Thanks to the construction of EMT networks which keeps important genes and meanwhile reduces the dimensionality, we are able to perform 30 times stratified 10-fold cross validation to evaluate the selected features. We have shown that due to the limits of sample size and heterogeneity, the overall performance still suffers from large variance. However, the FSFs from multiplex show significantly improved performance compared with the FSFs from single-level networks. This finding also agrees with our experiments in Chapter 3 where FSFs significantly improved the average prediction performance.

4.4.2 The Importance of Relevant Network-based Feature Selection

As we have shown, without feature selection it is difficult to stratify samples into different prognostic groups. However, selecting features using the whole dimensionality of the data is difficult, which often results in features that generalize poorly on independent data. By first constructing phenotype relevant networks and extracting these features from the original data, we can reduce the computational cost, decrease the randomness of selected features, and improve the predictive performance and biological interpretation of the molecular signatures. We successfully demonstrated that EMT network-based feature selection and data integration can provide advantages in selecting cancer prognostic signatures. We show that it is a good strategy to first use domain knowledge to extract relevant features and networks from the original data and then select molecular signatures in the low-dimensional space.

A similar idea to our strategy has been very implicitly and not systematically utilized in several recent studies, mainly as a remedy or trial-and-error approach to avoid dealing with many features. For example, [127] aims to predict clinical outcomes in ovarian carcinoma. They applied feature screening on each omics data type and then transformed the selected features into a pathway-based dataset using biological knowledge-base. Feature screening was applied again on this pathway dataset to select features, which were used for subsequent analysis. [107] aimed to select features for predicting distant metastases in lung cancer. They did not use all features as input but selected 4314 transcripts annotated as EMT-related according to the Ingenuity Knowledge Base. From these transcripts they selected 474 top-ranked transcripts according to their p-values of t-test between the two classes. On these 474 features an optimization approach was applied to select molecular signatures. In [299], where the algorithm RegCox was proposed, only a list of 2647 genes that were previously known to be related to cancer were used to identify features for cancer prognosis. In [35], where network-constrained support vector machine was evaluated, only 584 genes and 2280 interactions were used from all the features. These genes were either breast cancer related [76] or involved in estrogen signaling pathways according to Ingenuity Pathway Analysis. By giving these examples we show that recent studies tend to first shrink the scope of features by using biological database resources before selecting molecular signatures.

However, these studies follow their own methods to trim the search space. There is no consensus on how to do this systematically. This introduces much arbitrariness because the gene pre-selection can be highly dependent on the dataset. Therefore, it may not improve the robustness of the molecular signatures. Within our knowledge, our study is the first one to systematically construct a phenotype relevant network to identify molecular signatures. Based on the successful use case of EMT network-based feature selection in lung cancer prognosis prediction, we propose to construct an intelligent decision support system for a wide range of medical applications, as will be introduced in Chapter 6.

Chapter 5

Strategies for Handling Large Biological Data

In the previous chapters we have shown that integrative omics data analysis can significantly improve prognosis prediction. With the goal of individualized medicine, similar analysis can be performed on each individual patient to assist decision makings in medical treatments. Thanks to the decreasing cost of obtaining molecular profiles, medical treatments are becoming more and more individualized. As a result, this requires a powerful data infrastructure to support the storing and querying of large amount of omics data. So far this has not raised enough concerns because individual studies typically take several hundred samples which does not necessitate a data infrastructure. However, in the long term, we think it is necessary to investigate what could be a suitable data infrastructure to handle large amounts of omics data. Here we take proteomics data as an example because of their large size and complexity. In the following we will introduce proteomics data, relational and non-relational databases (NoSQL); and then we compare these database systems in their performance of storing and querying proteomics data.

5.1 Proteomics and Mass Spectrometry

In this section we will introduce the complexity of proteomics data, the analysis of proteomics using mass spectrometry (MS) technologies, and its potentials in biomarker discovery.

5.1.1 The Complexity of Proteomics Data

Proteomics is the large-scale experimental analysis of gene and cellular functions at the protein level. While the genome of an organism is more or less constant, protein expression varies in different cell types and different cellular conditions. Protein expression levels can hardly be predicted by genomic and transcriptomic analysis except in some cellular machines such as ribosome and cell adhesion complexes. For example, research has shown that there are only modest correlations between mRNA expression and protein abundance [58, 132, 206]. Multiple regulatory mechanisms can contribute to this phenomenon. For example, some gene transcripts are not translated to produce proteins [11]. Many mRNAs give rise to more than one proteins through alternative splicing. A mRNA produced in abundance may be degraded rapidly or translated inefficiently, resulting in a small amount of protein. Under distinct conditions such as cell cycle, cellular differentiation, or carcinogenesis, a cell can produce different sets of proteins.

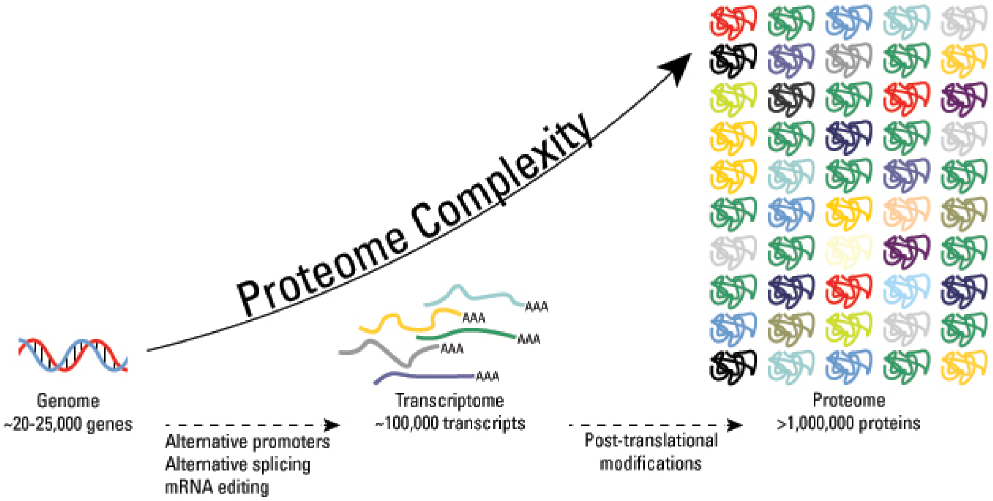


Figure 5.1: Post-translational modifications dramatically increase the diversity of proteome. While there are by estimation 19,000-20,000 human protein-coding genes [70], human proteome is estimated to encompass over 1 million proteins due to alternative splicing, single amino acid polymorphisms, and PTMs [197]. Figure source: Thermo Fisher Scientific.

The complexity does not end here. Besides translational regulations, various post-translational modifications (PTMs) are critical to protein functions. One of the most well-studied modification is phosphorylation. A phosphate is added to a particular amino acid mediated by kinases. It causes a protein to interact or bind other proteins that recognize the phosphorylated domain. This is necessary for many signal transduction pathways. Other PTMs includes ubiquitination, acetylation, etc. These PTMs can happen to some proteins in time-dependent combinations. Proteins are subject to degradation with different half-life times, by complex and temporally controlled mechanisms such as ubiquitin-mediated proteolysis [79, 82]. Figure 5.1 illustrates the complexity of proteomics analysis that is scaled up by PTMs.

Proteome therefore carries rich biological information that is not accessible by genomics or transcriptomics. Meanwhile, it is also more complex due to its molecular complexity and dynamic nature. Various techniques have been developed to measure proteome. Polyacrylamide gel electrophoresis (PAGE) technique resolves proteins by molecular mass. Isoelectric focusing (IEF) technique resolves proteins by isoelectric points. 2D PAGE separates proteins by both of these two properties in two dimensions [184]. Cell imaging by light and electron microscopy has been used to mark proteins, e.g., imaging specific proteins in live cells after binding fluorescently labeled antibodies to these proteins [248]. Protein microarray is developed based on DNA microarray to measure protein interactions and activities [94]. However, due to the complexity of cellular proteome and the low abundance of many proteins, these technologies are not sensitive enough.

5.1.2 MS and MS/MS Technologies for Protein Characterization

MS has been developed to offer sensitive protein characterization. Proteins or peptides are ionized and their mass-to-charge ratios (m/z) are measured by mass analyzer. The number of ions at each m/z value is detected by ion detector. The m/z values are compared with protein database to identify the peptides and proteins. Nowadays MS has becoming a powerful and the most commonly used techniques for protein quantification [162]. The performance of mass spectrometers in terms of sensitivity, speed, mass accuracy, and resolution has been improved rapidly [13]. MS-based proteomics can deliver three different types of experimental results [45]. One is to quantify the amount of proteins in a sample, which is analogous to transcriptomics. Another usage is for analyzing the presence and sites of PTM for >200 PTMs [165] although some low molecular weight PTMs are hard to be detected robustly [186]. Thirdly, MS is used for studying protein interactions [17]. Here we focus on the first experimental type - proteome profiling. The widely employed technique is the coupling of liquid chromatography (LC) with high-resolution tandem mass spectrometry (MS/MS) to identify and quantify peptides [45]. We are going to explain the basic steps of a LC-MS/MS experiment below. An illustration is provided in Figure 5.2.

1. Sample fractionation. Proteins are separated from samples by biochemical fractionation or affinity selection. Gel electrophoresis can be used to define the sub-proteome to be analyzed.

2. Trypsin digestion. Proteins are digested to peptides because MS of whole proteins is less sensitive than peptide MS and the mass of the intact protein by itself is insufficient for identification.

3. Peptide chromatography coupled with ionization. High-pressure liquid chromatography (HPLC) is used to separate the peptides in very fine capillaries and elute the peptides into an electrospray ion source (ESI) for ionization.

4. m/z determination. The ionized peptides enter the mass spectrometer and their m/z values are measured at this time point (MS1). In the mass analyzer, peptide ions are exposed to electrostatic and/or magnetic fields and their motions in such field can be written as a function of their m/z values. Different mass analyzers have been devised based on different technical ways to measure this ratio. For example, there are time-of-flight (TOF), quadrupole, ion traps, Orbitraps, and Fourier transform ion cyclotron resonance (FT-ICR) mass analyzers. Ion trap mass analyzer is often chosen for LC-MS with ESI ionization. The peaks for MS1 can be selected for sequencing. Inside the mass spectrometer peptides of particular m/z value are selected and fragmented using techniques such as collision induced dissociation (CID). The m/z values of the fragments are recorded in tandem mass spectrum (MS2).

5. Peptide mass fingerprinting (PMF). The peptide m/z values in MS1 and MS2 are searched against a database of known protein sequences. These proteins are digested *in silico* and the masses of produced peptides are computed. The searching process finds matches between the observed peptide masses and the ones in the database. While it is likely to find multiple peptides for a single m/z value, it becomes more reliable to find the peptide matches that exist in the same protein [115, 164].

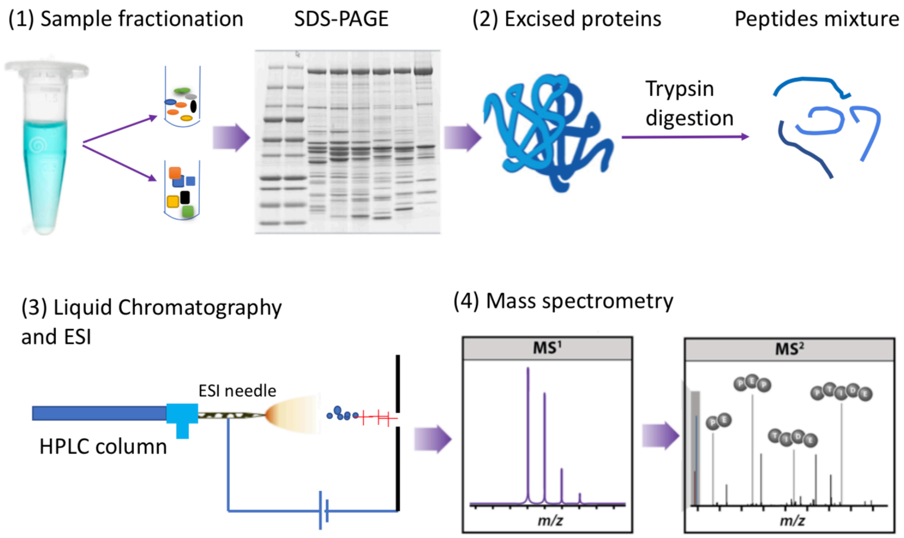


Figure 5.2: An illustration of a typical LC-MS/MS procedure. It combines the physical separation capabilities of HPLC with the mass analysis capabilities of MS.

5.1.3 The Potential of Proteomics in Biomarker Discovery

Numerous studies have demonstrated that molecular aberrations at different molecular levels can induce cancer [103, 290]. Let us take the example of the tumor suppressor gene TP53. Its somatic mutations [185] and PTMs [22] both play important roles in tumorigenesis. One can thus only partially capture its influence in cancer using genomic data alone. This simple example is in line with the key question in cancer research, which is to understand how the information flow from genome to proteome is altered in tumors [7,156]. Therefore, integrating proteomics with transcriptomics and genomics becomes indispensable to unravel this information flow. It has the potential to more accurately identify the disease driver molecules including their PTM status, which is necessary for biomarker translation [221]. However, compared with other types of genomic data, proteomic data have been rarely used to model, predict phenotypes, or discover biomarkers. There could be a few reasons or difficulties. First, the use of microarrays to measure mRNA expression and genomic profiling using NGS technologies are more ubiquitous than proteomic technologies [7]. Second, there are difficulties in mapping peptides to proteins and identifying coding genes of the proteins [273]. Third, data are generated and processed using different platforms with varying performance. This causes difficulties in the alignment of multiple datasets after data processing and feature detection [233]. Last but not least, the challenge of 'big data': it is nontrivial to effectively store and manage large amount of proteomics data using traditional databases. [273] developed the ProteomicsDB in-memory database for storing and analyzing MS data. Using a random-access memory (RAM) of 2TB, the data can be stored in memory all the time for processing. However, such computational infrastructure may not be feasible to build for every application.

TCGA has used reverse phase protein arrays to measure abundances of cancer-related proteins [151]. However, only a small fraction of the proteome was measured. This cannot provide a global profile, as that provided by transcriptomes such as RNA-Seq data. This could be one reason that biomarker discovery studies that use TCGA data usually did not consider protein expression profile. For this reason and also due to the limited number of samples, we could not conveniently include protein features in our experiments in the previous chapters, e.g., to map protein data on the EMT networks. In contrast, MS can provide good proteome coverage to support integrative studies. With MS-based proteomics, researchers started to compare transcriptomes with proteomes and investigated their relationships [177, 262]. A new area of research *Proteogenomics* has emerged. Several proteogenomic applications have been launched to study alternative splicing, discover novel protein coding regions, etc [180].

5.2 The Big Data Challenge

As mentioned above, a few challenges remain before a seamless integration of proteomics data with genomics and transcriptomics data. Here we address the last challenge - the big data challenge. Our scope here is to find the suitable data infrastructure to efficiently store and query MS data.

MS data are eligible for the name ”big data” which is characterized by the large volume, velocity and variety [141]. Traditional relational databases become unsuitable for the big data challenge in bioinformatics as they follow rigid table schema and lack scalability for data aggregation. NoSQL databases emerge in recent years to provide alternative, flexible and more scalable data stores. For example, there are key-value databases such as DynamoDB, column-oriented databases such as HBase and Cassandra, document-based databases such as CouchDB and MongoDB, and graph databases such as Allegro Graph and Neo4j [238]. Relational databases guarantee ACID properties (atomicity, consistency, isolation and durability) while NoSQL databases guarantee BASE properties (basically available, soft state, and eventual consistency) instead [178]. NoSQL databases are characterized with the ability to store large volume of data and support flexible data models. The storage of data is not restricted by fixed table schema as relational databases. Therefore, NoSQL databases can have an advantage in the applications where ACID is not essential, but scalability and flexibility are more important [12]. Biological data are commonly stored as flat files or in relational databases. Once the data are stored, most of the operations on the data are queries and not to modify the data. For example, many MS data are produced and stored. Later they need to be queried over and over again to be analyzed for biomarker identification and protein identification. Therefore, an ideal database should have low latencies in storing and querying data, while maintaining the consistency. Since relational model is not necessary for MS data, it is therefore interesting to investigate whether NoSQL techniques can provide benefits.

There have been a number of qualitative or conceptual studies to compare relational and NoSQL databases [178, 238, 239]. Databases are compared in terms of their data models, query models, consistency models, scalability, and maturity, etc. They show that each data store has its own data structures. NoSQL data stores are designed to manage large volumes of unstructured data. However, relational databases cannot be replaced because of their unique features such as the support of transactions, reliability (ACID), and the maturity of technology. There have also been a few quantitative studies to compare different data stores. They deployed certain data stores in concrete problems. [147] performed experiments to store and query clinical data with XML database. The results showed that XML database can store more flexible clinical data, but it has higher query latencies than NoSQL databases. [259] compared the usefulness of MySQL and graph database Neo4J on storing and querying graph data. The results show that Neo4J is much faster than MySQL for traversal queries because Neo4J has a built-in traversal framework. As Neo4J uses Lucene for query which treats data as text, MySQL has better query speed than Neo4J on integer databases. However, when the data are stored as text, MySQL and Neo4J have comparable performance and Neo4J has better scalability. [99] used Neo4J graph database to store STRING human protein interactions data. They pointed out that depending on the types of queries, graph database may not be the best choice for graph data.

Note that the suitability of data stores, especially NoSQL data stores, not only depend on their own features, but also depend on the individual problems and use cases. Therefore, we are going to conduct quantitative studies to compare the performance of different databases on proteomics mass spectrometry data. Since MS data do not have graph data structure, graph databases are not included in the study. We compared the latencies of one relational database (MySQL), three NoSQL databases and the flat file system on storing and querying MS and MS/MS data, as well as the disk or memory usage. The four data stores are the representatives from four main database categories. In addition, both in memory and disk-based configurations are considered, as listed below:

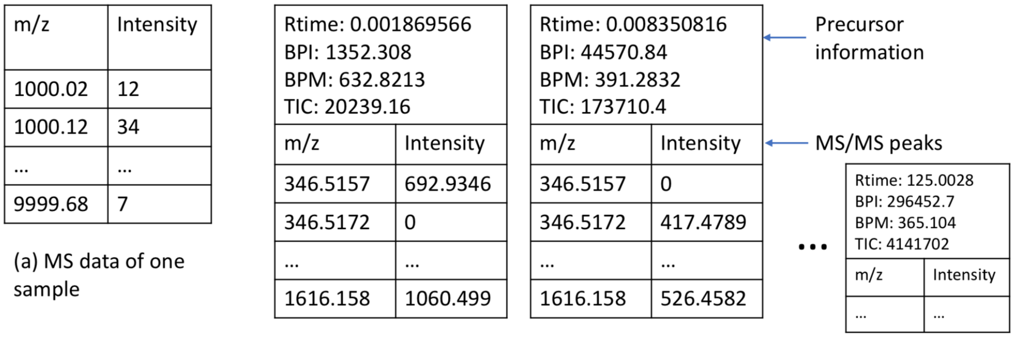
- Relational database (MySQL, both disk-based and in memory configurations)
- Document-oriented database (MongoDB, disk-based)
- Column-oriented database (HBase, disk-based)
- Key-value database (Redis, in memory)

5.3 Methods

5.3.1 Data for Testing the Databases

MS data

We used both MS and MS/MS data for testing these databases. The MS data are from [42], where they analyzed the proteome of blood samples for testicular germ cell cancer and thyroid disease detection. The MS data are produced using MALDI-TOF techniques. Matrix-assisted laser desorption/ionization (MALDI) is an ionization technique that uses a laser energy absorbing matrix to create ions from large molecules with minimal fragmentation. Time-of-flight (TOF) is a type of mass analyzer that measures the m/z values based on the time it takes for the ions to reach the detector. The data have been processed and are in the form of text files in the dat format. The dat format presents the data in a simple table with two columns for m/z values and intensity values. Each text file has 42,381 pairs of m/z and intensity values. The m/z values for each sample range between 1000 and 10,000 Da. The m/z values among different samples are aligned by finding the *masterpeaks* as proposed in [42]. The average size of one MS sample is about 440KB. Figure 5.3 provides an illustration of one MS sample.



m/z	Intensity
1000.02	12
1000.12	34
…	…
9999.68	7

(a) MS data of one sample

Rtime: 0.001869566
BPI: 1352.308
BPM: 632.8213
TIC: 20239.16

m/z	Intensity
346.5157	692.9346
346.5172	0
…	…
1616.158	1060.499

Rtime: 0.008350816
BPI: 44570.84
BPM: 391.2832
TIC: 173710.4

m/z	Intensity
346.5157	0
346.5172	417.4789
…	…
1616.158	526.4582

Rtime: 125.0028
BPI: 296452.7
BPM: 365.104
TIC: 4141702

m/z	Intensity
…	…

Precursor information
MS/MS peaks
…

(b) MS/MS data of one sample

Figure 5.3: An illustration of the data structures of a MS sample and a MS/MS sample.

MS/MS data

The MS/MS data were taken from the study [71], where they analyzed the proteome of paired malignant and non-malignant tissue samples from 38 early stage LUAD patients using LC-MS/MS approach. LC separation was performed using a Waters Nano Acquity UHPLC. Mass spectra were collected on an Orbitrap Q Exactive Plus mass spectrometer (Thermo Fisher Scientific). It separates precursor ions ”in time” using ion traps and performs MS2 analysis on the separated ion populations in high resolution. The precursor isolation is achieved by applying a quadrupole to the precursor ions which allows only a certain ion population (within the chosen mass window) to remain stable in the trap. The m/z values are measured by Orbitrap ion trap mass analyzer. We converted the raw files to MS1 format using ProteoWizard msConvert [126]. Different MS2 spectra are marked with different retention time (elution time from LC), base peak intensity (BPI), base peak mass (BPM), and TIC (total ion current). Each MS2 spectrum contains 283 pairs of m/z and intensity values. The m/z values range from 346.5157 to 1616.158 Da. The average size of one MS/MS sample is about 2GB. Figure 5.3 provides an illustration of the data for one MS/MS sample.

