INTEGRATIVE ANALYSIS OF MULTIMODAL BIOMEDICAL DATA WITH MACHINE LEARNING

by

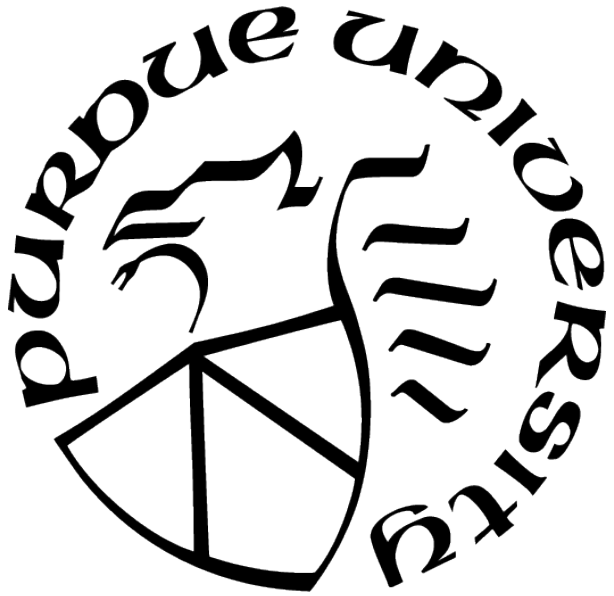
Zhi Huang

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy



School of Electrical and Computer Engineering

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Edward J. Delp, Co-chair

School of Electrical and Computer Engineering

Dr. Paul Salama, Co-chair

Department of Electrical and Computer Engineering, IUPUI

Dr. Kun Huang

Indiana University School of Medicine, Indiana University

Dr. Maher E. Rizkalla

Department of Electrical and Computer Engineering, IUPUI

Dr. Thomas M. Talavage

School of Electrical and Computer Engineering

Approved by:

Dr. Dimitrios Peroulis

Dedication

For my advisors, Dr. Paul Salama and Dr. Kun Huang, whose relentless support made this work possible.

ACKNOWLEDGMENTS

Upon the completion of this thesis, I realized that my limited words could not express my appreciation and gratitude toward those who have motivated and encouraged me through my doctorate study at Purdue University. I have had a distinct pleasure of being surrounded by my professors and colleagues.

I am grateful for having Professor Paul Salama who serves as my committee co-chair. This thesis would not be accomplished without his guidance and help. He shares numerous advice and comments on my research through weekly meetings. I would not be able to have current achievements without his instruction and inspiration in my academic research and life. He also puts a lot of time to shape my critical thinking and personality, which makes me become a better engineer. The doctoral study experience with him is the enduring treasure of my life.

I would like to thank my doctoral advisor Professor Kun Huang for his encouragement, guidance, and support for my Ph.D. program. This thesis would not be accomplished if he had not given me the opportunity to be a part of his laboratory. His attention to detail, thoughtfulness, and analytical skills helped me form my research attitude. He teaches me how to communicate, perform research, and make presentations. I am grateful that I could work with him.

I would like to thank Professor Edward J. Delp who serves as my committee co-chair, for his guidance and suggestions. I truly appreciate his involvement in my doctoral study plan and how he shapes my critical thinking.

I would like to thank Professor Maher E. Rizkalla for his involvement in my research projects. He provides valuable input and feedback throughout my doctoral study.

I would also like to thank Professor Thomas M. Talavage for his valuable suggestions, discussions, and supports.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\boldsymbol{X}$	Data matrix.
$\boldsymbol{W}$	Basis matrix.
$\boldsymbol{H}$	Coefficient matrix.
$\boldsymbol{I}$	Identity matrix.
P	Number of features/covariates.
N	Number of patients/samples.
K	Low-rank dimensions.
Y	Survival time (vector).
C	Survival event (vector). 1: deceased; 0: censored.
T	Survival time (random variable).
t	Survival time. Observation of random variable T.
Δ	A small (time) interval.
T_E	Random variable stands for time to event in survival analysis.
T_C	Random variable stands for time to censoring event in survival analysis.
t_E	Observation of random variable T_E.
t_C	Observation of random variable T_C.
δ_{i}	Indicator. Equals to 1 if $t_E^{(\mathrm{i})} < t_C^{(\mathrm{i})}$, 0 otherwise.
n_{i}	Number at risk at t_{i}.
c_{i}	Number of censored patients in the time interval $[t_{\mathrm{i}}, t_{\mathrm{i}+1})$.
d_{i}	Number of events at t_{i}.
h_{i}	Probability of death occurred at t_{i} given survived to $t_{\mathrm{i}}-$.
β	Coefficient for Cox proportional hazards regression.
$\tilde{\boldsymbol{W}}$	Weighted basis matrix.
L	General likelihood function.
ℓ	Log likelihood function.
$f(t)$	Probability density function in survival analysis.
$F(t)$	Cumulative distribution function in survival analysis.
$S(t)$	Survival function.

$h(t)$	Hazard function.
$H(t)$	Cumulative hazard function.
$\hat{\mathbb{S}}$	Kaplan-Meier estimator.
$\lambda(t)$	Hazard function in survival analysis.
$\lambda_0(t)$	Baseline hazard function in survival analysis.
$\mathcal{R}$	Risk set in survival analysis.
$\mathbb{1}$	Indicator function.
C-Index	Concordance index.
$\chi^2(\cdot)$	Chi-square test.
O	Observed value in Chi-square test.
E	Expected value in Chi-square test.
N	Number of censored samples in Chi-square test.
dof	Degree of freedom.
LR	Log-rank test statistics.
$D(\cdot)$	Kullback-Leibler divergence.
$Tr(\cdot)$	Trace.
M	Number of the iterations.
$\mathrm{i}ter$	Iteration.
tol	CoxNMF concordance index tolerance.
α	Step size for $\boldsymbol{H}$ in CoxNMF algorithm.
γ	L1 penalty ratio $\in [0, 1]$.
ε	Parameter in exponential distribution for CoxNMF simulation noise.
$\boldsymbol{E}$	Artificial noise introduced for synthetic data.
ζ	Parameter in exponential distribution for constructing synthetic data in CoxNMF simulation.
κ	Parameter for smoothed L1 norm.
ξ	Regularization weight.
$\mathcal{H}$	Hessian matrix. $K \times K$ matrix.
g	Gradient of log partial likelihood. $K \times 1$ vector.

m_{j}	Number of the death count at Y_{j}.
τ_1	Number of worse prognosis bases in CoxNMF simulation.
τ_2	Number of better prognosis bases in CoxNMF simulation.
Φ_{i}	Gene cluster indicator.
ϕ	Gene cluster/module.
$\odot$	Hadamard product.
∂	Partial derivative.
∇	Gradient.
η	Learning rate in non-negative matrix factorization.
$\left\|\cdot\right\|_1$	Norm 1.
$\left\|\cdot\right\|_2$	Norm 2, or Euclidean norm.
$\left\|\cdot\right\|_2^2$	Squared Euclidean norm.
$\left\|\cdot\right\|_F$	Frobenius norm.
Ω	Definition of cluster in silhouette score.
$\tilde{s}$	Silhouette score.
$\mathcal{U}$	Uniform distribution.
r_s	Spearman's Rank Correlation Coefficient.
ρ	Correlation coefficient.
Θ	Network weights for SALMON algorithm.
TP	True positive.
TN	True negative.
FP	False positive.
FN	False negative.
$Dice$	Dice coefficient.
J	Threshold for gene data filtering.
G	Undirected weighted network in lmQCM algorithm.
V_G	Vertice set of undirected weighted network G in lmQCM algorithm.
E_G	Edge set of undirected weighted network G in lmQCM algorithm.

W_G	Non-negative weight set of undirected weighted network G in lmQCM algorithm.
d_G	density of undirected weighted network G in lmQCM algorithm.
q_γ	Parameter γ in lmQCM algorithm.
q_λ	Parameter λ in lmQCM algorithm.
q_t	Parameter t in lmQCM algorithm.
q_β	Parameter β in lmQCM algorithm.
s_α	Parameter in Supervised NMF for linear regression.
s_β	Parameter in Supervised NMF for regularization.
s_γ	Parameter in Supervised NMF for regularization.
J_1	Threshold for gene expression mean values.
J_2	Threshold for gene expression variance values.

ABBREVIATIONS

CT	Computed tomography
MRI	Magnetic resonance imaging
NMF	Non-negative matrix factorization
TPM	Transcripts per million
RSEM	RNA-seq by expectation-maximization
PCA	Principal component analysis
SVD	Singular value decomposition
FA	Factor analysis
NNDSVD	Non-negative double singular value decomposition
CD	Coordinate descent
MU	Multiplicative update
SNMF	Supervised non-negative matrix factorization
TCGA	The cancer genome atlas
NGS	Next generation sequencing
SALMON	Survival Analysis Learning with Multi-Omics Neural Networks
OS	Overall survival
EFS	Event-free survival
ROC	Receiver operating characteristic
AUC	Area under the ROC curve
KL	Kullback-Leibler
KM	Kaplan-Meier
MLE	Maximum likelihood estimation
CNV	Copy number variation
CNB	Copy number burden
TMB	Tumor mutation burden
MAF	Mutation annotation format
CNB	Copy number burden
FDR	False discovery rate

PyPI	The python package index
TSUNAMI	Tools SUite for Network Analysis and MIning
GO	Gene ontology
BP	Biological process
GCN	Gene co-expression network
lmQCM	local maximal Quasi-Clique Merger
WGCNA	Weighted gene co-expression network analysis
chr	Chromosome
CDF	Cumulative distribution function
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger ribonucleic acid
miRNA	Micro ribonucleic acid
scRNA	Single cell ribonucleic acid
NCBI	National Center for Biotechnology Information
VMC	Vanderbilt Medical Center
NA	Not a number
NaN	Not a number
PCC	Pearson correlation coefficient
SCC	Spearman's rank correlation coefficient
CSV	Comma-separated values
TSV	Tab-separated values
hg	Human genomes
GRCh	Genome Reference Consortium for human
RFS	Random survival forest
ER	Estrogen receptor
PR	Progesterone receptor
CD4	Cluster of differentiation 4
CD8	Cluster of differentiation 8

KEGG	Kyoto Encyclopedia of Genes and Genomes
PWM	Position weight matrix
PPI	Protein-protein interaction
ENCODE	The Encyclopedia of DNA Elements
TF	Transcription factor
ChIP-seq	Chromatin immunoprecipitation sequencing
ChEA	Chromatin immunoprecipitation enrichment analysis
TRANSFAC	TRANScription FACtor database
miRTarBase	Database of MicroRNA-target interactions
ECM	Extracellular matrix
UCSC	University of California, Santa Cruz
P-value	The probability of test results observed, or more extreme, when the null hypothesis is true.
LU	Lower–upper
NAC	Neoadjuvant chemotherapy
TIL	Tumor infiltrating lymphocyte
HER2	Human epidermal growth factor receptor 2
TNBC	Triple-negative breast cancer
AI	Artificial intelligence
WSI	Whole slide image
pCR	pathologic complete response
IMPRESS	IMage-based Pathological REgistration and Segmentation Statistics
H&E	Hematoxylin and eosin
IHC	Immunohistochemistry
PD-L1	Programmed death-ligand 1
CD8	Cluster of differentiation 8
CD163	Cluster of differentiation 163
FISH	Fluorescence in situ hybridization
ASCO	American Society of Clinical Oncology

CAP	College of American Pathologist
RCB	Residual cancer burden
LASSO	Least Absolute Shrinkage and Selection Operator
CEP17	Chromosome 17
ECM	Extracellular matrix
TME	Tumor micro-environment
TIME	Tumor immune micro-environment
EMT	Epithelial-mesenchymal transition
FAO	Fatty acid beta-oxidation

ABSTRACT

With the rapid development in high-throughput technologies and the next generation sequencing (NGS) during the past decades, the bottleneck for advances in computational biology and bioinformatics research has shifted from data collection to data analysis. As one of the central goals in precision health, understanding and interpreting high-dimensional biomedical data is of major interest in computational biology and bioinformatics domains. Since significant effort has been committed to harnessing biomedical data for multiple analyses, this thesis is aiming for developing new machine learning approaches to help discover and interpret the complex mechanisms and interactions behind the high dimensional features in biomedical data. Moreover, this thesis also studies the prediction of post-treatment response given histopathologic images with machine learning.

Capturing the important features behind the biomedical data can be achieved in many ways such as network and correlation analyses, dimensionality reduction, image processing, etc. In this thesis, we accomplish the computation through co-expression analysis, survival analysis, and matrix decomposition in supervised and unsupervised learning manners. We use co-expression analysis as upfront feature engineering, implement survival regression in deep learning to predict patient survival and discover associated factors. By integrating Cox proportional hazards regression into non-negative matrix factorization algorithm, the latent clusters of human genes are uncovered. Using machine learning and automatic feature extraction workflow, we extract thirty-six image features from histopathologic images, and use them to predict post-treatment response. In addition, a web portal written by R language is built in order to bring convenience to future biomedical studies and analyses.

In conclusion, driven by machine learning algorithms, this thesis focuses on the integrative analysis given multimodal biomedical data, especially the supervised cancer patient survival prognosis, the recognition of latent gene clusters, and the application of predicting post-treatment response from histopathologic images. The proposed computational algorithms present its superiority comparing to other state-of-the-art models, provide new insights toward the biomedical and cancer studies in the future.

1. INTRODUCTION

Analyzing and interpreting complex mechanisms behind multimodal biomedical data can be conducted with numerous advanced computational approaches. In this thesis, we limit our scope to the survival analysis and image analysis. In survival analysis, we aim to interpret the latent gene interactions by integrating multimodal biomedical data and performing dimensionality reduction. In image analysis, we develop an automatic feature extraction workflow and use thirty-six extracted features to predict post-treatment outcome. In this chapter, we will briefly introduce the multimodal biomedical data and data acquisition. The introduction is then followed with an overview of this thesis. The main contributions of this thesis and related publications are also listed.

1.1 Multimodal Biomedical Data and Data Acquisition

With the evolutionary development of the next generation sequencing (NGS) technique as well as the advanced computational algorithms, unraveling the latent biological interpretation behind multimodal high-dimensional biomedical data becomes imperative in precision health. In Bioinformatics, significant effort has been committed to harness genomics and transcriptomics data for multiple analyses [1]–[9]. These multimodal biomedical data can be classified into "omics" data and clinical data. Omics data can be further categorized into genomics data, transcriptomics data, proteomics data, metabolomics data, etc. These data are also considered as "multi-omics" biomedical data. For clinical data, it includes electronic medical records and image data. Survival outcome, which is the major clinical outcome we are studying in this thesis, is belonging to electronic medical records. While image data includes histopathologic image, computed tomography (CT) image, magnetic resonance imaging (MRI), etc. Figure 1.1 summarizes these data into a diagram. As the increasing amount of multimodal biomedical data becomes available, mRNA-seq, also known as mRNA sequencing data which reveals the presence and quantity of RNA in biological samples at a given moment, has expeditiously become the standard approach for analyzing the transcriptomes of disease states, biological processes, and a wide range of study designs [10]. Benefit from the powerful machines such as Illumina NovaSeq Sequencing [10], as well

as the well-established data analysis & storage workflows, bioinformatician and data analysts can directly analyze biomedical data in 2D matrices, with units RPKM (Reads Per Kilobase Million), FPKM (Fragments Per Kilobase Million), or TPM (Transcripts Per Million). In this thesis, we will benefit from their sequencing analysis results, and directly face the 2D matrix data describing the gene expressions of samples/patients. Despite mRNA-seq data, some other multimodal biomedical data are also included in our study. They are: (1) miRNA-seq data, also known as microRNA sequencing data; (2) copy number variation (CNV), measured by total kilobase (kb) length, reflects the variation of a certain genome section. The total number CNV is also measured, and defined as copy number burden (CNB); (3) tumor mutation burden (TMB), calculated by the total number of mutated genes based on mutation annotation format (MAF) files; and (4) other demographical and clinical data. These multimodal biomedical data are presented as a data vector per sample, or a data matrix per cohort, describing the features of patients in a high-dimensional space. In addition, this thesis also analyzes breast cancer histopathologic image data[1]. Specifically, hematoxylin and eosin (H&E) stained images and immunohistochemistry (IHC) stained images are analyzed. In H&E stained histopathologic images, hematoxylin stains cell nucleus into dark blue color, while eosin stains extracellular matrix and cytoplasm into pink color. In IHC stained histopathologic images, programmed death-ligand 1 (PD-L1), CD8+ T cells, and CD163+ macrophages in tumor immune micro-environment are stained into brown color, green color, and red color, respectively.

Multi-omics genomics and transcriptomics data are collected from open-access databases, including The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) and NCBI Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/).

1.2 Overview of This Thesis

This thesis starts from survival analysis, and then links it to the gene co-expression analysis and low-rank decomposition methods. A review of basic concepts, prior work, and previous literature will be performed in Chapter 2. Specifically for survival analysis, like-

[1]↑Breast cancer histopathologic dataset was provided by Dr. Zaibo Li at The Ohio State University.

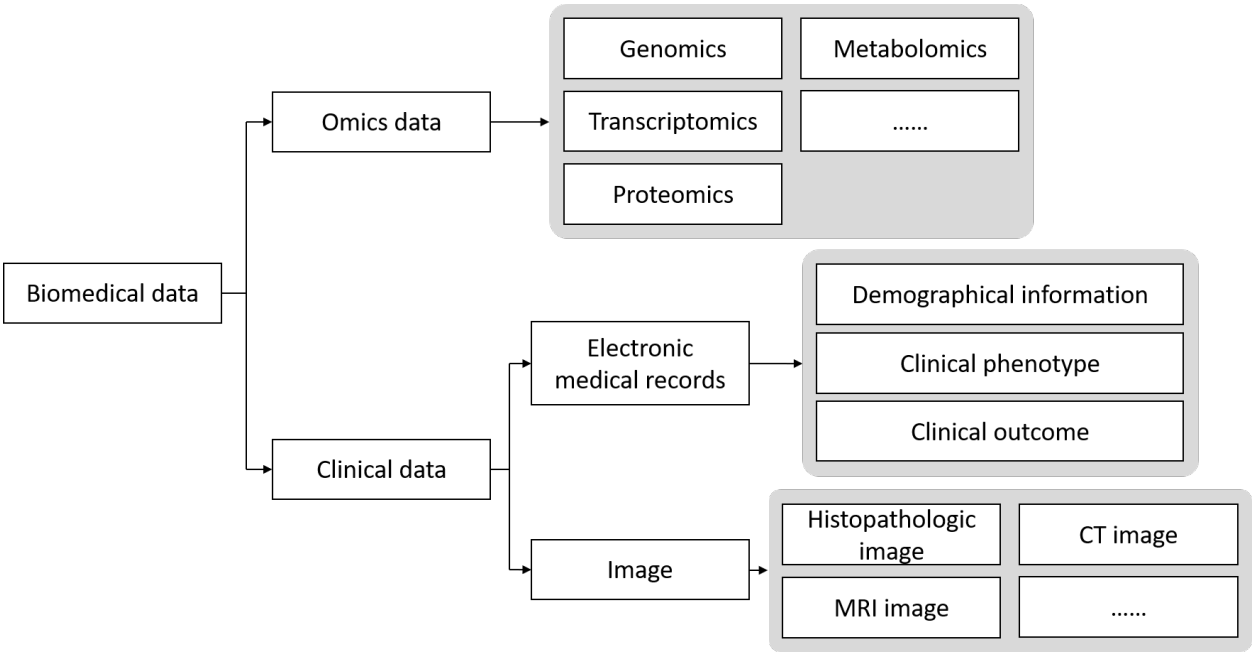


Figure 1.1. Introduction to biomedical data. MRI stands for magnetic resonance imaging, CT stands for computed tomography.

lihood and censorship of survival data, hazard function and partial likelihood, log partial likelihood function, Kaplan-Meier estimator, and Cox proportional hazards model will be explained. This is followed with a section discussing deep learning-based survival prediction using biomedical data. For gene co-expression network analysis, local maximal Quasi-Clique Merger will be introduced in the third section. In the fourth section, unsupervised low-rank approximation via non-negative matrix factorization (NMF) will be introduced, including its formulation, updating rules, and some other variations. Next, histopathologic image registration and segmentation approaches related to this thesis will be introduced and discussed in the fifth and sixth sections, respectively. Finally, related evaluation metrics and statistical tests will be introduced.

By first combining survival analysis with gene co-expression network construction as an upfront feature engineering technique, we propose an integrative algorithm SALMON (Survival Analysis Learning With Multi-Omics Neural Networks) to interpret densely connected gene clusters by deep learning in Chapter 3. In Chapter 3, performances in terms of concordance index and log-rank test P-value are improved when more omics data integrates

into the input of the SALMON algorithm. It also demonstrates a competitive performance compared to other deep learning survival prognosis models. SALMON algorithm further confirms certain mRNA-seq co-expression modules which play pivotal roles in breast cancer prognosis, with several biological functions revealed. This part of work was published in *Frontiers in Genetics* [2] in 2019.

In Chapter 4, by combining non-negative matrix factorization and Cox proportional hazards regression, a novel algorithm named CoxNMF is innovatively designed. CoxNMF aims to unravel latent features behind the high-dimensional transcriptomic data under the time-to-event survival constraints. The algorithm is achieved by decomposing original 2D matrix data into two low-rank matrices "basis" $\boldsymbol{W}$ and "coefficient" $\boldsymbol{H}$, subject to the non-negative constraint. Multiplicative updating rule is adopted for this algorithm. The results present its power of decomposing data into desired latent spaces and also emphasizing its survival information.

Chapter 5 focuses on integrative medical image analysis. In this chapter, an automatic, accurate, comprehensive, interpretable, and reproducible whole slide image (WSI) feature extraction workflow known as, IMage-based Pathological REgistration and Segmentation Statistics (IMPRESS), is described. Features are derived from tumor immune microenvironment and clinical information, and then used to train machine learning models to accurately predict the response to neoadjuvant chemotherapy in breast cancer patients. The results demonstrate that this method outperforms the results trained from features that are manually generated by pathologists.

Finally, bioinformatics and computational biology tool TSUNAMI: Tools SUite for Network Analysis and MIning will be described in Chapter 6. It is a one-stop tool that offers flexibility in parameter selections, comprehensive gene co-expression network mining, direct link to downstream gene set enrichment analysis, Circos plot visualization, and survival analysis. The proposed software tool can bring many conveniences to the bioinformatics community. This part of work was published in *Genomics, Proteomics & Bioinformatics* [11] in 2021.

With different topics introduced in Chapters 3, 4, 5, and 6, this thesis will draw conclusions in Chapter 7. This thesis will also discuss the future direction and potential work.

In the next chapter, we start with the previous work and literature review. Details on the derivation of Cox proportional hazards regression, deep learning-based survival prediction, co-expression network analysis, non-negative matrix factorization, histopathologic image registration and segmentation, and related evaluation metrics will be explained according to previous literature including articles and books.

1.3 Contributions of This Thesis

- We propose an integrative algorithm SALMON to predict breast cancer patient survival and interpret densely connected gene clusters by deep learning.
- We propose CoxNMF algorithm to unravel latent gene interactions under survival constraints by combining non-negative matrix factorization and Cox proportional hazards regression.
- We introduce a comprehensive feature extraction workflow to accurately predict the response to neoadjuvant chemotherapy in breast cancer patients given pre-treatment multimodal histopathologic images.
- We propose a bioinformatics and computational biology tool TSUNAMI for online gene co-expression network analysis.

1.4 Publications Result from This Work

1.4.1 Journal Papers

1. **Z. Huang**, Z. Han, T. Wang, W. Shao, S. Xiang, P. Salama, M. Rizkalla, K. Huang, and J. Zhang, "TSUNAMI: Translational Bioinformatics Tool Suite for Network Analysis and Mining," *Genomics, Proteomics and Bioinformatics*, *Accept, in press*, 2021. [Online]. Available: https://doi.org/10.1016/j.gpb.2019.05.006.

2. **Z. Huang**, T. S. Johnson, Z. Han, B. Helm, S. Cao, C. Zhang, P. Salama, M. Rizkalla, C. Y. Yu, J. Cheng, S. Xiang, X. Zhan, J. Zhang, and K. Huang, "Deep Learning-based Cancer Survival Prognosis from RNA-seq Data: Approaches and Evaluations," *BMC Medical Genomics*, vol. 13, no. 5, pp. 1–12, 2020. [Online]. Available: https://doi.org/10.1186/s12920-020-0686-1.

3. **Z. Huang**, X. Zhan, S. Xiang, T. S. Johnson, B. Helm, C. Y. Yu, J. Zhang, P. Salama, M. Rizkalla, Z. Han, and K. Huang, "SALMON: Survival Analysis Learning with Multi-omics Neural Networks on Breast Cancer," *Frontiers in Genetics*, vol. 10, p. 166, 2019. [Online]. Available: https://doi.org/10.3389/fgene.2019.00166.

4. **Z. Huang**, P. Salama, W. Shao, J. Zhang, and K. Huang, "Low-Rank Reorganization via Proportional Hazards Non-negative Matrix Factorization Unveils Survival Associated Gene Clusters," *arXiv preprint arXiv:2008.03776*, 2020. [Online]. Available: https://arxiv.org/abs/2008.03776.

5. **Z. Huang**, W. Shao, Z. Han, A. M. Alkashash, C. De la Sancha, A. V. Parwani, H. Nitta, Y. Hou, T. Wang, P. Salama, M. Rizkalla, J. Zhang, K. Huang, and Z. Li, "Artificial Intelligence Predicts Breast Cancer Neoadjuvant Chemotherapy Response from H&E and IHC histopathologic Images," *To be submitted*, 2021.

1.4.2 Conference Papers and Abstracts

1. **Z. Huang**, Z. Han, A. V. Parwani, K. Huang, and Z. Li, "Artificial Intelligence Driven Neoadjuvant Chemotherapy Response Prediction in Triple Negative Breast Cancer (TNBC) Unveils Non-linear Feature Interactions," *United States and Canadian Academy of Pathology (USCAP) 2020 Annual Meeting Abstracts*, Los Angeles, CA, USA, March 1–5, 2020.

2. **Z. Huang**, Z. Han, A. V. Parwani, K. Huang, and Z. Li, "Predicting Response to Neoadjuvant Chemotherapy in HER2-positive Breast Cancer using Machine Learning Models with Combined Tissue Imaging and Clinical Features," *United States and Canadian Academy of Pathology (USCAP) 2019 Annual Meeting Abstracts*, National Harbor, Maryland, USA, March 16–21, 2019.

3. **Z. Huang**, T. S. Johnson, Z. Han, B. Helm, S. Cao, C. Zhang, P. Salama, M. Rizkalla, C. Y. Yu, J. Cheng, S. Xiang, X. Zhan, J. Zhang, and K. Huang, "Deep Learning-based Cancer Survival Prognosis from RNA-seq Data: Approaches and Evaluations," *International Conference on Intelligent Biology and Medicine (ICIBM 2019)*, Columbus, OH, USA, June 9–11, 2019. [Online]. Available: https://doi.org/10.1186/s12920-020-0686-1.

1.5 Publications Not Include in This Work

1.5.1 Journal Papers

1. T. S. Johnson, S. Xiang, T. Dong, **Z. Huang**, M. Cheng, T. Wang, K. Yang, D. Ni, K. Huang, J. Zhang, "Combinatorial Analyses Reveal Cellular Composition Changes have Different Impacts on Transcriptomic Changes of Cell Type Specific Genes in Alzheimer's Disease," *Scientific Reports*, vol. 11, no. 1, pp. 1–19, 2021. [Online]. Available: https://doi.org/10.1038/s41598-020-79740-x.

2. W. Shao, T. Wang, L. Sun, T. Dong, Z. Han, **Z. Huang**, J. Zhang, D. Zhang, and K. Huang, "Multi-Task Multi-Modal Learning for Joint Diagnosis and Prognosis of Human Cancers," *Medical Image Analysis*, vol. 65, p. 101795, 2020. [Online]. Available: https://doi.org/10.1016/j.media.2020.101795.

3. S. Cong, X. Yao, **Z. Huang**, S. L. Risacher, K. Nho, A. J. Saykin, and L. Shen, "Volumetric GWAS of Medial Temporal Lobe Structures Identifies an ERC1 Locus using ADNI High-Resolution T2-weighted MRI Data," *Neurobiology of Aging*, vol. 95, pp. 81–93, 2020. [Online]. Available: https://doi.org/10.1016/j.neurobiolaging.2020.07.005.

4. J. Yan, V. V. Raja, **Z. Huang**, E. Amico, K. Nho, S. Fang, O. Sporns, Y. Wu, A. J. Saykin, J. Goñi, and L. Shen, "Brain-wide Structural Connectivity Alterations under the Control of Alzheimer Risk Genes," *International Journal of Computational Biology and Drug Design*, vol. 13, no. 1, pp. 58–70, 2020. [Online]. Available: https://doi.org/10.1504/IJCBDD.2020.105098.

5. S. Huang, **Z. Huang**, P. Chen, and C. Feng, "Aberrant Chloride Intracellular Channel 4 Expression is Associated with Adverse Outcome in Cytogenetically Normal Acute Myeloid Leukemia," *Frontiers in Oncology*, vol. 10, p. 1648, 2020. [Online]. Available: https://doi.org/10.3389/fonc.2020.01648.

6. S. Huang, **Z. Huang**, C. Ma, L. Luo, Y. Li, Y. Wu, Y. Ren, and C. Feng, "Acidic Leucine-rich Nuclear Phosphoprotein-32A Expression Contributes to Adverse Outcome in Acute Myeloid Leukemia," *Annals of Translational Medicine*, vol. 8, no. 6, p. 345, 2020. [Online]. Available: https://doi.org/10.21037/atm.2020.02.54.

7. C. Y. Yu, S. Xiang, **Z. Huang**, T. S. Johnson, X. Zhan, Z. Han, M. I. Abu Zaid, and K. Huang, "Gene Co-expression Network and Copy Number Variation Analyses Identify Transcription Factors Involved in Multiple Myeloma Progression," *Frontiers in Genetics*, vol. 10, p. 468, 2019. [Online]. Available: https://doi.org/10.3389/fgene.2019.00468.

8. T. S. Johnson, T. Wang, **Z. Huang**, C. Y. Yu, Y. Wu, Y. Han, Y. Zhang, K. Huang, and J. Zhang, "LAmbDA: Label Ambiguous Domain Adaptation Dataset Integration Reduces Batch Effects and Improves Subtype Detection," *Bioinformatics*, vol. 35, no. 22, pp. 4696–4706, 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz295.

9. X. Zhan, J. Cheng, **Z. Huang**, Z. Han, B. Helm, X. Liu, J. Zhang, T. Wang, D. Ni, and K. Huang, "Correlation Analysis of Histopathology and Proteogenomics Data for Breast Cancer," *Molecular & Cellular Proteomics*, vol. 18, no. 8, S37–S51, 2019. [Online]. Available: https://doi.org/10.1074/mcp.RA118.001232.

10. W. Shao, T. Wang, **Z. Huang**, J. Cheng, Z. Han, D. Zhang, and K. Huang, "Diagnosis-Guided Multi-Modal Feature Selection for Prognosis Prediction of Lung Squamous Cell Carcinoma," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 113–121, 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-32251-9_13.

11. T. S. Johnson, S. Li, E. Franz, **Z. Huang**, S. D. Li, M. J. Campbell, K. Huang, and Y. Zhang, "PseudoFuN: Deriving functional potentials of Pseudogenes from Integrative Relationships with Genes and microRNAs Across 32 Cancers," *GigaScience*, vol. 8, no. 5, p. giz046, 2019. [Online]. Available: https://doi.org/10.1093/gigascience/giz046.

12. S. Xiang, **Z. Huang**, T. Wang, Z. Han, C. Y. Yu, D. Ni, K. Huang, and J. Zhang, "Condition-specific Gene Co-expression Network Mining Identifies Key Pathways and Regulators in the Brain Tissue of Alzheimer's Disease Patients," *BMC Medical Genomics*, vol. 11, no. 6, pp. 39–51, 2018. [Online]. Available: https://doi.org/10.1186/s12920-018-0431-1.

13. C. Feng, H. Huang, S. Huang, Y. Zhai, J. Dong, L. Chen, **Z. Huang**, X. Zhou, B. Li, L. Wang, W. Chen, F. Lv, and T. Li, "Identification of Potential Key Genes Associated with Severe Pneumonia using mRNA-seq," *Experimental and Therapeutic Medicine*, vol. 16, no. 2, pp. 758–766, 2018. [Online]. Available: https://doi.org/10.3892/etm.2018.6262.

14. S. Huang, C. Feng, L. Chen, **Z. Huang**, X. Zhou, B. Li, L. Wang, W. Chen, F. Lv, and T. Li, "Molecular Mechanisms of Mild and Severe Pneumonia: Insights from RNA Sequencing," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 23, p. 1662, 2017. [Online]. Available: https://doi.org/10.12659/MSM.900782.

15. S. Huang, C. Feng, L. Chen, **Z. Huang**, X. Zhou, B. Li, L. Wang, W. Chen, F. Lv, and T. Li, "Identification of Potential Key Long Non-Coding RNAs and Target Genes Associated with Pneumonia using Long Non-Coding RNA Sequencing (lncRNA-Seq): A Preliminary Study," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 22, p. 3394, 2016. [Online]. Available: https://doi.org/10.12659/MSM.900783.

16. S. Huang, H. Yang, Y. Li, C. Feng, L. Gao, G. Chen, H. Gao, **Z. Huang**, Y. Li, and L. Yu, "Prognostic Significance of Mixed-Lineage Leukemia (MLL) Gene Detected by Real-Time Fluorescence Quantitative PCR Assay in Acute Myeloid Leukemia," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 22, p. 3009, 2016. [Online]. Available: https://doi.org/10.12659/MSM.900429.

17. T. Wang, W. Shao, **Z. Huang**, H. Tang, J. Zhang, Z. Ding, and K. Huang, "MOGONET Integrates Multi-omics Data using Graph Convolutional Networks allowing Patient Classification and Biomarker Identification," *Nature Communications*, vol. 12, p. 3445, 2021. [Online]. Available: https://doi.org/10.1038/s41467-021-23774-w.

18. N. Alghamdi, W. Chang, P. Dang, X. Lu, C. Wan, S. Gampala, **Z. Huang**, J. Wang, Q. Ma, Y. Zang, M. Fishel, S. Cao, and C. Zhang, "A Graph Neural Network Model to Estimate Cell-wise Metabolic Flux Using Single Cell RNA-seq Data," *arXiv preprint arXiv:10.1101/2020.09.23.310656*, 2021 [Online]. Available: https://doi.org/10.1101/2020.09.23.310656.

1.5.2 Conference Papers and Abstracts

1. **Z. Huang**, K. Tgavalekos, and C. Zhao, "AI-Driven Forecasting of Mean Pulmonary Artery Pressure for the Management of Cardiac Patients," *Society of Critical Care Medicine's 49th Critical Care Congress (SCCM 2020)*, Orlando, Florida, USA, February 16–19, 2020. [Online]. Available: http://doi.org/10.1097/01.ccm.0000619240.04761.13.

2. W. Shao, T. Wang, **Z. Huang**, J. Cheng, Z. Han, D. Zhang, and K. Huang, "Diagnosis-Guided Multi-Modal Feature Selection for Prognosis Prediction of Lung Squamous Cell Carcinoma," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2019)*, Shenzhen, China, October 13–17, 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-32251-9_13.

3. A. Hara, **Z. Huang**, Q. Sun, G. Maupomé, and L. Shen, "Machine-learning Identification of Dental Hard-tissue Conditions through Fully Convolutional Neural Networks," *American Public Health Association Annual Meeting and Expo (APHA 2019)*, Philadelphia, USA, November 2–6, 2019. [Online]. Available: https://apha.confex.com/apha/2019/meetingapp.cgi/Paper/436179.

4. J. Yan, V. V. Raja, **Z. Huang**, E. Amico, K. Nho, S. Fang, O. Sporns, Y. Wu, A. J. Saykin, J. Goñi, and L. Shen, "Brain-wide Structural Connectivity Alterations Under the Control of Alzheimer Risk Genes," *International Conference on Intelligent Biology and Medicine (ICIBM 2018)*, Los Angeles, CA, USA, June 10–12, 2018. [Online]. Available: https://doi.org/10.1504/IJCBDD.2020.105098.

5. J. Xue, **Z. Huang**, J. Zhou, Y. Chen, and S. Chien, "A Hierarchical Clustering Analysis (HCA) in Automatic Driving Regarding to Vehicle-to-Vehicle Pedestrian Position Identification," *25th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*, Detroit, MI, USA, June 5–8, 2017. [Online]. Available: https://www-esv.nhtsa.dot.gov/Proceedings/25/25ESV-000176.pdf.

6. B. Tang, S. Chien, **Z. Huang**, and Y. Chen, "Pedestrian Protection using the Integration of V2V and the Pedestrian Automatic Emergency Braking System," *19th International IEEE Conference on Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro, Brazil, November 1–4, 2016. [Online]. Available: https://doi.org/10.1109/ITSC.2016.7795913.

1.5.3 Thesis

1. **Z. Huang**, "Vehicle Sensor-based Pedestrian Position Identification in V2V Environment," *Purdue University Thesis, Master of Science in Electrical & Computer Engineering*, 2016. [Online]. Available: https://docs.lib.purdue.edu/dissertations/AAI10248822.

1.5.4 Patents

1. **Z. Huang**, X. Bei, K. Huang, K. Qian, and A. Zhang, "One Refers to Upper Mouse," China Patent No. CN103257723B, filed April 27, 2013, and issued July 06, 2016. [Online]. Available: https://patents.google.com/patent/CN103257723B/en.

2. PRIOR WORK AND LITERATURE REVIEW

2.1 Cox Proportional Hazards Model

Including genomics, transcriptomics, proteomics, metabolomics, and other type of data such as medical images, the multi-omics source of information consist the big family of biomedical data. Benefit from advanced biotechnology and next generation sequencing (NGS), acquiring and modeling biomedical data becomes achievable nowadays. As one of the central goal in precision health, analyzing the data associated with survival is of great interest to the biomedical researchers. We start from introducing survival analysis, then aim to unravel latent information behind the biomedical data by combining the survival analysis with other advanced machine learning techniques.

Survival analysis, also known as time-to-event analysis, aims to analyze the lifespan and estimate the time to an event of interest (especially death event), given the observed data of an individual or a population [12]. It was originally designed for medical researchers and data analysts to analyze the lifetimes [13]. Nowadays the time-to-event analysis can also be used for other applications such as predicting churning customers or employees when considering the event to be leaving the company, estimating the lifetime of a machine, etc [12].

Survival analysis considers relative time duration [14], thus a subject can enter the study at any time. However, not all subjects of the given population will experience the event of interest (such as death, churn, etc.) during the study. For those their survival times longer than the end time in the study were labelled as "censored" [12]. The censorship allows the analysis measure the lifetimes even when the subjects are not experiencing the event of interest.

Modeling the survival can be traced back to 1972 [15] when previous statisticians systematically summarized the survival function, hazard function, Kaplan-Meier estimator, survival regression, Cox model, etc. In this thesis, survival analysis will play an important role in conjunct with gene co-expression network analysis and non-negative matrix factorization. The combination of these analyses help us to unravel the functionality behind the gene expression data, as well as to determine the latent interactions.

2.1.1 Likelihood and Censorship of Survival Data

It is worth to know in biomedical applications, time-to-event data often has the property where some individuals are still alive (or because of insufficient follow-up) at the end of study, we refer it to "right censored" data. Examples of right censored data is shown in Figure 2.1. In contrary, "left censored" is when the event of interest has already occurred before the data is enrolled. In most cases and in this thesis, we are dealing with right censored data.

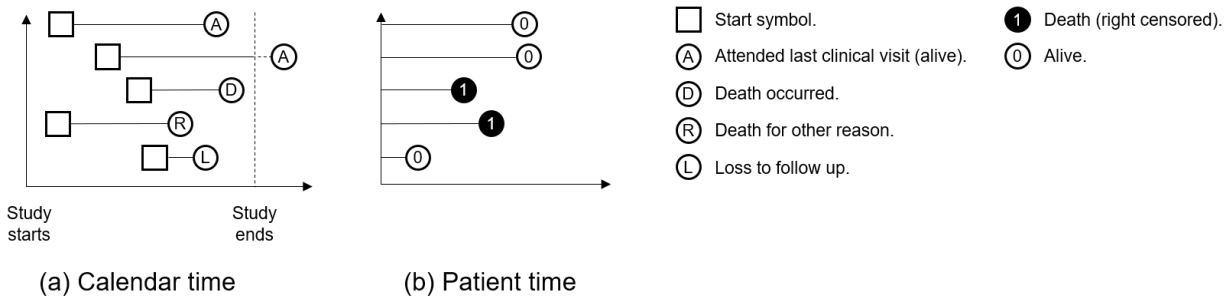


Figure 2.1. Calendar time (a) and patient time (b) with right censored data.

Given the cumulative distribution function $F(t) = \Pr(T \leq t)$ with the associated probability density function $f(t) = F(t)$, we define the survival function as

$$S(t) = \Pr(T > t) = 1 - F(t). \tag{2.1}$$

We also define the hazard function and cumulative hazard function be

$$h(t) = \lim_{\Delta t \to 0} \left(\frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \right) = \lim_{\Delta t \to 0} \left(\frac{S(t) - S(t + \Delta t)}{\Delta t S(t)} \right), \tag{2.2}$$

and

$$H(t) = \int_0^t h(s) ds, \tag{2.3}$$

respectively. It is worth to know that if T is discrete and positive integer-valued, then $h(t) = \Pr(T = t|T \geq t) = \frac{\Pr(T=t)}{S(t-1)}$. From equation 2.1, 2.2, and 2.3, we have:

(I)

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \left(\frac{S(t) - S(t + \Delta t)}{\Delta t S(t)} \right) \\ &= -\frac{S(t)}{S(t)} \\ &= -\frac{d}{dt} \log S(t) \\ &= \frac{d}{dt} H(t), \end{aligned} \tag{2.4}$$

(II)

$$S(t) = \exp(-H(t)), \tag{2.5}$$

and (III)

$$f(t) = h(t)S(t). \tag{2.6}$$

Since $S(0) = 1$ for (II). Table 2.1 shows some typical density functions.

Table 2.1. Typical density functions for survival analysis modeling. Note: α and ρ are parameters defined only in this table.

	$h(t)$	$S(t)$	$f(t)$
Weibull	$\exp(-(\rho t)^\alpha)$	$\alpha\rho^\alpha t^{\alpha-1}$	$\alpha\rho^\alpha t^{\alpha-1}\exp(-(\rho t)^\alpha)$
Log-logistic	$\frac{1}{1+(\rho t)^\alpha}$	$\frac{\alpha\rho^\alpha t^{\alpha-1}}{1+(\rho t)^\alpha}$	$\frac{\alpha\rho^\alpha t^{\alpha-1}}{(1+(\rho t)^\alpha)^2}$
Exponential	$\exp(-\rho t)$	ρ	$\rho e^{-\rho t}$

Suppose we have the time to event T_E and time to censoring event T_C by assuming all patients eventually will have event (such as death) or be censored, where T denotes the positive random variable representing time. If a pair of observations $(t_E^{(\mathrm{i})}, t_C^{(\mathrm{i})})$ is observed with respect to patient i, we have the likelihood

$$\begin{aligned} L &= \prod_{t_E^{(\mathrm{i})} < t_C^{(\mathrm{i})}} f(t_E^{(\mathrm{i})}) S_C(t_E^{(\mathrm{i})}) \prod_{t_C^{(\mathrm{i})} < t_E^{(\mathrm{i})}} S(t_C^{(\mathrm{i})}) f_C(t_C^{(\mathrm{i})}) \\ &= \prod_{\mathrm{i}} h(t^{(\mathrm{i})})^{\delta_\mathrm{i}} S(t^{(\mathrm{i})}) \prod_{\mathrm{i}} h_C(t^{(\mathrm{i})})^{1-\delta_\mathrm{i}} S_C(t^{(\mathrm{i})}), \end{aligned} \tag{2.7}$$

by assuming censoring mechanism is independent from the event time. Where $t^{(\mathrm{i})} = \min(t_E^{(\mathrm{i})}, t_C^{(\mathrm{i})})$, $\delta_{\mathrm{i}} = \begin{cases} 1 & \text{if } t_E^{(\mathrm{i})} < t_C^{(\mathrm{i})} \\ 0 & \text{otherwise} \end{cases}$. It is noteworthy that the second product in Equation 2.7 can be omitted if the censoring is independent. Thus we have

$$L = \prod_{\mathrm{i}} h(t^{(\mathrm{i})})^{\delta_{\mathrm{i}}} S(t^{(\mathrm{i})}). \tag{2.8}$$

2.1.2 Hazard Function, Partial Likelihood, and Cox Proportional Hazards Model

For each subject i it has a covariate vector $\boldsymbol{H}_{\mathrm{i}}$ and a scale parameter ρ_{i}. In Cox regression [15], $\rho_{\mathrm{i}} = \exp(\beta^T \boldsymbol{H}_{\mathrm{i}})$. We assume any two subjects have hazard functions where the ratio is a constant proportion which depends on the covariates

$$\lambda_{\mathrm{i}}(t|\boldsymbol{H}_{\mathrm{i}}) = \rho_{\mathrm{i}} \lambda_0(t|\boldsymbol{H}_{\mathrm{i}}), \tag{2.9}$$

where

$$\lambda_{\mathrm{i}}(t|\boldsymbol{H}_{\mathrm{i}}) = \lim_{\Delta t \to 0^+} \frac{\Pr(t \le Y_{\mathrm{i}} < t + \Delta t | Y_{\mathrm{i}} \ge t, \boldsymbol{H}_{\mathrm{i}})}{\Delta t}. \tag{2.10}$$

In the equation, λ_0 is the baseline hazard function, reflects the underlying hazard for subjects with all covariates equal to 0 (*i.e.*, the "reference group") [16]. β is the vector of regression coefficients to be estimated, and ρ_{i} depends on the linear predictor $\beta^T \boldsymbol{H}_{\mathrm{i}}$. As we mentioned in Cox regression, $\rho_{\mathrm{i}} = \exp(\beta^T \boldsymbol{H}_{\mathrm{i}})$ and the functional form of the baseline hazard is not given but is determined from the data, thus the Cox model is termed as semi-parametric model.

If we rewrite the survival time t to Y_{i} for i^{th} patient, then the hazard function for the Cox proportional hazards model with respect to i^{th} patient has the form

$$\begin{aligned} \lambda(Y_{\mathrm{i}}|\boldsymbol{H}_{\mathrm{i}}) &= \lambda_0(Y_{\mathrm{i}}) \exp(\beta_1 \boldsymbol{H}_{1,\mathrm{i}} + \beta_2 \boldsymbol{H}_{2,\mathrm{i}} + \cdots + \beta_K \boldsymbol{H}_{K,\mathrm{i}}) \\ &= \lambda_0(Y_{\mathrm{i}}) \exp(\beta^T \boldsymbol{H}). \end{aligned} \tag{2.11}$$

This expression gives the hazard function at time Y_{i} for subject i with covariate vector (explanatory variables) $\boldsymbol{H}_{\mathrm{i}}$. A value of $\beta_m \boldsymbol{H}_{m,\mathrm{i}}$ greater than zero, or equivalently a hazard ratio $\exp(\beta_m \boldsymbol{H}_{m,\mathrm{i}})$ greater than one, indicates that as the value of the m^{th} covariate increases, the event hazard increases and thus the length of survival decreases. A hazard ratio $\exp(\beta_m \boldsymbol{H}_{m,\mathrm{i}})$ above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.

In summary,

- hazard ratio $\exp(\beta_m \boldsymbol{H}_{m,\mathrm{i}}) = 1$: No effect.
- hazard ratio $\exp(\beta_m \boldsymbol{H}_{m,\mathrm{i}}) < 1$: Reduction in the hazard.
- hazard ratio $\exp(\beta_m \boldsymbol{H}_{m,\mathrm{i}}) > 1$: Increase in Hazard.

Suppose the event times are given by $0 < Y_1 < Y_2 < Y_3 < \cdots < Y_N$ by assuming no tied event times, and let C_{i} denote the event for subject i at time Y_{i}. If there is a death event associated at time Y_{i}, then the probability of the subject i has the death event is

$$\Pr(\text{subject i}|Y_{\mathrm{i}}) = \frac{\lambda_{\mathrm{i}}(Y_{\mathrm{i}})}{\lambda_1(Y_{\mathrm{i}}) + \cdots + \lambda_N(Y_{\mathrm{i}})}. \tag{2.12}$$

Under the proportional hazards assumption with the risk set $\mathcal{R}_{\mathrm{i}}$ where those subjects available for the death event at time Y_{i} (*i.e.*, denote the set of individuals who are "at risk" for death at time Y_{i}), we have

$$\begin{aligned} \Pr(C_{\mathrm{i}}|Y_{\mathrm{i}}) &= \frac{\rho_{\mathrm{i}}\lambda_0(Y_{\mathrm{i}})}{\sum_{\mathrm{j}\in\mathcal{R}_{\mathrm{i}}} \rho_{\mathrm{j}}\lambda_0(Y_{\mathrm{i}})} = \frac{\rho_{\mathrm{i}}\lambda_0(Y_{\mathrm{i}})}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_{\mathrm{i}}} \rho_{\mathrm{j}}\lambda_0(Y_{\mathrm{i}})} \\ &= \frac{\rho_{\mathrm{i}}}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_{\mathrm{i}}} \rho_{\mathrm{j}}} \\ &= \frac{\exp(\beta^T \boldsymbol{H}_{\mathrm{i}})}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_{\mathrm{i}}} \exp(\beta^T \boldsymbol{H}_{\mathrm{j}})}. \end{aligned} \tag{2.13}$$

2.1.3 Log Partial Likelihood Function

Given the data $\boldsymbol{H}$, the likelihood of the death event to be observed occurring for patient i at time Y_{i} can be written as

$$\begin{aligned} L_{\mathrm{i}}(\beta) &= \frac{\lambda(Y_{\mathrm{i}}|\boldsymbol{H}_{\mathrm{i}})}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_{\mathrm{i}}}\lambda(Y_{\mathrm{i}}|\boldsymbol{H}_{\mathrm{j}})} = \frac{\lambda_0(Y_{\mathrm{i}})\exp(\beta^T\boldsymbol{H}_{\mathrm{i}})}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_{\mathrm{i}}}\lambda_0(Y_{\mathrm{i}})\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})} \\ &= \frac{\exp(\beta^T\boldsymbol{H}_{\mathrm{i}})}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_{\mathrm{i}}}\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})}, \end{aligned} \tag{2.14}$$

where λ is the hazard function. The corresponding log partial likelihood is

$$\ell_{\boldsymbol{H},\beta}(C,Y) = \sum_{\mathrm{i}:C_{\mathrm{i}}=1}\left(\beta^T\boldsymbol{H}_{\mathrm{i}} - \log\left(\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_{\mathrm{i}}}\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})\right)\right). \tag{2.15}$$

The partial derivative of $\ell_{\boldsymbol{H},\beta}(C,Y)$ with respect to β is:

$$\frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\beta} = \sum_{\mathrm{i}:C_{\mathrm{i}}=1}\left(\boldsymbol{H}_{\mathrm{i}} - \frac{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_{\mathrm{i}}}\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})\boldsymbol{H}_{\mathrm{j}}}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_{\mathrm{i}}}\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})}\right). \tag{2.16}$$

The partial derivative of $\ell_{\boldsymbol{H},\beta}(C,Y)$ with respect to $\boldsymbol{H}$ is:

$$\begin{aligned} \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}} &= \begin{bmatrix} \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{1,1}} & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{1,2}} & \cdots & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{1,N}} \\ \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{2,1}} & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{2,2}} & \cdots & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{2,N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{K,1}} & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{K,2}} & \cdots & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{K,N}} \end{bmatrix} \\ &= \left[\underbrace{\left(C_r\beta - \sum_{s=r}^{N} C_s\frac{\mathbb{1}_{(Y_r\geq Y_s)}\beta\exp(\beta^T\boldsymbol{H}_r)}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_s}\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})}\right)}_{K\times 1 \text{ vector which repeats } N \text{ times for } r=1,2,\cdots,N.}\right], \end{aligned} \tag{2.17}$$

where $\mathbb{1}_{(Y_n\geq Y_s)}$ is the indicator function: $\mathbb{1}_{(Y_n\geq Y_s)} = \begin{cases} 1 & \text{if } Y_n \geq Y_s \\ 0 & \text{otherwise} \end{cases}$.

2.1.4 Kaplan-Meier Estimator

Kaplan-Meier estimator [17] is one of the frequently used non-parametric estimators of the survival function $\mathbb{S}(t)$ (the probability that survival time is longer than t). It usually compares two groups in a study. Suppose in the discrete time case, we let

$$\Pr(\text{death occurred at } t_{\mathrm{i}} | \text{survived to } t_{\mathrm{i}}-) = h_{\mathrm{i}}, \tag{2.18}$$

where $t_{\mathrm{i}}-$ is the time right before the time t_{i}. Suppose random variables $X_1, \cdots, X_N$ which represent independent observations from a distribution with cumulative distribution function (CDF) F. Consider there are observations $x_1, \cdots, x_N$ from a random sample, then we define the empirical distribution function

$$\hat{F}(x) = \frac{1}{N} \sum_{\mathrm{i}=1}^{N} (\mathbb{1}_{x_{\mathrm{i}} \leq x}). \tag{2.19}$$

The above formula is appropriate for no censoring occurs. If censoring occurs, the scenario will be different. Recall that in Equation 2.7, $t^{(\mathrm{i})} = \min(t_E^{(\mathrm{i})}, t_C^{(\mathrm{i})})$, $\delta_{\mathrm{i}} = \begin{cases} 1 & \text{if } t_E^{(\mathrm{i})} < t_C^{(\mathrm{i})} \\ 0 & \text{otherwise} \end{cases}$. Suppose we have N pairs of observations $(x_{\mathrm{i}}, \delta_{\mathrm{i}})$ for $\mathrm{i} = 1, 2, \cdots, N$, then

$$\begin{aligned} L &= \prod_{\mathrm{i}} f(x_{\mathrm{i}})^{\delta_{\mathrm{i}}} S(x_{\mathrm{i}})^{1-\delta_{\mathrm{i}}} \\ &= \prod_{\mathrm{i}} f(x_{\mathrm{i}})^{\delta_{\mathrm{i}}} \left(1 - F(x_{\mathrm{i}})\right)^{1-\delta_{\mathrm{i}}}. \end{aligned} \tag{2.20}$$

Suppose there are failure times $(0 <) t_1 < \cdots < t_{\mathrm{i}} < \cdots$. Let $s_{\mathrm{i}1}, s_{\mathrm{i}2}, \cdots, s_{\mathrm{i}c_{\mathrm{i}}}$ be the censoring times within the interval $[t_{\mathrm{i}}, t_{\mathrm{i}+1})$. We also suppose that there are d_{i} failures at time t_{i} by allowing tied failure times. By rewriting $f(t_{\mathrm{i}}) = F(t_{\mathrm{i}}) - F(t_{\mathrm{i}}-)$ and assuming $F(t_{\mathrm{i}})$ takes

fixed values at the failure time points, thus $F(t_{\mathrm{i}}-) = F(t_{\mathrm{i}-1})$ and $F(s_{\mathrm{i}k}) = F(t_{\mathrm{i}})$. Then the likelihood becomes

$$\begin{aligned} L &= \prod_{failures} f(t_{\mathrm{i}})^{d_{\mathrm{i}}} \prod_{\mathrm{i}} \left(\prod_{k=1}^{c_{\mathrm{i}}} (1 - F(s_{\mathrm{i}k})) \right) \\ &= \prod_{failures} (F(t_{\mathrm{i}}) - F(t_{\mathrm{i}}-))^{d_{\mathrm{i}}} \prod_{\mathrm{i}} \left(\prod_{k=1}^{c_{\mathrm{i}}} (1 - F(s_{\mathrm{i}k})) \right) \\ &= \prod_{failures} (F(t_{\mathrm{i}}) - F(t_{\mathrm{i}}-))^{d_{\mathrm{i}}} \prod_{\mathrm{i}} (1 - F(t_{\mathrm{i}}))^{c_{\mathrm{i}}} . \end{aligned} \tag{2.21}$$

As mentioned in Equation 2.18, Pr(death occurred at t_{i}|survived to $t_{\mathrm{i}}-$) $= h_{\mathrm{i}}$, we will have

$$\begin{aligned} S(t_{\mathrm{i}}) &= 1 - F(t_{\mathrm{i}}) = \prod_{1}^{\mathrm{i}} (1 - h_{\mathrm{j}}), \\ f(t_{\mathrm{i}}) &= h_{\mathrm{i}} \prod_{1}^{\mathrm{i}-1} (1 - h_{\mathrm{j}}), \end{aligned} \tag{2.22}$$

and

$$L = \prod_{t_{\mathrm{i}}} h_{\mathrm{i}}^{d_{\mathrm{i}}} (1 - h_{\mathrm{i}})^{n_{\mathrm{i}} - d_{\mathrm{i}}}. \tag{2.23}$$

The Kaplan-Meier estimator uses maximum likelihood estimation technique for h_{i}. Taking the log of the equation 2.23, we have:

$$\ell = \sum_{\mathrm{i}} d_{\mathrm{i}} \log h_{\mathrm{i}} + \sum_{\mathrm{i}} (n_{\mathrm{i}} - d_{\mathrm{i}}) \log(1 - h_{\mathrm{i}}). \tag{2.24}$$

By taking the derivative with respect to h_{i}, we have:

$$\begin{aligned} \frac{\partial \ell}{\partial h_{\mathrm{i}}} &= \frac{d_{\mathrm{i}}}{h_{\mathrm{i}}} - \frac{n_{\mathrm{i}} - d_{\mathrm{i}}}{1 - h_{\mathrm{i}}} = 0, \\ \hat{h}_{\mathrm{i}} &= \frac{d_{\mathrm{i}}}{n_{\mathrm{i}}}, \end{aligned} \tag{2.25}$$

which the formal formula Kaplan-Meier estimator becomes

$$\hat{\mathbb{S}}(t) = \prod_{t_{\mathrm{i}} \le t} \left(1 - \frac{d_{\mathrm{i}}}{n_{\mathrm{i}}} \right). \tag{2.26}$$

Note: $n_{i+1} + c_i + d_i = n_i$, where n_i is the number at risk at t_i, c_i is the number of censored patients in the time interval $[t_i, t_{i+1})$, and d_i is the number of events at t_i. To avoid confusion with other parts of this thesis, symbols and notations s, X, x, and δ are defined locally only in this section.

2.2 Deep Learning-based Survival Prediction using Omics Data

With the high prevalence of neural networks and deep learning-based algorithms in the computational biology studies, it is clear that the advantages of optimization in a highly non-linear space are welcomed improvements in biomedicine [18]–[22]. In bioinformatics, significant effort has been committed to harnessing transcriptomic data for multiple analyses [3], [4], [7]–[9] especially cancer survival prognosis [23]–[25]. Recent advances in kernel-based deep learning models have introduced a new era in medical research. Originally designed for pattern recognition and image processing, Deep learning models are now applied to survival prognosis of cancer patients [1]. Specifically, deep learning versions of the Cox proportional hazards models are trained with multi-omics data (especially transcriptomics data) to predict survival outcomes in cancer patients.

Among deep learning-based survival prediction studies, Faraggi and Simon [26] was the first study to use clinical information to predict prostate cancer survival through an artificial neural network model. Mobadersany *et al.* [27] integrated histological features, convolutional neural networks (CNN), and genomics data to predict cancer prognosis via Cox regression. Despite of various existed applications on survival analysis such as random survival forest (RFS) [28] or generalized linear model with Cox regression (GLMNET) [29], the use of deep learning Cox models was pioneered by Ching *et al.* [24], who applied Cox regression with neural networks (Cox-nnet) to predict survival using transcriptomic data became prevalent. Similarly, Katzman *et al.* [25] used DeepSurv with multi-layer neural networks for survival prognosis and developed a personalized treatment recommendation system. Meanwhile, AECOX (AutoEncoder with Cox regression network) [1] also attempted for cancer prognosis prediction with simultaneous learning of lower dimensional representation of inputs.

In Figure 2.2, we demonstrate three deep learning architectures for cancer survival prognosis. Cox-nnet (Figure 2.2A) is the most succinct model with only one hidden layer, while DeepSurv (Figure 2.2B) uses multiple hidden layers of consistent dimensions and treats the number of hidden layers as a hyper-parameter. Similarly, AECOX also treats the number of hidden layers as a hyper-parameter, but the hidden layers are set symmetrically in the encoder and decoder (Figure 2.2C). All three models employ the same Cox proportional hazards model.

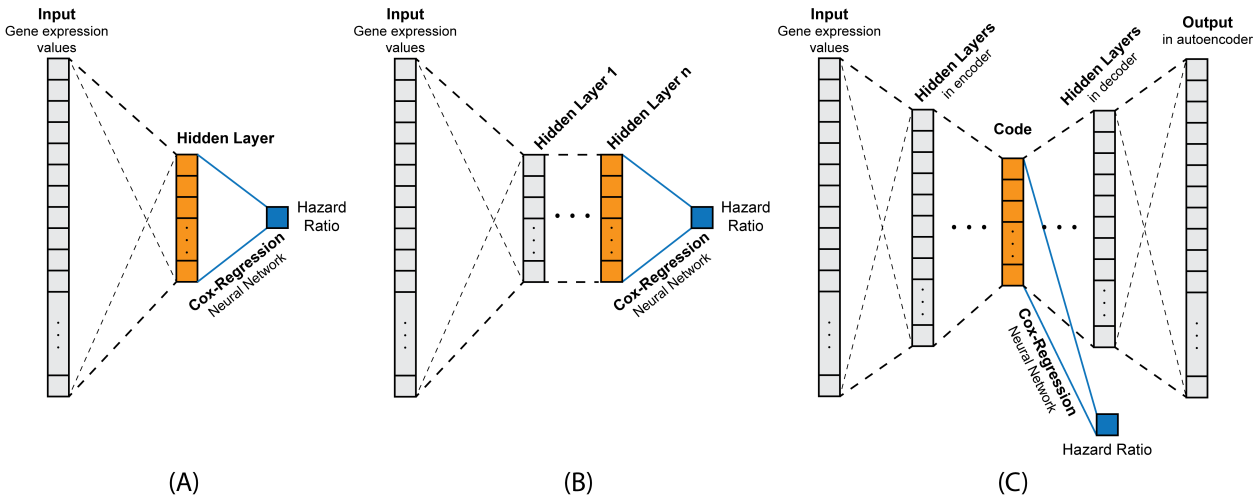


Figure 2.2. Neural network architectures of three deep learning-based models. (A) Cox-nnet with a single hidden layer; (B) DeepSurv with multiple hidden layers having consistent dimensions; (C) AECOX with multiple hidden layers in the both encoder and decoder part. Last hidden layers in all models are indicated in orange and were connect to a Cox regression neural networks with hazard ratios as the outputs.

Cox proportional hazards regression with partial log likelihood (Equation 2.15) is used in the objective functions of the aforementioned deep learning-based survival prediction models. It is proved that the Deep Learning architectures can be effectively applied for cancer prognosis prediction with Cox-proportional hazard model incorporated while demonstrating superior performances comparing to traditional machine learning models for survival analysis [1].

In order to achieve better performance of survival prognosis with deep learning, recent studies have made use of multi-omics data instead of single omics data. This was pioneered by Chaudhary *et al.* [30], who used a deep autoencoder to guide dimensionality reduction upfront, followed by survival prediction, for liver cancer. Similarly, [31] used an autoencoder to predict neuroblastoma patient survival by integrating the number of alterations and gene expression data. [32] also used an autoencoder to produce a new set of features, then used the Cox model to predict survival of bladder cancer patients. Lee *et al.* [33] similarly used a deep autoencoder to predict lung adenocarcinoma patients survival using four omics data.

These deep learning-based models exhibited the advantage of integrating multi-omics data for cancer survival prognosis. However, these methods did not completely exploit all available omics data, and the high-dimensional features were either directly used as input to the model or obtained via a deep autoencoder for dimensionality reduction that is separate from Cox regression. In addition, these deep learning-based approaches analyzed multi-omics features at the gene-level, which lack module-level analysis. To this end, we address all current limitations and propose the SALMON algorithm [2], a deep learning cancer survival prognosis model that utilizes multi-omics inputs, in Chapter 3.

2.3 Co-expression Network Analysis

After the Cox proportional hazards model and its related derivations are elucidated, We next introduce gene co-expression network analysis, which is a group of methods to obtain the underlying information behind gene expressions based on correlation analysis [34].

Correlation networks along with the algorithm of co-expression network analysis are increasingly being used in bioinformatics and biology domains for analyzing large and high-dimensional data, such as gene expressions [35]. Finding the clusters contain highly correlated and densely connected genes is one of the common practices in data mining with numerous applications including social network and biomedicine [34].

Narrowing down to the area of analyzing gene expressions, the concept of gene co-expression networks was initially introduced as relevance networks [36] by firstly calculating the Pearson correlation coefficients or other pairwise similarity measurements given a pop-

ulation of data and associated gene expressions, then constructing the gene co-expression networks according to certain pre-defined parameters or thresholds. The input data, typically gene expression matrix, is given by P genes/features and N samples. For example, the mRNA-seq expression matrix. Given the input data, calculating the pairwise similarity score will return a P by P similarity matrix.

Having the similarity matrix, the next step is to perform the gene-level co-expression clustering. The resulting gene clusters can be mutually exclusive (no overlapping) or with overlaps, varied by the chosen algorithm or parameters. For example, weighted gene co-expression network analysis (WGCNA) [35] algorithm will return gene clusters without overlapping. In contrary, local maximal Quasi-Clique Merger (lmQCM) [34] algorithm allowed module overlapping by setting a parameter q_β stands for maximum overlapping ratio.

Moreover, the similarity matrix can be replaced by regulatory network [37], metabolism network [38], and protein-protein interaction (PPI) network [39] for general types of co-expression analyses. Such analyses have been widely used to predict new gene functions [40]–[42], identify protein interactions [43], detect genetic variants in cancers [44], or use for upfront feature engineering [2] prior to the machine learning.

2.3.1 Local Maximal Quasi-Clique Merger

The algorithm of local maximal Quasi-Clique Merger (lmQCM) was proposed and explained in [34]. With P number of features, given an undirected weighted network $G = \{V_G, E_G, W_G\}$ with vertices $V_G = \{v_1, v_2, \cdots, v_P\}$, non-negative weights $W_G = \{w_{ij}\}$ on the edges e_{ij} where $w_{ij} = w_{ji} \geq 0$, $w_{ii} = 0$ (not allowing self-loop) for $i, j \in \{1, 2, \cdots, P\}$. The density is defined as

$$d_G = \frac{\sum_{j=i+1}^{P} \sum_{i=1}^{P-1} w_{ij}}{\frac{P(P-1)}{2}}. \tag{2.27}$$

In order to find densely connected network modules/subgraphs of G, the original QCM algorithm adopted a greedy based approach. The nodes which contribute to the network module density are aggregated to the highest weight edge in current graph [45], until the module density below certain threshold. The threshold is associated with module size. Since

the identified densely connected modules may have overlaps, a merging process is further applied.

The algorithm of lmQCM has four parameters: q_γ, q_λ, q_t, and q_β. q_γ controls the threshold for initialize each new module, q_λ and q_t are the parameters for the module density threshold. q_β is the threshold for the merging overlap ratio. Given two modules ϕ_U and ϕ_V, the module will be merged if $\frac{|\phi_U \bigcap \phi_V|}{\min(|\phi_U|,|\phi_V|)} > q_\beta$.

lmQCM has been successfully adopted for gene co-expression analyses with numerous biomedical applications [2], [4], [7], [46].

2.3.2 Generating Eigengene from Gene Co-expression Analysis Result

After gene co-expression analysis is done, eigengenes [47] are extracted given the gene expression matrix and the co-expression modules information. Specifically, the Singular Value Decomposition (SVD) [48], [49] is applied to each co-expressed gene module, and the first right singular vector of each SVD result is chosen.

Suppose a gene expression matrix is grouped into K modules, each with $P^{(1)}$, $P^{(2)}$, $\cdots$, and $P^{(K)}$ genes, respectively. Then for each i$^{\text{th}}$ group, the matrix $\boldsymbol{X}^{(\text{i})}$ of dimension $P^{(\text{i})} \times N$ will be decomposed into $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ by SVD, where $\boldsymbol{U}$ is a $P^{(\text{i})}$ by $P^{(\text{i})}$ matrix, $\boldsymbol{\Sigma}$ is a $P^{(\text{i})}$ by N non-negative, real valued diagonal matrix, and $\boldsymbol{V}$ is an N by N matrix. Then the eigengene of that module i is the first right singular vector $\boldsymbol{V}_{1,\cdot}$.

The eigengene can be treated as the patient summary for that gene module. It projects co-expressed genes to 1-D space and thus can be treated as the "super gene". Such gene module-level summaries can then be used for numerous analyses including module-level survival prognosis [2], [50] and enrichment analysis [46], [51].

In this thesis, ARPACK [52] solver is adopted to solve SVD. In addition, to avoid confusion with other part of this thesis, $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{V}$ are defined only in this section.

2.4 Feature Engineering and Overfitting in Survival Prediction

In machine learning, overfitting is one of the common issues that can lead to poor model performance [53]. The Principle of Parsimony, also known as Occam's Razor [54], suggests

that one should not introduce more predictors (or model parameters) if a simple model can sufficiently distinguish label classes, as a large model with many parameters may not improve the performance, and may also lead to undesirable results. Since the number of available training data is typically small in biomedical datasets, it is imperative to control the complexity of the machine learning model.

In this section, we will first introduce how to avoid overfitting by introducing eigengenes as neural networks input and help simplify the model, which is then adopted in Chapter 3. Moreover, other approaches to avoiding overfitting in neural networks will also be discussed.

2.4.1 Using Eigengene as Neural Networks Input to Avoid Overfitting

Previous work [55]–[58] suggests that large and complex neural networks tend to have insufficient learning capacity. Consider a neural network with N training samples, M total number of nodes, W total number of weights, and $\epsilon \in (0, 1/8]$ as the accuracy parameter, then if the distribution of training samples and testing samples are the same, Baum *et al.* [55] showed that when $N > O(\frac{W}{\epsilon} \log \frac{M}{\epsilon})$ samples are used for neural network training and at least $1 - \frac{\epsilon}{2}$ of the samples are classified correctly at the training stage, then the neural network will correctly classify $1 - \epsilon$ number of testing samples. When there is a small group of N training samples, a generally smaller M and W can help to classify more fraction of test samples correctly. In survival prognosis studies, using eigengenes instead of all gene expressions as input can greatly reduce the number of parameters M and W in a machine learning model, and thus can avoid overfitting, especially when few samples are available in the training dataset. Nevertheless, using eigengenes can also help to analyze survival associated genes in module-level.

To avoid confusion with other parts of this thesis, W, M, and ϵ are defined only in this section.

2.4.2 Other Approaches to Avoid Overfitting in Neural Networks

Instead of using eigengenes as neural network inputs to avoid overfitting, we discuss some other approaches to avoid overfitting, including: reducing number of nodes and weights in

neural networks, cross-validation, early stop, regularization, dropout, data augmentation, and feature selection. These practices applied to either data or model can improve model predictability in survival prognosis tasks.

Reducing Number of Layers and Nodes

According to the Principle of Parsimony [54], complex neural networks tend to overfit the data. So instead of using a complex and deep neural networks model, decreasing the number of layers and nodes can greatly reduce the neural network complexity thus avoiding overfitting.

Cross-validation

Cross-validation, or k-fold cross-validation [59], is a repeated learning and testing approach for improving model robustness and correcting learning bias. In practice, the training set will be first split into k groups with equal sizes. Next, for each iteration one of the k groups is treated as the hold-out validation set and the remaining samples from the $k-1$ groups are used for model training. Each sample is used only once for testing and $k-1$ times for training. Although cross-validation can help to determine the hyper-parameters used in the machine learning model, it can also be used to avoid overfitting.

Furthermore, leave-one-out cross-validation, a special case of k-fold cross-validation [60] where k equals the number of data samples N, can alternatively be used.

To avoid confusion with other parts of this thesis, symbol k is defined only in this section.

Early Stop

Early stop often works with cross-validation to improve model performance and avoid overfitting [61]. It helps to determine the optimal number of epochs of the neural networks. Essentially, when we split the dataset into training, validation, and testing sets, the optimal learning model is achieved when a minimum error is observed in the validation set at a certain iteration/epoch.

Regularization

One of the typical approaches for simplifying neural networks is regularization. Regularization of network parameters Θ includes L1 (or LASSO) regularization [62], L2 (or Ridge) regularization [63], and elastic net regularization [64].

L1 regularization minimizes the sum of absolute values of all model weights, $\|\Theta\|$. It helps to push the weights towards zero if the associated covariates are less predictive. The L2 regularization $\|\Theta\|_2^2$ minimizes the sums of the square of the values of all model weights, and reduces the parameter values especially those with a larger impact. Elastic net regularization combines both L1 and L2 regularization linearly with a parameter λ, $\lambda\|\Theta\| + (1-\lambda)\|\Theta\|_2^2$, which can either reduce or eliminate the non-important parameters.

In practice, regularization may be used partially, for example on the parameters of the last hidden layer of the neural networks, depending on the different study designs.

To avoid confusion with other parts of this thesis, symbol λ is defined only in this section.

Dropout

In addition to regularization, dropout [65] is another approach to avoid overfitting that manipulates network weights as well. It is used typically during the training stage, by discarding network units with a certain probability $\alpha \in [0, 1)$. When model is in the validation or testing stage, dropout will not affect the network units. It has been shown empirically that using dropout in neural networks can reduce the classification error, thus making the networks more robust and avoid overfitting [65].

To avoid confusion with other parts of this thesis, symbol α is defined only in this section.

Data Augmentation

It has been shown that the number of the training data samples is directly associated with model accuracy and overfitting [55]. When only a small training set is available, data augmentation [66] can be used to generate more data to avoid overfitting. It aims to increase the size of training data. In image-based learning tasks, data augmentation includes rotating,

flipping, rescaling, and shifting. In other machine learning tasks, perturbing the original data with noise is also one of the approaches to implement data augmentation.

Feature Selection

Similar to the data augmentation, feature selection is another approach to avoid overfitting [67]. However, instead of increasing the number of training samples, feature selection aims to reduce the input dimension (number of features) by choosing a subset of the input features. One of the feature selection practices is to remove each feature one by one, and re-train the model, thus eliminate redundant features. Feature selection is computationally expensive, since it requires repeated training.

In survival prognosis tasks, feature selection becomes necessary when the inputs have extremely high dimensions. Instead of selecting features one at a time, using pre-computed eigengene matrices as input can be an alternative feature selection approach.

2.5 Low-rank Approximation via Non-negative Matrix Factorization

Although co-expression analysis is one of the unsupervised approaches for investigating the densely connected gene modules, meanwhile, the family of low-rank approximation, including eigenvalue methods [68], [69] and non-negative matrix factorization (NMF) methods, presented the power on dimensionality reduction and clustering properties [70]. In Chapter 4 of this thesis, we are particularly interested in using NMF to identify connected gene clusters given the non-negative gene expression values.

NMF, studied since 1999 [71], is a family of algorithms that can decompose a non-negative matrix $\boldsymbol{X}$ into two low-rank matrices: a basis matrix $\boldsymbol{W}$ representing features, and a coefficient matrix $\boldsymbol{H}$ representing samples/patients. Initially aimed for decomposing images especially human faces [72]–[75], the use of NMF was then applied to biological analysis such as human gene clustering [76]–[80]. It is one of the general matrix factorization methods, with the power of better interpretability and the clustering property [70]. Different from other matrix factorization methods, the imposed non-negative property on $\boldsymbol{W}$ and $\boldsymbol{H}$ can reflect biological interpretations [71]. In this section, the concept of NMF will be introduced

first. Then, we demonstrate an update rule for optimizing NMF. The variations of NMF and other low-rank approaches will also be introduced.

2.5.1 Formulation of Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) decomposes a matrix $\boldsymbol{X}$ into two low-rank matrices $\boldsymbol{W}$ and $\boldsymbol{H}$, while all the elements in $\boldsymbol{X}$, $\boldsymbol{W}$, and $\boldsymbol{H}$ are non-negative. The rank K in the definition of NMF can be determined or undetermined. Suppose the target $\boldsymbol{X}$ is a P by N matrix, given a determined low-rank $K \ll N$, we aim to decompose the matrix $\boldsymbol{X}$ to a P by K basis matrix $\boldsymbol{W}$, and a K by N coefficient matrix $\boldsymbol{H}$, such that the estimation of the target matrix is then $\hat{\boldsymbol{X}} = \boldsymbol{W}\boldsymbol{H}$, and should be as similar to $\boldsymbol{X}$ as possible. The formal definition of NMF can be generalized as an optimization problem in the form of

$$\text{Minimize } \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F, \tag{2.28}$$

subject to $\boldsymbol{X}_{\mathrm{i,j}} \geq 0 \ \forall \mathrm{i} \in [0, P], \mathrm{j} \in [0, N]$, $\boldsymbol{W}_{\mathrm{i,j}} \geq 0 \ \forall \mathrm{i} \in [0, P], \mathrm{j} \in [0, K]$, and $\boldsymbol{H}_{\mathrm{i,j}} \geq 0$ $\forall \mathrm{i} \in [0, K], \mathrm{j} \in [0, N]$. Where $\|\cdot\|_F$ is the Frobenius norm, also known as Euclidean distance [81].

Another useful objective function uses Kullback-Leibler divergence (KL divergence) as the loss function by enforcing $\sum_{\mathrm{i,j}} \boldsymbol{X}_{\mathrm{i,j}} = 1$ and $\sum_{\mathrm{i,j}} \hat{\boldsymbol{X}}_{\mathrm{i,j}} = \sum_{\mathrm{i,j}} (\boldsymbol{W}\boldsymbol{H})_{\mathrm{i,j}} = 1$ and consider they are normally distributed [81]. The optimization problem is then written in

$$\text{Minimize } \mathrm{D}(\boldsymbol{X}\|\hat{\boldsymbol{X}}) = \sum_{\mathrm{i,j}} \boldsymbol{X}_{\mathrm{i,j}} \log\left(\frac{\boldsymbol{X}_{\mathrm{i,j}}}{\hat{\boldsymbol{X}}_{\mathrm{i,j}}}\right) - \boldsymbol{X}_{\mathrm{i,j}} + \hat{\boldsymbol{X}}_{\mathrm{i,j}}, \tag{2.29}$$

subject to $\boldsymbol{X}_{\mathrm{i,j}} \geq 0 \ \forall \mathrm{i} \in [0, P], \mathrm{j} \in [0, N]$, $\boldsymbol{W}_{\mathrm{i,j}} \geq 0 \ \forall \mathrm{i} \in [0, P], \mathrm{j} \in [0, K]$, and $\boldsymbol{H}_{\mathrm{i,j}} \geq 0$ $\forall \mathrm{i} \in [0, K], \mathrm{j} \in [0, N]$. Similar to the equation (2.28), the KL divergence $\mathrm{D}(\boldsymbol{X}\|\hat{\boldsymbol{X}})$ vanished if and only if $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H}$.

In this thesis, we are interested in minimizing Frobenius norm in equation (2.28) to solve the NMF problem.