5.3.2 Databases and Data Models

Each database has its unique features. When we store MS data in them the data models are different. In some cases, there are alternative ways to store the same data. We decide to choose the data model based on the distinguishing feature of a database while taking into account its constraints. For example, MongoDB is document-oriented, so we store each sample file in one document. HBase is column-oriented thus we store each sample file in one column. We deployed the standalone mode of all databases. In the following we use an example of student data including basic student information to show the schemas of different databases. Then we explain how MS and MS/MS data are stored in these databases.

MySQL

MySQL is the most widely used open-source relational database management system (RDBMS). It uses tables to store data. A table has rows for records and columns for the fields. When the table is created each field is assigned a data type such as *CHAR(30)* (can hold up to 30 characters), *DATE*, *FLOAT*, etc. MySQL supports a number of data types: numeric types, date and time, character and byte types, and spatial types. Queries in tables are done using SELECT statement. It can be combined with *Where*, *AND & OR*, *Order By*, and *Group By* clauses to specify query conditions and the output aggregates. SQL tables can be linked in queries using for example *JOIN*, and *UNION* clauses. *FOREIGN KEY* constraints can be added to the columns of a table that point to the columns of other tables, which makes sure that only valid data can be inserted into the foreign key columns. An SQL table needs to be created with a rigid schema - at least the data types and the primary key, before data can be inserted. The inserted data cannot violate the data types and constraints.

MySQL supports several storage engines, which are software modules that a database management system can employ to handle CRUD (create, read, update and delete) operations. *InnoDB* is the default engine and is most widely used. *Memory* storage engine is a fast one because it creates tables in memory. It does not support transactions. Since choosing the right storage engine is important for database performance, we included both *InnoDB* and *Memory* engines for testing.

The student data can be stored in a relational table as shown in Figure 5.4a. To store MS samples, we create a table with three columns: sample number, m/z value and intensity value, as shown in Figure 5.4b. An index is built on the sample number column to accelerate the search for multiple samples. As a table can contain a maximum of 1017 columns we could not store the data row-wise. As the *VARCHAR* data type can hold maximum 65,535 bytes, it is not convenient either to store a sample as a string. Thus, to store MS/MS data, we need two tables for one sample - one table for storing the spectra and one table for storing the metadata, including the retention time, BPI, BPM, and TIC. Since the two tables are related, the common schema design is to add a foreign key constraint to ensure data integrity. We built a foreign key on the spectrum ID column of the spectra table, which points to the primary key of the metadata table. The table schemas are shown in Figure 5.4c. To accelerate inserting data, we use the bulk load operation to insert one MS sample, or one MS/MS spectrum at a time.

MongoDB

MongoDB is a document-oriented NoSQL database. It does not have fixed table structures. A record in MongoDB is a document, which is a data structure composed of field and value pairs. Every document has an *_id* field as a primary key. MongoDB stores documents in collections, and collections in databases. Collections can be considered as an analog to tables in relational databases. Documents are stored as BSON (binary representation of JSON document) objects. The values can include strings, arrays, and other documents. This offers flexibility in data storage. For example, to store embedded data structure in relational database, one has to break the data into multiple cross-referenced tables. While in MongoDB, it can be stored in a single document. Compared to relational tables that

ID (primary key)	name (char (30))	age (smallint)	subject (char(20))	year (tinyint)
1	Ana	24	math	3
2	Bob	34	art	6
3	Tom	25	medicine	5

(a) Student table

Sample ID (smallint)	m/z value (float)	intensity (smallint)
1	1000.02	29
1	1000.12	21
...	...	...
1	9999.68	5
2	1000.02	26
...	...	...
n	9999.68	7

(b) MS table

Spectrum ID (smallint) foreign key	m/z value (float)	intensity (float)
1	346.5157	692.9346
1	346.5172	0
...	...	...
1	1616.158	1060.499
2	346.5157	0
...	...	...
m	1616.158	1181.356

Spectrum ID (smallint) primary key	RTime (float)	BPI (float)	BPM (float)	TIC (float)
1	0.001869566	1352.308	632.8213	20239.16
2	0.008350816	44570.84	391.2832	173710.4
...	...	...	...	...
m	125.0028	296452.7	365.104	4141702

(c) MS/MS tables

Figure 5.4: MySQL data models. (a) student data are stored in an SQL table where each field corresponds to one attribute and is assigned a data type. The student ID is the primary key. (b) MS data are stored in one SQL table in three columns. (c) One MS/MS sample is stored in two tables, one for spectra and one for metadata.

sample	document (field-value pairs)
1	name:Ana, age:24, subject:math, year:3
2	name:Bob, age:34, subject:art, year:6
3	name:Tom, age:25, subject: medicine, year:5

(a) Student collection

sample	document (field-value pairs)
1	1000_02:29, 1000_12:21, ..., 9999_68:5
2	1000_02:26, 1000_33:8, ..., 9999_68:30
...	...
n	1000_02:4, 1000_12:38, ..., 9999_68:7

(b) MS data collection

_id	document (metadata pairs and m/z - intensity value pairs)
...	sample:1, rtime:0.001869566, bpi:1352.308, bpm:632.8213, tic:20239.16, 346_5157:692.9346, ... 1616_158:1181.356
...	sample:1, rtime:0.008350816, bpi:44570.84, bpm: 391.2832, tic:173710.4, 346_5157:0, ...
...	...
...	sample:m, rtime:125.0028, bpi:296452.7,bpm:365.104, tic:4141702, ... 1616_158, 1181.356

(c) MS/MS data collection

Figure 5.5: MongoDB data models. (a) student data are stored in a collection with the student ID as the primary key. (b) MS data are stored in a collection with sample ID as the primary key. (c) MS/MS data are stored in a collection with automatic ID and an index on the retention time field.

have to be defined upon creation, collections do not need to be defined. Documents with different structures can be stored in the same collection.

In addition to being able to store data flexibly, complex queries and aggregations can be performed on the data. In queries, the *dot notation* is used to access the elements of an array or the fields of embedded documents. Query selectors are used to specify the conditions on a field. The selectors cover different categories such as comparison, logical operators, bitwise operators, etc. Conditions can be specified on multiple fields to have a compound query. In terms of aggregations, MongoDB is able to process documents and return computed results. It can group values from multiple documents and perform a variety of operations such as map-reduce function, single purpose aggregation operations (e.g., *count()*, *distinct()*), etc.

The student data can be stored in a collection as shown in Figure 5.5a. We store the MS sample files in one collection with one document for one sample. The key of the document is the sample number. Within each document, the field-value pairs are the m/z value-peak value pairs in the sample files. Since the field value of a field-value pair in MongoDB is a string and cannot contain dot, we could not store the m/z values as float values. We use strings with dots replaced by underscores, as shown in Figure 5.5b. For MS/MS data, we stored each spectrum as a document with the first a few fields storing sampleId, retention time, BPI, BPM, and TIC. Different from storing MS data, we used the automatically generated _id field. To accelerate spectrum query, we built an index on the retention time field. The data schema is shown in Figure 5.5c.

HBase

HBase is a column-based NoSQL database modeled after Google's Bigtable. It stores data in wide tables. A table in HBase does not have fixed column schema. Only the column families (CF) need to be declared at the schema definition time. A table is a collection of rows. A row is a collection of CFs. A column family is a collection of columns. Each column is identified by a collection of column qualifier-value pairs. A row key, a CF, a

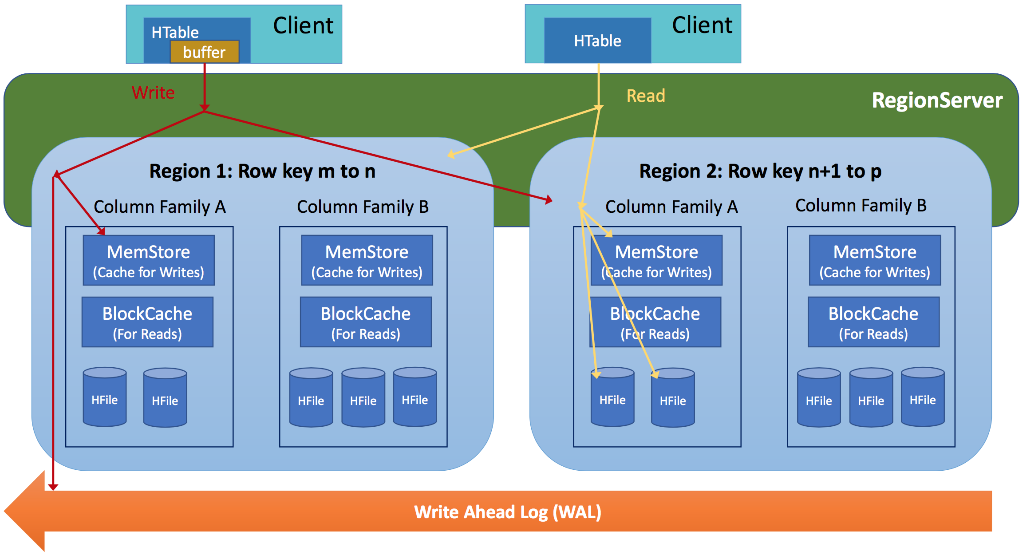


Figure 5.6: An illustration of Memstore usage in HBase read/write paths.

column qualifier and a timestamp can exactly specify a cell in a HBase table. The names of CFs and column qualifiers are copied on each row. The student data can be stored as in Figure 5.7a.

It is suggested in the HBase manual that the number of column families should be kept low (not more than two or three) to have a good performance. In many cases having one column family is the best schema design. This is due to the write and read paths in HBase. Figure 5.6 shows how HBase handles write requests. The RegionServer (RS) serves data for read and write. It directs the requests to a specific region. HBase tables are divided horizontally by the row key range into *Regions*. Each region stores a set of rows. The rows are separated in multiple column families. There is one MemStore for each column family. The main reason for using Memstore is to sort the data by the row key before writing it to the distributed file systems.

When we write data to HBase tables, the data are firstly written into MemStore. When the size of data exceeds certain thresholds the data in MemStore are flushed into HFiles. The Memstore flush creates one HFile per CF and all CFs are flushed together. This means that when there are multiple CFs and the data from one CF exceed the MemStore threshold, the other CFs in this region are flushed as well (in the default setting). Thus N (the number of CFs) HFiles are created per flush which causes needless i/o operations. In addition, to handle the many HFiles created by frequent flushes, HBase periodically compacts multiple small HFiles into a big one. Thus, having multiple CFs also increases the compaction cost. In our experiment, to avoid unnecessary i/o operations, we will define one CF. HBase is deployed in standalone mode. It uses local file system instead of hadoop distributed file system (HDFS).

In contrast to the design of CFs, HBase does not have a limit regarding the number of columns in a CF. All mutations on a row in HBase are atomic - it either completes entirely or not at all. This applies when mutations occur to multiple CFs of a row. When a row is updated, it is locked by the RS until the update is finished. HBase can scale

Row key	personal data [column family 1]		study [column family 2]	
ID	[column qualifier: value] pairs		[column qualifier: value] pairs	
1	name: Ana	age:24	subject: math	year:3
2	name: Bob	age:34	subject: art	year:6
3	name: Tom	age:25	subject: medicine	year:5

(a) Student table

Row key	MS [column family]			
Sample ID	m/z : intensity [column qualifier: value]	m/z: intensity	...	m/z: intensity
1	1000.02:29	1000.12:21	...	9999.68:5
2	1000.02:26	1000.33:8	...	9999.68:30
...				
n	1000.02:4	1000.12:38	...	9999.68:7

(b) MS table

Row key	MS/MS [column family]				
sampleID_RTime	BPI: value [column qualifier: value]	BPM: value	TIC: value	m/z: intensity	...
1_0.001869566	BPI:1352.308	BPM:632.8213	TIC:20239.16	346.5157:692.9346	...
1_0.008350816	BPI:44570.84	BPM: 391.2832	TIC: 173710.4	346.5157:0	...
...					
m_125.0028	BPI:296452.7	BPM: 365.104	TIC: 4141702	...	...

(c) MS/MS table

Figure 5.7: HBase data models. (a) student data are stored in two CFs based on data categories with student ID as the row key. (b) MS data are stored in one CF with the sample ID as the row key. (c) MS/MS data are stored with sampleID_RTime as the row key. The spectra and other metadata are stored in one CF.

to handle very large tables with billions of rows and millions of columns. To store MS data, we define the sample number as the row key and the m/z value-intensity value pairs as the column qualifier-value pairs, as shown in Figure 5.7b. To store MS/MS data, we take advantage of the HBase row keys, which are sorted to provide fast random find and contiguous scanning of rows, e.g., prefix-based row key scans. We store the sample ID and retention time information in the row keys and the other meta-data in columns. The table schema is illustrated in Figure 5.7c.

Redis

Redis is an in-memory NoSQL database. It can be used as a database, cache, or message broker. By using in-memory configurations, it can provide high read and write speed. Depending on the use case, the dataset in memory can be dumped to disk manually or at certain intervals. Data are stored in the form of key-value pairs. Keys are associated with

key	hash
1	(name,Ana) (age,24) (subject,math) (year,3)
2	(name,Bob) (age,34) (subject,art) (year,6)
3	(name,Tom) (age,25) (subject, medicine) (year,5)

(a) Student hash

Key	hash
1	(1000.02,29) (1000.12,21) ... (9999.68,5)
2	(1000.02,26) (1000.33,8) ... (9999.68,30)
...	...
n	(1000.02,4) (1000.12,38) ... (9999.68,7)

(b) MS data stored in hash

Key	string
1	1000.02 2 1000.12 21 ... 9999.68 5
2	1000.02 26 1000.33 8 ... 9999.68 30
...	...
n	1000.02 4 1000.12 38 ... 9999.68 7

(c) MS data stored in string

Key (rtime_sample)	Hash
0.001869566_1	(bpi,1352.308) (bpm,632.8213) (tic,20239.16) (346.5157, 692.9346) ... (1616.158,1181.356)
0.008350816_1	(bpi,44570.84) (bpm,391.2832) (tic,173710.4) (346.5157, 0) ...
...	...
125.0028_m	(bpi,296452.7) (bpm,365.104) (tic, 4141702) ... (1616.158,1181.356)

(d) MS/MS data stored in hash

Key (rtime_sample)	string
0.001869566_1	bpi 1352.308 bpm 632.8213 tic 20239.16 346.5157 692.9346 ... 1616.158 1181.356
0.008350816_1	bpi 44570.84 bpm 391.2832 tic 173710.4 346.5157 0, ...
...	...
125.0028_m	bpi 296452.7 bpm 365.104 tic 4141702 ... 1616.158 1181.356

(e) MS/MS data stored in string

Figure 5.8: Redis data models. (a) student data are stored using hash with student ID as the key. (b) MS data are stored using hash with the sample ID as the key. (c) MS data are stored using string with sample ID as the key. (d) MS/MS data are stored using hash with retention time and smaple ID as the key. (e) MS/MS data are stored using string with retention time and sample ID as the key.

string values. Keys are binary safe, which allows any binary sequence to be used as a key, ranging from a string to an image file. The values allow different data structure such as strings, hashes, lists, sets, sorted sets, bitmaps, etc. The student table can be stored using hashes (field-value pairs) as shown in Figure 5.8a.

We found that both hash and string data structures can be adapted to store MS data. Thus, we used both of them and compared their performance. With hash data structure we use field-value pairs to store the m/z value-intensity value pairs, as shown in Figure 5.8b. With string data structure, we append all lines of the MS sample file to a string, as shown in Figure 5.8c. The key of a hash or a string is the sample ID. List and sorted set are not suitable in our use case. List is implemented as linked list in Redis. This data structure is efficient for inserting data but requires sequential scanning to retrieve data. Sorted set can only keep unique intensity values, which does not apply to the data.

5.3.3 Experiments

We performed experiments to compare the latencies of MySQL, MongoDB, HBase, Redis, and flat file system on storing and querying MS and MS/MS data, as well as the disk or memory usage. Java clients and JDBC (Java database connectivity) API (application programming interface) were used to uniformly access the databases and local file systems. To evaluate the suitability of these databases for building a proteomics data infrastructure, we selected a few commonly needed operations when working with MS and MS/MS data to serve as proxy applications. We used a powerful local workstation for the experiment. The workstation is equipped with 12 Intel Xeon (R) CPUs running at 3.50GHz, 64GB RAM and 2 TB SATA hard-disk drive (Model: MegaRAID SAS 9341-8i). The operating system is Ubuntu 16.04.

We have 3 use cases for MS data:

1. Store new data: insert n MS samples. n ranges from 500 to 150,000. The total size of files on disk ranges from 200MB to 64.3GB (close to the size of the main memory). We choose the sequence of n to be n = (500, 2500, 5000, 10000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 110000, 120000, 130000, 140000, 150000), in total 18 measurements.

2. m/z range query: select all m/z - intensity value pairs from all samples where m/z values are within a certain range. Here we use the range from 1500 to 1800 Da.

3. Sample query: select all spectra of m samples from all samples by sample IDs. We generate the sample IDs randomly and the number of IDs is 10% of the total samples. For example, if we stored n samples in a database, $n/10$ sample IDs are generated randomly for the query.

We have 3 use cases for MS/MS data:

1. Store new data: insert n MS/MS samples. n ranges from 1 to 32. The total size of files on disk ranges from 2GB to 64GB. We choose the sequence of n to be n = (1, 2, 3, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32), in total 18 measurements.

2. Retention time query (RTime query): select all spectra from each sample where the retention time falls into a certain range. Here we choose the time window to be between 20 and 22 seconds.

3. RTime and m/z range query: select all m/z-intensity value pairs from each sample where the retention time falls into a certain range and the m/z values are within a certain range. Here we choose the retention time window to be between 20 and 25 seconds, and the m/z value range to be between 400 and 500 Da.

5.4 Results and Analysis

5.4.1 Storing and Querying MS Data

Figure 5.9 shows the latencies of storing MS data and the consumed space on disk or in memory. We observe the following:

- HBase and MySQL are slowest in storing data. It is mainly because they are disked based systems. Comparatively, in memory databases are very fast to store data.

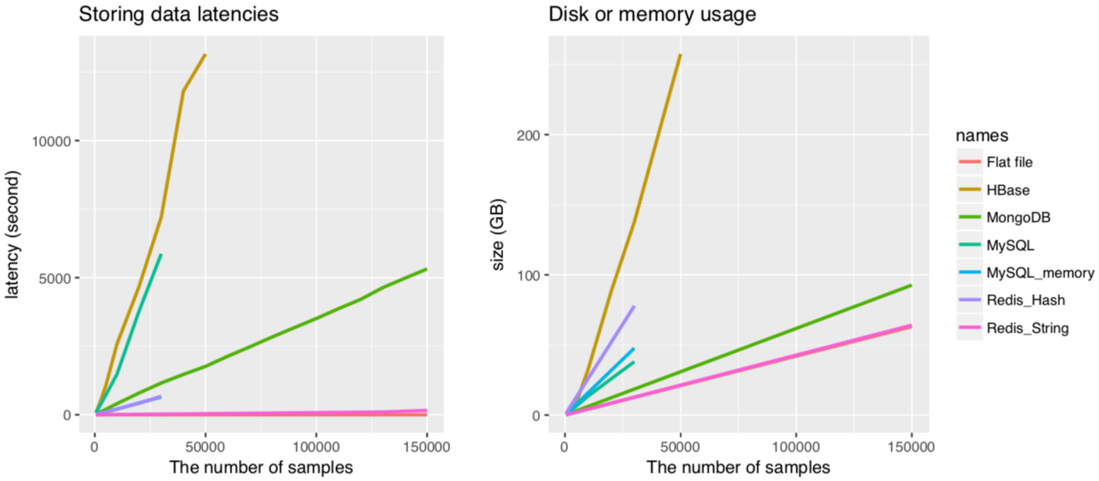


Figure 5.9: The latencies of storing MS data and the disk or memory usage.

MongoDB is also a disk-based database, but it is much faster than MySQL and HBase. We think this is due to the fact that MongoDB uses documents to store samples rather than tables. This not only accelerates the write operations but also saves storage space.

- HBase consumes the most disk space. It is because in each row (which stores one MS sample), the names of the column qualifiers and the CF need to be stored. Within each cell, a timestamp is also stored. This makes HBase more demanding on disk space. MongoDB uses the least space among disk-based databases.

- Redis database gives very different performance when using string data structure and when using hash data structure. Storing data with Hash is much more expensive than with Strings. The latter can store the data in memory using nearly the same space as the size of disk-based files.

Figure 5.10 shows the query latencies of the two use cases: m/z range query and sample query. We observe that in m/z value query, MySQL and HBase are slower than using flat files. MongoDB shows significantly lower latencies compared with flat files. It is even faster than MySQL with MEMORY storage engine. Redis is the fastest. Hash data structure leads to much lower latencies than string data structure because it is faster to filter the m/z value range based on hash keys. When the data in Redis reach the size of memory, the latencies increase dramatically. On sample queries, MySQL is extremely expensive compared with the other databases, which are able to store one sample in one record: a HBase row, a MongoDB document, or a Redis hash or string. Excluding MySQL, we observe that MongoDB has lower latencies than HBase, Redis hash, and MySQL with MEMORY engine.

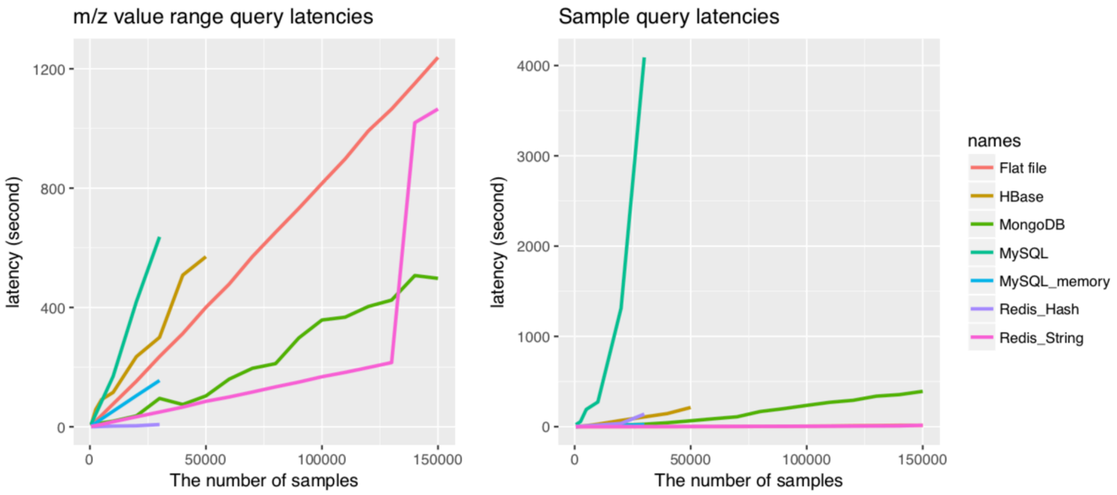


Figure 5.10: The latencies of m/z value range query and sample query on MS data.

5.4.2 Storing and Querying MS/MS Data

Figure 5.11 shows the latencies of storing MS/MS data and the consumed space on disk or in memory. It also gives the query latencies of two use cases: RTime query, RTime and m/z range query. We observe the following:

- Consistent with the observations on MS data, MongoDB achieved top performance among disk-based databases. It has low latencies in storing data. The memory consumption is efficient - on average 1.47 times the size of flat files. It shows much lower query latencies than the majority of other databases, even including MySQL with MEMORY engine and Redis with hash data structure. MongoDB shows good potentials for handling data that are in the structure of key-value pairs.

- MySQL database still shows very high latencies in storing and querying data. In both queries, especially the RTime and m/z range query where MySQL has to perform JOIN operations on tables, the latencies increase rapidly to a different scale compared with the other databases. It shows that querying MS/MS data is not a good application of relational databases.

- Some of the curves follow a linear trend but the shape has some variations, especially for MongoDB. We accredit it to the memory allocation of the operating system when we run consecutive testings, where we delete the current data and store data of the next (larger) size. As it is certain that storing and querying more samples requires longer time, we believe the results would look more regular if the testings on different data sizes were performed separately.

Above all, there are significant performance differences among the databases. Since all storage systems show a nearly linear dependency between the latencies and data sizes, we calculate the average latencies of storing or querying one MS or MS/MS sample, and the average disk or memory consumption of storing one MS or one MS/MS sample. The results are given in Table 5.1 and Table 5.2.

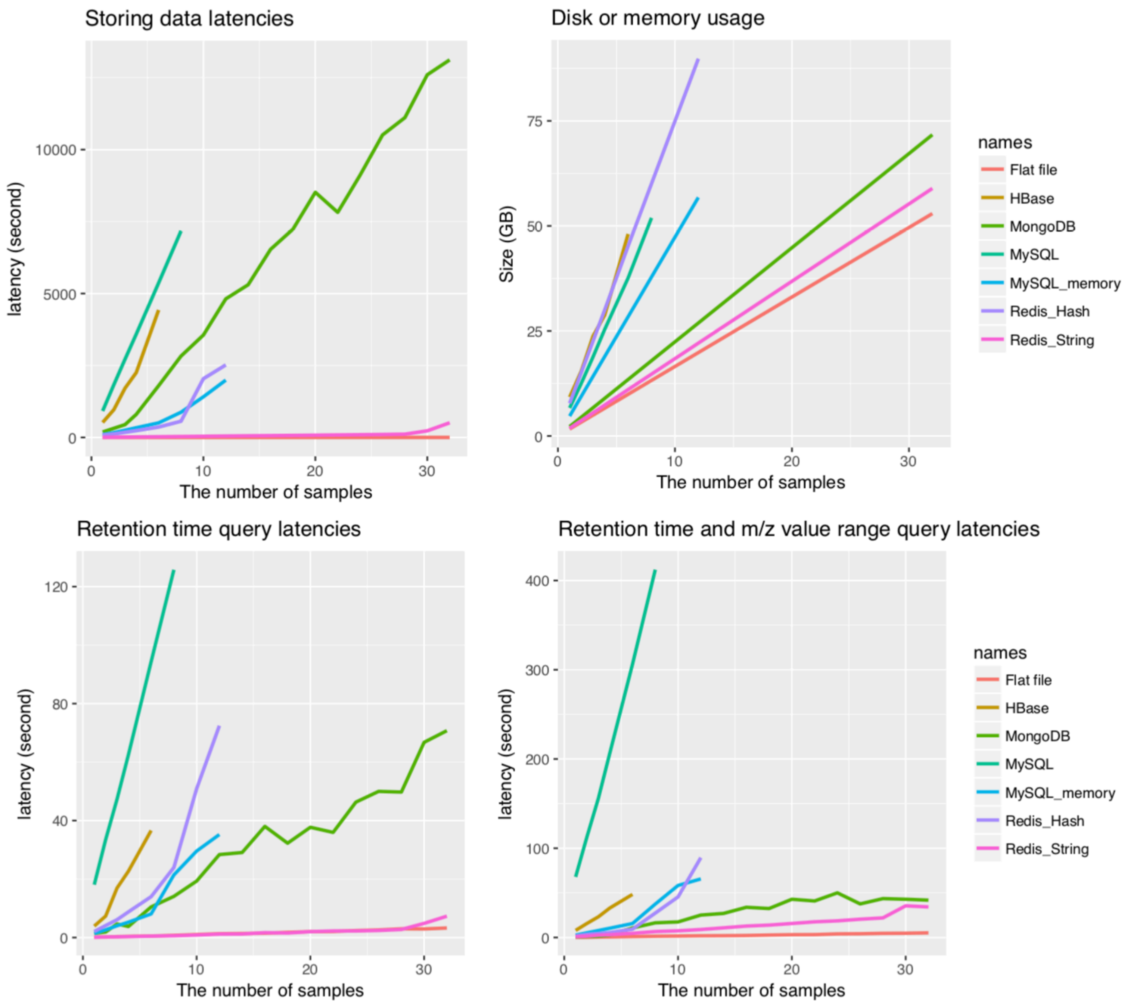


Figure 5.11: The performance of databases on MS/MS data. Upper figures: the latencies of storing MS/MS data and the disk or memory usage. Lower figures: The latencies of RTime query, and RTime and m/z range query.

Table 5.1: The average storing and querying latencies, and the disk or memory consumptions of the databases per MS sample.

Measurement	MongoDB	HBase	MySQL (InnoDB)	MySQL (MEMORY)	Redis (hash)	Redis (string)	Flat file	
Write latency	37.22	235.2	160.4	20.87	21.60	0.84	-	mean
(millisecond)	2.76	37.34	25.01	0.61	0.93	0.16	-	sd
Range query latency	2.82	13.80	18.33	5.19	0.31	1.70	7.97	mean
(millisecond)	0.65	4.50	2.13	0.08	0.19	0.13	0.26	sd
Sample query latency	1.70	3.01	55.80	0.79	2.28	0.054	0.040	mean
(millisecond)	0.69	0.89	42.31	0.27	1.16	0.014	0.039	sd
Size (KB)	647.8	3727	1361	1669	2722	448.2	440.8	mean
	0	1456.52	35.30	0.11	1.94	0.30	0	sd

Table 5.2: The average storing and querying latencies, and the disk or memory consumptions of the databases per MS/MS sample.

Measurement	MongoDB	HBase	MySQL (InnoDB)	MySQL (MEMORY)	Redis (hash)	Redis (string)	Flat file	
Write latency (second)	338.1	574.7	904.7	106.43	101.25	4.91	-	mean
	95.80	98.06	10.45	31.26	65.95	2.88	-	sd
RTime query (second)	1.80	5.00	16.27	1.91	3.10	0.10	0.10	mean
	0.41	1.12	0.98	0.79	1.57	0.036	1.43e-2	sd
RTime and m/z range query (second)	1.54	7.97	55.08	3.63	2.89	0.83	0.151	mean
	0.34	0.20	6.58	1.44	2.17	0.11	2.34e-2	sd
Size (GB)	2.24	8.03	6.39	4.73	7.55	1.84	1.65	mean
	3.10e-5	0.76	0.16	0	0.097	6.49e-4	0	sd

5.5 Discussion

Compared with our previous study [217] where only 7000 MS samples (around 3GB) were used for testing due to the technical limitations, in this study we experimented with 150,000 MS samples (around 63GB) and 32 MS/MS samples (around 53GB). As a result, we are able to compare the performance of the databases more objectively. Having looked at the individual performance above, we would like to address some common ground for all database systems involved.

Better memory utilization results in lower latencies. As expected, the two in memory databases (Redis and MySQL with MEMORY engine) have lower write latencies compared with disk-based databases. On querying data, MongoDB gives comparable performance as in memory databases. This is because MongoDB uses memory-mapped files which first utilizes all available memory before using the hard-disk. This contributes to its good performance in queries.

Accessing data from disk and from memory are intrinsically different [1]. Accessing data from disk is done through the serial ATA interface. It has a theoretical bandwidth of 4,800 MB/s. The theoretical bandwidth of memory is much higher. It is calculated as the product of base DRAM clock frequency, the number of data transfers per clock, memory bus width, and the number of memory interfaces, which is 2133 million clocks per second $\times$ 2 $\times$ 64 bits per line $\times$ 2 = 68.26GB/s. Thus, accessing data from disk is about 27 times slower than accessing data from the main memory. Besides, the latencies of seeking to the correct location on a disk takes about 4 milliseconds. This makes random access on disk much slower than in the memory.

Flat file storage can achieve comparable query latencies. Although querying from flat files involves disk reads, it achieves optimistic query latencies compared with certain databases. It has achieved the lowest latency on sample query of MS data and RTime query of MS/MS data. This shows that databases may not be necessary for some use cases. Nowadays, technical improvements such as operating system dependent page caching and hardware-based caching mechanisms can reduce the latencies of disk read, especially when the reads are sequential. In querying MS and MS/MS data, many operations are sequential because researchers are usually interested in a range of m/z values. Additionally, performing queries in individual sample files avoids the overhead of loading large volume

[1] The following data is based on the configurations of our workstation.

Table 5.3: Conditions in which the databases may be considered

If you have ...	
Unstructured or flexible data that require complex queries	MongoDB
Very large data volume applications	HBase
Data that require a relational model and ACID transactional properties	MySQL
Data that do not require complex queries and can fit in the memory	Redis
Data that only require limited operations	Flat files

of data to the memory, which can cause page faults and disk swaps if the data do not fit in the memory.

Range queries are more expensive. Our experiments show that querying data ranges is usually much more expensive than querying the entire records. This occurs because sequential access is usually faster than random access. Databases often implement range queries as first returning all data fulfilling the lower bound and then filtering on the upper bound. This can increase the query latencies. For example, on MS data range query always has higher latencies than sample query except MySQL (InnoDB) and Redis (hash). For MySQL, it is because the database engine has to query the entire table for sample query. Regarding Redis, its hash data structure is especially designed to retrieve key ranges which makes range query faster. On MS/MS data, RTime and m/z range query always requires longer time than RTime query except Redis (hash) and MongoDB. This is because in MongoDB we built an index on the RTime field so that the time to find the correct documents is significantly reduced. Within each document it only needs to pick the sequential key-value pairs within the required m/z value range. This can explain the results observed with MongoDB.

The trade-off between ACID compliance and other desired properties. MySQL and HBase have higher write and query latencies than other databases. At the same time, they are the most reliable because they guarantee higher level of data consistency. MySQL provides ACID properties. HBase can provide ACID properties within the same row. These inevitably require more disk writing. In comparison, MongoDB does not guarantee ACID properties. It trades off ACID compliance for higher availability which contributes to better speed. NoSQL databases relax the ACID compliance for other desired properties such as availability, horizontal scalability, etc [179, 213]. Our experimental results show that this is necessary to efficiently manage MS and MS/MS data.

The suitability of a database depends on the use case. We have shown that the combinations of databases and data models give different performance, depending on individual use cases. Nevertheless, we would like to give our general use experience in Table 5.3. Overall, we would recommend MongoDB as the default choice when storing data in the structure of key-value pairs.

5.6 Conclusion

As introduced in the beginning of this chapter, system-level investigations of cellular and molecular interactions produce large amounts of data. MS and MS/MS proteomics data are among these data types. Efficient database systems can assist data management and analysis. We have used an experimental approach to compare the performance of a relational database (MySQL) and three NoSQL databases (MongoDB, HBase, and Redis) on their latencies of storing and querying MS and MS/MS data with representative use cases. We also performed the same queries on a flat file system for comparison. To the best of our knowledge, this study was the first quantitative comparison among relational database, NoSQL databases, and flat file system in storing and querying both MS and MS/MS data, which can provide reference for researchers who would like to build a bioinformatics data infrastructure.

Our results show that NoSQL databases with suitable data models can achieve lower write and query latencies as well as less disk or memory consumption than relational databases. Overall, MongoDB achieved good performance compared with other disk-based databases. Depending on the use cases, flat file system can achieve comparable query performance as with using databases. Last but not least, the suitability of databases and data models need to be considered based on the application requirements. In the future, we would like to extend our study by comparing the performance of the databases in distributed mode, e.g., HBase with Hadoop and HDFS, MongoDB with sharding technique, to provide references for bioinformatics data centers.

Chapter 6

Summary

Given the high dimensionality of omics data, feature selection has become a prerequisite for building predictive models. In this thesis, we have covered comprehensively, from the understanding state-of-the-art feature selection methods, to the proposal and evaluation of phenotype relevant network-based feature selection (PRNFS) framework, and to the benchmark of large data applications. We have achieved data integration in cancer prognosis prediction based on epithelial mesenchymal transition (EMT) gene regulations, which have been demonstrated as highly relevant to the metastasis and prognosis of epithelial cancers. The results have shown the good predictive performance of this approach. In the following we will summarize our contributions, identify the limitations and propose to build an intelligent decision support system (IDSS) to support a wide variety of predictive tasks in personalized medicine.

6.1 Contributions

Our contributions are manifold. The major ones are summarized below:

1. A comprehensive literature review. We have reviewed the development of feature selection algorithms in cancer prognosis prediction, from using one type of omics data without network, to the integration of single omics data with network, and to the integrative analysis of multiple omics data. In each of the three categories we have discussed the advantages and disadvantages of several underlying methodologies. Based on our review and several other studies where feature selection algorithms are objectively evaluated, we identified the research gaps.

2. The proposal of RRNFS framework. Motivated by the research gaps, we proposed the novel RRNFS framework to selected robust features from omics data for phenotype predictions. We demonstrated the benefits of this approach with the application of prognosis prediction in lung adenocarcinoma, where we constructed EMT networks and selected molecular signatures from them. We have shown that even the dimensionality was reduced to less than 2.5 % of the original data, remarkable prediction performance was obtained.

3. The identification of EMT single-omics signatures. We mapped multiple types of omics data alternatively on EMT networks to identify molecular signatures and com-

pared their predictive performance. Further, we analyzed the frequently selected features (FSFs) from different data levels. We found out that the network properties of FSFs from individual data levels are significantly different. We derived prognostic association rules from the FSFs of different omics data and analyzed their biological interpretations. We showed that combining FSFs from different omics data gave rules of higher qualities.

4. The identification of EMT multi-omics signatures. We obtained multi-omics signatures first by combining single-omics signatures. We showed that combined features can separate all-stage samples and early-stage samples into more significantly different groups. Then we proposed an integrative feature selection approach using multiplex network. We can directly select multi-omics signatures with this multilayered network structure, where each layer was mapped with one type of omics data. We showed that the FSFs from multiplex network gave more optimistic prediction performance than any single-omics FSFs. To our knowledge, we for the first time employed multiplex for feature selection in cancer prognosis prediction.

5. The evaluation of EMT signatures on independent patient cohorts. To test the utilities of EMT-based feature selection, we tested both single-omics and multi-omics EMT signatures on a real-world clinical dataset consisting of both gene expression and DNA methylation data for a cohort of patients. Employing EMT signatures, we were able to stratify the samples into significantly different prognostic groups. Further, multi-omics signatures gave superior performance over single-omics signatures.

6. The benchmark of database systems for handling large biological data. An integrative omics data analysis requires efficient database systems. We chose proteomics data as an example due to its complexity and large volume. Both SQL and NoSQL databases were tested in terms of their latencies of storing and querying MS and MS/MS data, as well as the disk or memory consumption.

We have shown in both Chapter 3 and Chapter 4 that the selection of molecular signatures is sensitive to small changes in the training set, even though only the EMT features were employed. By using the FSFs, we have obtained remarkable improvements in prediction performance. As introduced in Chapter 1, methods for integrating multiple omics data have been developed in recent years for different predictive purposes. In this ongoing endeavor, one has to inevitably deal with even higher dimensionality of data. Thus, feature selection becomes indispensable to avoid overfitting and improve the robustness of the signatures. Our study shows convincingly that biological knowledge-based feature selection, both on individual data level and on multiple data levels, can improve the predictability of the models significantly. This is in line with the current research frontier where the utilities of knowledge-driven predictive models are being acknowledged.

6.2 Limitations

Much of the limitations of our study come from the data. The curse of dimensionality makes it hard to find the molecular signatures. Additionally, the heterogeneity of samples makes it more difficult to differentiate signals and noise. For the same reason, some variations in omics data are probably caused by other reasons unrelated to the phenotype of interest, e.g., whether the patient has other diseases. However, due to the limitation of the sample size and clinical data, this effect cannot be separated. As for most of the cancer omics measurements, the molecular data are generated at the time of diagnosis or surgery. The data are static and cannot capture the changes of molecular profiles in a later disease stage. With static data, it is not known how the disease develops and affects the prognosis. For example, we know that EMT is a gene regulatory process that can be found in different status. Studies show that the master regulators in each status can be different [214, 236]. When using static EMT data, the profile of how EMT evolves remains unknown.

The second limitation or challenge in our opinion is the complexity of molecular interaction networks. Given a large network such as PPI, it is hard to know which parts of it are important for the phenotype of interest. If we take the entire network, the irrelevant information can overwhelm the signals. Additionally, it is often hard to map the expression or modifications of the functional molecules - usually proteins, to the network nodes, because the data may not be available. Many post-translational modifications are still technically difficult to measure [162].

6.3 Outlook - an Intelligent Decision Support System

Although recent studies have proposed new methods for omics data integration, the potential clinical utility of integrating these data levels remains largely unknown. This is related to the randomness of results from different studies, which could be caused by the curse of dimensionality. Based on the good prediction performance of our proposed RRNFS framework that integrates multiple omics data, we propose to build an intelligent decision support system (IDSS) for not only prognosis prediction, but also for a variety of other phenotypes. A schematic view is given in Figure 6.1. Its major components are the data management system (DMS), model management system (MMS), knowledge management system (KMS), and user interface. They goal is to integrate multiple levels of omics data with biologically relevant features from each level, which can be identified with the assistance of domain knowledge. These phenotype relevant features will be used for feature selection and building predictive models. The models can be refined iteratively with increased sample size, updated domain knowledge, and new feature selection algorithms.

Building such an IDSS requires large amount of digital patient data, a rich variety of biological knowledge-base [138], and efficient computational infrastructures. All these requirements can nowadays be satisfied. To realize personalized medicine for complex diseases, which necessitates timely decision makings in diagnosis, prognosis, and treatment, a centralized IDSS is desirable. It can potentially:

- Improve the quality of decisions. IDSS can select phenotype relevant features and evaluate multiple predictive models to support its decision making. With the addition of more patient data, the MSS iteratively refines the models to give robust predictions.

- Facilitate data integration. While it is hard for individuals to process large amount of biological knowledge and employ a wide variety of models, this system can offer these options at one place.

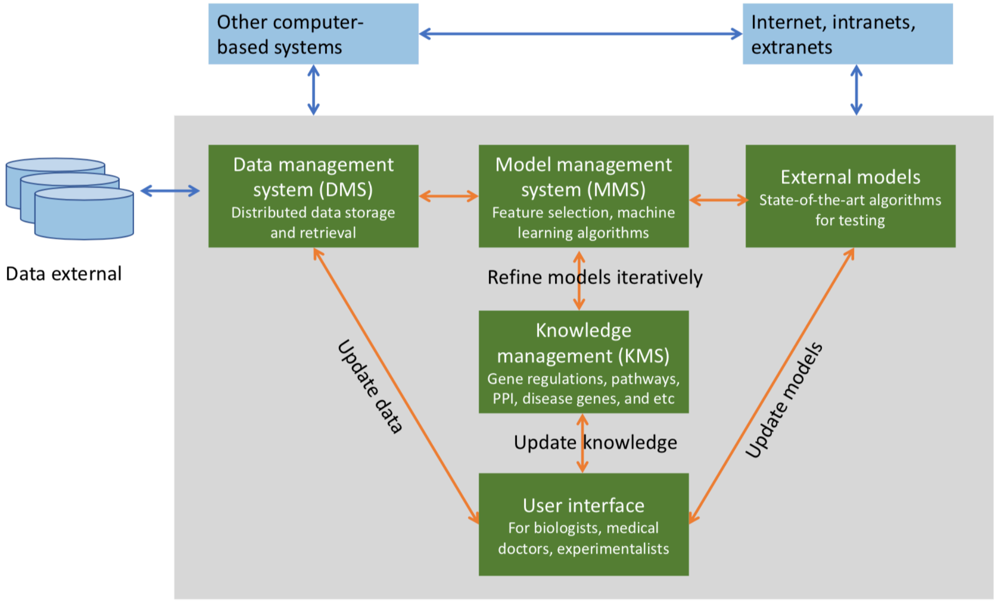

Other computer-based systems
Internet, intranets, extranets
Data external
Data management system (DMS)
Distributed data storage and retrieval
Model management system (MMS)
Feature selection, machine learning algorithms
External models
State-of-the-art algorithms for testing
Refine models iteratively
Knowledge management (KMS)
Gene regulations, pathways, PPI, disease genes, and etc
Update data
Update models
Update knowledge
User interface
For biologists, medical doctors, experimentalists

Figure 6.1: A schematic view of an IDSS for personalized medical decision making.

- Improve the communication of experts. IDSS can provide user interfaces to stakeholders at different locations. They can jointly contribute to the knowledge-base and predictive models.

- Reduce cost. Compared with building one intelligent system at every research institute and hospital, this system can benefit many medical centers. By improving the quality of decisions, unnecessary medical treatments could be avoided.