2.5.2 Optimization of NMF

Optimizing NMF in equation (2.28) is non-convex. However, the problem can be transformed into a convex problem if we fix either $\boldsymbol{H}$ and update $\boldsymbol{W}$ or vice versa. There are many ways of optimizing NMF, such as using coordinate descent [82] or multiplicative update [81]. Here we present the multiplicative update rule as it will be the foundation of our proposed algorithm in Chapter 4.

Multiplicative Update

Lemma 2.5.1. *For a matrix* $\boldsymbol{A}$*,*

$$\|\boldsymbol{A}\|^2 = Tr(\boldsymbol{A}^T\boldsymbol{A}). \tag{2.30}$$

Proof.

$$\begin{aligned} &\text{because } (\boldsymbol{A}^T\boldsymbol{A})_{\mathrm{i,j}} = \sum_k \boldsymbol{A}^T_{\mathrm{i},k}\boldsymbol{A}_{k,\mathrm{j}} = \sum_k \boldsymbol{A}_{k,\mathrm{i}}\boldsymbol{A}_{k,\mathrm{j}}, \\ &\text{thus, } Tr(\boldsymbol{A}^T\boldsymbol{A}) = \sum_{\mathrm{i}}(\boldsymbol{A}^T\boldsymbol{A})_{\mathrm{i,i}} = \sum_{\mathrm{i}}\sum_k(\boldsymbol{A}_{k,\mathrm{i}})^2 = \|\boldsymbol{A}\|^2. \end{aligned} \tag{2.31}$$

□

Definition 2.5.1. *Given the target non-negative matrix* $\boldsymbol{X}$*, two initialized non-negative matrices* $\boldsymbol{W}$ *and* $\boldsymbol{H}$*, the non-negative matrix factorization (NMF) multiplicative update rule can be written as an alternately iterative update algorithm and can guarantee both* $\boldsymbol{W}$ *and* $\boldsymbol{H}$ *be non-negative:*

$$\boldsymbol{W}^{(\mathrm{iter}+1)}_{\mathrm{i,j}} \leftarrow \boldsymbol{W}^{(\mathrm{iter})}_{\mathrm{i,j}} \odot \frac{\boldsymbol{X}\boldsymbol{H}^{(\mathrm{iter})^T}_{\mathrm{i,j}}}{\boldsymbol{W}^{(\mathrm{iter})}_{\mathrm{i,j}}\boldsymbol{H}^{(\mathrm{iter})}_{\mathrm{i,j}}\boldsymbol{H}^{(\mathrm{iter})^T}_{\mathrm{i,j}}}, \tag{2.32}$$

$$\boldsymbol{H}^{(\mathrm{iter}+1)}_{\mathrm{i,j}} \leftarrow \boldsymbol{H}^{(\mathrm{iter})}_{\mathrm{i,j}} \odot \frac{\boldsymbol{W}^{(\mathrm{iter}+1)^T}_{\mathrm{i,j}}\boldsymbol{X}}{\boldsymbol{W}^{(\mathrm{iter}+1)^T}_{\mathrm{i,j}}\boldsymbol{W}^{(\mathrm{iter}+1)}_{\mathrm{i,j}}\boldsymbol{H}^{(\mathrm{iter})}_{\mathrm{i,j}}}. \tag{2.33}$$

The value for the current iteration is denoted with (iter) *superscript. Note that the divisions is element-wised. For the sake of simplicity, we omit the iteration and element-wised notation, and rewrite the updating rules as*

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\boldsymbol{X}\boldsymbol{H}^T}{\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T}, \tag{2.34}$$

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \odot \frac{\boldsymbol{W}^T\boldsymbol{X}}{\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}}. \tag{2.35}$$

Note: $\odot$ *denotes the Hadamard product (element-wise product), and the divisions in equation* (2.32) *and* (2.33) *are element-wised.*

Proof.

Since the NMF problem is convex if we fix $\boldsymbol{W}$ and update $\boldsymbol{H}$ or vice versa, we need the derivatives of previous equations on $\boldsymbol{W}$ and $\boldsymbol{H}$. The update rule shall be written as

$$\begin{aligned} \boldsymbol{W} &\leftarrow \boldsymbol{W} - \eta_{\boldsymbol{W}} \cdot \nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}), \\ \boldsymbol{H} &\leftarrow \boldsymbol{H} - \eta_{\boldsymbol{H}} \cdot \nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}). \end{aligned} \tag{2.36}$$

Instead of minimizing $\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F$, we start from minimizing $\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2 = f(\boldsymbol{W}, \boldsymbol{H})$. According to Lemma (2.5.1),

$$\begin{aligned} \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2 &= Tr((\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H})^T(\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H})) \\ &= Tr((\boldsymbol{X}^T - \boldsymbol{H}^T\boldsymbol{W}^T)(\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H})) \\ &= Tr(\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{H} - \boldsymbol{H}^T\boldsymbol{W}^T\boldsymbol{X} + \boldsymbol{H}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}) \\ &= Tr(\boldsymbol{X}^T\boldsymbol{X}) - Tr(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{H} - Tr(\boldsymbol{H}^T\boldsymbol{W}^T\boldsymbol{X} + Tr(\boldsymbol{H}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}). \end{aligned} \tag{2.37}$$

Starting from $\boldsymbol{W}$, to get the equation of $\nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H})$, the partial derivative of each term in equation (2.37) with respect to $\boldsymbol{W}$ are:

$$\nabla_{\boldsymbol{W}} Tr(\boldsymbol{X}^T\boldsymbol{X}) = 0; \tag{2.38}$$

$$\nabla_{\boldsymbol{W}} Tr(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{H}) = \nabla_{\boldsymbol{W}} Tr(\boldsymbol{H}\boldsymbol{X}^T\boldsymbol{W}) = (\boldsymbol{H}\boldsymbol{X}^T)^T = \boldsymbol{X}\boldsymbol{H}^T; \tag{2.39}$$

$$\nabla_{\boldsymbol{W}} Tr(\boldsymbol{H}^T\boldsymbol{W}^T\boldsymbol{X}) = \nabla_{\boldsymbol{W}} Tr(\boldsymbol{X}\boldsymbol{H}^T\boldsymbol{W}^T) = \boldsymbol{X}\boldsymbol{H}^T; \tag{2.40}$$

$$\begin{aligned} \nabla_{\boldsymbol{W}} Tr(\boldsymbol{H}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}) &= \nabla_{\boldsymbol{W}} Tr(\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T\boldsymbol{W}^T) \\ &= \boldsymbol{W}((\boldsymbol{H}\boldsymbol{H}^T)^T + \boldsymbol{H}\boldsymbol{H}^T) \\ &= 2\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T. \end{aligned} \tag{2.41}$$

The partial derivative of $\nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H})$ can be computed similarly. Therefore, we have

$$\begin{aligned} \nabla_{\boldsymbol{W}} f(\boldsymbol{W}, \boldsymbol{H}) &= -2\boldsymbol{X}\boldsymbol{H}^T + 2\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T, \\ \nabla_{\boldsymbol{H}} f(\boldsymbol{W}, \boldsymbol{H}) &= -2\boldsymbol{W}^T\boldsymbol{X} + 2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}. \end{aligned} \tag{2.42}$$

Now, we can write the equation (2.36) as

$$\begin{aligned} \boldsymbol{W} &\leftarrow \boldsymbol{W} + \eta_{\boldsymbol{W}} \cdot (2\boldsymbol{X}\boldsymbol{H}^T - 2\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T), \\ \boldsymbol{H} &\leftarrow \boldsymbol{H} + \eta_{\boldsymbol{H}} \cdot (2\boldsymbol{W}^T\boldsymbol{X} - 2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}). \end{aligned} \tag{2.43}$$

In order to guarantee all elements in $\boldsymbol{W}$ and $\boldsymbol{H}$ are non-negative during the updates, it is important to remove the subtraction in (2.43). The way of doing this is by introducing adaptive learning rate $\eta_{\boldsymbol{W}} = \frac{\boldsymbol{W}}{2\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T}$ and $\eta_{\boldsymbol{H}} = \frac{\boldsymbol{H}}{2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}}$. Note that the division is element-wised as well.

Thus, passing the learning rate $\eta_{\boldsymbol{W}} = \frac{\boldsymbol{W}}{2\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T}$ and $\eta_{\boldsymbol{H}} = \frac{\boldsymbol{H}}{2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}}$ into equation (2.36), we will get the multiplicative update rule (2.32) and (2.33).

□

2.5.3 Variations of NMF

As non-negative matrix factorization becomes a hot topic, there are many variations on non-negative matrix factorization. Some researchers introduced orthogonal constraint on basis $\boldsymbol{W}$ and coefficient $\boldsymbol{H}$ to constraint the learning process, named it as orthogonal NMF [83], [84]. Others introduced Fisher discriminant [85] into NMF and enforced the separability into coefficient matrix $\boldsymbol{H}$ and named it as discriminant NMF [72], [74], etc.

Orthogonal NMF

The algorithm of Orthogonal NMF (ONMF) was first systematically carried out and summarized by Ding-Ti-Peng-Park (DTPP) algorithm [83]. Introducing orthogonality into NMF problem can have several advantages including the uniqueness of the solution and better interpretations on clustering.

Orthogonality constraint can be applied on either $\boldsymbol{W}$ or $\boldsymbol{H}$ (one-sided). Take the example of making orthogonal constraint on $\boldsymbol{W}$, the objective function is

$$\text{Minimize } \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F, \tag{2.44}$$

subject to $\boldsymbol{X}_{\text{i,j}} \geq 0; \boldsymbol{W}_{\text{i,j}} \geq 0; \boldsymbol{H}_{\text{i,j}} \geq 0$, $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$.

Alternatively, orthogonality constraint can also be applied on both $\boldsymbol{W}$ and $\boldsymbol{H}$:

$$\text{Minimize } \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F, \tag{2.45}$$

subject to $\boldsymbol{X}_{\text{i,j}} \geq 0; \boldsymbol{W}_{\text{i,j}} \geq 0; \boldsymbol{H}_{\text{i,j}} \geq 0$, $\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}$, and $\boldsymbol{H}^T\boldsymbol{H} = \boldsymbol{I}$. We refer the algorithms of ONMF and the proof of uniqueness solution to the original article [83].

Discriminant NMF

Discriminant NMF, or DNMF in short, was initially proposed in 2006 [72] by imposing fisher discriminant to the coefficient matrix $\boldsymbol{H}$. However, imposing fisher discriminant can be backtracked to Fisherfaces [86] in 1997 and Fisher NMF [73] in 2005. The idea of DNMF is to

enhance the separability between classes in a discriminant manner [72] as well as maintaining the NMF optimization constraint. Later work such as PGDNMF [74], dNMF [87], SDNMF [75], DPNMF [88], and NDMF [89] are also came out with their algorithms and derivations to solve this problem. Meanwhile, it was soon applied to biomedical problems such as gene ranking [90].

Although DNMF can apply to many biomedical problems, in Chapter 4 of this thesis, the CoxNMF algorithm we are going to propose is out of the scope of what DNMF can handle. Similar to the DNMF algorithm, CoxNMF algorithm has its own properties of regressing the survival data simultaneously with NMF updating process.

Supervised NMF

Supervised NMF (or SNMF) [91], is another NMF variant that introduced discriminant ability. It imposes linear regression into the NMF optimization process. Given data matrix $\boldsymbol{X}$ and label vector ϕ, N number of patients, K low-rank dimensions, initial basis matrix $\boldsymbol{W}$ and coefficient matrix $\boldsymbol{H}$, and initial K by 1 weight vector w and bias vector b, the SNMF can be formulated as

$$\begin{aligned}\text{Minimize } & \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\|_F^2 + s_\alpha \sum_{\mathrm{i}=1}^{N} \ln\left(1+\exp\left(-\phi_{\mathrm{i}}\sum_{\mathrm{j}=1}^{K}(w_{\mathrm{j}}\boldsymbol{H}_{\mathrm{j,i}}+b_{\mathrm{j}})\right)\right) \\ & + \frac{1}{2}s_\beta\sum_{\mathrm{j}=1}^{K}(w_{\mathrm{j}}^2+b_{\mathrm{j}}^2) + \frac{1}{2}s_\gamma\|\boldsymbol{H}\|_F^2\,.\end{aligned} \tag{2.46}$$

In Equation 2.46, linear regression is integrated into the NMF framework. $\boldsymbol{H}$ is the new representation they aimed to learn based on linear regression. s_α, s_β, and s_γ are used to balance the corresponding terms. We refer the details of the updating rule to the original article [91].

2.5.4 Other Low-Rank Approximation Methods

In this thesis, except the NMF approach will be used to design CoxNMF algorithm, other low-rank approximation algorithms will also be compared with CoxNMF. These low-

rank approximation algorithms including Truncated singular value decomposition (truncated SVD), non-negative double singular value decomposition (NNDSVD), principal component analysis (PCA), sparse PCA, and factor analysis.

When performing low-rank approximation, truncated SVD is one of the fundamental approach. In truncated SVD, the low-rank approximation of $\boldsymbol{X}$ is now becomes $\hat{\boldsymbol{X}} = \boldsymbol{U}_K \boldsymbol{\Sigma}_K \boldsymbol{V}_K^*$, where K is the rank, and the P by K matrix $\boldsymbol{U}$, K by N matrix $\boldsymbol{V}^*$, are corresponding to the K largest singular values $\boldsymbol{\Sigma}_K$. It can also be written as $\hat{\boldsymbol{X}} = \sum_{\mathrm{j}=1}^{K} \sigma_{\mathrm{j}} \boldsymbol{U}_{\cdot,\mathrm{j}} \boldsymbol{V}_{\mathrm{j},\cdot}^T$, where $\boldsymbol{U}_{\cdot,\mathrm{j}}$ and $\boldsymbol{V}_{\mathrm{j},\cdot}$ are the left and right singular vectors associate with the singular value σ_{j}.

Non-negative double singular value decomposition [92], or NNDSVD in short, is another variant of SVD. It is based on approximating the input matrix $\boldsymbol{X}$ and positive sections of the SVD results, and is widely used for initializing NMF solutions $\boldsymbol{W}$ and $\boldsymbol{H}$. Given non-negative input matrix $\boldsymbol{X}$, the NNDSVD first computes K leading singular triplets of $\boldsymbol{X}$, then forms the unit rank matrices $\{\boldsymbol{C}^{(\mathrm{j})}\}_{\mathrm{j}=1}^{K}$, where $\boldsymbol{C}^{(\mathrm{j})} = \boldsymbol{U}_{\cdot,\mathrm{j}} \boldsymbol{V}_{\mathrm{j},\cdot}^T$. Finally, NNDSVD will use the positive section of the $\boldsymbol{C}^{(\mathrm{j})}$ as the results.

Principal component analysis [93], [94], or PCA in short, is an orthogonal linear transformation method. It projects the input $\boldsymbol{X}$ into a new coordinate system, where the i^{th} highest variance of scalar projection lies on the i^{th} coordinate. Different from the truncated SVD which performs the factorization on the data matrix, PCA performs the factorization on the covariance matrix. In addition, sparse PCA [95] is a variant of PCA which enables sparse coding.

Factor analysis is a linear Gaussian latent variable model related to PCA [96]. In matrix notation, it can be written as $\boldsymbol{X} = \boldsymbol{M} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon}$, where $\boldsymbol{M} \in \mathcal{R}^{P \times N}$ is the mean matrix of $\boldsymbol{X}$, $\boldsymbol{L} \in \mathcal{R}^{P \times K}$ is the loading matrix, $\boldsymbol{F} \in \mathcal{R}^{K \times N}$ is the factor matrix, and $\boldsymbol{\epsilon} \in \mathcal{R}^{P \times N}$ is the error term matrix.

To avoid confusion with other parts of this thesis, symbols and notations $\boldsymbol{U}$, $\boldsymbol{\Sigma}$, $\boldsymbol{V}$, σ, r, $\boldsymbol{L}$, $\boldsymbol{F}$, $\boldsymbol{M}$, and $\boldsymbol{\epsilon}$ are defined only in this section.

Comparing to other low-rank approximation methods, NMF provides new insights about complex latent relationships in high-dimensional biomedical data [97]. It decomposes a non-negative matrix $\boldsymbol{X}$ into two low-rank matrices representing features and samples, provides a

well-established geometric and topological representation of the feature space by visualizing the basis matrix. Given the interpretable characteristics that NMF presented, we develop CoxNMF in Chapter 4, an algorithm based on NMF and Cox proportional hazards regression, to unveil latent gene clusters and interactions.

2.6 Histopathologic Image Registration

Histopathologic image is one of the common biomedical data. It is a type of digital microscopy tissue images that used for numerous analyses such as examining disease. Histopathologic tissue image can be stained in different methods for different research purposes, including hematoxylin and eosin (H&E) staining and immunohistochemistry (IHC) staining. When consecutive tissue biopsies stained with different staining approaches, extra information can be exploited from these multimodal tissue images. However, this requires a proper way to align those images before any analysis. In following sections, we will introduce common image registration approaches and the image registration algorithm we adopt for the application described in Chapter 5.

Image registration [98], describes a way to align two or more images geometrically, normally achieved by automatic computational algorithms in computer vision. This process often require one so called "fixed image" (or reference/target image) as a reference, and the remaining are so called "moving images" (or floating/source image). The objective function of image registration can be generally summarized as

$$\hat{T} = \underset{T \in \mathcal{T}}{\operatorname{argmax}}\, S(I_F, I_M, T), \tag{2.47}$$

where T is the transformation, $\mathcal{T}$ is the searching space, I_F is the fixed image, I_M is the moving image, and S is the similarity measure. Typical similarity measures including correlation [99], mutual information [100], etc. To optimize Equation 2.47, one can use gradient descent [101], conjugate gradient method [102], Newton–Raphson method [103], quasi-Newton method [104], or Levenberg-Marquardt method [105], [106], etc. To avoid confusion with other parts of this thesis, symbols T, $\mathcal{T}$, I_F, I_M, and S are defined only in Equation 2.47. Image registration approaches can be concluded into linear or non-linear approaches.

2.6.1 Linear Registration

Linear registration is a family of registration methods that perform image transformation on moving images with linear operations, such as rotation, scaling, shearing, translation, and affine transformations. Suppose given original 2-D data point $\begin{bmatrix} x \\ y \end{bmatrix}$, the linear transformation is formulated as

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2.48}$$

in homogeneous coordinates. Where $\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$ is the 2-D data coordinates after the transformation (with an extra 3$^{\text{rd}}$ coordinate), and we denote $\begin{bmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & 1 \end{bmatrix} = M$ as transformation matrix. The scaling transformation matrix can be represented by

$$M = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.49}$$

where s_x and s_y scale the point $\begin{bmatrix} x \\ y \end{bmatrix}$. The rotation transformation matrix can be represented by

$$M = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.50}$$

where θ is the rotation angle in counterclockwise. The shearing transformation matrix can be represented by

$$M = \begin{bmatrix} 1 & h_x & 0 \\ h_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.51}$$

where h_x, h_y are the shear factor in horizontal and vertical, respectively. Finally, the transformation matrix for translation can be represented by

$$M = \begin{bmatrix} 1 & 0 & \Delta x \\ 0 & 1 & \Delta y \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.52}$$

where Δx and Δy stand for the translation that applied on x and y directions, respectively. To avoid confusion with other parts of this thesis, symbols and notations M, x, y, x, y, a, b, c, d, θ, h, and s are defined only in this section.

2.6.2 Non-linear Registration

Although linear registration is the simplest image registration approaches, it aligns moving images to the fixed image globally without discerning local geometrical differences [107]. In contrast, non-linear registration can handle these local geometrical disparities through various methods.

Given the different research objectives and different kinds of image data, there are many different non-linear registration methods, and new methods are coming up every year.

Among non-linear registration algorithms, demons algorithm [108], [109] is one of the most widely used algorithms for multimodal biomedical image registration tasks. Inspired from Maxwell's demons, demons algorithm assumes the pixels in the fixed image (act as local forces) are able to displace the pixels in the moving image [110]. By applying a displacement

vector (or velocity) $v = (dx, dy)$, the moving image is deformed to each pixel iteratively. Here we introduce the optical flow equation

$$v \cdot \nabla I_F = I_M - I_F, \tag{2.53}$$

where v can be considered as "velocity" in optical flow between two successive image frames I_F and I_M. Inspired from optical flow equation, the original demons algorithm for the n^{th} iteration can be written as

$$v^{(n)} = \frac{(I_M^{(n-1)} - I_F^{(0)})\nabla I_F^{(0)}}{(I_M^{(n-1)} - I_F^{(0)})^2 + |\nabla I_F^{(0)}|^2}, \tag{2.54}$$

where ΔI_F is the gradient of the fixed image, $I_F^{(n)}$ and $I_M^{(n)}$ are the intensity of the fixed and moving images in n^{th} iteration, respectively.

Instead of original version of the demons implementation, there are many other demons variants by modifying Equation 2.54, such as combining passive and active forces [111], [112] (double force method), or adding a normalization factor [113]. In 2021, Wodzinski and Skalski proposed a multistep, automatic and non-linear image registration method [114], which incorporates modality independent neighbourhood descriptor (MIND) [115] into demons algorithm for multimodal histopathologic image deformable registration. After the initial affine registration for multimodal histopathologic images, The modified double force demons algorithm was used in the following non-linear registration. In the modified algorithm, calculating the velocity v for the n^{th} iteration can be formulated as:

$$v^{(n)} = \frac{\nabla I_{M,MIND}^{(n-1)} \cdot (I_{M,MIND}^{(n-1)} - I_{F,MIND}^{(0)})}{|\nabla I_{M,MIND}^{(n-1)}|^2 + (I_{M,MIND}^{(n-1)} - I_{F,MIND}^{(0)})^2} + \frac{\nabla I_{F,MIND}^{(0)} \cdot (I_{M,MIND}^{(n-1)} - I_{F,MIND}^{(0)})}{|\nabla I_{F,MIND}^{(0)}|^2 + (I_{M,MIND}^{(n-1)} - I_{F,MIND}^{(0)})^2}, \tag{2.55}$$

where $I_{M,MIND}$ and $I_{F,MIND}$ are the MIND descriptors of the moving and fixed images, respectively. The complete workflow for the non-linear registration method [114] can be summarized in following 4 steps:

- Step 1. Preprocessing I_M and I_F.

- Step 2. Initial alignment by using SIFT [116], SURF [117], and ORB [118] keypoints and features.
- Step 3. Performing affine registration.
- Step 4. Performing non-linear registration and return the deformation field.

Wodzinski and Skalski's multistep, multimodal image registration workflow [114] has many advantages, including the high accuracy of multimodal histopathologic images alignment, specially designed for whole slide tissue images, and with relatively efficient running time [119]. To avoid confusion with other parts of this thesis, symbols and notations v, I_F, I_M, and n are defined only in this section.

2.7 Histopathologic Image Segmentation

Histopathologic image segmentation is an approach to group regions or objects in the image at pixel-level. The objective of image segmentation is to understand and explain the global context of the image [120]. Given different learning tasks, different algorithms may apply to achieve different research goals.

2.7.1 Color-based K-means Algorithm for Object Segmentation

Color-based K-means clustering is a way to identify regions or objects that characterized by visually distinct colors for applications such as nucleus segmentation [121]. It is a simple and useful approach for segmenting cell markers in histopathologic images.

To perform color-based K-means clustering, the first step is to convert the histopathologic images from RGB color space to L*a*b* color space [122]. L*a*b* color space, or CIE LAB color space, is designed to approximate human vision, the L* component is related to lightness (L* = 0 indicates black and L* = 100 indicates diffuse white), a* component is related to red and green color, and b* component is related to yellow and blue color,

respectively. During the conversion, RGB color space will first need to convert to CIE XYZ space [123]:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = M \begin{bmatrix} R \\ G \\ B \end{bmatrix}, \tag{2.56}$$

where M is the linear transformation matrix. When using sRGB, the transformation matrix is

$$M = \begin{bmatrix} 0.4124564 & 0.3575761 & 0.1804375 \\ 0.2126729 & 0.7151522 & 0.0721750 \\ 0.0193339 & 0.1191920 & 0.9503041 \end{bmatrix}. \tag{2.57}$$

Next, the conversion from CIE XYZ space to L*a*b* space [123] can be formulated as

$$\begin{aligned} \mathrm{L}^* &= 116 f_y - 16, \\ \mathrm{a}^* &= 500 \left(f_x - f_y \right), \\ \mathrm{b}^* &= 200 \left(f_y - f_z \right), \end{aligned} \tag{2.58}$$

where

$$\begin{aligned} f_x &= \begin{cases} \sqrt[3]{\frac{X}{X_n}} & \text{if } \frac{X}{X_n} > \epsilon \\ \frac{\kappa \frac{X}{X_n} + 16}{116} & \text{otherwise} \end{cases}, \\ f_y &= \begin{cases} \sqrt[3]{\frac{Y}{Y_n}} & \text{if } \frac{Y}{Y_n} > \epsilon \\ \frac{\kappa \frac{Y}{Y_n} + 16}{116} & \text{otherwise} \end{cases}, \\ f_z &= \begin{cases} \sqrt[3]{\frac{Z}{Z_n}} & \text{if } \frac{Z}{Z_n} > \epsilon \\ \frac{\kappa \frac{Z}{Z_n} + 16}{116} & \text{otherwise} \end{cases}, \end{aligned} \tag{2.59}$$

and

$$\begin{aligned} \epsilon &= \frac{216}{24389}, \\ \kappa &= \frac{24389}{27}. \end{aligned} \tag{2.60}$$

Note that the reference white (X_n, Y_n, Z_n) for standard illuminant D65 is (95.0489, 100, 108.8840).

After converting each pixel of histopathologic images from RGB space to the L*a*b* space, objects with visually distinct colors are more easily to be discerned by the clustering algorithm. Finally, K-means algorithm will be used to cluster pixels with L*a*b* colors. To avoid confusion with other parts of this thesis, symbols X, Y, Z, R, G, B, M, L, a, b, f_x, f_y, f_z, X_n, Y_n, Z_n, ϵ, and κ are only defined in this section.

2.7.2 Deep Learning-based Algorithm for Semantic Segmentation

Pixel-wised semantic segmentation for objects or regions detection through deep learning can be traced back to fully convolutional neural network (FCN) [124]. It is a non-linear deep learning operation that assign every pixel in an image to different categorical labels [125]. Based on fully convolutional neural network, deep learning is currently the most prevalent and preferred approach for semantic segmentation, which has been validated in several benchmark datasets [125]. Since then, significant effort has been committed to study the semantic segmentation in biomedical images, which can be used to improve image-based interventions and post-treatment outcome predictions [120]. The proposed deep learning-based algorithms such as U-Net [126] and DeepLabV3 [127] are especially well-suited for biomedical image segmentation tasks.

The backbone of the U-Net deep learning architecture is based on an encoder-decoder like skeleton [120], [126] with the main idea of skip connections in the network. The skip connections can avoid vanishing gradient issue thus can achieve a higher segmentation accuracy.

Another deep learning algorithm "DeepLabV3" [127] adopts atrous convolution instead at multiple scales [120]. Atrous convolution [128], [129], also known as dilated convolution, is a convolutional kernel with a larger value of stride parameter. It increases model's field of view, thus enlarges the object encoding. The formulation of atrous convolution in 2-D Euclidean space is

$$y[\mathrm{i},\mathrm{j}] = \sum_{m=-M}^{M} \sum_{n=-N}^{N} x[\mathrm{i}+s\cdot m, \mathrm{j}+s\cdot n]\cdot w[m,n], \tag{2.61}$$

where s is the stride parameter, i, j are the coordinates in horizontal and vertical directions, w is the convolution kernel weight. Typically, a 3×3 atrous convolution has $M = N = 1$.

DeepLabV3 can be further constructed based on deep residual neural network (ResNet) backbone [130] for a better segmentation accuracy [127]. This implementation is also available in PyTorch deep learning package [131]. In the later part of this thesis, we adopt DeepLabV3 for H&E stained histopathologic image segmentation to identify tumoral region, stromal region, and lymphocytes aggregated region.

2.8 Evaluation Metrics and Statistical Tests

2.8.1 Concordance Index

The concordance index, also known as C-Index, is a generalization of the area under the ROC curve (AUC) which introduces the censorship information. It assesses the model discrimination power of the ability to correctly provide a reliable ranking of the survival times based on the individual risk scores. It can be computed with the formula

$$\text{C-Index} = \frac{\sum_{\text{i,j}} \mathbb{1}_{(Y_\text{j}<Y_\text{i})} \cdot \mathbb{1}_{(r_\text{j}>r_\text{i})} \cdot C_\text{j}}{\sum_{\text{i,j}} \mathbb{1}_{(Y_\text{j}<Y_\text{i})} \cdot C_\text{j}}, \tag{2.62}$$

where r_j is the risk score of a subject j. $\mathbb{1}_{(Y_\text{j}<Y_\text{i})}$ is the indicator function: $\mathbb{1}_{(Y_\text{j}<Y_\text{i})} = \begin{cases} 1 & \text{if } Y_\text{j} < Y_\text{i} \\ 0 & \text{otherwise} \end{cases}$. Similar to the AUC, C-Index = 1 corresponds to the best model prediction, and C-Index = 0.5 represents a random prediction.

2.8.2 Dice Coefficient

The Dice coefficient is formulated as

$$Dice = \frac{2 \times TP}{(TP + FP) + (TP + FN)}, \tag{2.63}$$

where TP stands for true positive, FP stands for false positive, FN stands for false negative. Similarly, TN stands for true negative, though TN is not used in $Dice$. Comparing to the

accuracy, Dice coefficient is more sensitive to true positives, which are the labels we are interested in.

2.8.3 Silhouette Score

The silhouette score measures the consistency within clusters of data. The score describes how well each element has been classified [132]. The formal definition is described as follows.

Definition 2.8.1. *Assuming the data* $\boldsymbol{W}$ *have been clustered into* K *clusters by an algorithm. For data point* $\mathrm{i} \in \Omega_k$ *in cluster* k*, let*

$$a(\mathrm{i}) = \frac{1}{|\Omega_k| - 1} \sum_{\mathrm{j} \in \Omega_k, \mathrm{i} \neq \mathrm{j}} d(\mathrm{i}, \mathrm{j}) \tag{2.64}$$

be the mean distance between i *and other data points within the same cluster* Ω_k*,* $d(\mathrm{i}, \mathrm{j})$ *is the distance between data points* i *and* j*.* $|\Omega_k|$ *is the number of elements in cluster* k*.*

Next, the mean dissimilarity of point i *to other cluster* $\Omega_{r, r \neq k}$ *is defined as the mean of the distance from* i *to all data points in* $\Omega_{r, r \neq k}$*. Then the smallest mean distance with respect to* i *to all data points in any other cluster is defined as*

$$b(\mathrm{i}) = \min_{r \neq k} \frac{1}{|\Omega_r|} \sum_{\mathrm{j} \in \Omega_r} d(\mathrm{i}, \mathrm{j}). \tag{2.65}$$

Then the silhouette value for data point i *in cluster* Ω_k *is defined as*

$$s(\mathrm{i}) = \frac{b(\mathrm{i}) - a(\mathrm{i})}{\max\{a(\mathrm{i}), b(\mathrm{i})\}}, \mathrm{i} f |\Omega_k| > 1, \tag{2.66}$$

and

$$s(\mathrm{i}) = 0, \mathrm{i} f |\Omega_k| > 1. \tag{2.67}$$

It can also be written as:

$$s(\mathrm{i}) = \begin{cases} 1 - \frac{a(\mathrm{i})}{b(\mathrm{i})} & \text{if } a(\mathrm{i}) < b(\mathrm{i}) \\ 0 & \text{if } a(\mathrm{i}) = b(\mathrm{i}) \\ \frac{b(\mathrm{i})}{a(\mathrm{i})} - 1 & \text{if } a(\mathrm{i}) > b(\mathrm{i}) \end{cases}. \tag{2.68}$$

Given the silhouette value for data point i*, the mean silhouette coefficient, or silhouette score for all samples, is then given by:*

$$\tilde{s} = \frac{1}{\sum_{k=1}^{K} |\Omega_k|} \sum_{k=1}^{K} \sum_{\mathrm{i} \in \Omega_k} s(\mathrm{i}). \tag{2.69}$$

In this thesis, Euclidean distance is adopted for distance metric $d(\mathrm{i}, \mathrm{j})$.

To avoid confusion with other parts of this thesis, symbols and notations d, r, k, a, b, and s are defined only in this section.

2.8.4 Log-rank Test and P-value

Log-rank test [133] is a non-parametric hypothesis testing to compare the estimates of the hazard functions of two groups [134]–[136]. The corresponding P-value is derived from the Chi-square test for independence. The Chi-square test for independence compares two groups of data in a contingency table and evaluate whether they are related or not. The larger the Chi-square test statistic, the less relationship between one group to another. The formula for Chi-square test is

$$\chi^2(dof) = \sum_{\mathrm{j}} \frac{(O_{\mathrm{j}} - E_{\mathrm{j}})^2}{E_{\mathrm{j}}}, \tag{2.70}$$

where the "dof" stands for the degrees of freedom, which is the number of classes minus 1. While O_{j} is the observed value (death or not) and E_{j} is the expected value for every single time t_{j}.

Suppose we divide the entire dataset into two groups A and B. Given the observed value O_{j}^A (death: 1; alive: 0) for group A and O_{j}^B for group B for every time t_{j}, and N_{j} stands for the number of censored (alive) patients at time t_{j}, the expected value for group A and B are given by:

$$\begin{aligned} E_{\mathrm{j}}^A &= N_{\mathrm{j}}^A \cdot \frac{O_{\mathrm{j}}}{N_{\mathrm{j}}}, \\ E_{\mathrm{j}}^B &= N_{\mathrm{j}}^B \cdot \frac{O_{\mathrm{j}}}{N_{\mathrm{j}}}, \end{aligned} \tag{2.71}$$

where

$$\begin{aligned} O_{\mathrm{j}} &= O_{\mathrm{j}}^A + O_{\mathrm{j}}^B, \\ N_{\mathrm{j}} &= N_{\mathrm{j}}^A + N_{\mathrm{j}}^B. \end{aligned} \tag{2.72}$$

It is also worth to know that $E_{\mathrm{j}}^A + E_{\mathrm{j}}^B = O_{\mathrm{j}}^A + O_{\mathrm{j}}^B$, thus $E_{\mathrm{j}}^B = (O_{\mathrm{j}}^A + O_{\mathrm{j}}^B) - E_{\mathrm{j}}^A$. The log-rank test statistic is then

$$LR = \frac{(\sum_{\mathrm{j}} O_{\mathrm{j}}^A - \sum_{\mathrm{j}} E_{\mathrm{j}}^A)^2}{\sum_{\mathrm{j}} E_{\mathrm{j}}^A} + \frac{(\sum_{\mathrm{j}} O_{\mathrm{j}}^B - \sum_{\mathrm{j}} E_{\mathrm{j}}^B)^2}{\sum_{\mathrm{j}} E_{\mathrm{j}}^B}. \tag{2.73}$$

If the two survival distributions for group A and group B are the same, *i.e.*, the null hypothesis is true, then the log-rank test statistic LR will have a Chi-square distribution with $dof = 1$:

$$LR \sim \chi^2(1), \tag{2.74}$$

and if the corresponding P-value from the Chi-square test is not significant (usually ≥ 0.05), then we cannot reject the null hypothesis, implying that group A and group B are statistically the same.

2.8.5 Student's t-test and P-value

The Student's t-test [137] is a statistical hypothesis. Under the null hypothesis with equal sample size, it tests whether the test statistic follows t-distribution, thus describe how significant between two different groups X_1 and X_2. The t statistic can be calculated with

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{2.75}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2}} \tag{2.76}$$

is the estimator for two sample groups' pooled standard deviation, $s_{X_1}^2$ and $s_{X_2}^2$ are the unbiased estimators for two sample groups' variances. $n_1 - 1$ and $n_2 - 1$ are the number of degrees of freedom for each group. $n_1 + n_2 - 2$ is the total number of degrees of freedom.

After calculating the t statistic and determining the degrees of freedom, the corresponding P-value can be found in the Student's t-distribution value table. In this thesis, two-sided P-value is adopted. To avoid confusion with other parts of this thesis, variables X, s, s_p, n, and t are defined only in this section.

2.8.6 Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient r_s is a nonparametric measure of rank correlation [138]. Similar to the Pearson correlation, the Spearman correlation between two variables, but using the rank instead of values. If no tied ranks are observed, a perfect Spearman correlation of $+1$ or -1 occurs if and only if one variable is a perfect monotone function of the other.

A high Spearman correlation between two variables will be observed when they tend to have a similar or identical rank (then it will close to $+1$). If two variables' ranks are opposite, then a low Spearman correlation between two variables will be observed. The Spearman correlation can be close to 0 if two variables are tend to be uncorrelated in terms of the rank.

Definition 2.8.2. ***Spearman Correlation Coefficient***

Given N samples, the values of $X_{\mathrm{i}}, Y_{\mathrm{i}}$ are converted to ranks $x_{\mathrm{i}}, y_{\mathrm{i}}$, and the Spearman's rank correlation coefficient r_s can be computed from

$$r_s = \rho_{x,y} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum_{\mathrm{i}}(x_{\mathrm{i}} - \bar{x})(y_{\mathrm{i}} - \bar{y})}{\sqrt{\sum_{\mathrm{i}}(x_{\mathrm{i}} - \bar{x})^2 \sum_{\mathrm{i}}(y_{\mathrm{i}} - \bar{y})^2}}, \tag{2.77}$$

where ρ stands for the Pearson correlation coefficient with ranks. $\mathrm{cov}(x, y)$ *is the covariance of the rank variables x and y. σ_x and σ_y are the standard deviations of the rank variables x and y.*

If among x and y there are all distinct integer values (no tied ranks), the Spearman's rank correlation coefficient can then be computed using the popular formula:

$$r_s = 1 - \frac{6 \sum_{\mathrm{i}} d_{\mathrm{i}}^2}{N(N^2 - 1)}, \tag{2.78}$$

where $d_{\mathrm{i}} = x_{\mathrm{i}} - y_{\mathrm{i}}$ is the difference between the two ranks of each observation.

To avoid confusion with other parts of this thesis, symbols and notations X, Y, x, y, and σ are defined locally only in this section.

2.8.7 Hypergeometric Test and P-value

The hypergeometric test is based on hypergeometric distribution [139]. Suppose a random variable X is distributed hypergeometrically, then $X \sim \mathrm{Hypergeometric}(N, K, n)$, the probability mass function is given by

$$\Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \tag{2.79}$$

where N is the population size (total number of genes in the background), K is the number of hits in the targeted gene ontology list population, n is the number of draws in the gene list of interests, and k is the number of observed successes. $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient.

Hypergeometric test is widely used for evaluating the gene set enrichment analysis. In gene set enrichment analysis, the variables are defined as in Table 2.2.

Table 2.2. Variable definitions in hypergeometric test.

Variable	Definition
K	The total number of genes in the targeted gene ontology list.
k	The total number of mutual genes in both gene list of interests and the targeted gene ontology list.
N	The total number of genes in the background.
n	The total number of genes in gene list of interests.

To calculate hypergeometric test P-value, we need to test the null hypothesis $\Pr(X \geq k)$ [140]. Given K, N, and n, the P-value measuring the significance of gene set enrichment is the tail probability when observing $\mathrm{i} \geq k$ genes annotated to the gene ontology term:

$$P\text{-value} = \sum_{\mathrm{i}=k}^{\min(K,n)} \frac{\binom{K}{\mathrm{i}}\binom{N-K}{n-\mathrm{i}}}{\binom{N}{n}}. \tag{2.80}$$

The hypergeometric test P-value used in gene set enrichment analysis is one of the most popular approaches to assess the enrichment results. To avoid confusion with other parts of this thesis, variables X, K, N, k, and n are defined only in this section.

2.8.8 False Discovery Rate

The false discovery rate (FDR) is a method to summarize the type I error rates in null hypothesis testing when performing multiple comparisons [141]. To describe FDR, We let the Q be the proportion of false discoveries (rejections of the null hypothesis) among all discoveries:

$$Q = \frac{V}{R} = \frac{V}{V+S}, \tag{2.81}$$

where V is when the test is declared significant under true null hypothesis, S is when the test is declared significant under true positives, and $R = V + S$ is the total number of rejected null hypotheses. The FDR is then given by

$$FDR = \mathrm{E}[Q], \tag{2.82}$$

which is the expectation of Q. Q is defined to be 0 when $R = 0$ [141]. To avoid confusion with other parts of this thesis, symbols and notations Q, V, R, S, and $E[Q]$ are defined only in this section.

2.8.9 q-value False Discovery Rate Benjamini-Hochberg Procedure

Instead of hypergeometric test P-value, q-value False discovery rate Benjamini-Hochberg procedure (or q-value FDR B&H in short) is the corrected P-value using the false discovery rate (FDR) method [141], [142]. It is a method to decrease the FDR and avoid false positives. In enrichment analysis, given M number of pathways or gene ontology terms, we would like to see whether the targeted gene ontology term is especially enriched.

To calculate each individual q-value FDR B&H from P-value, the formula

$$q\text{-value} = \frac{\mathrm{i}}{M}Q \tag{2.83}$$

will be used after assigning the ranks of P-values in ascending order, where i is the individual P-value's rank, M is the total number of tests, Q is the FDR. In this thesis, we choose FDR $= 0.05$ if not stated otherwise. To avoid confusion with other parts of this thesis, variables i, Q, and M are defined only in this section.

2.8.10 Mann-Whitney U Test and P-value

Mann-Whitney U Test [143] is a non-parametric test that can tell whether one group of observations is larger than the other group. Considering group A has N_A observations x_1, x_2, $\cdots$, x_{N_A}, group B has N_B observations y_1, y_2, $\cdots$, y_{N_B}. The total number of pairwised comparison is $N_A N_B$. We let the null hypothesis be

$$H_0 : \Pr(x_{\mathrm{i}} > y_{\mathrm{j}}) = \frac{1}{2}, \tag{2.84}$$

and the alternative hypothesis be

$$H_1 : \Pr(x_{\mathrm{i}} > y_{\mathrm{j}}) \neq \frac{1}{2}, \tag{2.85}$$

by assuming group A and group B have the same median values. Next, we let the number of occurrences that x_{i} is greater than y_{j} be U_x, and the number of occurrences that x_{i} is smaller than y_{j} be U_y. We would expect $U_x = U_y$ under the null hypothesis. Then we calculate $U = \min(U_x, U_y)$. The P-value is then determined by inspecting the Mann-Whitney U test statistical table. In this thesis, two-sided Mann-Whitney U test P-values are used. To avoid confusion with other parts of this thesis, variables A, B, x, y, and U are defined only in this section.

As these popular research domains and related evaluation metrics are elaborated within different sections respectively, combining Cox proportional hazards regression with gene co-expression analysis or NMF will be used extensively to solve the problem of our interests in Chapter 3 and Chapter 4.

3. SALMON: SURVIVAL ANALYSIS LEARNING WITH MULTI-OMICS NEURAL NETWORKS

With the rapid development on mRNA sequencing (mRNA-seq) and the next generation sequencing (NGS) technology, analyzing the transcriptomes for biological, cancer, and clinical research becomes feasible [144]–[146], especially for the precision health medicine.

The studies of using multivariate Cox proportional hazards regression to analyze and uncover important genes in survival analysis have been conducting over decades such as in breast cancer [147] and head & neck cancer [148]. These areas of research normally consider mRNA-seq expression values $\boldsymbol{X}$ as the data input. Rather than linear Cox model, other variants such as random survival forest (RSF) [28] used ensemble tree-based approaches to inference the survival. Though many models can predict survival from data, there is a strong need for sophisticated algorithms that can aggregate and filter relevant predictors from increasingly complex data inputs [2]. In turn, with a highly flexible model and account for data complexity in a non-linear fashion, deep learning-based neural networks offer a potential solution [2], [149], [150]. The advantages of learning non-linear functions and retrieving lower dimensional representations [24] reveal advances of deep learning models. It has also been proved by experiments that integrating Cox proportional hazards regression into last hidden layer of neural networks can achieve better performance in terms of concordance index [24].

Since the applications of survival prognosis that incorporates Cox proportional hazards regression with a single transcriptomics dataset [24], [25], [151] and with multi-omics data [30]–[32], [152], [153] is of major interest in precision health, in this chapter, we implement deep learning-based networks to determine how gene expression data predicts survival in breast cancer. We accomplish this through an algorithm called SALMON (Survival Analysis Learning with Multi-Omics Neural Networks), which aggregates and simplifies gene expression data and cancer biomarkers to enable prognosis prediction. The results reveal improved performance when more omics data are used in model construction. Our study shows the feasibility of discovering breast cancer related co-expression modules, sketches a blueprint of future endeavors on deep learning-based survival analysis. SALMON source code is available at https://github.com/huangzhii/SALMON/.

3.1 Dataset

There is a strong need to identify effective prognostic biomarkers to help optimize and personalize treatment [154]. Among cancers, breast invasive carcinoma is one of the most heterogeneous cancers with distinct prognoses based on morphological, phenological, and molecular stratifications [155], [156]. One study showed that breast invasive carcinoma patients have a 77% survival rate after 5 years and 44% survival rate after 15 years [157], so developing accurate prognostic models could significantly improve risk stratification after diagnosis. While most contemporary approaches incorporate one or few types of omics data, such as mRNA-seq data and miRNA-seq data [158], [159], we propose that integrating multimodal biomedical data may lead to improved modeling, especially when driven by machine learning. Moreover, classic cancer biomarkers can often stratify patients into risk groups, and these too should be integrated when available. Specifically, copy number burden (CNB) and tumor mutation burden (TMB) are important for predicting tumor progression [160], [161] and immunotherapy [162]–[164]. Other demographical and clinical information such as diagnosis age, estrogen receptors (ER) status, progesterone receptors (PR) status should also be considered during model construction. In this study, we integrate multi-omics data as input to the deep learning-based survival prognosis model. The data cohort we collect is reported in Table 3.1 from The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/) and cBioPortal [165], [166].

3.2 Study Design

In this study, 5-fold cross-validation is performed on the breast cancer dataset. In each fold, 80% of the data are used for model training and 20% of the data are used for model testing. mRNA and miRNA data are pre-processed by TSUNAMI online analysis suite (https://shiny.ph.iu.edu/TSUNAMI/) [11]. The pre-processing is 2-folded: It firstly removes genes with lowest 20% of mean expression values shared by all patients. Then it removes genes with lowest 20% of expression values' variance. These pre-processing steps are necessary to ensure the robustness for the downstream correlational computation in gene co-expression module analysis step.

Table 3.1. Demographical and clinical characteristics of 583 female breast invasive carcinoma (BRCA) patients. The status of estrogen receptor (ER) and progesterone receptor (PR) are derived from IHC (immunohistochemistry). Clinical information is collected from cBioPortal.

	Original	Co-expression module
mRNA-seq data size	13,132	57
miRNA-seq data size	530	12
	Median	**Range**
Overall survival months	31.70	0.00 - 216.59
Diagnostic age	57	26-90
ER positive ratio		76.16%
PR positive ratio		67.41%

3.2.1 Gene Co-expression Analysis as Upfront Feature Engineering

One of the challenges for such diverse multimodal data is high-dimensionality. Rather than using raw gene expression values as model inputs, we innovatively use eigengene modules from the result of gene co-expression network analysis. The corresponding high impact co-expression modules and other omics data are identified by feature selection technique, then examined by conducting enrichment analysis and exploiting biological functions, which escalates the interpretation of input feature from gene level to co-expression modules level.

Instead of feeding mRNA-seq and miRNA-seq data to the neural networks and analyzing results at the gene level, we use eigengene matrices of gene co-expression modules obtained from lmQCM algorithm [34] as the input to the SALMON algorithm. This reduces 99.46% of input features and greatly reduces the number of parameters in the neural networks. Using eigengenes as features can be considered as bias/variance (error/complexity) trade-off in machine learning [57], [167], which simplifies the networks significantly. The total number of neural network weights to be learned in the later study are then narrowed down from 107,193 to 521, ensuring the robustness of the learning procedure and alleviating the overfitting issue [53], [168].

There are many gene co-expression network analysis packages, such as the R package for weighted correlation network analysis (WGCNA) [35] and local maximal Quasi-Clique

Merger (lmQCM) [34]. These available methods enable the discovery of densely connected gene modules across samples/patients. Co-expression network analyses are used increasingly to reveal latent gene-gene interactions, biomarkers and novel gene functions [7], [46], [51], [169]–[171]. Comparing to WGCNA, weight normalization process in lmQCM is inspired by the spectral clustering [172] in machine learning. With efficient implementation of the revision from eQCM (edge-covering quasi-clique merger) algorithm [173], lmQCM allows module overlap, mining smaller densely co-expressed modules, and thus it is adopted in this chapter. The generally smaller size of mined modules can also generate more meaningful gene ontology (GO) enrichment results [40], [44], [174]–[176]. The implementation is performed on TSUNAMI (introduced in Chapter 6) and positive correlations are analyzed. For mRNA-seq data, we set lmQCM parameters $q_\gamma = 0.7$, $q_\lambda = 1$, $q_t = 1$, $q_\beta = 0.4$, minimum size of cluster = 10, and adopt Spearman's rank correlation coefficient [177] to calculate gene-wised correlations. The parameters setting of miRNA-seq data are the same except $q_\gamma = 0.4$, $q_\beta = 0.6$, and minimum size of cluster = 4.

After calculating gene co-expression modules with lmQCM, eigengene matrices are then determined. The eigengene matrix is the expression values of each gene co-expression module summarized into the first principal component using singular value decomposition (SVD) [68]. With the first right-singular vector of each module as the summarized expression values, it projects co-expressed genes to 1-D space and thus can be treated as the "super gene". In our experiment with breast invasive carcinoma, an eigengene matrix with 57 dimensions is derived from mRNA-seq data and an eigengene matrix with 12 dimensions is derived from miRNA-seq data. These eigengene matrices are treated as the substitution of the original expression inputs.

3.2.2 Neural Networks Design, Architecture, and Evaluation Metrics

SALMON is designed and implemented in Pytorch 1.0 [131]. mRNA-seq and miRNA-seq eigengene matrices are firstly connected to hidden layers with dimensions 8 and 4, respectively. They then connect to the final output (hazard ratio) with Cox proportional hazards regression networks. Alternatively, CNB, TMB, and demographical and clinical information

(diagnosis age, ER status, PR status) have no hidden layer and are connected to final output directly as covariates. This architecture is graphically explained in Figure 3.1. The rationale behind this network architecture instead of using simple fully connected networks such as Cox-nnet [24] is by assuming (1) each omic type may affect the hazard ratio independently; (2) downscale eigengene matrices by hidden layers can force multi-omics data contributed to hazard ratios in a relatively equal scale at Cox proportional hazards regression networks part. The performance comparison of our proposed model with a modified fully connected SALMON model is also conducted in the later section.

SALMON adopts Adaptive moment estimation (Adam) optimizer [178]. We set the number of epochs = 100 with fine-tuned learning rates for each 5-folds cross-validation experiments. LASSO (least absolute shrinkage and selection operator) regularization [62] is applied to the networks. Sigmoid activation function is also applied right after each forward propagation and Cox proportional hazards regression networks. The Sigmoid function

$$sigmoid(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{3.1}$$

forces the output range be within 0 to 1, introduces non-linearity to the system. In this model, we set the batch size = 256, and the batch normalization was not adopted. The number of the hidden layers and dimensions of hidden layers can be fine-tuned, in this study, two single hidden layers are attached to the transcriptomics data with size = 8 for mRNA-seq modules, and size = 4 for miRNA-seq modules, respectively.

Our algorithm SALMON, integrates Cox proportional hazards model, differs from previous work [153], [179] which used survival status (living or deceased) in a binary classification problem. In contrast, we take survival times (overall survival months) into account denote as Y_{i} and make our neural networks into a Cox regression learning task. Maximum likelihood estimation (MLE) is then applied to the log partial likelihood

$$\ell(\beta) = \sum_{\mathrm{i}:C_{\mathrm{i}}=1} \left(\sum_{k=1}^{K} \beta_k \boldsymbol{H}_{\mathrm{i}k} - \log \left(\sum_{\mathrm{j}:Y_{\mathrm{j}} \geq Y_{\mathrm{i}}} \exp \left(\sum_{k=1}^{K} \beta_k \boldsymbol{H}_{\mathrm{i}k} \right) \right) \right), \tag{3.2}$$

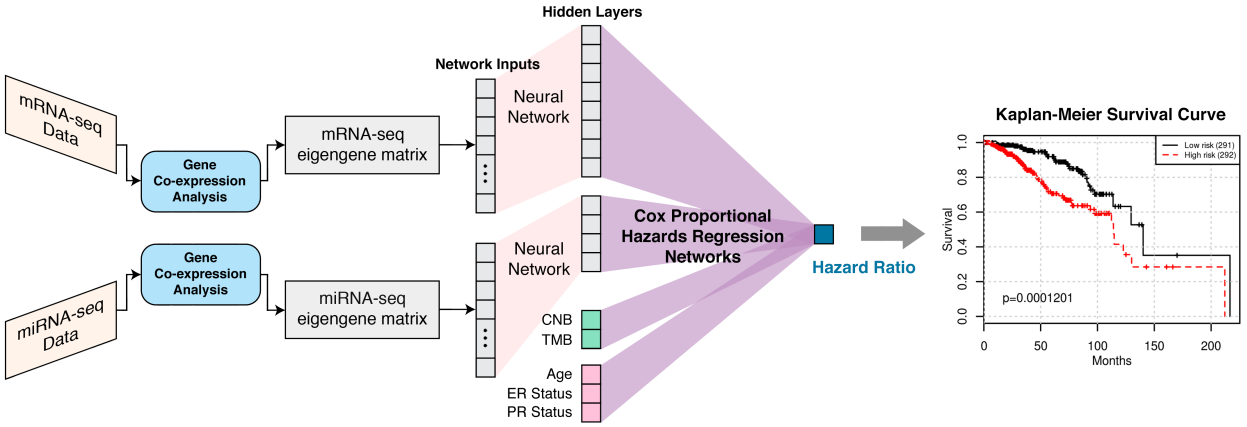


Figure 3.1. SALMON (Survival Analysis Learning with Multi-Omics Neural Networks) architecture with the implementation of Cox proportional hazards regression networks. Co-expression modules (eigengene matrices) are the inputs to the SALMON. The output is the hazard ratios which can be interpreted as the relative risks of patients.

where β are the parameters to be estimated. $C_{\mathrm{i}} = 1$ indicates the occurrence of the death events for patient i with K-dimensional vector $\boldsymbol{H}_{\mathrm{i}}$ represents the last hidden layer in the low-rank space.

Based on Cox proportional hazards regression networks we formulate the objective function of neural networks as

$$\underset{\Theta,\boldsymbol{H}}{\operatorname{argmin}} \left\{ - \sum_{\mathrm{i}:C_{\mathrm{i}}=1} \left(\sum_{k=1}^{K} \beta_k \boldsymbol{H}_{\mathrm{i}k} - \log \left(\sum_{\mathrm{j}:Y_{\mathrm{j}} \geq Y_{\mathrm{i}}} \exp \left(\sum_{k=1}^{K} \beta_k \boldsymbol{H}_{\mathrm{i}k} \right) \right) \right) + \xi \|\Theta\|_1 \right\}, \tag{3.3}$$

where Θ is the entire network weights (including β in Cox proportional hazard model) to be optimized via back-propagation, ξ is the weight multiplier of LASSO regularization. We set $\xi = 1 \times 10^{-5}$ in the experiments.

Concordance index (C-Index) [180], values from 0 to 1 mentioned earlier, is used in this thesis as the evaluation metric of survival prognosis. It is widely adopted to evaluate the performances of survival prognosis models [24], [25] and is equivalent to the area under the ROC curve (AUC) [181], which measures the model's distinguishability between living and deceased groups. A C-Index = 0.5 indicates the model makes ineffective prediction. A higher

C-Index > 0.5 indicates a better survival prognosis model. For breast invasive carcinoma cancer, we consider a C-Index > 0.7 indicates a good model performance.

Log-rank test [134] is used to inspect the performances of SALMON on 5-fold cross-validation testing set results. The Kaplan-Meier survival curves are generated by dichotomizing all testing patients to low risk and high risk groups via the median hazard ratio. The corresponding log-rank P-values imply the ability of the model to differentiate two risk groups. Lower P-values convey better model performances.

3.2.3 Experimental Settings

The experiments are performed with six different combinations of multi-omics data as input sources, they are: (i) mRNA-seq data (mRNA) (57 features); (ii) miRNA-seq data (miRNA) (12 features); (iii) integration of mRNA and miRNA (69 features); (iv) integration of mRNA, miRNA, copy number burden (CNB), and tumor mutation burden (TMB) (71 features); (v) integration of mRNA, miRNA, and demographical and clinical (diagnosis age, ER status, PR status) data (72 features); (vi) integration of mRNA, miRNA, CNB, TMB, and demographical and clinical (diagnosis age, ER status, PR status) data (74 features). Where both RNA-seq co-expression modules are required for all integrative combinations. The SALMON model architecture from Figure 3.1 removes certain network substructures which are not been used and performs 5-folds cross-validation with 583 patients. C-Index is used to evaluate the performances. SALMON then compares with several other survival prognosis algorithm Cox-nnet [24], DeepSurv [25], generalized linear model with Cox regression (GLMNET) [29], and random survival forest (RSF) [28] with all omics data fed in. Since their Cox regression model did not take multi-omics data sources into consideration, we modify their original framework to integrate multi-omics data (with co-expression modules) altogether as single input vector. The feature importance of all 74 covariates are also investigated by repeated feature deletion, then being ranked by the median of decreased C-Index, suggest and reveal certain biological interpretations. The performances of SALMON variations with different model architectures are also investigated.

3.2.4 Downstream Gene Ontology and Functional Enrichment Analysis

Co-expression modules generated by lmQCM are then exported to ToppGene Suite (https://toppgene.cchmc.org/) [182] and Enrichr (https://maayanlab.cloud/Enrichr/) [183]. Using ToppGene, we perform functional analysis including gene ontology (GO) and cytoband analysis. The false discovery rate (FDR) < 0.05 and FDR < 1.0 are considered to be significantly enriched for GO analysis and cytoband analysis, respectively. Human Gene Atlas [up regulated genes in human tissues from BioGPS (http://biogps.org/)] and ARCHS4 tissues are also investigated for some certain co-expression modules by Enrichr.

3.3 Results

3.3.1 Integrating Multi-Omics Features Increased the Performances

From Figure 3.2A, we observe an upward trend on median/mean concordance indices with more omics data are integrated. Moreover, integrating all omics data (74 features) give the optimal performances (C-Index: median = 0.7285; mean = 0.6918). Next, all hazard ratios from 5-fold testing sets are concatenated and performed the log-rank test as shown in Figures 3.2C–E and Figure 3.3. Another feature set without transcriptomics data is also considered as reference (5 features containing CNB, TMB, and demographical and clinical features) with median C-Index = 0.6949 and the Kaplan-Meier plot is shown in Figure 3.3F (log-rank test P-value $= 3.67 \times 10^{-3}$). We find that integrating all omics data (Figure 3.2E) gives the most significant P-value (1.201×10^{-4}) with respect to the log-rank test, suggesting that integrating more multi-omics data to SALMON can enhance the prediction.

We further perform pairwise paired t-test to the resulting concordance indices. As shown in Table 3.2, a negative t-statistic implies that the set 1 is lower than set 2. This concludes that integrating more omics data can generally increase the performance of survival prognosis in breast cancer dataset.

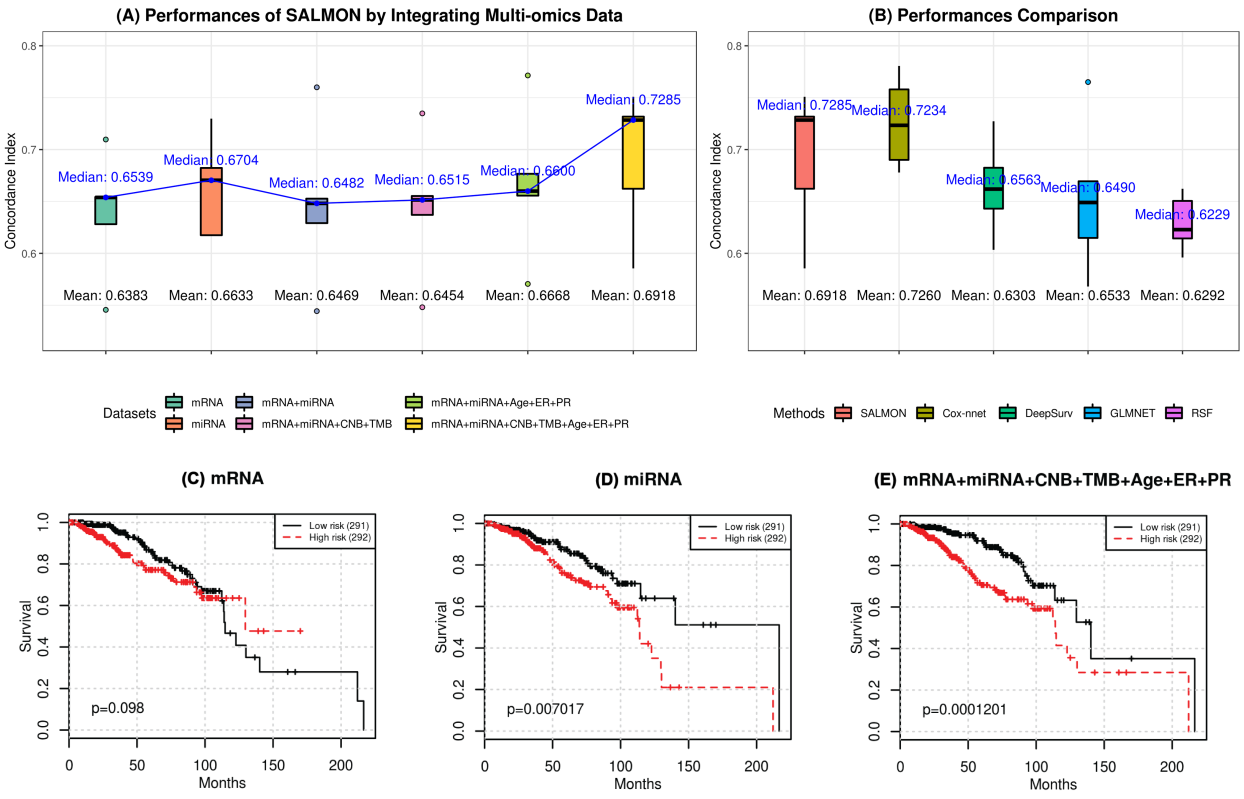


Figure 3.2. (A) Performances of SALMON with multi-omics data integrated in terms of C-Index. (B) Performance comparison between SALMON and the modified Cox-nnet, DeepSurv, GLMNET, and RSF in terms of C-Index with all omics data used for learning. (C–E) Kaplan-Meier plot of survival prognosis. Hazard ratios are derived from all five testing sets. Log-rank test is used to find the corresponding P-value with low risk and high risk groups dichotomized by the median hazard ratio. Omics data used for training and testing: (C) mRNA-seq data (mRNA); (D) miRNA-seq data (miRNA); (E) integration of mRNA, miRNA, CNB, TMB, and demographical & clinical (diagnosis age, ER status, PR status) data. All other combinations of multi-omics results are shown in Figure 3.3.

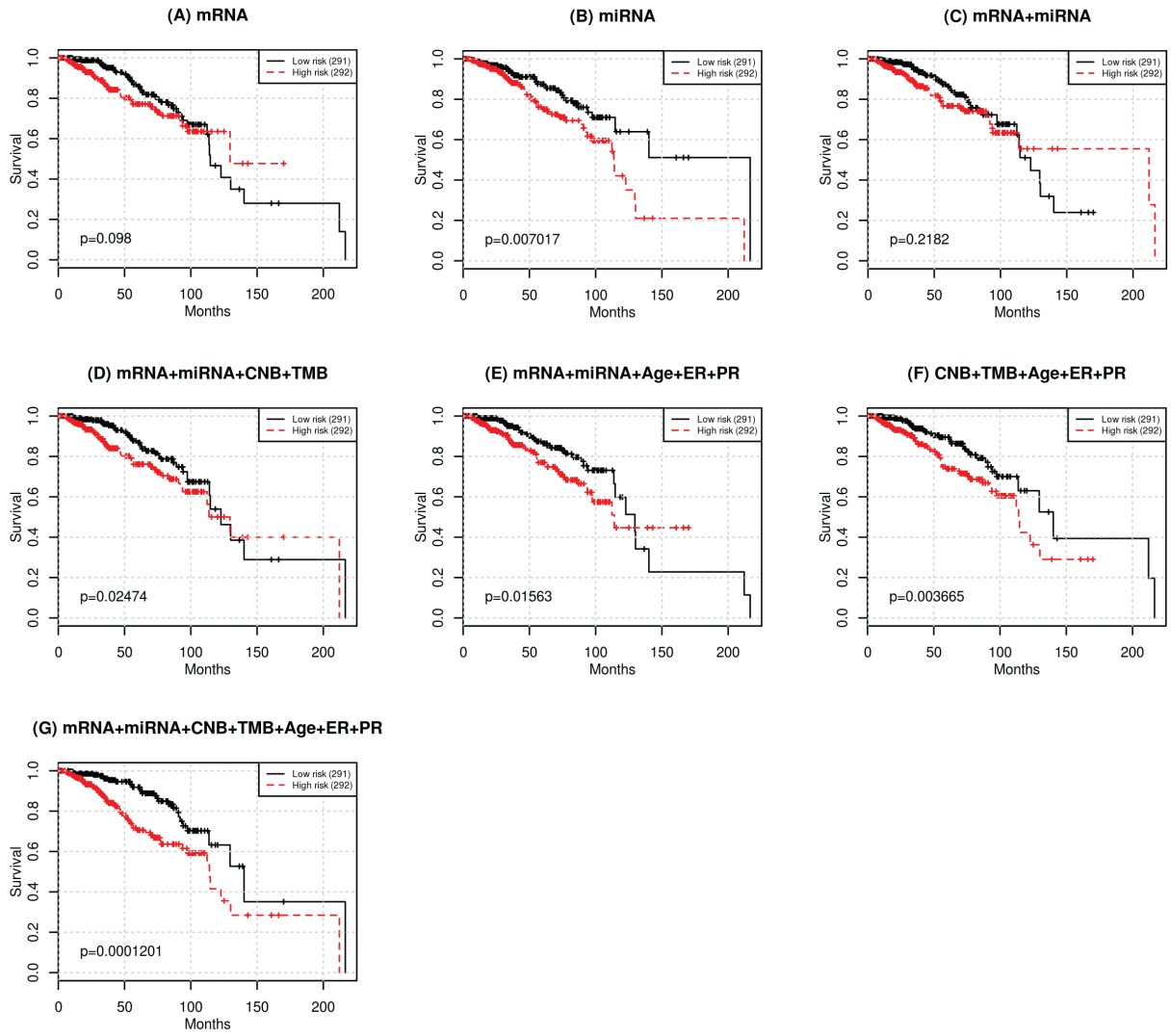


Figure 3.3. Kaplan-Meier plot of survival prognosis. Hazard ratios are derived from all five testing sets. Log-rank test is used to find the corresponding *P*-value with low risk and high risk groups dichotomized by the median hazard ratio. Omics data used for training and testing: (A) mRNA-seq data (mRNA) (57 features); (B) miRNA-seq data (miRNA) (12 features); (C) integration of mRNA and miRNA (69 features); (D) integration of mRNA, miRNA, copy number burden (CNB), and tumor mutation burden (TMB) (71 features); (E) integration of mRNA, miRNA, and demographical & clinical (diagnosis age, ER status, PR status) data (72 features); (F) integration of copy number burden (CNB), tumor mutation burden (TMB), demographical & clinical (diagnosis age, ER status, PR status) data (5 features); (G) integration of mRNA, miRNA, CNB, TMB, and demographical & clinical (diagnosis age, ER status, PR status) data (74 features).

Table 3.2. Performances comparison with different combinations of multi-omics data by pairwise paired t-test, according to C-Index among 5-folds cross-validation results. Note that a negative t-statistic indicates set 1 worse than set 2 in terms of performances. Multi-omics dataset applied as inputs: (i) mRNA-seq data (mRNA) (57 features); (ii) miRNA-seq data (miRNA) (12 features); (iii) integration of mRNA and miRNA (69 features); (iv) integration of mRNA, miRNA, copy number burden (CNB), and tumor mutation burden (TMB) (71 features); (v) integration of mRNA, miRNA, and demographical & clinical (diagnosis age, ER status, PR status) data (72 features); (vi) integration of mRNA, miRNA, CNB, TMB, and demographical & clinical (diagnosis age, ER status, PR status) data (74 features).

		Set 2									
		ii		iii		iv		v		vi	
		t	P	t	P	t	P	t	P	t	P
Set 1	i	−0.784	0.477	−0.676	0.536	−0.832	0.452	−2.928	0.043*	−3.315	0.030*
	ii	—	—	0.406	0.705	−0.487	0.652	−0.092	0.931	−0.652	0.550
	iii	—	—	—	—	0.247	0.817	−5.804	0.004*	−2.710	0.054
	iv	—	—	—	—	—	—	−4.168	0.014*	−3.603	0.023*
	v	—	—	—	—	—	—	—	—	−1.529	0.201

Notes: t denotes the pairwise paired Student's t-test statistic, P denotes the P-value obtained. P-value < 0.05 are considered to be significant and indicated with * symbol.

Next, we compare SALMON to the state-of-the-art deep learning-based cancer survival prognosis model Cox-nnet [24], as well as DeepSurv [25], and two traditional models generalized linear model with Cox regression (GLMNET) [29] and RSF [28]. We further modify their original implementation with all omics data as inputs. As shown in Figure 3.2B, the median C-Index of SALMON (0.7285) is reported higher than the modified Cox-nnet (0.7234), DeepSurv (0.6563), GLMNET (0.6490), and RSF (0.6229). Comparing to the modified Cox-nnet with similar performance in terms of C-Index, SALMON has a more significant result in log-rank test (P-value $= 1.201 \times 10^{-4}$) than the modified Cox-nnet (P-value $= 2.282 \times 10^{-4}$) with all testing sets and all 74 features as inputs (Figure 3.4). Between SALMON and the modified Cox-nnet the performance is insignificant (paired t-test statistic $= -2.105$, P-value $= 0.103$) suggesting these two methods are comparable. But from the neural network structure perspective, SALMON is more flexible since it separates forward propagation for each omics data, enables a scalable integration of multimodal biomedical data.

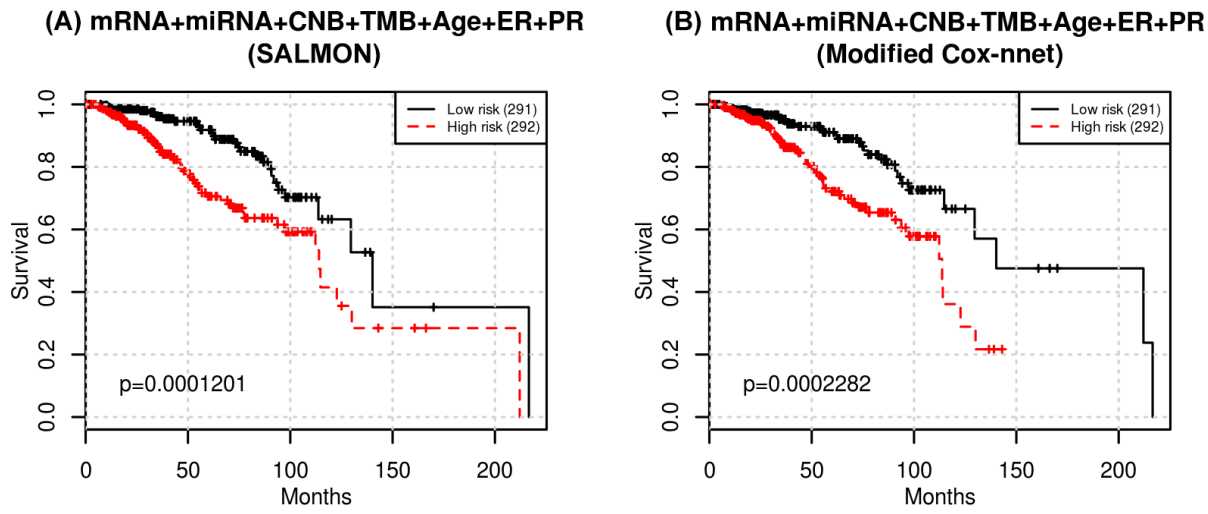


Figure 3.4. Performances comparison in terms of the P-value of the log-rank test between SALMON (A) and modified Cox-nnet (B) with all omics data as inputs.

3.3.2 Evaluation of SALMON Architecture

We further perform evaluations with some variations of the SALMON architecture, by introducing: (1) SALMON_Full_Gene: using original mRNA-seq and miRNA-seq data as input (Figure 3.5A). Using this architecture, the median concordance index in testing set is 0.7048. (2) SALMON_FC: using fully connected layers instead of the separate networks (Figure 3.5B). Using this architecture, the median concordance index in testing set is 0.6695. (3) SALMON_2_Layers: one additional hidden layer with number of nodes = 6 before Cox regression is introduced (Figure 3.5C). Using this architecture, the median concordance index in testing set is 0.6861. (4) SALMON_3_Layers: two additional hidden layers with number of nodes = 6 each before Cox regression are introduced (Figure 3.5D). Using this architecture, the median concordance index in testing set is 0.7029. Based on these modified architectures, we found the original SALMON model returns highest median concordance index (0.7285), suggesting the original SALMON model is the preferred architecture in breast cancer survival prognosis. One interesting finding is that SALMON_Full_Gene has an inferior median concordance index than our original architecture. It suggests that using high-dimensional

features as input (instead of using eigengene matrices as input) can result in many neural network parameters, which may not bring any additional benefit in survival prediction.

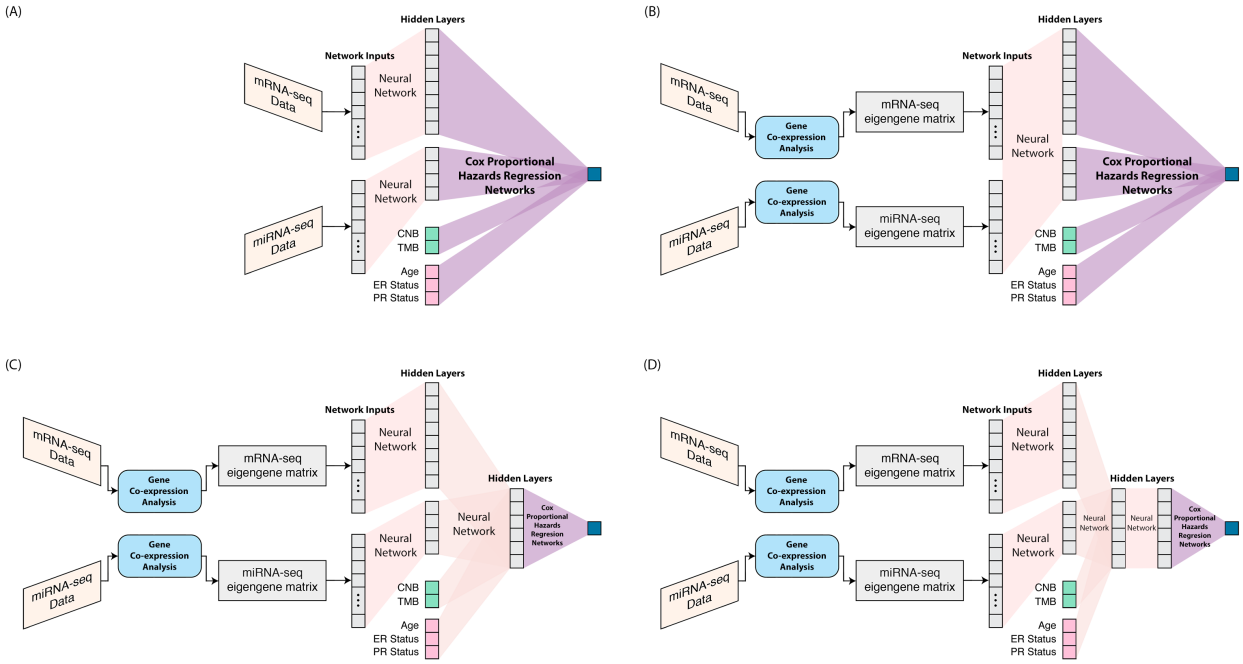


Figure 3.5. SALMON architecture variations. (A) SALMON_Full_Gene: SALMON using original mRNA-seq and miRNA-seq data as input. (B) SALMON_FC: SALMON architecture, but the mRNA-seq eigengene and miRNA-seq eigengene are fully connected. (C) SALMON_2_Layers: SALMON architecture, but has a second hidden layer (number of nodes = 6). (D) SALMON_3_Layers: SALMON architecture, but has second and third hidden layers (both number of nodes = 6).

3.3.3 Interpreting and Ranking the Importance of Co-expression Modules

Interpreting feature importance for neural networks has been studied over years. One way is to assign each feature be zero repeatedly, then the feature with lowest change of the resulting accuracy implies the least importance that affects to the prediction model. This approach is widely adopted for feature selection and ranking the importance of features in neural network [184]–[186]. Based on this approach, we analyze the contribution of each eigengene's module to the final hazard ratio by forcing each input feature of the testing sets be zero. By feeding the modified testing sets to the pre-trained SALMON networks, we rank the importance of features by inspecting how much the concordance indices decreased. Features

that decrease the testing concordance indices more are considered to be more important. At this moment, we integrate all omics data for training and testing. Table 3.3 presents top features that mostly reduced the C-Index. The leading two features are the diagnosis age and PR status, then five mRNA-seq co-expression modules are followed.

Table 3.3. Top features that reduce the C-Index, including two demographical and clinical features, and five mRNA-seq co-expression modules (eigengene matrices as inputs to the SALMON). C-Index changed: The median value of changed C-Index.

Ranks	Feature names	C-Index changed	Highlighted genes interpretations, enrichments, or notes
1	Diagnosis age	−0.1257	Age.
2	PR status	−0.0343	Progesterone receptors status.
3	Module 13	−0.0150	Genes MST1, CPT1B. CD8+, CD4+, Breast bulk tissue.
4	Module 47	−0.0071	Genes MAP3K7, CCNC. Cytoband chr6q14-q16 and chr6q21.
5	Module 5	−0.0059	Genes DDR2, FLNA, TCF4. Associated with extracellular matrix (ECM), cell adhesion, and cell migration.
6	Module 36	−0.0053	Gene SNW1. Cytoband chr14q23-q24 and chr14q31-q32.
7	Module 51	−0.0047	Genes TCP1, HDAC2. Cytoband chr6q14-q15and chr6q21-q26.

Next, we select those features (33 in total) of which their median values < 0 in Figure 3.6 and re-perform the machine learning with SALMON. Results show that, before and after feature selection, the performances are insignificant in terms of C-Index (before feature selection: mean = 0.6918, median = 0.7285; after feature selection: mean = 0.7108, median = 0.7200; paired t-test statistic = −0.861, P-value = 0.438) (Figure 3.7). This result suggests that training with selected "important" multi-omics features instead of all can still preserve the prognosis performances.