Bibliography

[1] O. O. Aalen, *A linear regression model for the analysis of life times*, Statistics in medicine, 8 (1989), pp. 907–925.

[2] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, et al., *Gene prioritization through genomic data fusion*, Nature biotechnology, 24 (2006), pp. 537–544.

[3] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, *On the surprising behavior of distance metrics in high dimensional spaces*, in ICDT, vol. 1, Springer, 2001, pp. 420–434.

[4] R. Agrawal, T. Imieliński, and A. Swami, *Mining association rules between sets of items in large databases*, in Acm sigmod record, vol. 22, ACM, 1993, pp. 207–216.

[5] R. Agrawal, R. Srikant, et al., *Fast algorithms for mining association rules*, in Proc. 20th int. conf. very large data bases, VLDB, vol. 1215, 1994, pp. 487–499.

[6] U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway, and D. Pe'er, *An integrated approach to uncover drivers of cancer*, Cell, 143 (2010), pp. 1005–1017.

[7] J. A. Alfaro, A. Sinha, T. Kislinger, and P. C. Boutros, *Onco-proteogenomics: cancer proteomics joins forces with genomics*, Nature methods, 11 (2014), pp. 1107–1113.

[8] A. Allahyar and J. de Ridder, *Feral: network-based classifier with application to breast cancer outcome prediction*, Bioinformatics, 31 (2015), pp. i311–i319.

[9] D. C. Altieri, *Survivin, cancer networks and pathway-directed drug discovery*, Nature reviews. Cancer, 8 (2008), p. 61.

[10] A. Alyass, M. Turcotte, and D. Meyre, *From big data analysis to personalized medicine for all: challenges and opportunities*, BMC medical genomics, 8 (2015), p. 33.

[11] P. P. Amaral, M. E. Dinger, T. R. Mercer, and J. S. Mattick, *The eukaryotic genome as an rna machine*, science, 319 (2008), pp. 1787–1789.

[12] P. Atzeni, F. Bugiotti, and L. Rossi, *Uniform access to NoSQL systems*, Information Systems, 43 (2014), pp. 117 – 133.

[13] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster, *Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present*, Analytical and bioanalytical chemistry, 404 (2012), pp. 939–965.

[14] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, *Network medicine: a network-based approach to human disease*, Nature Reviews Genetics, 12 (2011), pp. 56–68.

[15] E. Batlle, E. Sancho, C. Francí, D. Domínguez, M. Monfar, J. Baulida, and A. G. de Herreros, *The transcription factor snail is a repressor of e-cadherin gene expression in epithelial tumour cells*, Nature cell biology, 2 (2000), pp. 84–89.

[16] A. H. Beck, N. W. Knoblauch, M. M. Hefti, J. Kaplan, S. J. Schnitt, A. C. Culhane, M. S. Schroeder, T. Risch, J. Quackenbush, and B. Haibe-Kains, *Significance analysis of prognostic signatures*, PLoS computational biology, 9 (2013), p. e1002875.

[17] A. Bensimon, A. J. Heck, and R. Aebersold, *Mass spectrometry–based proteomics and network biology*, Annual review of biochemistry, 81 (2012), pp. 379–405.

[18] M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, and L. Milanesi, *Methods for the integration of multi-omics data: mathematical aspects*, BMC bioinformatics, 17 (2016), p. 15.

[19] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, *When is âĂIJnearest neighborâĂİ meaningful?*, in International conference on database theory, Springer, 1999, pp. 217–235.

[20] M. J. Bissell, D. C. Radisky, A. Rizki, V. M. Weaver, and O. W. Petersen, *The organizing principle: microenvironmental influences in the normal and malignant breast*, Differentiation, 70 (2002), pp. 537–546.

[21] M. M. Bjaanæs, T. Fleischer, A. R. Halvorsen, A. Daunay, F. Busato, S. Solberg, L. Jørgensen, E. Kure, H. Edvardsen, A.-L. Børresen-Dale, et al., *Genome-wide dna methylation analyses in lung adenocarcinomas: association with egfr, kras and tp53 mutation status, gene expression and prognosis*, Molecular oncology, 10 (2016), pp. 330–343.

[22] A. M. Bode and Z. Dong, *Post-translational modification of p53 in tumorigenesis*, Nature reviews. Cancer, 4 (2004), p. 793.

[23] K. Boyd, K. H. Eng, and C. D. Page, *Area under the precision-recall curve: Point estimates and confidence intervals*, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2013, pp. 451–466.

[24] G. B. Bram De Craene, *Regulatory networks defining emt during cancer initiation and progression*, Nature Reviews Cancer, 13 (2013), pp. 97 – 110.

[25] U. Burk, J. Schubert, U. Wellner, O. Schmalhofer, E. Vincan, S. Spaderna, and T. Brabletz, *A reciprocal repression between zeb1 and members of the mir-200 family promotes emt and invasion in cancer cells*, EMBO reports, 9 (2008), pp. 582–589.

[26] S. Bustin, *Molecular biology of the cell, sixth edition; isbn: 9780815344643; and molecular biology of the cell, sixth edition, the problems book; isbn 9780815344537*, International Journal of Molecular Sciences, 16 (2015), pp. 28123–28125.

[27] V. Byles, L. Zhu, J. Lovaas, L. Chmilewski, J. Wang, D. Faller, and Y. Dai, *Sirt1 induces emt by cooperating with emt transcription factors and enhances prostate cancer cell migration and metastasis*, Oncogene, 31 (2012), pp. 4619–4629.

[28] K. Campbell, G. Whissell, X. Franch-Marro, E. Batlle, and J. Casanova, *Specific gata factors act as conserved inducers of an endodermal-emt*, Developmental cell, 21 (2011), pp. 1051–1061.

[29] A. Cano, M. A. Pérez-Moreno, I. Rodrigo, A. Locascio, M. J. Blanco, M. G. del Barrio, F. Portillo, and M. A. Nieto, *The transcription factor snail controls epithelial–mesenchymal transitions by repressing e-cadherin expression*, Nature cell biology, 2 (2000), pp. 76–83.

[30] M. Ceccarelli, F. P. Barthel, T. M. Malta, T. S. Sabedot, S. R. Salama, B. A. Murray, O. Morozova, Y. Newton, A. Radenbaugh, S. M. Pagnotta, et al., *Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma*, Cell, 164 (2016), pp. 550–563.

[31] C.-J. Chang, C.-H. Chao, W. Xia, J.-Y. Yang, Y. Xiong, C.-W. Li, W.-H. Yu, S. K. Rehman, J. L. Hsu, H.-H. Lee, et al., *p53 regulates epithelial-mesenchymal transition and stem cell properties through modulating mirnas*, Nature cell biology, 13 (2011), pp. 317–323.

[32] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu, *Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods*, PloS one, 6 (2011), p. e17238.

[33] H. Chen and P. C. Boutros, *Venndiagram: a package for the generation of highly-customizable venn and euler diagrams in r*, BMC bioinformatics, 12 (2011), p. 35.

[34] H.-C. Chen, R. L. Kodell, K. F. Cheng, and J. J. Chen, *Assessment of performance of survival prediction models for cancer prognosis*, BMC medical research methodology, 12 (2012), p. 1.

[35] L. Chen, J. Xuan, R. B. Riggins, R. Clarke, and Y. Wang, *Identifying cancer biomarkers by network-constrained support vector machines*, BMC systems biology, 5 (2011), p. 161.

[36] Y. Chen, J. Hao, W. Jiang, T. He, X. Zhang, T. Jiang, and R. Jiang, *Identifying potential cancer driver genes by genomic data integration*, Scientific reports, 3 (2013), p. 3538.

[37] G. Z. Cheng, W. Zhang, M. Sun, Q. Wang, D. Coppola, M. Mansour, L. Xu, C. Costanzo, J. Q. Cheng, and L.-H. Wang, *Twist is transcriptionally induced by activation of stat3 and mediates stat3 oncogenic function*, Journal of Biological Chemistry, 283 (2008), pp. 14665–14673.

[38] Y. CHIKAISHI, H. Uramoto, and F. Tanaka, *The emt status in the primary tumor does not predict postoperative recurrence or disease-free survival in lung adenocarcinoma*, Anticancer research, 31 (2011), pp. 4451–4456.

[39] L. Chiu, I. Hsin, T. Yang, W. Sung, J. Chi, J. Chang, J. Ko, and G. Sheu, *The erk–zeb1 pathway mediates epithelial–mesenchymal transition in pemetrexed resistant lung cancer cells with suppression by vinca alkaloids*, Oncogene, 36 (2017), pp. 242–253.

[40] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, *Network-based classification of breast cancer metastasis*, Molecular systems biology, 3 (2007).

[41] J. Comijn, G. Berx, P. Vermassen, K. Verschueren, L. van Grunsven, E. Bruyneel, M. Mareel, D. Huylebroeck, and F. Van Roy, *The two-handed e box binding zinc finger protein sip1 downregulates e-cadherin and induces invasion*, Molecular cell, 7 (2001), pp. 1267–1278.

[42] T. Conrad, *New statistical algorithms for the analysis of mass spectrometry time-of-flight mass data with applications in clinical diagnostics*, PhD thesis, Freie Universität Berlin, 2008.

[43] D. R. Cox, *Regression models and life-tables*, Journal of the Royal Statistical Society. Series B (Methodological), 34 (1972), pp. 187–220.

[44] D. R. Cox, *Partial likelihood*, Biometrika, (1975), pp. 269–276.

[45] J. Cox and M. Mann, *Quantitative, high-resolution proteomics for data-driven systems biology*, Annual review of biochemistry, 80 (2011), pp. 273–299.

[46] A. P. Crijns, R. S. Fehrmann, S. de Jong, F. Gerbens, G. J. Meersma, H. G. Klip, H. Hollema, R. M. Hofstra, G. J. te Meerman, E. G. de Vries, et al., *Survival-related profile, pathways, and transcription factors in ovarian cancer*, PLoS Med, 6 (2009), p. e1000024.

[47] G. Csardi and T. Nepusz, *The igraph software package for complex network research*, InterJournal, Complex Systems, 1695 (2006), pp. 1–9.

[48] Y. Cun and H. Fröhlich, *Prognostic gene signatures for patient stratification in breast cancer-accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions*, BMC bioinformatics, 13 (2012), p. 69.

[49] ——, *Network and data integration for biomarker signature discovery via network smoothed t-statistics*, PloS one, 8 (2013), p. e73074.

[50] M.-P. Curado, B. Edwards, H. R. Shin, H. Storm, J. Ferlay, M. Heanue, P. Boyle, et al., *Cancer incidence in five continents, Volume IX.*, IARC Press, International Agency for Research on Cancer, 2007.

[51] P. Dao, R. Colak, R. Salari, F. Moser, E. Davicioni, A. SchÃűnhuth, and M. Ester, *Inferring cancer subnetwork markers using density-constrained biclustering*, Bioinformatics, 26 (2010), pp. i625–i631.

[52] N. Dave, S. Guaita-Esteruelas, S. Gutarra, À. Frias, M. Beltran, S. Peiró, and A. G. de Herreros, *Functional cooperation between snail1 and twist in the regulation of zeb1 expression during epithelial to mesenchymal transition*, Journal of Biological Chemistry, 286 (2011), pp. 12024–12032.

[53] J. Davis and M. Goadrich, *The relationship between precision-recall and roc curves*, in Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 233–240.

[54] T. De Bie, L.-C. Tranchevent, L. M. Van Oeffelen, and Y. Moreau, *Kernel-based data fusion for gene prioritization*, Bioinformatics, 23 (2007), pp. i125–i132.

[55] B. De Craene, B. Gilbert, C. Stove, E. Bruyneel, F. Van Roy, and G. Berx, *The transcription factor snail induces tumor cell invasion through modulation of the epithelial cell differentiation program*, Cancer research, 65 (2005), pp. 6237–6244.

[56] H. Deng and G. Runger, *Feature selection via regularized trees*, in Neural Networks (IJCNN), The 2012 International Joint Conference on, IEEE, 2012, pp. 1–8.

[57] ——, *Gene selection with guided regularized random forest*, Pattern Recognition, 46 (2013), pp. 3483–3489.

[58] V. Dhingra, M. Gupta, T. Andacht, and Z. F. Fu, *New frontiers in proteomics research: a perspective*, International journal of pharmaceutics, 299 (2005), pp. 1–18.

[59] R. Díaz-Uriarte and S. A. De Andres, *Gene selection and classification of microarray data using random forest*, BMC bioinformatics, 7 (2006), p. 3.

[60] L. Ding, G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. M. Muzny, M. B. Morgan, et al., *Somatic mutations affect key pathways in lung adenocarcinoma*, Nature, 455 (2008), p. 1069.

[61] P. Domingos, *A few useful things to know about machine learning*, Communications of the ACM, 55 (2012), pp. 78–87.

[62] Y. Drier and E. Domany, *Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes?*, PloS one, 6 (2011), p. e17795.

[63] C. Du, C. Zhang, S. Hassan, M. H. U. Biswas, and K. Balaji, *Protein kinase d1 suppresses epithelial-to-mesenchymal transition through phosphorylation of snail*, Cancer research, 70 (2010), pp. 7810–7819.

[64] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification and scene analysis part 1: Pattern classification*, Wiley, Chichester, (2000).

[65] A. Eger, K. Aigner, S. Sonderegger, B. Dampier, S. Oehler, M. Schreiber, G. Berx, A. Cano, H. Beug, and R. Foisner, *Deltaef1 is a transcriptional repressor of e-cadherin and regulates epithelial plasticity in breast cancer cells*, Oncogene, 24 (2005), pp. 2375–2385.

[66] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, *Outcome signature genes in breast cancer: is there a unique set?*, Bioinformatics, 21 (2004), pp. 171–178.

[67] L. Ein-Dor, O. Zuk, and E. Domany, *Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer*, Proceedings of the National Academy of Sciences, 103 (2006), pp. 5923–5928.

[68] J. Espada, H. Peinado, L. Lopez-Serra, F. Setién, P. Lopez-Serra, A. Portela, J. Renart, E. Carrasco, M. Calvo, A. Juarranz, et al., *Regulation of snail1 and e-cadherin function by dnmt1 in a dna methylation-independent context*, Nucleic acids research, 39 (2011), pp. 9194–9205.

[69] V. Evdokimova, C. Tognon, T. Ng, P. Ruzanov, N. Melnyk, D. Fink, A. Sorokin, L. P. Ovchinnikov, E. Davicioni, T. J. Triche, et al., *Translational activation of snail1 and other developmentally regulated transcription factors by yb-1 promotes an epithelial-mesenchymal transition*, Cancer cell, 15 (2009), pp. 402–415.

[70] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, and M. L. Tress, *Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes*, Human molecular genetics, 23 (2014), pp. 5866–5878.

[71] J. F. Fahrmann, D. Grapov, B. S. Phinney, C. Stroble, B. C. DeFelice, W. Rom, D. R. Gandara, Y. Zhang, O. Fiehn, H. Pass, et al., *Proteomic profiling of lung adenocarcinoma indicates heightened dna repair, antioxidant mechanisms and identifies lasp1 as a potential negative predictor of survival*, Clinical proteomics, 13 (2016), p. 31.

[72] T. Fawcett, *Roc graphs: Notes and practical considerations for researchers*, Machine learning, 31 (2004), pp. 1–38.

[73] J. Friedman, T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, Journal of statistical software, 33 (2010), p. 1.

[74] U. H. Frixen, J. Behrens, M. Sachs, G. Eberle, B. Voss, A. Warda, D. Lochner, and W. Birchmeier, *E-cadherin-mediated cell-cell adhesion prevents invasiveness of human carcinoma cells*, J cell Biol, 113 (1991), pp. 173–185.

[75] L. I. Furlong, *Human diseases through the lens of network biology*, Trends in Genetics, 29 (2013), pp. 150–159.

[76] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, *A census of human cancer genes*, Nature Reviews Cancer, 4 (2004), pp. 177–183.

[77] O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. D. Moor, *Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks*, Bioinformatics, 22 (2006), pp. e184–e190.

[78] C. Ghigna, S. Giordano, H. Shen, F. Benvenuto, F. Castiglioni, P. M. Comoglio, M. R. Green, S. Riva, and G. Biamonti, *Cell motility is controlled by sf2/asf through alternative splicing of the ron protooncogene*, Molecular cell, 20 (2005), pp. 881–890.

[79] M. H. Glickman and A. Ciechanover, *The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction*, Physiological reviews, 82 (2002), pp. 373–428.

[80] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, *The human disease network*, Proceedings of the National Academy of Sciences, 104 (2007), pp. 8685–8690.

[81] W. W. B. Goh, W. Wang, and L. Wong, *Why batch effects matter in omics data, and how to avoid them*, Trends in Biotechnology, (2017).

[82] A. Goldberg, *Functions of the proteasome: from protein degradation and immune surveillance to cancer therapy*, Biochemical Society Transactions, 35 (2007), pp. 12–17.

[83] J. Grau, I. Grosse, and J. Keilwagen, *Prroc: computing and visualizing precision-recall and receiver operating characteristic curves in r*, Bioinformatics, (2015), p. btv153.

[84] G. Greenburg and E. Hay, *Cytodifferentiation and tissue phenotype change during transformation of embryonic lens epithelium to mesenchyme-like cells in vitro*, Developmental biology, 115 (1986), pp. 363–379.

[85] G. Greenburg and E. D. Hay, *Epithelia suspended in collagen gels can lose polarity and express characteristics of migrating mesenchymal cells.*, The Journal of cell biology, 95 (1982), pp. 333–339.

[86] M. Greenwood et al., *A report on the natural duration of cancer.*, A Report on the Natural Duration of Cancer., (1926).

[87] P. A. Gregory, A. G. Bert, E. L. Paterson, S. C. Barry, A. Tsykin, G. Farshid, M. A. Vadas, Y. Khew-Goodall, and G. J. Goodall, *The mir-200 family and mir-205 regulate epithelial to mesenchymal transition by targeting zeb1 and sip1*, Nature cell biology, 10 (2008), pp. 593–601.

[88] Q. Gu, Z. Li, and J. Han, *Generalized fisher score for feature selection*, arXiv preprint arXiv:1202.3725, (2012).

[89] S. Guaita, I. Puig, C. FranciÌĄ, M. Garrido, D. DomiÌĄnguez, E. Batlle, E. Sancho, S. Dedhar, A. G. de Herreros, and J. Baulida, *Snail induction of epithelial to mesenchymal transition in tumor cells is accompanied by muc1 repression and zeb1 expression*, Journal of Biological Chemistry, 277 (2002), pp. 39209–39216.

[90] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Gene selection for cancer classification using support vector machines*, Machine learning, 46 (2002), pp. 389–422.

[91] F. Gwinner, G. Boulday, C. Vandiedonck, M. Arnould, C. Cardoso, I. Nikolayeva, O. Guitart-Pla, C. V. Denis, O. D. Christophe, J. Beghain, et al., *Network-based analysis of omics data: The lean method*, Bioinformatics, (2016), p. btw676.

[92] M. Hahsler, B. Grün, and K. Hornik, *A computational environment for mining association rules and frequent item sets*, (2005).

[93] K. M. Hajra, D. Y. Chen, and E. R. Fearon, *The slug zinc-finger protein represses e-cadherin in breast cancer*, Cancer research, 62 (2002), pp. 1613–1618.

[94] D. A. Hall, J. Ptacek, and M. Snyder, *Protein microarray technology*, Mechanisms of ageing and development, 128 (2007), pp. 161–167.

[95] M. A. Hall, *Correlation-based feature selection for machine learning*, (1999).

[96] D. Hanahan and R. A. Weinberg, *Hallmarks of cancer: the next generation*, cell, 144 (2011), pp. 646–674.

[97] A.-C. Haury, P. Gestraud, and J.-P. Vert, *The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures*, PloS one, 6 (2011), p. e28210.

[98] S. Hautaniemi, M. Ringnér, P. Kauraniemi, R. Autio, H. Edgren, O. Yli-Harja, J. Astola, A. Kallioniemi, and O.-P. Kallioniemi, *A strategy for identifying putative causes of gene expression variation in human cancers*, Journal of the Franklin Institute, 341 (2004), pp. 77–88.

[99] C. T. Have and L. J. Jensen, *Are graph databases ready for bioinformatics?*, Bioinformatics, (2013).

[100] E. Hay, *An overview of epithelio-mesenchymal transformation*, Cells Tissues Organs, 154 (1995), pp. 8–20.

[101] Z. M. Hira and D. F. Gillies, *A review of feature selection and feature extraction methods applied on microarray data*, Advances in bioinformatics, 2015 (2015).

[102] K. Horiguchi, K. Sakamoto, D. Koinuma, K. Semba, A. Inoue, S. Inoue, H. Fujii, A. Yamaguchi, K. Miyazawa, K. Miyazono, et al., *Tgf-β drives epithelial-mesenchymal transition through δef1-mediated downregulation of esrp*, Oncogene, 31 (2012), pp. 3190–3201.

[103] J. J. Hornberg, F. J. Bruggeman, H. V. Westerhoff, and J. Lankelma, *Cancer: a systems biology disease*, Biosystems, 83 (2006), pp. 81–90.

[104] N. Howlader, A. B. Mariotto, S. Woloshin, and L. M. Schwartz, *Providing clinicians and patients with actual prognosis: cancer in the context of competing causes of death*, Journal of the National Cancer Institute Monographs, 2014 (2014), pp. 255–264.

[105] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W.-T. Tsai, G.-Z. Chen, C.-J. Lee, C.-M. Chiu, et al., *mirtarbase: a database curates experimentally validated microrna–target interactions*, Nucleic acids research, 39 (2010), pp. D163–D169.

[106] J. Hua, W. D. Tembe, and E. R. Dougherty, *Performance of feature-selection methods in the classification of high-dimension data*, Pattern Recognition, 42 (2009), pp. 409–424.

[107] H.-L. Huang, Y.-C. Wu, L.-J. Su, Y.-J. Huang, P. Charoenkwan, W.-L. Chen, H.-C. Lee, W.-C. Chu, and S.-Y. Ho, *Discovery of prognostic biomarkers for predicting lung cancer metastasis using microarray and survival data*, BMC Bioinformatics, 16 (2015).

[108] J. Huang, J. L. Horowitz, and S. Ma, *Asymptotic properties of bridge estimators in sparse high-dimensional regression models*, The Annals of Statistics, (2008), pp. 587–613.

[109] R. Y.-J. Huang, P. Guilford, and J. P. Thiery, *Early events in cell adhesion and polarity during epithelial-mesenchymal transition*, 2012.

[110] S. Huang, K. Chaudhary, and L. X. Garmire, *More is better: Recent progress in multi-omics data integration methods*, Frontiers in Genetics, 8 (2017), p. 84.

[111] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, *Discovering regulatory and signalling circuits in molecular interaction networks*, Bioinformatics, 18 (2002), pp. S233–S240.

[112] J. Ikenouchi, M. Matsuda, M. Furuse, and S. Tsukita, *Regulation of tight junctions during the epithelium-mesenchyme transition: direct repression of the gene expression of claudins/occludin by snail*, Journal of cell science, 116 (2003), pp. 1959–1967.

[113] I. Inza, P. Larrañaga, R. Blanco, and A. J. Cerrolaza, *Filter versus wrapper gene selection approaches in dna microarray domains*, Artificial intelligence in medicine, 31 (2004), pp. 91–103.

[114] L. Jacob, G. Obozinski, and J.-P. Vert, *Group lasso with overlap and graph lasso*, in Proceedings of the 26th annual international conference on machine learning, ACM, 2009, pp. 433–440.

[115] P. James, M. Quadroni, E. Carafoli, and G. Gonnet, *Protein identification by mass profile fingerprinting*, Biochemical and biophysical research communications, 195 (1993), pp. 58–64.

[116] M. Jechlinger, S. Grünert, and H. Beug, *Mechanisms in epithelial plasticity and metastasis: insights from 3d cultures and expression profiling*, Journal of mammary gland biology and neoplasia, 7 (2002), pp. 415–432.

[117] Y.-G. Jiang, Y. Luo, D.-l. He, X. Li, L.-l. Zhang, T. Peng, M.-C. Li, and Y.-H. Lin, *Role of wnt/β-catenin signaling pathway in epithelial-mesenchymal transition of human prostate cancer induced by hypoxia-inducible factor-1α*, International Journal of Urology, 14 (2007), pp. 1034–1039.

[118] N. Jin, H. Wu, Z. Miao, Y. Huang, Y. Hu, X. Bi, D. Wu, K. Qian, L. Wang, C. Wang, et al., *Network-based survival-associated module biomarker and its crosstalk with cell death genes in ovarian cancer*, Scientific reports, 5 (2015).

[119] R. Kalluri, *Emt: when epithelial cells decide to become mesenchymal-like cells*, The Journal of clinical investigation, 119 (2009), pp. 1417–1419.

[120] R. Kalluri and E. G. Neilson, *Epithelial-mesenchymal transition and its implications for fibrosis*, The Journal of clinical investigation, 112 (2003), pp. 1776–1784.

[121] R. Kalluri and R. A. Weinberg, *The basics of epithelial-mesenchymal transition*, The Journal of Clinical Investigation, 119 (2009), pp. 1420–1428.

[122] A. Kalousis, J. Prados, and M. Hilario, *Stability of feature selection algorithms*, in Data Mining, Fifth IEEE International Conference on, IEEE, 2005, pp. 8–pp.

[123] ——, *Stability of feature selection algorithms: a study on high-dimensional spaces*, Knowledge and information systems, 12 (2007), pp. 95–116.

[124] E. L. Kaplan and P. Meier, *Nonparametric estimation from incomplete observations*, Journal of the American statistical association, 53 (1958), pp. 457–481.

[125] J. Keilwagen, I. Grosse, and J. Grau, *Area under precision-recall curves for weighted and unweighted data*, PLoS One, 9 (2014), p. e92209.

[126] D. Kessner, M. Chambers, R. Burke, D. Agus, and P. Mallick, *Proteowizard: open source software for rapid proteomics tools development*, Bioinformatics, 24 (2008), pp. 2534–2536.

[127] D. Kim, R. Li, A. Lucas, S. S. Verma, S. M. Dudek, and M. D. Ritchie, *Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma*, Journal of the American Medical Informatics Association, 24 (2016), pp. 577–587.

[128] D. Kim, H. Shin, Y. S. Song, and J. H. Kim, *Synergistic effect of different levels of genomic data for cancer clinical outcome prediction*, Journal of biomedical informatics, 45 (2012), pp. 1191–1198.

[129] D. Kim, H. Shin, Y. S. Song, and J. H. Kim, *Synergistic effect of different levels of genomic data for cancer clinical outcome prediction*, Journal of Biomedical Informatics, 45 (2012), pp. 1191 – 1198.

[130] M. Kim, Y. Nam, and H. Shin, *An inference method from multi-layered structure of biomedical data*, BMC medical informatics and decision making, 17 (2017), p. 52.

[131] T. Kim, A. Veronese, F. Pichiorri, T. J. Lee, Y.-J. Jeon, S. Volinia, P. Pineau, A. Marchio, J. Palatini, S.-S. Suh, et al., *p53 regulates epithelial–mesenchymal transition through micrornas targeting zeb1 and zeb2*, The Journal of experimental medicine, 208 (2011), pp. 875–883.

[132] T. Kislinger, B. Cox, A. Kannan, C. Chung, P. Hu, A. Ignatchenko, M. S. Scott, A. O. Gramolini, Q. Morris, M. T. Hallett, et al., *Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling*, Cell, 125 (2006), pp. 173–186.

[133] R. Kohavi et al., *A study of cross-validation and bootstrap for accuracy estimation and model selection*, in Ijcai, vol. 14, Stanford, CA, 1995, pp. 1137–1145.

[134] P. Koistinen and L. Holmström, *Kernel regression and backpropagation training with noise*, in Advances in Neural Information Processing Systems, 1992, pp. 1033–1039.

[135] D. Koller and M. Sahami, *Toward optimal feature selection*, tech. rep., Stanford InfoLab, 1996.

[136] K. Komurov, S. Dursun, S. Erdin, and P. T. Ram, *Netwalker: a contextual network analysis tool for functional genomics*, BMC genomics, 13 (2012), p. 282.

[137] P. K. Kreeger and D. A. Lauffenburger, *Cancer systems biology: a network modeling perspective*, Carcinogenesis, 31 (2009), pp. 2–8.

[138] I. Kuperstein, L. Grieco, D. P. Cohen, D. Thieffry, A. Zinovyev, and E. Barillot, *The shortest path is not the one you know: application of biological network resources in precision oncology research*, Mutagenesis, 30 (2015), pp. 191–204.

[139] S. Lamouille, J. Xu, and R. Derynck, *Molecular mechanisms of epithelial–mesenchymal transition*, Nature reviews. Molecular cell biology, 15 (2014), p. 178.

[140] D. R. Lamouille Samy, Xu Jian, *Molecular mechanisms of epithelialâĂŞmesenchymal transition*, NATURE REVIEWS MOLECULAR CELL BIOLOGY, 15 (2014), pp. 178–196.

[141] D. Laney, *The Importance of 'Big Data': A Definition*, (2012).

[142] A. N. Langville and C. D. Meyer, *Deeper inside pagerank*, Internet Mathematics, 1 (2004), pp. 335–380.

[143] J. Lapointe, C. Li, J. P. Higgins, M. Van De Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, et al., *Gene expression profiling identifies clinically relevant subtypes of prostate cancer*, Proceedings of the National Academy of Sciences of the United States of America, 101 (2004), pp. 811–816.

[144] M. Lauss, I. Visne, A. Kriegner, M. Ringnér, G. Jönsson, and M. Höglund, *Monitoring of technical variation in quantitative high-throughput datasets*, Cancer informatics, 12 (2013), p. 193.

[145] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, *A survey on filter techniques for feature selection in gene expression microarray analysis*, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 9 (2012), pp. 1106–1119.

[146] A. Le Bechec, E. Portales-Casamar, G. Vetter, M. Moes, P.-J. Zindy, A. Saumet, D. Arenillas, C. Theillet, W. Wasserman, C.-H. Lecellier, and E. Friederich, *Mir@nt@n: a framework integrating transcription factors, micrornas and their targets to identify sub-network motifs in a meta-regulation network model*, BMC Bioinformatics, 12 (2011), p. 67.

[147] K. K.-Y. Lee, W.-C. Tang, and K.-S. Choi, *Alternatives to Relational Database: Comparison of NoSQL and XML Approaches for Clinical Data Storage*, Comput. Methods Prog. Biomed., 110 (2013), pp. 99–109.

[148] B. Li and C. N. Dewey, *Rsem: accurate transcript quantification from rna-seq data with or without a reference genome*, BMC bioinformatics, 12 (2011), p. 1.

[149] C. Li and H. Li, *Network-constrained regularization and variable selection for analysis of genomic data*, Bioinformatics, 24 (2008), pp. 1175–1182.

[150] J. Li, A. E. Lenferink, Y. Deng, C. Collins, Q. Cui, E. O. Purisima, M. D. O'Connor-McCourt, and E. Wang, *Identification of high-quality cancer prognostic markers and metastasis network modules*, Nature communications, 1 (2010), p. 34.

[151] J. Li, Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, J.-Y. Yang, B. M. Broom, R. G. Verhaak, D. W. Kane, et al., *Tcpa: a resource for cancer functional proteomics data*, Nature methods, 10 (2013), pp. 1046–1047.

[152] J. Li, P. Roebuck, S. Grünewald, and H. Liang, *Survnet: a web server for identifying network-based biomarkers that most correlate with patient survival data*, Nucleic acids research, (2012), p. gks386.

[153] W. Li, S. Zhang, C.-C. Liu, and X. J. Zhou, *Identifying multi-layer gene regulatory modules from multi-dimensional genomic data*, Bioinformatics, 28 (2012), pp. 2458–2466.

[154] X. Li, W. Deng, C. D. Nail, S. K. Bailey, M. H. Kraus, J. M. Ruppert, and S. M. Lobo-Ruppert, *Snail induction is an early response to gli1 that determines the efficiency of epithelial transformation*, Oncogene, 25 (2006), pp. 609–621.

[155] Y. Li and J. C. Patra, *Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network*, Bioinformatics, 26 (2010), pp. 1219–1224.

[156] M. Locard-Paulet, O. Pible, A. Gonzalez de Peredo, B. Alpha-Bazin, C. Almunia, O. Burlet-Schiltz, and J. Armengaud, *Clinical implications of recent advances in proteogenomics*, Expert review of proteomics, 13 (2016), pp. 185–199.

[157] J. Loughrey and P. Cunningham, *Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets*, Research and Development in Intelligent Systems XXI, (2005), pp. 33–43.

[158] R. Louhimo and S. Hautaniemi, *Cnamet: an r package for integrating copy number, methylation and expression data*, Bioinformatics, 27 (2011), pp. 887–888.

[159] S. Ma and J. Huang, *Penalized feature selection and classification in bioinformatics*, Briefings in bioinformatics, 9 (2008), pp. 392–403.

[160] L. v. d. Maaten and G. Hinton, *Visualizing data using t-sne*, Journal of Machine Learning Research, 9 (2008), pp. 2579–2605.

[161] B. T. MacDonald, K. Tamai, and X. He, *Wnt/β-catenin signaling: components, mechanisms, and diseases*, Developmental cell, 17 (2009), pp. 9–26.

[162] P. Mallick and B. Kuster, *Proteomics: a pragmatic perspective*, Nature biotechnology, 28 (2010), pp. 695–709.

[163] P. K. Mankoo, R. Shen, N. Schultz, D. A. Levine, and C. Sander, *Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles*, PLoS One, 6 (2011), p. e24709.

[164] M. Mann, P. Højrup, and P. Roepstorff, *Use of mass spectrometric molecular weight information to identify proteins in sequence databases*, Biological mass spectrometry, 22 (1993), pp. 338–345.

[165] M. Mann and O. N. Jensen, *Proteomic analysis of post-translational modifications*, Nature biotechnology, 21 (2003), pp. 255–261.

[166] M. Marsan, G. Van den Eynden, R. Limame, P. Neven, J. Hauspy, P. A. Van Dam, I. Vergote, L. Y. Dirix, P. B. Vermeulen, and S. J. Van Laere, *A core invasiveness gene signature reflects epithelial-to-mesenchymal transition but not metastatic potential in breast cancer cell lines and tissue samples*, PloS one, 9 (2014), p. e89262.

[167] E. Martinez-Ledesma, R. G. Verhaak, and V. Treviño, *Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm*, Scientific reports, 5 (2015).

[168] A. MaâĂŹayan, *Introduction to network analysis in systems biology*, Science signaling, 4 (2011), p. tr5.

[169] J. E. McDermott, J. Wang, H. Mitchell, B.-J. Webb-Robertson, R. Hafen, J. Ramey, and K. D. Rodland, *Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data*, Expert opinion on medical diagnostics, 7 (2013), pp. 37–51.

[170] D. Medici, E. D. Hay, and B. R. Olsen, *Snail and slug promote epithelial-mesenchymal transition through β-catenin–t-cell factor-4-dependent expression of transforming growth factor-β3*, Molecular biology of the cell, 19 (2008), pp. 4875–4887.

[171] S. Michiels, S. Koscielny, and C. Hill, *Prediction of cancer outcome with microarrays: a multiple random validation strategy*, The Lancet, 365 (2005), pp. 488–492.

[172] T. M. Mitchell, *Machine learning. 1997*, Burr Ridge, IL: McGraw Hill, 45 (1997), pp. 870–877.

[173] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker, *Integrative approaches for finding modular structure in biological networks*, Nature Reviews Genetics, 14 (2013), pp. 719–732.

[174] G. Moreno-Bueno, E. Cubillo, D. Sarrió, H. Peinado, S. M. Rodríguez-Pinilla, S. Villa, V. Bolós, M. Jordá, A. Fabra, F. Portillo, et al., *Genetic profiling of epithelial cells expressing e-cadherin repressors reveals a distinct role for snail, slug, and e47 factors in epithelial-mesenchymal transition*, Cancer research, 66 (2006), pp. 9543–9556.

[175] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert, *Generank: using search engine technology for the analysis of microarray experiments*, BMC bioinformatics, 6 (2005), p. 233.

[176] S. Mounika Inavolu, J. Renbarger, M. Radovich, V. Vasudevaraja, G. Kinnebrew, S. Zhang, and L. Cheng, *Iodne: An integrated optimization method for identifying the deregulated subnetwork for precision medicine in cancer*, CPT: Pharmacometrics & Systems Pharmacology, (2017).

[177] N. Nagaraj, J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Pääbo, and M. Mann, *Deep proteome and transcriptome mapping of a human cancer cell line*, Molecular systems biology, 7 (2011), p. 548.

[178] A. Nayak, A. Poriya, and D. Poojary, *Type of NoSQL Databases and its Comparison with Relational Databases*, International Journal of Applied Information Systems, 5 (2013), pp. 16–19. Published by Foundation of Computer Science, New York, USA.

[179] A. Nayak, A. Poriya, and D. Poojary, *Type of nosql databases and its comparison with relational databases*, International Journal of Applied Information Systems, 5 (2013), pp. 16–19.

[180] A. I. Nesvizhskii, *Proteogenomics: concepts, applications and computational strategies*, Nature methods, 11 (2014), pp. 1114–1125.

[181] C. G. A. R. Network et al., *Integrated genomic analyses of ovarian carcinoma*, Nature, 474 (2011), pp. 609–615.

[182] R. Niu, L. Zhang, G. Xi, X. Wei, Y. Yang, Y. Shi, F. Zhang, and X. Hao, *Up-regulation of twist induces angiogenesis and correlates with metastasis in hepatocellular carcinoma*, 2007.

[183] D. G. Nowak, J. Woolard, E. M. Amin, O. Konopatskaya, M. A. Saleem, A. J. Churchill, M. R. Ladomery, S. J. Harper, and D. O. Bates, *Expression of pro-and anti-angiogenic isoforms of vegf is differentially regulated by splicing and growth factors*, Journal of cell science, 121 (2008), pp. 3487–3495.

[184] P. H. O'Farrell, *High resolution two-dimensional electrophoresis of proteins.*, Journal of biological chemistry, 250 (1975), pp. 4007–4021.

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

[186] S.-E. Ong, G. Mittler, and M. Mann, *Identifying and quantifying in vivo methylation sites by heavy methyl silac*, Nature methods, 1 (2004), p. 119.

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

[188] M. Y. Park and T. Hastie, *L1-regularization path algorithm for generalized linear models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69 (2007), pp. 659–677.

[189] S.-M. Park, A. B. Gaur, E. Lengyel, and M. E. Peter, *The mir-200 family determines the epithelial phenotype of cancer cells by targeting the e-cadherin repressors zeb1 and zeb2*, Genes & development, 22 (2008), pp. 894–907.

[190] V. N. Patel, G. Gokulrangan, S. A. Chowdhury, Y. Chen, A. E. Sloan, M. KoyutÃijrk, J. Barnholtz-Sloan, and M. R. Chance, *Network signatures of survival in glioblastoma multiforme*, PLoS Comput Biol, 9 (2013), p. e1003237.

[191] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos, *Using graph theory to analyze biological networks*, BioData mining, 4 (2011), p. 10.

[192] H. Peinado, E. Ballestar, M. Esteller, and A. Cano, *Snail mediates e-cadherin repression by the recruitment of the sin3a/histone deacetylase 1 (hdac1)/hdac2 complex*, Molecular and cellular biology, 24 (2004), pp. 306–319.

[193] H. Peinado, D. Olmeda, and A. Cano, *Snail, zeb and bhlh factors in tumour progression: an alliance against the epithelial phenotype?*, Nature Reviews Cancer, 7 (2007), pp. 415–428.

[194] H. Peinado, D. Olmeda, K. Csiszar, K. S. Fong, S. Vega, M. A. Nieto, A. Cano, F. Portillo, et al., *A molecular role for lysyl oxidase-like 2 enzyme in snail regulation and tumor progression*, The EMBO journal, 24 (2005), pp. 3446–3458.

[195] M. Perez and T. Marwala, *Microarray data feature selection using hybrid genetic algorithm simulated annealing*, in Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of, IEEE, 2012, pp. 1–5.

[196] M. A. Pérez-Moreno, A. Locascio, I. Rodrigo, G. Dhondt, F. Portillo, M. A. Nieto, and A. Cano, *A new role for e12/e47 in the repression ofe-cadherin expression and epithelial-mesenchymal transitions*, Journal of Biological Chemistry, 276 (2001), pp. 27424–27431.

[197] E. A. Ponomarenko, E. V. Poverennaya, E. V. Ilgisonis, M. A. Pyatnitskiy, A. T. Kopylov, V. G. Zgoda, A. V. Lisitsa, and A. I. Archakov, *The size of the human proteome: the width and depth*, International journal of analytical chemistry, 2016 (2016).

[198] F. Provost and T. Fawcett, *Robust classification systems for imprecise environments*, in AAAI/IAAI, 1998, pp. 706–713.

[199] F. J. Provost, T. Fawcett, et al., *Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions.*, in KDD, vol. 97, 1997, pp. 43–48.

[200] P. Pudil, J. Novovičová, and J. Kittler, *Floating search methods in feature selection*, Pattern recognition letters, 15 (1994), pp. 1119–1125.

[201] M. Raponi, Y. Zhang, J. Yu, G. Chen, G. Lee, J. M. Taylor, J. MacDonald, D. Thomas, C. Moskaluk, Y. Wang, et al., *Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung*, Cancer research, 66 (2006), pp. 7466–7472.

[202] L. M. Reinke, Y. Xu, and C. Cheng, *Snail represses the splicing regulator epithelial splicing regulatory protein 1 to promote epithelial-mesenchymal transition*, Journal of Biological Chemistry, 287 (2012), pp. 36435–36442.

[203] H. E. R. L. R. P. S. A. K. D. Ritchie, Marylyn D., *Methods of integrating data to uncover genotype-phenotype interactions*, Nature Review Genetics, 16 (2015), pp. 85–97.

[204] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, *proc: an open-source package for r and s+ to analyze and compare roc curves*, BMC bioinformatics, 12 (2011), p. 77.

[205] D. J. Rogers, T. T. Tanimoto, et al., *A computer program for classifying plants*, Science, 132 (1960), pp. 1115–1118.

[206] S. Rogers, M. Girolami, W. Kolch, K. M. Waters, T. Liu, B. Thrall, and H. S. Wiley, *Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models*, Bioinformatics, 24 (2008), pp. 2894–2900.

[207] M. Ruffalo, M. Koyutürk, and R. Sharan, *Network-based integration of disparate omic data to identify" silent players" in cancer*, PLoS Comput Biol, 11 (2015), p. e1004595.

[208] R. Ruiz, J. C. Riquelme, and J. S. Aguilar-Ruiz, *Incremental wrapper-based gene selection from microarray data for cancer classification*, Pattern Recognition, 39 (2006), pp. 2383–2392.

[209] Y. Saeys, I. Inza, and P. Larrañaga, *A review of feature selection techniques in bioinformatics*, bioinformatics, 23 (2007), pp. 2507–2517.

[210] N. Sah, Z. Khan, G. Khan, and P. Bisen, *Structural, functional and therapeutic biology of survivin*, Cancer letters, 244 (2006), pp. 164–171.

[211] C. Sahlgren, M. V. Gustafsson, S. Jin, L. Poellinger, and U. Lendahl, *Notch signaling mediates hypoxia-induced tumor cell migration and invasion*, Proceedings of the National Academy of Sciences, 105 (2008), pp. 6392–6397.

[212] E. Sanchez-Tillo, A. Lazaro, R. Torrent, M. Cuatrecasas, E. Vaquero, A. Castells, P. Engel, and A. Postigo, *Zeb1 represses e-cadherin and induces an emt by recruiting the swi/snf chromatin-remodeling protein brg1*, Oncogene, 29 (2010), pp. 3490–3500.

[213] M. C. Schatz, B. Langmead, and S. L. Salzberg, *Cloud computing and the dna data race*, Nature biotechnology, 28 (2010), pp. 691–693.

[214] M. J. Schliekelman, A. Taguchi, J. Zhu, X. Dai, J. Rodriguez, M. Celiktas, Q. Zhang, A. Chin, C.-H. Wong, H. Wang, et al., *Molecular portraits of epithelial, mesenchymal, and hybrid states in lung adenocarcinoma and their relevance to survival*, Cancer research, 75 (2015), pp. 1789–1800.