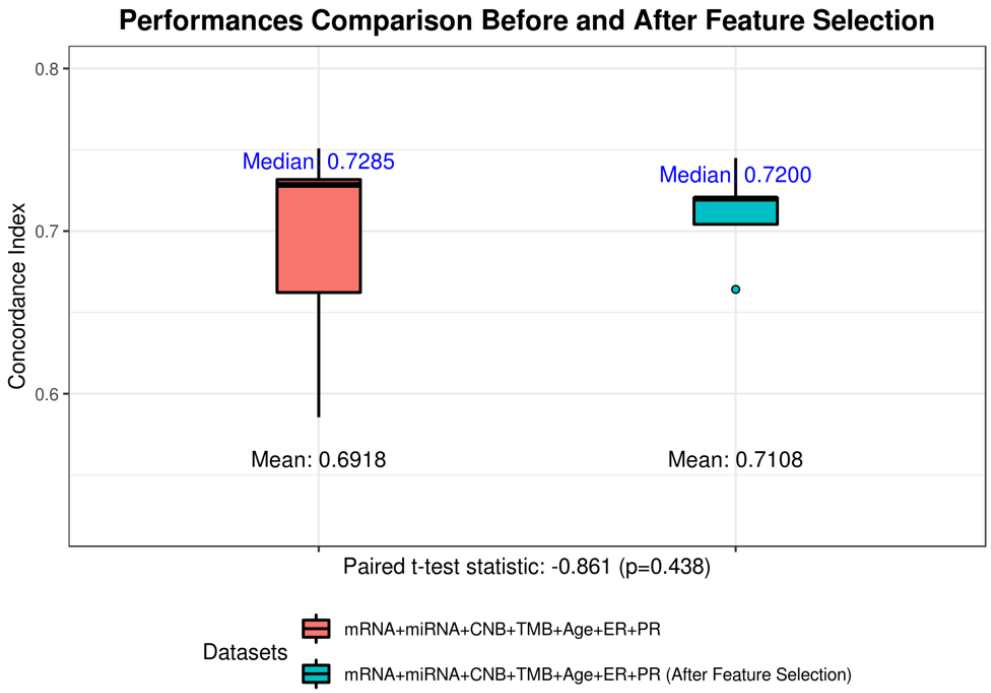


Figure 3.7. SALMON's performance comparison using all 74 multi-omics features and using selected 33 features (which their medians result in decrements to the C-Index in Figure 3.6). Selected 33 features are with ID 72, 74, 13, 47, 5, 36, 51, 19, 33, 29, 53, 20, 58, 66, 15, 16, 34, 70, 31, 42, 60, 11, 18, 71, 2, 10, 43, 44, 9, 32, 56, 62, 68 in Figure 3.6, where 24 of them are from mRNA co-expression modules, 5 of them are from miRNA co-expression modules, other 4 features are copy number burden (CNB), tumor mutation burden (TMB), diagnosis age, and progesterone receptors (PR) status, respectively. C-Index before feature selection: median = 0.7285, mean = 0.6918; after feature selection: median = 0.7200, mean = 0.7108. Paired t-test statistics = −0.861 (P-value = 0.438).

3.3.4 Investigating Feature Importance with Different Age Groups

As shown in Figure 3.6, we observe a strong predictive power of diagnosis age, which is consistent with previous studies demonstrating age is one of the most prominent cancer risk factors [187]. Thus, it is crucial to further investigate whether patients in different groups can be stratified using the same set of features. In this paper, we define three age groups: (1) age in range 26–50 (191 patients), (2) age in range 51–70 (280 patients), and (3) age in range 71–90 (112 patients) to represent younger, middle aged, and elderly patients. By

training and testing these three distinct groups with SALMON algorithm, we aim to answer two questions: (1) whether the diagnosis age still be a strong factor that affect prognosis performance after the stratification, and (2) how feature rankings differ among these three distinct groups.

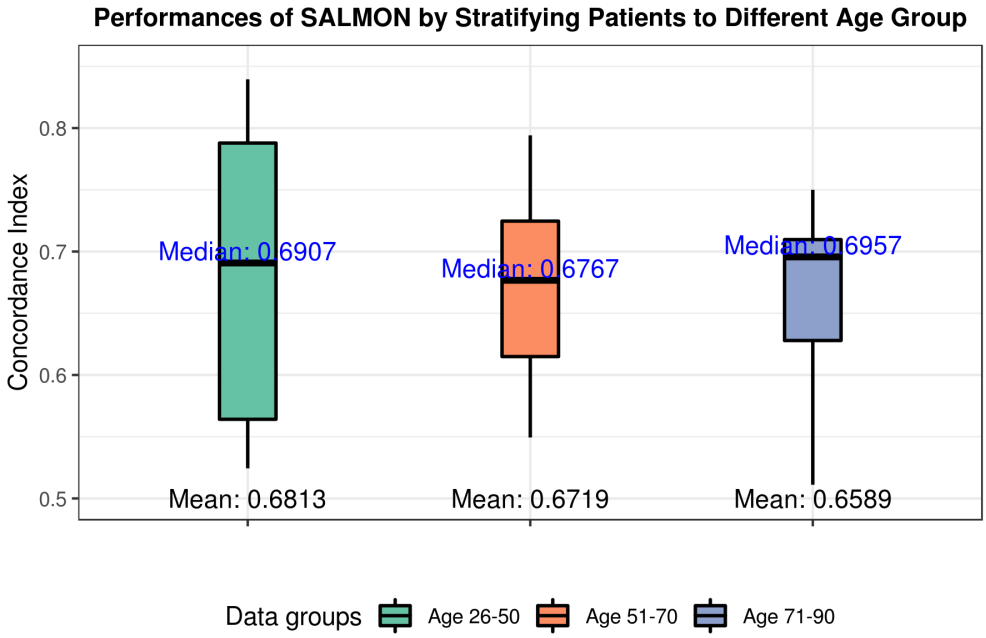


Figure 3.8. Performances of SALMON algorithm stratified by three age groups: 26–50 group; 51–70 group; 71–90 group by integrating all omics data (including mRNA, miRNA, CNB, TMB, diagnosis age, ER status, PR status).

Table 3.4. Top features that reduce the concordance indices. Experiments are performed separately with three age groups: 26–50 group; 51–70 group; 71–90 group, by integrating all omics data (including mRNA, miRNA, CNB, TMB, diagnosis age, ER status, PR status). Detailed feature rankings are shown in Figure 3.9, 3.10, and 3.11. C-Index changed: The median value of changed C-Index.

Ranks	Age group 26–50		Age group 51–70		Age group 71–90	
	Feature	C-Index changed	Feature	C-Index changed	Feature	C-Index changed
1	**PR status**	**−0.0247**	**ER status**	**−0.0807**	**Module 11**	**−0.0323**
2	Module 1	0	Module 13	−0.0221	**Module 1**	**−0.0233**
3	Module 2	0	Module 4	−0.0185	**Module 29**	**−0.0233**
4	Module 3	0	Module 5	−0.0150	**Module 35**	**−0.0233**
5	Module 4	0	Diagnosis age	−0.0150	**Module 4**	**−0.0222**

The performances in terms of C-Index by integrating all omics and clinical data (including mRNA, miRNA, CNB, TMB, diagnosis age, ER status, PR status) are shown in Figure 3.8. As expected they are all slightly inferior than the performance when not stratifying patients by age (median = 0.7285; mean = 0.6918), we do not observe a significant difference. When inspecting the feature rankings, as shown in Table 3.4, we observe that in the age group 26–50, PR status (Progesterone Receptors status) plays a pivotal role in prognosis, while other features do not have substantial contributions to the prognosis including the diagnosis age (we still list some modules). This situation changes in the age group 51–70 as ER status (Estrogen Receptors status) becomes the most important feature, while diagnosis age ranks at #5 with only marginal contribution. In age group 71–90, neither ER, PR status nor diagnosis age ranks in the front. Instead, mRNA-seq co-expression modules appear to have the major influence on prognosis. The top ranked modules are #11, #1, #29, #35, and #4. By performing enrichment analysis, we conclude that the module #11 is significantly enriched with epithelium development genes (GO:0060429, P-value $= 2.253 \times 10^{-9}$); module #1 is significantly enriched with chromosome organization genes (GO:0051276, P-value $= 5.344 \times 10^{-17}$) and two well-known breast cancer genes NCOA3 [188] and FOXA1 [189], [190] are identified in module #1; module #29 is enriched on cytoband 19q13.41 (P-value $= 1.517 \times 10^{-25}$) and is exclusively zinc-finger proteins; module #35 is enriched on cytoband 1q34 (P-value $= 1.252 \times 10^{-15}$) and contains multiple genes which have been previously

detected in multiple breast cancer studies including UQCRH, PSMB2, PPIH, and YBX1 [42], [191], [192]; and module #4 is highly enriched with mitotic cell cycle genes (GO:1903047, P-value $= 2.183 \times 10^{-70}$) including well-known breast cancer-related genes such as MKI67 [193] and AURKA [194]. Detailed feature rankings are reported in Figure 3.9, 3.10, and 3.11.

3.3.5 Identification of Breast Cancer Related Genes and Cytobands Associated with Important Modules

To inference the biological implication from the feature ranking, we perform gene ontology (GO) and cytoband enrichment from ToppGene Suite (https://toppgene.cchmc.org/) [182]. Gene ontology (GO) is one of the major bioinformatics initiatives with the aim of unifying the representation of gene across species [195]. The gene ontology covers three domains: (1) cellular component, (2) biological process, and (3) molecular function. Each GO term has a term name, a unique identifier, a definition, its ancestor, and the associated genes.

Specifically, we focus on analyzing top five mRNA co-expression modules (Table 3.3). We identify 10 genes such as MST1, CPT1B, MAP3K7, CCNC, etc. We also identify various enriched cytoband and other biological functions.

3.4 Discussion

As feature importance has been conveyed and ranked from SALMON, we find that keeping only top important features can still preserve the testing performances. Based on features ranking, we also investigate the biological interpretation behind each demographical feature, clinical feature, and co-expression module. For the leading two features, since the importance of diagnosis age and PR status have been widely examined and recognized in breast cancer [187], [196]–[198] and further confirmed by our results (Figure 3.2C), we focus on the top five mRNA-seq data co-expression modules ranked from 3 to 7. Those top five mRNA-seq data co-expression modules are: module #13, #47, #5, #36, #51.

Module #13 appears to be significantly associated with CD8+ T Cells (P-value $= 6.54 \times 10^{-6}$) and CD4+ T Cells (P-value $= 1.50 \times 10^{-2}$) based on Human Gene Atlas analysis.

CD8+ and CD4+ T cells are two important components of the immune system, which have been proved to have strong correlation with cancers [199], [200]. It contains multiple breast cancer related genes: (1) MST1 kinase, a core component of Hippo pathway, its phosphorylation can inhibit oncoproteins TAZ/YAP and regulate T-cell function. [201], [202]; and (2) CPT1B, which encodes the critical enzyme for fatty acid beta-oxidation (FAO), the inhibition of FAO can inhibit breast cancer stem cells, chemoresistance, and breast tumor growth [203]. In addition, tissues enrichment analysis using ARCHS4 [204] also reveals that nearly one third of genes (11 out of 36) in this module are associated with breast cancer bulk tissue (P-value $= 1.867 \times 10^{-3}$) (Figure 3.12).

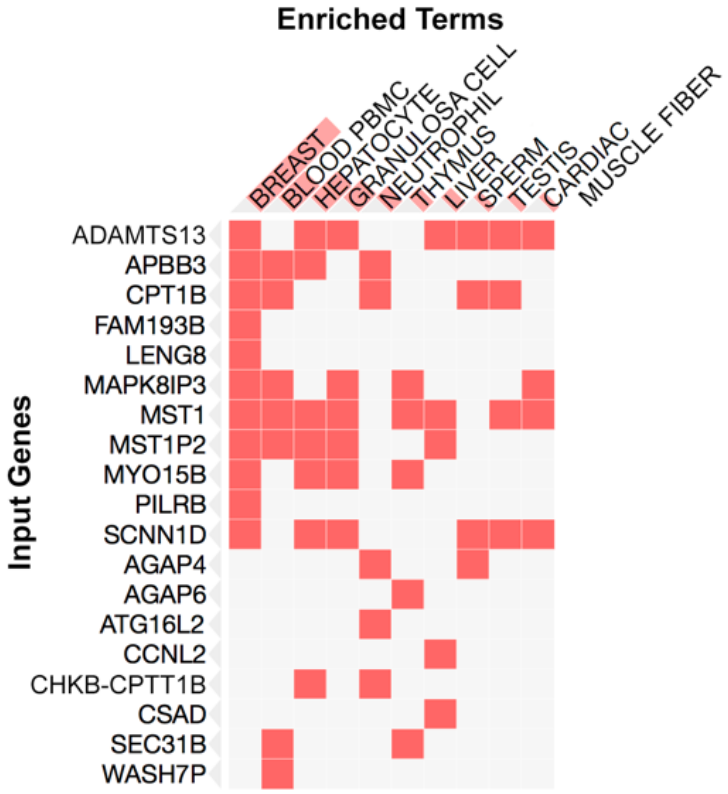


Figure 3.12. Enriched ARCHS4 Tissues terms with mRNA co-expression modules 13. nearly one third of genes (11 out of 36) in this module are associated with breast cancer bulk tissue (P-value $= 1.867 \times 10^{-3}$). Results are generated from the Enrichr online web server.

In module #47, two genes related to breast cancer are identified: (1) MAP3K7, also known as TAK1, is a key mediator between survival and cell death in TNF-α-mediated signaling [205]; and (2) CCNC, an important transcriptional regulator whose higher expression is associated with shorter relapse-free survival and impact the response to adjuvant therapy in breast cancer. Gene amplification of CCNC is also the most frequent type of genetic alterations in breast cancers [206]. Module #47 is also enriched in cytoband chr6q.

In module #5, genes are highly enriched on tumor micro-environment (TME) related processes such as extracellular matrix (ECM), cell adhesion, and cell migration. Among them, DDR2 played an indispensable role in a series of hypoxia-induced behaviors of breast cancer cells, such as migration, invasion, and epithelial-mesenchymal transition (EMT), the activated DDR2 can promote the metastasis of breast cancer [207]. In addition, the over-expression of FLNA is associated with the advanced stage, lymph node metastasis, and vascular or neural invasion of breast cancer [208]. It also contributes the development of breast cancer [209]. Finally, TCF4 is an important transcription factor, the loss of TCF4 related to breast cancer chemoresistance [210].

In module #36, SNW1 is a component of spliceosome in RNA splicing, its deletion can induce apoptosis, where the inhibition of SNW1 or its associating proteins may be a novel therapeutic strategy for cancer treatment [211]. Module #36 is also enriched in cytoband chr14q23–q24 and chr14q31–q32.

In module #51, TCP1 functions as a cytosolic chaperone in the biogenesis of tubulin [212], which has been proven to have an association with breast cancer [213]. HDAC2, its over-expression has a correlation with DNA-damage response and promote tumor progression [214]. Module #51 is also enriched on cytoband chr6q.

Instead of the identified breast cancer related genes, the enrichment analysis in selected modules also reveals important biological functions. Module #47 and #51 are enriched in chr6q. Not surprisingly, previous studies identified the frequent alterations at chr6q in archival breast cancer specimens [215], while chr6q21 is hotspots copy number alteration region [216]. The copy number alterations at chr6q26 can affect MAP3K4, play an important role of epidermal growth factor receptor pathway [215]. Module #36 is enriched in chr14q, the cytoband where the high-level alterations at 14q31.3–32.12 are found in breast cancer

[215]. Besides, the deletion of chr14q is a common feature of tumors with BRCA2 mutations [217]. Modules #5 is specifically associated with TME related biological process such as extracellular matrix (ECM), cell adhesion and cell migration. All these GO biological processes have been shown to play pivotal roles in TME development in cancers while TME has now been widely recognized as a critical participant in tumor progression [218]. Abnormal ECM in tumors can promote the aggressiveness of breast cancer [219]. Cell adhesion as a common event in cancer can promote cell growth as well as tumor dissemination [220], [221]. All these discoveries not only confirm the existed breast cancer studies, but also justify the feature importance that SALMON generated.

Another interesting finding is that no miRNA-seq module is ranked in top features although miRNA-seq modules show a better prognosis performance than mRNA-seq modules. This could due to the modules within miRNA-seq are more dependent with each other than the modules within mRNA-seq, thus simply knock out one module/feature may not reduce the performance too much. Indeed, by performing pair-wised Pearson correlation analysis, we find 3.03% miRNA-seq modules have strong correlations (Pearson $\rho > 0.8$), while in mRNA-seq modules this ratio is down to 0.94%. It leads us a new perspective to inspect modules dependency in the future.

Since we confirm that diagnosis age is the most powerful predictor, we examine the feature rankings with three different age groups, namely, younger group (age 26–50), middle aged group (age 51–70), and elderly group (age 71–90). We confirm that by separating the 583 patients into three distinct age groups, the diagnosis age becomes less important to the prognosis outcome. In younger group, PR status is the most important feature. In middle aged group, ER status is the most important feature. When we inspect the elderly group with age ranged 71–90, we find that only mRNA-seq co-expression modules are ranked at the top and the five most conspicuous ones are modules #11, #1, #29, #35, and #4. These observations suggest that specific biological processes may play different roles in breast cancer patients of different ages, while biomarkers and predictive models may alter in different age groups. Further inspection of the modules finds that three of these modules are related to several breast cancer related processes such as epithelium development [222], chromosome organization [223], and mitotic cell cycle [224] including well-known breast cancers genes

such as NCOA3, AURKA, MKI67, and FOXA1. The other two modules are highly enriched on specific cytobands on different chromosomes, implying potential copy number variations on these regions. Indeed, both cytobands (19q13.41 and 1q34) are known to be associated with breast cancer outcomes [225], [226]. For module #35, while most of the genes locate on 1q34, many of the genes such as UQCRH, PSMB2, PPIH, and YBX1 are involved in RNA processing and have been identified with breast cancer in multiple studies [42], [191], [192]. Interestingly, all genes identified from module #29 are zinc finger transcription factors. although it is not clear if any of them are specifically related to breast cancer, it is of great interest to further investigate the roles of the ZNF family genes in breast cancer development.

3.5 Conclusion

We perform survival prognosis on breast cancer, propose a deep learning-based algorithm SALMON (Survival Analysis Learning with Multi-Omics Network) by integrating Cox proportional hazards model and adopting gene co-expression network analysis results as input. The model was able to predict patient survival. The performances (C-Index and log-rank test P-value) improve when more omics data integrates to the input of SALMON. Besides, SALMON also shows a competitive performance compared to other deep learning survival prognosis model. By inspecting how each feature contributed to the hazard ratios, SALMON confirms certain mRNA-seq co-expression modules and clinical information, which play pivotal roles in breast cancer prognosis and revealed several biological functions. By further stratifying patients with diagnosis age, SALMON confirms that different age groups have different main features that control survival prognosis performance. To sum up, SALMON fuses the gene co-expression network analysis, deep learning technique, feature selection, Cox proportional hazard model, integrative analysis, and module-level enrichment analysis altogether, offers a new avenue for the future integrative analysis and deep learning-based cancer survival prognosis.

In this work, SALMON demonstrated the feasibility of breast cancer survival prognosis by integrating multi-omics data using a deep learning-based approach. SALMON also pro-

vided new prognostic biomarkers for breast cancer. Compared to other related algorithms, SALMON has several advantages:

- Previous studies either use fully connected neural networks with high dimensional omics input [24], [25], or use autoencoders to reduce input feature dimension [30], [32], but the survival prediction was a separate Cox regression task. Instead of using an autoencoder to reduce feature dimension or using gene level mRNA-seq or miRNA-seq data directly, we perform gene co-expression as upfront feature engineering, which summarizes highly co-expressed genes into modules and also help to understand the survival associated features at the module-level. Using a fewer number of features as input can also avoid overfitting, especially when the training data size is small.

- By bridging the gap between gene co-expression analysis and deep learning, it allows us to backtrack and identify the module/feature that affects the performances. The detected modules reveal cancer related biological processes and functions or structural variations allowing for further biomedical investigations. This is an added advantage.

- Unlike other fully connected models such as Cox-nnet [24], SALMON performs forward propagation separately with respect to each type of omics or clinical data. The separation of forward propagation prevents the interactions across omics data types thus enabling easier examination of the module/feature importance for interpretability, and also simplifies the number of neural network weights. Moreover, it demonstrated good prognosis results in terms of concordance index and log-rank test.

- Different from other studies which use the entire cohort of cancer patients for survival analysis, SALMON further stratifies the breast patient cohort into three subgroups according to the diagnosis age. By retraining different age groups, SALMON confirms the different prognostic markers that are associated with survival prediction.

Although SALMON has been successfully demonstrated for breast cancer patients, including patients with specific biomarkers such as ER and PR status, yet it does have several limitations. When SALMON is applied to other cancer data for survival prognosis, the performance tends be different. For example, [1] compared 12 different TCGA cancer types

and suggested that deep learning-based cancer survival prediction using gene expressions is well suited for breast cancer compared to other cancer types such as lung squamous cell carcinoma (LUSC) and stomach adenocarcinoma (STAD). [1] further suggested that there is a direct relationship between overall survival statistics and survival prediction performances and concluded that if a cancer data cohort has shorter overall survival times, then an inferior survival prediction performance is observed.

Furthermore, similar to all other deep learning models, SALMON may not achieve an acceptable performance when the dataset has a small number of samples. When multi-omics data are used, this sample shortage is further aggravated when one of the omics data types has fewer samples than others. For example, the TCGA ovarian serous cystadenocarcinoma (OV) cohort has 453 miRNA-seq data, but only has 304 mRNA-seq data, and hence fewer samples which have all multi-omics are qualified for neural networks training, which in turn leads to weakened performance.

We have demonstrated that the survival analysis can be incorporate with deep learning and co-expression network analysis in integrative analysis approach, and uncovered rich biological interpretation. In the next chapter, low-rank approximation will be introduced in conjunct with survival analysis. Specifically, Cox proportional hazards regression will be inserted into non-negative matrix factorization algorithm. The proposed algorithm demonstrates its superiority for both synthetic datasets and human cancer datasets in terms of C-Index, accuracy, and Dice coefficient. The power of unraveling latent interactions behind human cancer data is also elucidated.

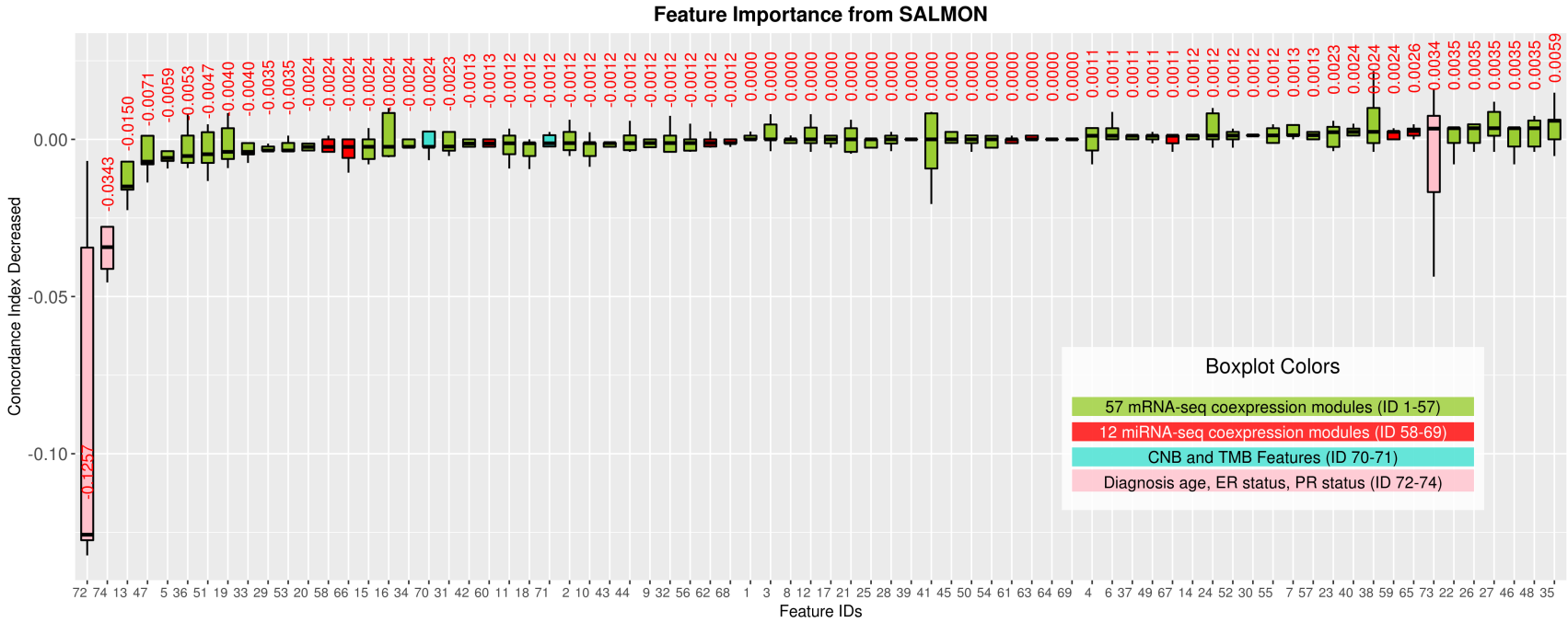


Figure 3.6. Features importance evaluated by the decrease of C-Index, based on median values and sorted in ascending order. Boxplots in Green: 57 mRNA co-expression module features (ID from 1 to 57); boxplots in red: 12 miRNA co-expression module features (ID from 58 to 69); boxplots in turquoise: copy number burden (CNB) and tumor mutation burden (TMB) features (ID from 70 to 71); boxplots in pink: demographical and clinical features (ID from 72 to 74).

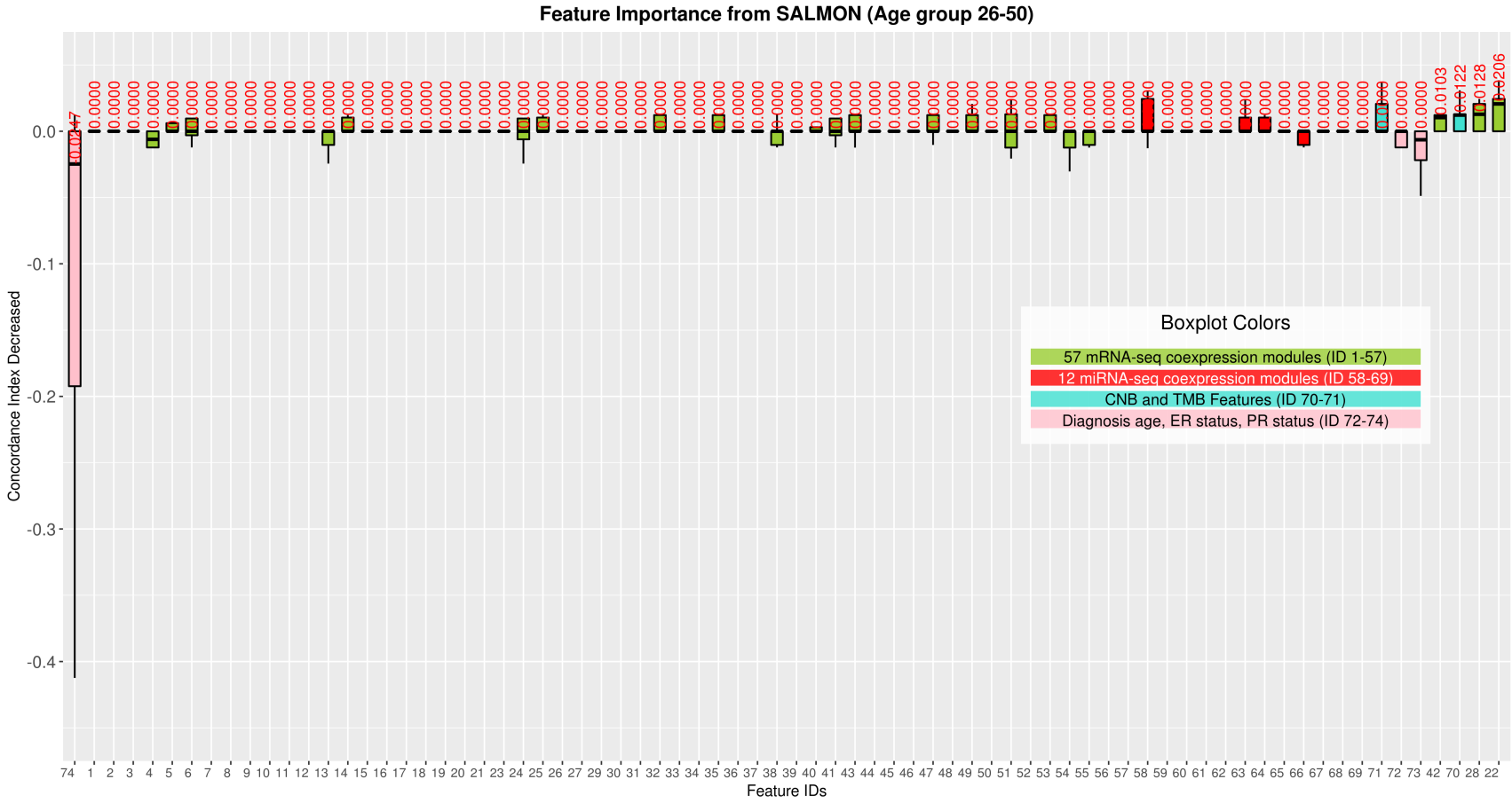


Figure 3.9. Feature importance with the diagnosis age in range 26–50, evaluated by the decrease of C-Index, sorted based on median values. Boxplots in Green: 57 mRNA co-expression module features (ID from 1 to 57); boxplots in red: 12 miRNA co-expression module features (ID from 58 to 69); boxplots in turquoise: copy number burden (CNB) and tumor mutation burden (TMB) features (ID from 70 to 71); boxplots in pink: demographical and clinical features (ID from 72 to 74).

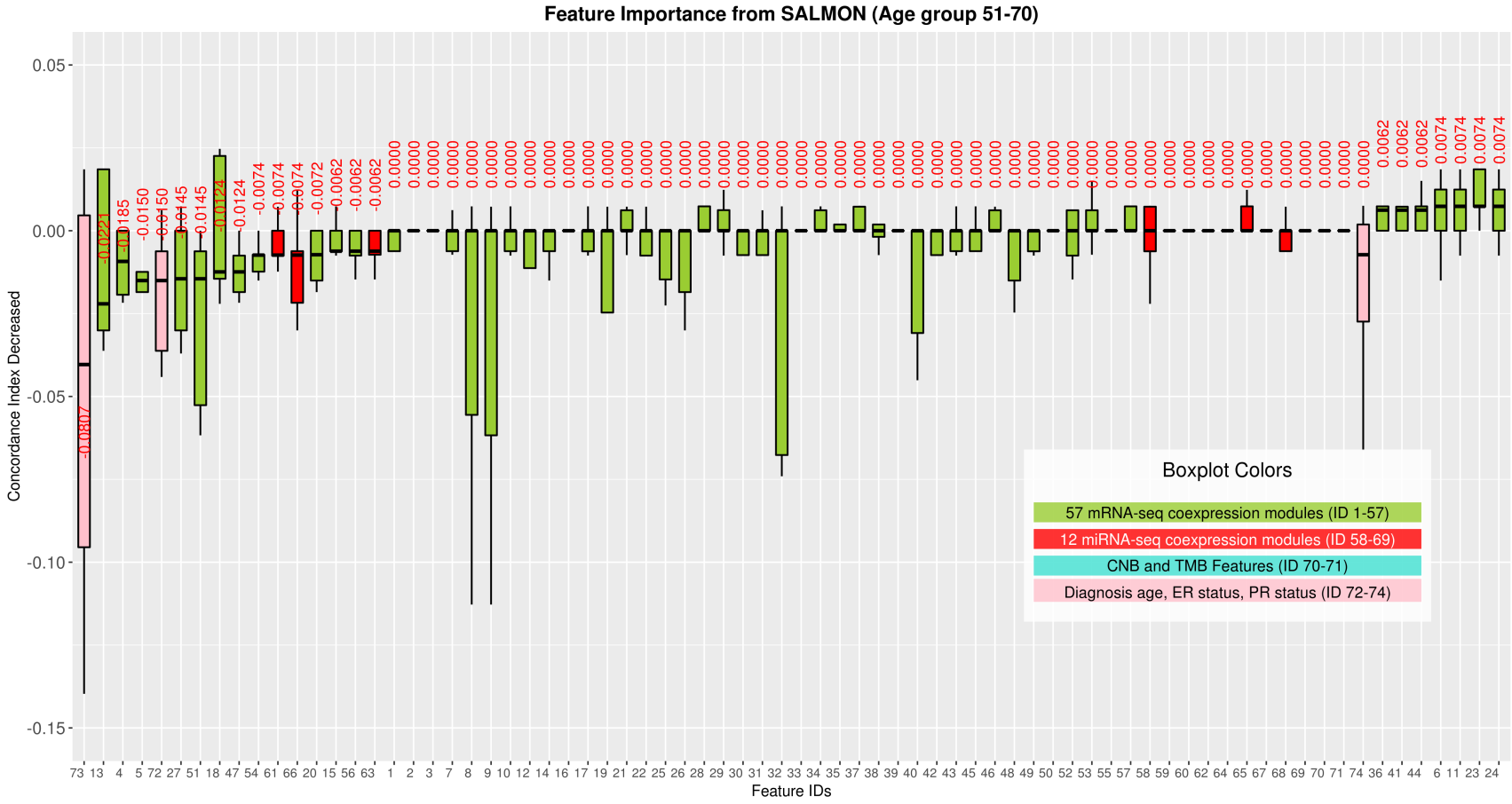


Figure 3.10. Feature importance with the diagnosis age in range 51–70, evaluated by the decrease of C-Index, sorted based on median values. Boxplots in Green: 57 mRNA co-expression module features (ID from 1 to 57); boxplots in red: 12 miRNA co-expression module features (ID from 58 to 69); boxplots in turquoise: copy number burden (CNB) and tumor mutation burden (TMB) features (ID from 70 to 71); boxplots in pink: demographical and clinical features (ID from 72 to 74).

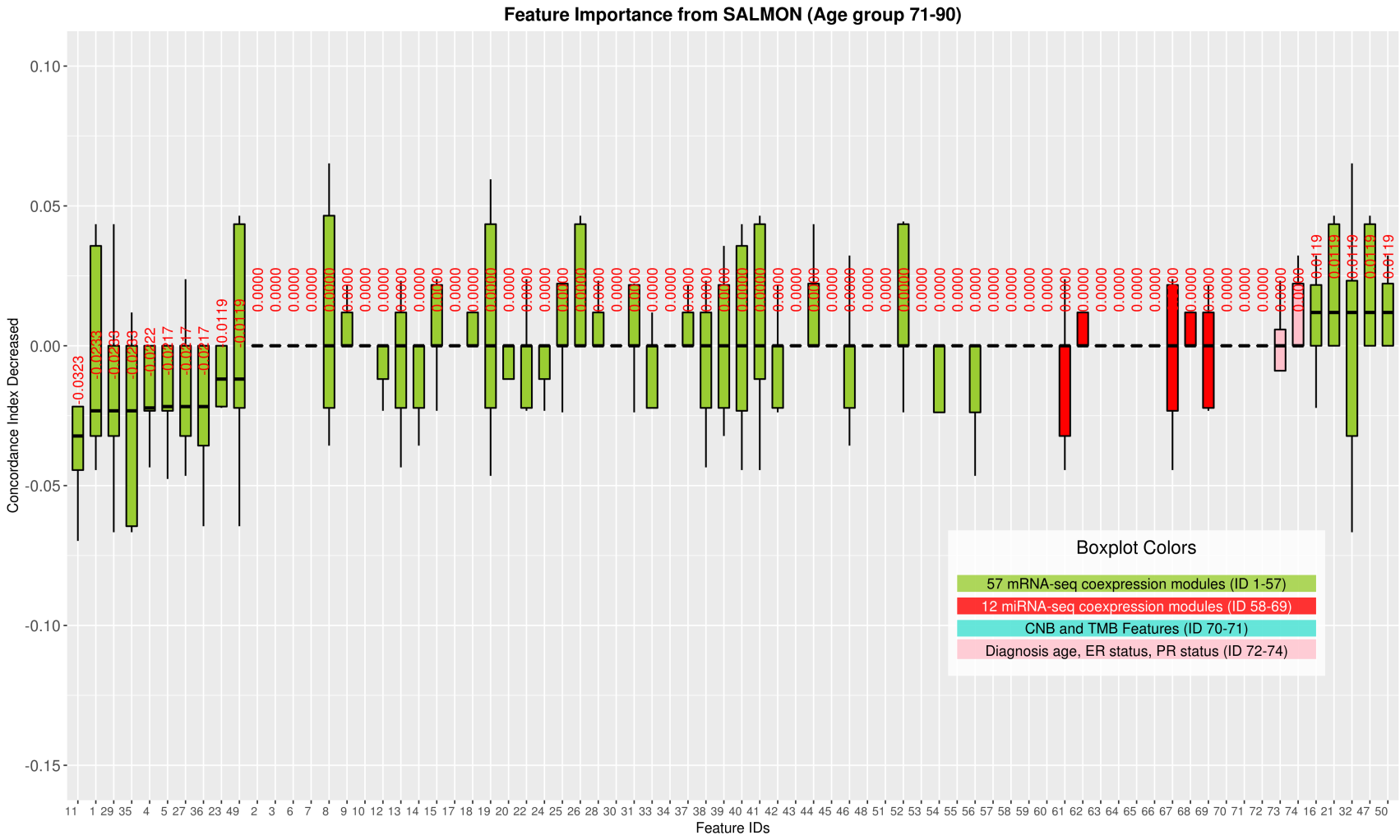


Figure 3.11. Feature importance with the diagnosis age in range 71–90, evaluated by the decrease of C-Index, sorted based on median values. Boxplots in Green: 57 mRNA co-expression module features (ID from 1 to 57); boxplots in red: 12 miRNA co-expression module features (ID from 58 to 69); boxplots in turquoise: copy number burden (CNB) and tumor mutation burden (TMB) features (ID from 70 to 71); boxplots in pink: demographical and clinical features (ID from 72 to 74).

4. COXNMF: A PROPORTIONAL HAZARDS NON-NEGATIVE MATRIX FACTORIZATION METHOD FOR IDENTIFYING SURVIVAL ASSOCIATED GENE CLUSTERS

4.1 Background and Introduction

In this chapter, we incorporate Cox proportional hazard regression into non-negative matrix factorization. Our goal is to unveil latent gene interactions under survival information constraints. We investigate the model performances and results in synthetic datasets and human gene expression datasets.

There are many ways to study and analyze high dimensional biological data. One common approach is to decompose original data into low-dimensional space, such as forming eigengene from co-expression analysis [34], or decompose data matrix into low-rank approximations. Matrix decomposition is a big family for solving many problems, including principal component analysis (PCA) [69], LU decomposition [227], QR decomposition [228], singular value decomposition (SVD) [68], etc. At this stage, we are especially interested in non-negative matrix decomposition (NMF). It has been mentioned earlier in Chapter 2 that the imposed non-negative property on NMF low-rank approximations can reflect biological interpretations [71]. NMF also has the attributes of better data interpretability and the intrinsic clustering property [70]. Furthermore, The non-negativity constraint applied to the low-rank approximations, preventing the cancellation among elements, provides abundant information with rich biological meaning.

NMF has an inherent clustering property and is equivalent to K-means clustering when imposing an orthogonality constraint [70], hence applying NMF in biological studies for unveiling gene interactions and clusters has received much attention in the last decade. For example, Brunet *et al.* [229] used NMF to decompose a gene expression matrix into metagenes to discover molecular patterns of those metagenes in leukemia and brain cancer datasets. This method is the analogue to that of discovering facial features as defined in [71]. Another study conducted by Liu *et al.* [230] compared the use of PCA and NMF in achieving dimensionality reduction of microarray data, followed by K-means clustering of 11

real gene expression datasets. The study suggested NMF can detect natural clusters, and demonstrated the superiority of NMF compared to PCA. Zheng *et al.* [231] similarly used NMF to identify the type of tumors on three public gene expression datasets, while both Wang *et al.* [232] and Gao *et al.* [233] performed cancer clustering with NMF algorithms. More recently, Zhu *et al.* [78] suggested that NMF is well-suited to analyze heterogeneous single-cell RNA-Seq data, and Jiang *et al.* [79] used NMF to unravel disease-related genes. Alternatively, Jia *et al.* [90] used discriminant NMF to get gene rankings, whereas Lai *et al.* [80] performed survival prediction after NMF pre-selection. These studies applied the original NMF algorithm [71] and focused on interpreting the gene clustering results but did not seek algorithm improvement.

In contrast, other studies designed new NMF algorithms with different loss functions or introduced new discrimination power. For example, Wang *et al.* [234] proposed LS-NMF to identify functionally related genes, which incorporated the gene expression uncertainty measurements into the objective function. Zafeirious *et al.* [72] proposed discriminant NMF (DNMF) for simultaneous low-rank estimation and classification by imposing the fisher discriminant in the NMF objective function. The fisher discriminant was then widely adopted in later NMF studies [74], [88] and applied to biomedical applications such as [90]. Alternatively, Chao *et al.* [91] added linear regression into the NMF updating rules, and proposed Supervised NMF (SNMF) for ICU mortality risk prediction. Nonetheless, these methods do not take survival information into the NMF updating rules, thus they are not well-suited for exploiting the latent survival gene clusters (or metagenes) given high-dimensional gene expression data and survival labels. To address this gap, we aim to find an ideal solution to the NMF along with the associate survival data simultaneously. The derived algorithm should both find the ideal low-rank decomposition according to survival information and reflect biological interpretations.

In this chapter, the algorithm "CoxNMF" will be proposed and elucidated, including the objective function, updating rule, and time & space complexity analyses. We also carry out the settings for model comparison, simulation with forty-two different combinations of time-to-event synthetic data, and the corresponding evaluation metrics. The analysis for real human cancer data is performed after the simulation studies, aims to discover the clusters

of genes which may play pivotal roles to the survival. To the best of our knowledge, we are the first work that unifying non-negative matrix factorization with Cox proportional hazards regression to discover cancer gene clusters.

4.2 Variables, Inputs, and Outputs

It is assumed that the input gene expression data $\boldsymbol{X}$ with shape P by N to be non-negative, real-valued 2-D matrix, which may contain several zero-valued entries. The rows indicate P features/genes, and the columns indicate N samples/patients. The output produced by CoxNMF is consisted of three parts: a low-rank K by N coefficient matrix $\hat{\boldsymbol{H}}$ which is learned from NMF and Cox proportional hazards regression, a low-rank P by K basis matrix $\hat{\boldsymbol{W}}$, associates with $\hat{\boldsymbol{H}}$ that minimizes the Frobenius norm $\left\|\boldsymbol{X} - \hat{\boldsymbol{W}}\hat{\boldsymbol{H}}\right\|_F$, and the simultaneously learned K by 1 Cox proportional hazards regression weight $\hat{\beta}$.

4.3 Objective Function

Given the target non-negative matrix $\boldsymbol{X}$, two initialized non-negative matrices $\boldsymbol{W}^{(0)}$ and $\boldsymbol{H}^{(0)}$, the objective function of CoxNMF is

$$\begin{aligned}\text{Minimize } \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2 - \alpha \sum_{\mathrm{i}:C_\mathrm{i}=1} \Big(\beta^T \boldsymbol{H}_\mathrm{i} - \log\Big(\sum_{\mathrm{j}:Y_\mathrm{j}\geq Y_\mathrm{i}} \exp(\beta^T \boldsymbol{H}_\mathrm{j})\Big)\Big) \\ + \frac{N}{2}\xi\Big(\gamma\|\beta\|_1 + (1-\gamma)\|\beta\|_2^2\Big),\end{aligned} \tag{4.1}$$

subject to $\boldsymbol{X} \in \mathcal{R}_{\geq 0}^{P\times N}$, $\boldsymbol{W} \in \mathcal{R}_{\geq 0}^{P\times K}$, $\boldsymbol{H} \in \mathcal{R}_{\geq 0}^{K\times N}$. Where

$$\|\beta\|_1 \triangleq \frac{1}{\kappa}\Big[\log\left(1+\exp(-\kappa\beta)\right) + \log\left(1+\exp(\kappa\beta)\right)\Big] \tag{4.2}$$

is the smooth approximation of the $\|\beta\|_1$ [235], [236], $\alpha > 0$ is the positive weight imposing Cox proportional hazards regression, $\|\cdot\|_F$ is the Frobenius norm, also known as Euclidean distance [81], C stands for the death events, Y stands for survival times. Weight $\frac{N}{2}\xi$ imposes an elastic net penalty [64] for the objective function, $\xi \geq 0$. $\gamma \in [0,1]$ balances the L1 and L2 ratio in the elastic net penalty. For a better smooth approximation of $\|\beta\|_1$, a higher κ is recommended. In this study, we set $\kappa = 1\times 10^{10}$. Through the Equation 4.2, we can impose the L1 penalty ratio as well as calculating its second order derivatives $\frac{\partial^2\|\beta\|_1}{\partial\beta\otimes\partial\beta}$.

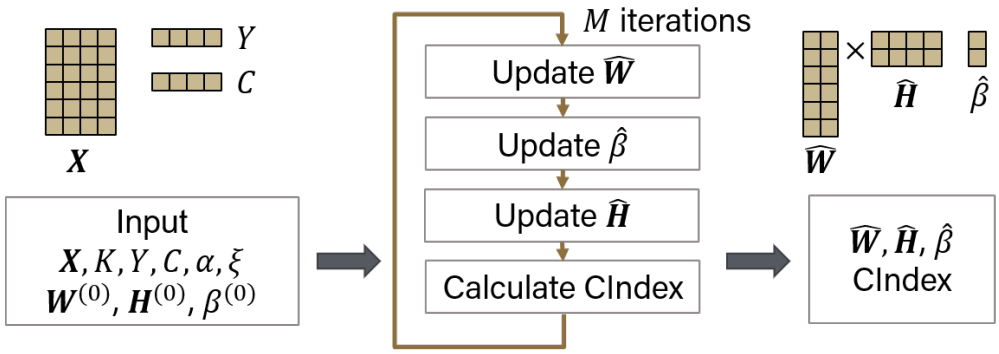


Figure 4.1. Flowchart of the proposed CoxNMF algorithm.

4.4 CoxNMF Update Rule

Given the target non-negative matrix $\boldsymbol{X}$, two initialized non-negative matrices $\boldsymbol{W}^{(0)}$ and $\boldsymbol{H}^{(0)}$, an initialized parameter $\beta^{(0)}$ for Cox proportional hazards regression, survival time vector Y and survival event vector C, and the maximum number of iterations M, we propose the CoxNMF alternately iterative update algorithm

$$\boldsymbol{W}_{\mathrm{i,j}}^{(iter+1)} \leftarrow \boldsymbol{W}_{\mathrm{i,j}}^{(iter)} \odot \frac{\boldsymbol{X}\boldsymbol{H}_{\mathrm{i,j}}^{(iter)^T}}{\boldsymbol{W}_{\mathrm{i,j}}^{(iter)}\boldsymbol{H}_{\mathrm{i,j}}^{(iter)}\boldsymbol{H}_{\mathrm{i,j}}^{(iter)^T}}, \tag{4.3}$$

$$\beta^{(iter+1)} \leftarrow \beta^{(iter)} - {\mathcal{H}_{\boldsymbol{H},\beta,\xi}^{(iter)}}^{-1} g_{\boldsymbol{H},\beta,\xi}^{(iter)}, \tag{4.4}$$

$$\beta^{(iter+1)} \leftarrow \frac{\beta^{(iter+1)}}{\max(\beta^{(iter+1)}) - \min(\beta^{(iter+1)})}, \tag{4.5}$$

$$\boldsymbol{H}_{\mathrm{i,j}}^{(iter+1)} \leftarrow \left(\boldsymbol{H}_{\mathrm{i,j}}^{(iter)} + \frac{\alpha}{2}\max\left\{\mathbf{0}, \frac{\partial \ell_{\boldsymbol{H}^{(iter)},\beta}(C,Y)}{\partial \boldsymbol{H}^{(iter)}}\right\}\right) \odot \frac{\boldsymbol{W}_{\mathrm{i,j}}^{(iter+1)^T}\boldsymbol{X}}{\boldsymbol{W}_{\mathrm{i,j}}^{(iter+1)^T}\boldsymbol{W}_{\mathrm{i,j}}^{(iter+1)}\boldsymbol{H}_{\mathrm{i,j}}^{(iter)}}. \tag{4.6}$$

Equation 4.5 is to normalize β. In the following context, we will omit $^{(iter)}$ superscript. Newton-Raphson [103] as maximum partial likelihood estimator (MPLE) [16] is used for updating Equation 4.4, where

$$\begin{aligned}\mathcal{H}_{\boldsymbol{H},\beta,\xi} &= \frac{\partial^2\left(-\alpha\ell_{\boldsymbol{H},\beta}(C,Y)+\xi\|\beta\|_1\right)}{\partial\beta\otimes\partial\beta}\bigg|_{\beta=\hat{\beta}^{(iter)}} \\ &= \alpha\sum_{\mathrm{i}:C_\mathrm{i}=1}\left(\frac{\sum_{\mathrm{j}:Y_\mathrm{j}\geq Y_\mathrm{i}}\exp(\beta^T\boldsymbol{H})_\mathrm{j}\boldsymbol{H}_\mathrm{j}\boldsymbol{H}_\mathrm{j}^T}{\sum_{\mathrm{j}:Y_\mathrm{j}\geq Y_\mathrm{i}}\exp(\beta^T\boldsymbol{H})_\mathrm{j}}\right. \\ &\quad\left.-\frac{\left[\sum_{\mathrm{j}:Y_\mathrm{j}\geq Y_\mathrm{i}}\exp(\beta^T\boldsymbol{H})_\mathrm{j}\boldsymbol{H}_\mathrm{j}\right]\left[\sum_{\mathrm{j}:Y_\mathrm{j}\geq Y_\mathrm{i}}\exp(\beta^T\boldsymbol{H})_\mathrm{j}\boldsymbol{H}_\mathrm{j}^T\right]}{\left[\sum_{\mathrm{j}:Y_\mathrm{j}\geq Y_\mathrm{i}}\exp(\beta^T\boldsymbol{H})_\mathrm{j}\right]^2}\right)+\xi\mathrm{diag}\left(\frac{\partial^2\|\beta\|_1}{\partial\beta\otimes\partial\beta}\right)\end{aligned} \tag{4.7}$$

is the Hessian matrix of Equation 2.15,

$$\begin{aligned}g_{\boldsymbol{H},\beta,\xi} &= \frac{\partial\left(-\alpha\ell_{\boldsymbol{H},\beta}(C,Y)+\xi\|\beta\|_1\right)}{\partial\beta}\bigg|_{\beta=\hat{\beta}^{(iter)}} \\ &= -\alpha\sum_{\mathrm{i}:C_\mathrm{i}=1}\left(\boldsymbol{H}_\mathrm{i}-\frac{\sum_{\mathrm{j}:Y_\mathrm{j}\geq Y_\mathrm{i}}\exp(\beta^T\boldsymbol{H})_\mathrm{j}\boldsymbol{H}_\mathrm{j}}{\sum_{\mathrm{j}:Y_\mathrm{j}\geq Y_\mathrm{i}}\exp(\beta^T\boldsymbol{H})_\mathrm{j}}\right)+\xi\frac{\partial\|\beta\|_1}{\partial\beta}\end{aligned} \tag{4.8}$$

is the partial gradient of Equation 2.15 with respect to the β, and

$$\begin{aligned}\frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}} &= \begin{bmatrix}\frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{1,1}} & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{1,2}} & \cdots & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{1,N}} \\ \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{2,1}} & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{2,2}} & \cdots & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{2,N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{K,1}} & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{K,2}} & \cdots & \frac{\partial\ell_{\boldsymbol{H},\beta}(C,Y)}{\partial\boldsymbol{H}_{K,N}}\end{bmatrix} \\ &= \left[\underbrace{\left(C_r\beta-\sum_{s=r}^{N}C_s\frac{\mathbb{1}_{(Y_r\geq Y_s)}\beta\exp(\beta^T\boldsymbol{H}_r)}{\sum_{\mathrm{j}:Y_\mathrm{j}\geq Y_s}\exp(\beta^T\boldsymbol{H}_\mathrm{j})}\right)}_{K\times 1\text{ vector which repeats }N\text{ times for }r=1,2,\cdots,N.}\right]\end{aligned} \tag{4.9}$$

is the partial derivative of Equation 2.15 with respect to the $\boldsymbol{H}$, $\mathbb{1}_{(Y_\mathrm{i}\geq Y_s)}=\begin{cases}1 & \text{if } Y_\mathrm{i}\geq Y_s\\ 0 & \text{otherwise}\end{cases}$ is the indicator function. The $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{H}}$ are returned where the algorithm achieved maximum concordance index (C-Index) during the optimization. If at any step the difference between current C-Index and maximum C-Index is higher or equal to the concordance index tolerance $tol=0.2$, the CoxNMF update will be terminated. The derivation of Equation 4.6 is further elaborated in Definition 4.4.1.

Algorithm 1: CoxNMF

Input : $\boldsymbol{X}$, K, Y, C, α, M, tol.
Output : $\boldsymbol{W}$, $\boldsymbol{H}$, β, and CI.
Initialization: Initialize $\boldsymbol{W}^{(0)}$, $\boldsymbol{H}^{(0)}$, $\beta^{(0)}$, empty list CI, $MaxCI = 0$.
for $iter = 0 : M - 1$ **do**

$\boldsymbol{W}^{(iter+1)} \leftarrow \boldsymbol{W}^{(iter)} \odot \frac{\boldsymbol{X}\boldsymbol{H}^{(iter)^T}}{\boldsymbol{W}^{(iter)}\boldsymbol{H}^{(iter)}\boldsymbol{H}^{(iter)^T}}$

$\beta^{(iter+1)} \leftarrow \beta^{(iter)} - \mathcal{H}^{(iter)^{-1}}_{\boldsymbol{H},\beta,\xi} g^{(iter)}_{\boldsymbol{H},\beta,\xi}$

$\beta^{(iter+1)} \leftarrow \frac{\beta^{(iter+1)}}{\max(\beta^{(iter+1)}) - \min(\beta^{(iter+1)})}$

$\boldsymbol{H}^{(iter+1)} \leftarrow \left(\boldsymbol{H}^{(iter)} + \frac{\alpha}{2}\max\left\{\boldsymbol{0}, \frac{\partial \ell_{\boldsymbol{H}^{(iter)},\beta}(C,Y)}{\partial \boldsymbol{H}^{(iter)}}\right\}\right) \odot \frac{\boldsymbol{W}^{(iter+1)^T}\boldsymbol{X}}{\boldsymbol{W}^{(iter+1)^T}\boldsymbol{W}^{(iter+1)}\boldsymbol{H}^{(iter)}}$

$CI^{(iter+1)} = \text{ConcordanceIndex}(\beta^{(iter+1)^T}\boldsymbol{H}^{(iter+1)}, Y, C)$

$imax = \text{argmax}_{(iter)}\, CI$

$MaxCI = CI^{(imax)}$

if $(CI^{(iter+1)} - MaxCI) \geq tol$ **then**

Break

end

end

$\hat{\boldsymbol{W}} = \boldsymbol{W}^{(imax)}$

$\hat{\boldsymbol{H}} = \boldsymbol{H}^{(imax)}$

$\hat{\beta} = \beta^{(imax)}$

return $\hat{\boldsymbol{W}}, \hat{\boldsymbol{H}}, \hat{\beta}, MaxCI$

4.4.1 Definition of Updating Coefficient Matrix for CoxNMF Algorithm

Definition 4.4.1. *We define the Equation 4.6 to update $\boldsymbol{H}$ during CoxNMF updating*

$$\boldsymbol{H}^{(iter+1)}_{\text{i,j}} \leftarrow \left(\boldsymbol{H}^{(iter)}_{\text{i,j}} + \frac{\alpha}{2} max\left\{\boldsymbol{0}, \frac{\partial \ell_{\boldsymbol{H}^{(iter)},\beta}(C,Y)}{\partial \boldsymbol{H}^{(iter)}}\right\}\right) \odot \frac{\boldsymbol{W}^{(iter+1)^T}_{\text{i,j}}\boldsymbol{X}}{\boldsymbol{W}^{(iter+1)^T}_{\text{i,j}}\boldsymbol{W}^{(iter+1)}_{\text{i,j}}\boldsymbol{H}^{(iter)}_{\text{i,j}}}, \tag{4.10}$$

where

$$\begin{aligned}\frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}} &= \begin{bmatrix} \frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_{1,1}} & \frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_{1,2}} & \cdots & \frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_{1,N}} \\ \frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_{2,1}} & \frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_{2,2}} & \cdots & \frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_{2,N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_{K,1}} & \frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_{K,2}} & \cdots & \frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_{K,N}} \end{bmatrix} \\ &= \begin{bmatrix} \underbrace{\left(C_r\beta - \sum_{s=r}^{N} C_s \frac{\mathbb{1}_{(Y_r \geq Y_s)}\beta exp(\beta^T \boldsymbol{H}_r)}{\sum_{\text{j}:Y_\text{j} \geq Y_s} exp(\beta^T \boldsymbol{H}_\text{j})}\right)}_{K \times 1 \text{ vector which repeats } N \text{ times for } r = 1, 2, \cdots, N.} \end{bmatrix}\end{aligned} \tag{4.11}$$

and $\mathbb{1}_{(Y_{\mathrm{i}} \geq Y_s)} = \begin{cases} 1 & \text{if } Y_{\mathrm{i}} \geq Y_s \\ 0 & \text{otherwise} \end{cases}$ *is the indicator function.*

Proof. Since the partial derivative $\frac{\partial\left(\|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\|_F^2-\alpha\ell_{\boldsymbol{H},\beta}(C,Y)\right)}{\partial \boldsymbol{H}}$ is

$$\frac{\partial\left(\|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\|_F^2-\alpha\ell_{\boldsymbol{H},\beta}(C,Y)\right)}{\partial \boldsymbol{H}} = -2\boldsymbol{W}^T\boldsymbol{X} + 2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H} - \alpha\frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}}. \tag{4.12}$$

Thus we have the update rule for $\boldsymbol{H}$:

$$\begin{aligned}
\boldsymbol{H} &\leftarrow \boldsymbol{H} - \eta_{\boldsymbol{H}} \odot \frac{\partial\left(\|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}\|_F^2-\alpha\ell_{\boldsymbol{H},\beta}(C,Y)\right)}{\partial \boldsymbol{H}} \\
&\leftarrow \boldsymbol{H} - \eta_{\boldsymbol{H}} \odot \left[-2\boldsymbol{W}^T\boldsymbol{X} + 2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H} - \alpha\frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}}\right] \\
&\leftarrow \boldsymbol{H} - \eta_{\boldsymbol{H}} \odot \left[-2\boldsymbol{W}^T\boldsymbol{X} + 2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}\right] + \alpha\eta_{\boldsymbol{H}} \odot \frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}} \\
&\leftarrow \boldsymbol{H} + \eta_{\boldsymbol{H}} \odot \left[2\boldsymbol{W}^T\boldsymbol{X} - 2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}\right] \\
&\qquad + \alpha\eta_{\boldsymbol{H}} \odot \left[\left(C_1\beta - \sum_{s=1}^{N} C_s \frac{\mathbb{1}_{(Y_1 \geq Y_s)}\beta\exp(\beta^T\boldsymbol{H}_1)}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_s}\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})}\right)\cdots \right. \\
&\qquad\qquad \left.\left(C_N\beta - \sum_{s=1}^{N} C_s \frac{\mathbb{1}_{(Y_n \geq Y_s)}\beta\exp(\beta^T\boldsymbol{H}_N)}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_s}\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})}\right)\right].
\end{aligned} \tag{4.13}$$

Given the same $\eta_{\boldsymbol{H}} = \frac{\boldsymbol{H}}{2\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}}$, as long as we project $C_r\beta - \sum_{s=r}^{N} C_s \frac{\mathbb{1}_{(Y_r \geq Y_s)}\beta\exp(\beta^T\boldsymbol{H}_r)}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_s}\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})}$ into the first orthant of K-dimensional space, make sure all the elements of

$$C_r\beta - \sum_{s=r}^{N} C_s \frac{\mathbb{1}_{(Y_r \geq Y_s)}\beta\exp(\beta^T\boldsymbol{H}_r)}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_s}\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})} \geq 0, \tag{4.14}$$

then we can guarantee the updating is non-negative with respect to $\boldsymbol{H}$.

According this projection rule, we get

$$\boldsymbol{H}_{\mathrm{i,j}} \leftarrow \left(\boldsymbol{H}_{\mathrm{i,j}} + \frac{\alpha}{2}\max\left\{\boldsymbol{0}, \frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}}\right\}\right) \odot \frac{{\boldsymbol{W}_{\mathrm{i,j}}}^T\boldsymbol{X}}{{\boldsymbol{W}_{\mathrm{i,j}}}^T\boldsymbol{W}_{\mathrm{i,j}}\boldsymbol{H}_{\mathrm{i,j}}}, \tag{4.15}$$

where

$$\begin{aligned}
\frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}} = &\left[\left(C_1\beta - \sum_{s=1}^{N} C_s \frac{\mathbb{1}_{(Y_1 \geq Y_s)}\beta\exp(\beta^T\boldsymbol{H}_1)}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_s}\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})}\right)\cdots \right. \\
&\left.\left(C_N\beta - \sum_{s=1}^{N} C_s \frac{\mathbb{1}_{(Y_N \geq Y_s)}\beta\exp(\beta^T\boldsymbol{H}_N)}{\sum_{\mathrm{j}:Y_{\mathrm{j}}\geq Y_s}\exp(\beta^T\boldsymbol{H}_{\mathrm{j}})}\right)\right].
\end{aligned} \tag{4.16}$$

□

4.4.2 Example of Updating Coefficient Matrix

An example data is provided to better understand the Equation 4.13. Suppose our $\boldsymbol{H}$ is with dimension K by 4 (4 patients), we have survival time $Y = [1, 3, 2, 4]$ and survival event $C = [1, 0, 1, 1]$. Then the partial derivative $\frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}}$ becomes

$$
\begin{aligned}
\frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}} &= \frac{\partial}{\partial \boldsymbol{H}} \sum_{\mathrm{i}:C_\mathrm{i}=1} \Big(\beta^T \boldsymbol{H}_\mathrm{i} - \log\Big(\sum_{\mathrm{j}:Y_\mathrm{j}\geq Y_\mathrm{i}} \exp(\beta^T \boldsymbol{H}_\mathrm{j})\Big)\Big) \\
&= \frac{\partial}{\partial \boldsymbol{H}} \Bigg(\beta^T \boldsymbol{H}_1 - \log\Big[\exp(\beta^T \boldsymbol{H}_1) + \exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_3) + \exp(\beta^T \boldsymbol{H}_4)\Big] \\
&\qquad + 0 \cdot \Big[\beta^T \boldsymbol{H}_2 - \log\Big[\exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_4)\Big]\Big] \\
&\qquad + \beta^T \boldsymbol{H}_3 - \log\Big[\exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_3) + \exp(\beta^T \boldsymbol{H}_4)\Big] \\
&\qquad + \beta^T \boldsymbol{H}_4 - \log\Big[\exp(\beta^T \boldsymbol{H}_4)\Big] \Bigg) \\
&= \frac{\partial}{\partial \boldsymbol{H}} \Bigg(\beta^T \boldsymbol{H}_1 - \log\Big[\exp(\beta^T \boldsymbol{H}_1) + \exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_3) + \exp(\beta^T \boldsymbol{H}_4)\Big] \\
&\qquad + \beta^T \boldsymbol{H}_3 - \log\Big[\exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_3) + \exp(\beta^T \boldsymbol{H}_4)\Big] \Bigg).
\end{aligned}
\tag{4.17}
$$

That is,

$$
\begin{aligned}
\frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_1} &= \beta - \frac{\beta \exp(\beta^T \boldsymbol{H}_1)}{\exp(\beta^T \boldsymbol{H}_1) + \exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_3) + \exp(\beta^T \boldsymbol{H}_4)}, \\
\frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_2} &= -\frac{\beta \exp(\beta^T \boldsymbol{H}_2)}{\exp(\beta^T \boldsymbol{H}_1) + \exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_3) + \exp(\beta^T \boldsymbol{H}_4)} \\
&\quad - \frac{\beta \exp(\beta^T \boldsymbol{H}_2)}{\exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_3) + \exp(\beta^T \boldsymbol{H}_4)}, \\
\frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_3} &= \beta - \frac{\beta \exp(\beta^T \boldsymbol{H}_3)}{\exp(\beta^T \boldsymbol{H}_1) + \exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_3) + \exp(\beta^T \boldsymbol{H}_4)} \\
&\quad - \frac{\beta \exp(\beta^T \boldsymbol{H}_3)}{\exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_3) + \exp(\beta^T \boldsymbol{H}_4)}, \\
\frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_4} &= -\frac{\beta \exp(\beta^T \boldsymbol{H}_4)}{\exp(\beta^T \boldsymbol{H}_1) + \exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_3) + \exp(\beta^T \boldsymbol{H}_4)} \\
&\quad - \frac{\beta \exp(\beta^T \boldsymbol{H}_4)}{\exp(\beta^T \boldsymbol{H}_2) + \exp(\beta^T \boldsymbol{H}_3) + \exp(\beta^T \boldsymbol{H}_4)}.
\end{aligned}
\tag{4.18}
$$

And for each patient r, the partial derivative $\frac{\partial \ell_{\boldsymbol{H},\beta}(C,Y)}{\partial \boldsymbol{H}_r}$ is

$$C_r\beta - \sum_{s=1}^{N} C_s \frac{\mathbb{1}_{(Y_r \geq Y_s)}\beta \exp(\beta^T \boldsymbol{H}_r)}{\sum_{\mathrm{j}:Y_\mathrm{j} \geq Y_s} \exp(\beta^T \boldsymbol{H}_\mathrm{j})}. \tag{4.19}$$

4.4.3 Time and Space Complexities

The time complexity of CoxNMF for one iteration consists of Equation 4.3, 4.4, 4.5, and 4.6 is $O(PNK + N^2K + K^2\max(P,N))$, the space complexity is $O(PN + N^2)$.

Theorem 4.4.1. *The time complexity of CoxNMF update for one iteration consists of Equation 4.3, 4.4, 4.5, and 4.6 is* $O(PNK + N^2K + K^2\max(P,N))$.

Proof. For Equation 4.3, since in general $K < \min(P,N)$, calculating the denominator in the form of $(\boldsymbol{W}\boldsymbol{H})\boldsymbol{H}^T$ would cost $O(PNK)$ operations, but calculating $\boldsymbol{W}(\boldsymbol{H}\boldsymbol{H}^T)$ costs $O(\max(P,N)K^2)$, thus the latter method is better [237]. For the numerator, $\boldsymbol{X}\boldsymbol{H}^T$ takes $O(PNK)$ operations. Thus, Equation 4.3 takes $O(PNK + \max(P,N)K^2 + PK)$ operations.

For Equation 4.4, since multivariate Newton-Raphson method (number of independent variables $= K$) is used to solve MPLE problem [238], and notice that we can pre-calculate some common structures. For example, the indication matrix $\sum_\mathrm{i}\sum_{\mathrm{j}:Y_\mathrm{j} \geq Y_\mathrm{i}}$ will cost $O(N^2)$, the $\exp(\beta^T\boldsymbol{H})$ will cost $O(NK^2)$. Thus, we can take $O(NK^2 + N^2 + N + K^2N + NK + K^2) = O(NK^2 + N^2)$ operations to get $\mathcal{H}$. Similarly, we can get g with $O(NK^2 + N^2 + NK)$ operations. Since the inverse of Hessian matrix $\mathcal{H}^{-1}$ can be $O(K^2\log(K))$ theoretically, and $\mathcal{H}^{-1}g$ will cost $O(K^2)$, thus the time complexity for Equation 4.4 is $O(N^2 + K^2(N + \log(K)))$.

For Equation 4.6, since the partial derivative of the partial log likelihood $\ell_{\boldsymbol{H},\beta}(C,Y)$ with respect to $\boldsymbol{H}$ in Equation 4.9 costs $O(NK^2 + N^2 + N^2K + N^2) = O(NK^2 + N^2K)$ operations, thus the Equation 4.6 takes $O(PNK + \max(P,N)K^2 + NK^2 + N^2K + NK)$ operations.

To sum up, the time complexity of CoxNMF update for one iteration consists of Equation 4.3, 4.4, 4.5, and 4.6 is $O(PNK + \max(P,N)K^2 + PK + N^2 + K^2(N + \log(K)) + PNK + \max(P,N)K^2 + NK^2 + N^2K + NK) = O(PNK + N^2K + K^2\max(P,N))$. □

Theorem 4.4.2. *The space complexity of CoxNMF update for one iteration consists of Equation 4.3, 4.4, 4.5, and 4.6 is* $O(PN + N^2)$.

Proof. For Equation 4.3, since in general $K < \min(P, N)$, the denominator in the form of $(\boldsymbol{W}\boldsymbol{H})\boldsymbol{H}^T$ would consume $O(PK+NK+K^2) = O(PK+NK)$ spaces, the numerator $\boldsymbol{X}\boldsymbol{H}^T$ consume $O(PN+NK)$ spaces. Thus, Equation 4.3 consume $O(PN+PK+NK) = O(PN)$ spaces.

For Equation 4.4, notice that the common structure $\exp(\beta^T \boldsymbol{H})$ consume $O(KN)$ spaces, and the indication matrix $\sum_{\mathrm{i}} \sum_{\mathrm{j}:Y_{\mathrm{j}} \geq Y_{\mathrm{i}}}$ will consume $O(N^2)$ spaces, thus it consumes $O(N^2 + NK + K^2) = O(N^2)$ spaces to get $\mathcal{H}$. Similarly, we can get g with $O(K + KN)$ spaces consumed. Thus, the space complexity for Equation 4.4 is $O(KN+N^2+K+KN) = O(N^2)$.

For Equation 4.6, since the partial derivative of the partial log likelihood $\ell_{\boldsymbol{H},\beta}(C, Y)$ with respect to $\boldsymbol{H}$ in Equation 4.9 consumes $O(N^2+KN)$ spaces, thus the Equation 4.6 consumes $O(PN + N^2)$ operations.

To sum up, the space complexity of CoxNMF update for one iteration consists of Equation 4.3, 4.4, 4.5, and 4.6 is $O(PN + N^2)$. □

4.5 Model Setup and Comparisons

4.5.1 Unconstrained Low-Rank Approaches

Starting from unconstrained matrix factorization approaches, truncated singular value decomposition (SVD) [239], principal component analysis (PCA) [240], sparse PCA [241], factor analysis (FA) [242], and non-negative double singular value decomposition (NNDSVD) [92] are adopted as baseline methods to perform dimensionality reduction upfront, then use the decomposed $\hat{\boldsymbol{W}}$ to cluster genes (features), and the decomposed $\hat{\boldsymbol{H}}$ for survival analysis. All unconstrained methods impose L1 norm on β with weight searching $\xi \in \{0, 0.01, 0.1, 1\}, \gamma = 1$. For the sparse PCA, we tune the sparsity controlling parameter in range $\{0.1, 1, 10, 100, 1000\}$ [243]. We denoted these low-rank approaches as "unconstrained" since they do not impose non-negativity in basis and coefficient matrices, and the Cox regression is performed afterwards.

4.5.2 Non-negative Matrix Factorization Approaches

CoxNMF is further compared with two vanilla NMF with Coordinate Descent (CD) solver and Mutiplicative Update (MU) solver [71]. In addition, CoxNMF is compared to supervised non-negative matrix factorization (SNMF) [91] which imposes linear regression into the NMF optimization process. In SNMF, survival time vector Y is regressed. Unlike CoxNMF which simultaneously minimizes the Frobenius norm and maximizes the partial log likelihood, methods with NMF (CD), NMF (MU), and SNMF perform Cox proportional hazards regression on $\hat{\boldsymbol{H}}$ afterwards. All NMF-based methods impose L1 norm on β with weight searching $\xi \in \{0, 0.01, 0.1, 1\}$, $\gamma = 1$. We tune $s_\alpha = \frac{v}{P}$, $s_\beta = \frac{v}{K}$, $s_\gamma = \frac{v}{N \times K}$, $v \in \{0.001, 0.01, 0.1\}$ for SNMF according to [91], $\alpha \in \{0.5, 1, 2, 5, 10, 20\}$ for CoxNMF, and use NMF(CD), NMF (MU), NNDSVD, or random as initializer for $\boldsymbol{W}^{(0)}$ and $\boldsymbol{H}^{(0)}$ in CoxNMF among all experiments, where P, N, and K stands for number of features, patients, and low-rank dimensions, respectively. In this paper, all algorithms including CoxNMF set the maximum number of iterations $M = 500$. In CoxNMF, the results where the C-Index achieved highest during the optimization will be used.

4.5.3 Hyper-parameters Choosing Criteria

Determining hyper-parameters of each algorithm along with estimating the $\hat{K}$ are two-folded. The first step is choosing the optimal pair of hyper-parameters for each $\hat{K}$ where it achieves highest C-Index (so each $\hat{K}$ will then have only one result). The second step is to choose $\hat{K}$ which has the highest silhouette score as the desired estimation.

The silhouette score, or mean silhouette coefficient, measures the consistency within clusters of data. It describes how well each element has been classified [132]. In this study, Euclidean distance is adopted as the distance metric. After row normalization on the resulting basis matrix $\bar{\boldsymbol{W}}_p = \hat{\boldsymbol{W}}_p / \left\| \hat{\boldsymbol{W}}_p \right\|$ for each row p, The hierarchical agglomerative clustering measures in Euclidean distance with Ward linkage [244] is used to cluster the basis matrix $\bar{\boldsymbol{W}}$. By performing this step we are able to get the gene clusters and the corresponding silhouette score.

4.5.4 Experimental Hardware and Running Time

All experiments are conducted in an 8-core Intel Xeon processor with 32 GB of RAM. All codes are written in Python version 3.8. All running times are measured in seconds. Specific python package versions are: lifelines v0.25.4, numpy v1.18.5, pandas v1.0.5, matplotlib v3.3.2, sklearn v0.23.2, scipy v1.5.0, seaborn v0.11.0, and tqdm v4.47.0.

4.6 Experimental Settings for Synthetic Data

In the simulation, we consider the dataset contains $N = 100$ patients. The survival times Y_{i} of patient i are sorted in ascending order follows exponential distribution $\lambda \mathrm{e}^{-\lambda Y}$ with $\lambda = 100$, $\mathrm{i} \in \{1, 2, \cdots, N\}$ according to [245]. All patients will have a complete event ($C_{\mathrm{i}} = 1$) in this ideal situation.

With K number of latent gene clusters each consists of 50 genes (ground truth), we have the ground truth basis matrix $\boldsymbol{W}$ in block diagonal with dimension P by K, each block is non-negative and is assumed to follow *i.i.d.* exponential distribution $\zeta \mathrm{e}^{-\zeta w}$ with $\zeta = 1$ [15], [246]. In this case, $P = 50 \cdot K$. We denote $\boldsymbol{W}_{[k]}$ as the k^{th} block consists of 50 genes in k^{th} cluster.

The ground truth coefficient matrix $\boldsymbol{H}$ with dimension K by N, each row $k \in \{1, \cdots, K\}$ of $\boldsymbol{H}_{k,\cdot}$ follows *i.i.d.* uniform distribution $\mathcal{U}(0, 1)$ except certain rows carry survival information. we suppose the $\boldsymbol{H}_{\mathrm{i},\cdot}$ for $\mathrm{i} \in \{1, 2, \cdots, \tau_1\}$ are associated with better prognosis (concord with survival time), and $\boldsymbol{H}_{\mathrm{j},\cdot}$ for $\mathrm{j} \in \{K - \tau_2 + 1, \cdots, K\}$ are associated with worse prognosis (discord with survival time), $\tau_1 + \tau_2 < K$.

4.6.1 Univariate Underlying Features

In this setting, the first/last rows are artificially replaced ($\tau_1 = \tau_2 = 1$). Specifically, the values of first row $\boldsymbol{H}_{k=1,\mathrm{i}}$ are constructed from 0.5 to 0.698 with step size = 0.02 (C-Index = 0). The values of last row $\boldsymbol{H}_{k=K,\mathrm{i}}$ are constructed from 0.698 to 0.5 with step size = -0.02 (C-Index = 1). This simulation setting hypothesizes that only first/last row carries survival

information which concord/discord with the survival. In this case, our goal is to unravel the gene clusters $\boldsymbol{W}_{[1]}$ and $\boldsymbol{W}_{[K]}$ associate with better/worse survival prognosis.

4.6.2 Multivariate Underlying Features

In reality, the underlying group of features may depend on each other. To simulate this scenario, we set $\tau_1 = \tau_2 = 2$. Initially all rows of the $\boldsymbol{H}$ follow *i.i.d.* uniform distribution $\mathcal{U}(0, 1)$. Then the second row $\boldsymbol{H}_{k=2,\mathrm{i}}$ is replaced by the array from 0.5 to 0.698 with step size = 0.02, further minus $\frac{1}{10}\boldsymbol{H}_{k=1,\mathrm{i}}$. Meanwhile, the $(K-1)^{\mathrm{th}}$ row $\boldsymbol{H}_{k=K-1,\mathrm{i}}$ is replaced by the array from 0.698 to 0.5 with step size = -0.02, further minus $\frac{1}{10}\boldsymbol{H}_{k=K,\mathrm{i}}$. Thus the first and second rows are depended on each other and both lead to the better prognosis, also vise versa for the K^{th} and $(K-1)^{\mathrm{th}}$ rows.

The ground truth data matrix $\boldsymbol{X}$ is then constructed as $\boldsymbol{X} = \boldsymbol{W}\boldsymbol{H} + \boldsymbol{E}$. The matrix $\boldsymbol{E}$ suggests an artificial noise introduced into the system, which follows exponential distribution $\varepsilon \mathrm{e}^{-\varepsilon w}$ with $\varepsilon \in \{0, 0.05, 0.1\}$.

Previous work suggests that either the Poisson distribution [247], [248] or the Exponential distribution [246] be used to model the error (or noise) of gene expression data. Although Poisson distribution can be used to model gene expression variation [249], [246] suggested a strictly decreasing probability of increasing gene expression TPM (Transcripts Per Million) values should be considered. In our human gene expression experiments, the gene expression data is normalized by RSEM (RNA-seq by Expectation-Maximization) [250], which provides a means to obtain TPM. Furthermore, a lot of entries in gene expression matrix are zero, which may not be appropriately modeled by a Poisson distribution. Since using the Exponential distribution can capture gene distribution with low read counts [246], we adopt it for modeling gene expression noise.

As K remains unknown in reality, the first step is required to determine the optimal $\hat{K}$ according to the silhouette score. We perform experiments with all combinations of $K \in \{6, 7, 8, 9, 10, 11, 12\}$ and search $\hat{K} = K \pm \{0, 1, 2\}$. All experiments perform 5 times each with different random seeds for the initialization and optimizations. Later from Table 4.7 and Table 4.8 we observe almost all algorithms can find the optimal $\hat{K} = K$ based on their

highest silhouette score especially the proposed CoxNMF. This step helps us to determine the number of the latent dimension K in the simulation study and in human cancer datasets as well.

4.7 Experimental Settings for TCGA Cancer Data

To improve precision health and cancer treatments, we are particularly interested in discovering gene clusters behind the gene expression matrix and the corresponding survival information. The goal of discovering latent cancer gene interaction groups can help biologist reveal gene functions, set up biological experiments, or help develop drugs based on targeted genes. Gene expression data (mRNA-seq) was downloaded from Broad GDAC Firehose (https://gdac.broadinstitute.org/). Since gene expressions remain a considerable amount of noises, 20% of genes with lowest expression mean and 20% of genes with lowest expression variance were excluded according to [2]. All expressions are normalized in $\log_2$ scale: $\boldsymbol{X} = \log_2(\boldsymbol{X} + 1)$ [2]. We end up with $P = 13{,}140$ genes for all cancers.

We search $\hat{K}$ from 10 to 30 with step size = 1, and determine the hyper-parameters using the same criteria described in the simulation study. We start searching $\hat{K}$ from 10 to ensure generally smaller sizes of the survival gene clusters. The optimal $\hat{K}$ is determined where the silhouette score is highest. All experiments set random seed = 1. In this study, we use ten cancer types to provide the guidance of choosing various hyper-parameters including α, ξ, and CoxNMF initialization method (Figure 4.5). The ten cancer types including BLCA: Bladder urothelial carcinoma; BRCA: Breast invasive carcinoma; COAD: Colon adenocarcinoma; KIRC: Kidney renal clear cell carcinoma; KIRP: Kidney renal papillary cell carcinoma; LIHC: Liver hepatocellular carcinoma; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; OV: Ovarian serous cystadenocarcinoma; PAAD: Pancreatic adenocarcinoma. We find that $\alpha = 5$, $\xi = 0.01$, and CoxNMF initialization method = random initialization will result in higher C-Index. Thus, we use this hyper-parameter pair for all following human cancer experiments.

4.8 Evaluation Metrics

We apply the Cox proportional hazards regression parameter weights on the normalized basis matrix: $\tilde{\boldsymbol{W}}_p = \hat{\beta}^T \odot \bar{\boldsymbol{W}}_p$ (*i.e.*, the $\hat{\beta}$ parameter will multiply each row p of $\bar{\boldsymbol{W}}$ element-wisely), and sort the columns of $\tilde{\boldsymbol{W}}$ and rows of $\hat{\boldsymbol{H}}$ according to $\hat{\beta}$ in ascending order, respectively. Note that $\hat{\beta}_k < 0$ suggests a reduction in hazard on k^{th} latent space, and the elements of $\tilde{\boldsymbol{W}}$ can be negative. In this case, a negative value in $\tilde{\boldsymbol{W}}$ suggests an association with better prognosis and vise versa.

4.8.1 Relative Error of Frobenius Norm

We introduce the Frobenius norm $\|\cdot\|_F$ in Equation 2.28 and Equation 4.1, which is the metric to evaluate the performance of NMF algorithms. Furthermore, we use the relative error [251], which is the ratio of $\left\|\boldsymbol{X} - \hat{\boldsymbol{W}}\hat{\boldsymbol{H}}\right\|_F$ to $\|\boldsymbol{X}\|_F$ in percentage, to evaluate whether a low-rank decomposition is adequately learned.

4.8.2 Silhouette Score for Determining Optimal Number of Latent Dimension K

The hierarchical agglomerative clustering is performed on $\bar{\boldsymbol{W}}$ to determine the optimal $\hat{K}$ according to the highest silhouette score and gene clusters $\phi_{\mathrm{j}}, \mathrm{j} \in \{1, 2, \cdots, \hat{K}\}$. Note that all models can only produce a low-rank basis matrix $\hat{\boldsymbol{W}}$, the hierarchical clustering is to further generate gene clusters given the normalized $\bar{\boldsymbol{W}}$.

4.8.3 Quantitative Measurements of CoxNMF Optimization Results and Label Accuracy

To tune the hyper-parameters as well as to evaluate the performance of optimization results along with survival information, the concordance index (C-Index) is adopted and defined in Equation 2.62. It is a generalization of the area under the ROC curve (AUC) which introduces the censorship information. Similar to the AUC, C-Index = 1 corresponds to the best model prediction, and C-Index = 0.5 represents a random prediction.

In simulation studies, to quantify whether models can identify the survival-associated gene clusters correctly, we adopt two measurements to evaluate the results, namely, accuracy and Dice coefficient [252].

To find the survival associated gene clusters, we focus on the τ_1 smallest $\hat{\beta}_1, \cdots, \hat{\beta}_{\tau_1}$ and τ_2 largest $\hat{\beta}_{K-\tau_2+1}, \cdots, \hat{\beta}_K$ associate with $\tilde{\boldsymbol{W}}$. The true labels ϕ_- (or ϕ_+) for better (or worse) prognosis genes are indicated as 1 at where the genes reside at $\boldsymbol{W}_{[1,\cdots,\tau_1]}$ (or at $\boldsymbol{W}_{[K-\tau_2+1,\cdots,K]}$). The estimated labels $\hat{\phi}_- = \sum_{\mathrm{i}=1}^{\tau_1} \mathrm{argmin}_{\Phi_\mathrm{j}} \frac{\sum_\mathrm{j} \tilde{\boldsymbol{W}}_{\Phi_\mathrm{j}=\mathrm{i},\mathrm{i}}}{\sum_{\mathrm{j}:\Phi_\mathrm{j}=\mathrm{i}} 1}$ and $\hat{\phi}_+ = \sum_{\mathrm{i}=K-\tau_2+1}^{K} \mathrm{argmax}_{\Phi_\mathrm{j}} \frac{\sum_\mathrm{j} \tilde{\boldsymbol{W}}_{\Phi_\mathrm{j}=\mathrm{i},\mathrm{i}}}{\sum_{\mathrm{j}:\Phi_\mathrm{j}=\mathrm{i}} 1}$ are determined by the highest mean absolute value on the associated low-rank dimensions (Operation $\mathrm{argmin}_{\Phi_\mathrm{j}}$ and $\mathrm{argmax}_{\Phi_\mathrm{j}}$ return binary label arrays where genes at Φ_j are indicated as 1). We concatenate ϕ_- and ϕ_+ as ϕ, and concatenate $\hat{\phi}_-$ and $\hat{\phi}_+$ as $\hat{\phi}$. Then we compare our true labels ϕ and the estimated labels $\hat{\phi}$ via two metrics: accuracy and Dice coefficient, to evaluate the performances of finding survival-associated gene clusters. Note that these two metrics are only valid in simulation study due to the absence of the ground truth labels in human cancer dataset. C-Index is determined during the model learning, while accuracy and Dice coefficient are unseen until the model with optimal C-Index is determined.

4.9 Results

Our method is first validated on forty-two different combinations of simulation study with univariate and multivariate underlying features setup, seven different sizes of synthetic data ($K \in \{6, 7, 8, 9, 10, 11, 12\}$), further perturbed by three different artificially induced noises ($\varepsilon \in \{0, 0.05, 0.10\}$). We then apply the proposed method on ten different TCGA human cancer datasets.

4.9.1 Simulation Results

From Table 4.1 we observe that the proposed CoxNMF can achieve highest C-Index (which is used for model selection) consistently among all K in univariate underlying features simulation when $\varepsilon = 0$. We also report the rest of univariate/multivariate simulation results with $\varepsilon \in \{0, 0.05, 0.10\}$ in Table 4.2, 4.3, 4.4, 4.5, and 4.6. The corresponding accuracy and relative error also lead other models among all experiments: The relationships

between C-Index and accuracy, Dice coefficient, and relative error are reported in Figure 4.2. Two simulation results where $K = \hat{K} = 10$, $\varepsilon = 0.05$ are presented for univariate experiment (Figure 4.3, C-Index = 1.0, accuracy = 0.9800, Dice coefficient = 0.8936) and multivariate experiment (Figure 4.4, C-Index = 1.0, accuracy = 0.9335, Dice coefficient = 0.7302), respectively. The complete univariate/multivariate simulation results including $\hat{K}$, relative errors, accuracy, Dice coefficient, running time in seconds, etc. are further reported in Table 4.7 and 4.8.

Table 4.1. Simulation results with univariate underlying features setup for $K \in \{6, 7, 8, 9, 10, 11, 12\}$, $\varepsilon = 0$. $\hat{K}$ is searched around $K \pm \{0, 1, 2\}$ and is determined by highest silhouette score. Experiments repeat 5 times each with random seed $\in \{1, 2, 3, 4, 5\}$. Mean values ± standard deviations are reported, best performed mean values among models are highlighted in bold font.

Metrics	K / Model	6	7	8	9	10	11	12
Accuracy	TruncatedSVD	**0.9667±0.07**	0.9429±0.08	0.9250±0.07	**0.9556±0.06**	0.9200±0.04	0.9273±0.04	**0.9500±0.05**
	PCA	0.7847±0.08	0.7740±0.05	0.7490±0.11	0.8073±0.02	0.9016±0.03	0.8689±0.02	0.8822±0.09
	SparsePCA	0.7657±0.03	0.7951±0.03	0.7933±0.01	0.8102±0.07	0.8304±0.07	0.7033±0.04	0.8327±0.08
	NNDSVD	0.9000±0.09	0.9143±0.08	0.8750±0.00	0.9111±0.05	0.9200±0.04	0.9273±0.04	**0.9500±0.05**
	FactorAnalysis	0.8533±0.12	0.8946±0.04	0.9235±0.07	0.8991±0.07	0.9114±0.07	0.9202±0.04	0.9240±0.04
	NMF (CD)	0.8000±0.07	0.8571±0.00	0.8750±0.09	0.8889±0.08	0.8800±0.11	0.8727±0.05	**0.9500±0.05**
	NMF (MU)	0.8667±0.07	0.7714±0.08	0.8250±0.07	0.8667±0.09	0.8400±0.05	0.8545±0.05	0.8833±0.05
	SNMF	0.9000±0.09	0.8857±0.06	0.8750±0.00	0.8889±0.00	0.9200±0.04	0.9273±0.04	0.9333±0.04
	CoxNMF	0.9333±0.09	**0.9714±0.06**	**0.9500±0.07**	**0.9556±0.06**	**0.9600±0.05**	**1.0000±0.00**	0.9333±0.07
Dice coefficient	TruncatedSVD	**0.9000±0.22**	0.8000±0.27	0.7000±0.27	**0.8000±0.27**	0.6000±0.22	0.6000±0.22	**0.7000±0.27**
	PCA	0.3765±0.11	0.1878±0.14	0.1948±0.14	0.0845±0.12	0.1606±0.16	0.2362±0.06	0.1029±0.12
	SparsePCA	0.1560±0.14	0.1704±0.16	0.0000±0.00	0.1167±0.11	0.1861±0.17	0.1641±0.09	0.0225±0.05
	NNDSVD	0.7000±0.27	0.7000±0.27	0.5000±0.00	0.6000±0.22	0.6000±0.22	0.6000±0.22	**0.7000±0.27**
	FactorAnalysis	0.6370±0.20	0.5831±0.19	0.6746±0.27	0.5282±0.29	0.5745±0.29	0.5304±0.21	0.5137±0.23
	NMF (CD)	0.4000±0.22	0.5000±0.00	0.5000±0.35	0.5000±0.35	0.4000±0.55	0.3000±0.27	**0.7000±0.27**
	NMF (MU)	0.6000±0.22	0.2000±0.27	0.3000±0.27	0.4000±0.42	0.2000±0.27	0.2000±0.27	0.3000±0.27
	SNMF	0.7000±0.27	0.6000±0.22	0.5000±0.00	0.5000±0.00	0.6000±0.22	0.6000±0.22	0.6000±0.22
	CoxNMF	0.8000±0.27	**0.9000±0.22**	**0.8000±0.27**	**0.8000±0.27**	**0.8000±0.27**	**1.0000±0.00**	0.6000±0.42

Additional Analyses on Synthetic Data

To investigate whether different results can be obtained when synthetically generated survival information is located on different low-rank positions, we perform additional simulation analyses on univariate and multivariate synthetic data, which we refer to as secondary univariate and multivariate simulation results, respectively.

In the secondary univariate simulation, the first and second rows of the matrix are replaced with synthetic data (we set $\tau_1 = \tau_2 = 1$). Specifically, the values of first row $\boldsymbol{H}_{k=1,\mathrm{i}}$ are

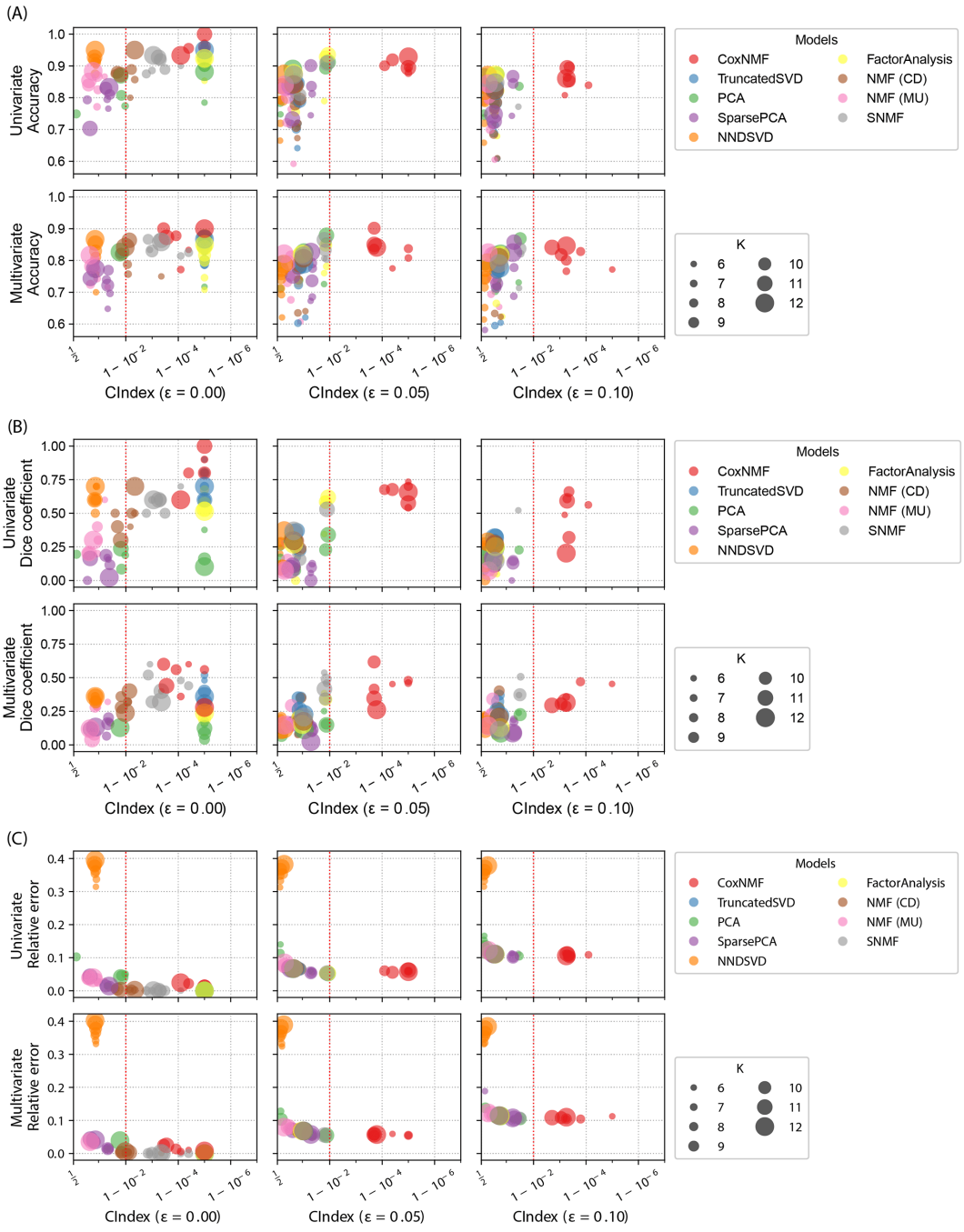


Figure 4.2. (A) C-Index and accuracy; (B) C-Index and Dice coefficient; and (C) C-Index and relative error among five unconstrained low-rank approaches and four NMF-based approaches across $K \in \{6, 7, 8, 9, 10, 11, 12\}$, and three different levels of artificial noise $\boldsymbol{E}$ for $\varepsilon \in \{0, 0.05, 0.10\}$ in both univariate and multivariate simulations. Mean values from 5 random seeds results are used for presenting this figure. X-axes are in logit scale. Figure best viewed in color. C-Index = 0.99 are indicated with red dashed lines.

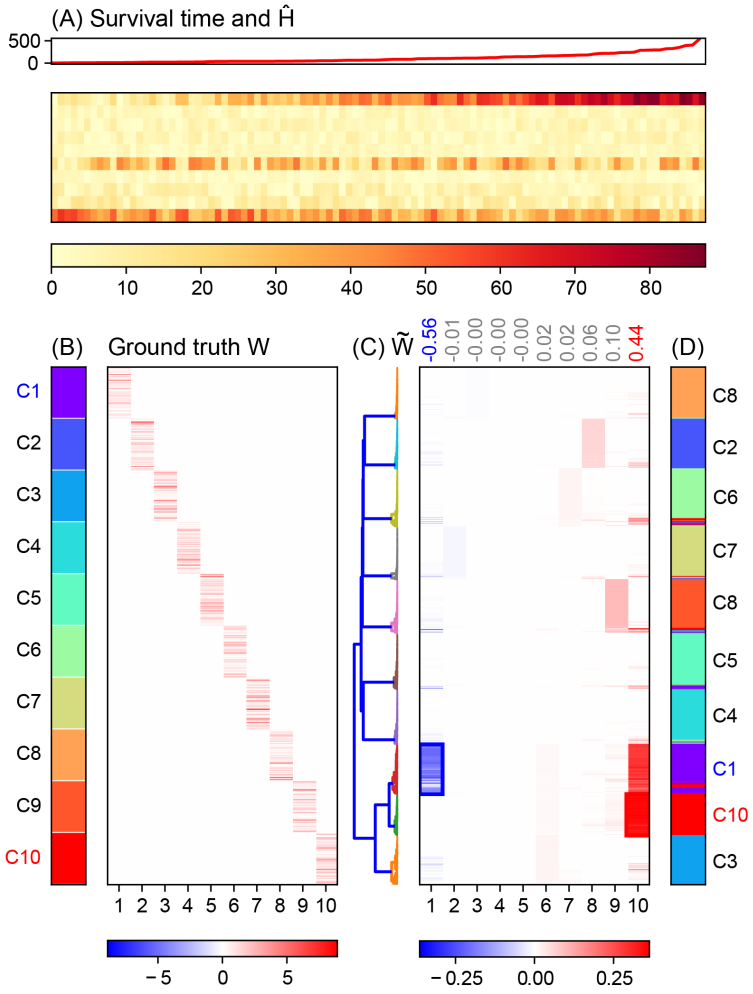


Figure 4.3. An univariate underlying features simulation result with CoxNMF model ($K = \hat{K} = 10$, $\varepsilon = 0.05$, $\alpha = 5$, $\xi = 0.1$, CoxNMF initialization = NNDSVD). (A) Survival time and $\hat{\boldsymbol{H}}$. (B) ground truth $\boldsymbol{W}$, note that $\boldsymbol{W}_{[1]}$ associate with better prognosis (longer survival time), $\boldsymbol{W}_{[K]}$ associate with worse prognosis. (C) $\tilde{\boldsymbol{W}}$ and hierarchical agglomerative clustering results (highlighted by most distinct colors, but do not relate with colors in (B) and (D)) with $\hat{K}$ number of clusters. Columns of $\tilde{\boldsymbol{W}}$ and rows of $\hat{\boldsymbol{H}}$ are sorted in ascending order of $\hat{\beta}$. Cluster with highest mean absolute value on the smallest $\hat{\beta}_1$ and cluster with highest mean absolute value on the largest $\hat{\beta}_K$ are highlighted with blue rectangle and red rectangle. (D) Ground truth labels in panel B with row permutation according to the hierarchical clustering result. Original and identified cluster ID associated with better and worse survival are highlighted in blue and red colors, respectively. In this figure, C-Index = 1.0, accuracy = 0.9800, dice coefficient = 0.8936, relative error = 5.4464%, and running time = 3.8185 seconds.

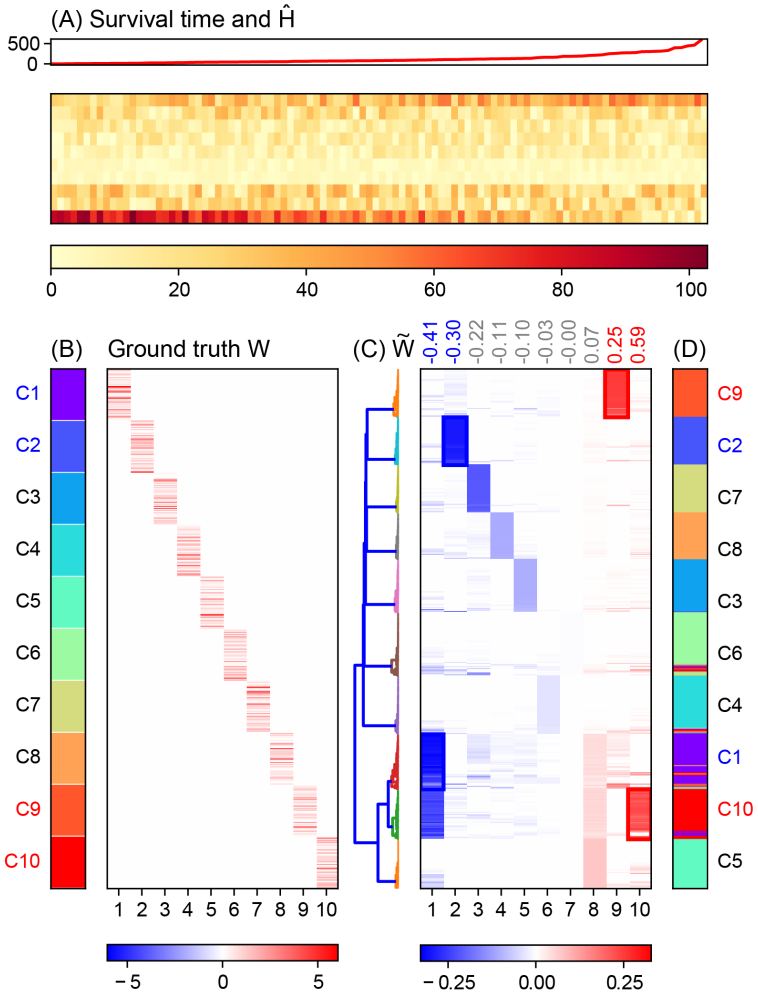


Figure 4.4. An multivariate underlying features simulation result with CoxNMF model ($K = \hat{K} = 10$, $\varepsilon = 0.05$, $\alpha = 5$, $\xi = 0.1$, CoxNMF initialization = NNDSVD). (A) Survival time and $\hat{\boldsymbol{H}}$. (B) ground truth $\boldsymbol{W}$, note that $\boldsymbol{W}_{[1]}$ associate with better prognosis (longer survival time), $\boldsymbol{W}_{[K]}$ associate with worse prognosis. (C) $\tilde{\boldsymbol{W}}$ and hierarchical agglomerative clustering results (highlighted by most distinct colors, but do not relate with colors in (B) and (D)) with $\hat{K}$ number of clusters. Columns of $\tilde{\boldsymbol{W}}$ and rows of $\hat{\boldsymbol{H}}$ are sorted in ascending order of $\hat{\beta}$. Cluster with highest mean absolute value on the smallest $\hat{\beta}_1$ and cluster with highest mean absolute value on the largest $\hat{\beta}_K$ are highlighted with blue rectangle and red rectangle. (D) Ground truth labels in panel B with row permutation according to the hierarchical clustering result. Original and identified cluster ID associated with better and worse survival are highlighted in blue and red colors, respectively. In this figure, C-Index = 1.0, accuracy = 0.9335, dice coefficient = 0.7302, relative error = 5.8849%, and running time = 3.8726 seconds.

Table 4.2. Simulation results with univariate underlying features setup for $K \in \{6, 7, 8, 9, 10, 11, 12\}$, $\varepsilon = 0.05$. $\hat{K}$ is searched around $K \pm \{0, 1, 2\}$ and is determined by highest silhouette score. Experiments repeat 5 times each with random seed $\in \{1, 2, 3, 4, 5\}$. Mean values ± standard deviations are reported, best performed mean values among models are highlighted in bold font.

Metrics	K / Model	6	7	8	9	10	11	12
Accuracy	TruncatedSVD	0.6413±0.09	0.7183±0.08	0.7012±0.06	0.7938±0.09	0.7888±0.05	0.8455±0.01	0.8497±0.05
	PCA	0.7477±0.06	0.7814±0.06	0.7853±0.05	0.8109±0.02	0.8932±0.03	0.9120±0.04	0.8905±0.02
	SparsePCA	0.7447±0.02	0.8123±0.09	0.7170±0.08	0.8362±0.02	0.9012±0.01	0.7333±0.06	0.8380±0.08
	NNDSVD	0.6650±0.11	0.7194±0.01	0.7932±0.06	0.7913±0.06	0.8308±0.05	0.8371±0.04	0.8750±0.04
	FactorAnalysis	0.7893±0.05	**0.9017±0.06**	0.7550±0.01	**0.9104±0.06**	**0.9284±0.04**	**0.9353±0.05**	0.8768±0.05
	NMF (CD)	0.6730±0.14	0.7197±0.07	0.7068±0.06	0.7767±0.06	0.7958±0.05	0.8396±0.01	0.8418±0.03
	NMF (MU)	0.5920±0.03	0.7640±0.10	0.7220±0.06	0.8211±0.07	0.8014±0.03	0.8296±0.01	0.8400±0.01
	SNMF	0.7070±0.09	0.7754±0.08	0.7927±0.08	0.8344±0.07	0.8092±0.04	0.9091±0.03	0.8723±0.04
	CoxNMF	**0.8897±0.08**	0.8771±0.07	**0.8947±0.05**	0.9007±0.04	0.9192±0.06	0.8945±0.02	**0.9288±0.05**
Dice coefficient	TruncatedSVD	0.1666±0.20	0.2425±0.20	0.1011±0.17	0.3156±0.29	0.2255±0.19	0.3697±0.01	0.3539±0.23
	PCA	0.2438±0.12	0.2050±0.17	0.1660±0.16	0.0737±0.10	0.2285±0.08	0.3400±0.12	0.3005±0.13
	SparsePCA	0.0552±0.12	0.1248±0.19	0.1164±0.13	0.0975±0.13	0.0000±0.00	0.1133±0.10	0.0839±0.12
	NNDSVD	0.1995±0.28	0.0645±0.14	0.2823±0.28	0.1147±0.22	0.2750±0.27	0.0980±0.21	0.3705±0.24
	FactorAnalysis	0.3454±0.08	**0.6045±0.24**	0.0000±0.00	0.5716±0.24	0.5855±0.22	**0.6186±0.27**	0.2708±0.31
	NMF (CD)	0.2400±0.33	0.2427±0.20	0.0931±0.17	0.2416±0.19	0.2382±0.19	0.2937±0.16	0.2952±0.15
	NMF (MU)	0.0757±0.14	0.3061±0.32	0.0916±0.17	0.2815±0.25	0.1506±0.20	0.0768±0.16	0.0749±0.17
	SNMF	0.1798±0.24	0.1980±0.27	0.1970±0.27	0.2515±0.35	0.1565±0.21	0.5293±0.18	0.3683±0.26
	CoxNMF	**0.7349±0.18**	0.5411±0.14	**0.7002±0.14**	**0.6760±0.10**	**0.6769±0.23**	0.5758±0.11	**0.6584±0.20**

Table 4.3. Simulation results with univariate underlying features setup for $K \in \{6, 7, 8, 9, 10, 11, 12\}$, $\varepsilon = 0.10$. $\hat{K}$ is searched around $K \pm \{0, 1, 2\}$ and is determined by highest silhouette score. Experiments repeat 5 times each with random seed $\in \{1, 2, 3, 4, 5\}$. Mean values ± standard deviations are reported, best performed mean values among models are highlighted in bold font.

Metrics	K / Model	6	7	8	9	10	11	12
Accuracy	TruncatedSVD	0.6083±0.08	0.7140±0.07	0.6890±0.06	0.7844±0.08	0.7834±0.05	0.8173±0.01	0.8297±0.04
	PCA	0.7390±0.07	0.7514±0.04	0.8360±0.03	0.8158±0.02	0.8470±0.02	0.8735±0.02	0.8633±0.02
	SparsePCA	0.7430±0.01	0.7849±0.05	0.7145±0.07	0.8449±0.01	**0.8674±0.08**	0.7260±0.07	0.7478±0.07
	NNDSVD	0.6583±0.11	0.7194±0.02	0.7878±0.05	0.7871±0.01	0.8322±0.05	0.8169±0.02	0.8712±0.04
	FactorAnalysis	0.6793±0.05	0.7531±0.09	0.7540±0.05	0.8009±0.05	0.8226±0.05	0.8633±0.01	**0.8722±0.05**
	NMF (CD)	0.6097±0.07	0.7440±0.10	0.6847±0.06	0.7658±0.04	0.7854±0.05	0.8295±0.01	0.8233±0.03
	NMF (MU)	0.6050±0.07	0.7606±0.10	0.7000±0.05	0.7944±0.07	0.8022±0.03	0.8380±0.01	0.8440±0.00
	SNMF	0.7717±0.03	0.7914±0.08	0.7637±0.07	0.7991±0.07	0.8286±0.05	0.8382±0.01	0.8473±0.01
	CoxNMF	**0.8080±0.00**	**0.8394±0.01**	**0.8645±0.01**	**0.8973±0.04**	0.8526±0.02	**0.8942±0.01**	0.8605±0.01
Dice coefficient	TruncatedSVD	0.0971±0.18	0.2363±0.19	0.1061±0.15	0.3088±0.26	0.2217±0.17	0.3281±0.01	**0.3172±0.20**
	PCA	0.2081±0.17	0.1475±0.15	0.2268±0.15	0.0832±0.11	0.1452±0.11	0.2441±0.05	0.1847±0.15
	SparsePCA	0.0000±0.00	0.1338±0.19	0.0994±0.10	0.1489±0.15	0.1291±0.20	0.1159±0.08	0.1632±0.05
	NNDSVD	0.2569±0.25	0.0631±0.14	0.2643±0.25	0.0041±0.01	0.2734±0.27	0.1195±0.16	0.2871±0.28
	FactorAnalysis	0.0469±0.10	0.2031±0.24	0.0608±0.14	0.1112±0.15	0.1604±0.15	0.2676±0.15	0.2516±0.27
	NMF (CD)	0.0872±0.17	0.3012±0.28	0.0957±0.15	0.1652±0.18	0.2241±0.18	0.2791±0.16	0.2690±0.14
	NMF (MU)	0.0933±0.16	0.3104±0.30	0.1056±0.15	0.1813±0.23	0.0776±0.17	0.1473±0.20	0.0716±0.15
	SNMF	**0.5217±0.12**	0.3150±0.31	0.1047±0.20	0.2210±0.31	0.2090±0.31	0.2217±0.19	0.1431±0.20
	CoxNMF	0.4883±0.11	**0.5623±0.10**	**0.6080±0.02**	**0.6612±0.11**	**0.3198±0.21**	**0.5926±0.02**	0.2033±0.19

filled with values ranging from 0.5 to 0.698 with step size = 0.02 (C-Index = 0). Similarly, the values of second row $\boldsymbol{H}_{k=2,\mathrm{i}}$ are filled with values ranging from 0.698 to 0.5 with step size

Table 4.4. Simulation results with multivariate underlying features setup for $K \in \{6, 7, 8, 9, 10, 11, 12\}$, $\varepsilon = 0$. $\hat{K}$ is searched around $K \pm \{0, 1, 2\}$ and is determined by highest silhouette score. Experiments repeat 5 times each with random seed $\in \{1, 2, 3, 4, 5\}$. Mean values ± standard deviations are reported, best performed mean values among models are highlighted in bold font.

Metrics	K / Model	6	7	8	9	10	11	12
Accuracy	TruncatedSVD	0.7833±0.05	**0.8286±0.04**	0.7875±0.03	0.8000±0.03	0.8500±0.04	0.8455±0.02	0.8667±0.03
	PCA	0.7173±0.02	0.8273±0.05	0.8602±0.01	0.8673±0.01	0.8743±0.01	**0.8745±0.01**	0.8263±0.02
	SparsePCA	0.6480±0.02	0.6964±0.06	0.7394±0.02	0.7694±0.06	0.7220±0.04	0.7435±0.05	0.7741±0.06
	NNDSVD	0.7000±0.05	0.7571±0.04	0.8000±0.03	0.8111±0.03	0.8300±0.04	0.8545±0.02	0.8667±0.03
	FactorAnalysis	0.7078±0.05	0.7459±0.04	0.7943±0.02	0.7993±0.04	0.8170±0.02	0.8409±0.01	0.8507±0.03
	NMF (CD)	0.7500±0.06	0.7571±0.06	0.7875±0.07	0.8222±0.02	0.8200±0.04	0.8636±0.05	0.8417±0.03
	NMF (MU)	0.7167±0.07	0.7429±0.04	0.7250±0.03	0.8000±0.05	0.7700±0.03	0.7818±0.02	0.8167±0.02
	SNMF	**0.8333±0.06**	0.8143±0.06	0.8250±0.07	0.8667±0.06	0.8300±0.04	0.8636±0.03	0.8583±0.02
	CoxNMF	**0.8333±0.06**	0.7714±0.06	**0.8625±0.05**	**0.8778±0.05**	**0.9000±0.00**	0.8727±0.06	**0.9002±0.03**
Dice coefficient	TruncatedSVD	0.4800±0.11	**0.5200±0.11**	0.3200±0.11	0.2800±0.11	0.4000±0.14	0.3200±0.11	**0.3600±0.17**
	PCA	0.2207±0.06	0.1554±0.09	0.1750±0.08	0.0412±0.05	0.0853±0.08	0.1242±0.11	0.1283±0.09
	SparsePCA	0.1735±0.14	0.1644±0.06	0.0657±0.10	0.2024±0.06	0.1579±0.05	0.1171±0.07	0.1324±0.06
	NNDSVD	0.2800±0.11	0.3200±0.11	0.3600±0.09	0.3200±0.11	0.3200±0.18	0.3600±0.09	**0.3600±0.17**
	FactorAnalysis	0.2656±0.12	0.2532±0.07	0.2728±0.10	0.2443±0.10	0.1939±0.06	0.2561±0.08	0.2380±0.13
	NMF (CD)	0.4000±0.14	0.3200±0.18	0.3200±0.23	0.3600±0.09	0.2800±0.18	0.4000±0.20	0.2400±0.17
	NMF (MU)	0.3200±0.18	0.2800±0.11	0.1200±0.11	0.2800±0.18	0.0800±0.11	0.0400±0.09	0.1200±0.11
	SNMF	**0.6000±0.14**	0.4800±0.18	0.4400±0.22	0.5200±0.23	0.3200±0.18	0.4000±0.14	0.3200±0.11
	CoxNMF	**0.6000±0.14**	0.3600±0.17	**0.5600±0.17**	**0.5600±0.17**	**0.6000±0.00**	**0.4400±0.26**	0.2783±0.20

Table 4.5. Simulation results with multivariate underlying features setup for $K \in \{6, 7, 8, 9, 10, 11, 12\}$, $\varepsilon = 0.05$. $\hat{K}$ is searched around $K \pm \{0, 1, 2\}$ and is determined by highest silhouette score. Experiments repeat 5 times each with random seed $\in \{1, 2, 3, 4, 5\}$. Mean values ± standard deviations are reported, best performed mean values among models are highlighted in bold font.

Metrics	K / Model	6	7	8	9	10	11	12
Accuracy	TruncatedSVD	0.6208±0.09	0.6027±0.05	0.7230±0.09	0.7737±0.02	0.7895±0.02	0.7820±0.06	0.7890±0.05
	PCA	0.7257±0.03	0.7113±0.02	0.7374±0.06	0.8546±0.00	0.7951±0.01	**0.8793±0.01**	0.8248±0.01
	SparsePCA	0.6728±0.02	0.6851±0.03	0.7752±0.03	0.7919±0.05	0.7361±0.04	0.7514±0.06	0.8264±0.01
	NNDSVD	0.6208±0.05	0.6924±0.04	0.7119±0.03	0.7424±0.03	0.7580±0.04	0.7724±0.01	0.7889±0.02
	FactorAnalysis	0.7602±0.03	0.7814±0.02	0.7398±0.06	0.7769±0.04	0.8481±0.03	0.8148±0.05	0.8201±0.04
	NMF (CD)	0.6405±0.08	0.6350±0.04	0.7396±0.07	0.7983±0.04	0.7754±0.02	0.7869±0.03	0.8071±0.04
	NMF (MU)	0.6077±0.07	0.6761±0.00	0.7305±0.05	0.7502±0.03	0.8114±0.03	0.7925±0.03	0.8214±0.02
	SNMF	**0.8067±0.03**	0.8043±0.04	0.7288±0.10	**0.8563±0.05**	0.8325±0.02	0.8654±0.02	0.8148±0.04
	CoxNMF	0.7752±0.09	**0.8079±0.04**	**0.8374±0.05**	0.8358±0.05	**0.9016±0.04**	0.8507±0.02	**0.8412±0.02**
Dice coefficient	TruncatedSVD	0.2442±0.17	0.0857±0.08	0.2899±0.25	0.3561±0.06	0.3498±0.06	0.2620±0.22	0.2279±0.19
	PCA	0.2512±0.09	0.1077±0.05	0.1377±0.06	0.1545±0.02	0.0707±0.04	0.1490±0.03	0.1130±0.09
	SparsePCA	0.1700±0.14	0.1422±0.10	0.1430±0.15	0.0972±0.09	0.1115±0.08	0.1583±0.03	0.0246±0.05
	NNDSVD	0.1506±0.08	0.1857±0.12	0.1964±0.13	0.2052±0.07	0.1657±0.15	0.0999±0.09	0.1179±0.08
	FactorAnalysis	0.3854±0.08	0.3414±0.05	0.1671±0.18	0.1862±0.13	0.3387±0.10	0.2007±0.21	0.1525±0.16
	NMF (CD)	0.2187±0.19	0.0775±0.09	0.2473±0.20	0.3494±0.12	0.2076±0.07	0.1738±0.12	0.1756±0.17
	NMF (MU)	0.1477±0.15	0.1807±0.00	0.1839±0.18	0.1431±0.15	0.2926±0.10	0.1447±0.14	0.1468±0.08
	SNMF	**0.5377±0.08**	0.4590±0.11	0.2168±0.29	**0.4894±0.18**	0.3424±0.08	**0.4164±0.08**	0.2070±0.18
	CoxNMF	0.4528±0.22	**0.4596±0.12**	**0.4795±0.17**	0.4158±0.17	**0.6171±0.12**	0.3462±0.10	**0.2612±0.14**

$= -0.02$ (C-Index $= 1$). A complete report of all secondary univariate simulation results is shown in Table 4.9.

Table 4.6. Simulation results with multivariate underlying features setup for $K \in \{6, 7, 8, 9, 10, 11, 12\}$, $\varepsilon = 0.10$. $\hat{K}$ is searched around $K \pm \{0, 1, 2\}$ and is determined by highest silhouette score. Experiments repeat 5 times each with random seed $\in \{1, 2, 3, 4, 5\}$. Mean values ± standard deviations are reported, best performed mean values among models are highlighted in bold font.

Metrics	K / Model	6	7	8	9	10	11	12
Accuracy	TruncatedSVD	0.6042±0.09	0.5951±0.06	0.7184±0.10	0.7681±0.03	0.7694±0.06	0.7666±0.07	0.7759±0.04
	PCA	0.7307±0.03	0.7031±0.03	0.7390±0.07	0.7858±0.03	**0.8688±0.01**	0.8215±0.03	0.8209±0.01
	SparsePCA	0.5813±0.06	0.6876±0.03	0.7779±0.08	0.7151±0.06	0.7275±0.06	**0.8577±0.04**	0.8248±0.03
	NNDSVD	0.6142±0.08	0.6517±0.02	0.7161±0.03	0.7404±0.03	0.7475±0.03	0.7730±0.01	0.7934±0.02
	FactorAnalysis	0.6233±0.07	0.6649±0.02	0.7532±0.05	0.7930±0.03	0.8125±0.01	0.8197±0.05	0.8164±0.03
	NMF (CD)	0.6228±0.07	0.6333±0.03	0.7213±0.06	**0.8141±0.05**	0.7898±0.06	0.7838±0.03	0.8098±0.04
	NMF (MU)	0.6540±0.04	0.6939±0.05	0.7551±0.04	0.7431±0.01	0.8241±0.03	0.8169±0.04	0.8248±0.02
	SNMF	0.7133±0.03	**0.8211±0.03**	0.7426±0.06	0.7783±0.02	0.8371±0.03	0.7947±0.02	0.7907±0.02
	CoxNMF	**0.7717±0.09**	0.7661±0.05	**0.8283±0.05**	0.8008±0.05	0.8182±0.02	0.8413±0.04	**0.8462±0.02**
Dice coefficient	TruncatedSVD	0.2293±0.17	0.1005±0.09	0.2897±0.24	0.3706±0.08	0.3234±0.16	0.2574±0.20	0.2174±0.16
	PCA	0.2385±0.10	0.0659±0.09	0.1562±0.06	0.1692±0.09	0.2253±0.05	0.1708±0.12	0.0875±0.09
	SparsePCA	0.2203±0.11	0.1911±0.10	0.1733±0.08	0.1313±0.06	0.0948±0.08	0.0886±0.09	0.0886±0.08
	NNDSVD	0.2269±0.16	0.2119±0.06	0.1991±0.14	0.2015±0.06	0.1316±0.12	0.1096±0.09	0.1493±0.07
	FactorAnalysis	0.1156±0.12	0.0631±0.09	0.2053±0.15	0.2573±0.07	0.2693±0.04	0.2146±0.21	0.1291±0.13
	NMF (CD)	0.1801±0.16	0.0799±0.08	0.2080±0.18	**0.4023±0.15**	0.2705±0.18	0.1688±0.11	0.2074±0.16
	NMF (MU)	0.2495±0.09	0.2185±0.08	0.2406±0.15	0.1434±0.08	0.3413±0.12	0.2118±0.18	0.1409±0.14
	SNMF	0.3709±0.05	**0.5069±0.07**	0.2443±0.18	0.2772±0.09	**0.3732±0.12**	0.2017±0.07	0.1084±0.08
	CoxNMF	**0.4538±0.18**	0.3565±0.18	**0.4700±0.17**	0.2883±0.17	0.3058±0.10	**0.2939±0.14**	**0.3155±0.12**

In the secondary multivariate simulation, we set $\tau_1 = \tau_2 = 2$. Initially all rows of the $\boldsymbol{H}$ are *i.i.d.* uniformly distributed over $\mathcal{U}(0, 1)$. The second row $\boldsymbol{H}_{k=2,\mathrm{i}}$ is then replaced by the difference of an array of values ranging from 0.5 to 0.698 (step size = 0.02) with $\frac{1}{10}\boldsymbol{H}_{k=1,\mathrm{i}}$. Finally, the 3$^{\mathrm{rd}}$ row $\boldsymbol{H}_{k=3,\mathrm{i}}$ is replaced by difference of an array of values from 0.698 to 0.5 (step size = -0.02) with $\frac{1}{10}\boldsymbol{H}_{k=4,\mathrm{i}}$. Thus, the first and second rows depend on each other and both lead to better prognosis, whereas the 3$^{\mathrm{rd}}$ and 4$^{\mathrm{th}}$ rows depend on each other and lead to adverse prognosis. A complete report of all secondary multivariate simulation results is shown in Table 4.10.

From the results it is observed that CoxNMF still outperforms the other baseline algorithms in most of the simulations. For example, CoxNMF achieved highest accuracy and Dice coefficient in 5 out of 7 K, and highest C-Index in 7 out of 7 K, among $K \in$ {6,7,8,9,10,11,12} and $\varepsilon = 0.10$, in both the secondary univariate and multivariate simulations. These results suggested that CoxNMF is still a preferred model for unveiling latent survival gene clusters, even when the artificially introduced survival information is at a different location of the ground truth $\boldsymbol{H}$.

Table 4.7. Simulation results with univariate underlying features setup among all combinations of $K \in \{6, 7, 8, 9, 10, 11, 12\}$ and $\varepsilon \in \{0, 0.05, 0.10\}$. Experiments repeat 5 times each with random seed $\in \{1, 2, 3, 4, 5\}$. Mean values $\pm$ standard deviations are reported, best performed mean values among models are highlighted in bold font.

		$\varepsilon = 0.00$						$\varepsilon = 0.05$						$\varepsilon = 0.10$					
K	Model	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime
6	TruncatedSVD	**6**	**0.0000±0.00**	**1.0000±0.00**	**0.9667±0.07**	**0.9000±0.22**	**0.1414±0.01**	5	0.0783±0.00	0.8526±0.17	0.6413±0.09	0.1666±0.20	**0.0846±0.02**	5	0.1232±0.01	0.8035±0.15	0.6083±0.08	0.0971±0.18	**0.1032±0.02**
	PCA	**6**	**0.0000±0.00**	**1.0000±0.00**	0.7847±0.08	0.3765±0.11	0.1430±0.01	4	0.1402±0.01	0.5669±0.02	0.7477±0.06	0.2438±0.12	0.1851±0.03	4	0.1661±0.01	0.5656±0.02	0.7390±0.07	0.2081±0.17	0.1763±0.03
	SparsePCA	**6**	0.0156±0.00	0.9551±0.01	0.7657±0.03	0.1560±0.14	119.9614±2.78	**6**	0.0608±0.00	0.9494±0.01	0.7447±0.02	0.0552±0.12	119.9969±6.02	**6**	0.1131±0.01	0.9362±0.01	0.7430±0.01	0.0000±0.00	120.8642±2.97
	NNDSVD	**6**	0.3145±0.01	0.8773±0.01	0.9000±0.09	0.7000±0.27	0.3854±0.09	4	0.3128±0.01	0.5573±0.02	0.6650±0.11	0.1995±0.28	0.3132±0.10	4	0.3149±0.01	0.5570±0.02	0.6583±0.11	0.2569±0.25	0.1488±0.00
	FactorAnalysis	**6**	**0.0000±0.00**	**1.0000±0.00**	0.8533±0.12	0.6370±0.20	0.1593±0.01	**6**	**0.0588±0.00**	0.9840±0.00	0.7893±0.05	0.3454±0.08	0.1241±0.02	5	0.1230±0.01	0.7999±0.16	0.6793±0.05	0.0469±0.10	0.1161±0.03
	NMF (CD)	**6**	0.0035±0.00	0.9932±0.01	0.8000±0.07	0.4000±0.22	0.1940±0.01	5	0.0783±0.00	0.8524±0.17	0.6730±0.14	0.2400±0.33	0.1433±0.01	5	0.1232±0.01	0.8036±0.15	0.6097±0.07	0.0872±0.17	0.1457±0.02
	NMF (MU)	**6**	0.0295±0.01	0.9391±0.05	0.8667±0.07	0.6000±0.22	0.1675±0.02	5	0.0807±0.01	0.8049±0.18	0.5920±0.03	0.0757±0.14	0.1472±0.02	5	0.1250±0.01	0.7639±0.16	0.6050±0.07	0.0933±0.16	0.1526±0.02
	SNMF	**6**	0.0004±0.00	0.9999±0.00	0.9000±0.09	0.7000±0.27	31.1359±3.82	5	0.0784±0.00	0.8595±0.15	0.7070±0.09	0.1798±0.24	35.8905±1.30	**6**	**0.1125±0.01**	0.9630±0.00	0.7717±0.03	**0.5217±0.12**	37.4042±2.63
	CoxNMF	**6**	0.0094±0.01	**1.0000±0.00**	0.9333±0.09	0.8000±0.27	5.3220±0.10	**6**	0.0616±0.00	**1.0000±0.00**	**0.8897±0.08**	**0.7349±0.18**	5.5224±0.27	**6**	0.1141±0.01	**0.9994±0.00**	**0.8080±0.00**	0.4883±0.11	5.4748±0.07
7	TruncatedSVD	**7**	**0.0000±0.00**	**1.0000±0.00**	0.9429±0.08	0.8000±0.27	0.1452±0.01	6	0.0747±0.00	0.8583±0.07	0.7183±0.08	0.2425±0.20	**0.0798±0.00**	6	0.1180±0.01	0.7729±0.10	0.7140±0.07	0.2363±0.19	**0.1003±0.02**
	PCA	6	0.0505±0.00	0.9894±0.00	0.7740±0.05	0.1878±0.14	**0.0836±0.00**	6	0.0746±0.00	0.8494±0.07	0.7814±0.06	0.2050±0.17	0.0815±0.00	5	0.1519±0.01	0.5596±0.03	0.7514±0.04	0.1475±0.15	0.2528±0.07
	SparsePCA	**7**	0.0150±0.00	0.9506±0.01	0.7951±0.03	0.1704±0.16	139.8523±4.29	8	0.0574±0.00	0.9549±0.01	0.8123±0.09	0.1248±0.19	145.2228±3.91	6	0.1185±0.01	0.7316±0.12	0.7849±0.05	0.1338±0.19	120.4231±2.11
	NNDSVD	**7**	0.3373±0.01	0.8819±0.02	0.9143±0.08	0.7000±0.27	0.4090±0.13	5	0.3328±0.01	0.5646±0.03	0.7194±0.01	0.0645±0.14	0.1621±0.00	5	0.3325±0.01	0.5640±0.03	0.7194±0.02	0.0631±0.14	0.1592±0.00
	FactorAnalysis	**7**	**0.0000±0.00**	**1.0000±0.00**	0.8946±0.04	0.5831±0.19	0.1672±0.01	**7**	**0.0563±0.00**	0.9861±0.00	**0.9017±0.06**	**0.6045±0.24**	0.1318±0.01	6	0.1181±0.01	0.7661±0.11	0.7531±0.09	0.2031±0.24	0.1313±0.01
	NMF (CD)	**7**	0.0024±0.00	0.9954±0.01	0.8571±0.00	0.5000±0.00	0.2118±0.01	6	0.0748±0.00	0.8577±0.07	0.7197±0.07	0.2427±0.20	0.1534±0.00	6	0.1182±0.01	0.7701±0.10	0.7440±0.10	0.3012±0.28	0.1657±0.02
	NMF (MU)	**7**	0.0276±0.02	0.8876±0.10	0.7714±0.08	0.2000±0.27	0.1643±0.01	6	0.0773±0.00	0.7722±0.10	0.7640±0.10	0.3061±0.32	0.1527±0.01	6	0.1196±0.01	0.7302±0.09	0.7606±0.10	0.3104±0.30	0.1562±0.02
	SNMF	**7**	0.0003±0.00	0.9990±0.00	0.8857±0.06	0.6000±0.22	32.6721±3.97	6	0.0748±0.00	0.8573±0.07	0.7754±0.08	0.1980±0.27	32.9375±4.06	6	0.1181±0.01	0.7726±0.10	0.7914±0.08	0.3150±0.31	35.7420±1.03
	CoxNMF	**7**	0.0057±0.00	**1.0000±0.00**	**0.9714±0.06**	**0.9000±0.22**	4.3984±0.20	8	0.0590±0.00	**1.0000±0.00**	0.8771±0.07	0.5411±0.14	4.3419±0.08	**7**	**0.1087±0.01**	**0.9999±0.00**	**0.8394±0.01**	**0.5623±0.10**	4.3185±0.07
8	TruncatedSVD	**8**	**0.0000±0.00**	**1.0000±0.00**	0.9250±0.07	0.7000±0.27	0.1448±0.01	7	0.0711±0.00	0.8399±0.07	0.7012±0.06	0.1011±0.17	0.0834±0.00	7	0.1144±0.00	0.7485±0.07	0.6890±0.06	0.1061±0.15	**0.1068±0.02**
	PCA	6	0.1025±0.01	0.5590±0.04	0.7490±0.11	0.1948±0.14	**0.0652±0.00**	6	0.1134±0.01	0.5594±0.04	0.7853±0.05	0.1660±0.16	**0.0655±0.00**	**8**	**0.1051±0.01**	0.9654±0.01	0.8360±0.03	0.2268±0.15	0.2298±0.11
	SparsePCA	7	0.0477±0.00	0.7700±0.12	0.7933±0.01	0.0000±0.00	138.8714±1.52	7	0.0721±0.00	0.7629±0.09	0.7170±0.08	0.1164±0.13	142.5608±9.15	7	0.1149±0.00	0.7204±0.10	0.7145±0.07	0.0994±0.10	139.2578±1.10
	NNDSVD	**8**	0.3547±0.01	0.8697±0.01	0.8750±0.00	0.5000±0.00	0.1926±0.01	6	0.3488±0.01	0.5564±0.03	0.7932±0.06	0.2823±0.28	0.5259±0.03	6	0.3468±0.01	0.5568±0.03	0.7878±0.05	0.2643±0.25	0.5157±0.01
	FactorAnalysis	**8**	**0.0000±0.00**	**1.0000±0.00**	0.9235±0.07	0.6746±0.27	0.1700±0.01	7	0.0716±0.00	0.8309±0.13	0.7550±0.01	0.0000±0.00	0.1090±0.01	7	0.1145±0.00	0.7556±0.09	0.7540±0.05	0.0608±0.14	0.1182±0.02
	NMF (CD)	**8**	0.0060±0.01	0.9747±0.04	0.8750±0.09	0.5000±0.35	0.2506±0.10	7	0.0712±0.00	0.8389±0.07	0.7068±0.06	0.0931±0.17	0.1676±0.01	7	0.1147±0.00	0.7484±0.07	0.6847±0.06	0.0957±0.15	0.1930±0.02
	NMF (MU)	**8**	0.0256±0.01	0.8955±0.05	0.8250±0.07	0.3000±0.27	0.1772±0.02	7	0.0759±0.00	0.6924±0.12	0.7220±0.06	0.0916±0.17	0.1571±0.01	7	0.1162±0.00	0.6862±0.07	0.7000±0.05	0.1056±0.15	0.1769±0.02
	SNMF	**8**	0.0003±0.00	0.9983±0.00	0.8750±0.00	0.5000±0.00	32.3563±5.42	7	0.0713±0.00	0.8440±0.07	0.7927±0.08	0.1970±0.27	42.8982±4.39	7	0.1147±0.00	0.7573±0.08	0.7637±0.07	0.1047±0.20	42.5548±3.44
	CoxNMF	**8**	0.0098±0.01	**1.0000±0.00**	**0.9500±0.07**	**0.8000±0.27**	4.3291±0.08	**8**	**0.0592±0.01**	**1.0000±0.00**	**0.8947±0.05**	**0.7002±0.14**	5.4263±0.29	**8**	0.1112±0.00	**0.9995±0.00**	**0.8645±0.01**	**0.6080±0.02**	4.2325±0.07
9	TruncatedSVD	**9**	**0.0000±0.00**	**1.0000±0.00**	**0.9556±0.06**	**0.8000±0.27**	0.1499±0.01	8	0.0692±0.00	0.8229±0.15	0.7938±0.09	0.3156±0.29	**0.0874±0.01**	8	0.1105±0.00	0.7616±0.16	0.7844±0.08	0.3088±0.26	**0.0981±0.02**
	PCA	8	0.0459±0.00	0.9855±0.01	0.8073±0.02	0.0845±0.12	**0.0919±0.01**	8	0.0691±0.00	0.8313±0.15	0.8109±0.02	0.0737±0.10	0.0960±0.01	7	0.1347±0.00	0.5713±0.05	0.8158±0.02	0.0832±0.11	0.2139±0.04
	SparsePCA	**9**	0.0138±0.00	0.9603±0.01	0.8102±0.07	0.1167±0.11	173.3591±6.95	**9**	0.0545±0.00	0.9522±0.01	0.8362±0.02	0.0975±0.13	174.3828±3.62	**9**	**0.1021±0.00**	0.9427±0.01	0.8449±0.01	0.1489±0.15	174.7893±4.25
	NNDSVD	**9**	0.3637±0.01	0.8745±0.02	0.9111±0.05	0.6000±0.22	0.5577±0.03	8	0.3538±0.01	0.6129±0.05	0.7913±0.06	0.1147±0.22	0.1856±0.00	7	0.3534±0.01	0.5767±0.02	0.7871±0.01	0.0041±0.01	0.5060±0.04
	FactorAnalysis	**9**	**0.0000±0.00**	**1.0000±0.00**	0.8991±0.07	0.5282±0.29	0.1777±0.01	**9**	**0.0528±0.00**	0.9855±0.00	**0.9104±0.06**	0.5716±0.24	0.2698±0.10	8	0.1106±0.00	0.7654±0.15	0.8009±0.05	0.1112±0.15	0.1233±0.02
	NMF (CD)	**9**	0.0028±0.00	0.9939±0.01	0.8889±0.08	0.5000±0.35	0.2294±0.01	8	0.0694±0.00	0.8251±0.15	0.7767±0.06	0.2416±0.19	0.1805±0.01	8	0.1109±0.00	0.7608±0.16	0.7658±0.04	0.1652±0.18	0.1792±0.02
	NMF (MU)	**9**	0.0373±0.01	0.8888±0.09	0.8667±0.09	0.4000±0.42	0.1731±0.02	8	0.0750±0.00	0.6880±0.16	0.8211±0.07	0.2815±0.25	0.1606±0.02	8	0.1148±0.00	0.6604±0.15	0.7944±0.07	0.1813±0.23	0.1601±0.02
	SNMF	**9**	0.0002±0.00	0.9997±0.00	0.8889±0.00	0.5000±0.00	39.8020±7.20	8	0.0693±0.00	0.8223±0.15	0.8344±0.07	0.2515±0.35	45.4949±4.14	8	0.1108±0.00	0.7613±0.16	0.7991±0.07	0.2210±0.31	47.4637±1.83
	CoxNMF	**9**	0.0215±0.01	**1.0000±0.00**	**0.9556±0.06**	**0.8000±0.27**	4.4475±0.16	**9**	0.0605±0.01	**0.9999±0.00**	0.9007±0.04	**0.6760±0.10**	4.3819±0.11	**9**	0.1041±0.01	**0.9996±0.00**	**0.8973±0.04**	**0.6612±0.11**	3.9993±0.88
10	TruncatedSVD	**10**	**0.0000±0.00**	**1.0000±0.00**	0.9200±0.04	0.6000±0.22	**0.1499±0.01**	9	0.0658±0.00	0.8779±0.05	0.7888±0.05	0.2255±0.19	**0.0933±0.02**	9	0.1088±0.00	0.7789±0.08	0.7834±0.05	0.2217±0.17	**0.1245±0.00**
	PCA	12	**0.0000±0.00**	**1.0000±0.00**	0.9016±0.03	0.1606±0.16	0.1627±0.01	12	**0.0518±0.00**	0.9858±0.00	0.8932±0.03	0.2285±0.08	0.1386±0.01	8	0.1263±0.00	0.5724±0.03	0.8470±0.02	0.1452±0.11	0.2698±0.08
	SparsePCA	**10**	0.0146±0.00	0.9447±0.01	0.8304±0.07	0.1861±0.17	194.6709±1.98	11	0.0543±0.00	0.9498±0.01	0.9012±0.01	0.0000±0.00	198.7960±7.38	11	**0.1017±0.00**	0.9419±0.01	**0.8674±0.08**	0.1291±0.20	204.1633±8.10
	NNDSVD	**10**	0.3777±0.02	0.8659±0.01	0.9200±0.04	0.6000±0.22	0.3671±0.08	8	0.3675±0.02	0.5680±0.03	0.8308±0.05	0.2750±0.27	0.3680±0.06	8	0.3643±0.02	0.5691±0.03	0.8322±0.05	0.2734±0.27	0.3625±0.04
	FactorAnalysis	**10**	**0.0000±0.00**	**1.0000±0.00**	0.9114±0.07	0.5745±0.29	0.1792±0.01	**10**	0.0530±0.00	0.9854±0.00	**0.9284±0.04**	0.5855±0.22	0.1376±0.01	9	0.1089±0.00	0.8119±0.08	0.8226±0.05	0.1604±0.15	0.1409±0.00
	NMF (CD)	**10**	0.0052±0.00	0.9792±0.01	0.8800±0.11	0.4000±0.55	0.3355±0.12	9	0.0660±0.00	0.8777±0.05	0.7958±0.05	0.2382±0.19	0.1918±0.00	9	0.1093±0.00	0.7781±0.08	0.7854±0.05	0.2241±0.18	0.2290±0.01
	NMF (MU)	**10**	0.0373±0.01	0.7887±0.08	0.8400±0.05	0.2000±0.27	0.1871±0.01	9	0.0712±0.00	0.7052±0.07	0.8014±0.03	0.1506±0.20	0.1764±0.02	9	0.1128±0.00	0.6696±0.07	0.8022±0.03	0.0776±0.17	0.1856±0.03
	SNMF	**10**	0.0003±0.00	0.9995±0.00	0.9200±0.04	0.6000±0.22	44.2806±4.22	9	0.0660±0.00	0.8779±0.05	0.8092±0.04	0.1565±0.21	53.4899±5.82	9	0.1092±0.00	0.7786±0.08	0.8286±0.05	0.2090±0.31	52.4478±4.76
	CoxNMF	**10**	0.0148±0.01	**1.0000±0.00**	**0.9600±0.05**	**0.8000±0.27**	3.8390±1.57	**10**	0.0558±0.00	**1.0000±0.00**	0.9192±0.06	**0.6769±0.23**	5.5532±0.18	**10**	0.1035±0.00	**0.9996±0.00**	0.8526±0.02	**0.3198±0.21**	4.5925±0.13
11	TruncatedSVD	**11**	**0.0000±0.00**	**1.0000±0.00**	0.9273±0.04	0.6000±0.22	0.1562±0.00	10	0.0671±0.00	0.8475±0.15	0.8455±0.01	0.3697±0.01	**0.1021±0.02**	10	0.1098±0.00	0.7808±0.15	0.8173±0.01	0.3281±0.01	**0.1143±0.03**
	PCA	10	0.0420±0.00	0.9845±0.00	0.8689±0.02	0.2362±0.06	**0.0986±0.01**	13	**0.0520±0.00**	0.9885±0.00	0.9120±0.04	0.3400±0.12	0.1410±0.00	10	**0.1096±0.00**	0.7957±0.14	0.8735±0.02	0.2441±0.05	0.1182±0.03
	SparsePCA	10	0.0439±0.00	0.8096±0.16	0.7033±0.04	0.1641±0.09	198.1901±5.38	10	0.0682±0.00	0.7891±0.15	0.7333±0.06	0.1133±0.10	192.5488±5.78	10	0.1103±0.00	0.7562±0.14	0.7260±0.07	0.1159±0.08	198.3129±3.85
	NNDSVD	**11**	0.3847±0.02	0.8643±0.01	0.9273±0.04	0.6000±0.22	0.5954±0.02	10	0.3746±0.01	0.6010±0.05	0.8371±0.04	0.0980±0.21	0.2094±0.00	9	0.3732±0.01	0.5758±0.03	0.8169±0.02	0.1195±0.16	0.4396±0.07
	FactorAnalysis	**11**	**0.0000±0.00**	**1.0000±0.00**	0.9202±0.04	0.5304±0.21	0.1899±0.00	**11**	0.0533±0.00	0.9886±0.00	**0.9353±0.05**	**0.6186±0.27**	0.1454±0.01	10	0.1099±0.00	0.7794±0.16	0.8633±0.01	0.2676±0.15	0.1326±0.03
	NMF (CD)	**11**	0.0040±0.00	0.9843±0.01	0.8727±0.05	0.3000±0.27	0.3172±0.10	10	0.0675±0.00	0.8383±0.17	0.8396±0.01	0.2937±0.16	0.2133±0.02	10	0.1105±0.00	0.7800±0.15	0.8295±0.01	0.2791±0.16	0.2279±0.03
	NMF (MU)	**11**	0.0395±0.01	0.7952±0.16	0.8545±0.05	0.2000±0.27	0.2151±0.03	10	0.0758±0.01	0.6580±0.07	0.8296±0.01	0.0768±0.16	0.1889±0.03	10	0.1151±0.00	0.6631±0.07	0.8380±0.01	0.1473±0.20	0.1838±0.01
	SNMF	**11**	0.0006±0.00	0.9994±0.00	0.9273±0.04	0.6000±0.22	51.9193±1.74	**11**	0.0537±0.00	0.9874±0.00	0.9091±0.03	0.5293±0.18	57.3671±3.01	10	0.1104±0.00	0.7846±0.14	0.8382±0.01	0.2217±0.19	58.6151±2.18
	CoxNMF	**11**	0.0105±0.00	**1.0000±0.00**	**1.0000±0.00**	**1.0000±0.00**	5.3712±0.68	**11**	0.0626±0.01	**1.0000±0.00**	0.8945±0.02	0.5758±0.11	4.6256±0.15	**11**	0.1118±0.01	**0.9995±0.00**	**0.8942±0.01**	**0.5926±0.02**	4.7213±0.19
12	TruncatedSVD	**12**	**0.0000±0.00**	**1.0000±0.00**	**0.9500±0.05**	**0.7000±0.27**	**0.1583±0.00**	11	0.0668±0.00	0.8059±0.20	0.8497±0.05	0.3539±0.23	0.1136±0.02	11	0.1105±0.01	0.7619±0.18	0.8297±0.04	**0.3172±0.20**	**0.1043±0.03**
	PCA	13	**0.0000±0.00**	**1.0000±0.00**	0.8822±0.09	0.1029±0.12	0.1648±0.00	11	0.0667±0.00	0.7937±0.21	0.8905±0.02	0.3005±0.13	**0.1109±0.02**	10	0.1250±0.01	0.5758±0.03	0.8633±0.02	0.1847±0.15	0.2237±0.03
	SparsePCA	**12**	0.0143±0.00	0.9581±0.01	0.8327±0.08	0.0225±0.05	228.7477±4.67	11	0.0679±0.00	0.7594±0.18	0.8380±0.08	0.0839±0.12	207.9066±5.07	11	0.1110±0.01	0.7444±0.16	0.7478±0.07	0.1632±0.05	211.4625±5.54
	NNDSVD	**12**	0.3953±0.01	0.8701±0.02	**0.9500±0.05**	**0.7000±0.27**	0.3910±0.05	11	0.3826±0.01	0.6433±0.06	0.8750±0.04	0.3705±0.24	0.2314±0.01	11	0.3788±0.01	0.6383±0.05	0.8712±0.04	0.2871±0.28	0.2299±0.01
	FactorAnalysis	**12**	**0.0000±0.00**	**1.0000±0.00**	0.9240±0.04	0.5137±0.23	0.1954±0.00	11	0.0673±0.00	0.8055±0.20	0.8768±0.05	0.2708±0.31	0.1568±0.01	11	0.1106±0.01	0.7724±0.18	**0.8722±0.05**	0.2516±0.27	0.1244±0.03
	NMF (CD)	**12**	0.0017±0.00	0.9954±0.01	**0.9500±0.05**	**0.7000±0.27**	0.2800±0.01	11	0.0672±0.00	0.8054±0.20	0.8418±0.03	0.2952±0.15	0.2517±0.03	11	0.1113±0.01	0.7635±0.18	0.8233±0.03	0.2690±0.14	0.2499±0.03
	NMF (MU)	**12**	0.0395±0.01	0.8493±0.11	0.8833±0.05	0.3000±0.27	0.2109±0.03	11	0.0843±0.01	0.6504±0.13	0.8400±0.01	0.0749±0.17	0.2102±0.03	11	0.1219±0.00	0.6525±0.11	0.8440±0.00	0.0716±0.15	0.2141±0.03
	SNMF	**12**	0.0005±0.00	0.9991±0.00	0.9333±0.04	0.6000±0.22	48.3502±3.68	11	0.0671±0.00	0.8060±0.20	0.8723±0.04	0.3683±0.26	61.8516±2.57	11	0.1113±0.01	0.7703±0.17	0.8473±0.01	0.1431±0.20	61.2557±6.35
	CoxNMF	**12**	0.0245±0.02	0.9999±0.00	0.9333±0.07	0.6000±0.42	4.7449±0.12	**12**	**0.0577±0.00**	**1.0000±0.00**	**0.9288±0.05**	**0.6584±0.20**	5.7699±0.14	**12**	**0.1055±0.01**	**0.9994±0.00**	0.8605±0.01	0.2033±0.19	5.8308±0.14

4.9.2 Human Cancer Gene Expression Results

To demonstrate the CoxNMF algorithm, in this thesis, we present ten cancer results. The detailed optimization results $\tilde{\boldsymbol{W}}$ and $\hat{\boldsymbol{H}}$ for those cancers are reported in Figure 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14, and 4.15.

Label estimations $\hat{\phi}_-$ which has negative β_{i} at column i of $\tilde{\boldsymbol{W}}$ are highlighted in blue rectangles. Similarly, label estimations $\hat{\phi}_+$ which has positive β_{i} at column i of $\tilde{\boldsymbol{W}}$ are highlighted in red rectangles. We assume that gene clusters $\hat{\phi}_-$ are associated with better survival outcomes, while gene clusters $\hat{\phi}_+$ are associated with worse survival outcomes.

Table 4.8. Simulation results with multivariate underlying features setup among all combinations of $K \in \{6, 7, 8, 9, 10, 11, 12\}$ and $\varepsilon \in \{0, 0.05, 0.10\}$. Experiments repeat 5 times each with random seed $\in \{1, 2, 3, 4, 5\}$. Mean values $\pm$ standard deviations are reported, best performed mean values among models are highlighted in bold font.