[215] J. A. Seoane, I. N. Day, T. R. Gaunt, and C. Campbell, *A pathway-based data integration framework for prediction of disease progression*, Bioinformatics, 30 (2014), pp. 838–845.

[216] B. Shao, C. V. Cannistraci, and T. O. Conrad, *Epithelial mesenchymal transition network-based feature engineering in lung adenocarcinoma prognosis prediction using multiple omic data*, Genomics and Computational Biology, 3 (2017), p. 57.

[217] B. Shao and T. Conrad, *Are nosql data stores useful for bioinformatics researchers?,âĂİJ*, International Journal on Recent and Innovation Trends in Computing and Communication, 3 (2015), pp. 1704–1708.

[218] K. Shedden, J. M. Taylor, S. A. Enkemann, M.-S. Tsao, T. J. Yeatman, W. L. Gerald, S. Eschrich, I. Jurisica, T. J. Giordano, D. E. Misek, et al., *Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study*, Nature medicine, 14 (2008), pp. 822–827.

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

[220] S. Shin, C. A. Dimitri, S.-O. Yoon, W. Dowdle, and J. Blenis, *Erk2 but not erk1 induces epithelial-to-mesenchymal transformation via def motif-dependent signaling events*, Molecular cell, 38 (2010), pp. 114–127.

[221] H. D. Shukla, J. Mahmood, and Z. Vujaskovic, *Integrated proteo-genomic approach for early diagnosis and prognosis of cancer*, Cancer letters, 369 (2015), pp. 28–36.

[222] R. L. Siegel, K. D. Miller, and A. Jemal, *Cancer statistics, 2016*, CA: a cancer journal for clinicians, 66 (2016), pp. 7–30.

[223] H. Siemens, R. Jackstadt, S. Hünten, M. Kaller, A. Menssen, U. Götz, and H. Hermeking, *mir-34 and snail form a double-negative feedback loop to regulate epithelial-mesenchymal transitions*, Cell cycle, 10 (2011), pp. 4256–4271.

[224] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, *Regularization paths for coxâĂŹs proportional hazards model via coordinate descent*, Journal of statistical software, 39 (2011), p. 1.

[225] Y. Song, T. Tian, X. Fu, W. Wang, S. Li, T. Shi, A. Suo, Z. Ruan, H. Guo, and Y. Yao, *Gata6 is overexpressed in breast cancer and promotes breast cancer cell epithelial–mesenchymal transition by upregulating slug expression*, Experimental and molecular pathology, 99 (2015), pp. 617–627.

[226] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*, Proceedings of the National Academy of Sciences, 98 (2001), pp. 10869–10874.

[227] N. K. Speicher and N. Pfeifer, *Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery*, Bioinformatics, 31 (2015), pp. i268–i275.

[228] C. Staiger, S. Cadot, B. Györffy, L. F. Wessels, and G. W. Klau, *Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis*, Frontiers in genetics, 4 (2013).

[229] C. Staiger, S. Cadot, R. Kooter, M. Dittrich, T. Müller, G. W. Klau, and L. F. Wessels, *A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer*, PloS one, 7 (2012), p. e34796.

[230] C. Stephan, S. Wesseling, T. Schink, and K. Jung, *Comparison of eight computer programs for receiver-operating characteristic analysis*, Clinical Chemistry, 49 (2003), pp. 433–439.

[231] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, *Bias in random forest variable importance measures: Illustrations, sources and a solution*, BMC bioinformatics, 8 (2007), p. 25.

[232] J. Subramanian and R. Simon, *Gene expression–based prognostic signatures in lung cancer: ready for clinical use?*, Journal of the National Cancer Institute, 102 (2010), pp. 464–474.

[233] M. Sugimoto, M. Kawakami, M. Robert, T. Soga, and M. Tomita, *Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis*, Current bioinformatics, 7 (2012), pp. 96–108.

[234] Z. Sun, H. S. Chai, Y. Wu, W. M. White, K. V. Donkena, C. J. Klein, V. D. Garovic, T. M. Therneau, and J.-P. A. Kocher, *Batch effect correction for genome-wide methylation data with illumina infinium platform*, BMC medical genomics, 4 (2011), p. 84.

[235] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, et al., *String v10: protein–protein interaction networks, integrated over the tree of life*, Nucleic acids research, 43 (2014), pp. D447–D452.

[236] H. Tanaka and S. Ogishima, *Network biology approach to epithelial–mesenchymal transition in cancer metastasis: three stage theory*, Journal of molecular cell biology, 7 (2015), pp. 253–266.

[237] J. Tang, S. Alelyani, and H. Liu, *Feature selection for classification: A review*, Data Classification: Algorithms and Applications, (2014), p. 37.

[238] C. J. Tauro, B. R. Patil, and K. Prashanth, *A Comparative Analysis of Different NoSQL Databases on Data Model, Query Model and Replication Model*, (2013).

[239] C. J. M. Tauro, A. S, and S. A.b, *Comparative Study of the New Generation, Agile, Scalable, High Performance NOSQL Databases*, International Journal of Computer Applications, 48 (2012), pp. 1–4.

[240] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, S. Bull, T. Pawson, Q. Morris, and J. L. Wrana, *Dynamic modularity in protein interaction networks predicts breast cancer outcome*, Nature biotechnology, 27 (2009), pp. 199–204.

[241] N. L. S. T. R. Team et al., *Reduced lung-cancer mortality with low-dose computed tomographic screening*, N Engl J Med, 2011 (2011), pp. 395–409.

[242] T. Therneau, B. Atkinson, B. Ripley, and M. B. Ripley, *Package âĂŸrpartâĂŹ*, Available online: cran. ma. ic. ac. uk/web/packages/rpart/rpart. pdf (accessed on 20 April 2016), (2017).

[243] T. M. Therneau, E. J. Atkinson, et al., *An introduction to recursive partitioning using the rpart routines*, tech. rep., Technical report Mayo Foundation, 1997.

[244] E. Theveneau and R. Mayor, *Cadherins in collective cell migration of mesenchymal cells*, Current opinion in cell biology, 24 (2012), pp. 677–684.

[245] J. P. Thiery, *Epithelial–mesenchymal transitions in tumour progression*, Nature Reviews Cancer, 2 (2002), pp. 442–454.

[246] J. P. Thiery, H. Acloque, R. Y. Huang, and M. A. Nieto, *Epithelial-mesenchymal transitions in development and disease*, cell, 139 (2009), pp. 871–890.

[247] J. P. Thiery and J. P. Sleeman, *Complex networks orchestrate epithelial–mesenchymal transitions*, Nature reviews Molecular cell biology, 7 (2006), pp. 131–142.

[248] K. Thorn, *A quick guide to light microscopy in cell biology*, Molecular biology of the cell, 27 (2016), pp. 219–222.

[249] S. Thuault, E.-J. Tan, H. Peinado, A. Cano, C.-H. Heldin, and A. Moustakas, *Hmga2 and smads co-regulate snail1 expression during induction of epithelial-to-mesenchymal transition*, Journal of Biological Chemistry, 283 (2008), pp. 33437–33446.

[250] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological), (1996), pp. 267–288.

[251] L. Toloşi and T. Lengauer, *Classification with correlated features: unreliability of feature ranking and solutions*, Bioinformatics, 27 (2011), pp. 1986–1994.

[252] K. Tsuda, H. Shin, and B. Schölkopf, *Fast protein classification with multiple networks*, Bioinformatics, 21 (2005), pp. ii59–ii65.

[253] G. A. Turenne and B. D. Price, *Glycogen synthase kinase3 beta phosphorylates serine 33 of p53 and activates p53's transcriptional activity*, BMC cell biology, 2 (2001), p. 12.

[254] C. Vandewalle, J. Comijn, B. De Craene, P. Vermassen, E. Bruyneel, H. Andersen, E. Tulchinsky, F. Van Roy, and G. Berx, *Sip1/zeb2 induces emt by repressing genes of different epithelial cell–cell junctions*, Nucleic acids research, 33 (2005), pp. 6566–6578.

[255] L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. Van Der Kooy, M. J. Marton, A. T. Witteveen, et al., *Gene expression profiling predicts clinical outcome of breast cancer*, nature, 415 (2002), pp. 530–536.

[256] S. Varambally, Q. Cao, R.-S. Mani, S. Shankar, X. Wang, B. Ateeq, B. Laxman, X. Cao, X. Jing, K. Ramnarayanan, et al., *Genomic loss of microrna-101 leads to overexpression of histone methyltransferase ezh2 in cancer*, science, 322 (2008), pp. 1695–1699.

[257] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, *Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm*, Bioinformatics, 26 (2010), pp. i237–i245.

[258] D. Venet, J. E. Dumont, and V. Detours, *Most random gene expression signatures are significantly associated with breast cancer outcome*, PLoS Comput Biol, 7 (2011), p. e1002240.

[259] C. Vicknair, M. Macias, Z. Zhao, X. Nan, Y. Chen, and D. Wilkins, *A Comparison of a Graph Database and a Relational Database: A Data Provenance Perspective*, in Proceedings of the 48th Annual Southeast Regional Conference, ACM SE '10, New York, NY, USA, 2010, ACM, pp. 42:1–42:6.

[260] A. F. Vieira and J. Paredes, *P-cadherin and the journey to cancer metastasis*, Molecular cancer, 14 (2015), p. 178.

[261] A. Vincent-Salomon and J. P. Thiery, *Host microenvironment in breast cancer development: epithelial–mesenchymal transition in breast cancer development*, Breast Cancer Research, 5 (2003), p. 101.

[262] C. Vogel and E. M. Marcotte, *Insights into the regulation of protein abundance from proteomic and transcriptomic analyses*, Nature reviews. Genetics, 13 (2012), p. 227.

[263] K. Vuoriluoto, H. Haugen, S. Kiviluoto, J. Mpindi, J. Nevo, C. Gjerdrum, C. Tiron, J. Lorens, and J. Ivaska, *Vimentin regulates emt induction by slug and oncogenic h-ras and migration by governing axl expression in breast cancer*, Oncogene, 30 (2011), pp. 1436–1448.

[264] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, *Similarity network fusion for aggregating data types on a genomic scale*, Nature methods, 11 (2014), pp. 333–337.

[265] D. Wang, A. E. Berglund, R. S. Kenchappa, R. J. MacAulay, J. J. Mulé, and A. B. Etame, *Birc3 is a biomarker of mesenchymal habitat of glioblastoma, and a mediator of survival adaptation in hypoxia-driven glioblastoma habitats.*, Scientific reports, 7 (2017), p. 9350.

[266] S. Wang, H. Sun, J. Ma, C. Zang, C. Wang, J. Wang, Q. Tang, C. A. Meyer, Y. Zhang, and X. S. Liu, *Target analysis by integration of transcriptome and chip-seq data with beta*, Nature protocols, 8 (2013), p. 2502.

[267] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, et al., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer*, The Lancet, 365 (2005), pp. 671–679.

[268] M. Y. A. I. S. A. L. M. Y. X. V. R. T. M. R. H. Wang J, Zuo Y, *Pathway and network approaches for identification of cancer signature markers from omics data*, vol. 6, 2015, pp. 54–65.

[269] C. C. Warzecha, P. Jiang, K. Amirikian, K. A. Dittmar, H. Lu, S. Shen, W. Guo, Y. Xing, and R. P. Carstens, *An esrp-regulated splicing programme is abrogated during the epithelial–mesenchymal transition*, The EMBO journal, 29 (2010), pp. 3286–3300.

[270] S. Watanabe, Y. Ueda, S.-i. Akaboshi, Y. Hino, Y. Sekita, and M. Nakao, *Hmga2 maintains oncogenic ras-induced epithelial-mesenchymal transition in human pancreatic cancer cells*, The American journal of pathology, 174 (2009), pp. 854–868.

[271] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, et al., *The cancer genome atlas pan-cancer analysis project*, Nature genetics, 45 (2013), pp. 1113–1120.

[272] M. J. Wheelock, Y. Shintani, M. Maeda, Y. Fukumoto, and K. R. Johnson, *Cadherin switching*, J Cell Sci, 121 (2008), pp. 727–735.

[273] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, et al., *Mass-spectrometry-based draft of the human proteome*, Nature, 509 (2014), p. 582.

[274] C. Winter, G. Kristiansen, S. Kersting, J. Roy, D. Aust, T. Knösel, P. Rümmele, B. Jahnke, V. Hentrich, F. Rückert, et al., *Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes*, PLoS Comput Biol, 8 (2012), p. e1002511.

[275] D. H. Wolpert, *The lack of a priori distinctions between learning algorithms*, Neural computation, 8 (1996), pp. 1341–1390.

[276] C.-I. Wu, J. A. Hoffman, B. R. Shy, E. M. Ford, E. Fuchs, H. Nguyen, and B. J. Merrill, *Function of wnt/β-catenin in counteracting tcf3 repression through the tcf3–β-catenin interaction*, Development, 139 (2012), pp. 2118–2129.

[277] G. Wu and L. Stein, *A network module-based method for identifying cancer prognostic signatures*, Genome biology, 13 (2012), p. R112.

[278] Y. Wu, J. Deng, P. G. Rychahou, S. Qiu, B. M. Evers, and B. P. Zhou, *Stabilization of snail by nf-κb is required for inflammation-induced cell migration and invasion*, Cancer cell, 15 (2009), pp. 416–428.

[279] J. Xia, M. J. Benner, and R. E. W. Hancock, *Networkanalyst - integrative approaches for proteinâĂŞprotein interaction network analysis and visual exploration*, Nucleic Acids Research, (2014).

[280] J. Xia, E. E. Gill, and R. E. Hancock, *Networkanalyst for statistical, visual and network-based meta-analysis of gene expression data*, Nature protocols, 10 (2015), pp. 823–844.

[281] M. Xiong, X. Fang, and J. Zhao, *Biomarker identification by feature wrappers*, Genome Research, 11 (2001), pp. 1878–1887.

[282] Y. Xu, S. Lee, H. Kim, N. Kim, S. Piao, S. Park, Y. Jung, J. Yook, B. Park, and N. Ha, *Role of ck1 in gsk3β-mediated phosphorylation and degradation of snail*, Oncogene, 29 (2010), pp. 3124–3133.

[283] A. Yadav, B. Kumar, J. Datta, T. N. Teknos, and P. Kumar, *Il-6 promotes head and neck tumor metastasis by inducing epithelial–mesenchymal transition via the jak-stat3-snail signaling pathway*, Molecular Cancer Research, 9 (2011), pp. 1658–1667.

[284] M. Yanagisawa, D. Huveldt, P. Kreinest, C. M. Lohse, J. C. Cheville, A. S. Parker, J. A. Copland, and P. Z. Anastasiadis, *A p120 catenin isoform switch affects rho activity, induces tumor cell invasion, and predicts metastatic disease*, Journal of Biological Chemistry, 283 (2008), pp. 18344–18354.

[285] J. Yang and R. A. Weinberg, *Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis*, Developmental cell, 14 (2008), pp. 818–829.

[286] M.-H. Yang, D. S.-S. Hsu, H.-W. Wang, H.-J. Wang, H.-Y. Lan, W.-H. Yang, C.-H. Huang, S.-Y. Kao, C.-H. Tzeng, S.-K. Tai, et al., *Bmi1 is essential in twist1-induced epithelial-mesenchymal transition*, Nature cell biology, 12 (2010), pp. 982–992.

[287] S. Yang, L. Yuan, Y.-C. Lai, X. Shen, P. Wonka, and J. Ye, *Feature grouping and selection over an undirected graph*, in Graph Embedding for Pattern Analysis, Springer, 2013, pp. 27–43.

[288] Z. Yang, S. Rayala, D. Nguyen, R. K. Vadlamudi, S. Chen, and R. Kumar, *Pak1 phosphorylation of snail, a master regulator of epithelial-to-mesenchyme transition, modulates snail's subcellular localization and functions*, Cancer research, 65 (2005), pp. 3179–3184.

[289] A. S. Yap, W. M. Brieher, and B. M. Gumbiner, *Molecular and functional analysis of cadherin-based adherens junctions*, Annual review of cell and developmental biology, 13 (1997), pp. 119–146.

[290] S. Yi, S. Lin, Y. Li, W. Zhao, G. B. Mills, and N. Sahni, *Functional variomics and network perturbation: connecting genotype to phenotype in cancer*, Nature Reviews Genetics, 18 (2017), pp. 395–410.

[291] M. Yilmaz and G. Christofori, *Emt, the cytoskeleton, and cancer cell invasion*, Cancer and Metastasis Reviews, 28 (2009), pp. 15–33.

[292] Y. J. Yitan Zhu, Peng Qiu, *Tcga-assembler: open-source software for retrieving and processing tcga data*, Nature Methods, (2014), pp. 599 – 600.

[293] J. I. Yook, X.-Y. Li, I. Ota, E. R. Fearon, and S. J. Weiss, *Wnt-dependent regulation of the e-cadherin repressor snail*, Journal of Biological Chemistry, 280 (2005), pp. 11740–11748.

[294] J. I. Yook, X.-Y. Li, I. Ota, C. Hu, H. S. Kim, N. H. Kim, S. Y. Cha, J. K. Ryu, Y. J. Choi, J. Kim, et al., *A wnt–axin2–gsk3β cascade regulates snail1 activity in breast cancer cells*, Nature cell biology, 8 (2006), pp. 1398–1406.

[295] L. Yu and H. Liu, *Efficient feature selection via analysis of relevance and redundancy*, Journal of machine learning research, 5 (2004), pp. 1205–1224.

[296] Y. Yuan, E. M. Van Allen, L. Omberg, N. Wagle, A. Amin-Mansour, A. Sokolov, L. A. Byers, Y. Xu, K. R. Hess, L. Diao, et al., *Assessing the clinical utility of cancer genomic and proteomic data across tumor types*, Nature biotechnology, 32 (2014), pp. 644–652.

[297] O. L. e. a. Yuan Y, Van Allen EM, *Assessing the clinical utility of cancer genomic and proteomic data across tumor types*, Nature Biotechnology, (2014), pp. 644 – 652.

[298] K. Zhang, E. Rodriguez-Aznar, N. Yabuta, R. J. Owen, J. M. Mingot, H. Nojima, M. A. Nieto, and G. D. Longmore, *Lats2 kinase potentiates snail1 activity by promoting nuclear retention upon phosphorylation*, The EMBO journal, 31 (2012), pp. 29–43.

[299] W. Zhang, T. Ota, V. Shridhar, J. Chien, B. Wu, and R. Kuang, *Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment*, PLoS Comput Biol, 9 (2013), p. e1002975.

[300] C. R. R. H. Zhang Y, Xuan J, *Module-based breast cancer classification*, International Journal of Data Mining and Bioinformatics, 7 (2013), pp. 284–302.

[301] J. Zhao, D. Dong, L. Sun, G. Zhang, and L. Sun, *Prognostic significance of the epithelial-to-mesenchymal transition markers e-cadherin, vimentin and twist in bladder cancer*, International braz j urol, 40 (2014), pp. 179–189.

[302] P. Zhao and B. Yu, *On model selection consistency of lasso*, Journal of Machine learning research, 7 (2006), pp. 2541–2563.

[303] Q. Zhao, X. Shi, Y. Xie, J. Huang, B. Shia, and S. Ma, *Combining multidimensional genomic measurements for predicting cancer prognosis: observations from tcga*, Briefings in bioinformatics, 16 (2014), pp. 291–303.

[304] X.-P. Zhao, H. Zhang, J.-Y. Jiao, D.-X. Tang, Y.-l. Wu, and C.-B. Pan, *Overexpression of hmga2 promotes tongue cancer metastasis through emt pathway*, Journal of translational medicine, 14 (2016), p. 26.

[305] S. Zheng, A. D. Cherniack, N. Dewal, R. A. Moffitt, L. Danilova, B. A. Murray, A. M. Lerario, T. Else, T. A. Knijnenburg, G. Ciriello, et al., *Comprehensive pan-genomic characterization of adrenocortical carcinoma*, Cancer cell, 29 (2016), pp. 723–736.

[306] B. P. Zhou, J. Deng, W. Xia, J. Xu, Y. M. Li, M. Gunduz, and M.-C. Hung, *Dual regulation of snail by gsk-3β-mediated phosphorylation in control of epithelial–mesenchymal transition*, Nature cell biology, 6 (2004), pp. 931–940.

[307] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, *Learning with local and global consistency*, in Advances in neural information processing systems, 2004, pp. 321–328.

[308] J. Zhu, P. Sova, Q. Xu, K. M. Dombek, E. Y. Xu, H. Vu, Z. Tu, R. B. Brem, R. E. Bumgarner, and E. E. Schadt, *Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation*, PLoS biology, 10 (2012), p. e1001301.