		$\varepsilon = 0.00$						$\varepsilon = 0.05$						$\varepsilon = 0.10$					
K	Model	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime
6	TruncatedSVD	**6**	**0.0000±0.00**	**1.0000±0.00**	0.7833±0.05	0.4800±0.11	**0.1409±0.01**	5	0.0778±0.01	0.9209±0.04	0.6208±0.09	0.2442±0.17	**0.0847±0.01**	5	0.1226±0.01	0.8462±0.07	0.6042±0.09	0.2293±0.17	0.1027±0.01
	PCA	**6**	**0.0000±0.00**	**1.0000±0.00**	0.7173±0.02	0.2207±0.06	0.1411±0.01	**6**	**0.0585±0.00**	0.9838±0.00	0.7257±0.03	0.2512±0.09	0.2774±0.10	**6**	**0.1116±0.01**	0.9638±0.00	0.7307±0.03	0.2385±0.10	**0.0851±0.01**
	SparsePCA	**6**	0.0158±0.00	0.9530±0.01	0.6480±0.02	0.1735±0.14	118.9442±2.46	**6**	0.0604±0.00	0.9490±0.01	0.6728±0.02	0.1700±0.14	120.0660±1.70	4	0.1888±0.01	0.5793±0.04	0.5813±0.06	0.2203±0.11	16.4808±4.48
	NNDSVD	**6**	0.3314±0.02	0.8780±0.01	0.7000±0.05	0.2800±0.11	0.5235±0.02	5	0.3238±0.02	0.5998±0.04	0.6208±0.05	0.1506±0.08	0.1504±0.00	4	0.3307±0.01	0.5574±0.03	0.6142±0.08	0.2269±0.16	0.2623±0.02
	FactorAnalysis	**6**	**0.0000±0.00**	**1.0000±0.00**	0.7078±0.05	0.2656±0.12	0.1578±0.01	**6**	**0.0585±0.00**	0.9839±0.00	0.7602±0.03	0.3854±0.08	0.2786±0.10	5	0.1224±0.01	0.8624±0.06	0.6233±0.07	0.1156±0.12	0.1204±0.00
	NMF (CD)	**6**	0.0012±0.00	0.9996±0.00	0.7500±0.06	0.4000±0.14	0.1983±0.01	5	0.0779±0.01	0.9212±0.04	0.6405±0.08	0.2187±0.19	0.1369±0.00	5	0.1227±0.01	0.8459±0.06	0.6228±0.07	0.1801±0.16	0.1569±0.02
	NMF (MU)	**6**	0.0218±0.01	0.9523±0.03	0.7167±0.07	0.3200±0.18	0.1504±0.01	5	0.0801±0.01	0.8837±0.04	0.6077±0.07	0.1477±0.15	0.1464±0.01	5	0.1236±0.01	0.8277±0.06	0.6540±0.04	0.2495±0.09	0.1565±0.01
	SNMF	**6**	0.0006±0.00	0.9988±0.00	**0.8333±0.06**	**0.6000±0.14**	24.6514±6.07	**6**	0.0587±0.00	0.9840±0.00	**0.8067±0.03**	**0.5377±0.08**	29.9696±3.20	**6**	0.1119±0.01	0.9650±0.01	0.7133±0.03	0.3709±0.05	36.7547±2.28
	CoxNMF	**6**	0.0109±0.01	**1.0000±0.00**	**0.8333±0.06**	**0.6000±0.14**	5.3217±0.19	**6**	0.0593±0.00	**1.0000±0.00**	0.7752±0.09	0.4528±0.22	5.3334±0.17	**6**	0.1127±0.01	**1.0000±0.00**	**0.7717±0.09**	**0.4538±0.18**	4.2085±0.10
7	TruncatedSVD	**7**	**0.0000±0.00**	**1.0000±0.00**	**0.8286±0.04**	**0.5200±0.11**	**0.1406±0.01**	6	0.0755±0.00	0.8585±0.07	0.6027±0.05	0.0857±0.08	**0.0790±0.00**	6	0.1161±0.00	0.7678±0.09	0.5951±0.06	0.1005±0.09	**0.1131±0.00**
	PCA	9	**0.0000±0.00**	**1.0000±0.00**	0.8273±0.05	0.1554±0.09	0.1561±0.01	5	0.1278±0.01	0.5685±0.02	0.7113±0.02	0.1077±0.05	0.3222±0.10	6	0.1160±0.00	0.7720±0.09	0.7031±0.03	0.0659±0.09	0.1156±0.00
	SparsePCA	**7**	0.0143±0.00	0.9611±0.00	0.6964±0.06	0.1644±0.06	141.3539±4.20	**7**	0.0560±0.00	0.9554±0.01	0.6851±0.03	0.1422±0.10	137.6613±1.85	**7**	0.1048±0.00	0.9455±0.00	0.6876±0.03	0.1911±0.10	137.8560±1.70
	NNDSVD	**7**	0.3418±0.01	0.8748±0.00	0.7571±0.04	0.3200±0.11	0.4648±0.10	6	0.3326±0.01	0.5876±0.03	0.6924±0.04	0.1857±0.12	0.1551±0.00	5	0.3350±0.01	0.5728±0.04	0.6517±0.02	0.2119±0.06	0.1561±0.00
	FactorAnalysis	**7**	**0.0000±0.00**	**1.0000±0.00**	0.7459±0.04	0.2532±0.07	0.1656±0.01	**7**	**0.0543±0.00**	0.9871±0.00	0.7814±0.02	0.3414±0.05	0.1311±0.01	6	0.1162±0.00	0.7890±0.08	0.6649±0.02	0.0631±0.09	0.1258±0.01
	NMF (CD)	**7**	0.0041±0.00	0.9918±0.01	0.7571±0.06	0.3200±0.18	0.2106±0.01	6	0.0756±0.00	0.8583±0.07	0.6350±0.04	0.0775±0.09	0.1503±0.00	6	0.1163±0.00	0.7656±0.09	0.6333±0.03	0.0799±0.08	0.1708±0.02
	NMF (MU)	**7**	0.0354±0.01	0.8886±0.04	0.7429±0.04	0.2800±0.11	0.1752±0.03	6	0.0793±0.00	0.7653±0.07	0.6761±0.00	0.1807±0.00	0.1570±0.01	6	0.1187±0.00	0.7028±0.07	0.6939±0.05	0.2185±0.08	0.1643±0.02
	SNMF	**7**	0.0001±0.00	0.9999±0.00	0.8143±0.06	0.4800±0.18	23.3543±9.48	**7**	0.0545±0.00	0.9870±0.00	0.8043±0.04	0.4590±0.11	31.9815±2.43	**7**	**0.1043±0.00**	0.9695±0.00	**0.8211±0.03**	**0.5069±0.07**	34.3212±2.88
	CoxNMF	**7**	0.0058±0.00	0.9999±0.00	0.7714±0.06	0.3600±0.17	5.6157±0.23	**7**	0.0550±0.00	**1.0000±0.00**	**0.8079±0.04**	**0.4596±0.12**	4.2549±0.14	**7**	0.1051±0.00	**0.9994±0.00**	0.7661±0.05	0.3565±0.18	5.3210±0.12
8	TruncatedSVD	**8**	**0.0000±0.00**	**1.0000±0.00**	0.7875±0.03	0.3200±0.11	**0.1515±0.01**	7	0.0698±0.00	0.8642±0.13	0.7230±0.09	0.2899±0.25	**0.0954±0.02**	7	0.1123±0.00	0.7984±0.13	0.7184±0.10	0.2897±0.24	**0.0973±0.02**
	PCA	10	**0.0000±0.00**	**1.0000±0.00**	0.8602±0.01	0.1750±0.08	0.1607±0.01	7	0.0697±0.00	0.8690±0.11	0.7374±0.06	0.1377±0.06	0.0968±0.02	7	0.1121±0.00	0.8005±0.12	0.7390±0.07	0.1562±0.06	0.1064±0.02
	SparsePCA	**8**	0.0146±0.00	0.9473±0.02	0.7394±0.02	0.0657±0.10	160.3633±12.83	**9**	**0.0550±0.00**	0.9563±0.01	0.7752±0.03	0.1430±0.15	160.6093±2.02	9	**0.1033±0.00**	0.9467±0.01	0.7779±0.08	0.1733±0.08	162.0228±1.15
	NNDSVD	**8**	0.3569±0.02	0.8686±0.01	0.8000±0.03	0.3600±0.09	0.1953±0.01	6	0.3503±0.02	0.5682±0.03	0.7119±0.03	0.1964±0.13	0.3752±0.07	6	0.3489±0.02	0.5654±0.03	0.7161±0.03	0.1991±0.14	0.5308±0.01
	FactorAnalysis	**8**	**0.0000±0.00**	**1.0000±0.00**	0.7943±0.02	0.2728±0.10	0.1778±0.01	7	0.0703±0.00	0.8612±0.15	0.7398±0.06	0.1671±0.18	0.1055±0.02	7	0.1123±0.00	0.8040±0.12	0.7532±0.05	0.2053±0.15	0.1165±0.02
	NMF (CD)	**8**	0.0031±0.00	0.9913±0.01	0.7875±0.07	0.3200±0.23	0.2453±0.07	7	0.0699±0.00	0.8636±0.13	0.7396±0.07	0.2473±0.20	0.1769±0.02	7	0.1125±0.00	0.7973±0.13	0.7213±0.06	0.2080±0.18	0.1799±0.02
	NMF (MU)	**8**	0.0344±0.01	0.8160±0.12	0.7250±0.03	0.1200±0.11	0.1646±0.02	7	0.0764±0.01	0.7074±0.10	0.7305±0.05	0.1839±0.18	0.1825±0.03	7	0.1174±0.01	0.6967±0.08	0.7551±0.04	0.2406±0.15	0.1702±0.02
	SNMF	**8**	0.0001±0.00	**1.0000±0.00**	0.8250±0.07	0.4400±0.22	31.7027±8.34	7	0.0699±0.00	0.8635±0.13	0.7288±0.10	0.2168±0.29	31.9221±1.40	7	0.1125±0.00	0.7978±0.13	0.7426±0.06	0.2443±0.18	34.9501±3.68
	CoxNMF	**8**	0.0087±0.01	**1.0000±0.00**	**0.8625±0.05**	**0.5600±0.17**	5.3822±0.86	**8**	**0.0550±0.00**	**1.0000±0.00**	**0.8374±0.05**	**0.4795±0.17**	4.3600±0.07	**8**	0.1044±0.00	**0.9998±0.00**	**0.8283±0.05**	**0.4700±0.17**	5.3985±0.07
9	TruncatedSVD	**9**	**0.0000±0.00**	**1.0000±0.00**	0.8000±0.03	0.2800±0.11	**0.1498±0.01**	8	0.0688±0.00	0.8876±0.11	0.7737±0.02	0.3561±0.06	**0.0967±0.02**	8	0.1121±0.00	0.8288±0.13	0.7681±0.03	0.3706±0.08	**0.1119±0.01**
	PCA	11	**0.0000±0.00**	**1.0000±0.00**	0.8673±0.01	0.0412±0.05	0.1591±0.01	10	**0.0535±0.00**	0.9863±0.00	0.8546±0.00	0.1545±0.02	0.1345±0.01	7	0.1381±0.01	0.5863±0.04	0.7858±0.03	0.1692±0.09	0.2304±0.03
	SparsePCA	**9**	0.0147±0.00	0.9554±0.01	0.7694±0.06	0.2024±0.06	173.9682±5.85	10	0.0554±0.00	0.9534±0.01	0.7919±0.05	0.0972±0.09	177.8056±2.69	8	0.1125±0.00	0.7869±0.15	0.7151±0.06	0.1313±0.06	157.3929±3.77
	NNDSVD	**9**	0.3659±0.01	0.8764±0.01	0.8111±0.03	0.3200±0.11	0.5066±0.09	7	0.3587±0.01	0.5790±0.02	0.7424±0.03	0.2052±0.07	0.3862±0.06	7	0.3570±0.01	0.5798±0.02	0.7404±0.03	0.2015±0.06	0.1820±0.01
	FactorAnalysis	**9**	**0.0000±0.00**	**1.0000±0.00**	0.7993±0.04	0.2443±0.10	0.1784±0.01	8	0.0691±0.00	0.8569±0.14	0.7769±0.04	0.1862±0.13	0.1255±0.01	8	0.1120±0.00	0.8099±0.15	0.7930±0.03	0.2573±0.07	0.1229±0.02
	NMF (CD)	**9**	0.0054±0.00	0.9846±0.02	0.8222±0.02	0.3600±0.09	0.2723±0.07	8	0.0690±0.00	0.8869±0.11	0.7983±0.04	0.3494±0.12	0.1785±0.02	8	0.1124±0.00	0.8284±0.13	**0.8141±0.05**	**0.4023±0.15**	0.1967±0.01
	NMF (MU)	**9**	0.0394±0.01	0.8668±0.06	0.8000±0.05	0.2800±0.18	0.1958±0.02	8	0.0786±0.01	0.7508±0.15	0.7502±0.03	0.1431±0.15	0.1714±0.02	8	0.1178±0.01	0.7324±0.15	0.7431±0.01	0.1434±0.08	0.1663±0.03
	SNMF	**9**	0.0005±0.00	0.9986±0.00	0.8667±0.06	0.5200±0.23	37.5492±6.23	**9**	0.0544±0.00	0.9863±0.00	**0.8563±0.05**	**0.4894±0.18**	46.9898±5.30	8	0.1124±0.00	0.8282±0.13	0.7783±0.02	0.2772±0.09	41.6692±7.59
	CoxNMF	**9**	0.0139±0.01	0.9999±0.00	**0.8778±0.05**	**0.5600±0.17**	5.1952±0.80	**9**	0.0553±0.00	**0.9998±0.00**	0.8358±0.05	0.4158±0.17	5.4012±0.06	**9**	**0.1053±0.00**	**0.9994±0.00**	0.8008±0.05	0.2883±0.17	4.4741±0.11
10	TruncatedSVD	**10**	**0.0000±0.00**	**1.0000±0.00**	0.8500±0.04	0.4000±0.14	**0.1507±0.01**	9	0.0703±0.00	0.8599±0.16	0.7895±0.02	0.3498±0.06	0.0902±0.02	9	0.1146±0.01	0.8092±0.13	0.7694±0.06	0.3234±0.16	0.1040±0.03
	PCA	12	**0.0000±0.00**	**1.0000±0.00**	0.8743±0.01	0.0853±0.08	0.1608±0.01	8	0.1035±0.01	0.5843±0.02	0.7951±0.01	0.0707±0.04	**0.0875±0.04**	11	**0.1052±0.01**	0.9683±0.00	**0.8688±0.01**	0.2253±0.05	**0.0981±0.00**
	SparsePCA	**10**	0.0149±0.00	0.9528±0.01	0.7220±0.04	0.1579±0.05	195.7634±5.99	**10**	0.0574±0.00	0.9545±0.01	0.7361±0.04	0.1115±0.08	192.7114±4.63	9	0.1151±0.01	0.7946±0.12	0.7275±0.06	0.0948±0.08	176.7760±5.09
	NNDSVD	**10**	0.3788±0.01	0.8742±0.01	0.8300±0.04	0.3200±0.18	0.3687±0.09	8	0.3694±0.01	0.5829±0.03	0.7580±0.04	0.1657±0.15	0.2411±0.10	8	0.3666±0.00	0.5826±0.03	0.7475±0.03	0.1316±0.12	0.1944±0.00
	FactorAnalysis	**10**	**0.0000±0.00**	**1.0000±0.00**	0.8170±0.02	0.1939±0.06	0.1791±0.01	**10**	**0.0557±0.00**	0.9859±0.00	0.8481±0.03	0.3387±0.10	0.1397±0.01	9	0.1147±0.01	0.8144±0.14	0.8125±0.01	0.2693±0.04	0.1263±0.02
	NMF (CD)	**10**	0.0039±0.00	0.9851±0.02	0.8200±0.04	0.2800±0.18	0.2518±0.01	9	0.0705±0.00	0.8592±0.16	0.7754±0.02	0.2076±0.07	0.1988±0.02	9	0.1151±0.01	0.8101±0.13	0.7898±0.06	0.2705±0.18	0.2087±0.03
	NMF (MU)	**10**	0.0353±0.01	0.8602±0.07	0.7700±0.03	0.0800±0.11	0.2066±0.02	9	0.0760±0.00	0.7732±0.13	0.8114±0.03	0.2926±0.10	0.1860±0.02	9	0.1177±0.01	0.7317±0.10	0.8241±0.03	0.3413±0.12	0.1848±0.02
	SNMF	**10**	0.0007±0.00	0.9990±0.00	0.8300±0.04	0.3200±0.18	45.1362±7.52	**10**	0.0561±0.00	0.9853±0.00	0.8325±0.02	0.3424±0.08	53.0563±4.25	**10**	0.1072±0.01	0.9676±0.00	0.8371±0.03	**0.3732±0.12**	53.1634±2.04
	CoxNMF	**10**	0.0252±0.03	0.9996±0.00	**0.9000±0.00**	**0.6000±0.00**	4.1739±0.86	**10**	0.0590±0.01	**0.9998±0.00**	**0.9016±0.04**	**0.6171±0.12**	5.6865±0.17	**10**	0.1087±0.01	**0.9991±0.00**	0.8182±0.02	0.3058±0.10	5.5462±0.16
11	TruncatedSVD	**11**	**0.0000±0.00**	**1.0000±0.00**	0.8455±0.02	0.3200±0.11	**0.1540±0.01**	10	0.0691±0.00	0.8739±0.10	0.7820±0.06	0.2620±0.22	**0.1001±0.02**	10	0.1139±0.01	0.8020±0.14	0.7666±0.07	0.2574±0.20	**0.1174±0.02**
	PCA	12	**0.0000±0.00**	**1.0000±0.00**	**0.8745±0.01**	0.1242±0.11	0.1601±0.00	12	**0.0550±0.00**	0.9867±0.00	**0.8793±0.01**	0.1490±0.03	0.1319±0.01	10	0.1138±0.01	0.7897±0.14	0.8215±0.03	0.1708±0.12	0.1177±0.02
	SparsePCA	10	0.0439±0.00	0.8133±0.15	0.7435±0.05	0.1171±0.07	201.5104±7.20	10	0.0701±0.00	0.7935±0.16	0.7514±0.06	0.1583±0.03	191.2154±5.97	12	**0.1065±0.01**	0.9469±0.01	**0.8577±0.04**	0.0886±0.09	218.9479±5.73
	NNDSVD	**11**	0.3940±0.01	0.8717±0.01	0.8545±0.02	0.3600±0.09	0.3531±0.01	9	0.3832±0.01	0.5795±0.03	0.7724±0.01	0.0999±0.09	0.2110±0.01	9	0.3793±0.01	0.5822±0.03	0.7730±0.01	0.1096±0.09	0.2106±0.01
	FactorAnalysis	**11**	**0.0000±0.00**	**1.0000±0.00**	0.8409±0.01	0.2561±0.08	0.1834±0.01	10	0.0696±0.00	0.8706±0.11	0.8148±0.05	0.2007±0.21	0.1310±0.02	10	0.1140±0.01	0.7949±0.14	0.8197±0.05	0.2146±0.21	0.1355±0.02
	NMF (CD)	**11**	0.0031±0.00	0.9926±0.01	0.8636±0.05	0.4000±0.20	0.2750±0.01	10	0.0693±0.00	0.8737±0.10	0.7869±0.03	0.1738±0.12	0.2208±0.01	10	0.1146±0.01	0.8030±0.14	0.7838±0.03	0.1688±0.11	0.2365±0.02
	NMF (MU)	**11**	0.0357±0.00	0.8345±0.10	0.7818±0.02	0.0400±0.09	0.2196±0.03	10	0.0803±0.01	0.7260±0.14	0.7925±0.03	0.1447±0.14	0.1981±0.02	10	0.1205±0.01	0.7103±0.14	0.8169±0.04	0.2118±0.18	0.2002±0.03
	SNMF	**11**	0.0005±0.00	0.9996±0.00	0.8636±0.03	0.4000±0.14	47.4345±4.82	**11**	0.0562±0.00	0.9842±0.00	0.8654±0.02	**0.4164±0.08**	57.3563±3.78	10	0.1146±0.01	0.8027±0.14	0.7947±0.02	0.2017±0.07	52.8029±6.38
	CoxNMF	**11**	0.0247±0.01	0.9997±0.00	0.8727±0.06	**0.4400±0.26**	4.6407±0.13	**11**	0.0572±0.01	**0.9998±0.00**	0.8507±0.02	0.3462±0.10	4.8930±1.74	**11**	0.1085±0.01	**0.9980±0.00**	0.8413±0.04	**0.2939±0.14**	4.7068±0.11
12	TruncatedSVD	**12**	**0.0000±0.00**	**1.0000±0.00**	0.8667±0.03	**0.3600±0.17**	0.1562±0.01	11	0.0680±0.00	0.9113±0.04	0.7890±0.05	0.2279±0.19	**0.0998±0.02**	11	0.1138±0.01	0.8368±0.07	0.7759±0.04	0.2174±0.16	**0.1273±0.00**
	PCA	11	0.0391±0.00	0.9834±0.00	0.8263±0.02	0.1283±0.09	**0.0979±0.01**	11	0.0679±0.00	0.9130±0.04	0.8248±0.01	0.1130±0.09	0.1086±0.02	11	0.1137±0.01	0.8463±0.06	0.8209±0.01	0.0875±0.09	0.1303±0.01
	SparsePCA	11	0.0415±0.00	0.8724±0.05	0.7741±0.06	0.1324±0.06	210.8450±4.94	**12**	0.0578±0.00	0.9497±0.01	0.8264±0.01	0.0246±0.05	227.0586±4.65	**12**	**0.1079±0.01**	0.9408±0.01	0.8248±0.03	0.0886±0.08	233.4639±10.97
	NNDSVD	**12**	0.4024±0.01	0.8693±0.01	0.8667±0.03	**0.3600±0.17**	0.5454±0.10	11	0.3895±0.01	0.6351±0.04	0.7889±0.02	0.1179±0.08	0.2286±0.00	11	0.3846±0.01	0.6335±0.04	0.7934±0.02	0.1493±0.07	0.2358±0.01
	FactorAnalysis	**12**	**0.0000±0.00**	**1.0000±0.00**	0.8507±0.03	0.2380±0.13	0.1935±0.01	11	0.0684±0.00	0.8992±0.07	0.8201±0.04	0.1525±0.16	0.1416±0.03	11	0.1139±0.01	0.8417±0.07	0.8164±0.03	0.1291±0.13	0.1513±0.01
	NMF (CD)	**12**	0.0061±0.01	0.9895±0.01	0.8417±0.03	0.2400±0.17	0.2810±0.01	11	0.0684±0.00	0.9080±0.05	0.8071±0.04	0.1756±0.17	0.2417±0.02	11	0.1147±0.01	0.8279±0.08	0.8098±0.04	0.2074±0.16	0.2761±0.01
	NMF (MU)	**12**	0.0358±0.01	0.8093±0.06	0.8167±0.02	0.1200±0.11	0.2007±0.02	11	0.0801±0.01	0.6430±0.04	0.8214±0.02	0.1468±0.08	0.1944±0.00	11	0.1228±0.01	0.6472±0.05	0.8248±0.02	0.1409±0.14	0.2021±0.01
	SNMF	**12**	0.0004±0.00	0.9996±0.00	0.8583±0.02	0.3200±0.11	50.1545±4.19	11	0.0684±0.00	0.9089±0.05	0.8148±0.04	0.2070±0.18	55.9900±4.78	11	0.1147±0.01	0.8271±0.08	0.7907±0.02	0.1084±0.08	58.1586±4.51
	CoxNMF	14	0.0079±0.00	**1.0000±0.00**	**0.9002±0.03**	0.2783±0.20	4.8752±0.09	**12**	**0.0574±0.00**	**0.9998±0.00**	**0.8412±0.02**	**0.2612±0.14**	5.8957±0.16	**12**	0.1100±0.01	**0.9994±0.00**	**0.8462±0.02**	**0.3155±0.12**	4.9122±0.25

In addition, Spearman's rank correlation coefficient matrices of original data $\boldsymbol{X}$ are also reported and organized by the associated clusters that concord/discord with the survival, suggest the effectiveness of low-rank reorganization in helping to find survival associated clusters which is hard to exploit from the original data $\boldsymbol{X}$.

By performing gene ontology (GO) analysis with ToppGene analysis suite [182], certain GO biological process terms are discovered. The measurement of the enrichment results is based on P-value using the hypergeometric distribution [253]. A smaller P-value indicates a more significant association of gene cluster to a particular GO term. GO terms are sorted in ascending order based on P-value. Top GO terms and other important GO terms are

Table 4.9. Simulation results in secondary univariate simulation setup among all combinations of $K \in \{6, 7, 8, 9, 10, 11, 12\}$ and $\varepsilon \in \{0, 0.05, 0.10\}$. Experiments repeat 5 times each with random seed $\in \{1, 2, 3, 4, 5\}$. Mean values $\pm$ standard deviations are reported, best performed mean values among models are highlighted in bold font.

		$\varepsilon = 0.00$						$\varepsilon = 0.05$						$\varepsilon = 0.10$					
K	Model	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime
6	TruncatedSVD	**6**	**0.0000±0.00**	**1.0000±0.00**	0.8333±0.12	0.5000±0.35	0.1417±0.00	5.0	0.0774±0.00	0.8776±0.03	0.7163±0.04	0.2359±0.22	**0.0775±0.00**	5.0	0.1228±0.01	0.7733±0.04	0.7123±0.03	0.2277±0.21	**0.1065±0.02**
	PCA	**6**	**0.0000±0.00**	**1.0000±0.00**	0.7350±0.09	0.1228±0.12	0.1439±0.01	**6**	**0.0590±0.00**	0.9837±0.00	0.6873±0.12	0.2520±0.04	0.1208±0.01	**6**	**0.1126±0.01**	0.9617±0.01	**0.7743±0.03**	0.2240±0.03	0.1870±0.08
	SparsePCA	**6**	0.0158±0.00	0.9493±0.01	0.7080±0.10	0.1847±0.13	119.2778±1.77	5.0	0.0782±0.00	0.8459±0.02	0.6273±0.02	0.2015±0.07	101.0216±2.33	5.0	0.1232±0.01	0.7824±0.04	0.7503±0.11	0.3109±0.22	103.3369±3.87
	NNDSVD	**6**	0.3238±0.00	0.8828±0.01	0.7333±0.09	0.2000±0.27	0.3236±0.10	5.0	0.3158±0.00	0.5762±0.03	0.7053±0.07	0.1096±0.21	0.1428±0.00	4.0	0.3223±0.00	0.5619±0.02	0.6983±0.08	0.2246±0.21	0.2807±0.04
	FactorAnalysis	**6**	**0.0000±0.00**	**1.0000±0.00**	0.7637±0.04	0.2964±0.17	0.1571±0.00	**6**	0.0591±0.00	0.9835±0.00	0.7573±0.04	0.3020±0.18	0.3128±0.09	5.0	0.1227±0.01	0.8079±0.04	0.7433±0.07	0.2492±0.17	0.1205±0.01
	NMF (CD)	**6**	0.0024±0.00	0.9968±0.01	0.7667±0.09	0.3000±0.27	0.1825±0.01	5.0	0.0776±0.00	0.8769±0.02	0.7157±0.03	0.2337±0.21	0.1306±0.00	5.0	0.1233±0.01	0.7697±0.04	0.7137±0.02	0.2278±0.21	0.1582±0.01
	NMF (MU)	**6**	0.0303±0.01	0.9159±0.04	0.7000±0.07	0.1000±0.22	**0.1378±0.00**	5.0	0.0792±0.00	0.8303±0.05	0.7890±0.07	**0.4827±0.16**	0.1383±0.01	5.0	0.1239±0.01	0.7507±0.03	0.7510±0.05	**0.4026±0.05**	0.1583±0.01
	SNMF	**6**	0.0001±0.00	**1.0000±0.00**	**0.8667±0.07**	**0.6000±0.22**	18.0335±5.68	5.0	0.0775±0.00	0.8777±0.03	0.6773±0.00	0.0000±0.00	22.4225±2.80	5.0	0.1229±0.01	0.7732±0.04	0.6930±0.03	0.0790±0.17	26.9871±3.11
	CoxNMF	**6**	0.0091±0.01	**1.0000±0.00**	0.8000±0.07	0.4000±0.22	4.2283±0.14	7.0	0.0623±0.01	**1.0000±0.00**	**0.8583±0.06**	0.4536±0.26	5.2925±0.52	**6**	0.1783±0.14	**0.9997±0.00**	0.7657±0.04	0.3616±0.06	3.9821±0.71
7	TruncatedSVD	**7**	**0.0000±0.00**	**1.0000±0.00**	0.8286±0.06	0.4000±0.22	**0.1446±0.01**	6.0	0.0748±0.01	0.7326±0.12	0.7394±0.03	0.1520±0.21	0.0939±0.01	6.0	0.1186±0.01	0.6493±0.10	0.7603±0.06	0.2397±0.22	0.0983±0.03
	PCA	**7**	**0.0000±0.00**	**1.0000±0.00**	0.7409±0.10	0.2113±0.12	0.1461±0.01	5.0	0.1234±0.00	0.5625±0.03	0.7654±0.02	0.1486±0.09	**0.0674±0.00**	5.0	0.1515±0.01	0.5623±0.03	0.7729±0.04	0.2072±0.18	**0.0702±0.00**
	SparsePCA	**7**	0.0155±0.00	0.9460±0.01	0.8103±0.04	0.1467±0.20	138.0505±2.56	**7**	0.0585±0.00	0.9448±0.01	0.7946±0.04	0.0747±0.17	142.1372±4.00	**7**	**0.1093±0.01**	0.9274±0.02	0.7483±0.05	0.0330±0.07	144.8767±5.81
	NNDSVD	**7**	0.3390±0.02	0.8776±0.01	0.8286±0.06	0.4000±0.22	0.1838±0.00	5.0	0.3340±0.02	0.5659±0.03	0.7103±0.01	0.1302±0.17	0.5367±0.02	5.0	0.3340±0.01	0.5674±0.03	0.7146±0.02	0.1317±0.17	0.5653±0.03
	FactorAnalysis	**7**	**0.0000±0.00**	**1.0000±0.00**	0.7849±0.05	0.2237±0.22	0.1644±0.01	**7**	**0.0567±0.00**	0.9838±0.00	0.7991±0.04	0.3894±0.06	0.1360±0.01	6.0	0.1187±0.01	0.6705±0.11	0.7309±0.03	0.0785±0.18	0.1131±0.03
	NMF (CD)	**7**	0.0018±0.00	0.9980±0.00	**0.8571±0.10**	**0.5000±0.35**	0.2046±0.01	6.0	0.0749±0.01	0.7338±0.11	0.7963±0.06	0.3490±0.20	0.1500±0.00	6.0	0.1188±0.01	0.6480±0.10	0.7934±0.08	0.3127±0.31	0.1556±0.02
	NMF (MU)	**7**	0.0276±0.02	0.8754±0.11	0.8000±0.13	0.3000±0.45	0.1565±0.02	6.0	0.0786±0.01	0.6602±0.08	0.7286±0.02	0.0758±0.17	0.1478±0.01	6.0	0.1204±0.01	0.6372±0.06	0.7266±0.01	0.0760±0.16	0.1496±0.02
	SNMF	**7**	0.0006±0.00	0.9986±0.00	**0.8571±0.00**	**0.5000±0.00**	29.0828±3.37	6.0	0.0749±0.01	0.7333±0.11	0.7749±0.07	0.1965±0.27	30.1428±2.89	6.0	0.1188±0.01	0.6545±0.09	0.7526±0.06	0.0957±0.20	35.8031±1.47
	CoxNMF	**7**	0.0057±0.00	0.9995±0.00	0.8286±0.12	0.4000±0.42	5.4608±0.06	**7**	0.0590±0.00	**0.9999±0.00**	**0.8346±0.06**	**0.4791±0.16**	5.7804±0.20	**7**	0.1097±0.01	**0.9998±0.00**	**0.8134±0.06**	**0.3875±0.24**	5.9892±0.23
8	TruncatedSVD	**8**	**0.0000±0.00**	**1.0000±0.00**	0.8500±0.06	0.4000±0.22	**0.1540±0.00**	7.0	0.0707±0.00	0.8318±0.15	0.7820±0.02	0.2288±0.21	**0.0992±0.02**	7.0	0.1144±0.00	0.7684±0.13	0.7732±0.01	0.2188±0.20	0.1031±0.03
	PCA	6.0	0.0983±0.01	0.5676±0.04	0.8165±0.04	0.2067±0.15	0.4319±0.03	7.0	0.0706±0.00	0.8356±0.14	0.8125±0.01	0.1288±0.12	0.1014±0.02	6.0	0.1405±0.01	0.5676±0.04	0.8170±0.04	0.2008±0.15	**0.0693±0.00**
	SparsePCA	**8**	0.0151±0.00	0.9448±0.03	0.8265±0.03	0.1117±0.16	158.8928±1.49	**8**	0.0570±0.00	0.9439±0.02	0.8205±0.05	0.1084±0.24	155.1393±3.18	7.0	0.1149±0.00	0.7416±0.13	0.7988±0.01	0.0000±0.00	144.1265±4.91
	NNDSVD	**8**	0.3632±0.01	0.8781±0.01	0.8250±0.07	0.3000±0.27	0.4970±0.11	6.0	0.3543±0.01	0.5622±0.04	0.7520±0.01	0.1294±0.18	0.5093±0.02	6.0	0.3517±0.01	0.5628±0.04	0.7730±0.03	0.1125±0.25	0.3313±0.08
	FactorAnalysis	**8**	**0.0000±0.00**	**1.0000±0.00**	0.7815±0.05	0.1506±0.16	0.1819±0.01	**8**	**0.0553±0.00**	0.9845±0.01	**0.8678±0.07**	**0.4822±0.33**	0.1433±0.01	7.0	0.1145±0.00	0.7768±0.13	0.8135±0.04	0.2709±0.16	0.1238±0.03
	NMF (CD)	**8**	0.0053±0.00	0.9814±0.02	0.8750±0.00	0.5000±0.00	0.2558±0.11	7.0	0.0708±0.00	0.8304±0.14	0.8045±0.07	0.3001±0.31	0.1639±0.02	7.0	0.1147±0.00	0.7650±0.13	0.8232±0.08	0.3619±0.36	0.1708±0.02
	NMF (MU)	**8**	0.0256±0.01	0.8776±0.08	0.8000±0.11	0.2000±0.45	0.1903±0.02	7.0	0.0733±0.00	0.7689±0.19	0.7980±0.05	0.2542±0.23	0.1510±0.01	7.0	0.1162±0.00	0.7347±0.15	0.8177±0.05	0.3271±0.18	0.1599±0.02
	SNMF	**8**	0.0002±0.00	0.9999±0.00	0.8750±0.09	0.5000±0.35	29.2263±5.00	7.0	0.0708±0.00	0.8319±0.14	0.7945±0.05	0.2484±0.23	34.3897±1.43	7.0	0.1146±0.00	0.7661±0.13	0.7783±0.02	0.2198±0.20	37.2294±4.14
	CoxNMF	**8**	0.0107±0.01	**1.0000±0.00**	**0.9000±0.06**	**0.6000±0.22**	4.6612±0.04	**8**	0.0562±0.00	**0.9996±0.00**	0.8212±0.06	0.2745±0.25	4.5423±0.15	**8**	**0.1069±0.00**	**0.9993±0.00**	**0.8627±0.06**	**0.4689±0.23**	4.5471±0.25
9	TruncatedSVD	**9**	**0.0000±0.00**	**1.0000±0.00**	0.8889±0.08	0.5000±0.35	**0.1573±0.01**	8.0	0.0682±0.00	0.7268±0.17	0.8296±0.05	0.3736±0.26	0.0856±0.02	8.0	0.1102±0.00	0.6761±0.13	0.8167±0.05	0.3502±0.24	**0.0851±0.03**
	PCA	10.0	**0.0000±0.00**	**1.0000±0.00**	0.8136±0.13	0.0865±0.12	0.1653±0.01	7.0	0.1006±0.00	0.5699±0.04	0.8244±0.02	0.1075±0.10	**0.0682±0.01**	7.0	0.1315±0.00	0.5699±0.04	0.8273±0.02	0.1080±0.10	0.2357±0.05
	SparsePCA	**9**	0.0147±0.00	0.9504±0.01	0.7424±0.08	0.1239±0.12	177.8470±9.40	**9**	0.0549±0.00	0.9441±0.01	**0.8589±0.04**	0.1785±0.25	172.3570±10.08	10.0	**0.1017±0.00**	0.9453±0.01	**0.8700±0.07**	0.1575±0.25	178.3705±5.29
	NNDSVD	**9**	0.3690±0.01	0.8797±0.01	0.8444±0.10	0.3000±0.45	0.5423±0.07	8.0	0.3579±0.01	0.5907±0.03	0.7907±0.02	0.0751±0.17	0.1897±0.01	7.0	0.3570±0.01	0.5727±0.03	0.7718±0.02	0.1207±0.17	0.3880±0.09
	FactorAnalysis	**9**	**0.0000±0.00**	**1.0000±0.00**	0.8427±0.05	0.3514±0.11	0.1874±0.01	**9**	**0.0530±0.00**	0.9866±0.00	0.8578±0.03	0.3838±0.06	0.1391±0.01	8.0	0.1103±0.00	0.6761±0.16	0.7924±0.02	0.0661±0.15	0.0995±0.03
	NMF (CD)	**9**	0.0039±0.00	0.9873±0.01	0.8222±0.06	0.2000±0.27	0.2415±0.01	8.0	0.0684±0.00	0.7275±0.17	0.8318±0.03	**0.4034±0.06**	0.1567±0.01	8.0	0.1106±0.00	0.6790±0.13	0.7978±0.02	0.2174±0.20	0.1598±0.03
	NMF (MU)	**9**	0.0392±0.02	0.7897±0.12	0.8222±0.06	0.2000±0.27	0.1710±0.01	8.0	0.0779±0.01	0.6029±0.05	0.8113±0.05	0.1684±0.23	0.1462±0.01	8.0	0.1173±0.01	0.6004±0.04	0.8298±0.05	0.2528±0.23	0.1419±0.01
	SNMF	**9**	0.0007±0.00	0.9954±0.00	0.8667±0.09	0.4000±0.42	38.5410±5.61	8.0	0.0683±0.00	0.7273±0.17	0.8131±0.07	0.1516±0.34	42.5121±0.79	8.0	0.1106±0.00	0.6778±0.13	0.8173±0.05	0.2109±0.31	41.9616±3.51
	CoxNMF	**9**	0.0238±0.02	0.9992±0.00	**0.9111±0.05**	**0.6000±0.22**	4.5483±1.71	**9**	0.0602±0.01	**0.9998±0.00**	0.8358±0.04	0.3215±0.18	5.5385±0.50	**9**	0.1071±0.00	**0.9987±0.00**	0.8304±0.01	**0.3688±0.01**	5.4091±0.38
10	TruncatedSVD	**10**	**0.0000±0.00**	**1.0000±0.00**	0.8800±0.08	0.4000±0.42	**0.1455±0.00**	9.0	0.0649±0.00	0.8564±0.15	0.8412±0.05	0.3631±0.25	**0.1068±0.02**	9.0	0.1084±0.00	0.8088±0.13	0.8136±0.01	0.2755±0.15	**0.1130±0.02**
	PCA	12.0	**0.0000±0.00**	**1.0000±0.00**	0.8340±0.11	0.1327±0.17	0.1588±0.00	12.0	**0.0519±0.00**	0.9855±0.00	0.8758±0.09	0.1674±0.18	0.3168±0.09	9.0	0.1082±0.00	0.8246±0.11	0.8486±0.03	0.1962±0.17	0.1225±0.00
	SparsePCA	**10**	0.0144±0.00	0.9496±0.01	0.8646±0.03	0.0917±0.21	186.8570±2.77	**10**	0.0549±0.00	0.9384±0.00	0.7846±0.07	0.0788±0.11	192.0732±5.06	**10**	**0.1027±0.00**	0.9329±0.01	0.8244±0.06	0.0305±0.07	193.4766±4.39
	NNDSVD	**10**	0.3779±0.01	0.8627±0.01	0.8200±0.08	0.6000±0.42	0.3656±0.08	9.0	0.3672±0.01	0.6152±0.08	0.8490±0.07	0.2842±0.38	0.2062±0.00	9.0	0.3644±0.01	0.6131±0.07	0.8156±0.02	0.1455±0.19	0.2072±0.00
	FactorAnalysis	**10**	**0.0000±0.00**	**1.0000±0.00**	0.9172±0.07	0.5416±0.39	0.1749±0.00	9.0	0.0653±0.00	0.8941±0.09	0.8080±0.03	0.2181±0.20	0.1343±0.02	9.0	0.1084±0.00	0.8149±0.14	0.7924±0.03	0.2491±0.15	0.1289±0.02
	NMF (CD)	**10**	0.0063±0.01	0.9899±0.01	0.8800±0.04	0.4000±0.22	0.2906±0.09	9.0	0.0651±0.00	0.8553±0.15	0.8256±0.01	0.2970±0.17	0.2089±0.02	9.0	0.1088±0.00	0.8084±0.12	0.8190±0.01	0.2857±0.16	0.2231±0.01
	NMF (MU)	**10**	0.0378±0.01	0.7896±0.09	0.8400±0.05	0.2000±0.27	0.1714±0.01	9.0	0.0725±0.01	0.6987±0.15	0.8356±0.04	0.2487±0.23	0.1742±0.03	9.0	0.1137±0.00	0.6872±0.12	0.8344±0.06	0.1586±0.33	0.1746±0.02
	SNMF	**10**	0.0003±0.00	0.9997±0.00	0.8000±0.00	0.0000±0.00	38.0857±1.72	9.0	0.0650±0.00	0.8570±0.15	0.8174±0.02	0.1513±0.21	52.8744±5.31	9.0	0.1088±0.00	0.8087±0.12	0.8276±0.01	0.2920±0.16	52.4272±4.51
	CoxNMF	**10**	0.0255±0.02	0.9996±0.00	**0.9400±0.05**	**0.7000±0.27**	5.3164±0.10	**10**	0.0661±0.02	**0.9998±0.00**	**0.8766±0.04**	**0.4658±0.18**	3.9035±1.35	**10**	0.1102±0.00	**0.9994±0.00**	**0.8542±0.05**	**0.3621±0.26**	5.1742±0.96
11	TruncatedSVD	**11**	**0.0000±0.00**	**1.0000±0.00**	**0.9091±0.00**	**0.5000±0.00**	0.1542±0.01	10.0	0.0675±0.00	0.8794±0.08	0.8527±0.05	0.3587±0.26	**0.0967±0.02**	10.0	0.1102±0.00	0.8107±0.09	0.8385±0.04	0.3256±0.22	**0.1173±0.01**
	PCA	10.0	0.0423±0.00	0.9848±0.00	0.8787±0.03	0.2311±0.18	**0.0974±0.01**	13.0	**0.0522±0.00**	0.9883±0.00	**0.9258±0.02**	0.3450±0.08	0.1514±0.01	9.0	0.1272±0.01	0.5777±0.03	0.8555±0.01	0.1608±0.09	0.2572±0.07
	SparsePCA	**11**	0.0143±0.00	0.9583±0.01	0.8798±0.02	0.1461±0.20	210.2446±6.55	**11**	0.0551±0.00	0.9559±0.01	0.8320±0.07	0.0658±0.10	208.8611±1.91	**11**	**0.1030±0.00**	0.9495±0.01	0.8700±0.02	0.1288±0.18	212.1568±7.08
	NNDSVD	**11**	0.3823±0.01	0.8690±0.01	0.8909±0.04	0.4000±0.22	0.5929±0.04	10.0	0.3714±0.01	0.6101±0.02	0.8175±0.01	0.0754±0.16	0.2173±0.00	9.0	0.3691±0.01	0.5837±0.02	0.8191±0.04	0.2283±0.23	0.5975±0.01
	FactorAnalysis	**11**	**0.0000±0.00**	**1.0000±0.00**	0.9067±0.01	0.4588±0.04	0.1877±0.01	10.0	0.0682±0.00	0.9123±0.06	0.8331±0.05	0.2197±0.13	0.1307±0.02	10.0	0.1103±0.01	0.8223±0.09	0.8296±0.04	0.1795±0.17	0.1432±0.00
	NMF (CD)	**11**	0.0018±0.00	0.9956±0.01	**0.9091±0.06**	**0.5000±0.35**	0.2778±0.02	10.0	0.0678±0.00	0.8791±0.09	0.8627±0.04	0.3741±0.26	0.2187±0.02	10.0	0.1108±0.00	0.8118±0.09	0.8425±0.03	0.3313±0.23	0.2403±0.01
	NMF (MU)	**11**	0.0426±0.01	0.8316±0.08	0.8364±0.04	0.1000±0.22	0.2047±0.02	10.0	0.0769±0.01	0.7091±0.15	0.8344±0.01	0.1493±0.20	0.1971±0.03	10.0	0.1172±0.01	0.6674±0.15	0.8315±0.00	0.0689±0.15	0.1885±0.02
	SNMF	**11**	0.0015±0.00	0.9975±0.00	0.8909±0.10	0.4000±0.55	47.0463±7.15	**11**	0.0538±0.00	0.9855±0.00	0.8884±0.09	0.3641±0.50	54.0821±3.79	10.0	0.1108±0.00	0.8142±0.09	0.8289±0.01	0.1437±0.19	56.5522±1.29
	CoxNMF	**11**	0.0140±0.01	0.9999±0.00	**0.9091±0.06**	**0.5000±0.35**	5.7317±0.40	**11**	0.0565±0.00	**0.9999±0.00**	0.8913±0.03	**0.4680±0.16**	5.6968±0.19	**11**	0.1127±0.02	**0.9994±0.00**	**0.8902±0.05**	**0.4140±0.29**	5.8073±0.11
12	TruncatedSVD	**12**	**0.0000±0.00**	**1.0000±0.00**	0.8833±0.05	0.3000±0.27	**0.1592±0.00**	11.0	0.0660±0.00	0.8946±0.07	0.8533±0.00	0.3615±0.01	**0.1065±0.02**	11.0	0.1100±0.01	0.8179±0.11	0.8375±0.01	0.3369±0.02	**0.1266±0.00**
	PCA	14.0	**0.0000±0.00**	**1.0000±0.00**	0.8827±0.10	0.0242±0.04	0.1719±0.01	11.0	0.0659±0.00	0.8924±0.07	0.8787±0.01	0.2120±0.06	0.1106±0.02	**12**	**0.1032±0.01**	0.9672±0.00	0.8322±0.10	0.0869±0.09	0.2830±0.12
	SparsePCA	11.0	0.0411±0.00	0.8210±0.11	0.8682±0.01	0.0505±0.11	210.9011±1.89	13.0	0.0552±0.00	0.9544±0.00	**0.9193±0.02**	0.0742±0.17	235.0052±5.16	**12**	0.1040±0.01	0.9419±0.01	0.8662±0.00	0.0000±0.00	224.2796±11.16
	NNDSVD	**12**	0.3935±0.00	0.8621±0.01	0.8667±0.05	0.2000±0.27	0.5998±0.02	11.0	0.3820±0.00	0.6489±0.07	0.8400±0.01	0.1452±0.19	0.3896±0.01	11.0	0.3781±0.00	0.6480±0.07	0.8488±0.04	0.0968±0.20	0.4005±0.04
	FactorAnalysis	**12**	**0.0000±0.00**	**1.0000±0.00**	0.8915±0.03	0.2546±0.23	0.1976±0.01	11.0	0.0665±0.00	0.8550±0.14	0.8555±0.05	0.3640±0.06	0.1368±0.02	11.0	0.1100±0.01	0.8183±0.11	0.8317±0.04	0.2969±0.05	0.1470±0.01
	NMF (CD)	**12**	0.0020±0.00	0.9945±0.01	0.8667±0.05	0.2000±0.27	0.2896±0.01	11.0	0.0664±0.00	0.8943±0.07	0.8692±0.04	**0.4295±0.15**	0.2476±0.02	11.0	0.1108±0.01	0.8168±0.11	0.8593±0.04	**0.4102±0.15**	0.2650±0.01
	NMF (MU)	**12**	0.0491±0.01	0.7388±0.10	0.8500±0.04	0.1000±0.22	0.2048±0.03	11.0	0.0788±0.01	0.6606±0.13	0.8465±0.01	0.1502±0.21	0.2114±0.03	11.0	0.1182±0.01	0.6664±0.12	0.8450±0.00	0.1379±0.19	0.1983±0.02
	SNMF	**12**	0.0005±0.00	0.9977±0.00	0.8500±0.04	0.1000±0.22	51.9233±4.94	**12**	**0.0545±0.00**	0.9858±0.00	0.8782±0.04	0.2716±0.25	59.3599±1.96	11.0	0.1107±0.01	0.8141±0.11	0.8417±0.01	0.1503±0.18	58.9192±4.82
	CoxNMF	**12**	0.0197±0.01	0.9998±0.00	**0.9000±0.04**	**0.4000±0.22**	3.9004±1.70	**12**	0.0585±0.00	**0.9999±0.00**	0.8775±0.03	0.3120±0.18	6.0485±0.20	13.0	0.1051±0.01	**0.9999±0.00**	**0.9138±0.03**	0.3073±0.29	5.9808±0.13

reported. Among real human cancer data experiments, we investigate some gene clusters which have conspicuous signal to the survival.

Colon Adenocarcinoma

For Colon adenocarcinoma (COAD), we have $P = 13,140$ filtered genes, and $N = 298$ samples. We find $\hat{K} = 10$ returns highest silhouette score. The optimization results C-Index $= 1.0$, relative error $= 8.8071\%$. In the results, we verified several important biological processes. Clusters in Figure 4.6A which may play important roles to survival are performed GO enrichment analysis in Table 4.11. We focus on two clusters C7 and C5. The cluster

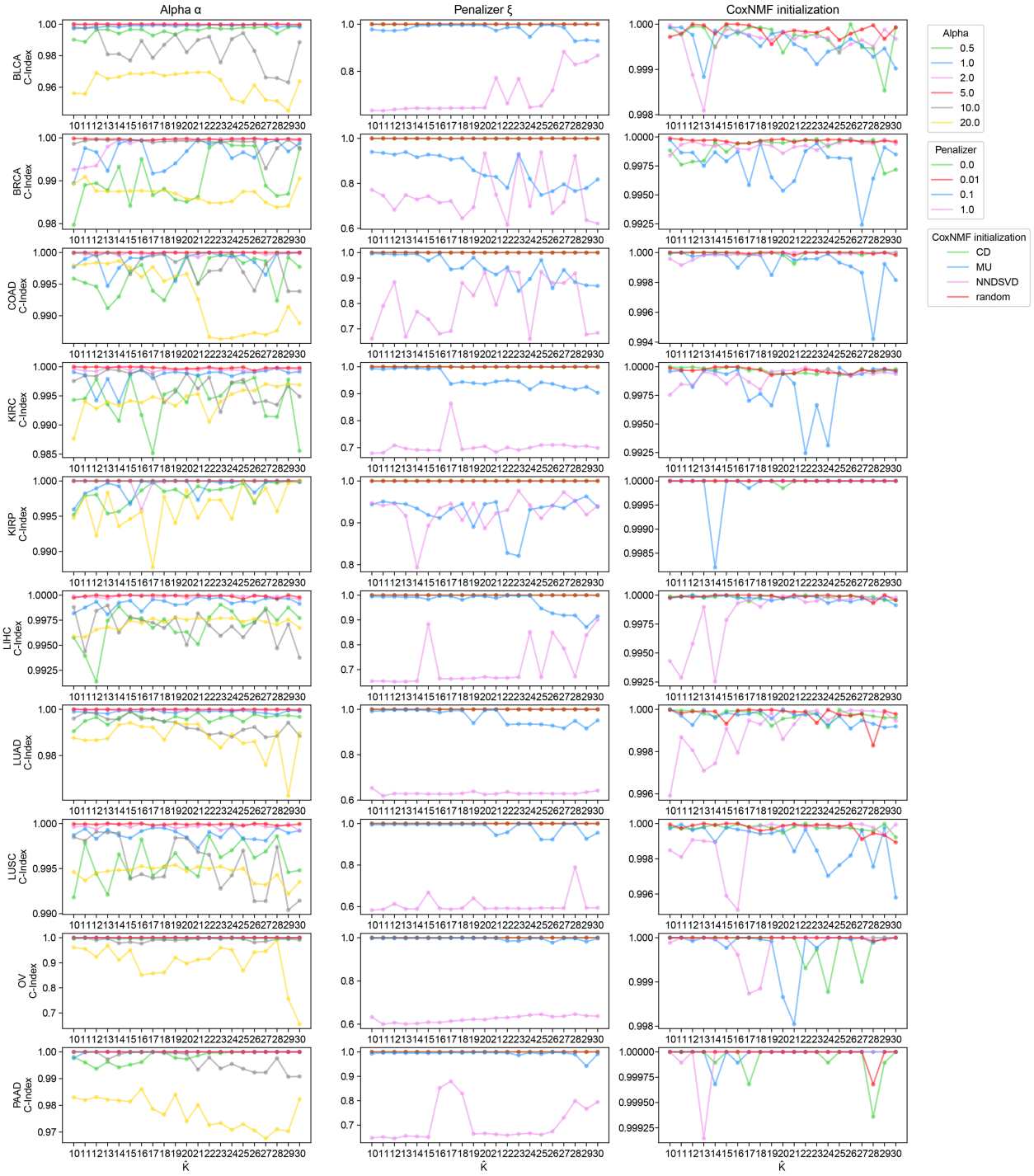


Figure 4.5. CoxNMF hyper-parameter guidance in TCGA human cancer. When certain parameter fixed, the highest C-Index are reported. X-axis: $\hat{K}$, Y-axis: C-Index. BLCA: Bladder urothelial carcinoma; BRCA: Breast invasive carcinoma; COAD: Colon adenocarcinoma; KIRC: Kidney renal clear cell carcinoma; KIRP: Kidney renal papillary cell carcinoma; LIHC: Liver hepatocellular carcinoma; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; OV: Ovarian serous cystadenocarcinoma; PAAD: Pancreatic adenocarcinoma.

Table 4.10. Simulation results in secondary multivariate simulation setup among all combinations of $K \in \{6, 7, 8, 9, 10, 11, 12\}$ and $\varepsilon \in \{0, 0.05, 0.10\}$. Experiments repeat 5 times each with random seed $\in \{1, 2, 3, 4, 5\}$. Mean values $\pm$ standard deviations are reported, best performed mean values among models are highlighted in bold font.

		$\varepsilon = 0.00$						$\varepsilon = 0.05$						$\varepsilon = 0.10$					
K	Model	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime	$\hat{K}$	Relative error	C-Index	Accuracy	Dice coefficient	Runtime
6	TruncatedSVD	**6**	**0.0000±0.00**	**1.0000±0.00**	0.7667±0.04	0.4400±0.09	**0.1409±0.00**	5.0	0.0802±0.00	0.8749±0.07	0.6892±0.05	0.3087±0.18	**0.0848±0.02**	5.0	0.1241±0.01	0.7913±0.10	0.6692±0.05	0.2738±0.18	**0.1238±0.03**
	PCA	4.0	0.1400±0.01	0.5664±0.04	0.6528±0.04	0.1628±0.06	0.1852±0.03	**6**	**0.0585±0.00**	0.9842±0.00	0.7543±0.04	0.2833±0.13	0.1111±0.02	7.0	**0.1101±0.01**	0.9641±0.00	**0.8065±0.04**	0.2608±0.14	0.2758±0.14
	SparsePCA	**6**	0.0158±0.00	0.9545±0.01	0.6608±0.05	0.2030±0.12	120.1996±1.83	**6**	0.0604±0.00	0.9495±0.00	0.6483±0.06	0.2172±0.13	119.0627±3.43	**6**	0.1125±0.01	0.9394±0.01	0.6550±0.04	0.2486±0.10	128.8546±15.10
	NNDSVD	**6**	0.3292±0.01	0.8695±0.00	0.7000±0.05	0.2800±0.11	0.4933±0.02	4.0	0.3260±0.01	0.5567±0.02	0.6188±0.07	0.2357±0.14	0.5098±0.03	4.0	0.3268±0.01	0.5555±0.02	0.6030±0.08	0.2310±0.16	0.5576±0.09
	FactorAnalysis	**6**	**0.0000±0.00**	**1.0000±0.00**	0.7592±0.03	0.4100±0.06	0.1575±0.01	**6**	**0.0585±0.00**	0.9842±0.00	0.7042±0.06	0.2362±0.17	0.1193±0.02	**6**	0.1117±0.01	0.9636±0.00	0.7432±0.09	0.3288±0.22	0.2626±0.08
	NMF (CD)	**6**	0.0033±0.00	0.9932±0.01	0.7500±0.06	0.4000±0.14	0.1905±0.01	5.0	0.0803±0.00	0.8750±0.07	0.7178±0.05	0.3318±0.14	0.1392±0.01	5.0	0.1241±0.01	0.7905±0.10	0.6993±0.04	0.2859±0.14	0.1669±0.03
	NMF (MU)	**6**	0.0203±0.00	0.9527±0.02	0.6667±0.06	0.2000±0.14	0.1505±0.02	5.0	0.0826±0.00	0.8275±0.08	0.6348±0.04	0.1485±0.08	0.1399±0.01	5.0	0.1255±0.01	0.7658±0.09	0.6367±0.03	0.1403±0.08	0.1746±0.03
	SNMF	**6**	0.0003±0.00	0.9995±0.00	0.7333±0.07	0.3600±0.17	20.8056±4.07	**6**	0.0586±0.00	0.9841±0.00	0.7245±0.03	0.3396±0.09	27.0718±3.34	5.0	0.1241±0.01	0.7906±0.10	0.6852±0.02	0.2513±0.09	27.5283±6.42
	CoxNMF	**6**	0.0073±0.00	**1.0000±0.00**	**0.8167±0.04**	**0.5600±0.09**	4.3176±0.13	**6**	0.0607±0.00	**0.9996±0.00**	**0.8323±0.10**	**0.6033±0.22**	5.3732±0.24	**6**	0.1126±0.01	**0.9998±0.00**	0.7953±0.08	**0.5085±0.21**	4.6531±0.77
7	TruncatedSVD	**7**	**0.0000±0.00**	**1.0000±0.00**	0.7286±0.06	0.2400±0.17	**0.1537±0.03**	6.0	0.0738±0.00	0.8388±0.11	0.7274±0.03	0.3338±0.10	**0.0885±0.00**	6.0	0.1155±0.00	0.7570±0.11	0.7216±0.03	0.3341±0.11	**0.1079±0.02**
	PCA	9.0	**0.0000±0.00**	**1.0000±0.00**	**0.8281±0.06**	0.0926±0.11	0.1606±0.03	8.0	**0.0540±0.00**	0.9857±0.00	**0.8147±0.03**	0.1376±0.11	0.1390±0.01	6.0	0.1153±0.00	0.7670±0.11	0.7289±0.02	0.1303±0.06	0.1096±0.02
	SparsePCA	**7**	0.0149±0.00	0.9537±0.01	0.7026±0.05	0.1304±0.12	148.4373±22.10	**7**	0.0564±0.00	0.9528±0.01	0.7010±0.04	0.2425±0.09	142.8247±1.59	**7**	**0.1055±0.00**	0.9478±0.01	0.6824±0.03	0.1828±0.13	141.4388±4.34
	NNDSVD	**7**	0.3435±0.01	0.8768±0.01	0.7714±0.08	0.3600±0.22	0.5365±0.06	5.0	0.3379±0.01	0.5574±0.02	0.6570±0.03	0.2032±0.07	0.5361±0.01	5.0	0.3367±0.01	0.5580±0.02	0.6579±0.03	0.2063±0.07	0.5352±0.03
	FactorAnalysis	**7**	**0.0000±0.00**	**1.0000±0.00**	0.7250±0.03	0.2512±0.05	0.1756±0.03	**7**	0.0547±0.00	0.9858±0.00	0.7526±0.07	0.2997±0.17	0.2587±0.11	6.0	0.1155±0.00	0.7654±0.12	0.7333±0.06	0.2309±0.16	0.1177±0.02
	NMF (CD)	**7**	0.0026±0.00	0.9965±0.00	0.7714±0.08	0.3600±0.22	0.2192±0.04	6.0	0.0738±0.00	0.8380±0.11	0.7101±0.05	0.2493±0.16	0.1506±0.00	6.0	0.1157±0.00	0.7574±0.12	0.7346±0.05	0.3059±0.19	0.1694±0.02
	NMF (MU)	**7**	0.0374±0.01	0.8368±0.09	0.7143±0.07	0.2000±0.20	0.1767±0.03	6.0	0.0767±0.01	0.7488±0.14	0.7199±0.03	0.2573±0.09	0.1554±0.02	6.0	0.1173±0.01	0.7014±0.12	0.7419±0.05	0.3101±0.18	0.1601±0.03
	SNMF	**7**	0.0005±0.00	0.9992±0.00	0.7286±0.06	0.2400±0.17	25.0171±4.37	6.0	0.0738±0.00	0.8378±0.11	0.6987±0.04	0.2130±0.15	27.7422±1.96	6.0	0.1156±0.00	0.7566±0.12	0.6811±0.02	0.1776±0.01	31.5182±5.41
	CoxNMF	**7**	0.0142±0.02	0.9998±0.00	0.8000±0.06	**0.4400±0.17**	5.8786±1.19	**7**	0.0587±0.01	**0.9999±0.00**	0.7814±0.09	**0.3876±0.24**	5.8843±0.16	**7**	0.1058±0.00	**0.9998±0.00**	**0.8124±0.08**	**0.4748±0.23**	4.6515±0.30
8	TruncatedSVD	**8**	**0.0000±0.00**	**1.0000±0.00**	0.7750±0.03	0.2800±0.11	**0.1453±0.00**	7.0	0.0726±0.00	0.8717±0.09	0.7264±0.03	0.1383±0.14	**0.0930±0.02**	7.0	0.1143±0.00	0.7903±0.13	0.7264±0.03	0.1362±0.14	0.1054±0.02
	PCA	10.0	**0.0000±0.00**	**1.0000±0.00**	**0.8269±0.08**	0.1928±0.09	0.1518±0.00	7.0	0.0724±0.00	0.8724±0.09	0.7767±0.02	0.1045±0.07	0.0947±0.02	10.0	**0.1014±0.00**	0.9700±0.00	**0.8649±0.01**	0.0725±0.08	**0.1014±0.01**
	SparsePCA	**8**	0.0140±0.00	0.9605±0.01	0.7365±0.05	0.1776±0.05	154.4487±3.67	9.0	0.0554±0.00	0.9604±0.00	**0.7851±0.05**	0.2056±0.08	158.1613±2.68	9.0	0.1040±0.00	0.9461±0.01	0.7736±0.05	0.0983±0.10	162.0641±6.00
	NNDSVD	**8**	0.3559±0.00	0.8661±0.01	0.7750±0.03	0.2800±0.11	0.5617±0.01	6.0	0.3494±0.00	0.5632±0.02	0.7006±0.03	0.1020±0.15	0.4878±0.03	6.0	0.3480±0.00	0.5595±0.02	0.7272±0.06	0.2036±0.19	0.4115±0.07
	FactorAnalysis	**8**	**0.0000±0.00**	**1.0000±0.00**	0.7491±0.04	0.1793±0.14	0.1662±0.00	**8**	**0.0543±0.00**	0.9873±0.00	0.7370±0.03	0.2072±0.07	0.2138±0.08	7.0	0.1144±0.00	0.8007±0.12	0.7348±0.05	0.1605±0.13	0.1238±0.02
	NMF (CD)	**8**	0.0031±0.00	0.9916±0.01	0.7750±0.03	0.2800±0.11	0.2065±0.01	7.0	0.0727±0.00	0.8707±0.10	0.7559±0.06	0.2226±0.23	0.1658±0.02	7.0	0.1145±0.00	0.7873±0.13	0.7701±0.05	0.2532±0.19	0.1818±0.02
	NMF (MU)	**8**	0.0285±0.01	0.9198±0.03	0.7250±0.03	0.1200±0.11	0.1696±0.02	7.0	0.0764±0.00	0.7432±0.12	0.7458±0.03	0.1899±0.13	0.1611±0.02	7.0	0.1169±0.00	0.7111±0.12	0.7470±0.04	0.1915±0.13	0.1579±0.02
	SNMF	**8**	0.0001±0.00	**1.0000±0.00**	0.7625±0.05	0.2400±0.17	28.7052±3.82	**8**	0.0545±0.00	0.9868±0.00	0.7595±0.05	0.2383±0.16	36.4144±4.74	7.0	0.1146±0.00	0.7915±0.13	0.7501±0.06	0.2202±0.18	37.1345±2.49
	CoxNMF	**8**	0.0151±0.01	**1.0000±0.00**	0.8000±0.03	**0.3600±0.09**	4.3617±0.05	**8**	0.0581±0.00	**0.9997±0.00**	0.7746±0.03	**0.2903±0.11**	5.3298±0.19	**8**	0.1051±0.00	**0.9994±0.00**	0.7785±0.05	**0.2778±0.16**	5.5859±0.30
9	TruncatedSVD	**9**	**0.0000±0.00**	**1.0000±0.00**	0.7889±0.07	0.2400±0.26	**0.1505±0.00**	8.0	0.0734±0.00	0.8352±0.13	0.7694±0.05	0.2563±0.24	**0.0924±0.01**	8.0	0.1148±0.00	0.7712±0.12	0.7656±0.04	0.2469±0.23	**0.1059±0.02**
	PCA	11.0	**0.0000±0.00**	**1.0000±0.00**	**0.8738±0.01**	0.0668±0.11	0.1600±0.00	**9**	**0.0543±0.00**	0.9882±0.00	0.7733±0.04	0.1935±0.09	0.1597±0.01	8.0	0.1146±0.00	0.7747±0.11	**0.8123±0.03**	0.2272±0.20	0.1134±0.02
	SparsePCA	8.0	0.0525±0.00	0.8088±0.14	0.7733±0.02	0.0671±0.06	156.7241±2.82	**9**	0.0560±0.00	0.9600±0.00	0.7821±0.04	0.1319±0.14	173.9911±4.98	**9**	0.1048±0.00	0.9534±0.01	0.7863±0.04	0.1545±0.05	172.3750±3.33
	NNDSVD	**9**	0.3594±0.01	0.8648±0.01	0.7667±0.02	0.1600±0.09	0.5523±0.02	8.0	0.3493±0.01	0.5967±0.02	0.7388±0.01	0.1075±0.10	0.3418±0.04	7.0	0.3492±0.01	0.5659±0.03	0.7733±0.04	0.2646±0.13	0.4986±0.11
	FactorAnalysis	**9**	**0.0000±0.00**	**1.0000±0.00**	0.7878±0.04	0.1730±0.17	0.1770±0.00	**9**	**0.0543±0.00**	0.9881±0.00	0.7881±0.06	**0.2988±0.11**	0.1712±0.01	8.0	0.1149±0.00	0.7770±0.10	0.8071±0.05	0.2913±0.23	0.1263±0.02
	NMF (CD)	**9**	0.0052±0.00	0.9853±0.01	0.7667±0.02	0.1600±0.09	0.2425±0.00	8.0	0.0735±0.00	0.8353±0.13	0.7507±0.02	0.1459±0.08	0.1692±0.01	8.0	0.1151±0.00	0.7714±0.12	0.7577±0.03	0.1787±0.12	0.1702±0.02
	NMF (MU)	**9**	0.0307±0.01	0.9091±0.04	0.8000±0.06	0.2800±0.23	0.1823±0.02	8.0	0.0766±0.00	0.7216±0.11	0.7823±0.04	0.2634±0.11	0.1769±0.02	8.0	0.1171±0.00	0.7028±0.12	0.7511±0.02	0.1435±0.08	0.1600±0.02
	SNMF	**9**	0.0008±0.00	0.9988±0.00	0.7778±0.06	0.2000±0.20	41.7896±2.25	**9**	0.0546±0.00	0.9879±0.00	0.7840±0.02	0.2327±0.09	44.1778±2.20	**9**	**0.1046±0.00**	0.9718±0.00	0.7993±0.05	**0.2938±0.16**	45.6264±1.09
	CoxNMF	**9**	0.0073±0.00	0.9999±0.00	0.8333±0.04	**0.4000±0.14**	4.5128±0.11	**9**	0.0558±0.00	**0.9999±0.00**	**0.7981±0.04**	0.2567±0.16	4.5279±0.05	**9**	0.1055±0.00	**0.9998±0.00**	0.7772±0.05	0.2020±0.18	4.4788±0.12
10	TruncatedSVD	**10**	**0.0000±0.00**	**1.0000±0.00**	0.8200±0.06	0.2800±0.23	0.1716±0.01	9.0	0.0730±0.00	0.9182±0.03	0.7965±0.03	0.2844±0.19	**0.0967±0.02**	9.0	0.1164±0.01	0.8446±0.06	0.7910±0.03	0.2776±0.18	0.1143±0.02
	PCA	9.0	0.0482±0.00	0.9856±0.00	0.8234±0.03	0.1759±0.13	**0.1080±0.01**	9.0	0.0729±0.00	0.9222±0.03	0.8249±0.02	0.1931±0.13	0.1043±0.02	8.0	0.1390±0.01	0.5717±0.03	0.8047±0.01	0.0477±0.07	**0.0713±0.00**
	SparsePCA	**10**	0.0146±0.00	0.9592±0.01	0.7794±0.06	0.1700±0.06	211.2021±13.47	**10**	0.0574±0.00	0.9603±0.01	0.7601±0.06	0.0956±0.09	191.0190±4.46	**10**	**0.1074±0.01**	0.9502±0.01	0.7865±0.04	0.1337±0.08	191.0090±3.95
	NNDSVD	**10**	0.3819±0.01	0.8696±0.01	0.8200±0.06	0.2800±0.23	0.3774±0.06	8.0	0.3724±0.01	0.5663±0.02	0.7789±0.05	0.1704±0.19	0.2021±0.01	8.0	0.3693±0.01	0.5639±0.02	0.7745±0.05	0.1602±0.17	0.2000±0.01
	FactorAnalysis	**10**	**0.0000±0.00**	**1.0000±0.00**	0.8028±0.02	0.1505±0.11	0.2109±0.02	**10**	**0.0558±0.00**	0.9879±0.00	0.8211±0.02	0.2386±0.08	0.2077±0.07	9.0	0.1165±0.01	0.8552±0.06	0.8070±0.04	0.2141±0.14	0.1333±0.02
	NMF (CD)	**10**	0.0012±0.00	0.9987±0.00	0.8200±0.03	0.2800±0.11	0.2866±0.02	9.0	0.0732±0.00	0.9173±0.03	0.7905±0.03	0.2095±0.14	0.1972±0.01	9.0	0.1168±0.01	0.8391±0.06	0.7699±0.02	0.1339±0.14	0.2079±0.02
	NMF (MU)	**10**	0.0372±0.01	0.8555±0.08	0.7600±0.02	0.0400±0.09	0.2313±0.03	9.0	0.0779±0.01	0.8074±0.06	0.8000±0.03	0.2265±0.08	0.2068±0.02	9.0	0.1201±0.01	0.7453±0.08	0.8055±0.03	0.2491±0.08	0.1973±0.03
	SNMF	**10**	0.0005±0.00	0.9992±0.00	0.8300±0.04	0.3200±0.18	46.1037±8.28	**10**	0.0561±0.00	0.9876±0.00	**0.8268±0.04**	**0.3087±0.17**	51.6823±1.58	9.0	0.1168±0.01	0.8402±0.06	0.7774±0.02	0.1379±0.14	48.7015±7.53
	CoxNMF	**10**	0.0051±0.00	0.9999±0.00	**0.8400±0.04**	**0.3600±0.17**	5.2114±0.33	**10**	0.0568±0.00	**1.0000±0.00**	0.8154±0.03	0.2867±0.11	4.6317±0.21	**10**	0.1083±0.01	**0.9997±0.00**	**0.8286±0.06**	**0.3066±0.25**	5.6290±0.33
11	TruncatedSVD	**11**	**0.0000±0.00**	**1.0000±0.00**	0.8091±0.04	0.1600±0.17	**0.1587±0.00**	10.0	0.0695±0.00	0.8159±0.19	0.7964±0.02	0.2124±0.20	**0.1121±0.03**	10.0	0.1142±0.01	0.7891±0.18	0.7991±0.02	0.2340±0.22	**0.1088±0.04**
	PCA	**11**	**0.0000±0.00**	**1.0000±0.00**	0.8107±0.04	0.1033±0.07	0.1599±0.01	**11**	**0.0558±0.00**	0.9858±0.00	0.8345±0.01	0.1072±0.07	0.3294±0.09	10.0	0.1141±0.01	0.7846±0.18	0.8285±0.01	0.1488±0.10	0.1114±0.04
	SparsePCA	**11**	0.0147±0.00	0.9540±0.01	0.8044±0.03	0.0878±0.08	213.4510±5.92	12.0	0.0570±0.00	0.9541±0.01	0.8245±0.04	0.0858±0.09	221.0057±2.95	10.0	0.1148±0.01	0.7697±0.17	0.8040±0.03	0.0687±0.07	198.3327±15.49
	NNDSVD	**11**	0.3904±0.01	0.8657±0.01	0.8182±0.03	0.2000±0.14	0.4134±0.07	9.0	0.3807±0.01	0.5826±0.02	0.7876±0.02	0.1545±0.15	0.5984±0.03	9.0	0.3768±0.01	0.5787±0.02	0.7874±0.03	0.1555±0.10	0.5948±0.04
	FactorAnalysis	**11**	**0.0000±0.00**	**1.0000±0.00**	0.8017±0.02	0.1305±0.07	0.1901±0.01	**11**	**0.0558±0.00**	0.9857±0.00	**0.8424±0.03**	0.2657±0.09	0.1479±0.01	10.0	0.1143±0.01	0.7846±0.18	0.8199±0.02	0.1437±0.14	0.2355±0.10
	NMF (CD)	**11**	0.0036±0.00	0.9892±0.01	0.8273±0.05	0.2400±0.22	0.3284±0.09	10.0	0.0697±0.00	0.8152±0.19	0.8120±0.04	0.2162±0.19	0.2370±0.03	10.0	0.1149±0.01	0.7840±0.18	0.8009±0.02	0.1754±0.12	0.2327±0.04
	NMF (MU)	**11**	0.0445±0.01	0.8000±0.12	0.8091±0.04	0.1600±0.17	0.2135±0.03	10.0	0.0764±0.00	0.7461±0.16	0.8091±0.04	0.1781±0.18	0.2089±0.03	10.0	0.1191±0.01	0.7290±0.15	0.8012±0.02	0.1367±0.14	0.2106±0.04
	SNMF	**11**	0.0003±0.00	**1.0000±0.00**	0.8273±0.06	0.2400±0.26	51.3815±3.43	**11**	0.0562±0.00	0.9862±0.00	0.8143±0.03	0.1871±0.13	62.4776±4.43	10.0	0.1149±0.01	0.7865±0.17	0.7951±0.02	0.1062±0.09	57.4164±2.38
	CoxNMF	**11**	0.0078±0.01	0.9998±0.00	**0.8545±0.04**	**0.3600±0.17**	4.8397±2.20	**11**	0.0570±0.00	**0.9998±0.00**	0.8402±0.03	**0.3124±0.15**	5.9937±0.22	12.0	**0.1082±0.01**	**1.0000±0.00**	**0.8836±0.04**	**0.3298±0.23**	6.0695±0.60
12	TruncatedSVD	**12**	**0.0000±0.00**	**1.0000±0.00**	0.8333±0.04	0.2000±0.20	0.1652±0.01	11.0	0.0667±0.00	0.8593±0.14	0.8173±0.02	**0.2788±0.17**	0.1064±0.02	11.0	0.1127±0.01	0.7985±0.11	0.8175±0.03	**0.3084±0.10**	**0.1162±0.02**
	PCA	11.0	0.0372±0.00	0.9815±0.01	0.8436±0.01	0.1695±0.04	**0.0966±0.00**	11.0	0.0666±0.00	0.8572±0.15	**0.8437±0.02**	0.1866±0.07	**0.0928±0.01**	11.0	0.1126±0.01	0.8032±0.12	0.8433±0.02	0.1815±0.08	0.1195±0.02
	SparsePCA	**12**	0.0151±0.00	0.9413±0.01	0.8218±0.05	0.1692±0.12	228.1820±6.38	**12**	0.0576±0.00	0.9362±0.01	0.8268±0.03	0.0921±0.13	226.9736±1.68	**12**	**0.1074±0.01**	0.9304±0.01	0.8346±0.02	0.0839±0.12	225.0878±6.60
	NNDSVD	**12**	0.4069±0.01	0.8687±0.01	0.8417±0.03	0.2400±0.17	0.2514±0.00	10.0	0.3950±0.01	0.5819±0.02	0.8172±0.02	0.1718±0.12	0.2236±0.00	10.0	0.3905±0.01	0.5823±0.02	0.8170±0.01	0.1609±0.15	0.2275±0.01
	FactorAnalysis	**12**	**0.0000±0.00**	**1.0000±0.00**	**0.8500±0.02**	0.2197±0.08	0.2019±0.01	**12**	**0.0558±0.00**	0.9849±0.00	0.8381±0.03	0.1989±0.11	0.3379±0.09	11.0	0.1128±0.01	0.8076±0.12	**0.8468±0.02**	0.2275±0.08	0.1362±0.02
	NMF (CD)	**12**	0.0051±0.01	0.9918±0.01	**0.8500±0.02**	**0.2800±0.11**	0.2958±0.01	11.0	0.0675±0.00	0.8287±0.13	0.8078±0.01	0.1423±0.08	0.2389±0.01	11.0	0.1136±0.01	0.7931±0.11	0.8023±0.00	0.1397±0.08	0.2568±0.02
	NMF (MU)	**12**	0.0365±0.01	0.8565±0.08	0.8417±0.05	0.2400±0.22	0.2233±0.03	11.0	0.0725±0.00	0.6964±0.07	0.8189±0.02	0.1772±0.12	0.2017±0.02	11.0	0.1173±0.01	0.6745±0.06	0.8252±0.02	0.1764±0.12	0.1985±0.02
	SNMF	**12**	0.0008±0.00	0.9985±0.00	**0.8500±0.02**	**0.2800±0.11**	44.8740±12.97	**12**	0.0564±0.00	0.9827±0.00	0.8302±0.03	0.1902±0.14	58.2628±3.47	11.0	0.1136±0.01	0.7998±0.09	0.8171±0.02	0.1979±0.13	56.3340±7.89
	CoxNMF	**12**	0.0084±0.00	**1.0000±0.00**	0.8417±0.02	0.2400±0.09	4.9341±0.18	**12**	0.0651±0.01	**0.9995±0.00**	0.8384±0.03	0.2440±0.11	5.0465±2.12	**12**	0.1089±0.01	**0.9988±0.00**	0.8423±0.02	0.2532±0.09	5.3910±1.34

C7, which is negatively associated with survival, is enriched with many digestion and epithelial related GO terms. For example, digestion (GO:0007586, P-value $= 6.00 \times 10^{-11}$), digestive system process (GO:0022600, P-value $= 8.20 \times 10^{-9}$), maintenance of gastrointestinal epithelium (GO:0030227, P-value $= 2.01 \times 10^{-10}$), and epithelial structure maintenance (GO:0010669, P-value $= 3.86 \times 10^{-9}$), etc. The CoxNMF results suggest that higher gene expressions in C7 cluster will generally result in shorter survival time.

The cluster C5, which is both positively and negatively associated with survival, is highly enriched with many immune system process GO terms. For example, regulation of immune system process (GO:0002682, P-value $= 9.51 \times 10^{-75}$), regulation of immune re-

sponse (GO:0050776, P-value $= 1.24 \times 10^{-65}$), positive regulation of immune system process (GO:0002684, P-value $= 9.99 \times 10^{-61}$), and lymphocyte activation (GO:0046649, P-value $= 3.48 \times 10^{-57}$). These enriched GO terms suggest that gene expressions in C5 cluster are highly related to immune system response and related activation.

Kidney Renal Clear Cell Carcinoma

For Kidney renal clear cell carcinoma (KIRC), we have $P = 13,140$ filtered genes, and $N = 533$ samples. We find $\hat{K} = 11$ returns highest silhouette score. The optimization results C-Index $= 0.9997$, relative error $= 8.7583\%$. Clusters in Figure 4.7A which may play important roles to survival are performed GO enrichment analysis in Table 4.12. We focus on two clusters C5 and C9. In the cluster C5, which is associated with both better and worse survival, we find several noticeable biological process terms, such as urate metabolic process (GO:0046415, P-value $= 4.19 \times 10^{-9}$), uronic acid metabolic process (GO:0006063, P-value $= 8.82 \times 10^{-9}$), and urate transport (GO:0015747, P-value $= 4.19 \times 10^{-9}$).

In the contrary, cluster C9 is associated with better survival. Cluster C9 is enriched with several immune response related biological process, including acute inflammatory response (GO:0002526, P-value $= 3.79 \times 10^{-16}$), humoral immune response (GO:0006959, P-value $= 2.57 \times 10^{-9}$), and inflammatory response (GO:0006954, P-value $= 1.55 \times 10^{-8}$).

Pancreatic Adenocarcinoma

For Pancreatic adenocarcinoma (PAAD), we have $P = 13,140$ filtered genes, and $N = 178$ samples. We find $\hat{K} = 12$ returns highest silhouette score. The optimization returns C-Index $= 1.0$, relative error $= 8.0890\%$. Clusters in Figure 4.8A which may play important roles to survival are performed GO enrichment analysis in Table 4.13. We focus on two clusters C1 and C12. In the cluster C1, which is associated with worse survival, we find several interesting biological process terms related to potassium ion and insulin, such as potassium ion transmembrane transport (GO:0071805, P-value $= 9.49 \times 10^{-15}$), potassium ion transport (GO:0006813, P-value $= 2.58 \times 10^{-14}$), insulin secretion (GO:0030073, P-value $= 5.79 \times 10^{-14}$), regulation of insulin secretion (GO:0050796, P-value $= 2.15 \times 10^{-10}$).

In the contrary, cluster C12 is associated with better survival. Cluster C12 is enriched with digestion related biological processes, including digestion (GO:0007586, P-value $= 4.69 \times 10^{-19}$), digestive system process (GO:0022600, P-value $= 3.08 \times 10^{-11}$). Furthermore, cluster C12 is also enriched with glycosylation related biological processes, including O-glycan processing (GO:0016266, P-value $= 7.40 \times 10^{-12}$), protein O-linked glycosylation (GO:0006493, P-value $= 4.89 \times 10^{-8}$), macromolecule glycosylation (GO:0043413, P-value $= 2.86 \times 10^{-7}$), etc.

Lung Squamous Cell Carcinoma

For Lung squamous cell carcinoma (LUSC), we have $P = 13,140$ filtered genes, and $N = 495$ samples. We find $\hat{K} = 10$ returns highest silhouette score. The optimization results C-Index $= 0.9995$, relative error $= 10.2462\%$. Clusters in Figure 4.9A which may play important roles to survival are performed GO enrichment analysis in Table 4.14. We are especially interested in gene cluster C6, which is associated with worse survival. From the gene ontology enrichment analysis results, we identify that cluster C6 is highly associated with cornification and keratinization. For example, cornified evelope assembly (GO:1903575, P-value $= 2.33 \times 10^{-46}$), cornification (GO:0070268, P-value $= 4.16 \times 10^{-46}$), keratinocyte differentiation (GO:0030216, P-value $= 2.98 \times 10^{-35}$), and keratinization (GO:0031424, P-value $= 1.44 \times 10^{-31}$).

Bladder Urothelial Carcinoma

For Bladder urothelial carcinoma (BLCA), we have $P = 13,140$ filtered genes, and $N = 406$ samples. We find $\hat{K} = 11$ returns highest silhouette score. The optimization results C-Index $= 0.9998$, relative error $= 11.3984\%$. Clusters in Figure 4.10A which may play important roles to survival are performed GO enrichment analysis in Table 4.15. We focus on two clusters C5 and C7. In the cluster C5, which associated with worse survival, we find several noticeable biological process terms related to digestion and epithelium development, such as digestion (GO:0007586, P-value $= 2.76 \times 10^{-6}$), epithelial cell differentia-

tion (GO:0030855, P-value $= 5.76 \times 10^{-11}$), epithelium development (GO:0060429, P-value $= 6.48 \times 10^{-9}$), epidermis development (GO:0008544, P-value $= 2.88 \times 10^{-6}$).

Similarly, cluster C7 is enriched with epithelium development, including epithelium development (GO:0060429, P-value $= 5.81 \times 10^{-17}$), epithelial cell differentiation (GO:0030855, P-value $= 9.98 \times 10^{-15}$), epidermis development (GO:0008544, P-value $= 1.27 \times 10^{-12}$), skin epidermis development (GO:0098773, P-value $= 3.40 \times 10^{-10}$), regulation of epidermis development (GO:0045682, P-value $= 1.51 \times 10^{-7}$), epithelial cell development (GO:0002064, P-value $= 6.97 \times 10^{-7}$), morphogenesis of an epithelium (GO:0002009, P-value $= 9.18 \times 10^{-7}$), etc.

Breast Invasive Carcinoma

For Breast invasive carcinoma (BRCA), we have $P = 13,140$ filtered genes, and $N = 1,092$ samples. We find $\hat{K} = 11$ returns highest silhouette score. The optimization results C-Index $= 0.9989$, relative error $= 13.6543\%$. Clusters in Figure 4.11A which may play important roles to survival are performed GO enrichment analysis in Table 4.16. We focus on two clusters C2 and C3. In the cluster C2, which is associated with worse survival, we find several interesting biological process terms related to macroautophagy. For example, regulation of macroautophagy (GO:0016241, P-value $= 8.84 \times 10^{-6}$), macroautography (GO:0016236, P-value $= 1.92 \times 10^{-5}$), etc.

In the contrary, cluster C3 relates to reproductive structure development (GO:0048608, P-value $= 9.31 \times 10^{-7}$), and reproductive system development (GO:0061458, P-value $= 1.12 \times 10^{-6}$), etc.

Kidney Renal Papillary Cell Carcinoma

For Kidney renal papillary cell carcinoma (KIRP), we have $P = 13,140$ filtered genes, and $N = 289$ samples. We find $\hat{K} = 10$ returns highest silhouette score. The optimization results C-Index $= 1.0$, relative error $= 9.9297\%$. Clusters in Figure 4.12A which may play important roles to survival are performed GO enrichment analysis in Table 4.17. We find cluster C8, which is associated with better survival, is highly enriched in immune related

biological process, such as regulation of immune system process (GO:0002682, P-value = 1.82×10^{-86}), positive regulation of immune system process (GO:0002684, P-value = 1.01×10^{-77}), regulation of immune response (GO:0050776, P-value = 5.14×10^{-70}), etc.

Liver Hepatocellular Carcinoma

For Liver hepatocellular carcinoma (LIHC), we have $P = 13,140$ filtered genes, and $N = 370$ samples. We find $\hat{K} = 10$ returns highest silhouette score. The optimization results C-Index = 0.9994, relative error = 12.3848%. Clusters in Figure 4.13A which may play important roles to survival are performed GO enrichment analysis in Table 4.18. We find that cluster C5, which is associated with worse survival, is enriched in alcohol metabolic process (GO:0006066, P-value = 1.29×10^{-6}) and oxidation-reduction process (GO:0055114, P-value = 2.33×10^{-6}).

Lung Adenocarcinoma

For Lung adenocarcinoma (LUAD), we have $P = 13,140$ filtered genes, and $N = 507$ samples. We find $\hat{K} = 11$ returns highest silhouette score. The optimization results C-Index = 0.9994, relative error = 10.6221%. Clusters in Figure 4.14A which may play important roles to survival are performed GO enrichment analysis in Table 4.19. We find that cluster C8 is associated with better survival, while cluster C6 is associated with worse survival. From the enrichment analysis results in Table 4.19, we identify that cluster C8 enriched in protein modification by small protein conjugation or removal (GO:0070647, P-value = 2.17×10^{-10}), regulation of catabolic process (GO:0009894, P-value = 1.64×10^{-8}), and protein modification by small protein conjugation (GO:0032446, P-value = 1.69×10^{-8}). We also identify that cluster C6 enriched in intracellular protein transport (GO:0006886, P-value = 2.01×10^{-40}), cellular macromolecule catabolic process (GO:0044265, P-value = 3.56×10^{-35}), and macromolecule catabolic process (GO:0009057, P-value = 2.53×10^{-34}).

Ovarian Serous Cystadenocarcinoma

For Ovarian serous cystadenocarcinoma (OV), we have $P = 13,140$ filtered genes, and $N = 411$ samples. We find $\hat{K} = 11$ returns highest silhouette score. The optimization results C-Index = 1.0, relative error = 10.1709%. Clusters in Figure 4.15A which may play important roles to survival are performed GO enrichment analysis in Table 4.20. We are interested in gene clusters C5 and C7, which both reflected better and worse survival. Cluster C5 is enriched in oxidation-reduction process (GO:0055114, P-value $= 6.53 \times 10^{-11}$), carbohydrate derivative metabolic process (GO:1901135, P-value $= 5.99 \times 10^{-10}$), and intracellular protein transport (GO:0006886, P-value $= 2.98 \times 10^{-9}$).

Immune related responses are highly enriched in gene cluster C7. For example, regulation of immune system process (GO:0002682, P-value $= 2.00 \times 10^{-56}$), positive regulation of immune system process (GO:0002684, P-value $= 6.21 \times 10^{-50}$), inflammatory response (GO:0006954, P-value $= 5.38 \times 10^{-48}$), immune effector process (GO:0002252, P-value $= 1.01 \times 10^{-43}$), etc.

These findings demonstrate that CoxNMF can unravel survival associated gene clusters precisely, which can greatly help researchers identify cancer-specific and survival-related gene modules as well as critical gene signatures.

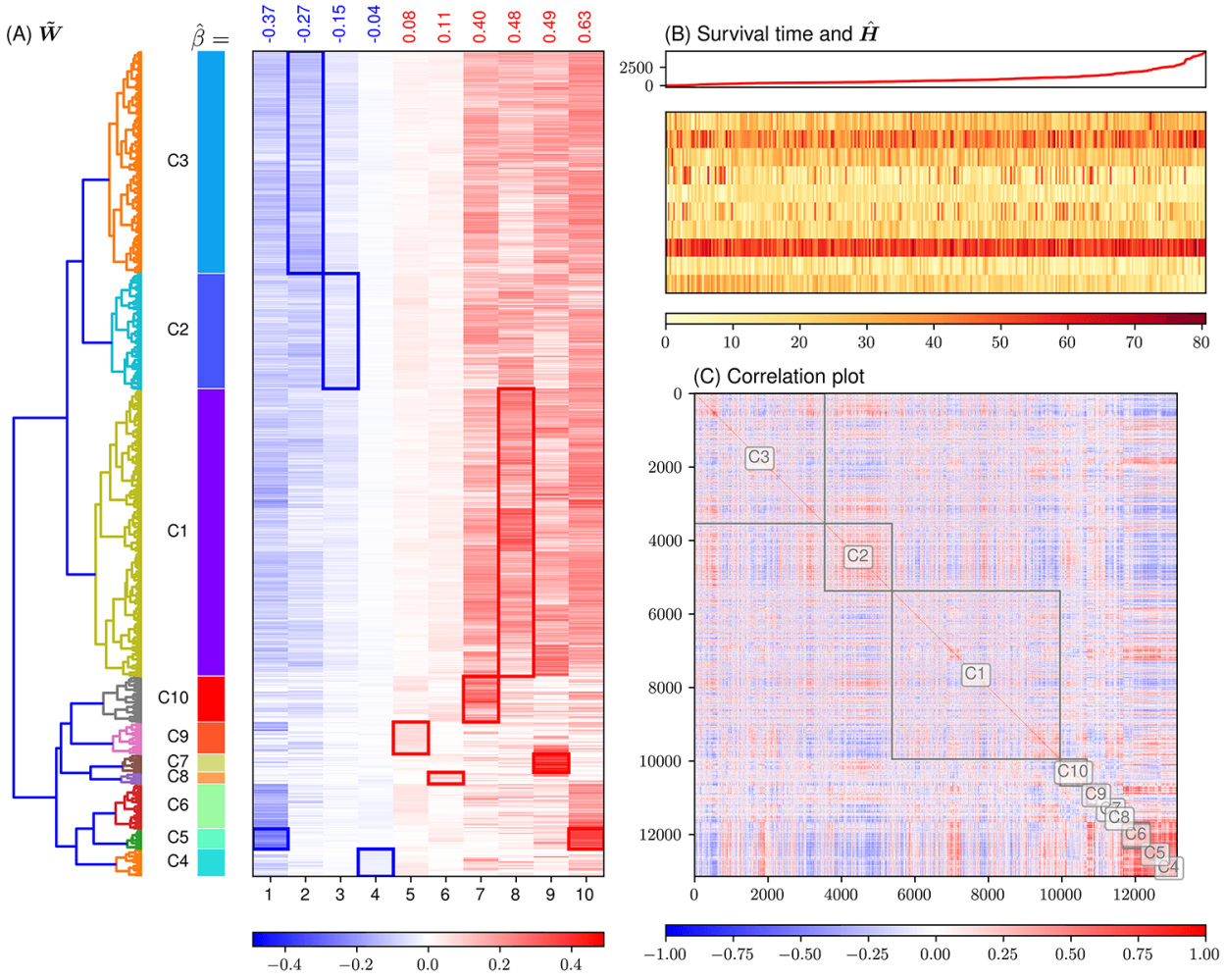


Figure 4.6. Experimental results on Colon Adenocarcinoma (COAD). (A) hierarchical agglomerative clustering with $\hat{K} = 11$ labels and the derived $\tilde{\boldsymbol{W}}$ are sorted by $\hat{\beta}$ in columns. $\hat{\phi}_-$ and $\hat{\phi}_+$ are highlighted in blue and red rectangles, respectively. (B) Survival time, and the corresponding $\hat{\boldsymbol{H}}$ are sorted by survival time in columns, and are sorted by $\hat{\beta}$ in rows. (C) Spearman's rank correlation plot of $\boldsymbol{X}$. Cluster labels are highlighted and resides at the block diagonal. Rectangles in blue and red colors indicate the true location on $\boldsymbol{X}$ with respect to the clusters which are positively/negatively associated with survival, respectively.

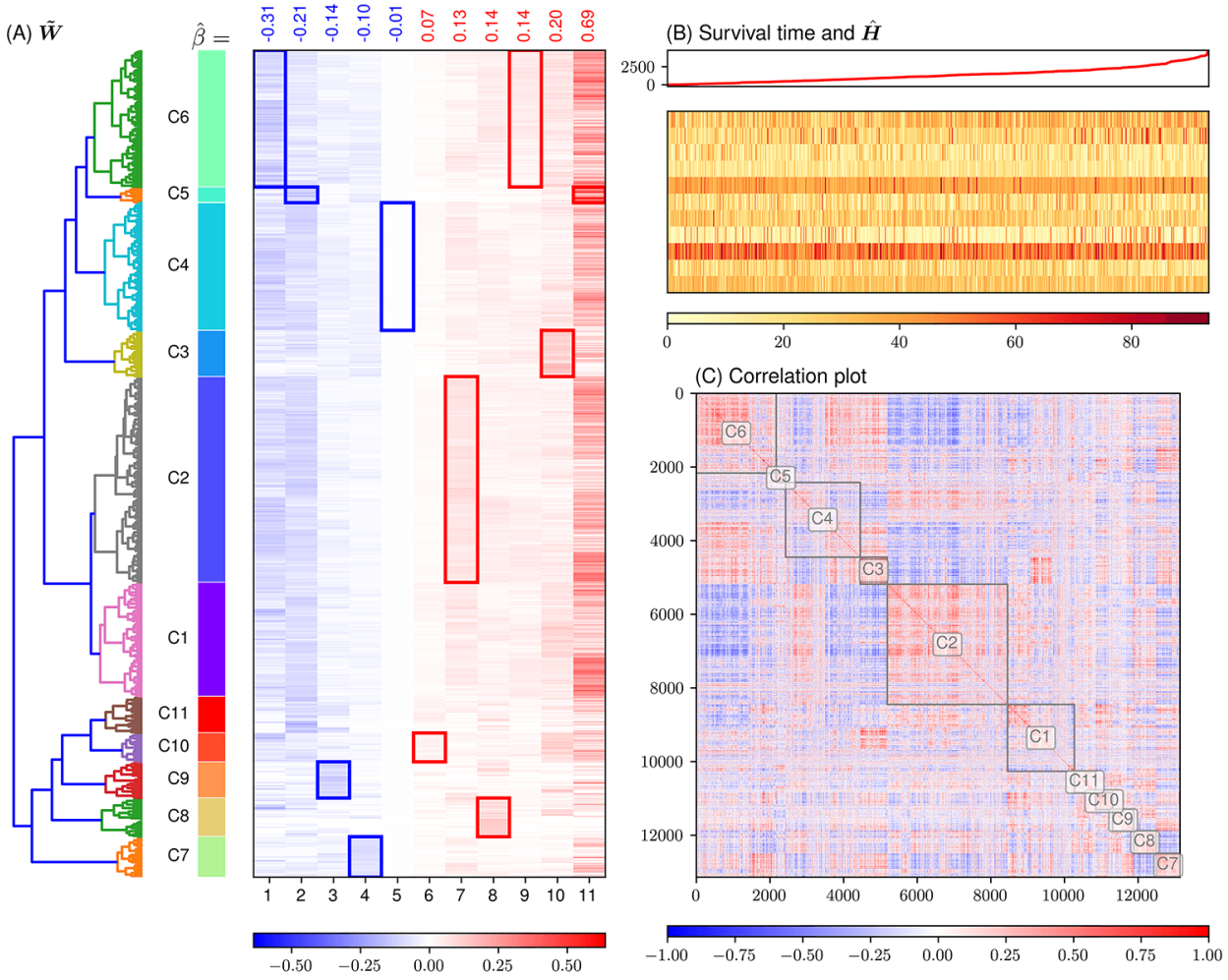


Figure 4.7. Experimental results on Kidney Renal Clear Cell Carcinoma (KIRC). (A) hierarchical agglomerative clustering with $\hat{K} = 11$ labels and the derived $\tilde{\boldsymbol{W}}$ are sorted by $\hat{\beta}$ in columns. $\hat{\phi}_-$ and $\hat{\phi}_+$ are highlighted in blue and red rectangles. (B) Survival time, and the corresponding $\hat{\boldsymbol{H}}$ are sorted by survival time in columns, and are sorted by $\hat{\beta}$ in rows. (C) Spearman's rank correlation plot of $\boldsymbol{X}$. Cluster labels are highlighted and resides at the block diagonal. Rectangles in blue and red colors indicate the true location on $\boldsymbol{X}$ with respect to the clusters which are positively/negatively associated with survival, respectively.

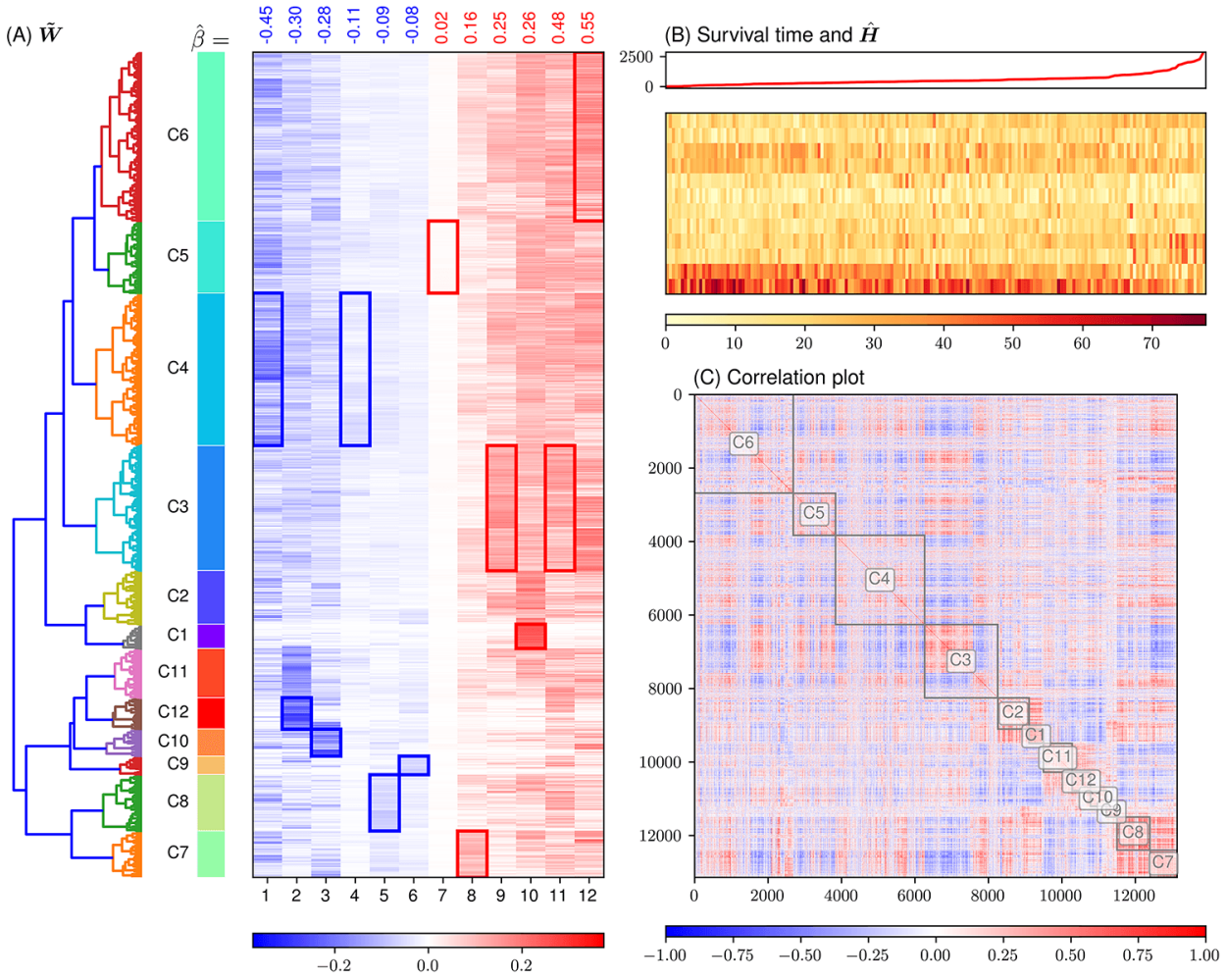


Figure 4.8. Experimental results on Pancreatic Adenocarcinoma (PAAD). (A) hierarchical agglomerative clustering with $\hat{K} = 11$ labels and the derived $\tilde{\boldsymbol{W}}$ are sorted by $\hat{\beta}$ in columns. $\hat{\phi}_-$ and $\hat{\phi}_+$ are highlighted in blue and red rectangles. (B) Survival time, and the corresponding $\hat{\boldsymbol{H}}$ are sorted by survival time in columns, and are sorted by $\hat{\beta}$ in rows. (C) Spearman's rank correlation plot of $\boldsymbol{X}$. Cluster labels are highlighted and resides at the block diagonal. Rectangles in blue and red colors indicate the true location on $\boldsymbol{X}$ with respect to the clusters which are positively/negatively associated with survival, respectively.

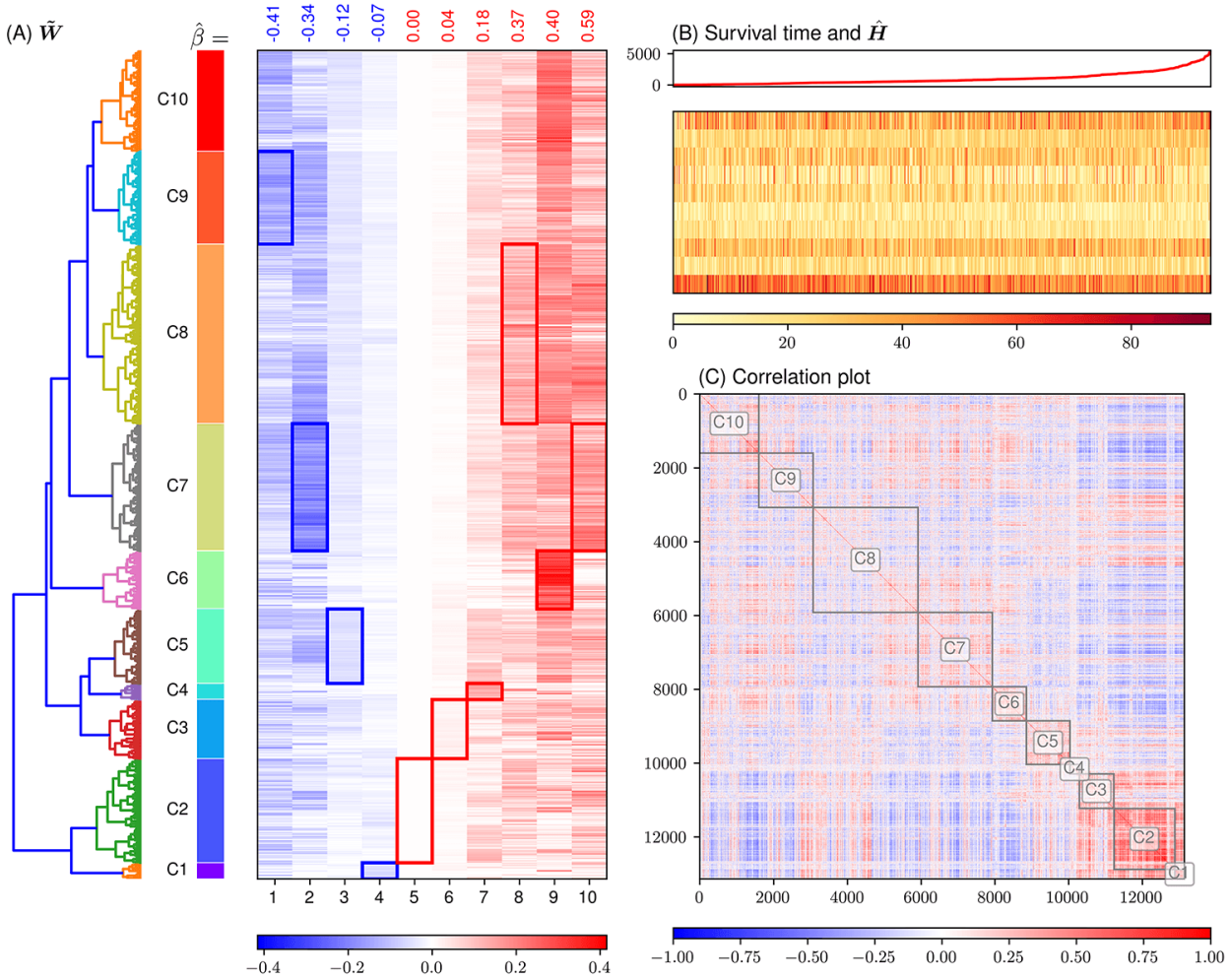


Figure 4.9. Experimental results on Lung squamous cell carcinoma (LUSC). (A) hierarchical agglomerative clustering with $\hat{K} = 11$ labels and the derived $\tilde{\boldsymbol{W}}$ are sorted by $\hat{\beta}$ in columns. $\hat{\phi}_-$ and $\hat{\phi}_+$ are highlighted in blue and red rectangles, respectively. (B) Survival time, and the corresponding $\hat{\boldsymbol{H}}$ are sorted by survival time in columns, and are sorted by $\hat{\beta}$ in rows. (C) Spearman's rank correlation plot of $\boldsymbol{X}$. Cluster labels are highlighted and resides at the block diagonal. Rectangles in blue and red colors indicate the true location on $\boldsymbol{X}$ with respect to the clusters which are positively/negatively associated with survival, respectively.

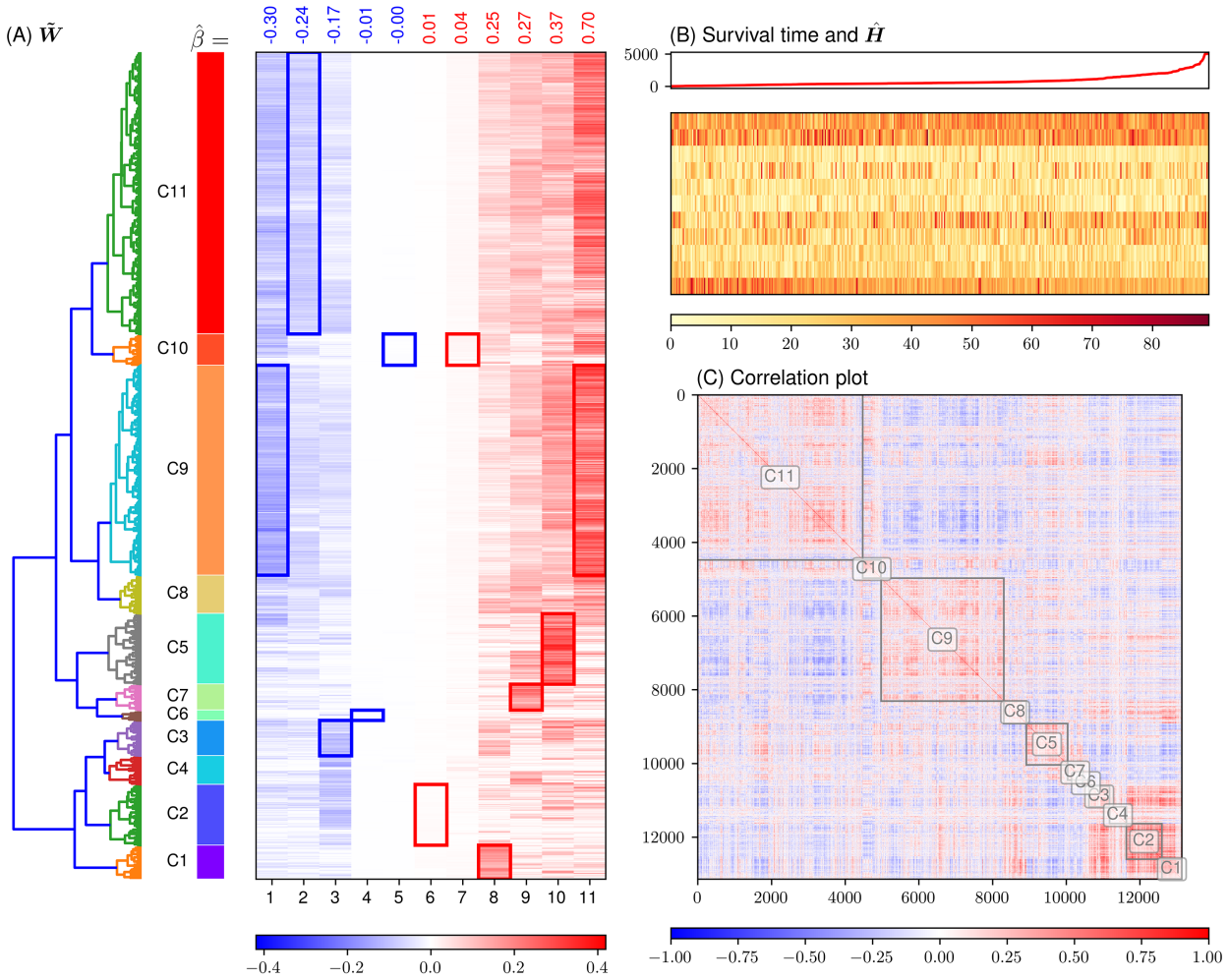


Figure 4.10. Experimental results on Bladder urothelial carcinoma (BLCA). (A) hierarchical agglomerative clustering with $\hat{K} = 11$ labels and the derived $\tilde{\boldsymbol{W}}$ are sorted by $\hat{\beta}$ in columns. $\hat{\phi}_-$ and $\hat{\phi}_+$ are highlighted in blue and red rectangles, respectively. (B) Survival time, and the corresponding $\hat{\boldsymbol{H}}$ are sorted by survival time in columns, and are sorted by $\hat{\beta}$ in rows. (C) Spearman's rank correlation plot of $\boldsymbol{X}$. Cluster labels are highlighted and resides at the block diagonal. Rectangles in blue and red colors indicate the true location on $\boldsymbol{X}$ with respect to the clusters which are positively/negatively associated with survival, respectively.

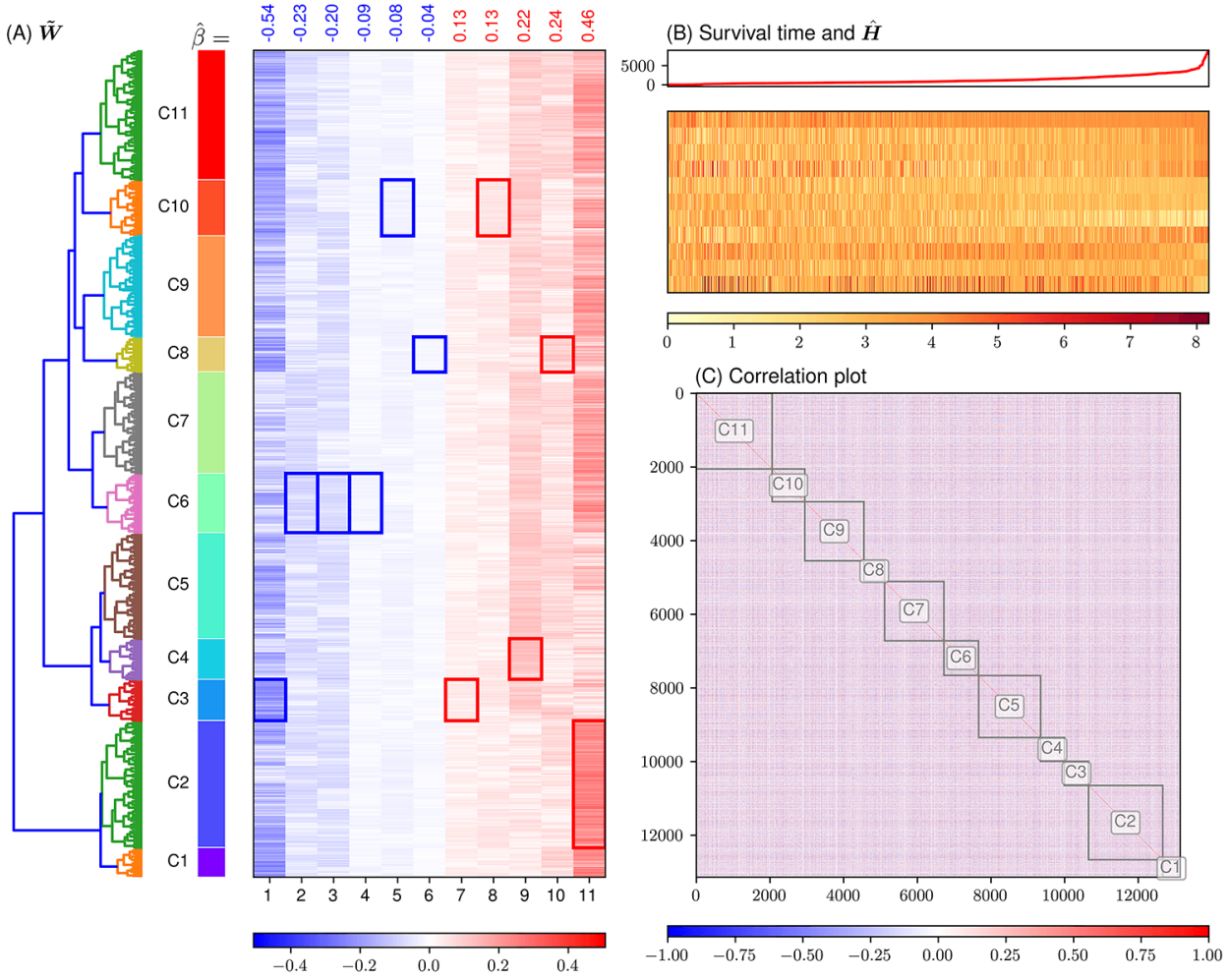


Figure 4.11. Experimental results on Breast Invasive Carcinoma (BRCA). (A) hierarchical agglomerative clustering with $\hat{K} = 11$ labels and the derived $\tilde{\boldsymbol{W}}$ are sorted by $\hat{\beta}$ in columns. $\hat{\phi}_-$ and $\hat{\phi}_+$ are highlighted in blue and red rectangles, respectively. (B) Survival time, and the corresponding $\hat{\boldsymbol{H}}$ are sorted by survival time in columns, and are sorted by $\hat{\beta}$ in rows. (C) Spearman's rank correlation plot of $\boldsymbol{X}$. Cluster labels are highlighted and resides at the block diagonal. Rectangles in blue and red colors indicate the true location on $\boldsymbol{X}$ with respect to the clusters which are positively/negatively associated with survival, respectively.

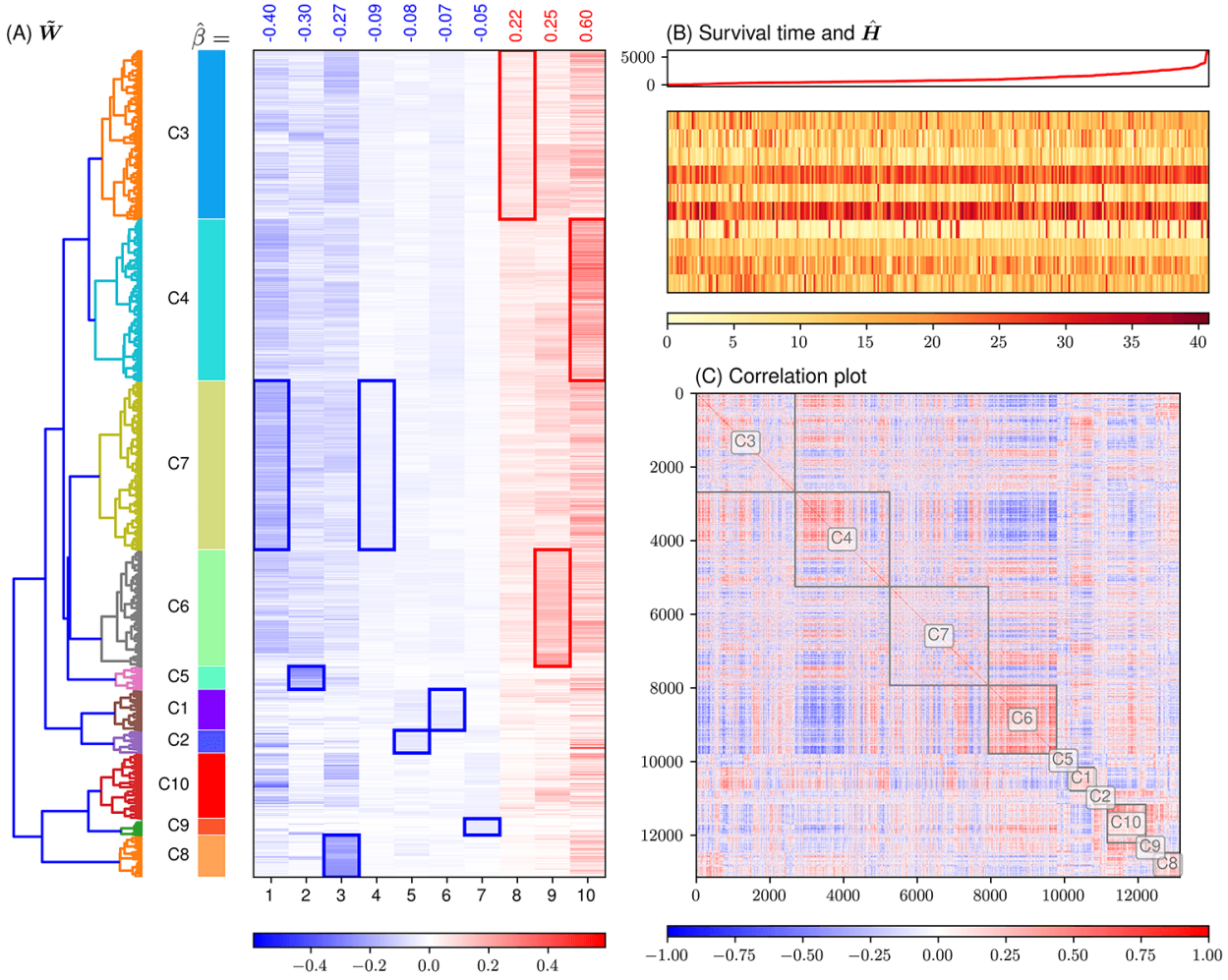


Figure 4.12. Experimental results on Kidney renal papillary cell carcinoma (KIRP). (A) hierarchical agglomerative clustering with $\hat{K} = 11$ labels and the derived $\tilde{\boldsymbol{W}}$ are sorted by $\hat{\beta}$ in columns. $\hat{\phi}_-$ and $\hat{\phi}_+$ are highlighted in blue and red rectangles. (B) Survival time, and the corresponding $\hat{\boldsymbol{H}}$ are sorted by survival time in columns, and are sorted by $\hat{\beta}$ in rows. (C) Spearman's rank correlation plot of $\boldsymbol{X}$. Cluster labels are highlighted and resides at the block diagonal. Rectangles in blue and red colors indicate the true location on $\boldsymbol{X}$ with respect to the clusters which are positively/negatively associated with survival, respectively.

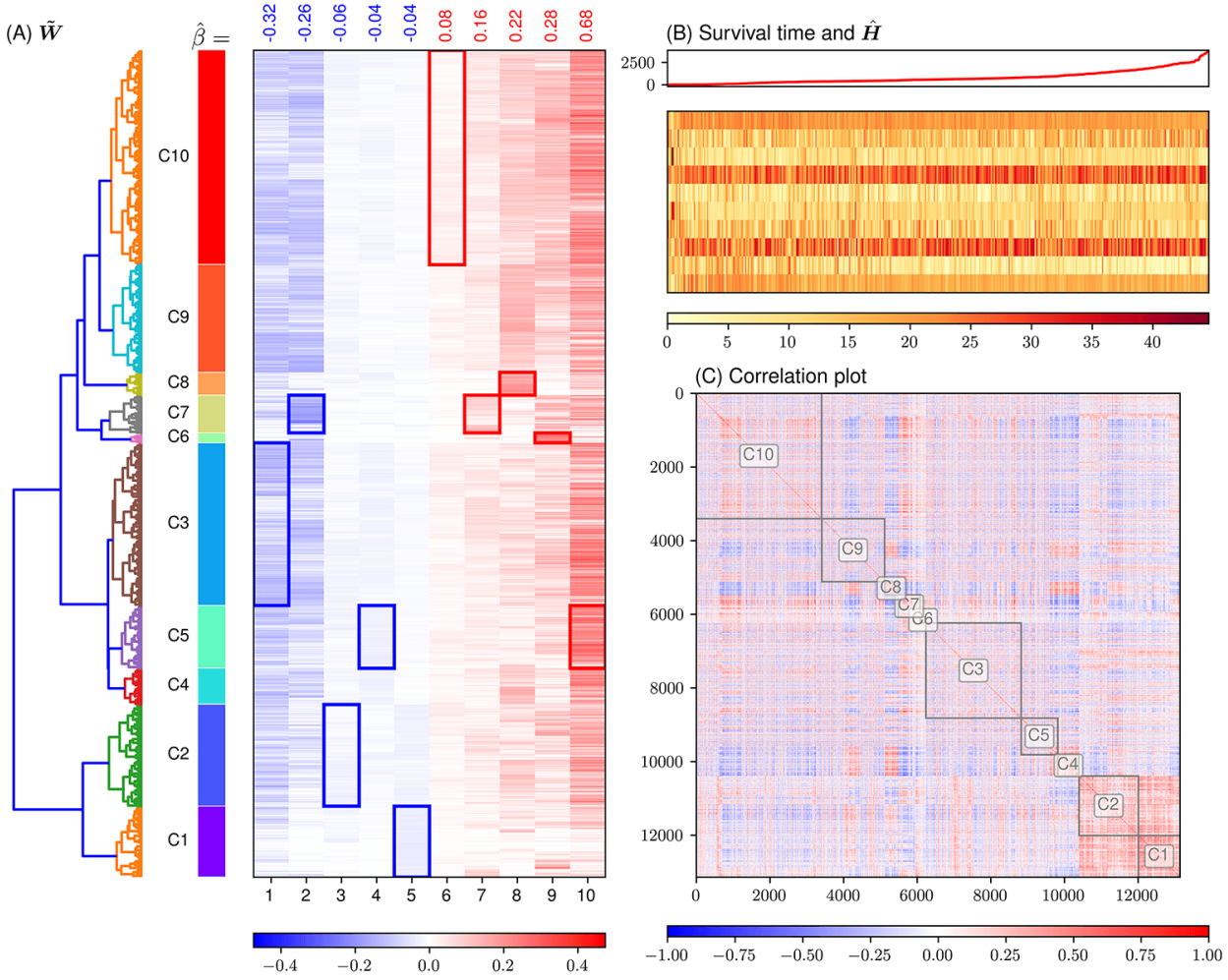


Figure 4.13. Experimental results on Liver hepatocellular carcinoma (LIHC). (A) hierarchical agglomerative clustering with $\hat{K} = 11$ labels and the derived $\tilde{\boldsymbol{W}}$ are sorted by $\hat{\beta}$ in columns. $\hat{\phi}_-$ and $\hat{\phi}_+$ are highlighted in blue and red rectangles, respectively. (B) Survival time, and the corresponding $\hat{\boldsymbol{H}}$ are sorted by survival time in columns, and are sorted by $\hat{\beta}$ in rows. (C) Spearman's rank correlation plot of $\boldsymbol{X}$. Cluster labels are highlighted and resides at the block diagonal. Rectangles in blue and red colors indicate the true location on $\boldsymbol{X}$ with respect to the clusters which are positively/negatively associated with survival, respectively.

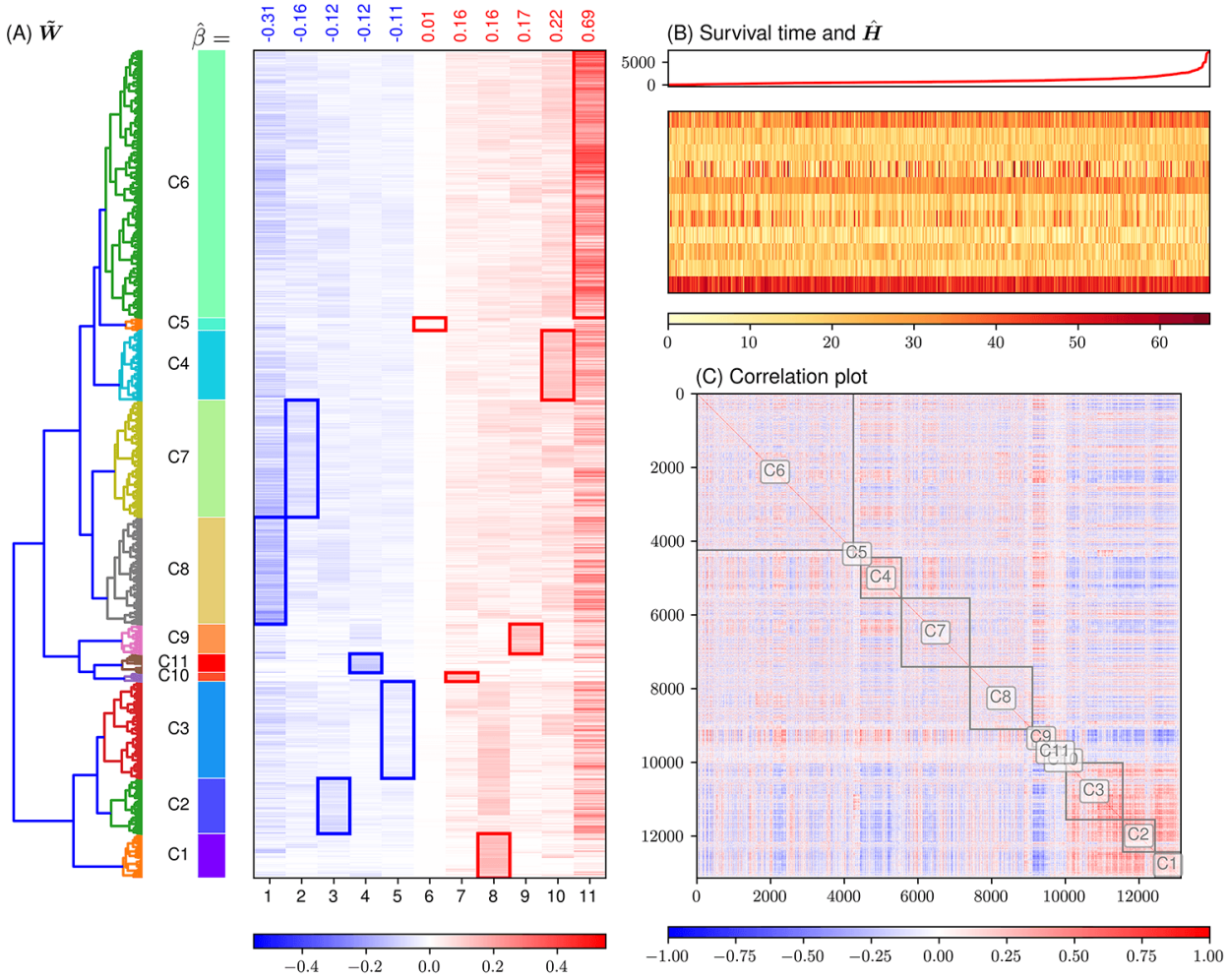


Figure 4.14. Experimental results on Lung Adenocarcinoma (LUAD). (A) hierarchical agglomerative clustering with $\hat{K} = 11$ labels and the derived $\tilde{\boldsymbol{W}}$ are sorted by $\hat{\beta}$ in columns. $\hat{\phi}_-$ and $\hat{\phi}_+$ are highlighted in blue and red rectangles, respectively. (B) Survival time, and the corresponding $\hat{\boldsymbol{H}}$ are sorted by survival time in columns, and are sorted by $\hat{\beta}$ in rows. (C) Spearman's rank correlation plot of $\boldsymbol{X}$. Cluster labels are highlighted and resides at the block diagonal. Rectangles in blue and red colors indicate the true location on $\boldsymbol{X}$ with respect to the clusters which are positively/negatively associated with survival, respectively.

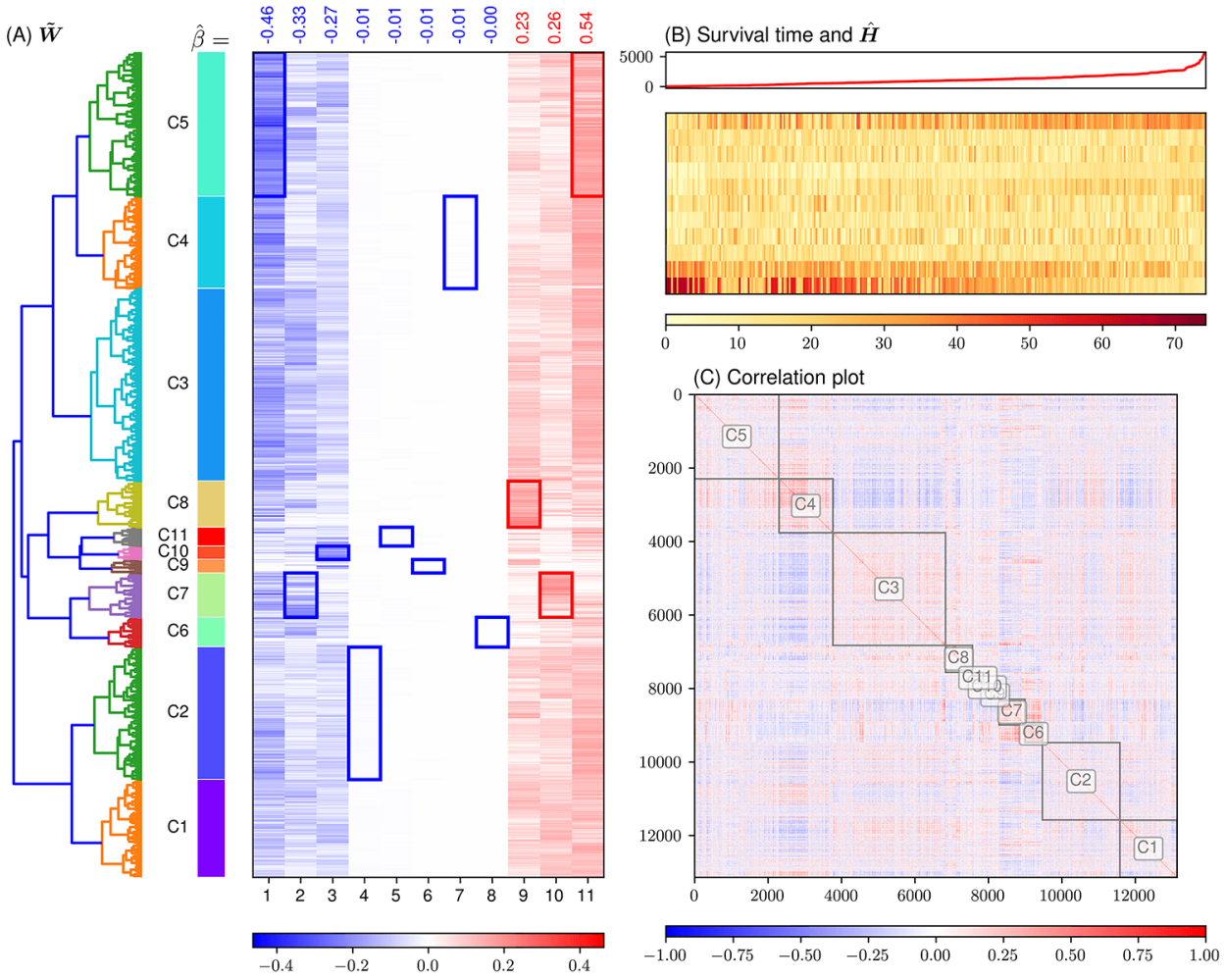


Figure 4.15. Experimental results on Ovarian serous cystadenocarcinoma (OV). (A) hierarchical agglomerative clustering with $\hat{K} = 11$ labels and the derived $\tilde{\boldsymbol{W}}$ are sorted by $\hat{\beta}$ in columns. $\hat{\phi}_-$ and $\hat{\phi}_+$ are highlighted in blue and red rectangles, respectively. (B) Survival time, and the corresponding $\hat{\boldsymbol{H}}$ are sorted by survival time in columns, and are sorted by $\hat{\beta}$ in rows. (C) Spearman's rank correlation plot of $\boldsymbol{X}$. Cluster labels are highlighted and resides at the block diagonal. Rectangles in blue and red colors indicate the true location on $\boldsymbol{X}$ with respect to the clusters which are positively/negatively associated with survival, respectively.

Table 4.11. Gene ontology (GO) enrichment analysis results for COAD. Top ranked GO terms and important GO terms are reported according to the P-values, which are associated with better/worse survival prognosis.

Gene cluster $C7$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0016266	O-glycan processing.	1.73×10^{-12}	6.97×10^{-09}
2	GO:0007586	Digestion.	6.00×10^{-11}	1.21×10^{-07}
3	GO:0030277	Maintenance of gastrointestinal epithelium.	2.01×10^{-10}	2.71×10^{-07}
4	GO:0010669	Epithelial structure maintenance.	3.86×10^{-09}	3.46×10^{-06}
6	GO:0022600	Digestive system process.	8.20×10^{-09}	5.06×10^{-06}
Gene cluster $C5$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0045321	Leukocyte activation.	6.62×10^{-78}	3.11×10^{-74}
2	GO:0001775	Cell activation.	2.04×10^{-77}	4.80×10^{-74}
3	GO:0002682	Regulation of immune system process.	9.51×10^{-75}	1.49×10^{-71}
4	GO:0050776	Regulation of immune response.	1.24×10^{-65}	1.46×10^{-62}
6	GO:0002684	Positive regulation of immune system process.	9.99×10^{-61}	7.83×10^{-58}
7	GO:0046649	Lymphocyte activation.	3.48×10^{-57}	2.34×10^{-54}

4.10 Discussion

In this work, we propose CoxNMF algorithm, which enables low-rank representation analysis and Cox regression simultaneously. In the simulation results, we show that the CoxNMF is able to achieve higher accuracy and Dice coefficient. The corresponding C-Index is also robust to the noise. The relative error is competitive and the running time is relative efficient. Different from other gene network mining algorithms such as WGCNA [35], lmQCM [34], or other low-rank approaches such as NMF or PCA, CoxNMF not only demonstrates the ability to unravel survival related gene clusters, but also helps to decode how the target gene clusters are associated with survival.

In the TCGA cancer results, we demonstrate that the CoxNMF algorithm can unravel important gene modules that are associated with survival. Among the TCGA cancer results and the downstream gene ontology analysis, we find that immune system related biological processes play important roles among cancers. For example, immune system related biolog-

Table 4.12. Gene ontology (GO) enrichment analysis results for KIRC. Top ranked GO terms and important GO terms are reported according to the P-values, which are associated with better/worse survival prognosis.

Gene cluster $C5$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0006082	Organic acid metabolic process.	4.61×10^{-36}	1.73×10^{-32}
2	GO:0019752	Carboxylic acid metabolic process.	2.02×10^{-30}	2.53×10^{-27}
3	GO:0032787	Monocarboxylic acid metabolic process.	1.19×10^{-21}	1.12×10^{-18}
33	GO:0046415	Urate metabolic process.	4.19×10^{-09}	4.76×10^{-07}
36	GO:0006063	Uronic acid metabolic process.	8.82×10^{-09}	9.18×10^{-07}
53	GO:0015747	Urate transport.	2.84×10^{-07}	2.01×10^{-05}
Gene cluster $C6$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0006518	Peptide metabolic process.	3.94×10^{-25}	3.71×10^{-21}
2	GO:0006605	Protein targeting.	1.13×10^{-23}	5.32×10^{-20}
3	GO:0043603	Cellular amide metabolic process.	1.84×10^{-23}	5.79×10^{-20}
Gene cluster $C9$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0043062	Extracellular structure organization.	1.63×10^{-23}	9.86×10^{-20}
2	GO:0030198	Extracellular matrix organization.	3.24×10^{-21}	9.79×10^{-18}
3	GO:0006953	Acute-phase response.	1.61×10^{-17}	3.24×10^{-14}
4	GO:0002526	Acute inflammatory response.	3.79×10^{-16}	5.72×10^{-13}
14	GO:0006959	Humoral immune response.	2.57×10^{-09}	1.11×10^{-06}
25	GO:0006954	Inflammatory response.	1.55×10^{-08}	3.75×10^{-06}

ical processes are highly enriched in cluster C5 in Colon adenocarcinoma (COAD), which is both positively and negatively associated with survival. Similarly, cluster C9 in Kidney renal clear cell carcinoma (KIRC) also highly enriched with several immune response related biological processes, it is also positively associated with survival. In Kidney renal papillary cell carcinoma (KIRP) and Ovarian serous cystadenocarcinoma (OV), clusters which is positively associated with survival can be found highly enriched with immune system processes as well as inflammatory responses. It is well known that the immune response is positively correlated with survival. For example, Moller *et al.* [254] listed twelve literature which showed the statistical test results between immune response and survival. Nevertheless, House and Watt [255] studied 107 colorectum carcinoma patients and showed that there was a Chi-squared difference in survival at three years between immune patients and the non-immune patients.

Our study in COAD further confirms the importance of immune related biological process, and demonstrates that those patients who expressed higher immune related genes, would generally result in longer survival times ($\hat{\beta} > 0$).

The identified gene clusters from CoxNMF results are also enriched in some cancer-specific biological processes. For example, in COAD, gene cluster C7 which is associated with worse survival, is enriched in digestion and digestive system related processes. As COAD is one of the digestive cancers [256], genes which are related to the digestion would be extremely helpful to further study the survival of the COAD cancer. Moreover, cluster C7 is also related to the maintenance of gastrointestinal epithelium [257].

In KIRC, gene cluster C5 is enriched with urate metabolic process, uronic acid metabolic process, and urate transport. It has been found that the uric acid metabolism is mainly related to kidney and electrolyte disorders [258].

In Pancreatic adenocarcinoma (PAAD), gene cluster C1 which is highly enriched with potassium ion and insulin secretion, is associated with worse survival. However, even cluster C1 is associated with shorter survival (since $\hat{\beta} > 0$), there are few samples which has longer survival time in row 10 of $\hat{\boldsymbol{H}}$ are highly valued. We believe that gene cluster C1 is associated with worse survival, but few samples are with longer survival times. These findings confirm previous literature that potassium ion transport pathways are found to be significantly enriched in PAAD [259], [260]. Similarly, for insulin secretion, Gullo *et al.* [261] found that most pancreatic cancer patients had higher insulin secretion response than non-pancreatic cancer patients or healthy controls. Meanwhile in cluster C12, since the association between pancreatic cancer and diabetes was well recognized [261], we further confirm that higher expressions of genes which are enriched with digestion and glycosylation related biological processes are associated with better survival. Furthermore, this finding also supports the idea that glycans could help guide precision medicine strategies in pancreatic cancer [262].

In Lung squamous cell carcinoma (LUSC), findings from CoxNMF results reflect rich biological interpretation and can be further validated in previous literature. For example, Ferone *et al.* [263] confirmed that the cornification was accompanied by infiltration of inflammatory cells and large areas of necrosis. Park *et al.* [264] confirmed that the keratinization of LUSC was associated with poor clinical outcome.

Nevertheless, epithelium related biological processes have been found enriched among several cancers. For example, in LUSC, epidermis development (P-value $= 2.95 \times 10^{-48}$) is enriched in gene cluster C6, associated with worse survival. In Bladder urothelial carcinoma (BLCA), both gene cluster C5 and C7 (associated with worse survival) are enriched with epithelium development.

In Breast invasive carcinoma (BRCA), macroautophagy may play important role related to breast cancer survival in gene cluster C2. Previous literature showed that the macroautophagy links to the cellular response to anticancer therapies [265], and it can associate with both life and death functions [266]. In our study, we find macroautophagy is enriched in gene cluster C2, which is associated with worse survival. In Lung Adenocarcinoma (LUAD), autophagy also plays important role in gene cluster C6 (associated with worse survival).

In Liver hepatocellular carcinoma (LIHC), gene cluster C5 is associated with worse survival and enriched with alcohol metabolic process (P-value $= 1.29 \times 10^{-6}$) and secondary alcohol biosynthetic process (P-value $= 3.13 \times 10^{-4}$), The alcohol related biological processes are proven to be associated with liver cancer [267] and can be further validated from our results.

4.11 Conclusion

In this chapter, a novel algorithm CoxNMF is proposed by simultaneously learning the non-negative matrix factorization and the Cox proportional hazards regression. We design the novel objective function and update rules for CoxNMF. To the best of our knowledge, this is the first work that performs non-negative matrix factorization and clustering driven by survival regression, accomplished by joint optimization of the Frobenius norm and partial log likelihood.

The proposed CoxNMF algorithm presents new contributions to biomedical data analysis and has several advantages:

- First, to the best of our knowledge this is the first attempt at fully utilizing and integrating survival data, gene expression matrices, and hazards information in the updating rule. Previous studies either applied NMF followed with a Cox regression separately, or

imposed linear regression onto NMF algorithm. When using NMF to find latent clusters from gene expression data, one should consider using survival information of the patient cohort as well.

- Second, CoxNMF can help find a coefficient matrix $\boldsymbol{H}$ constrained on a K-dimensional hyperspace where survival labels can be well-distinguished. Since NMF is ill-posed and non-convex [268], [269], the solution to the original NMF cost function is not unique and is strongly deponent on the updating algorithm and initialization points [270], [271]. Imposing survival information onto $\boldsymbol{H}$, helps the algorithm find solutions in a supervised manner. This can be observed from the C-Index values that approach 100% in experiments.

- Third, CoxNMF is computationally efficient and robust to noise. Our experiments using synthetic datasets show that both the running time is low while having a performance that is very competitive to that of the traditional NMF methods. Moreover, CoxNMF demonstrates robustness to noise for large values of ϵ.

Nonetheless, CoxNMF does have several limitations. First of all, when applied to real cancer datasets, resulted in a number of clusters that are relatively large in size. Such large gene clusters are too general to be defined. Thus, a further gene expression analysis on those clusters is recommended. Second, Although CoxNMF is efficient in running time, finding the hyper-parameters especially the optimal low-rank dimension K, is computationally expensive. For the TCGA human cancer datasets, we performed a hyper-parameter search that suggested a reasonable hyper-parameter pair for 10 analyzed TCGA datasets. When applying CoxNMF to other types of data, such as microarray datasets from NCBI GEO, the hyper-parameter values many need to be re-evaluated. Future analyses are recommended to inspect and address these limitations.

To sum up, the proposed CoxNMF algorithm successfully demonstrates its superiority of identifying survival-associated gene clusters than other algorithms across forty-two different synthetic data. The experiments conduct in human cancer datasets help unravel latent gene clusters which reflect rich biological interpretations, achieve the goal of understanding and interpretation of high-dimensional biological data in precision health.

In previous chapters, we focus on integrating multimodal biomedical data to unravel the latent gene interactions, study the relationship between multimodal biomedical data and patient survival. In the next chapter, we are interested in histopathologic imaging data, and two breast cancer subtypes are studied. Specifically, we would like to predict the post-treatment pathologic complete response (pCR) given the pre-treatment histopathologic images, through an automatic feature extraction workflow.

Table 4.13. Gene ontology (GO) enrichment analysis results for PAAD. Top ranked GO terms and important GO terms are reported according to the P-values, which are associated with better/worse survival prognosis.

Gene cluster $C6$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0009057	Macromolecule catabolic process.	1.50×10^{-28}	1.51×10^{-24}
2	GO:0044265	Cellular macromolecule catabolic process.	9.67×10^{-27}	4.88×10^{-23}
3	GO:0030163	Protein catabolic process.	5.06×10^{-20}	1.70×10^{-16}
Gene cluster $C1$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0099536	Synaptic signaling.	4.68×10^{-36}	7.39×10^{-33}
2	GO:0098916	Anterograde trans-synaptic signaling.	5.25×10^{-36}	7.39×10^{-33}
3	GO:0007268	Chemical synaptic transmission.	5.25×10^{-36}	7.39×10^{-33}
18	GO:0071805	Potassium ion transmembrane transport.	9.49×10^{-15}	2.23×10^{-12}
21	GO:0006813	Potassium ion transport.	2.58×10^{-14}	5.19×10^{-12}
23	GO:0030073	Insulin secretion.	5.79×10^{-14}	1.06×10^{-11}
53	GO:0050796	Regulation of insulin secretion.	2.15×10^{-10}	1.71×10^{-08}
Gene cluster $C4$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0016071	mRNA metabolic process.	4.04×10^{-30}	3.81×10^{-26}
2	GO:0006397	mRNA processing.	2.68×10^{-28}	1.27×10^{-24}
3	GO:0009057	Macromolecule catabolic process.	6.73×10^{-25}	2.12×10^{-21}
Gene cluster $C12$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0007586	Digestion.	4.69×10^{-19}	2.26×10^{-15}
2	GO:0006805	Xenobiotic metabolic process.	1.83×10^{-15}	4.42×10^{-12}
3	GO:0016266	O-glycan processing.	7.40×10^{-12}	1.19×10^{-08}
5	GO:0022600	Digestive system process.	3.08×10^{-11}	2.97×10^{-08}
7	GO:0042445	Hormone metabolic process.	7.67×10^{-10}	5.28×10^{-07}
17	GO:0006493	Protein O-linked glycosylation	4.89×10^{-08}	1.39×10^{-05}
23	GO:0043413	Macromolecule glycosylation	2.86×10^{-07}	5.75×10^{-05}
24	GO:0006486	Protein glycosylation	2.86×10^{-07}	5.75×10^{-05}
26	GO:0070085	Glycosylation	6.52×10^{-07}	1.21×10^{-04}
34	GO:0010817	Regulation of hormone levels.	3.32×10^{-06}	4.55×10^{-04}
35	GO:0052695	Cellular glucuronidation	3.35×10^{-06}	4.55×10^{-04}
37	GO:0009101	Glycoprotein biosynthetic process	4.19×10^{-06}	5.46×10^{-04}
Gene cluster $C10$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0043588	Skin development.	6.21×10^{-23}	9.15×10^{-20}
2	GO:0000070	Mitotic sister chromatid segregation.	6.67×10^{-23}	9.15×10^{-20}
3	GO:0008544	Epidermis development.	7.80×10^{-23}	9.15×10^{-20}

Table 4.14. Gene ontology (GO) enrichment analysis results for LUSC. Top ranked GO terms and important GO terms are reported according to the P-values, which are associated with better/worse survival prognosis.

Gene cluster $C6$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0008544	Epidermis development.	2.95×10^{-48}	2.02×10^{-44}
2	GO:0043588	Skin development.	3.25×10^{-47}	1.11×10^{-43}
3	GO:1903575	Cornified envelope assembly.	2.33×10^{-46}	5.32×10^{-43}
4	GO:0070268	Cornification.	4.16×10^{-46}	7.13×10^{-43}
7	GO:0060429	Epithelium development.	2.62×10^{-40}	2.57×10^{-37}
8	GO:0030855	Epithelial cell differentiation.	4.63×10^{-37}	3.97×10^{-34}
9	GO:0009913	Epidermal cell differentiation.	3.56×10^{-36}	2.71×10^{-33}
10	GO:0030216	Keratinocyte differentiation.	2.98×10^{-35}	2.04×10^{-32}
12	GO:0031424	Keratinization.	1.44×10^{-31}	8.23×10^{-29}
Gene cluster $C7$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0140053	Mitochondrial gene expression.	7.56×10^{-39}	6.24×10^{-35}
2	GO:0032543	Mitochondrial translation.	3.18×10^{-34}	1.31×10^{-30}
3	GO:0006414	Translational elongation.	1.19×10^{-32}	2.46×10^{-29}
Gene cluster $C9$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0009057	Macromolecule catabolic process.	1.57×10^{-19}	7.13×10^{-16}
2	GO:0044265	Cellular macromolecule catabolic process.	1.75×10^{-19}	7.13×10^{-16}
3	GO:0070647	Protein modification by small protein conjugation or removal.	3.06×10^{-19}	8.33×10^{-16}

Table 4.15. Gene ontology (GO) enrichment analysis results for BLCA. Top ranked GO terms and important GO terms are reported according to the P-values, which are associated with better/worse survival prognosis.

Gene cluster $C9$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0070647	Protein modification by small protein conjugation or removal.	1.02×10^{-51}	1.05×10^{-47}
2	GO:0032446	Protein modification by small protein conjugation.	6.60×10^{-40}	3.41×10^{-36}
3	GO:0009894	Regulation of catabolic process.	4.71×10^{-37}	1.62×10^{-33}
Gene cluster $C5$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0006629	Lipid metabolic process.	6.12×10^{-14}	4.26×10^{-10}
2	GO:0030855	Epithelial cell differentiation.	5.76×10^{-11}	1.38×10^{-07}
3	GO:0048871	Multicellular organismal homeostasis.	5.95×10^{-11}	1.38×10^{-07}
8	GO:0010817	Regulation of hormone levels.	6.03×10^{-09}	5.00×10^{-06}
9	GO:0060429	Epithelium development.	6.48×10^{-09}	5.00×10^{-06}
30	GO:0007586	Digestion.	2.76×10^{-06}	6.40×10^{-04}
31	GO:0008544	Epidermis development.	2.88×10^{-06}	6.45×10^{-04}
47	GO:0042445	Hormone metabolic process.	1.13×10^{-05}	1.67×10^{-03}
Gene cluster $C7$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0060429	Epithelium development.	5.81×10^{-17}	3.04×10^{-13}
2	GO:0030855	Epithelial cell differentiation.	9.98×10^{-15}	2.61×10^{-11}
3	GO:0043588	Skin development.	5.83×10^{-13}	1.02×10^{-09}
4	GO:0008544	Epidermis development.	1.27×10^{-12}	1.67×10^{-09}
6	GO:0042445	Hormone metabolic process.	3.36×10^{-10}	1.98×10^{-07}
7	GO:0098773	Skin epidermis development.	3.40×10^{-10}	1.98×10^{-07}
11	GO:0034754	Cellular hormone metabolic process.	1.07×10^{-09}	5.09×10^{-07}
23	GO:0045682	Regulation of epidermis development.	1.51×10^{-07}	3.44×10^{-05}
34	GO:0002064	Epithelial cell development.	6.97×10^{-07}	1.07×10^{-04}
38	GO:0002009	Morphogenesis of an epithelium.	9.18×10^{-07}	1.25×10^{-04}

Table 4.16. Gene ontology (GO) enrichment analysis results for BRCA. Top ranked GO terms and important GO terms are reported according to the P-values, which are associated with better/worse survival prognosis.

Gene cluster $C2$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0009057	Macromolecule catabolic process.	4.06×10^{-09}	3.82×10^{-05}
2	GO:0044265	Cellular macromolecule catabolic process.	3.97×10^{-08}	1.87×10^{-04}
3	GO:0006886	Intracellular protein transport.	1.34×10^{-07}	4.20×10^{-04}
6	GO:0016241	Regulation of macroautophagy.	8.84×10^{-06}	1.09×10^{-02}
13	GO:0016236	Macroautophagy.	1.92×10^{-05}	1.09×10^{-02}
Gene cluster $C3$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0048608	Reproductive structure development.	9.31×10^{-07}	3.56×10^{-03}
2	GO:0061458	Reproductive system development.	1.12×10^{-06}	3.56×10^{-03}

Table 4.17. Gene ontology (GO) enrichment analysis results for KIRP. Top ranked GO terms and important GO terms are reported according to the P-values, which are associated with better/worse survival prognosis.

Gene cluster $C4$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0006886	Intracellular protein transport.	8.97×10^{-33}	8.47×10^{-29}
2	GO:0016071	mRNA metabolic process.	5.54×10^{-31}	2.62×10^{-27}
3	GO:0044265	Cellular macromolecule catabolic process.	1.52×10^{-29}	4.77×10^{-26}
Gene cluster $C6$				
1	GO:0070647	Protein modification by small protein conjugation or removal.	3.20×10^{-22}	2.84×10^{-18}
2	GO:0051276	Chromosome organization.	2.51×10^{-21}	1.12×10^{-17}
3	GO:0032446	Protein modification by small protein conjugation.	1.63×10^{-15}	4.83×10^{-12}
Gene cluster $C7$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0016071	mRNA metabolic process.	5.23×10^{-29}	5.04×10^{-25}
2	GO:0006397	mRNA processing.	4.29×10^{-21}	1.34×10^{-17}
3	GO:0006886	Intracellular protein transport.	4.40×10^{-21}	1.34×10^{-17}
Gene cluster $C5$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0006082	Organic acid metabolic process.	6.74×10^{-44}	2.99×10^{-40}
2	GO:0019752	Carboxylic acid metabolic process.	5.30×10^{-39}	1.17×10^{-35}
3	GO:0043436	Oxoacid metabolic process.	1.22×10^{-38}	1.80×10^{-35}
4	GO:0044282	Small molecule catabolic process.	1.38×10^{-34}	1.53×10^{-31}
5	GO:0016054	Organic acid catabolic process.	7.51×10^{-32}	5.55×10^{-29}
Gene cluster $C8$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0001775	Cell activation.	1.23×10^{-98}	7.83×10^{-95}
2	GO:0045321	Leukocyte activation.	4.37×10^{-95}	1.39×10^{-91}
3	GO:0002682	Regulation of immune system process.	1.82×10^{-86}	3.86×10^{-83}
4	GO:0006952	Defense response.	8.77×10^{-79}	1.39×10^{-75}
5	GO:0002684	Positive regulation of immune system process.	1.01×10^{-77}	1.28×10^{-74}
6	GO:0050776	Regulation of immune response.	5.14×10^{-70}	5.43×10^{-67}
7	GO:0002252	Immune effector process.	2.67×10^{-68}	2.42×10^{-65}
12	GO:0098542	Defense response to other organism.	3.41×10^{-59}	1.80×10^{-56}
13	GO:0002443	Leukocyte mediated immunity.	1.92×10^{-58}	9.38×10^{-56}
14	GO:0046649	Lymphocyte activation.	1.76×10^{-55}	7.59×10^{-53}
15	GO:0006954	Inflammatory response.	1.79×10^{-55}	7.59×10^{-53}
16	GO:0002366	Leukocyte activation involved in immune response.	1.86×10^{-53}	7.38×10^{-51}
17	GO:0002263	Cell activation involved in immune response.	3.45×10^{-53}	1.29×10^{-50}
18	GO:0042110	T-cell activation.	1.09×10^{-52}	3.84×10^{-50}
19	GO:0045087	Innate immune response.	4.90×10^{-52}	1.64×10^{-49}
20	GO:0050778	Positive regulation of immune response.	8.71×10^{-51}	2.76×10^{-48}

Table 4.18. Gene ontology (GO) enrichment analysis results for LIHC. Top ranked GO terms and important GO terms are reported according to the P-values, which are associated with better/worse survival prognosis.

Gene cluster $C5$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0009057	Macromolecule catabolic process.	4.79×10^{-08}	3.55×10^{-04}
2	GO:1901565	Organonitrogen compound catabolic process.	4.22×10^{-07}	9.67×10^{-04}
3	GO:0007034	Vacuolar transport.	5.07×10^{-07}	9.67×10^{-04}
4	GO:0030163	Protein catabolic process.	5.21×10^{-07}	9.67×10^{-04}
6	GO:0006066	Alcohol metabolic process.	1.29×10^{-06}	1.59×10^{-03}
7	GO:0055114	Oxidation-reduction process.	2.33×10^{-06}	2.47×10^{-03}
60	GO:1902653	Secondary alcohol biosynthetic process.	3.13×10^{-04}	3.87×10^{-02}
Gene cluster $C6$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0044278	Cell wall disruption in other organism.	1.02×10^{-06}	2.22×10^{-03}
2	GO:0007506	Gonadal mesoderm development.	1.40×10^{-05}	1.53×10^{-02}
Gene cluster $C3$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0016071	mRNA metabolic process.	6.32×10^{-21}	6.07×10^{-17}
2	GO:0044265	Cellular macromolecule catabolic process.	4.27×10^{-18}	2.05×10^{-14}
3	GO:0006886	Intracellular protein transport.	2.51×10^{-17}	7.33×10^{-14}
6	GO:0009057	Macromolecule catabolic process.	1.71×10^{-16}	2.74×10^{-13}
8	GO:0009894	Regulation of catabolic process.	1.89×10^{-14}	2.27×10^{-11}

Table 4.19. Gene ontology (GO) enrichment analysis results for LUAD. Top ranked GO terms and important GO terms are reported according to the P-values, which are associated with better/worse survival prognosis.

Gene cluster $C6$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0006886	Intracellular protein transport.	2.01×10^{-40}	2.29×10^{-36}
2	GO:0044265	Cellular macromolecule catabolic process.	3.56×10^{-35}	2.03×10^{-31}
3	GO:0009057	Macromolecule catabolic process.	2.53×10^{-34}	9.60×10^{-31}
31	GO:0006914	Autophagy.	1.13×10^{-18}	4.16×10^{-16}
Gene cluster $C4$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0006414	Translational elongation.	3.79×10^{-35}	1.12×10^{-31}
2	GO:0006412	Translation.	3.79×10^{-35}	1.12×10^{-31}
3	GO:0043043	Peptide biosynthetic process.	1.56×10^{-33}	3.06×10^{-30}
Gene cluster $C8$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0070647	Protein modification by small protein conjugation or removal.	2.17×10^{-10}	1.82×10^{-06}
2	GO:0009894	Regulation of catabolic process.	1.64×10^{-08}	3.35×10^{-05}
3	GO:0032446	Protein modification by small protein conjugation.	1.69×10^{-08}	3.35×10^{-05}

Table 4.20. Gene ontology (GO) enrichment analysis results for OV. Top ranked GO terms and important GO terms are reported according to the P-values, which are associated with better/worse survival prognosis.

Gene cluster $C5$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0055114	Oxidation-reduction process.	6.53×10^{-11}	6.18×10^{-07}
2	GO:1901135	Carbohydrate derivative metabolic process.	5.99×10^{-10}	2.84×10^{-06}
3	GO:0006886	Intracellular protein transport.	2.98×10^{-09}	9.41×10^{-06}
Gene cluster $C7$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0001775	Cell activation.	3.96×10^{-72}	2.84×10^{-68}
2	GO:0045321	Leukocyte activation.	2.23×10^{-66}	8.01×10^{-63}
3	GO:0002682	Regulation of immune system process.	2.00×10^{-56}	4.78×10^{-53}
4	GO:0006952	Defense response.	1.70×10^{-55}	3.05×10^{-52}
5	GO:0002684	Positive regulation of immune system process.	6.21×10^{-50}	8.91×10^{-47}
6	GO:0006954	Inflammatory response.	5.38×10^{-48}	6.44×10^{-45}
8	GO:0002252	Immune effector process.	1.01×10^{-43}	9.08×10^{-41}
9	GO:0051707	Response to other organism.	1.87×10^{-43}	1.49×10^{-40}
10	GO:0043207	Response to external biotic stimulus.	2.21×10^{-43}	1.59×10^{-40}
11	GO:0009607	Response to biotic stimulus.	5.05×10^{-43}	3.30×10^{-40}
14	GO:0050776	Regulation of immune response.	3.20×10^{-39}	1.64×10^{-36}
15	GO:0002366	Leukocyte activation involved in immune response.	2.56×10^{-38}	1.22×10^{-35}
16	GO:0002443	Leukocyte mediated immunity.	3.21×10^{-38}	1.44×10^{-35}
17	GO:0002263	Cell activation involved in immune response.	4.24×10^{-38}	1.79×10^{-35}
18	GO:0098542	Defense response to other organism.	6.22×10^{-38}	2.48×10^{-35}
19	GO:0050900	Leukocyte migration.	5.82×10^{-37}	2.20×10^{-34}
21	GO:0046649	Lymphocyte activation.	2.54×10^{-35}	8.68×10^{-33}
22	GO:0034097	Response to cytokine.	2.79×10^{-35}	9.12×10^{-33}
23	GO:0002694	Regulation of leukocyte activation.	3.06×10^{-35}	9.54×10^{-33}
Gene cluster $C8$				
Rank	Term	Description	P-value	q-value FDR B&H
1	GO:0007017	Microtubule-based process.	4.85×10^{-50}	2.73×10^{-46}
2	GO:0000226	Microtubule cytoskeleton organization.	2.20×10^{-42}	6.19×10^{-39}
3	GO:0007018	Microtubule-based movement.	8.87×10^{-41}	1.67×10^{-37}

5. IMPRESS: PREDICTING BREAST CANCER NEOADJUVANT CHEMOTHERAPY RESPONSE FROM MULTIMODAL HISTOPATHOLOGIC IMAGES

Advances in computational algorithms and tools have made the prediction of cancer patient outcomes using computational pathology feasible. However, predicting clinical outcomes from pre-treatment histopathologic images remains a challenging task, hindered by the limited understanding of tumor immune micro-environments.

In this chapter, an automatic, accurate, comprehensive, interpretable, and reproducible whole slide image (WSI) feature extraction workflow known as, IMage-based Pathological REgistration and Segmentation Statistics (IMPRESS), is described. We aim to investigate whether machine learning algorithms using automatic feature extraction methods can predict neoadjuvant chemotherapy (NAC) outcomes in HER2-positive (HER2+) and triple-negative breast cancer (TNBC) patients. Features are derived from tumor immune micro-environment and clinical data and used to train machine learning models to accurately predict the response to NAC in breast cancer patients. The results demonstrate that this method outperforms the results trained from features that manually generated by pathologists.

5.1 Introduction

Predicting patient outcomes based on features or grades derived from tumor histopathologic images has become a cornerstone of modern cancer care and precision medicine [272]. In contrast to traditional image analysis, artificial intelligence (AI)-based computational pathology utilizes multimodal histopathologic images and automatic feature calculation approaches to extract patterns and analyze features [273]. One of the objectives of such AI-based computational pathology approaches is to predict the clinical outcome including survivability. This has been recently demonstrated by "so called" end-to-end deep learning approaches [272], [274] and interpretable machine learning approaches that employ morphologic feature extraction [176], [275], [276]. These studies facilitated the applications of computational pathology for clinical diagnosis and prognosis, as well as the interpretation of the roles of

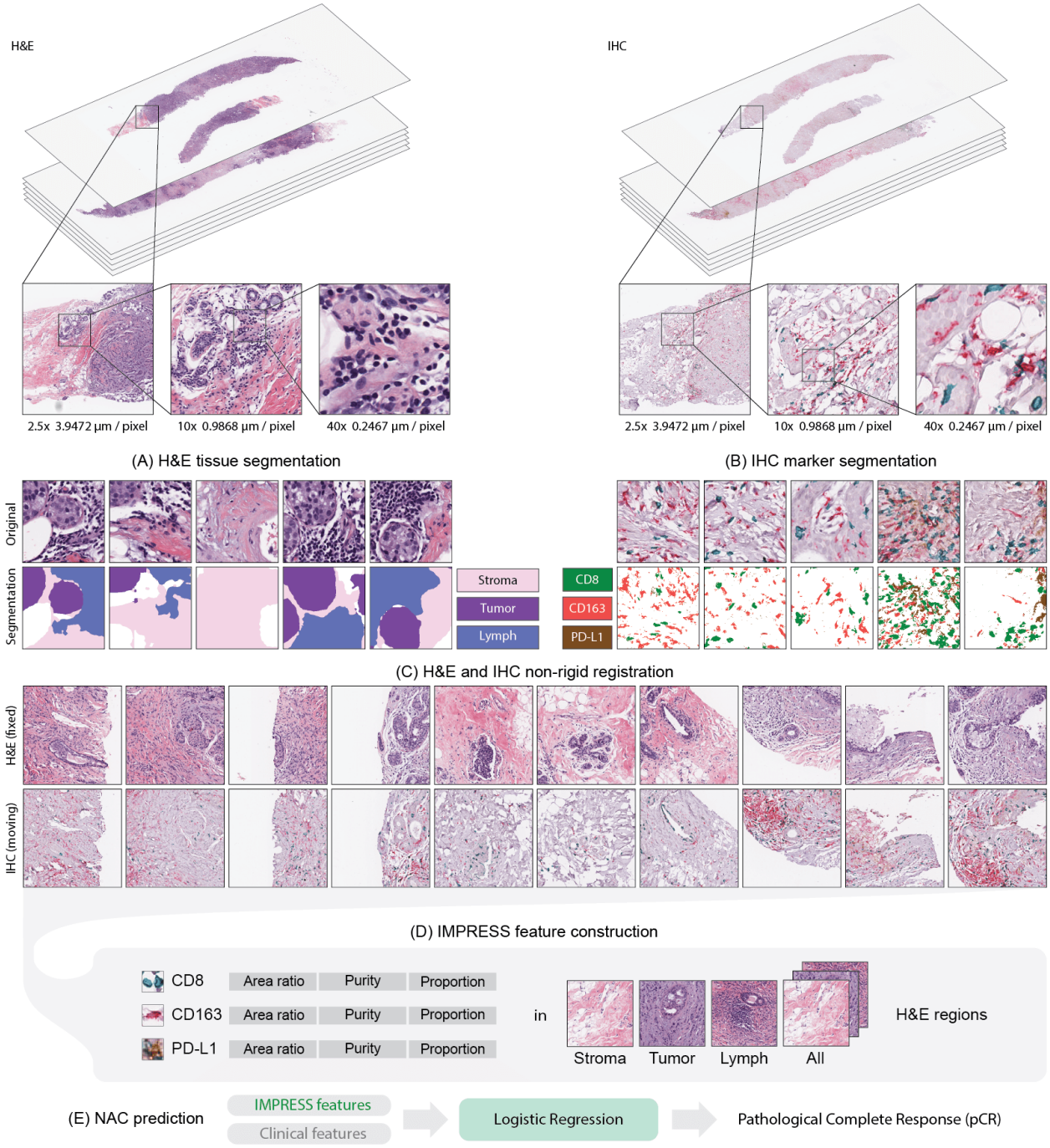


Figure 5.1. Overview of the IMPRESS workflow. (A) H&E tissue segmentation based on DeepLabV3 model. The segmentation generates stroma region, tumor region, and lymphocytes aggregated (lymph) region. (B) IHC markers segmentation. CD8, CD163, and PD-L1 are segmented. (C) H&E and IHC non-linear registration. First row: representative H&E patches; second row: corresponding IHC patches after registration. (D) IMage-based Pathological REgistration and Segmentation Statistics (IMPRESS) feature construction. Totally 36 IMPRESS features are constructed. (E) Neoadjuvant chemotherapy (NAC) prediction with logistic regression.

different cellular components in the tumor immune micro-environment such as tumor infiltrating lymphocytes (TILs), which has been discovered to play important roles in clinical outcomes of cancers [277].