Chapter 7

Appendix

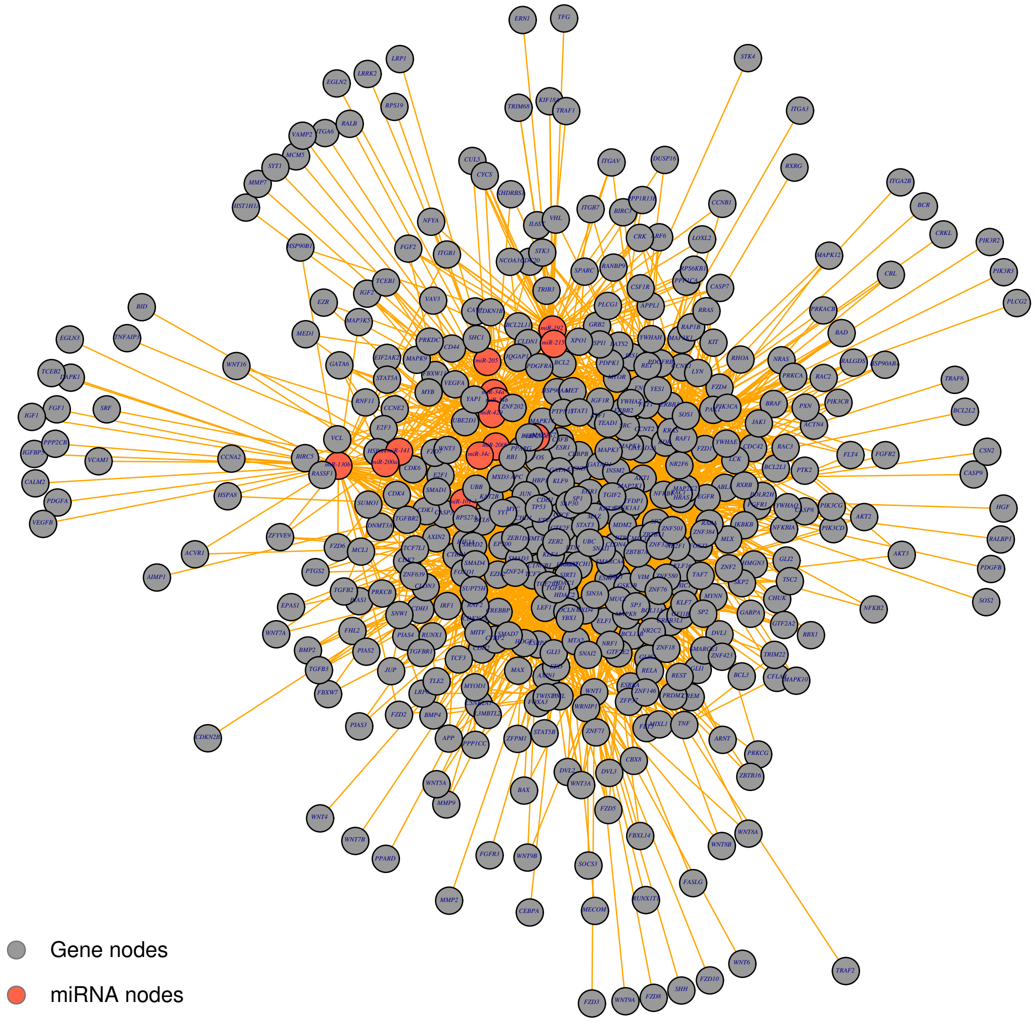


Figure 7.1: Extended EMT Network. We extended the EMT core network to incorporate the molecules that interact with or being regulated by the genes and miRNAs in the core network. We referred to three databases: STRING protein-protein interactions, ENCODE transcription factor - gene regulations, and miRTarBase miRNA-gene regulations.

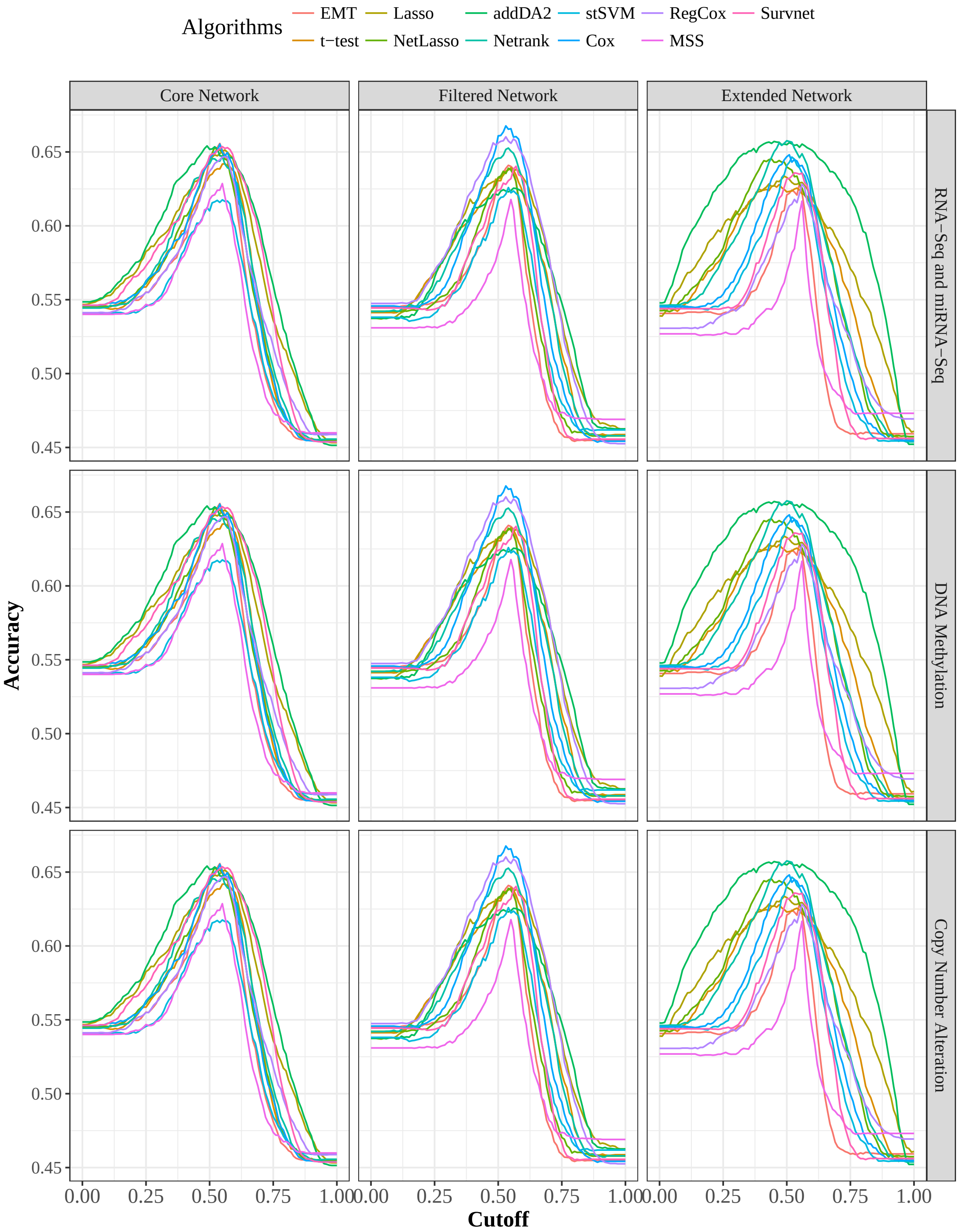


Figure 7.2: The accuracies at different cutoffs of each feature selection algorithm for each combination of data levels and networks using SVM classifier.

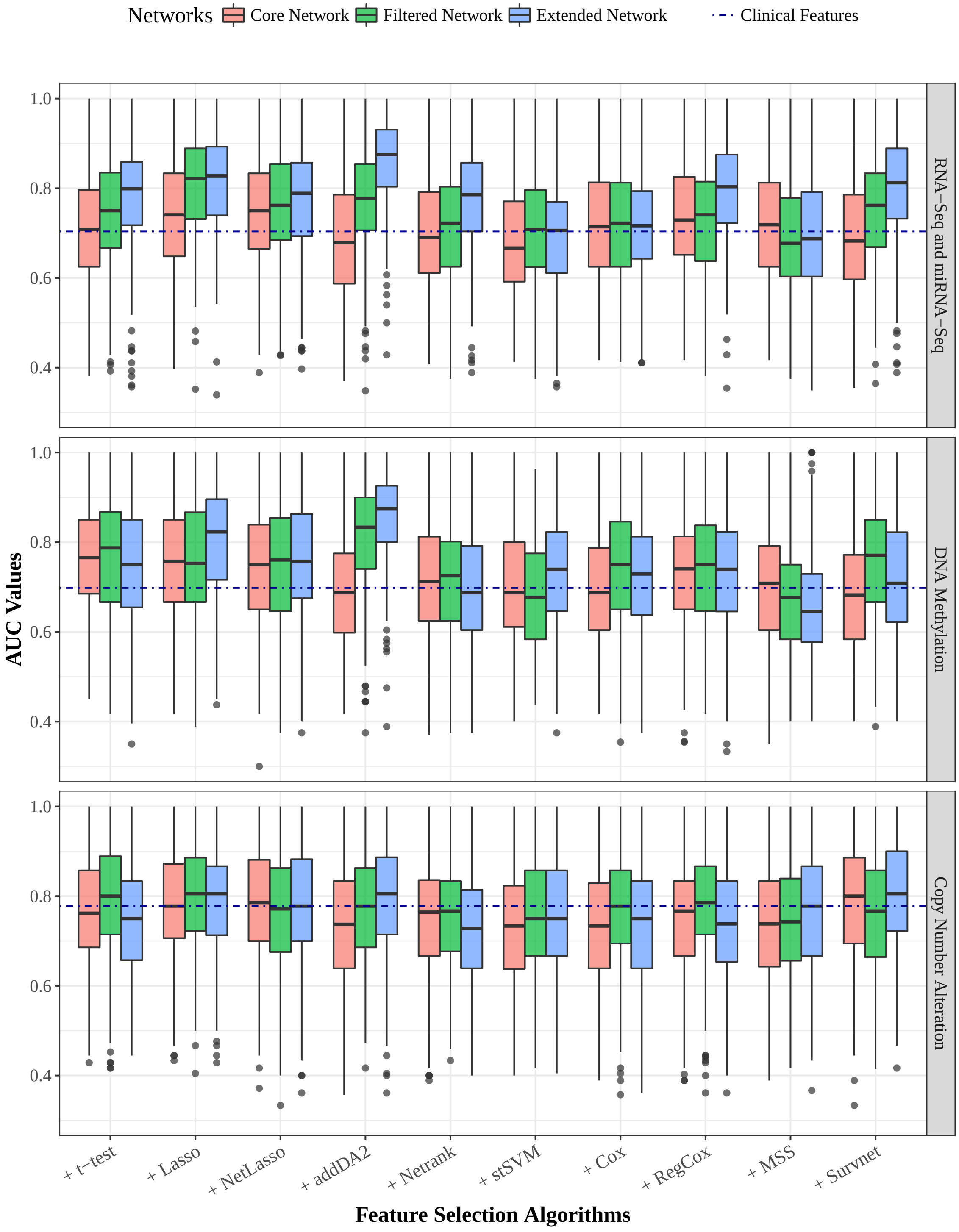


Figure 7.3: The average AUC values of FSFs combined with clinical features for three data levels and three EMT networks using random forest classifier. The blue dotted lines show the median AUC values of using only clinical features.

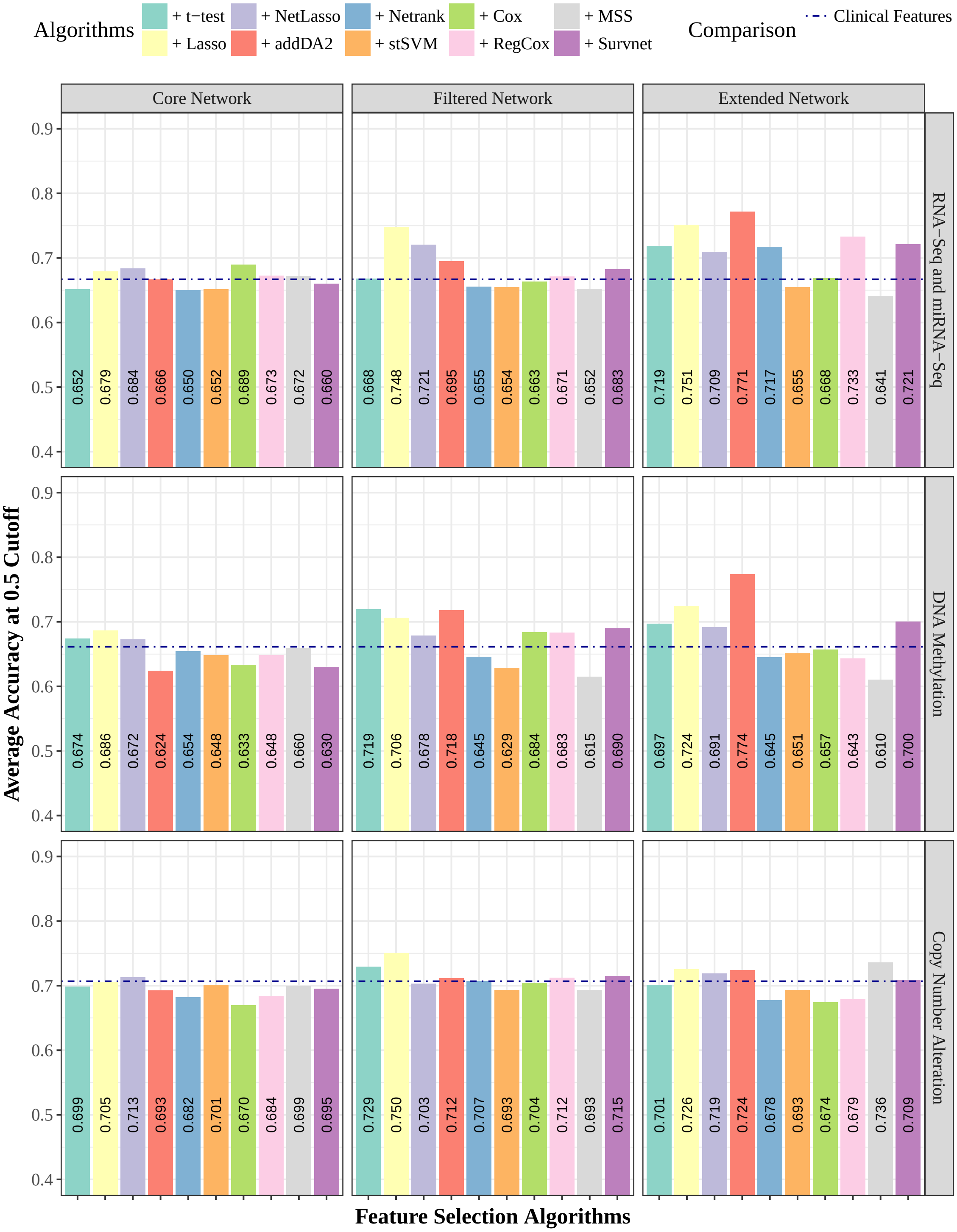


Figure 7.4: The average prediction accuracy at 0.5 cutoff of FSFs combined with clinical features for three data levels and three EMT networks using random forest classifier. The blue dotted lines are the median accuracy values of using only clinical features.

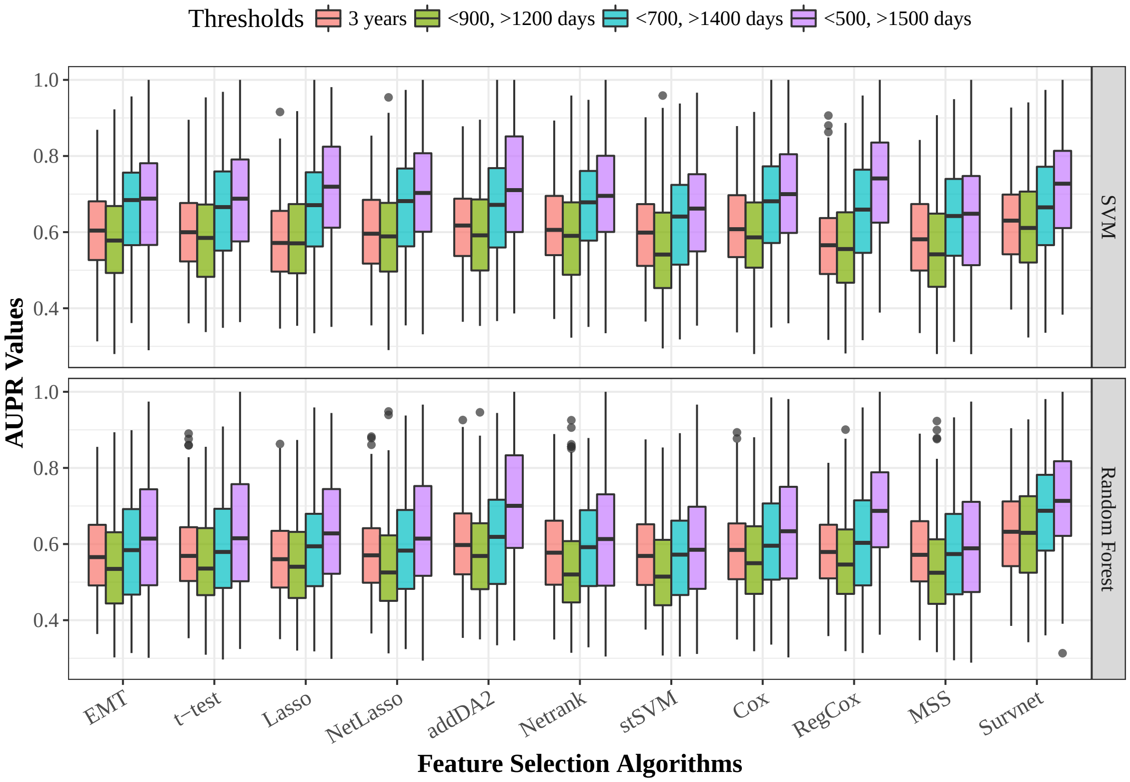


Figure 7.5: The AUPR values of 10 feature selection algorithms using different thresholds for good prognosis class and poor prognosis class. The data level is DNA methylation data. The network is EMT core network.

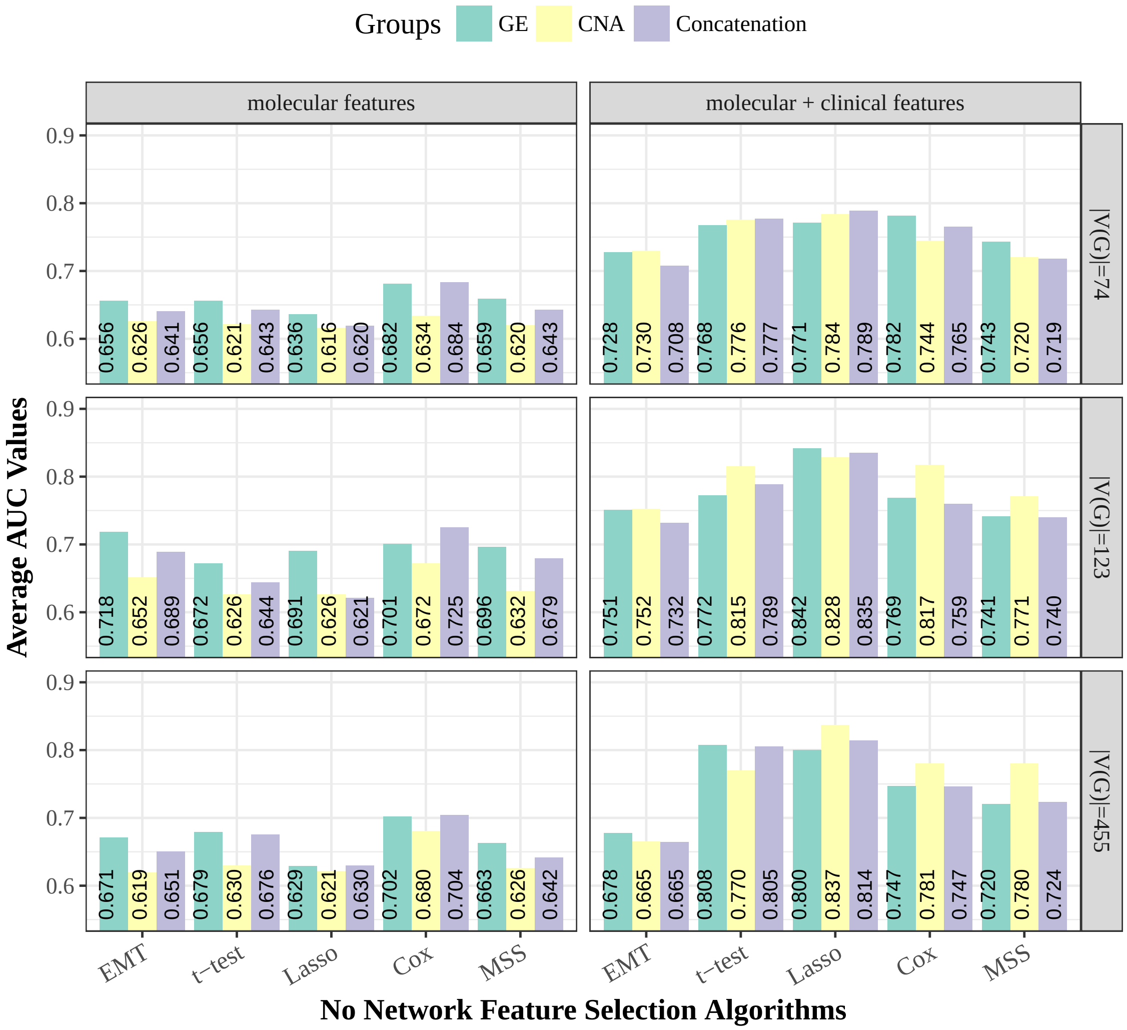


Figure 7.6: The average AUC values of concatenation-based feature selection using GE and CNA data. The left side shows the performance of molecular features. The right side shows the performance of FSFs combined with clinical features. On each side we show the results of using 3 EMT networks.