In this work, we are interested in investigating whether machine learning-based algorithms using automatic feature extraction methods can predict neoadjuvant chemotherapy (NAC) outcomes in HER2-positive (HER2+) and triple-negative breast cancer (TNBC) patients[1]. In particular, we aim to predict the pathologic complete response (pCR), which is a presumptive surrogate for disease-free survival in HER2-positive (HER2+) and triple-negative breast cancer (TNBC) patients who have received neoadjuvant chemotherapy (NAC) [278], [279]. Potential factors associated with pCR have been widely investigated. It is known that higher pCR rates were found in hormone receptor (HR)-negative tumors in multiple trials [280]–[282], and a high Ki-67 index ($\geq 50\%$) was observed to be an independent predictive factor for pCR in HER2-positive BC patients [283], [284]. Besides these common histopathologic features, studies have also suggest a positive association between pCR and tumor immune micro-environment, especially TILs [285]–[294]. Wimberly *et al.* [295] also found PD-L1 expression was correlated with TILs and was a significant factor in predicting pCR .

Albeit predicting patient response to treatments or survivability to NAC in BC patients using imaging analysis has been explored in both the areas of radiology [296]–[299] and pathology [291], [300]–[302], predicting pCR from pre-neoadjuvant chemotherapy (pre-NAC) biopsies has tremendous clinical impact. However, existing studies (Qu *et al.* [299]) found that pre-NAC images are more challenging to predict pCR than post-NAC images. In fact, when they used a deep learning method to predict breast cancer pCR via MRI images, they observed AUC (area under the receiver operating characteristic curve) values of 0.553 for pre-NAC and 0.968 for post-NAC, respectively [299].

When predicting the pCR with machine learning model, one can also study the features that closely related to pCR. It has been reported over the past several years that cellular components of tumor immune micro-environment such as TILs are associated with response to NAC in breast cancer [291], [293], [302]–[304]. Hwang *et al.* [301] reported that high

[1]↑Breast cancer histopathologic dataset was provided by Dr. Zaibo Li at The Ohio State University.

pre-NAC TILs is a strong prognostic marker for pCR. Ali *et al.* [300] extracted lymphocyte density from pre-treatment biopsies and confirmed it is one of the strongest predictors in logistic regression. To systematically study how TILs and other image-based features in pre-NAC images can explain pCR outcomes, it is imperative to build an automatic and reproducible image-based feature extraction procedure. In Hwang *et al.* [301], pre-NAC TILs were calculated from the percentage of all mononuclear cells (including lymphocytes and plasma cells) in stromal areas, and were scored as a categorical variable in 10% increments [305]. Similarly, Denkert *et al.* [306] also assessed stromal TILs only, while Zhang *et al.* [307] evaluated the lymphocyte to monocyte ratio in pre-NAC to predict pCR. Denkert *et al.* [291] instead showed that the percentage of intratumoral lymphocytes (iTu-Ly) was the most significant independent parameter for pCR in breast cancer NAC rather than the percentage of stromal lymphocytes (str-Ly). However, in most of these studies, TILs and other histopathologic features were evaluated manually. Furthermore, the accuracy of NAC response prediction by machine learning algorithm and its comparison to human assessments are usually not reported and image-based statistical features are not exploited completely via multiplexed histopathologic images.

It is known that multiplexed histopathologic images can identify multiple markers simultaneously from a single tissue section [308]. In order to advance the understanding of whether pre-NAC histopathologic images (both H&E-stained and IHC-stained) features including those derived from tumor immune micro-environment (PD-L1, CD8+ T cells, and CD163+ macrophages) can predict NAC response, as well as to address the aforementioned limitations, an automatic whole slide image (WSI) feature extraction workflow is constructed, followed with a machine learning NAC prediction model. To improve feature extraction procedure, we take the advantages of multiplexed histopathologic images, extract 36 interpretable and meaningful histopathological features. Specifically, we establish three categories of quantitative features to characterize different cellular components — namely the "area ratio", "proportion", and "purity" — in our proposed workflow, and formally designate our workflow as "IMage-based Pathological REgistration and Segmentation Statistics", or "IMPRESS" in short. Sixty-two HER2+ and sixty-four TNBC patients are included in

our cohort to examine whether machine learning model using IMPRESS would be able to predict pCR for NAC.

To summarize, we investigate whether machine learning model using quantitative features automatically extracted by AI-based methods can predict response to neoadjunvant chemotherapy in breast cancer patients. We also compare the prediction accuracy between the model learned from IMPRESS and the model learned from features that manually generated by pathologists. Additionally, we comprehensively assess those automatically extracted features by feature importance analysis and residual cancer burden analysis. We find that the developed machine learning models utilize IMPRESS and clinical features can accurately predict the response to NAC in breast cancer patients and outperform the results learned by features that manually generated by pathologists.

5.2 Methods

5.2.1 Hardware and Software

We take advantage of four NVIDIA V100 graphics processing units (GPUs) and 1.6TB local storage. We use OpenSlide [309] v1.1.2 to access the WSI files, and PyTorch [131] v1.6.0, torchvision v0.7.0 for data loading, model training and testing. Machine learning and statistical analyses are performed in python with scikit-learn v0.23.2 [310]. We use pillow v7.2.0 and OpenCV [311] v4.4.0 for image processing in python. We use pandas v1.0.5 for data processing.

5.2.2 Study Cohorts

This study is approved by the Ohio State University Institutional Research Board and included 62 HER2-positive breast cancer (HER2+) patients and 64 triple-negative breast cancer (TNBC) patients treated with neoadjuvant chemotherapy (NAC) and follow-up surgical excision. Patients with histopathologically confirmed invasive breast carcinoma who underwent NAC from January 2011 to December 2016, those who had underwent surgery after completing NAC were included. HER2 status was determined on biopsy specimens using HER2 IHC and/or fluorescence in situ hybridization (FISH) in accordance with the

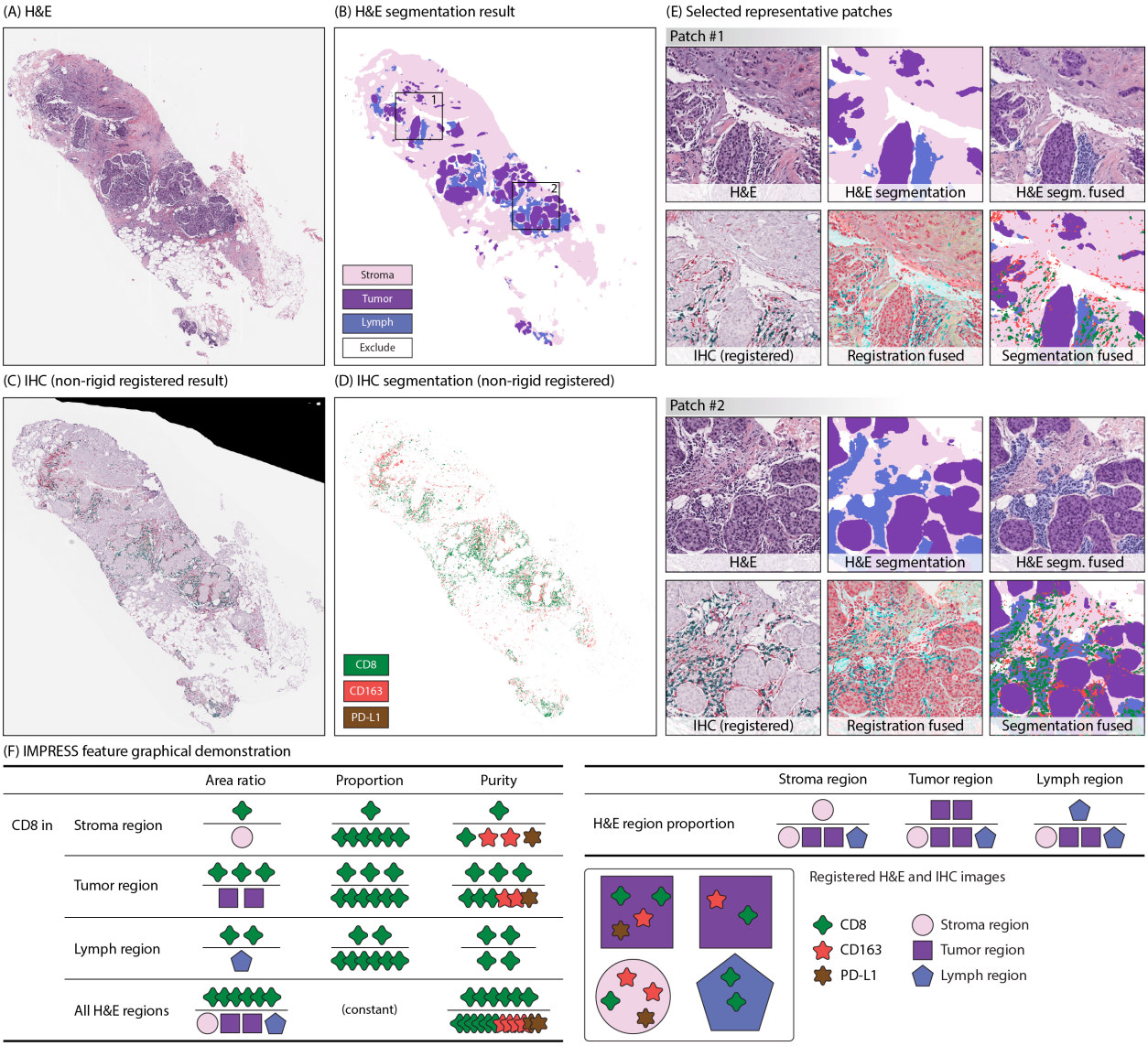


Figure 5.2. Tissue segmentation and image-level features extraction from registered H&E and IHC segmentation. (A) An example H&E tissue; (B) H&E tissue segmentation result; (C) IHC tissue (aligned to (A)) after non-linear registration; (D) IHC segmentation results, after non-linear registration. (E) Selected representative patches from (B) including (1) H&E patch, (2) H&E segmentation, (3) H&E segmentation (segm. in short) fused with original patch, (4) IHC patch after registration, (5) IHC patch after registration fused with H&E patch, and (6) H&E, IHC segmentation fused patch; (F) IMPRESS feature graphical demonstration. In (F), each IHC marker produces 11 features (CD8 was shown as an example), H&E region produces 3 features, totally 36 IMPRESS features. Figure best viewed in color.

criteria of American Society of Clinical Oncology (ASCO)/College of American Pathologist (CAP) guidelines updated guidelines [312].

5.2.3 Pathologic Assessment of the Response to Breast Cancer Neoadjuvant Chemotherapy

For neoadjuvant chemotherapy, all HER2+ patients received four cycles of AC (doxorubicin/cyclophosphamide) together with Taxol (paclitaxel/docetaxel) and trastuzumab except 7 patients (3 with residual tumor, 4 without residual tumor) who received four cycles of AC together with PTD (pertuzumab + trastuzumab + docetaxel). Triple-negative breast cancer patients received AC (doxorubicin/cyclophosphamide) together with Taxol (paclitaxel/docetaxel).

After NAC, all study cohort patients underwent surgery and the resection specimens are examined grossly and microscopically. A pathologic complete response (pCR) is recognized if no detectable residual invasive carcinoma (except in situ carcinoma) and absence of any metastatic tumor in lymph node, whereas the presence of residual invasive carcinoma in breast or in lymph node designated the incomplete response.

Residual cancer burden (RCB) [313], calculated from primary tumor bed and lymph nodes information, is a continuous variable describing the effectiveness after chemotherapy treatment. RCB is evaluated in all breast cancer cases with incomplete chemotherapy response, by comparing the pre-treatment core needle biopsy with the post-treatment resection specimen. RCB value is calculated based on tumor cellularity, tumor size change, and lymph node metastasis as described in [313].

5.2.4 Multi-color Multiplex Immunohistochemistry with CD8, CD163, PD-L1, and Assessment by Pathologists

Multi-color multiplex immunohistochemistry (IHC) with CD8 for cytotoxic T lymphocytes (clone MRQ26, mouse, Ventana), CD163 for macrophages (clone SP57, rabbit, Ventana), and PD-L1 (clone SP263, rabbit, Ventana) are performed on freshly cut pre-treatment biopsies as described in [314], [315]. Specific staining is considered in membranous PD-L1 staining in tumor cells or immune cells. The immunohistochemistry is evaluated with

consensus viewing by two pathologists (Drs. Yanjun Hou and Zaibo Li). The percentage of PD-L1 positively-stained cells are recorded and used for machine learning models (features that generated by pathologists). Pathologists manually assessed parameters include: PD-L1 expression in tumor cells (PD-L1 TC), PD-L1 expression in immune cells (PD-L1 IC), PD1 expression in immune cells, intratumoral CD8+ immune cells (IT-CD8+), peritumoral CD8+ immune cells (PT-CD8+), intratumoral CD163+ macrophages (IT-CD163+), and peritumoral CD163+ macrophages (PT-CD163+).

5.2.5 Non-Linear Image Registration

All H&E-stained and IHC-stained slides are scanned into WSIs using Hamamatsu scanner with 20× magnification. Although H&E-stained slides and IHC-stained slides from each case are continuous sections from paraffin-embedded tissue blocks, they are not always well aligned in the same space (2-D Euclidean space). In order to correctly assemble CD8 cytotoxic T-cells, CD163 macrophages, and PD-L1-expressing cells on H&E stained images, non-linear image registration is applied on IHC-stained images using H&E-stained images as templates.

Specifically, we adopt a multi-step, automatic, and non-linear histological image registration method [114], [119] and applies it to our dataset. First, the images are converted into grayscale, downsampled, and histogram equalized images. Then an initial affine registration is performed, followed with the non-linear registration. A few tissues in WSIs which had visually bad registration results are excluded.

5.2.6 H&E Region Segmentation

Training Data

The H&E region segmentation aims to automatically identify the stromal tissue region, tumoral tissue region, and lymphocytes aggregated tissue region. In this thesis, we fully utilize the breast cancer dataset from The Cancer Genome Atlas (TCGA) [316] consisting of 151 images [317] as training data, where each image has a segmentation map with 22 region classes labelled by multiple pathologists. We slice those images with 10% horizontal and

vertical overlapping, and generated 900 patches in total. Each patch is in 20× magnification (around $0.5\mu m$ per pixel) with 1024×1024 pixels in size.

We define four segmentation classes, including (1) stromal region (Stroma), (2) tumoral region (Tumor), (3) lymphocytes aggregated region (Lymph), and (4) excluded region (Exclude). The tumoral region includes invasive carcinoma and angioinvasion regions. The lymphocytes aggregated region includes lymphocytic infiltration, lymphatics, and other immune infiltrate, as well as considering the inflammation-rich area. The excluded region contains background or other regions not of our interest (*e.g.*, adipocytes).

Deep Learning Model, Hyper-parameters, and Evaluation Metrics

The deep learning model "DeepLabV3" [127] is adopted to learn the segmentation of the H&E regions. In DeepLabV3, atrous convolution was introduced and has the ability to capture larger field-of-view as well as control the resolution of feature responses [127]. In detail, the residual network ResNet-101 [318] is employed into DeepLabV3 and is implemented in PyTorch and torchvision [131] with auxiliary loss weight = 0.5. During the training, weighted mean squared error loss criterion is used by adopting the inverse of number of the pixels as class weights. Adaptive moment estimation (Adam) optimizer [178] is adopted with learning rate = 1×10^{-4} and batch size = 2 throughout the experiments. We search the number of epochs as the hyper-parameter [319] with five-fold cross-validation training scheme. Dice coefficient [252] is adopted to evaluate the model performance. A higher dice coefficient suggests a better segmentation performance.

Training, Validation, and Testing Schemes

For 900 image patches, we firstly hold out 10% of the image patches for testing (these patches are not used for any training purposes). Next, five-fold cross-validation training scheme is applied to the rest of the 810 patches. Namely, in each fold, 80% of the data are used for training, and 20% of the data are used for validation (*i.e.*, tuning the hyper-parameter). Patches cropped from the same image will not be separated into different sets. Models are evaluated every 20 epochs, the optimal number of epochs is chosen according to

the optimal mean dice coefficients among the five folds. We find the number of epochs = 280 results in optimal validation performance.

After the optimal number of epochs is determined, the deep learning model is applied on the entire training set for model training, and the testing performances on the 10% hold out testing set is reported.

Performances in TCGA dataset

All performances are measured in dice coefficients. The final training performances are 0.9881 for stromal region, 0.9941 for tumoral region, 0.9876 for lymph region, and 0.9911 for excluded region. The mean dice coefficient for training is 0.9902. The final testing performances are 0.8314 for stromal region, 0.8880 for tumoral region, 0.7065 for lymph region, and 0.7996 for excluded region. The mean dice coefficient for testing is 0.8064.

Applying Trained Deep Learning Model to Study Cohorts

The trained DeepLabV3 model is then applied to our study cohorts HER2+ and TNBC. The trained TCGA images and the targeted HER2+ and TNBC WSIs are in same magnifications (20× objective lens). We firstly slice H&E WSIs into 1024 × 1024 pixels patches with 200 pixels horizontal and vertical overlapping. Then, during the feed-forward process in deep neural networks, the predicted class probabilities in each pixel at overlapped regions are averaged, and the class with highest probability in each pixel is voted as the prediction result. The performances are shown in Result section.

5.2.7 Immunohistochemistry Markers Segmentation

Segmenting the IHC markers including CD8, CD163, and PD-L1, which are amplified by several visually distinctive colors, is one of the essential step for acquiring final image-based features. In this study, color-based K-means clustering [320] is performed to segment CD8, CD163, PD-L1, and other areas (background and area not of interest).

Color-based K-means Segmentation

Firstly, at most 10 image patches with 512×512 pixels in size are selected from each IHC tissue with lowest excluded region ratio (from H&E segmentation results) in HER2+ and TNBC cohorts, respectively. Secondly, we convert all selected image patches from RGB color space to L*a*b* color space [321], [322], which ensures the highest color contrast across three different IHC markers. Thirdly, a K-means clustering is performed. It aggregates each pixels of selected patches to different clusters in L*a*b* color space. In detail, we set $K = 15$, number of initialization $= 3$, and maximum number of iteration $= 300$ with tolerance $= 1 \times 10^{-4}$.

Next, two pathologists (Drs. Ahmad Mahmoud Alkashash and Carlo De la Sancha) help to identify and confirm the clustering centers of CD8, CD163, and PD-L1. Each IHC markers may contain several clustering centers. Finally, according to the learned K-means clustering centers, we apply the rest of the IHC WSIs to this model and obtain the IHC markers segmentation results.

We compare each four 1024×1024 patches from HER2+ and TNBC cohorts with two pathologists manually labeled IHC markers, the dice coefficients are then reported. The performances are shown in Result section.

5.2.8 IMPRESS Feature Extraction

In total 36 image-based features are extracted from the proposed IMPRESS workflow (Figure 5.2F). All features are calculated based on the WSI from each patient. Basically, each of CD8, CD163, and PD-L1 IHC markers will produce 11 features, which are the combination of "area ratio" (or "ratio" in short), "proportion", "purity" in *stroma*, *tumor*, *lymph*, and *all H&E* regions. Here *lymph* stands for lymphocytes aggregated region. The proportions in *all H&E* regions are excluded as it equals to 1. In addition, 3 features from H&E region proportions are also exploited: (1) the ratio of stromal region to all H&E regions; (2) the ratio of tumoral region to all H&E regions; and (3) the ratio of lymphocytes region to all H&E regions. So the total number of IMPRESS features is $3 \times 11 + 3 = 36$.

The definition of the area ratio (*e.g.*, *Lymph: CD8 ratio*) is the ratio of the total number of pixels of an IHC marker (CD8) on a certain H&E region (*Lymph*) to the total number of pixels of that H&E region (*Lymph*). The area ratio can be interpreted as how much of an IHC marker can be expressed on a certain type of tumor micro-environments. The definition of the proportion (*e.g.*, *Lymph: CD8 proportion*) is the ratio of the total number of pixels of an IHC marker (CD8) on a certain H&E region (*Lymph*) to the total number of pixels of that marker (CD8) on all valid H&E regions. The definition of the purity (*e.g.*, *Lymph: CD8 purity*) is the ratio of the total number of pixels of an IHC marker (CD8) to the total number of pixels of all IHC markers (CD8, CD163, and PD-L1) on a certain H&E region (*Lymph*).

The definition of "all H&E regions" (*All*) is the pixel sum of *Stroma*, *Tumor*, and *Lymph* regions. The full list of features is also presented in Table 5.1.

5.2.9 Machine Learning Predicts NAC Outcome

Training, Validation, and Testing Setting

Due to the sample size, leave-one-out training and testing scheme [60] is adopted. Given N patients in the data cohort, each time 1 patient is held out for testing, and the remained $N-1$ patients are used for training and validation. For the $N-1$ patients during training, five-fold cross-validation is adopted. For each fold, 80% of the data is used for training, and the rest 20% data is used for model validation (*i.e.*, finding the hyper-parameters of the model). All features in training & validation sets are standardized to follow standard normal distribution, and the standardization parameters are also applied to the testing set.

LASSO-regularized Logistic Regression

Logistic regression model implemented in scikit-learn [310] (version 0.23.2) is adopted to predict NAC outcome. The objective function for LASSO-regularized logistic regression is

$$\text{Minimize } L(\theta) = \frac{1}{N-1}\left[\sum_{\mathrm{i}=1}^{N-1} -\Big(\alpha_1 y^{(\mathrm{i})}\log(h_{\theta,b}(x^{(\mathrm{i})}))+\alpha_2(1-y^{(\mathrm{i})})\log(1-h_{\theta,b}(x^{(\mathrm{i})}))\Big)+\lambda\sum_{\mathrm{j}=1}^{K}|\theta_{\mathrm{j}}|\right], \tag{5.1}$$

Table 5.1. List of 36 IMPRESS features constructed from H&E and IHC images. 3 features can be constructed only from H&E image, 3 features can be only constructed from IHC image.

	Feature name	Explanation	Data source
1	*Stroma: CD8 ratio*	The area ratio of CD8 to stroma region.	H&E + IHC
2	*Stroma: CD163 ratio*	The area ratio of CD163 to stroma region.	H&E + IHC
3	*Stroma: PD-L1 ratio*	The area ratio of PD-L1 to stroma region.	H&E + IHC
4	*Stroma: CD8 proportion*	The area ratio of CD8 in stroma region to all H&E regions.	H&E + IHC
5	*Stroma: CD163 proportion*	The area ratio of CD163 in stroma region to all H&E regions.	H&E + IHC
6	*Stroma: PD-L1 proportion*	The area ratio of PD-L1 in stroma region to all H&E regions.	H&E + IHC
7	*Stroma: CD8 purity*	In stroma region, the area ratio of CD8 to all IHC markers.	H&E + IHC
8	*Stroma: CD163 purity*	In stroma region, the area ratio of CD163 to all IHC markers.	H&E + IHC
9	*Stroma: PD-L1 purity*	In stroma region, the area ratio of PD-L1 to all IHC markers.	H&E + IHC
10	*Tumor: CD8 ratio*	The area ratio of CD8 to tumoral region.	H&E + IHC
11	*Tumor: CD163 ratio*	The area ratio of CD163 to tumoral region.	H&E + IHC
12	*Tumor: PD-L1 ratio*	The area ratio of PD-L1 to tumoral region.	H&E + IHC
13	*Tumor: CD8 proportion*	The area ratio of CD8 in tumoral region to all H&E regions.	H&E + IHC
14	*Tumor: CD163 proportion*	The area ratio of CD163 in tumoral region to all H&E regions.	H&E + IHC
15	*Tumor: PD-L1 proportion*	The area ratio of PD-L1 in tumoral region to all H&E regions.	H&E + IHC
16	*Tumor: CD8 purity*	In tumoral region, the area ratio of CD8 to all IHC markers.	H&E + IHC
17	*Tumor: CD163 purity*	In tumoral region, the area ratio of CD163 to all IHC markers.	H&E + IHC
18	*Tumor: PD-L1 purity*	In tumoral region, the area ratio of PD-L1 to all IHC markers.	H&E + IHC
19	*Lymph: CD8 ratio*	The area ratio of CD8 to lymphocytes aggregated region.	H&E + IHC
20	*Lymph: CD163 ratio*	The area ratio of CD163 to lymphocytes aggregated region.	H&E + IHC
21	*Lymph: PD-L1 ratio*	The area ratio of PD-L1 to lymphocytes aggregated region.	H&E + IHC
22	*Lymph: CD8 proportion*	The area ratio of CD8 in lymphocytes aggregated region to all H&E regions.	H&E + IHC
23	*Lymph: CD163 proportion*	The area ratio of CD163 in lymphocytes aggregated region to all H&E regions.	H&E + IHC
24	*Lymph: PD-L1 proportion*	The area ratio of PD-L1 in lymphocytes aggregated region to all H&E regions.	H&E + IHC
25	*Lymph: CD8 purity*	In lymphocytes aggregated region, the area ratio of CD8 to all IHC markers.	H&E + IHC
26	*Lymph: CD163 purity*	In lymphocytes aggregated region, the area ratio of CD163 to all IHC markers.	H&E + IHC
27	*Lymph: PD-L1 purity*	In lymphocytes aggregated region, the area ratio of PD-L1 to all IHC markers.	H&E + IHC
28	*All: CD8 ratio*	The area ratio of CD8 to all H&E region.	H&E + IHC
29	*All: CD163 ratio*	The area ratio of CD163 to all H&E region.	H&E + IHC
30	*All: PD-L1 ratio*	The area ratio of PD-L1 to all H&E region.	H&E + IHC
31	*All: CD8 purity*	In all H&E region, the area ratio of CD8 to all IHC markers.	IHC
32	*All: CD163 purity*	In all H&E region, the area ratio of CD163 to all IHC markers.	IHC
33	*All: PD-L1 purity*	In all H&E region, the area ratio of PD-L1 to all IHC markers.	IHC
34	*Stroma: H&E proportion*	The area ratio of stromal region to all H&E regions.	H&E
35	*Tumor: H&E proportion*	The area ratio of tumoral region to all H&E regions.	H&E
36	*Lymph: H&E proportion*	The area ratio of lymphocytes aggregated region to all H&E regions.	H&E

where x represents the feature values, y is the response (pCR), $h_{\theta,b}(x) = x^T\theta + b$ is the linear function with weight θ and bias b. $N-1$ is the number of training samples (1 sample for held out testing), K is the number of features, λ is the LASSO regularization penalty weight, α_1 and α_2 imposed the class weights. To avoid confusion with other parts of this thesis, symbols θ, α, y, h, b, x, and λ are only defined in this section.

We set the number of maximum iteration = 100, optimization tolerance = 1×10^{-4}. The hyper-parameters to be searched is the LASSO regularization [62] penalty weight λ from

0.1 to 1.0 with step = 0.1. The class weights α_1 and α_2 are used for balanced learning by adjusting weights inversely proportional to pCR frequencies in the training label.

Evaluation Metrics

We adopt four measurements to evaluate the results, namely, AUC (area under the ROC curve) [181], F1 score [323], precision, and recall. AUC is calculated in scikit-learn v0.23.2 [310]. F1 score, precision, and recall are calculated in scikit-learn with "macro" average method.

A well-discriminated model would have an AUC close to 1. We consider an AUC $>$ 0.85 being a well-performed prediction, and an AUC $>$ 0.75 being an adequate prediction. Precision is the ratio of true positives to all positive classifications, it suggests how likely a patient will have pCR. A higher precision indicates a more reliable performance generated from the algorithm. Recall is the ratio of true positives to all true samples and shows the ability of the model to correctly identify the patients with pCR. Note that during this calculation, "positive" stands for a pCR. The F1 score is formulated as

$$\text{F1} = \frac{2(\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}, \tag{5.2}$$

which is the harmonic mean of the precision and recall, reflecting the learning accuracy.

To compare the performances between IMPRESS features and pathologists' assessed features (both include clinical features), LASSO-regularized logistic regression is used for both feature sets. The model and training schemes as well as evaluation metrics are remained same as before.

5.2.10 Statistical Analyses

We compare the distributions of IMPRESS and clinical features between HER2+ and TNBC cohorts using Mann-Whitney U test [143]. The fold change is calculated by the ratio of the median feature values between HER2+ and TNBC cohorts. Student's t-test [137] is adopted for comparing pair-wised AUCs from different trials. Spearman's rank correlation

coefficients [324] is adopted for calculating the relationships between features and pCR, the relationships among IMPRESS features, and the relationships between IMPRESS features and residual tumor sizes. It provides a correlation coefficient ρ and a P-value. All P-values are two-sided; P-values < 0.05 are deemed statistically significant.

5.3 Results

5.3.1 Clinical and Histopathological Characteristics of the Study Cohort

62 HER2-positive (HER2+) BC and 64 TNBC patients treated with NAC and surgical excision are included. HER2+ BC patients are treated with doxorubicin/cyclophosphamide/taxol together with anti-HER2 targeted therapy, including 24 patients (39%) with residual tumor and the other 36 patients (61%) with pCR. TNBC patients are treated with standard NAC (doxorubicin/cyclophosphamide/taxol) including 37 patients (58%) with residual tumor and the other 27 patients (32%) with pCR. The clinical and histopathologic characteristics of these patients are summarized in Table 5.2.

5.3.2 Workflow and Feature Construction

The workflow of this paper is presented in Figure 5.1, including H&E image acquisition and segmentation, IHC image acquisition and segmentation, and H&E – IHC image registration. Given the input paired H&E and IHC WSI, the automatic non-linear registration is performed on each IHC WSI using the corresponding H&E WSI as fixed references. With deep neural network "DeepLabV3" trained from pathologists labelled TCGA breast cancer H&E images [317], H&E tissue segmentation is performed and four region of interests are identified including stromal region (*Stroma*), tumoral region (*Tumor*), lymphocytes aggregated region (*Lymph*), and excluded region. All included regions (*Stroma*, *Tumor*, and *Lymph*) are defined as all H&E region (*All*). Figure 5.2AB shows an example of H&E image and its segmentation result. The multiplexing IHC markers including CD8 (green), CD163 (red), and PD-L1 (brown) are identified via color-based K-means segmentation. Figure 5.2CD shows an IHC image and its segmentation result. All results are reviewed and confirmed by two pathologists (Drs. Ahmad Mahmoud Alkashash and Carlo De la Sancha).

Table 5.2. Clinical and histopathological results of HER2-positive and TNBC cases with neoadjuvant chemotherapy (NAC).

Cohort	Characteristics		Case # / median	% / range
HER2+	Total case number		62	—
	Cases with residual tumor		24	38.71%
	Cases with pCR		38	61.29%
	Age (years)		56	30 – 76
	Nottingham grade	I	1	1.61%
		II	27	43.55%
		III	34	54.84%
	Nuclear grade	I	0	0.00%
		II	10	16.13%
		III	52	83.87%
	Estrogen receptor (ER) positive		30	48.39%
	Progesterone receptor (PR) positive		19	30.65%
	HER2/CEP17 ratio		6.73	1.23 – 22.98
	Residual cancer burden for non-pCR patients		1.39	0.91 – 4.14
TNBC	Total case number		64	—
	Cases with residual tumor		37	57.81%
	Cases with pCR		27	42.19%
	Age (years)		51	26 – 74
	Nottingham grade	I	0	0%
		II	15	23.4%
		III	49	76.6%
	Nuclear grade	I	0	0%
		II	9	14.1%
		III	55	85.9%
	Residual cancer burden for non-pCR patients		2.01	0.80 – 4.27

Next, an automatic, accurate, comprehensive, interpretable, and reproducible WSI feature extraction workflow is constructed, and generates 36 IMage-based Pathological REgistration and Segmentation Statistics (IMPRESS) features. Figure 5.2F demonstrates how IMPRESS features are calculated by using CD8 as an example. The full list of features is shown in Table 5.1. The distributions of IMPRESS features' expressions are demonstrated in Figure 5.3 in violin plots with values ranging from 0 to 1.

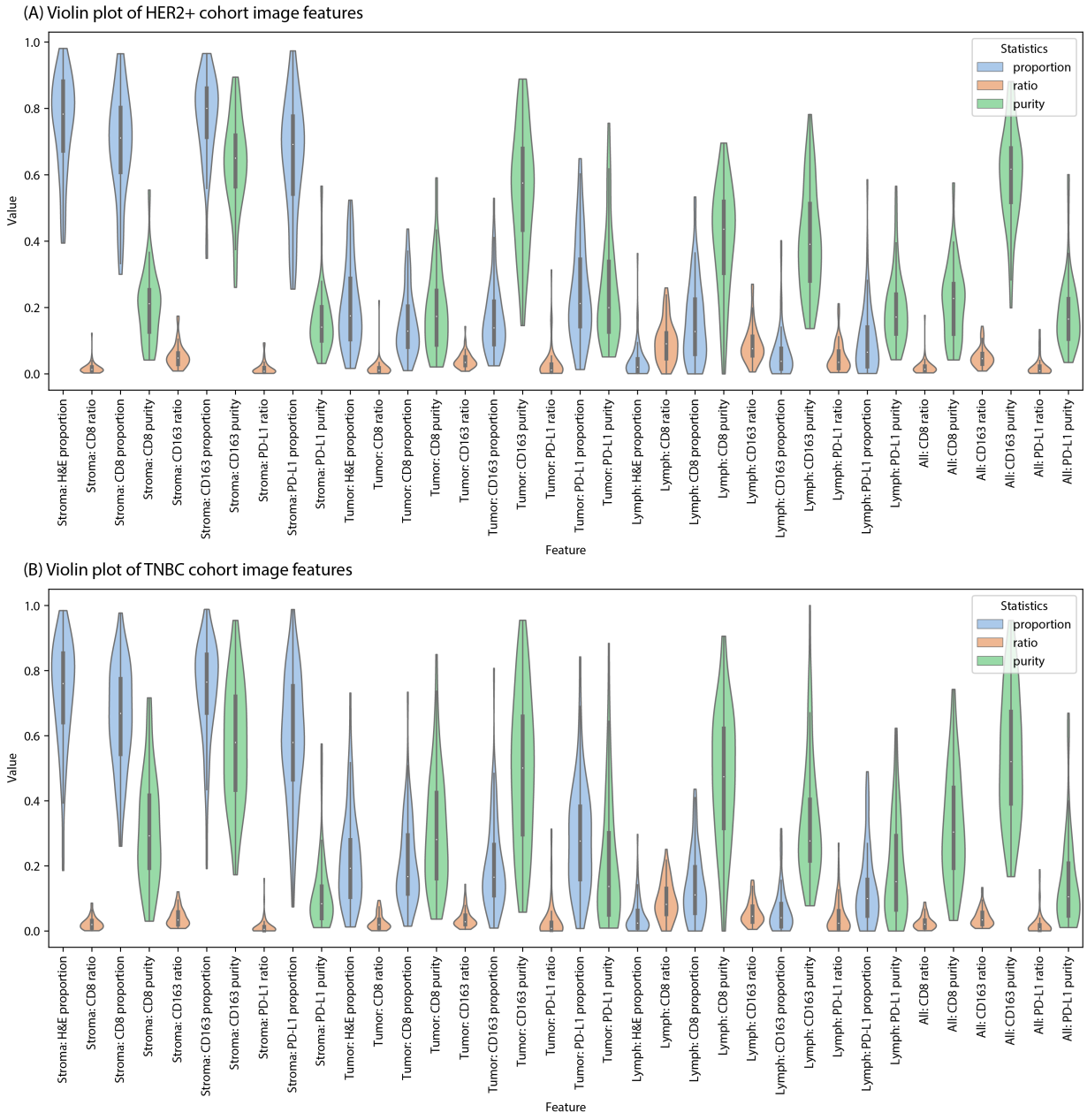


Figure 5.3. Violin plot of IMPRESS feature expressions in HER2+ cohort (A) and TNBC cohort (B).

In addition to IMPRESS features, clinical features and status of molecular markers (ER, PR, and HER2) are exploited. In the HER2+ cohort, six features are adopted including age, estrogen receptor status (ER+/−), estrogen receptor percentage (ER%), progesterone receptor status (PR+/−), progesterone receptor percentage (PR%), and the ratio of HER2

expression to chromosome 17 (HER2/CEP17). In TNBC cohort, age is the only available clinical feature since ER, PR, and HER2 are all negative.

5.3.3 Machine Learning Model using IMPRESS Features to Predict NAC Outcomes

LASSO-regularized logistic regression is adopted to evaluate the prediction power of the proposed IMPRESS features. In this study, four groups of features are compared, including all 36 IMPRESS plus clinical features (IMPRESS), IMPRESS H&E image features plus clinical features [IMPRESS (H&E only)] (Table 5.1), IMPRESS IHC image features plus clinical features [IMPRESS (IHC only)] (Table 5.1), and pathologists assessed IHC image features plus clinical features (Pathologists).

We first compare IMPRESS with IMPRESS (H&E only) and IMPRESS (IHC only). From Table 5.3 and Figure 5.4A, we find IMPRESS achieves significantly higher AUC than IMPRESS (H&E only) (t-test statistics = 62.69, P-value = 5.68×10^{-40}) and IMPRESS (IHC only) (t-test statistics = 79.97, P-value = 5.83×10^{-44}) in HER2 cases. Similarly, from Table 5.3 and Figure 5.4B, we find IMPRESS achieves significantly higher AUC than IMPRESS (H&E only) (t-test statistics = 16.87, P-value = 3.04×10^{-19}) and IMPRESS (IHC only) (t-test statistics = 33.60, P-value = 7.23×10^{-30}) in TNBC cases. The results suggest that combining H&E and IHC histopathologic images can extract additional features which benefits to the pCR prediction.

5.3.4 IMPRESS Features Outperformed Pathologists' Assessed Features for Predicting NAC Outcomes

Furthermore, IMPRESS features are compared to pathologists' manually assessed IHCs image features for CD8, CD163, and PD-L1.

In HER2+ cohort, we find IMPRESS achieves better performances (AUC = 0.8975 ± 0.0038) than pathologists' assessed features (AUC = 0.7880 ± 0.0065) significantly with t-test statistics = 64.59 (P-value = 1.84×10^{-40}) (Figure 5.4A). In TNBC cohort, we find IMPRESS achieves slightly better performances (AUC = 0.7674 ± 0.0209) than pathologists' assessed features (AUC = 0.7626 ± 0.0095) with t-test statistics = 0.94 (P-value = 3.54×10^{-1})

(Figure 5.4B). The detailed performances are summarized in Table 5.3. The results suggest that the AI-based features extracted from H&E and IHC histopathologic images can achieve equal or better performances than pathologists' assessed features, and are the preferred input to develop machine learning algorithms and to predict response to NAC in breast cancer patients.

Table 5.3. LASSO-regularized logistic regression performances in HER2+ and TNBC cohorts. Experiments are repeated 20 times with different random seeds in leave-one-out cross-validation setting. mean value ± standard deviation are reported. Best performed mean values are highlighted in bold face.

Cohort	Features	AUC	F1 score	Precision	Recall
HER2+	IMPRESS	**0.8975±0.0038**	**0.8687±0.0077**	0.8716±0.0115	**0.8658±0.0081**
	IMPRESS (H&E only)	0.8118±0.0048	0.8269±0.0052	**0.9059±0.0009**	0.7605±0.0081
	IMPRESS (IHC only)	0.7746±0.0057	0.7775±0.0085	0.8454±0.0023	0.7197±0.0129
	Pathologists' features	0.7880±0.0065	0.7820±0.0025	0.8696±0.0061	0.7105±0.0000
TNBC	IMPRESS	**0.7674±0.0209**	**0.7017±0.0377**	**0.6903±0.0286**	0.7148±0.0552
	IMPRESS (H&E only)	0.6795±0.0103	0.5882±0.0000	0.6250±0.0000	0.5556±0.0000
	IMPRESS (IHC only)	0.5975±0.0087	0.5915±0.0103	0.5637±0.0061	0.6222±0.0152
	Pathologists' features	0.7626±0.0095	0.6897±0.0077	0.6454±0.0135	**0.7407±0.0000**

5.3.5 Feature Importance Analysis in Machine Learning Model

We summarize the feature coefficients produced from LASSO-regularized logistic regression in Figure 5.4C (HER2+ cohort) and Figure 5.4D (TNBC cohort). The top important features are also summarized in Table 5.5. For the HER2+ cohort, three out of the top five favorable prognostic markers (positively associated with pCR) are related to lymphocyte aggregated region, including CD8 ratio, CD163 ratio, and PD-L1 ratio. The favorable clinical prognostic marker of HER2/CEP17 ratio is ranked as the third, which echos the finding in [325] that suggested high HER2/CEP17 ratio is significantly associated with pCR. In contrast, four out of the top five adverse prognostic markers (negatively associated with pCR) are related to clinical variables including age, ER ratio, PR positivity, and PR ratio. The second strongest adverse prognostic marker is *Stroma: CD8 proportion*. For the TNBC cohort, the top five favorable prognostic markers are *Lymph: PD-L1 ratio*, *Lymph: PD-L1 proportion*, *Tumor: CD8 proportion*, *Tumor: CD8 purity*, and *Lymph: CD163 proportion*.

The top five adverse prognostic markers are *Stroma: CD8 proportion*, age, *Tumor: PD-L1 ratio*, *Stroma: CD8 ratio*, and *Lymph: CD8 purity*. Detailed feature importance ranking and coefficients are listed in Table 5.5. From these results, we observe that features related to lymphocytes aggregated region (*Lymph*) are the most favorable prognostic markers to pCR. In addition, age, which plays an adverse role, is more critical in the HER2+ cohort than in the TNBC cohort. Interestingly, we find *Stroma: CD8 proportion* is one of the most adverse prognostic markers in both cohorts, suggesting more CD8 in the stromal region than in other regions is not a favorable sign for pCR.

The comparison of coefficient importance between HER2+ and TNBC cohorts is shown in Figure 5.5A. Some IMPRESS features are agreed between HER2+ and TNBC cohorts. For example, *Lymph: PD-L1 ratio* and *Tumor: CD8 proportion* act as common favorable features to pCR; Age and *Stroma: CD8 proportion* act as common adverse features to pCR. However, we also observe some disparities between the HER2+ and TNBC cohorts. CD8 and CD163 play more essential roles in HER2+ cohort (*e.g.*, *Lymph: CD8 ratio* and *Lymph: CD163 ratio*), whereas PD-L1 is more informative in the TNBC cohort. Similar results can also be observed in the following univariate analysis (Figure 5.4BC).

5.3.6 Univariate Analyses with pCR Response

As IMPRESS features outperformed pathologists' assessed features in predicting pCR and are correlated with RCB, we further perform univariate analysis to investigate the relationships between IMPRESS features and NAC responses, and to identify specific IMPRESS features which show significant differences in predicting NAC response between the HER2+ and TNBC cohorts.

We compare each feature by using pCR cases against residual tumor cases using the two-sided Student's t-test [137]. The top five favorable/adverse features with the most significant differences are presented in Figure 5.5B for the HER2+ cohort and in Figure 5.5C for the TNBC cohort. Complete results are further presented in Table 5.4. We find that the most significantly different features in pCR cases against residual tumor cases are highly consistent with those identified by machine learning methods, such as *Lymph: CD163 ratio* and *Lymph:*

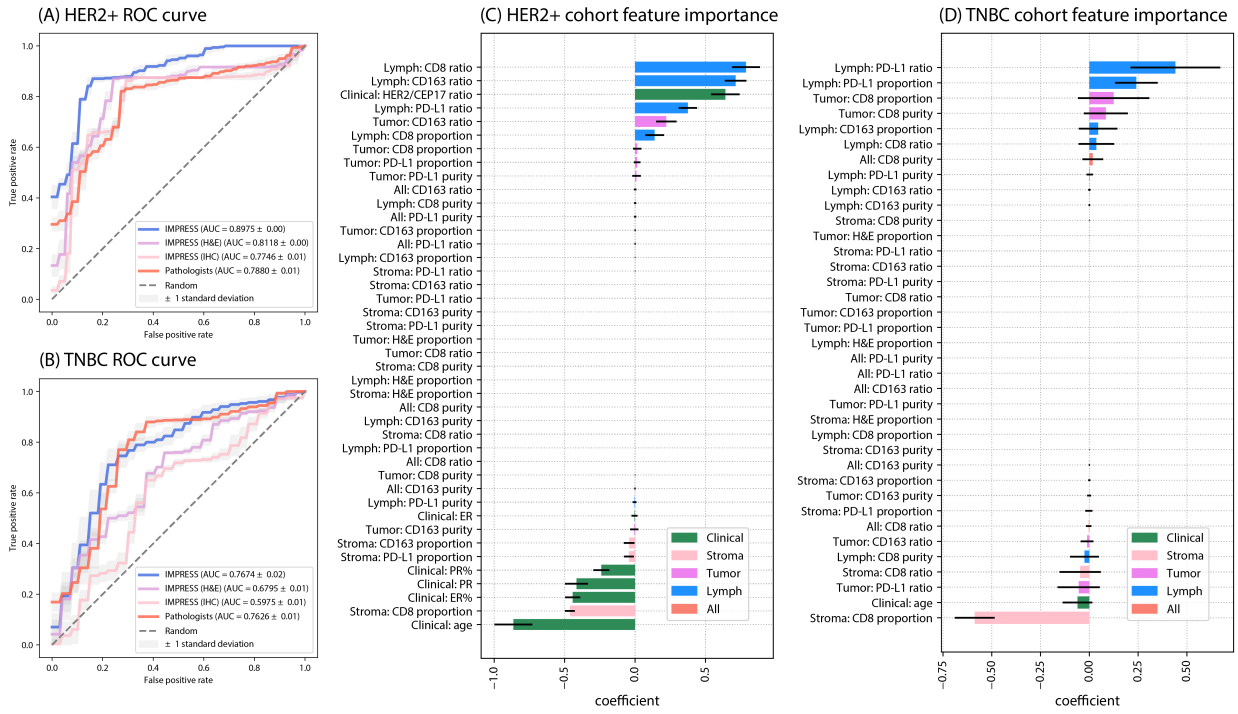


Figure 5.4. (A–B) Receiver operating characteristic (ROC) curve for HER2+ (A) and TNBC (B) cohorts in the logistic regression results. Blue line: IMPRESS plus clinical features; Purple line: IMPRESS (H&E features only) plus clinical features; Pink line: IMPRESS (IHC features only) plus clinical features; Red line: pathologists assessed plus clinical features. (C–D) Feature importance generated by logistic regression. Positive coefficients are associated with better prognosis (pCR) and vice versa. Horizontal line in each bar stands for standard deviation. (C) HER2+ cohort; (D) TNBC cohort. Figure best viewed in colors.

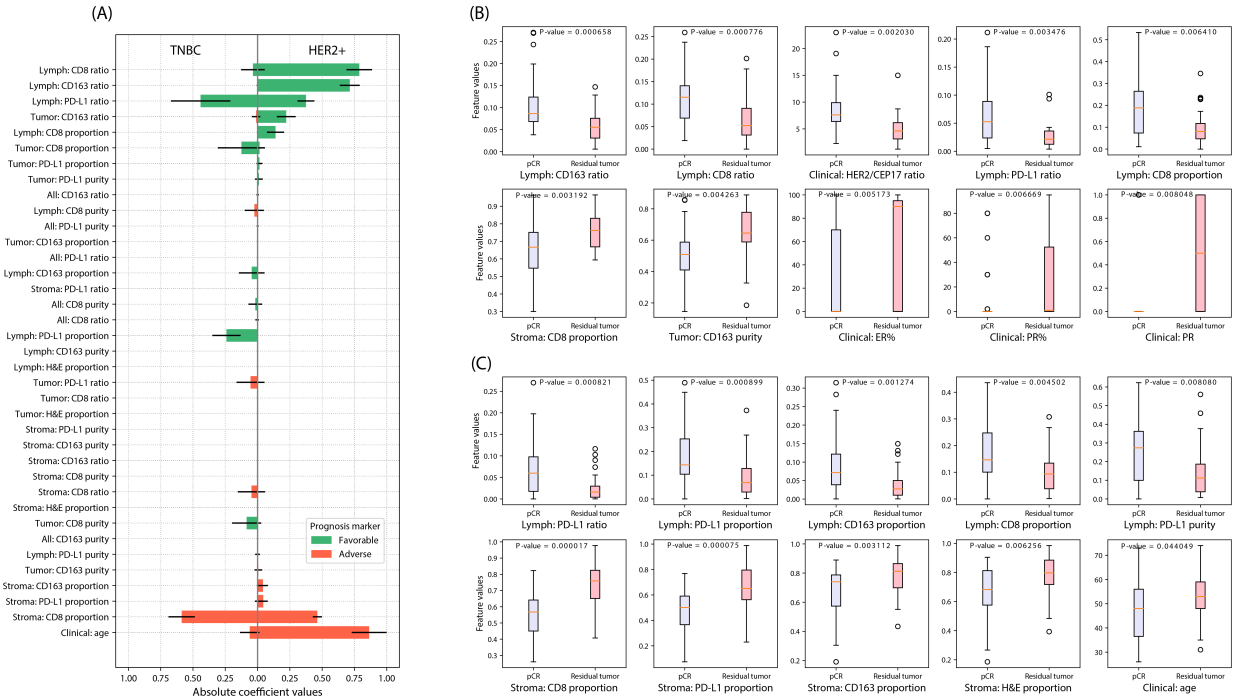


Figure 5.5. (A) Comparison of IMPRESS and clinical coefficient importance in machine learning results between HER2+ and TNBC cohorts, organized by HER2+ coefficients in descending order. Coefficients in the horizontal bar plot are reported in absolute values, the positive values are defined as "favorable" prognostic marker and vise versa for negative values. Figure best viewed in colors. (B–C) Univariate feature analysis in HER2+ cohort (B) and TNBC cohort (C) by comparing pCR cases against residual tumor cases. In (B) and (C), top row showed five most favorable features, bottom row showed five most adverse features. Two-sided P-values are calculated based on Student's t-test.

CD8 ratio, two top-ranked favorable features for HER2+ cases, which are identified by both univariate analysis and machine learning model. Nevertheless, a few features identified by the univariate analysis are not agreed with machine learning results. For example, *Tumor: CD163 purity* (P-value = 0.0043), one of the adverse features in HER2+ cases, is not identified in machine learning (Figure 5.4C). Similar inconsistency is also found in TNBC cases, such as *Lymph: CD8 proportion*.

To present an alternative point of view of the relationship between IMPRESS features and pCR, Spearman's rank correlation coefficient (SCC) is used to evaluate the differences among the features regarding their relationship to pCR. The results are shown in Figure 5.6A (HER2+) and Figure 5.6B (TNBC). The SCC results are generally consistent with the machine learning feature importance results (in Figure 5.4CD) and the univariate analysis results (in Figure 5.5BC), especially for the features related to lymphocytes aggregated regions (*Lymph*) and tumoral regions (*Tumor*). These results confirm the important roles of pre-NAC TILs in predicting pCR.

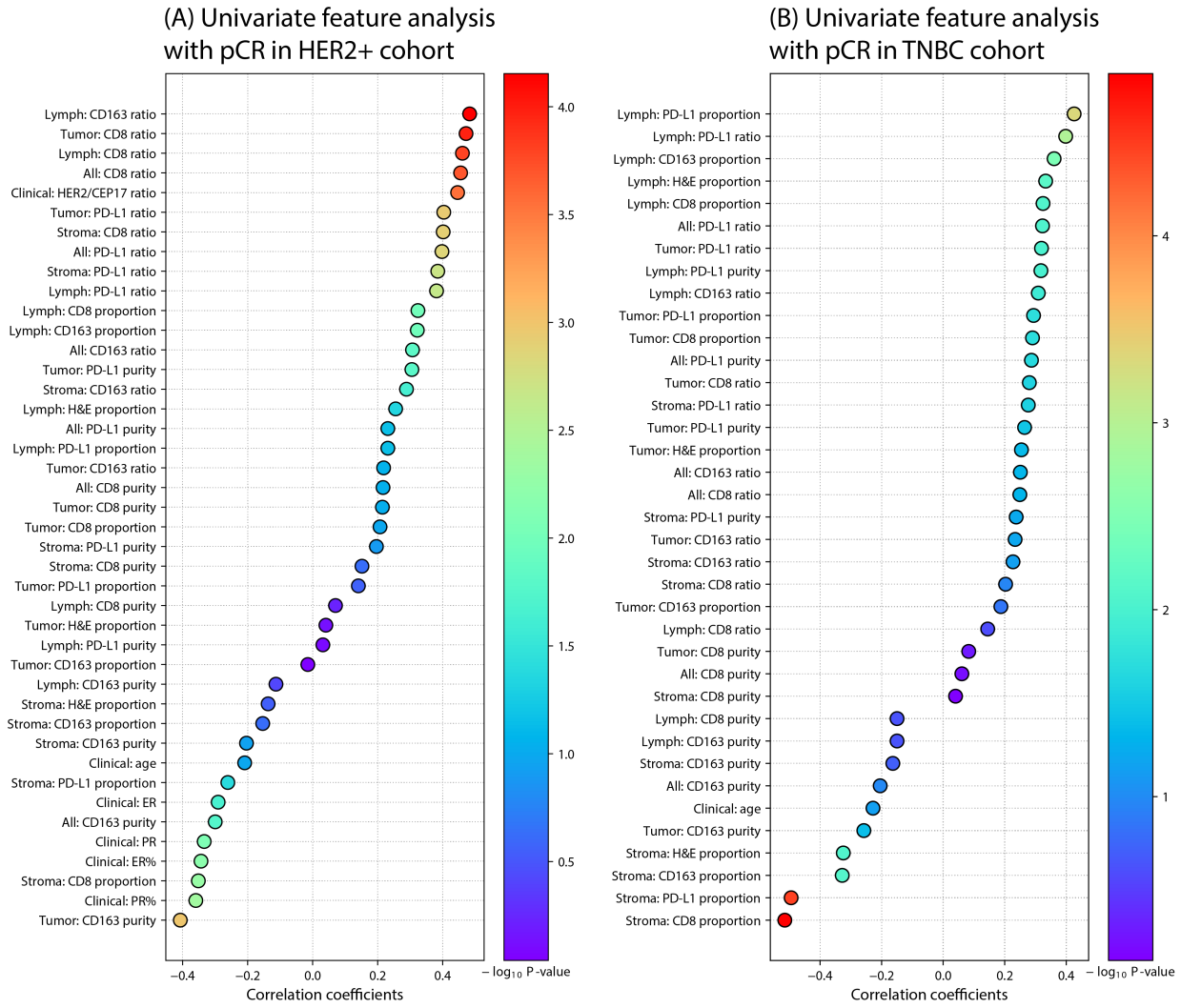


Figure 5.6. (A) Univariate feature analysis with pCR in HER2+ cohort. (B) Univariate feature analysis with pCR in TNBC cohort. Two-sided P-values are calculated based on Spearman's rank correlations. Figure best viewed in color.

Furthermore, we find that several IMPRESS features expressed significantly higher in HER2+ than in TNBC based on the Mann-Whitney U test results in Figure 5.7, such as *Stroma: PD-L1 purity* (P-value $= 5.06\times10^{-5}$), *Lymph: CD163 ratio* (P-value $= 5.50\times10^{-5}$), *etc.* Some features express significantly lower in HER2+ than TNBC, such as *Tumor: CD8 purity* (P-value $= 2.50 \times 10^{-4}$), *Stroma: CD8 purity* (P-value $= 2.23 \times 10^{-4}$), *etc.* These

results suggest that IMPRESS features distributed differently among different breast cancer cohorts, providing a different perspective between two breast cancer subtypes.

5.3.7 Relationships between IMPRESS and Residual Cancer Burden

For those patients who do not have pCR outcomes, residual cancer burden (RCB) is calculated in patients with residual tumor. The median RCB in the HER2+ cohort is 1.39 with a range of 0.91 – 4.14. The median RCB in the TNBC cohort is 2.01 with a range of 0.80 – 4.27. RCB is defined as 0 for patients with pCR. The non-parametric statistics from SCC ρ with two-sided P-values are used to examine the relationships between IMPRESS features and RCBs. The top 5 most favorable and most adverse IMPRESS prognostic features from machine learning analyses list in Table 5.5 are further compared with RCBs [Figure 5.8A (HER2+) and Figure 5.8B (TNBC)]. The complete list is showed in Table 5.6.

Table 5.5. Feature importance (top 5 are reported) in HER2+ and TNBC cohorts. Experiments are repeated 20 times with different random seeds in leave-one-out cross-validation setting. Top 5 favorable and adverse prognosis marker for HER2+ and TNBC cohorts are reported respectively. Values are reported in mean ± standard deviation.

	Favorable prognostic markers				**Adverse prognostic markers**			
Cohort	Rank	H&E region	Feature	Coefficients	Rank	H&E region	Feature	Coefficients
HER2+	1	Lymph	CD8 ratio	0.7879 ± 0.0992	1	Clinical	age	-0.8638 ± 0.1346
	2	Lymph	CD163 ratio	0.7141 ± 0.0770	2	Stroma	CD8 proportion	-0.4627 ± 0.0352
	3	Clinical	HER2/CEP17 ratio	0.6414 ± 0.1010	3	Clinical	ER%	-0.4429 ± 0.0536
	4	Lymph	PD-L1 ratio	0.3748 ± 0.0653	4	Clinical	PR+/−	-0.4158 ± 0.0818
	5	Tumor	CD163 ratio	0.2219 ± 0.0726	5	Clinical	PR%	-0.2396 ± 0.0568
TNBC	1	Lymph	PD-L1 ratio	0.4412 ± 0.2300	1	Stroma	CD8 proportion	-0.5878 ± 0.1024
	2	Lymph	PD-L1 proportion	0.2409 ± 0.1095	2	clinical	age	-0.0608 ± 0.0762
	3	Tumor	CD8 proportion	0.1249 ± 0.1827	3	Tumor	PD-L1 ratio	-0.0545 ± 0.1084
	4	Tumor	CD8 purity	0.0847 ± 0.1129	4	Stroma	CD8 ratio	-0.0472 ± 0.1059
	5	Lymph	CD163 proportion	0.0452 ± 0.0994	5	Lymph	CD8 purity	-0.0249 ± 0.0739

As demonstrated in Figure 5.8A for HER2+ cases, *Lymph: CD8 ratio*, *Lymph: CD163 ratio*, *Lymph: PD-L1 ratio*, and *Lymph: CD8 proportion* are negatively correlated with RCB significantly. In contrast, *Stroma: CD8 proportion* and *Tumor: CD163 purity* are positively correlated with RCB significantly. From the TNBC cohort in Figure 5.8B, *Lymph: PD-L1 ratio*, *Lymph PD-L1 proportion*, and *Lymph: CD163 proportion* are negatively correlated

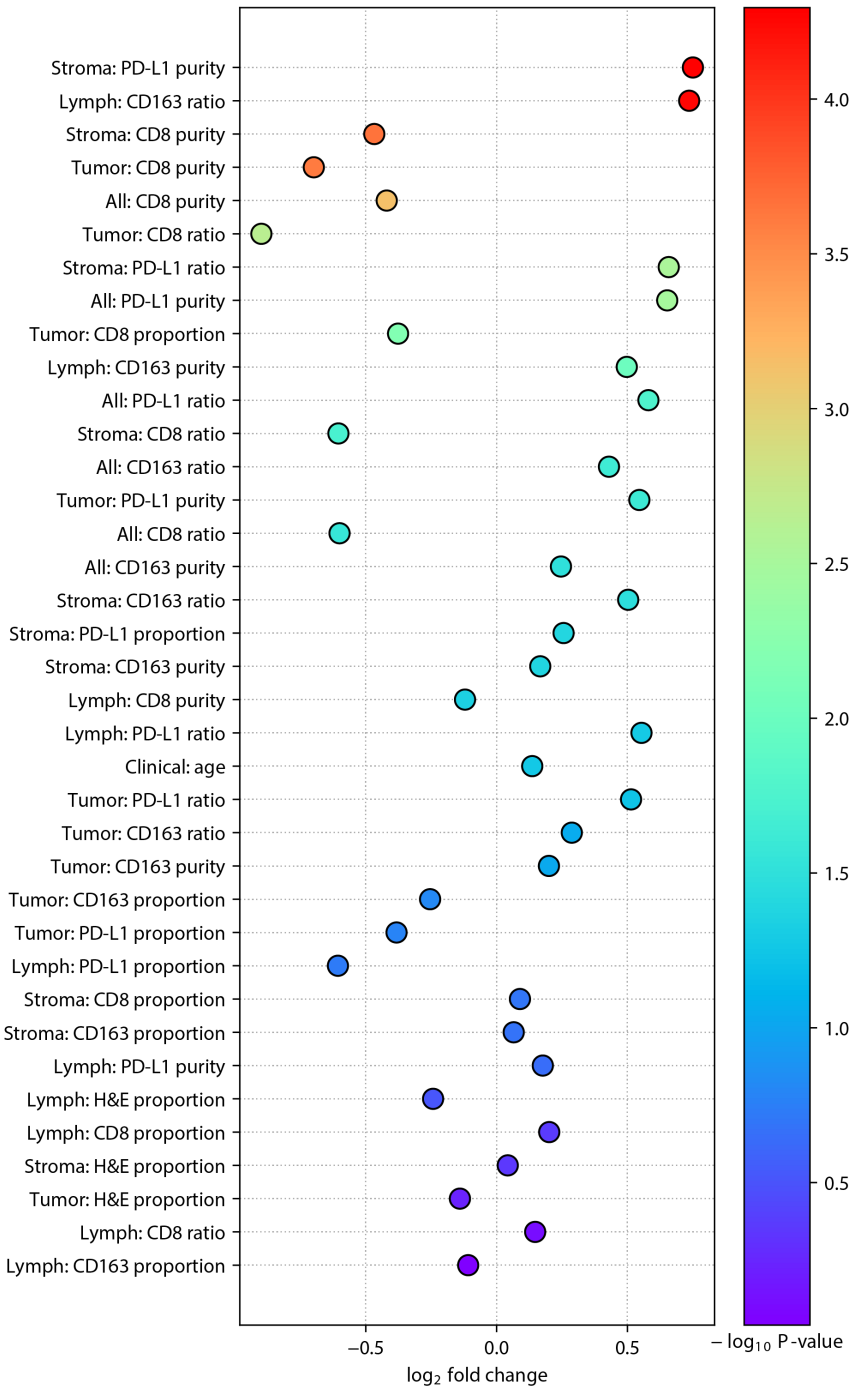


Figure 5.7. Comparison between HER2+ and TNBC cohorts among extracted IMPRESS and clinical features. Two-sided P-values are calculated based on Mann-Whitney U test. The fold changes are calculated by the ratio of the median feature values between HER2+ and TNBC cohorts. Figure best viewed in color.

with RCB significantly. In contrast, *Stroma: CD8 proportion* is positively correlated with RCB.

One study by Meisel *et al.* [325] suggested that TILs associated with RCB in HER2+ breast cancer NAC patients. In our results, we further demonstrate the association between TILs and RCB using correlation analysis, especially the *Lymph: PD-L1 ratio*. In addition, the inverse relations are detected between RCB scores and CD8+ TIL in Miyashita *et al.* [326], which also agreed with our findings such as *Lymph: CD8 ratio* in the HER2+ cohort and *Lymph: CD8 proportion* (Table 5.6).

These results suggest that IMPRESS features from pre-NAC images can also infer RCB values in a quantitative manner. For example, *Lymph: PD-L1 ratio* (favorable marker) and *Stroma: CD8 proportion* (adverse marker) are two common features that are significantly correlated with RCB in both the HER2+ and TNBC cohorts.

5.3.8 Reliability Results of IMPRESS Feature Extraction Workflow

Tissue Segmentation Results for H&E and IHC

The H&E tissue segmentation produced four regions of interests: stromal region (*Stroma*), tumoral region (*Tumor*), lymphocytes aggregated region (*Lymph*), and exclude region (*Exclude*). Each cohort has pathologist labeled 25 patches in 20× magnification, each with 512×512 pixels. The dice coefficients in HER2+ cohort for each class are 0.9312 (stromal region), 0.8413 (tumoral region), 0.7035 (lymphocytes aggregated region), and 0.8482 (exclude region). The mean dice coefficient in HER2+ cohort is 0.8311. The dice coefficients in TNBC cohort for each class are 0.9140 (stromal region), 0.7576 (tumoral region), 0.7323 (lymphocytes aggregated region), and 0.8752 (exclude region). The mean dice coefficient in TNBC cohort is 0.8198. The confusion matrix for HER2+ and TNBC cohorts are also reported in Table 5.7.

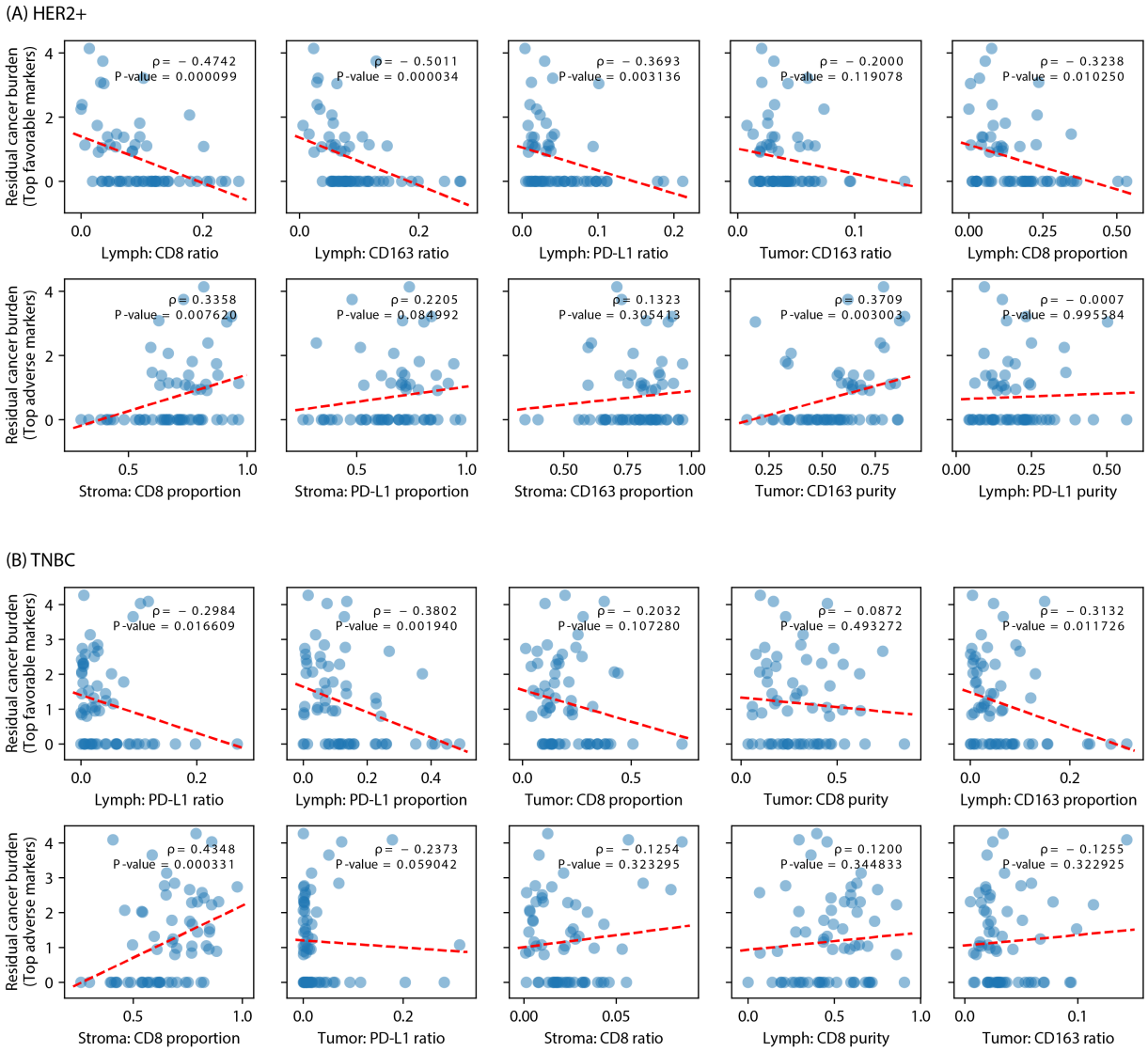


Figure 5.8. Scatter plot with Spearman's rank correlation coefficient ρ and P-value between IMPRESS features and residual cancer burden (RCB). (A) HER2+ cohort, first row: top 5 favorable IMPRESS features in Figure 5.4A; second row: top 5 adverse IMPRESS features in Figure 5.4A. (B) TNBC cohort, first row: top 5 favorable IMPRESS features in Figure 5.4B; second row: top 5 adverse IMPRESS features in Figure 5.4B. Dashed red lines represent the fitted linear regression curves.

Table 5.7. Confusion matrix in H&E segmentation results for HER2+ (A) and TNBC (B). *Exclude*: excluded region; *Stroma*: stromal region; *Tumor*: tumoral region; *Lymph*: lymphocytes aggregated region.

(A) HER2+

		Label predicted			
		Exclude	**Stroma**	**Tumor**	**Lymph**
Label truth	**Exclude**	1013128	246813	20645	196
	Stroma	51214	3966505	52017	8137
	Tumor	43656	195743	853359	2008
	Lymph	11	32339	7855	59974

(B) TNBC

		Label predicted			
		Exclude	**Stroma**	**Tumor**	**Lymph**
Label truth	**Exclude**	1204848	140546	73137	375
	Stroma	113180	3596604	175049	2014
	Tumor	16424	213614	822769	17049
	Lymph	0	32729	31204	114058

The IHC marker segmentation produced four regions of interests: CD8 region, CD163 region, PD-L1 region, and exclude region. Each cohort has pathologists labeled 5 patches in 20× magnification, each with 512 × 512 pixels. The dice coefficients in HER2+ cohort for each class are 0.8422 (CD8 region), 0.7379 (CD163 region), 0.7669 (PD-L1 region), and 0.9506 (exclude region). The mean dice coefficient in HER2+ cohort is 0.7823. The dice coefficients in TNBC cohort for each class are 0.8608 (CD8 region), 0.7500 (CD163 region), 0.7237 (PD-L1 region), and 0.9693 (exclude region). The mean dice coefficient in TNBC cohort is 0.7782. The confusion matrix for HER2+ and TNBC cohorts are also reported in Table 5.8.

Table 5.8. Confusion matrix in IHC segmentation results for HER2+ (A) and TNBC (B). Exclude: excluded background region; CD8: CD8 region; CD163: CD163 region; PD-L1: PD-L1 region.

(A) HER2+

		Label predicted			
		Exclude	**CD8**	**CD163**	**PD-L1**
Label truth	**Exclude**	980285	8799	17195	16166
	CD8	15168	83190	120	2785
	CD163	19796	111	54759	1068
	PD-L1	24832	4186	607	81653

(B) TNBC

		Label predicted			
		Exclude	**CD8**	**CD163**	**PD-L1**
Label truth	**Exclude**	1067720	4508	6006	9970
	CD8	11948	59050	37	1520
	CD163	16463	50	49179	7365
	PD-L1	18702	1037	2867	54298

Non-Linear Registration Results

A non-linear registration process is performed on each tissue for every pair of tissues in WSI. The total number of pathologists labelled landmark correspondences are 516 for HER2+ cohort and 304 for TNBC cohort. The evaluation performances including mean and median of the distance (in μm) and median $rTRE$ [119] before and after registration are reported in Table 5.9. We consider the registration is adequate if the median distance (μm) is $\leq 50\mu m$. From the results, we find the distances before and after registration in HER2+ cohort are $374.01\mu m$ and $33.31\mu m$ in mean, or $278.73\mu m$ and $18.23\mu m$ in median. The distances before and after registration in TNBC cohort are $627.66\mu m$ and $47.78\mu m$ in mean, or $48.14\mu m$ and $27.13\mu m$ in median. Both results in HER2+ and TNBC cohorts suggest the paired pathology images are aligned adequately. An example H&E tissue and the corresponding IHC tissue with 36 landmark correspondence pairs are demonstrated in Figure 5.9AB.

Table 5.9. Non-linear registration performances of HER2+ and TNBC cohorts. The median $rTRE$ is aggregated within each tissue image according to [119].

		Before registration		**After registration**	
Cohort	Metric	Mean	Median	Mean	Median
HER2+	distance (μm)	374.01	278.73	33.31	18.23
	median $rTRE$	6.04	5.35	0.59	0.40
TNBC	distance (μm)	627.66	482.14	47.78	27.13
	median $rTRE$	6.46	5.85	0.44	0.27

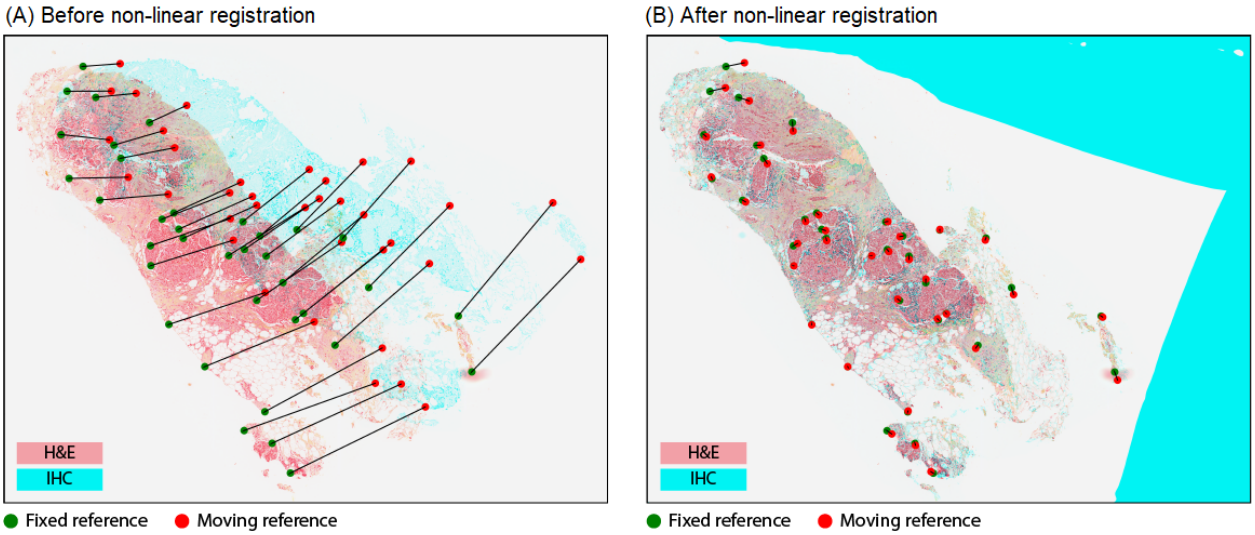


Figure 5.9. An example H&E tissue (fixed reference) and the corresponding IHC tissue (moving reference) before the non-linear registration (A) and after the non-linear registration (B). Figure best viewed in color.

5.3.9 Visualization of Representative Patches

The top IMPRESS features with favorable or adverse prognostic values are further exploited with their associated image patches. In each cohort, the cases with highest image feature value are selected, and their WSIs are sliced into patches with 1024×1024 pixels. Representative patch are presented within that specific patient's WSI and are shown in Figure 5.10. Figure 5.10A presents representative patches in HER2+ with top important

features; Figure 5.10B presents representative patches in TNBC with top importance features. The adverse prognostic markers are highlighted in grey backgrounds. These results help to visualize typical image patches where the top important features are enriched.

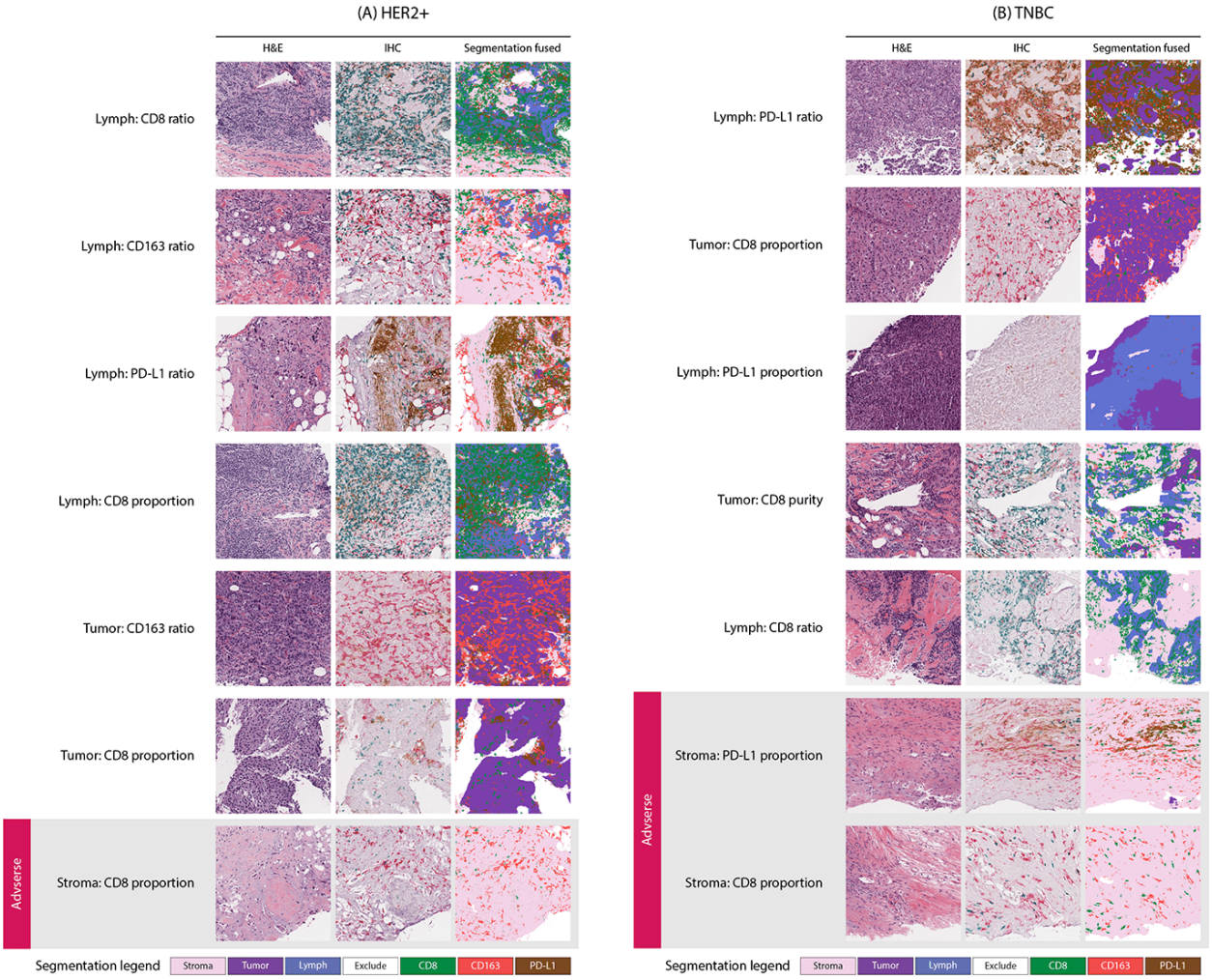


Figure 5.10. Selected representative patches in HER2+ cohort (A) and TNBC cohort (B). Patches are derived from patient WSIs which achieved highest IMPRESS feature values among cohorts. Adverse prognostic markers are highlighted. Figure best viewed in color.