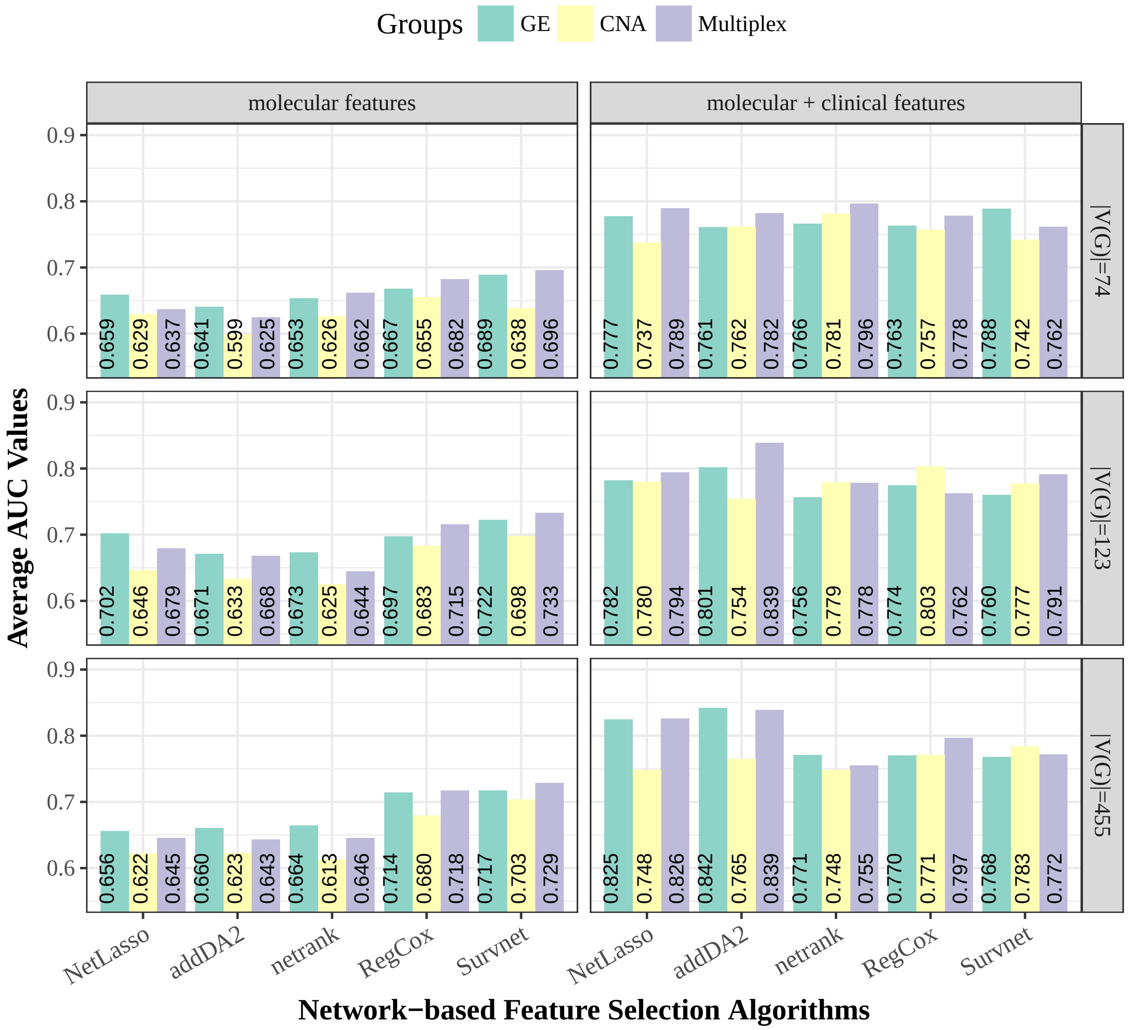


Figure 7.7: The average AUC values of multiplex-based feature selection using GE and CNA data. The left side shows the performance of molecular features. The right side shows the performance of FSFs combined with clinical features. On each side we show the results of using 3 EMT networks.

Table 7.1: Gene regulatory interactions in the core EMT network.

SNAI1 represses CDH1	miR-141 represses ZEB1	TGFβ1 increases DAB2
SNAI2 represses CDH1	miR-429 represses ZEB1	CK1 phosphorylates SNAI1
ZEB1 represses CDH1	miR-200a represses ZEB2	GSK3β phosphorylates SNAI1
ZEB2 represses CDH1	miR-200b represses ZEB2	βTRCP1 degrades SNAI1
TCF3 represses CDH1	miR-200c represses ZEB2	TNFa stabilizes SNAI1
SNAI1 induces CDH2	miR-141 represses ZEB2	MDM2 degrades SNAI1
SNAI2 induces CDH2	miR-429 represses ZEB2	MDM2 degrades SNAI2
ZEB1 induces CDH2	miR-205 represses ZEB1	FBXL14 degrades SNAI1
SNAI1 induces FN1	miR-205 represses ZEB2	FBXL14 degrades SNAI2
SNAI1 represses MUC1	miR130b represses ZEB1	PPA1 degrades SNAI1
ZEB1 represses MUC1	ZEB1 represses miR-141	PPA1 degrades SNAI2
SNAI1 induces LEF1	ZEB1 represses miR-200c	PPA1 degrades TWIST1
SNAI1 induces ZEB1	ZEB1 represses miR-200a	PPA1 degrades ZEB2
SNAI2 induces VIM	ZEB1 represses miR-200b	PPA2 degrades SNAI1
HRAS induces VIM	ZEB1 represses miR-429	PPA2 degrades SNAI2
SOS interacts HRAS	ZEB2 represses miR-141	PPA2 degrades TWIST1
SOS interacts KRAS	ZEB2 represses miR-200c	PPA2 degrades ZEB2
KRAS interacts RAF1	ZEB2 represses miR-200a	PAK1 stabilizes SNAI1
HRAS interacts RAF1	ZEB2 represses miR-200b	LOXL2 stabilizes SNAI1
RAF1 interacts MEK1	ZEB2 represses miR-429	LATS2 stabilizes SNAI1
RAF1 interacts MEK2	SNAI1 represses miR-34a	PRKD1 nucleus_exports SNAI1
MEK1 interacts ERK2	SNAI1 represses miR-34b	β-catenin activates LEF1
MEK2 interacts ERK2	SNAI1 represses miR-34c	TGFβ1 phosphorylates SMAD2
ERK2 induces ZEB1	miR128-1 represses BMI1	TGFβ1 phosphorylates SMAD3
ERK2 induces ZEB2	miR128-2 represses BMI1	SMAD2 interacts SMAD4
ERK2 induces SNAI1	miR-200c target BMI1	SMAD3 interacts SMAD4
ERK2 induces SNAI2	miR-203 target BMI1	SMAD3 activates HMGA2
ERK2 interacts EGR1	MIR101-1 represses EZH2	SMAD4 activates HMGA2
EGR1 induces SNAI1	SNAI1 represses ESRP1	HMGA2 induces SNAI1
SNAI1 represses cytokeratin_18	ZEB1 represses ESRP1	SMAD3 interacts ETS1
SNAI1 represses claudin_3	ZEB2 represses ESRP1	ETS1 increases ZEB1
SNAI1 represses claudin_4	SNAI1 represses ESRP2	SMAD2 inhibits ID2
SNAI1 represses occludin	ZEB1 represses ESRP2	SMAD3 inhibits ID2
ZEB2 represses claudin_4	ZEB2 represses ESRP2	ID2 inhibits TCF3
ZEB2 represses P_cadherin	SNAI1 interacts DNMT1	TCF3 increases SNAI1
p53 induces miR-34a	DNMT1 represses CDH1	miR-200b inhibits JAG1
p53 induces miR-34b	SNAI1 interacts HDAC1	miR-200b inhibits JAG2
p53 induces miR-34c	HDAC1 represses CDH1	JAG1 activates NOTCH
miR-34a represses SNAI1	SNAI1 interacts HDAC2	JAG2 activates NOTCH
miR-34b represses SNAI1	HDAC2 represses CDH1	NOTCH increases SNAI1
miR-34c represses SNAI1	ZEB1 interacts SIRT	NOTCH increases SNAI2
p53 induces miR-200a	SIRT represses CDH1	NOTCH increases LOXL2
p53 induces miR-200b	ZEB1 interacts BRG1	EGF activates JAK
p53 induces miR-200c	BRG1 represses CDH1	EGF activates PI3K
p53 induces miR-141	TWIST1 interacts EZH2	PI3K activates AKT2
p53 induces miR-429	EZH2 represses CDH1	AKT2 increases SNAI1
p53 induces miR-192	TWIST1 interacts BMI1	AKT2 increases SNAI2
p53 induces miR-215	BMI1 represses CDH1	JAK phosphorylates STAT3
miR-192 represses ZEB2	YB1 increases SNAI1	STAT3 increases TWIST1
miR-215 represses ZEB2	YB1 increases ZEB2	AKT2 inhibits GSK3β
miR-200a represses ZEB1	YB1 increases LEF1	GSK3β phosphorylates β-catenin
miR-200b represses ZEB1	YB1 increases TWIST1	wnt1 inhibits GSK3β
miR-200c represses ZEB1	TGFβ1 increases ILEI	

Table 7.2: The average AUC and AUPR values of the 10 feature selection algorithms using random forest classifier. For each algorithm, we evaluated its prediction performance using 3 data levels. Within each data level, we used 3 different sizes of EMT networks.

Data Level	Gene expression			DNA Methylation			CNA			Metric
$\lvert V(G)\rvert$	74	123	455	74	123	455	70	117	445	
EMT	0.631	0.709	0.700	0.634	0.643	0.655	0.621	0.626	0.617	AUC
	0.590	0.687	0.672	0.586	0.599	0.629	0.595	0.615	0.590	AUPR
t-test	0.630	0.714	0.704	0.635	0.647	0.655	0.621	0.637	0.606	AUC
	0.583	0.696	0.677	0.587	0.590	0.617	0.597	0.618	0.551	AUPR
Lasso	0.617	0.688	0.639	0.640	0.633	0.649	0.625	0.612	0.614	AUC
	0.562	0.662	0.597	0.591	0.584	0.621	0.577	0.567	0.543	AUPR
NetLasso	0.628	0.697	0.695	0.634	0.642	0.660	0.627	0.631	0.616	AUC
	0.579	0.675	0.680	0.589	0.594	0.636	0.591	0.598	0.572	AUPR
addDA2	0.640	0.685	0.665	0.655	0.655	0.704	0.609	0.631	0.614	AUC
	0.583	0.637	0.630	0.611	0.606	0.676	0.547	0.612	0.584	AUPR
Netrank	0.640	0.697	0.700	0.637	0.639	0.644	0.632	0.623	0.615	AUC
	0.586	0.679	0.660	0.592	0.581	0.599	0.596	0.601	0.570	AUPR
stSVM	0.630	0.674	0.642	0.627	0.644	0.638	0.623	0.606	0.615	AUC
	0.595	0.646	0.596	0.567	0.570	0.579	0.593	0.588	0.570	AUPR
Cox	0.648	0.694	0.704	0.653	0.677	0.697	0.626	0.642	0.646	AUC
	0.607	0.659	0.655	0.605	0.628	0.653	0.599	0.601	0.626	AUPR
RegCox	0.667	0.694	0.721	0.668	0.679	0.661	0.636	0.636	0.635	AUC
	0.631	0.661	0.690	0.610	0.633	0.589	0.603	0.602	0.596	AUPR
MSS	0.634	0.685	0.656	0.633	0.634	0.631	0.615	0.626	0.619	AUC
	0.598	0.667	0.619	0.580	0.576	0.581	0.585	0.588	0.580	AUPR
Survnet	0.667	0.679	0.663	0.711	0.683	0.659	0.631	0.677	0.670	AUC
	0.652	0.661	0.632	0.681	0.649	0.618	0.596	0.652	0.633	AUPR

Table 7.3: The average AUC and AUPR values of FSFs together with clinical features using random forest classifier. For each algorithm, we evaluated its prediction performance using 3 data levels. On each data level, we used 3 different sizes of EMT networks.

Data Level	Gene expression			DNA Methylation			CNA			Metric
$\|V(G)\|$	74	123	455	74	123	455	70	117	445	
clinical	0.710			0.696			0.771			AUC
	0.641			0.612			0.750			AUPR
t-test	0.706	0.741	0.782	0.759	0.769	0.750	0.761	0.789	0.745	AUC
	0.677	0.717	0.763	0.671	0.705	0.680	0.724	0.746	0.691	AUPR
Lasso	0.733	0.802	0.812	0.749	0.756	0.803	0.781	0.800	0.788	AUC
	0.700	0.779	0.798	0.683	0.697	0.743	0.727	0.757	0.744	AUPR
NetLasso	0.740	0.764	0.771	0.742	0.745	0.755	0.783	0.765	0.779	AUC
	0.713	0.725	0.732	0.666	0.675	0.699	0.746	0.722	0.754	AUPR
addDA2	0.686	0.773	0.858	0.689	0.813	0.854	0.732	0.771	0.791	AUC
	0.627	0.727	0.853	0.573	0.776	0.799	0.713	0.729	0.770	AUPR
Netrank	0.700	0.716	0.773	0.710	0.712	0.700	0.753	0.760	0.725	AUC
	0.653	0.694	0.755	0.599	0.629	0.600	0.719	0.703	0.688	AUPR
stSVM	0.685	0.704	0.699	0.700	0.680	0.729	0.728	0.748	0.752	AUC
	0.652	0.698	0.661	0.611	0.583	0.633	0.656	0.710	0.714	AUPR
Cox	0.720	0.717	0.717	0.701	0.737	0.721	0.725	0.772	0.733	AUC
	0.682	0.672	0.653	0.588	0.668	0.673	0.678	0.720	0.713	AUPR
RegCox	0.728	0.725	0.791	0.733	0.735	0.727	0.750	0.781	0.737	AUC
	0.679	0.668	0.746	0.662	0.657	0.635	0.706	0.743	0.693	AUPR
MSS	0.717	0.687	0.695	0.709	0.678	0.659	0.739	0.744	0.769	AUC
	0.682	0.648	0.650	0.642	0.565	0.549	0.700	0.727	0.751	AUPR
Survnet	0.688	0.753	0.799	0.683	0.760	0.714	0.778	0.756	0.806	AUC
	0.637	0.717	0.792	0.562	0.699	0.677	0.744	0.723	0.787	AUPR

Table 7.4: The summary of association rules inferred from the FSFs of 10 feature selection algorithms and the ensemble algorithm. We have counted the total number of rules with $confidence \geq 0.8$ and $support \geq 0.1$, and measured their average confidence, lift, and length.

Data levels	t-test	Lasso	NetLasso	addDA2	Netrank	stSVM	Cox	RegCox	MSS	Survnet	Ensemble	Metric
GE	343	179	95	38	231	105	209	81	162	212	212	#rules
	0.93	0.91	0.85	0.84	0.92	0.89	0.88	0.86	0.89	0.9	0.92	confidence
	1.85	1.8	1.69	1.66	1.84	1.79	1.78	1.72	1.77	1.78	1.79	lift
	3.78	3.49	4.21	3.24	3.75	3.98	4.02	3.48	3.87	3.97	3.53	length
DM	219	86	58	9	364	158	359	139	41	2	157	#rules
	0.9	0.86	0.84	0.85	0.94	0.89	0.95	0.9	0.84	0.83	0.91	confidence
	1.79	1.7	1.69	1.71	1.85	1.76	1.92	1.81	1.7	1.66	1.8	lift
	3.87	3.52	3.57	3.56	3.82	4.26	3.94	3.57	3.63	4.00	3.53	length
CNA	204	58	64	48	183	37	60	77	22	34	109	#rules
	0.9	0.84	0.87	0.85	0.89	0.84	0.84	0.86	0.83	0.85	0.87	confidence
	1.77	1.7	1.71	1.67	1.75	1.65	1.67	1.7	1.66	1.69	1.71	lift
	3.29	3.97	3.25	3.02	3.31	3.19	3.43	3.64	3.27	3.47	3.46	length
GE+DM	184	217	86	57	247	29	181	215	243	31	229	#rules
	0.91	0.92	0.87	0.87	0.91	0.84	0.91	0.91	0.89	0.84	0.91	confidence
	1.8	1.83	1.74	1.77	1.8	1.68	1.85	1.81	1.82	1.67	1.82	lift
	3.73	3.54	3.66	3.61	3.68	3.52	3.59	3.42	4.30	3.45	3.65	length
GE+CNA	334	284	68	92	362	9	204	139	73	48	303	#rules
	0.94	0.93	0.86	0.87	0.94	0.82	0.89	0.89	0.85	0.83	0.9	confidence
	1.84	1.89	1.73	1.73	1.84	1.65	1.76	1.78	1.7	1.65	1.79	lift
	3.70	3.69	3.37	3.49	3.65	3.22	3.50	3.53	3.86	3.69	3.61	length
DM+CNA	351	198	62	32	298	22	267	212	61	51	218	#rules
	0.94	0.89	0.84	0.85	0.92	0.85	0.93	0.92	0.85	0.85	0.92	confidence
	1.84	1.76	1.68	1.7	1.8	1.67	1.84	1.82	1.72	1.68	1.82	lift
	3.56	3.43	3.16	3.28	3.47	3.32	3.53	3.51	3.75	3.10	3.44	length
GE+DM +CNA	382	427	167	84	318	31	188	363	250	32	524	#rules
	0.95	0.95	0.89	0.86	0.95	0.85	0.9	0.94	0.91	0.83	0.95	confidence
	1.87	1.91	1.81	1.72	1.86	1.7	1.77	1.89	1.86	1.66	1.89	lift
	3.54	3.61	3.54	3.44	3.60	3.55	3.43	3.48	4.03	3.06	3.56	length

Table 7.5: The p-values of log-rank tests based on the clustering of spectral clustering algorithm for different data level combinations using extended EMT network. We highlighted all p-values that are lower than 10e-5.

	GE	DM	CNA	GE+DM	GE+CNA	DM+CNA	GE+DM +CNA
t-test	1.19e-5	1.18e-1	6.39e-2	8.16e-1	**9.51e-7**	3.85e-1	1.86e-5
Lasso	2.34e-4	6.78e-1	7.67e-1	4.92e-1	1.76e-5	**1.85e-6**	**2.69e-6**
NetLasso	3.36e-2	8.61e-1	5.25e-1	2.96e-1	2.46e-2	2.61e-1	8.03e-1
addDA2	2.16e-5	4.55e-1	3.00e-4	**8.44e-13**	**3.09e-6**	**1.61e-7**	1.45e-5
Netrank	4.89e-03	1.82e-01	8.85e-01	5.30e-02	2.11e-02	1.62e-02	1.15e-01
stSVM	1.64e-01	5.22e-01	5.77e-01	5.62e-01	2.20e-01	7.45e-01	7.92e-01
Cox	3.49e-03	1.36e-04	1.04e-02	4.45e-05	1.34e-04	1.49e-05	2.10e-01
RegCox	1.41e-04	6.53e-03	1.60e-03	1.20e-01	7.63e-02	2.89e-01	4.07e-01
MSS	1.72e-03	8.71e-01	1.12e-01	4.51e-03	6.32e-04	1.59e-01	1.44e-02
Survnet	3.89e-05	2.48e-02	2.30e-01	9.20e-04	6.13e-04	1.23e-03	2.14e-05
Ensemble	1.40e-03	9.75e-01	8.37e-02	3.95e-03	**8.12e-07**	**2.42e-09**	**1.68e-08**
allemt	5.78e-03	8.37e-01	3.46e-01	9.58e-01	9.36e-03	7.79e-01	4.02e-01

Zusammenfassung

Netzwerke können als Vorwissen verwendet werden, um molekulare Signaturen zu identifizieren. Die meisten Studien verwenden PPI-Netzwerke und Genexpressionsdaten, wobei PPI-Netzwerke häufig verwendet werden, um die wichtige Features oder Unternetzwerke als Signaturen zu finden. Da immer mehr der zahlreichen Arten von Omics-Daten zur Verfügung stehen, stellt sich die Frage, welche Art von Daten bessere molekulare Signaturen hervorbringen können und ob es vorteilhaft ist, mehrere Omics-Daten gleichzeitig zu verwenden, um bessere Signaturen zu identifizieren. Dies sind die zwei Hauptfragen, die in dieser Dissertation behandelt werden.

Wir untersuchen dieses Thema mit dem Anwendungsfall der Vorhersage von Krebsprognosen. Nach einer umfassenden Literaturrecherche zu Feature Selection Algorithmen wählen wir 10 Algorithmen aus. Fünf von diesen integrieren Netzwerkinformationen, die anderen betrachten die Omics-Daten ohne Zunahme eines Netzwerks. Auf Einzeldatenebene führen wir die Feature Selection alternativ auf drei Datenebenen durch - mRNA- und miRNA-Expression, DNA-Methylierung und Kopienzahlvariation aus. Dann analysieren wir die Vorhersageleistung dieser Features auf ihre Stabilität, Netzwerkeigenschaften und biologische Interpretation. Um Multi-Omics-Signaturen zu erhalten, kombinieren wir zuerst die ausgewählten Features aus einzelnen Datenebenen und testen deren Vorhersageleistung. Dann erweiterten wir netzwerkbasierte Feature Selection Algorithmen, um mehrere Datenebenen unter Verwendung einer Multiplexstruktur zu integrieren. Die Feature Selection Algorithmen ohne Netzwerk hingegen werden auf kombinierten Daten angewendet. Schließlich untersuchen wir die Vorhersageleistung von Single-Omics- und Multi-Omics-Signaturen an einer Reihe unabhängiger Stichproben mit zwei Omics-Datenebenen.

Das kritische Thema bei der Entdeckung von Biomarkern ist sog. Curse of dimensionality, der eine geringe Reproduzierbarkeit der molekularen Signaturen verursachen kann. Selbst mit netzwerkbasierten Feature Selection Algorithmen wurden signifikante Verbesserungen nicht erreicht, da die Netzwerke oft groß sind. Um dieses Problem zu lösen, schlagen wir vor, Phänotyp-relevante Genregulationsnetzwerke basierend auf epithelial-mesenchymale Transition (EMT) zu benutzen, die sich als äußerst relevant für Krebsmetastasen und -prognosen erweisen. Dann haben wir verschiedene Arten von Omics-Daten in das Netzwerk integriert, um prognostische Signaturen zu finden. Obwohl die Dimensionalität auf weniger als 2,5% des Originals reduziert ist, bieten EMT-Funktionen eine bessere Vorhersageleistung als die aus den Originaldaten ausgewählten Features. Häufig ausgewählte Features erreichen durchschnittliche Vorhersage-AUC-Werte von über 0,8. Diese Features sind in der Lage, Stichproben in signifikant unterschiedliche prognostische Gruppen sowohl auf den Trainingsdaten als auch auf den unabhängigen Testdaten zu stratifizieren. Die Verwendung kombinierter Features aus mehreren Omics-Ebenen und die Verwendung einer multiplexbasierten Feature-Auswahl verbessern die Vorhersage weiter.

Darüber hinaus haben wir relationale und NoSQL Datenbanken zur Integration in die Multi-Omics-Datenanalyse verglichen. Da biologische Daten große Volumen, hohe Dynamik und große Vielfalt aufweisen, ist es notwendig, ein Datenbanksystem zu haben, das Daten effizient speichern und abrufen kann. Basierend auf den Ergebnissen haben wir notwendige Bedingungen zum Aufbau eine skalierbaren Omics-Dateninfrastrukturen abgeleitet.