5.3.10 Correlation Analyses Disclose Latent Dependencies among IMPRESS Features

To fully investigate the relationships and unveil the latent dependencies among IMPRESS features, pair-wised SCC are further conducted on both breast cancer cohorts (Figure 5.11). These pair-wised SCC ρ demonstrate the latent relationships between each pair of IMPRESS features.

All SCC ρ for area ratio features are positive. We are interested in those highly correlated area ratio features from different IHC markers. For area ratio in HER2+ (Figure 5.11B), the most correlated ratio statistics from different IHC markers are *Stroma: PD-L1 ratio* and *All: CD163 ratio* ($\rho = 0.73$, P-value $= 2.29 \times 10^{-11}$); *Stroma: CD163 ratio* and *Stroma: PD-L1 ratio* ($\rho = 0.72$, P-value $= 6.64 \times 10^{-11}$). For area ratio in TNBC (Figure 5.11F), the most correlated ratio statistics from different IHC markers are *Stroma: PD-L1 ratio* and *Stroma: CD8 ratio* ($\rho = 0.71$, P-value $= 3.82 \times 10^{-11}$); *All: PD-L1 ratio* and *Stroma: CD8 ratio* ($\rho = 0.71$, P-value $= 3.64 \times 10^{-11}$); *All: PD-L1 ratio* and *All: CD8 ratio* ($\rho = 0.71$, P-value $= 6.88 \times 10^{-11}$). The results of area ratio statistics suggest that the area ratio of PD-L1 has the strongest association with CD163 in HER2+, but has the strongest association with CD8 in TNBC.

For the proportion statistics in IMPRESS features, positive correlations are observed within same H&E regions. In contrast, negative correlations are observed across different H&E regions (Figure 5.11CD). We are interested in those features that from different H&E regions with most negative correlations. In HER2+ (Figure 5.11C), the most negatively correlated proportion statistics are *Tumor: H&E proportion* and *Stroma: H&E proportion* ($\rho = -0.92$, P-value $= 1.57 \times 10^{-25}$); *Tumor: CD163 proportion* and *Stroma: CD163 proportion* ($\rho = -0.83$, P-value $= 4.18 \times 10^{-17}$). In TNBC (Figure 5.11G), the most negatively correlated proportion statistics are *Tumor: H&E proportion* and *Stroma: H&E proportion* ($\rho = -0.95$, P-value $= 3.01 \times 10^{-33}$); *Tumor: CD163 proportion* and *Stroma: CD163 proportion* ($\rho = -0.88$, P-value $= 2.52 \times 10^{-21}$). The results of proportion statistics suggest that CD163 is the most negatively correlated IHC marker populated at either tumoral or stromal region.

For the purity statistics in IMPRESS features, positive correlations are observed within same IHC markers. In contrast, negative correlations are observed across different IHC markers (Figure 5.11DH). We are interested in those features that from different IHC markers with most negative correlations. In HER2+ (Figure 5.11D), the most negatively correlated purity statistics from different IHC markers are *Lymph: CD163 purity* and *Lymph: CD8 purity* ($\rho = -0.79$, P-value $= 1.96 \times 10^{-14}$); *Stroma: CD163 purity* and *All: CD8 purity* ($\rho = -0.73$, P-value $= 1.05 \times 10^{-11}$); *Tumor: CD163 purity* and *Tumor: PD-L1 purity* ($\rho = -0.73$, P-value $= 2.46 \times 10^{-11}$). In TNBC (Figure 5.11H), the most negatively correlated purity statistics from different IHC markers are *Stroma: CD163 purity* and *Stroma: CD8 purity* ($\rho = -0.77$, P-value $= 7.85 \times 10^{-14}$); *Stroma: CD163 purity* and *All: CD8 purity* ($\rho = -0.75$, P-value$= 7.18 \times 10^{-13}$). The results of purity statistics suggest that CD163 and CD8 are two most distinct IHC markers that populated against each other among various H&E regions.

5.4 Discussion and Conclusion

Recently, AI-based computational pathology methods based on tumor morphology have been developed to predict the clinical outcome including survival [176], [275]. Additionally, evaluating cell-level features in tumor immune micro-environment such as tumor-infiltrating lymphocyte (TIL) in pre-treatment breast cancer biopsies to predict NAC outcomes is also imperative and can contribute to potential clinical guidelines and treatment intervention. To the best of our knowledge, this is the first study to provide an automatic, accurate, comprehensive, interpretable, and reproducible whole slide image (WSI) feature extraction workflow, and to quantitatively evaluate the histopathological features extracted from H&E-stained and IHC-stained WSIs and predict NAC outcomes using machine learning model based on features derived from tumor itself and tumor immune micro-environment.

This study has several strength and advantages. First, analyzing histopathologic images is one of the most difficult machine learning tasks, hindered by the large size of the microscopy images [327]. Studies usually sliced WSIs into several small patches [327], while different choices of patch sizes can increase the uncertainties of models and performances.

In this study, the IMPRESS features are assessed on WSI level, which is a more robust and reproducible feature extraction workflow.

Second, the dataset was provided by collaborators at The Ohio State University, with two breast cancer subtypes including 62 HER2-positive breast cancer and 64 triple negative breast cancer patients. This dataset fully utilizes two consecutive whole slide images for each patient that are further stained with H&E and IHC. Multicolor multiplex immmunohistochemistry staining is a recent and well-established technology [328], which offers tremendous insights and understanding into complex tumor-immune microenvironment and cancer heterogeneity [329]. The multicolor multiplex immunohistochemical assays [314] enabled us to co-localize CD8, CD163, and PD-L1 markers, and study the relationships between those cell markers and NAC response. Since the image registration process is quantitatively robust and accurate, the extra information derived from IHC-stained WSIs can provide detailed tumor immune micro-environment information complimentary to the tumor H&E images. Different from other methods that only relied on H&E WSIs to extract lymphocytes, the identification of CD8, CD163, and PD-L1 provided extra information, which helped to better characterize the tumor immune micro-environment.

Third, AI is suggested to be an automatic approach for providing a potential clinical guideline. However, many AI-based methods are limited by their poor interpretability and unpredictable performance, especially when the sophisticated end-to-end learning methods are used. Our experiments not only demonstrated that the AI-based automatic feature extraction workflow has the capacity to generate interpretable IMPRESS features, but can also predict NAC outcome equally or more accurate than the model based on pathologists' assessed features. Last but not the least, many feature extraction methods are based on pathologists' manual assessments (*e.g.,* Ali *et al.* [300]; Hwang *et al.* [301]). The features assessed by pathologists conveyed rich interpretable explanations, however, they are difficult to reproduce with consistent quality. Instead, our automatic feature extraction workflow provides abundant reproducible interpretable features (36 IMPRESS features), and is also proved to outperform pathologists' assessment in HER2+ cohort (or have equal performances in TNBC cohort) using the logistic regression model.

In current study, we also investigate the association of clinicopathologic features from pre-treatment biopsies with response to NAC in two different breast cancer subtypes, HER2+ BC and TNBC. Previous study [306] found that the increased TIL concentration can predict response to neoadjuvant chemotherapy and survival but differences were observed between HER2+ and TNBC subtypes. In our results, we find several common and different feature behaviors across those two breast cancer subtypes, suggesting that breast cancer is immunogenic [306] and TILs might target differently in different breast cancer subtypes.

Our study has also demonstrated the relationship between several tumor immune micro-environment features and pCR. One of the most interesting findings is PD-L1 expression in pre-treatment tumor immune micro-environment, especially in TNBC cohort. It has been reported that the upregulation of PD-L1 is involved in various cellular processes in cancer cells as well as interactions between cancer cells and immune cells [330]–[332]. It has been conflicting whether PD-L1 expression is a favorable or adverse prognostic factor for breast cancer patients' survival [333]–[340]. The conflicting conclusions may result from the differences in composition of cohorts, PD-L1 antibody clones, or assessment methods (most studies used manual assessment). Kong *et al.* [341] suggests that PD-L1 expression at different locations had different impact on survival in colorectal cancer (CRC) patients, and shows that the total PD-L1 expression is a favorable prognostic marker. In our study, we observed similar behavior of high TIL and PD-L1 expression. For example, PD-L1 in lymphocytes aggregated region is found to associate with a favorable response to NAC.

Furthermore, our data has also demonstrated that the most important IMPRESS features identified from the logistic regression model to predict pCR (such as CD8, CD163, and PD-L1 ratios in lymphocytes aggregated region, and CD8 proportion in lymphocytes aggregated region) also correlated with RCB, at least partially.

In summary, we constructed an automatic, accurate, comprehensive, interpretable, and reproducible WSI feature extraction workflow (IMPRESS) and used these IMPRESS features to develop machine learning model to accurately predict the response to NAC in breast cancer patients.

With the aim of integrating multimodal biomedical data to unveil latent gene interactions and predict post-treatment response based on imaging data through machine learn-

ing, three research practices are performed. Specifically, survival analysis is combined with gene co-expression network analysis and deep learning in Chapter 3, and is combined with non-negative matrix factorization in Chapter 4. An automatic, accurate, comprehensive, interpretable, and reproducible workflow is proposed to predict pCR given histopathologic images in Chapter 5. In the next chapter, a software tool "TSUNAMI" based on R language will be introduced.

Table 5.4. Student's t-test results by comparing IMPRESS and clinical features of pCR cases against residual tumor cases. Features are sorted by P-values in ascending order.

	HER2+			**TNBC**		
Rank	Feature names	t statistic	P-value	Feature names	t statistic	P-value
1	*Lymph: CD163 ratio*	3.594455	0.000658	*Stroma: CD8 proportion*	−4.666563	0.000017
2	*Lymph: CD8 ratio*	3.541561	0.000776	*Stroma: PD-L1 proportion*	−4.242431	0.000075
3	*Clinical: HER2/CEP17 ratio*	3.226637	0.002030	*Lymph: PD-L1 ratio*	3.517819	0.000821
4	*Stroma: CD8 proportion*	−3.072326	0.003192	*Lymph: PD-L1 proportion*	3.488683	0.000899
5	*Lymph: PD-L1 ratio*	3.042724	0.003476	*Lymph: CD163 proportion*	3.376036	0.001274
6	*Tumor: CD163 purity*	−2.971149	0.004263	*Stroma: CD163 proportion*	−3.076857	0.003112
7	*Clinical: ER%*	−2.902422	0.005173	*Lymph: CD8 proportion*	2.948188	0.004502
8	*Lymph: CD8 proportion*	2.825048	0.006410	*Stroma: H&E proportion*	−2.830533	0.006256
9	*Clinical: PR%*	−2.810583	0.006669	*Lymph: PD-L1 purity*	2.736953	0.008080
10	*Clinical: PR*	−2.741490	0.008048	*Tumor: CD8 proportion*	2.712844	0.008624
11	*Lymph: CD163 proportion*	2.370234	0.021008	*Lymph: H&E proportion*	2.552162	0.013184
12	*Clinical: ER*	−2.353331	0.021901	*Tumor: PD-L1 proportion*	2.166371	0.034135
13	*All: CD163 ratio*	2.341825	0.022527	*Tumor: H&E proportion*	2.160806	0.034582
14	*Stroma: CD8 ratio*	2.278235	0.026288	*Clinical: age*	−2.055501	0.044049
15	*Tumor: PD-L1 purity*	2.255415	0.027767	*Tumor: CD163 purity*	−1.975628	0.052652
16	*Stroma: CD163 ratio*	2.249187	0.028183	*Stroma: PD-L1 ratio*	1.920414	0.059408
17	*All: CD8 ratio*	2.247796	0.028277	*All: PD-L1 purity*	1.890408	0.063379
18	*All: CD163 purity*	−2.205629	0.031252	*Lymph: CD163 ratio*	1.827774	0.072396
19	*All: PD-L1 ratio*	2.204862	0.031309	*Tumor: PD-L1 purity*	1.819611	0.073647
20	*Clinical: age*	−2.078504	0.041948	*All: PD-L1 ratio*	1.788630	0.078560
21	*Stroma: PD-L1 proportion*	−2.035503	0.046224	*Stroma: PD-L1 purity*	1.695521	0.094993
22	*Stroma: PD-L1 ratio*	2.034231	0.046356	*Tumor: CD163 proportion*	1.685473	0.096925
23	*Tumor: CD8 ratio*	1.770354	0.081748	*All: CD163 purity*	−1.659165	0.102133
24	*Tumor: PD-L1 ratio*	1.665460	0.101035	*Stroma: CD163 ratio*	1.549523	0.126346
25	*Lymph: PD-L1 proportion*	1.661584	0.101813	*All: CD163 ratio*	1.512692	0.135439
26	*Tumor: CD163 ratio*	1.656927	0.102755	*Lymph: CD8 ratio*	1.496247	0.139663
27	*Lymph: H&E proportion*	1.505547	0.137430	*Tumor: CD8 ratio*	1.354115	0.180615
28	*Stroma: CD163 purity*	−1.481817	0.143622	*Stroma: CD163 purity*	−1.161706	0.249808
29	*All: PD-L1 purity*	1.449248	0.152475	*Lymph: CD8 purity*	−1.148333	0.255242
30	*Stroma: CD163 proportion*	−1.328562	0.189024	*Tumor: PD-L1 ratio*	1.132209	0.261907
31	*Stroma: H&E proportion*	−1.312893	0.194218	*All: CD8 ratio*	1.015053	0.314025
32	*All: CD8 purity*	1.286521	0.203203	*Lymph: CD163 purity*	−0.878047	0.383309
33	*Tumor: CD8 proportion*	1.242422	0.218915	*Tumor: CD163 ratio*	0.695981	0.489041
34	*Tumor: CD8 purity*	1.179454	0.242873	*Tumor: CD8 purity*	0.561242	0.576655
35	*Tumor: PD-L1 proportion*	1.156670	0.251992	*Stroma: CD8 ratio*	0.550337	0.584066
36	*Stroma: PD-L1 purity*	1.005040	0.318914	*All: CD8 purity*	0.349839	0.727645
37	*Stroma: CD8 purity*	0.938257	0.351874	*Stroma: CD8 purity*	0.148199	0.882666
38	*Tumor: H&E proportion*	0.830702	0.409435			
39	*Lymph: CD163 purity*	−0.817853	0.416677			
40	*Lymph: CD8 purity*	0.630844	0.530539			
41	*Lymph: PD-L1 purity*	0.246588	0.806068			
42	*Tumor: CD163 proportion*	−0.043845	0.965174			

Table 5.6. Spearman's rank correlation coefficient statistics between IMPRESS features and residual cancer burden (RCB) values in HER2+ and TNBC cohorts. Features are sorted by P-values in ascending order.

HER2+			**TNBC**		
Feature name	Spearman ρ	P-value	Feature name	Spearman ρ	P-value
Lymph: CD163 ratio	−0.50105	3.36×10^{-5}	*Stroma: CD8 proportion*	0.434782	0.000331
Lymph: CD8 ratio	−0.47416	9.90×10^{-5}	*Stroma: PD-L1 proportion*	0.398219	0.001119
Tumor: CD8 ratio	−0.4599	0.00017	*Lymph: PD-L1 proportion*	−0.38025	0.00194
All: CD8 ratio	−0.44273	0.000314	*Lymph: CD8 proportion*	−0.31836	0.010355
Stroma: CD8 ratio	−0.38132	0.00223	*Lymph: CD163 proportion*	−0.31324	0.011726
Tumor: PD-L1 ratio	−0.37483	0.002686	*Lymph: H&E proportion*	−0.30888	0.013013
Tumor: CD163 purity	0.370868	0.003003	*Lymph: PD-L1 ratio*	−0.29843	0.016609
Lymph: PD-L1 ratio	−0.36932	0.003136	*All: PD-L1 ratio*	−0.25149	0.045004
All: PD-L1 ratio	−0.36553	0.003484	*All: PD-L1 purity*	−0.24423	0.051788
Stroma: PD-L1 ratio	−0.36338	0.003697	*Stroma: H&E proportion*	0.240374	0.055717
Stroma: CD8 proportion	0.335825	0.00762	*Tumor: PD-L1 ratio*	−0.23728	0.059042
Lymph: CD8 proportion	−0.3238	0.01025	*Lymph: PD-L1 purity*	−0.23628	0.06015
Lymph: CD163 proportion	−0.32058	0.011074	*Lymph: CD163 ratio*	−0.22412	0.075025
Tumor: PD-L1 purity	−0.28861	0.02291	*Stroma: PD-L1 ratio*	−0.2205	0.079975
All: CD163 purity	0.28293	0.025867	*Tumor: PD-L1 purity*	−0.21162	0.09322
All: CD163 ratio	−0.27395	0.031196	*Stroma: PD-L1 purity*	−0.20862	0.098063
Stroma: CD163 ratio	−0.24766	0.052293	*Tumor: CD163 purity*	0.207548	0.099839
Tumor: CD8 purity	−0.24134	0.058798	*Tumor: CD8 proportion*	−0.20322	0.10728
All: CD8 purity	−0.2375	0.063072	*Stroma: CD163 proportion*	0.19617	0.120289
Lymph: H&E proportion	−0.22949	0.072772	*Tumor: PD-L1 proportion*	−0.196	0.12061
Stroma: PD-L1 proportion	0.220535	0.084992	*Tumor: H&E proportion*	−0.18429	0.144898
Lymph: PD-L1 proportion	−0.21755	0.089406	*Tumor: CD8 ratio*	−0.17189	0.174417
All: PD-L1 purity	−0.21502	0.09328	*All: CD163 purity*	0.167914	0.184748
Stroma: CD163 purity	0.200301	0.118538	*All: CD8 ratio*	−0.16539	0.19153
Tumor: CD163 ratio	−0.20001	0.119078	*All: CD163 ratio*	−0.15682	0.215895
Stroma: PD-L1 purity	−0.19823	0.122466	*Stroma: CD163 ratio*	−0.14011	0.269483
Tumor: CD8 proportion	−0.19614	0.126548	*Tumor: CD163 ratio*	−0.12554	0.322925
Stroma: CD8 purity	−0.17019	0.185999	*Stroma: CD8 ratio*	−0.12545	0.323295
Stroma: CD163 proportion	0.132281	0.305413	*Lymph: CD8 ratio*	−0.12516	0.324406
Tumor: PD-L1 proportion	−0.11584	0.369959	*Lymph: CD8 purity*	0.12002	0.344833
Lymph: CD163 purity	0.107283	0.406566	*Stroma: CD163 purity*	0.119401	0.347345
Stroma: H&E proportion	0.095343	0.46104	*Lymph: CD163 purity*	0.110618	0.384205
Lymph: CD8 purity	−0.09072	0.483141	*Tumor: CD163 proportion*	−0.09493	0.455567
Tumor: CD163 proportion	0.022932	0.859577	*Tumor: CD8 purity*	−0.08719	0.493272
Tumor: H&E proportion	−0.01271	0.921869	*All: CD8 purity*	−0.04556	0.720725
Lymph: PD-L1 purity	−0.00072	0.995584	*Stroma: CD8 purity*	−0.01326	0.917181

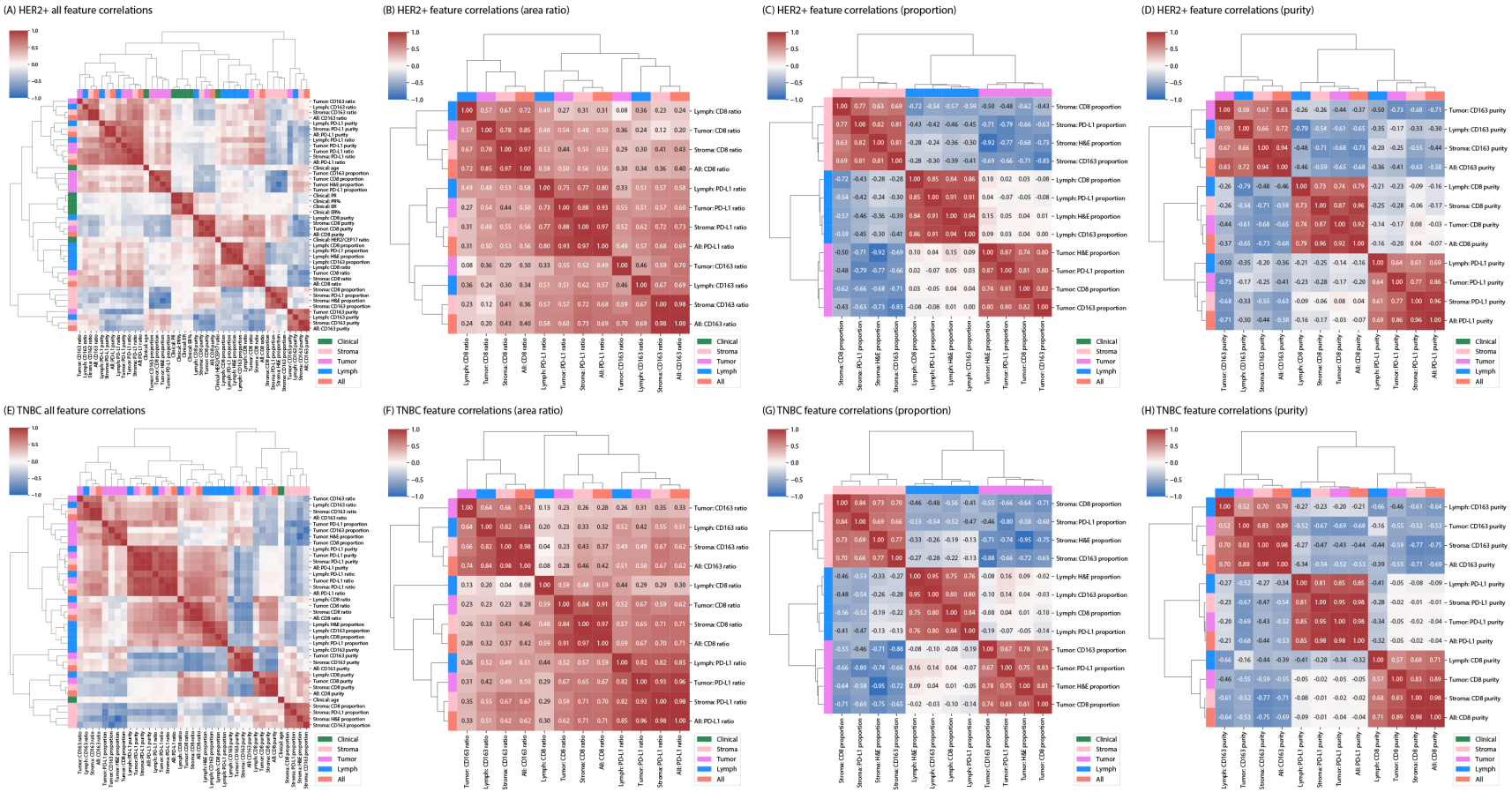


Figure 5.11. Correlation analyses for IMPRESS features in HER2+ (A–D) and TNBC (E–H) cohorts. (A) HER2+ all IMPRESS feature correlation matrix; (B) HER2+ area ratio correlation matrix; (C) HER2+ proportion correlation matrix; (D) HER2+ purity correlation matrix; (E) TNBC all IMPRESS feature correlation matrix; (F) TNBC area ratio correlation matrix; (G) TNBC proportion correlation matrix; (H) TNBC purity correlation matrix. Figure best viewed in color.

6. TSUNAMI: TRANSLATIONAL BIOINFORMATICS TOOL SUITE FOR NETWORK ANALYSIS AND MINING

Gene co-expression network (GCN) mining identifies gene modules with highly correlated expression profiles across samples/conditions. It enables researchers to discover latent gene or molecule interactions, identify novel gene functions, and extract molecular features from certain disease/condition groups, thus helping to identify disease biomarkers. However, there lacks an easy-to-use tool package for users to mine GCN modules that are relatively small in size with tightly connected genes that can be convenient for downstream gene set enrichment analysis, as well as modules that may share common members. To address this need, we develop an online GCN mining tool package: TSUNAMI (Tools SUite for Network Analysis and MIning). TSUNAMI incorporates lmQCM algorithm to mine GCN modules for both public and user-input data (microarray, RNA-seq, or any other numerical omics data), and then performs downstream gene set enrichment analysis for the identified modules. It has several features and advantages: 1) a user-friendly interface and real-time co-expression network mining through a web server; 2) direct access and search of NCBI GEO and TCGA databases, as well as user-input gene expression matrices for GCN module mining; 3) multiple co-expression analysis tools to choose from, all of which are highly flexible in regards to parameter selection options; 4) identified GCN modules are summarized to eigengenes, which are convenient for users to check their correlation with other clinical traits; 5) integrated downstream Enrichr enrichment analysis and links to other gene set enrichment tools; and 6) visualization of gene loci by Circos plot in any step of the process. The web service is freely accessible through URL: https://shiny.ph.iu.edu/TSUNAMI/.

6.1 Background and Introduction

Gene co-expression network (GCN) mining is a popular bioinformatics approach to identify densely connected gene modules, which are linked by their highly correlated expression profiles. It helps biologists discover latent gene/molecule interactions and identify novel gene functions, disease pathways, biomarkers, and insights for disease mechanisms. GCN mining

approaches such as WGCNA [35] and lmQCM [34] have been increasingly used [46], [51], [169]–[171]. Compared to the more popularly used WGCNA package, lmQCM is capable of mining smaller densely connected GCN modules. It also allows overlapping membership in the output modules. Such features are more consistent with biological networks in which the same genes may participate in multiple pathways, where a small group of genes are more likely to be synergistically regulated in local pathway functions. In addition, gene modules with smaller size derived from lmQCM usually generate more meaningful gene set enrichment results, which have been successfully applied to many diseases and cancer types [2], [40], [44], [174]–[176], [342], [343].

Currently, several online databases exist that curate transcriptomic data. For instance, PanglaoDB (https://panglaodb.se/) collects single-cell RNA-seq (scRNA-seq) data from mice and humans; Cao *et al.* scRNASeqDB [344] provides an scRNA-seq database for gene expression profiling in humans; and Recount2 [345] provides publicly available analysis-ready gene and exon counts datasets. However, all of these databases focus on data collection and curation. To the best of our knowledge, there is no tool offering the complete workflow that can directly process transcriptomic data, mine GCN modules, carry out gene set enrichment analysis, and provide visualization for the results. To meet such needs, we implement our web-based analysis tool suite TSUNAMI (Tools SUite for Network Analysis and MIning).

For users' convenience, mRNA-seq data from The Cancer Genome Atlas (TCGA, Illumina HiSeq RSEM genes normalized from https://gdac.broadinstitute.org/) and NCBI Gene Expression Omnibus (GEO) are directly incorporated into TSUNAMI. GEO hosts a large number of transcriptomic datasets generated from multiple platforms, including microarray and RNA-seq data. Other data types, such as miRNA-seq and DNA methylation, are also compatible with TSUNAMI. In fact, TSUNAMI can handle any numerical matrix data regardless of the omics data type. TSUNAMI not only incorporates the lmQCM algorithm, but it also includes the WGCNA package for users to explore and compare GCN modules generated from two different algorithms. We offer highly flexible parameter choices in each step to users who want to fine tune each algorithm to suit their own data and goal.

Prior to data mining, a data pre-processing interface has been designed to address differences in the input data formats and to filter the data in order to remove noise for GCN

mining. Each step of pre-processing is transparent to users and can be adjusted according to their preferences and needs.

Furthermore, our website directly incorporates enrichment analysis of the gene modules and Circos plot function for researchers to explore the enriched biological terms and gene locations in the output GCN modules. It also provides a tool for survival analysis with respect to each GCN module's eigengene values. All the aforementioned functions only require button clicks from users. The design of such a user-friendly interface seen in our TSUNAMI workflow provides a one-stop comprehensive analysis tool suite for biological researchers and clinicians to perform transcriptomic data analysis without any programming skill or data mining knowledge.

6.2 Functionality

6.2.1 Data Input

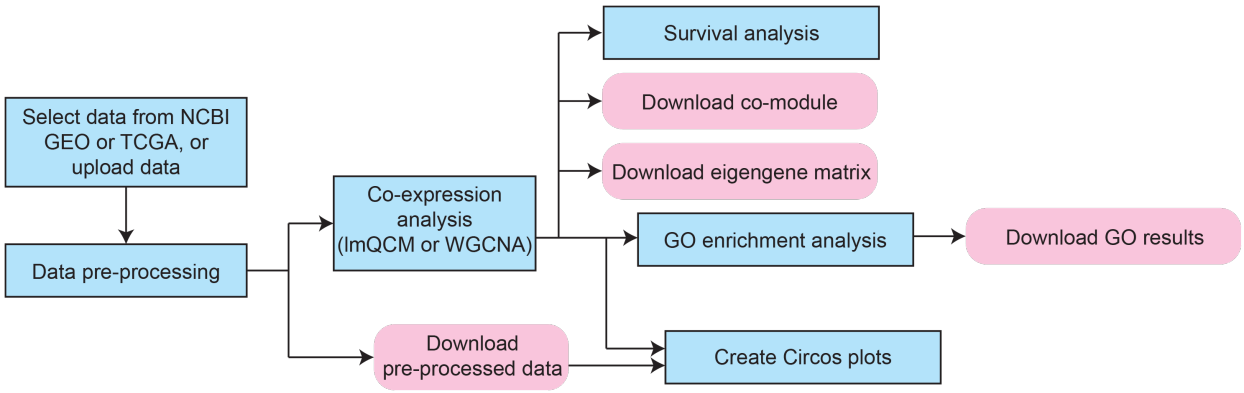


Figure 6.1. Flowchart of TSUNAMI. In this flowchart for TSUNAMI workflow, blue rectangles represent workflow operations; rounded rectangles in pink represent download processes.

A flowchart of the TSUNAMI workflow is presented in Figure 6.1. The entire workflow is implemented in R language with Shiny server pages. Some front-end interfaces and functions are implemented using JavaScript. With TSUNAMI, users can choose to use multiple types of data formats, including TCGA RNA-seq data, gene expression microarray data from GEO (in the format of GSE series matrix data), RNA-seq data from GEO, and user-defined numerical

matrix data, such as microarray, RNA-seq, scRNA-seq, and DNA methylation data. Instead of searching the GEO database manually, TSUNAMI provides a friendly interface for users to retrieve data from GEO by utilizing keywords and offers a flexible selection tool to retrieve a relevant GSE dataset to perform GCN analysis. Users can also choose a specific omics data type on the GEO database if keywords are entered in the search window to indicate the desired data type. On the website, a variety of example datasets ranging from microarray to scRNA-seq data are listed on TSUNAMI for users' reference. TSUNAMI also provides an upload bar for users to upload local files in various formats (e.g., CSV, TSV, XLSX, TXT). The data uploading interface is shown in 6.2A. In this paper, one microarray dataset (GSE17537 from GEO) is chosen as an example to demonstrate the features of TSUNAMI. GSE17537 contains gene expression data of 55 colorectal cancer patients from the Vanderbilt Medical Center (VMC) generated from the Affymetrix HU133 2.0 Plus Genechip with 54,675 probesets [346], [347].

6.2.2 Online Data Pre-processing

One issue of the microarray dataset from GEO is that different platforms adopted different rules when converting probeset IDs to gene symbols. To make this step easier for users, probeset IDs in GSE data matrix from GEO can be converted to gene symbols using R package "BiocGenerics" [348] by only one click. For instance, in the GSE17537 dataset, the annotation platform is GPL570. TSUNAMI then automatically identifies the annotation platforms of the data from GEO. During the conversion, TSUNAMI: (i) removes rows with empty gene symbols; and (ii) selects the rows with the largest mean expression value when multiple probesets are matched to the same gene symbol. The user interface of the data pre-processing step is shown in Figure 6.2B.

Additional data filtering steps include: (i) converting "NA" value (not a number value) to 0 in expression data, to ensure all the values are numeric and can be processed by co-expression algorithms; (ii) performing $\log_2(\boldsymbol{X}+1)$ transformation of the expression values $\boldsymbol{X}$ if the original values have not been previously transformed; (iii) removing lowest J_1 percentile rows (genes) with respect to mean expression values; and (iv) removing lowest J_2 percentile

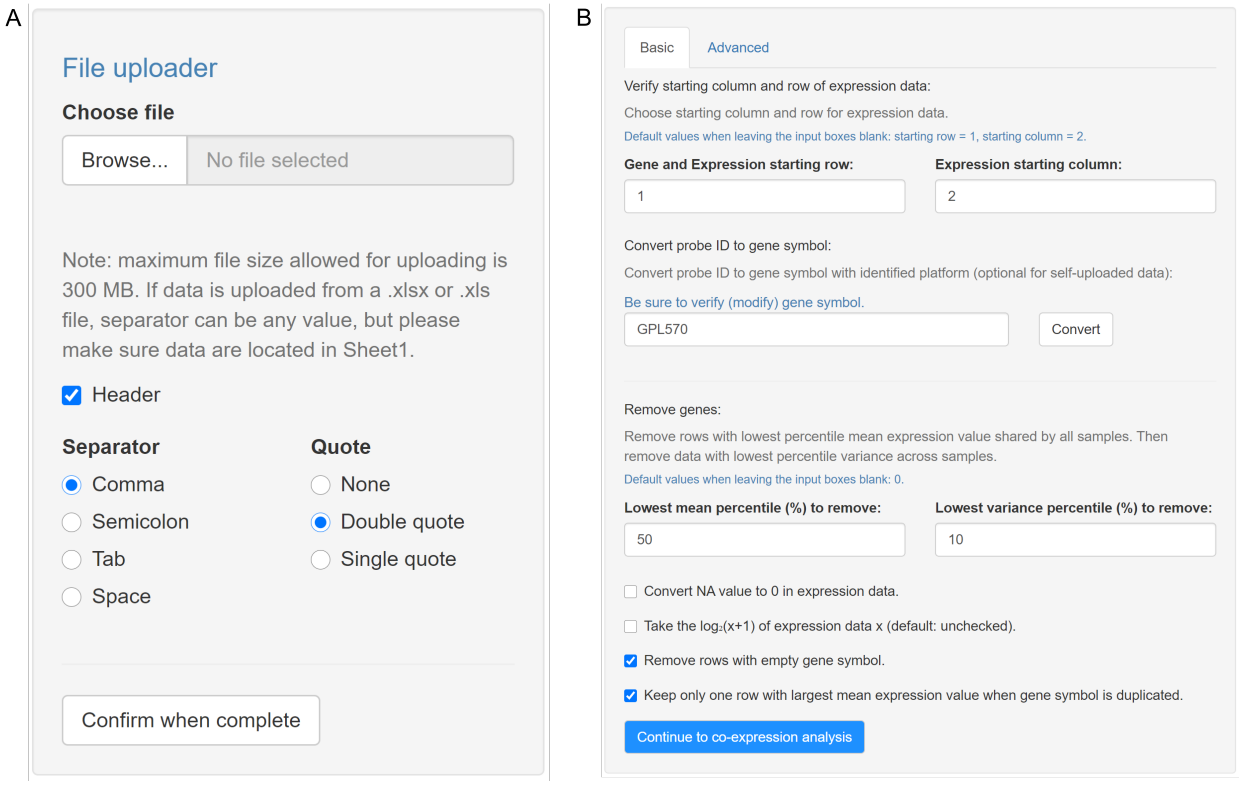


Figure 6.2. Dataset Selection and Pre-processing Panel. (A) Data can be uploaded manually or chosen from the NCBI GEO database. When uploading the data, the maximum file size that TSUNAMI allows is 300 Megabytes. Header, separators and quote methods can be adjusted by users. (B) The Data Pre-processing Panel includes several pre-processing steps.

rows with respect to expression values' variance. These data filtering steps are necessary to reduce noise and to ensure the robustness for the downstream correlational computation in the lmQCM algorithm. The default settings are $J_1 = 20$ and $J_2 = 20$, by which genes with low expression means and variances across samples are filtered out. In our example with GSE17537, we deselect logarithm conversion and "NA" value to 0 conversion and set $J_1 = 50$, and $J_2 = 10$, as shown in Figure 6.2B. However, users can always adjust these parameters based on their own needs and preferences. In the data pre-processing section, we further provide an "Advanced" panel to allow users to select a subgroup of samples of their interest. After the data pre-processing finishes, a dialog box appears to indicate how many genes are preserved after the filtering process.

6.2.3 Weighted Network Co-expression Analysis

After data pre-processing, users can directly download pre-processed data or further proceed to GCN analysis. In GCN analysis, we implement the lmQCM algorithm as well as the WGCNA workflow. The R package "WGCNA" from Bioconductor (https://bioconductor.org/) is adopted to integrate the WGCNA workflow. We keep the mining steps concise and simple with default parameter settings, while preserving the flexibility for users to select parameters in each step. Guidelines for parameter selection are in the Method pages of the website. In addition, we also release the lmQCM R package to CRAN (https://CRAN.R-project.org/package=lmQCM/).

In the lmQCM method panel, users can adjust parameters such as initial edge weight q_γ, weight threshold controlling parameters q_λ, q_t, q_β, and minimum cluster size (Figure 6.3). Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SCC) are implemented separately for users to select. SCC is recommended for analysing RNA-seq data due to the large range of data values, and it is more robust than PCC to outliers. In our example with GSE17537, positive correlations are analyzed and the default settings are used (unchecked weight normalization, $q_\gamma = 0.7$, $q_\lambda = 1$, $q_t = 1$, $q_\beta = 0.4$, minimum cluster size = 10, and PCC for correlation measure). The running time of lmQCM depends on the number of genes present after the filtering process. A progress bar is provided to show the program progress. Note that lmQCM will not work if the data contain no clustering structures or the gene pair correlations are so poor that none is above the initial mining starting threshold (q_γ). In those cases, the program will stop running and generate a warning message. However, this should not happen if the data contain enough highly correlated gene pairs after filtering and the default program settings are used.

The WGCNA method panel is a two-step analysis. Step 1 helps users to specify the hyper-parameter "power" in step 2, *i.e.*, the soft thresholding in [35] by visualizing the resulting plot (Figure 6.4A). Step 2 allows users to select the remaining parameters. TSUNAMI allows users to customize the parameters of power, reassign threshold, merge cut height, and indicate minimum module size. After applying WGCNA, a hierarchical clustering plot for the result modules is also shown in this panel (Figure 6.4B). The resulting plot in Figure

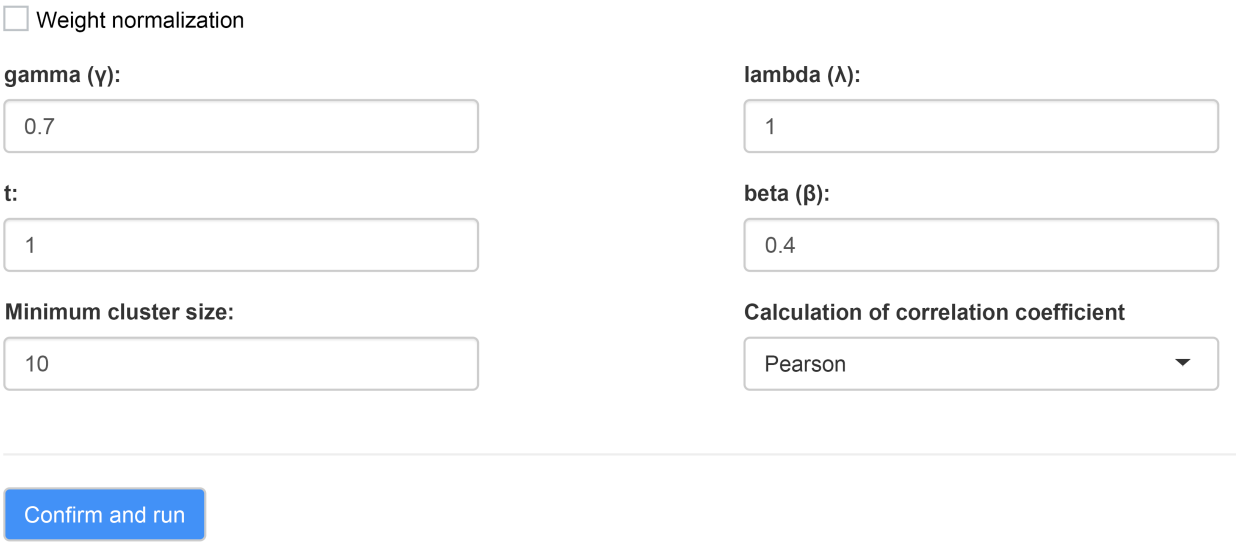


Figure 6.3. lmQCM Method Panel Data Pre-processing Panel. The lmQCM algorithm panel that allows users to choose a variety of parameters. In this paper, experiment runs with no weight normalization, $q_\gamma = 0.7$, $q_\lambda = 1$, $q_t = 1$, $q_\beta = 0.4$, minimum cluster size = 10, and Pearson correlation coefficient.

6.4B is from the example data GSE17537 with power = 10, set reassign threshold = 0, merge cut height = 0.25, and minimum module size = 10.

In the last step of GCN mining, two outputs are provided by TSUNAMI: (i) merged gene clusters sorted by their sizes in descending order (Figure 6.5A with lmQCM algorithm); and (ii) an eigengene matrix, which is the summarized expression values of genes in each GCN using the first principal component from singular value decomposition (Figure 6.5B with lmQCM algorithm). Eigengene values can be regarded as the weighted average expressions of each GCN. Such values are very useful for users to correlate GCN modules' expression profiles with various clinical and phenotypic traits in the downstream analysis, such as survival analysis. All results can be downloaded as files in CSV or TXT format.

6.2.4 Downstream Enrichment Analysis

Enrichr [183], [349] is used as the tool for downstream gene set enrichment analysis implementation. By default, a total of 14 types of frequently used enrichment analyses are

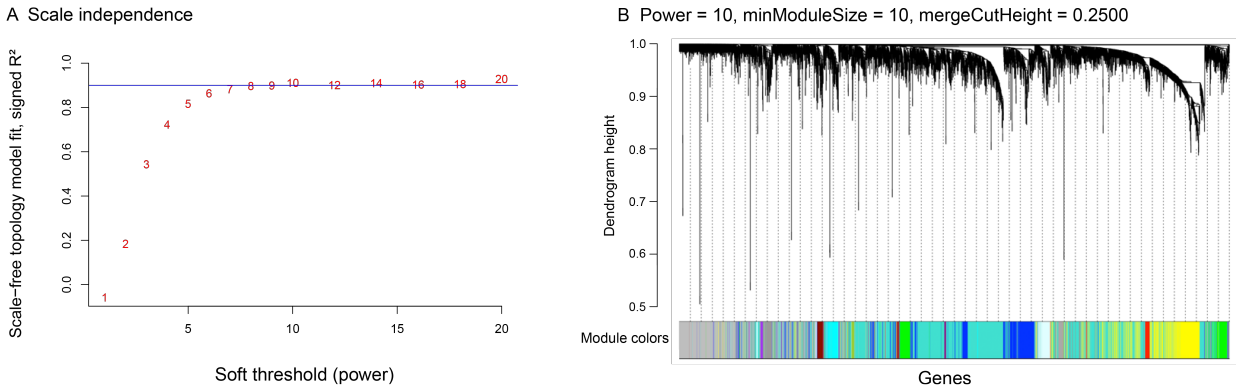


Figure 6.4. Choosing the Power in WGCNA and the Hierarchical Clustering Graph of WGCNA. (A) The hyper-parameter "power" chosen from the value above the blue horizontal line. (B) The hierarchical clustering graph with color bar indicating modules with GSE17537 dataset as an example; parameters for WGCNA are power = 10, reassign threshold = 0, merge cut height = 0.25, and minimum module size = 10.

performed: (1) Biological Process; (2) Molecular Function; (3) Cellular Component; (4) Jensen DISEASES; (5) Reactome; (6) KEGG; (7) Transcription Factor PPIs; (8) Genome Browser PWMs; (9) TRANSFAC and JASPAR PWMs; (10) ENCODE TF ChIP-seq; (11) Chromosome Location (Cytoband); (12) miRTarBase; (13) TargetScan microRNA; and (14) ChEA. Users can further customize the enrichment result categories from the open source code available in Github (https://github.com/huangzhii/TSUNAMI/).

To access Enrichr results, users can simply click the blue "GO" button in each row adjacent to the GCN mining results (as shown in Figure 6.5A). In each enrichment analysis, its output includes multiple results, such as the enriched term (*e.g.*, GO term or pathway), *P*-value, z-score, and overlapped genes. Users can download multiple analysis results that are bundled in a ZIP file. In addition, other popular gene set enrichment analysis websites are also directly linked in TSUNAMI to enhance convenience for users. In our example with GSE17537, we select the 36$^{\text{th}}$ GCN module with 15 genes generated by lmQCM to be analyzed for enrichment, and each result table is sorted based on the *P*-value generated by Enrichr. From the result in Table 6.1, we can see the 36$^{\text{th}}$ GCN module is highly enriched

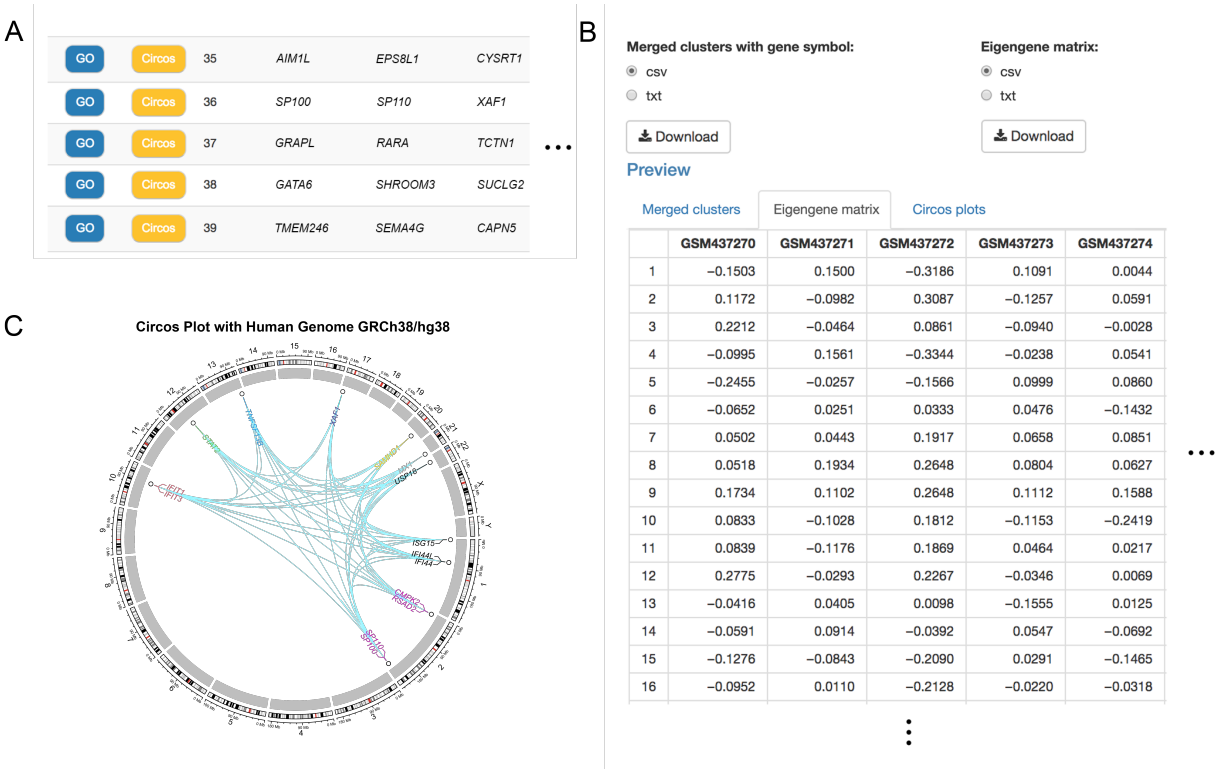


Figure 6.5. Merged clusters result generated by lmQCM. (A) The merged GCN modules, sorted in descending order based on the length of each cluster. Figure only shows part of the results (cluster 35 – 39) with part of genes. (B) The screenshot of the eigengene matrix (rounded to 4 decimal places for better visualization). Figure only shows part of the results (cluster 1 – 16) with part of samples (GSM437270 – GSM437274). (C) The Circos plot is resulted from the 36$^{\text{th}}$ GCN module with 15 genes. All modules in these subfigures are generated using the lmQCM algorithm with default parameters (unchecked weight normalization, $q_\gamma = 0.7$, $q_\lambda = 1$, $q_t = 1$, $q_\beta = 0.4$, minimum cluster size = 10, and Pearson correlation coefficient) with the GSE17537 dataset as an example.

in GO Biological Process term "type I interferon signaling pathway (GO:0060337)" (9 out of 148 genes).

6.2.5 Circos Plot

TSUNAMI provides Circos plots [350] through intermediate results or inputs in the cases of human transcriptomic data. Circos plots are very useful graphs for visualizing the positions

of genes on chromosomes and gene-gene relationships/interactions. The Circos plot function from the R package "circlize" [350] is adopted in this package for users to locate and visualize mined GCNs of human genes.

In TSUNAMI, users can visualize the Circos plot via "Circos Plots" section, either by typing their own gene list separated by the carriage return character ("\n") directly, or by using the calculated GCN modules (for example, by clicking the yellow button right next to the "GO" button in Figure 6.5A). TSUNAMI supports both human genomes hg38 (GRCh38) and hg19 (GRCh37). To match the gene symbol to starting and ending sites on a chromosome, we use the *refGene* database downloaded from the UCSC genome browser [351]. If multiple starting/ending sites are matched, we choose the longest one with length calculated by

$$\text{length} = |\text{ending\_site} - \text{starting\_site}| + 1. \tag{6.1}$$

By updating the plots, users can also choose the size of the plots and decide whether gene symbols and pair-wised links should be shown on the graph.

An example output of Circos plot is shown in Figure 6.5C used the 36^{th} GCN module with 15 genes from our previously discussed example (use a color set for texts to get a clear visual effect), annotated by gene symbols of human genome hg38 (GRCh38). The link between a pair of genes indicates that they belong to the same co-expressed GCN module.

Circos plots can help users visualize the locations of genes in a GCN on human chromosomes, thus enable them to identify GCNs due to copy number variation and other structural changes. In the future, genome from mice and other species will be incorporated for Circos plots.

6.2.6 Survival Analysis with respect to GCN Modules

An optional step of survival analysis follows the generation of the eigengene matrix. It allows users to correlate the GCN modules' eigengene values with patient survival time (or event-free survival). This extension of the tool can be further customized to correlate module eigengene values with other clinical traits in the future version. In our current version, we only implemented survival analysis as a starting point.

In the survival analysis, users can perform Overall Survival/Event-Free Survival (OS/EFS) analysis based on the GCN modules' eigengene values and look for GCN modules that are significantly associated with prognosis. Although, depending on the group of patients specified by users, such GCNs may not exist all the time. To carry out the analysis, users first select an eigengene (corresponding to a GCN module) in TSUNAMI. The program then splits the patients into two groups by the median of eigengene values. Next, it tests the two groups against OS/EFS by calculating the P-value of the log-rank test [352], [353]. Before doing so, users need to input the numerical survival time of OS/EFS (either in months or in days) with categorical events OS/EFS status (1: deceased; 0: censored). The "survdiff" function from R package "survival" is adopted to calculate the P-value and plot the Kaplan-Meier survival curves.

Taking GSE17537 with full survival information as an example, the Kaplan-Meier survival plot is generated according to the OS information by dichotomizing the 36^{th} GCN module's eigengene values at its median to high and low groups, as shown in Figure 6.6. Such a GCN module is generated from lmQCM method with default settings, as shown in 6.3. This survival analysis offers researchers a tool to immediately identify any GCN modules that are associated with patients' survival time, thus allowing researchers to further study their roles as potential prognosis biomarkers, as well as the biological pathways that differentiate the patients.

6.3 Conclusion

We release the online TSUNAMI tool package for gene co-expression modules identification with direct link to the TCGA RNA-seq datasets and the NCBI GEO database, while also accommodating users' input data. It is a one-stop comprehensive tool package with several advantages, such as flexibility in parameter selections, comprehensive GCN mining tools, direct link to downstream gene set enrichment analysis, Circos plot visualization, and survival analysis, with downloadable results in each step. All of these features bring tremendous convenience to study biomedical data.

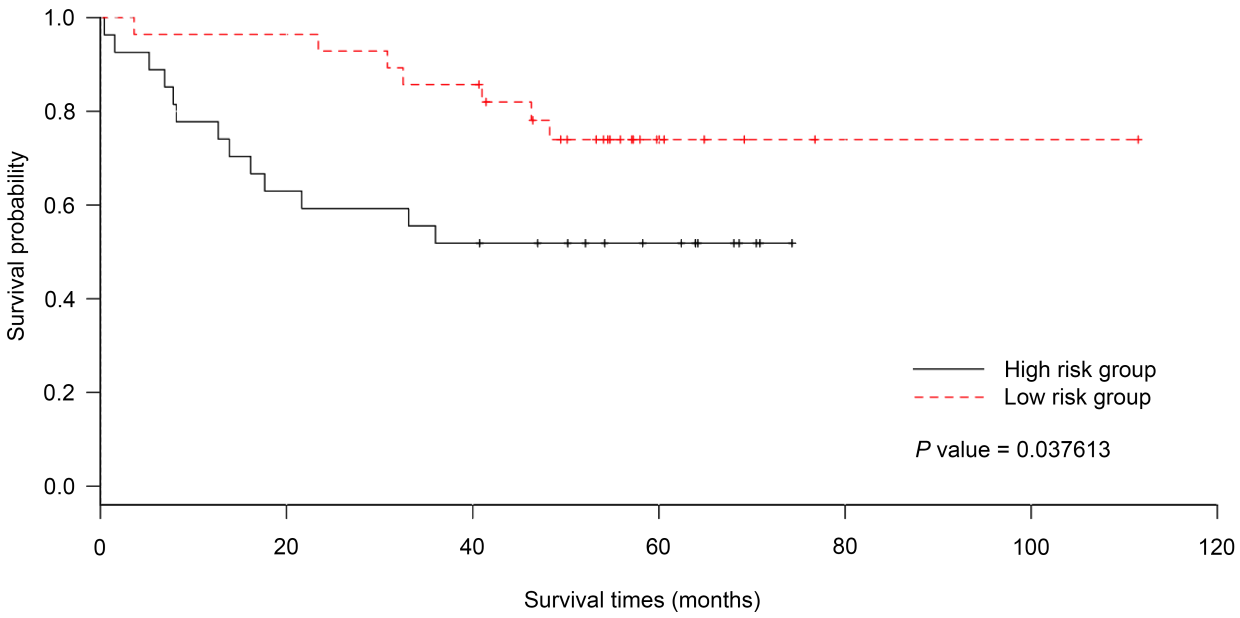


Figure 6.6. Survival analysis using the 36$^{\text{th}}$ GCN module eigenvalues generated from lmQCM algorithm, with default parameters (unchecked weight normalization, $q_\gamma = 0.7$, $q_\lambda = 1$, $q_t = 1$, $q_\beta = 0.4$, minimum cluster size = 10, and Pearson correlation coefficient) with GSE17537 series matrix as an example. 55 samples are used with Overall Survival information.

Table 6.1. The partial results of GO enrichment analysis. Note: This table contains partial rows and columns from original result (active panel: GO Biological Process) from the 36^{th} GCN module with 15 genes generated by lmQCM with GSE17537 series matrix as data. GO terms are sorted by P-value. We refer readers to explore other P-values and scores from TSUNAMI webpage and Enrichr package.

ID	Term	Overlap	P-value	Z-score	Genes
1	Type I interferon signaling pathway (GO:0060337).	9/148	2.51×10^{-16}	−3.2821	SP100; RSAD2; STAT2; MX1; ISG15; SAMHD1; XAF1; IFIT1; IFIT3
2	Cellular response to type I interferon (GO:0071357).	4/23	1.80×10^{-9}	−2.7766	SP100; MX1; ISG15; IFIT1
3	Negative regulation of single stranded viral RNA replication via double stranded DNA intermediate (GO:0045869).	4/44	2.73×10^{-8}	−2.6829	RSAD2; MX1; ISG15; IFIT1
4	Negative regulation of viral genome replication (GO:0045071).	4/40	1.84×10^{-8}	−2.4940	RSAD2; MX1; ISG15; IFIT1
5	Negative regulation by host of viral genome replication (GO:0044828).	4/51	5.01×10^{-8}	−2.6224	RSAD2; MX1; ISG15; IFIT1
6	Response to type I interferon (GO:0034340).	3/35	2.20×10^{-6}	−2.7155	SP100; MX1; ISG15
7	Regulation of type I interferon-mediated signaling pathway (GO:0060338).	3/43	4.14×10^{-6}	−2.6859	STAT2; SAMHD1; USP18
8	Negative regulation of type I interferon-mediated signaling pathway (GO:0060339).	2/43	4.66×10^{-4}	−2.5488	STAT2; USP18
9	Positive regulation of type I interferon-mediated signaling pathway (GO:0060340).	2/52	6.81×10^{-4}	−2.5122	STAT2; USP18
10	Positive regulation of Fas signaling pathway (GO:1902046).	1/7	5.24×10^{-3}	−2.9563	SP100

7. CONCLUSION AND FUTURE WORK

7.1 Overview

In this thesis, a deep learning based survival prognosis algorithm "SALMON", a non-negative matrix factorization in conjunct with Cox proportional hazards regression "CoxNMF", and an automatic feature extraction workflow "IMPRESS" are developed. In SALMON algorithm, the experimental results demonstrate the superiority of the proposed method in achieving higher concordance index as well as the power of identifying survival associated gene modules. The CoxNMF demonstrates its power of reconstructing the latent clusters with optimal silhouette score comparing to other baseline non-negative matrix factorization algorithms. The IMPRESS workflow can predict breast cancer neoadjuvant chemotherapy outcomes given pre-treatment biopsies. In addition, the bioinformatics and computational biology tool "TSUNAMI" is also introduced and explained. In Chapter 1, an introduction to biological data and a detailed overview of our problems is stated.

In Chapter 2, a brief introduction to survival analysis, co-expression network analysis, non-negative matrix factorization, histopathological image process, and related evaluation metrics is elaborated. The survival analysis consists of the likelihood and censorship of survival data, hazard function and partial likelihood, log partial likelihood function, Kaplan-Meier estimator, and two evaluation metrics. This introduction provides some basic concepts of Cox proportional hazards model, followed with an introduction of deep learning-based survival prediction, which is helpful to understand the basic concepts that are used in the later chapters. Next, the definition of co-expression network analysis and non-negative matrix factorization (NMF) are also explained. For NMF, the objective function of NMF with two different losses, and the optimization for NMF with the derivation of multiplicative update rule are explained. Some NMF variants and other low-rank approximation methods are also presented. In addition, histopathologic image registration and segmentation approaches are also elaborated. Finally, related evaluation metrics are also presented.

In Chapter 3, the algorithm SALMON: survival analysis learning with multi-omics neural networks [2] is introduced. This work has been published in *Frontiers in Genetics* in 2019. With the introduction of the background of human cancers and survival prognosis,

this chapter explains study design, neural networks design, architecture, evaluation metric, experimental settings, downstream gene ontology, and functional enrichment analysis. The final results demonstrate the superiority of the proposed algorithms with multiple new findings.

In Chapter 4, a novel algorithm that simultaneously performs non-negative matrix factorization and Cox proportional hazards regression has been developed. With the explained background and problems, the developed algorithm "CoxNMF" outperforms other baseline algorithms in the simulation study, and further demonstrates its power of finding latent gene interactions in human cancer datasets.

In Chapter 5, we construct an automatic, accurate, comprehensive, interpretable, and reproducible WSI feature extraction workflow (IMPRESS). We use these IMPRESS features to develop machine learning model to accurately predict the response to neoadjuvant chemotherapy (NAC) outcome in breast cancer patients. We demonstrate that machine learning using IMPRESS features can be a preferred method to predict post-treatment neoadjuvant chemotherapy outcomes.

In Chapter 6, TSUNAMI: Translational Bioinformatics Tool Suite For Network Analysis And Mining [11] has been elucidated. This work has been published in *Genomics, Proteomics, and Bioinformatics* in 2021. We use a dataset to demonstrated the functionalities of the web portal, which can help biologists discover latent gene interactions.

7.2 Future Work

With the rapid development of advanced technologies in computational biology and bioinformatics domains, integrating traditional statistical methods (Cox proportional hazards model) and computational analyses (deep learning, non-negative matrix factorization) becomes a great interest to interpret multimodal biomedical data. In this thesis, we present two survival analysis work, one image analysis work, and one web tool. As the expected or desired results have been successfully demonstrated in Chapter 3, 4, 5, and 6, further studies in the future becomes necessary to provide more insight in computational biology and bioinformatics.

For example, SALMON can be trained on multiple cancers to unravel the gene modules that associated with survival. These further analyses will help to understand how gene co-expression modules interacted or behaved among various cancers. Similarly, CoxNMF can be used to analyze the datasets which endpoints are relapse-free survival (time until recurrence) [354]. In general, CoxNMF is suitable for all time-to-event dataset analyses.

The IMPRESS workflow can be further studied if additional data is available. First, due to the limitation of the data source, results have not been further validated on an independent external validation dataset and the sizes of the cohorts are relatively small, but we are able to achieve the good accuracy and robustness of machine learning model based on standard practice. Second, the markers from IHC-stained WSIs are limited to CD8, CD163, and PD-L1, which may not fully represent the entire tumor immune micro-environment. More IHC markers could provide more information for prediction. In light of our results demonstrating that IMPRESS features derived from pre-treatment H&E and IHC histopathologic images can predict pCR outcome, we would expect to see advanced machine learning studies with additional immune IHC markers in the future.

Finally, integrative analysis for post-treatment outcome prediction can be further conducted if both genomics data and multimodal histopathologic images are available. These future practices could help to study the associations between genomics information and histopathologic markers, and contribute to the precision health and clinical intervention.

7.3 Publications Result from This Work

7.3.1 Journal Papers

1. **Z. Huang**, Z. Han, T. Wang, W. Shao, S. Xiang, P. Salama, M. Rizkalla, K. Huang, and J. Zhang, "TSUNAMI: Translational Bioinformatics Tool Suite for Network Analysis and Mining," *Genomics, Proteomics and Bioinformatics*, *Accept, in press*, 2021. [Online]. Available: https://doi.org/10.1016/j.gpb.2019.05.006.

2. **Z. Huang**, T. S. Johnson, Z. Han, B. Helm, S. Cao, C. Zhang, P. Salama, M. Rizkalla, C. Y. Yu, J. Cheng, S. Xiang, X. Zhan, J. Zhang, and K. Huang, "Deep Learning-based Cancer Survival Prognosis from RNA-seq Data: Approaches and Evaluations," *BMC Medical Genomics*, vol. 13, no. 5, pp. 1–12, 2020. [Online]. Available: https://doi.org/10.1186/s12920-020-0686-1.

3. **Z. Huang**, X. Zhan, S. Xiang, T. S. Johnson, B. Helm, C. Y. Yu, J. Zhang, P. Salama, M. Rizkalla, Z. Han, and K. Huang, "SALMON: Survival Analysis Learning with Multi-omics Neural Networks on Breast Cancer," *Frontiers in Genetics*, vol. 10, p. 166, 2019. [Online]. Available: https://doi.org/10.3389/fgene.2019.00166.

4. **Z. Huang**, P. Salama, W. Shao, J. Zhang, and K. Huang, "Low-Rank Reorganization via Proportional Hazards Non-negative Matrix Factorization Unveils Survival Associated Gene Clusters," *arXiv preprint arXiv:2008.03776*, 2020. [Online]. Available: https://arxiv.org/abs/2008.03776.

5. **Z. Huang**, W. Shao, Z. Han, A. M. Alkashash, C. De la Sancha, A. V. Parwani, H. Nitta, Y. Hou, T. Wang, P. Salama, M. Rizkalla, J. Zhang, K. Huang, and Z. Li, "Artificial Intelligence Predicts Breast Cancer Neoadjuvant Chemotherapy Response from H&E and IHC histopathologic Images," *To be submitted*, 2021.

7.3.2 Conference Papers and Abstracts

1. **Z. Huang**, Z. Han, A. V. Parwani, K. Huang, and Z. Li, "Artificial Intelligence Driven Neoadjuvant Chemotherapy Response Prediction in Triple Negative Breast Cancer (TNBC) Unveils Non-linear Feature Interactions," *United States and Canadian Academy of Pathology (USCAP) 2020 Annual Meeting Abstracts*, Los Angeles, CA, USA, March 1–5, 2020.

2. **Z. Huang**, Z. Han, A. V. Parwani, K. Huang, and Z. Li, "Predicting Response to Neoadjuvant Chemotherapy in HER2-positive Breast Cancer using Machine Learning Models with Combined Tissue Imaging and Clinical Features," *United States and Canadian Academy of Pathology (USCAP) 2019 Annual Meeting Abstracts*, National Harbor, Maryland, USA, March 16–21, 2019.

3. **Z. Huang**, T. S. Johnson, Z. Han, B. Helm, S. Cao, C. Zhang, P. Salama, M. Rizkalla, C. Y. Yu, J. Cheng, S. Xiang, X. Zhan, J. Zhang, and K. Huang, "Deep Learning-based Cancer Survival Prognosis from RNA-seq Data: Approaches and Evaluations," *International Conference on Intelligent Biology and Medicine (ICIBM 2019)*, Columbus, OH, USA, June 9–11, 2019. [Online]. Available: https://doi.org/10.1186/s12920-020-0686-1.

7.4 Publications Not Include in This Work

7.4.1 Journal Papers

1. T. S. Johnson, S. Xiang, T. Dong, **Z. Huang**, M. Cheng, T. Wang, K. Yang, D. Ni, K. Huang, J. Zhang, "Combinatorial Analyses Reveal Cellular Composition Changes have Different Impacts on Transcriptomic Changes of Cell Type Specific Genes in Alzheimer's Disease," *Scientific Reports*, vol. 11, no. 1, pp. 1–19, 2021. [Online]. Available: https://doi.org/10.1038/s41598-020-79740-x.

2. W. Shao, T. Wang, L. Sun, T. Dong, Z. Han, **Z. Huang**, J. Zhang, D. Zhang, and K. Huang, "Multi-Task Multi-Modal Learning for Joint Diagnosis and Prognosis of Human Cancers," *Medical Image Analysis*, vol. 65, p. 101795, 2020. [Online]. Available: https://doi.org/10.1016/j.media.2020.101795.

3. S. Cong, X. Yao, **Z. Huang**, S. L. Risacher, K. Nho, A. J. Saykin, and L. Shen, "Volumetric GWAS of Medial Temporal Lobe Structures Identifies an ERC1 Locus using ADNI High-Resolution T2-weighted MRI Data," *Neurobiology of Aging*, vol. 95, pp. 81–93, 2020. [Online]. Available: https://doi.org/10.1016/j.neurobiolaging.2020.07.005.

4. J. Yan, V. V. Raja, **Z. Huang**, E. Amico, K. Nho, S. Fang, O. Sporns, Y. Wu, A. J. Saykin, J. Goñi, and L. Shen, "Brain-wide Structural Connectivity Alterations under the Control of Alzheimer Risk Genes," *International Journal of Computational Biology and Drug Design*, vol. 13, no. 1, pp. 58–70, 2020. [Online]. Available: https://doi.org/10.1504/IJCBDD.2020.105098.

5. S. Huang, **Z. Huang**, P. Chen, and C. Feng, "Aberrant Chloride Intracellular Channel 4 Expression is Associated with Adverse Outcome in Cytogenetically Normal Acute Myeloid Leukemia," *Frontiers in Oncology*, vol. 10, p. 1648, 2020. [Online]. Available: https://doi.org/10.3389/fonc.2020.01648.

6. S. Huang, **Z. Huang**, C. Ma, L. Luo, Y. Li, Y. Wu, Y. Ren, and C. Feng, "Acidic Leucine-rich Nuclear Phosphoprotein-32A Expression Contributes to Adverse Outcome in Acute Myeloid Leukemia," *Annals of Translational Medicine*, vol. 8, no. 6, p. 345, 2020. [Online]. Available: https://doi.org/10.21037/atm.2020.02.54.

7. C. Y. Yu, S. Xiang, **Z. Huang**, T. S. Johnson, X. Zhan, Z. Han, M. I. Abu Zaid, and K. Huang, "Gene Co-expression Network and Copy Number Variation Analyses Identify Transcription Factors Involved in Multiple Myeloma Progression," *Frontiers in Genetics*, vol. 10, p. 468, 2019. [Online]. Available: https://doi.org/10.3389/fgene.2019.00468.

8. T. S. Johnson, T. Wang, **Z. Huang**, C. Y. Yu, Y. Wu, Y. Han, Y. Zhang, K. Huang, and J. Zhang, "LAmbDA: Label Ambiguous Domain Adaptation Dataset Integration Reduces Batch Effects and Improves Subtype Detection," *Bioinformatics*, vol. 35, no. 22, pp. 4696–4706, 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz295.

9. X. Zhan, J. Cheng, **Z. Huang**, Z. Han, B. Helm, X. Liu, J. Zhang, T. Wang, D. Ni, and K. Huang, "Correlation Analysis of Histopathology and Proteogenomics Data for Breast Cancer," *Molecular & Cellular Proteomics*, vol. 18, no. 8, S37–S51, 2019. [Online]. Available: https://doi.org/10.1074/mcp.RA118.001232.

10. W. Shao, T. Wang, **Z. Huang**, J. Cheng, Z. Han, D. Zhang, and K. Huang, "Diagnosis-Guided Multi-Modal Feature Selection for Prognosis Prediction of Lung Squamous Cell Carcinoma," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 113–121, 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-32251-9_13.

11. T. S. Johnson, S. Li, E. Franz, **Z. Huang**, S. D. Li, M. J. Campbell, K. Huang, and Y. Zhang, "PseudoFuN: Deriving functional potentials of Pseudogenes from Integrative Relationships with Genes and microRNAs Across 32 Cancers," *GigaScience*, vol. 8, no. 5, p. giz046, 2019. [Online]. Available: https://doi.org/10.1093/gigascience/giz046.

12. S. Xiang, **Z. Huang**, T. Wang, Z. Han, C. Y. Yu, D. Ni, K. Huang, and J. Zhang, "Condition-specific Gene Co-expression Network Mining Identifies Key Pathways and Regulators in the Brain Tissue of Alzheimer's Disease Patients," *BMC Medical Genomics*, vol. 11, no. 6, pp. 39–51, 2018. [Online]. Available: https://doi.org/10.1186/s12920-018-0431-1.

13. C. Feng, H. Huang, S. Huang, Y. Zhai, J. Dong, L. Chen, **Z. Huang**, X. Zhou, B. Li, L. Wang, W. Chen, F. Lv, and T. Li, "Identification of Potential Key Genes Associated with Severe Pneumonia using mRNA-seq," *Experimental and Therapeutic Medicine*, vol. 16, no. 2, pp. 758–766, 2018. [Online]. Available: https://doi.org/10.3892/etm.2018.6262.

14. S. Huang, C. Feng, L. Chen, **Z. Huang**, X. Zhou, B. Li, L. Wang, W. Chen, F. Lv, and T. Li, "Molecular Mechanisms of Mild and Severe Pneumonia: Insights from RNA Sequencing," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 23, p. 1662, 2017. [Online]. Available: https://doi.org/10.12659/MSM.900782.

15. S. Huang, C. Feng, L. Chen, **Z. Huang**, X. Zhou, B. Li, L. Wang, W. Chen, F. Lv, and T. Li, "Identification of Potential Key Long Non-Coding RNAs and Target Genes Associated with Pneumonia using Long Non-Coding RNA Sequencing (lncRNA-Seq): A Preliminary Study," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 22, p. 3394, 2016. [Online]. Available: https://doi.org/10.12659/MSM.900783.

16. S. Huang, H. Yang, Y. Li, C. Feng, L. Gao, G. Chen, H. Gao, **Z. Huang**, Y. Li, and L. Yu, "Prognostic Significance of Mixed-Lineage Leukemia (MLL) Gene Detected by Real-Time Fluorescence Quantitative PCR Assay in Acute Myeloid Leukemia," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 22, p. 3009, 2016. [Online]. Available: https://doi.org/10.12659/MSM.900429.

17. T. Wang, W. Shao, **Z. Huang**, H. Tang, J. Zhang, Z. Ding, and K. Huang, "MOGONET Integrates Multi-omics Data using Graph Convolutional Networks allowing Patient Classification and Biomarker Identification," *Nature Communications*, vol. 12, p. 3445, 2021. [Online]. Available: https://doi.org/10.1038/s41467-021-23774-w.

18. N. Alghamdi, W. Chang, P. Dang, X. Lu, C. Wan, S. Gampala, **Z. Huang**, J. Wang, Q. Ma, Y. Zang, M. Fishel, S. Cao, and C. Zhang, "A Graph Neural Network Model to Estimate Cell-wise Metabolic Flux Using Single Cell RNA-seq Data," *arXiv preprint arXiv:10.1101/2020.09.23.310656*, 2021 [Online]. Available: https://doi.org/10.1101/2020.09.23.310656.

7.4.2 Conference Papers and Abstracts

1. **Z. Huang**, K. Tgavalekos, and C. Zhao, "AI-Driven Forecasting of Mean Pulmonary Artery Pressure for the Management of Cardiac Patients," *Society of Critical Care Medicine's 49th Critical Care Congress (SCCM 2020)*, Orlando, Florida, USA, February 16–19, 2020. [Online]. Available: http://doi.org/10.1097/01.ccm.0000619240.04761.13.

2. W. Shao, T. Wang, **Z. Huang**, J. Cheng, Z. Han, D. Zhang, and K. Huang, "Diagnosis-Guided Multi-Modal Feature Selection for Prognosis Prediction of Lung Squamous Cell Carcinoma," *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2019)*, Shenzhen, China, October 13–17, 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-32251-9_13.

3. A. Hara, **Z. Huang**, Q. Sun, G. Maupomé, and L. Shen, "Machine-learning Identification of Dental Hard-tissue Conditions through Fully Convolutional Neural Networks," *American Public Health Association Annual Meeting and Expo (APHA 2019)*, Philadelphia, USA, November 2–6, 2019. [Online]. Available: https://apha.confex.com/apha/2019/meetingapp.cgi/Paper/436179.

4. J. Yan, V. V. Raja, **Z. Huang**, E. Amico, K. Nho, S. Fang, O. Sporns, Y. Wu, A. J. Saykin, J. Goñi, and L. Shen, "Brain-wide Structural Connectivity Alterations Under the Control of Alzheimer Risk Genes," *International Conference on Intelligent Biology and Medicine (ICIBM 2018)*, Los Angeles, CA, USA, June 10–12, 2018. [Online]. Available: https://doi.org/10.1504/IJCBDD.2020.105098.

5. J. Xue, **Z. Huang**, J. Zhou, Y. Chen, and S. Chien, "A Hierarchical Clustering Analysis (HCA) in Automatic Driving Regarding to Vehicle-to-Vehicle Pedestrian Position Identification," *25th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*, Detroit, MI, USA, June 5–8, 2017. [Online]. Available: https://www-esv.nhtsa.dot.gov/Proceedings/25/25ESV-000176.pdf.

6. B. Tang, S. Chien, **Z. Huang**, and Y. Chen, "Pedestrian Protection using the Integration of V2V and the Pedestrian Automatic Emergency Braking System," *19th International IEEE Conference on Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro, Brazil, November 1–4, 2016. [Online]. Available: https://doi.org/10.1109/ITSC.2016.7795913.

7.4.3 Thesis

1. **Z. Huang**, "Vehicle Sensor-based Pedestrian Position Identification in V2V Environment," *Purdue University Thesis, Master of Science in Electrical & Computer Engineering*, 2016. [Online]. Available: https://docs.lib.purdue.edu/dissertations/AAI10248822.

7.4.4 Patents

1. **Z. Huang**, X. Bei, K. Huang, K. Qian, and A. Zhang, "One Refers to Upper Mouse," China Patent No. CN103257723B, filed April 27, 2013, and issued July 06, 2016. [Online]. Available: https://patents.google.com/patent/CN103257723B/en.

REFERENCES

[1] Z. Huang, T. S. Johnson, Z. Han, B. Helm, S. Cao, C. Zhang, P. Salama, M. Rizkalla, Y. Y. Christina, J. Cheng, S. Xiang, X. Zhan, J. Zhang, and K. Huang, "Deep learning-based cancer survival prognosis from rna-seq data: Approaches and evaluations," *BMC Medical Genomics*, vol. 13, no. 5, pp. 1–12, 2020. [Online]. Available: https://doi.org/10.1186/s12920-020-0686-1.

[2] Z. Huang, X. Zhan, S. Xiang, T. S. Johnson, B. Helm, C. Y. Yu, J. Zhang, P. Salama, M. Rizkalla, Z. Han, and K. Huang, "Salmon: Survival analysis learning with multi-omics neural networks on breast cancer," *Frontiers in Genetics*, vol. 10, p. 166, 2019. [Online]. Available: https://doi.org/10.3389/fgene.2019.00166.

[3] T. S. Johnson, S. Li, E. Franz, Z. Huang, S. Dan Li, M. J. Campbell, K. Huang, and Y. Zhang, "Pseudofun: Deriving functional potentials of pseudogenes from integrative relationships with genes and micrornas across 32 cancers," *GigaScience*, vol. 8, no. 5, giz046, 2019. [Online]. Available: https://doi.org/10.1093/gigascience/giz046.

[4] C. Y. Yu, S. Xiang, Z. Huang, T. S. Johnson, X. Zhan, Z. Han, M. I. Abu Zaid, and K. Huang, "Gene co-expression network and copy number variation analyses identify transcription factors involved in multiple myeloma progression," *Frontiers in Genetics*, vol. 10, p. 468, 2019. [Online]. Available: https://doi.org/10.3389/fgene.2019.00468.

[5] C. Feng, H. Huang, S. Huang, Y.-Z. Zhai, J. Dong, L. Chen, Z. Huang, X. Zhou, B. Li, L.-L. Wang, *et al.*, "Identification of potential key genes associated with severe pneumonia using mrna-seq," *Experimental and Therapeutic Medicine*, vol. 16, no. 2, pp. 758–766, 2018. [Online]. Available: https://doi.org/10.3892/etm.2018.6262.

[6] S. Huang, C. Feng, L. Chen, Z. Huang, X. Zhou, B. Li, L.-L. Wang, W. Chen, F.-Q. Lv, and T.-S. Li, "Molecular mechanisms of mild and severe pneumonia: Insights from rna sequencing," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 23, p. 1662, 2017. [Online]. Available: https://doi.org/10.12659/MSM.900782.

[7] S. Xiang, Z. Huang, T. Wang, Z. Han, Y. Y. Christina, D. Ni, K. Huang, and J. Zhang, "Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of alzheimer's disease patients," *BMC Medical Genomics*, vol. 11, no. 6, pp. 39–51, 2018. [Online]. Available: https://doi.org/10.1186/s12920-018-0431-1.

[8] X. Zhan, J. Cheng, Z. Huang, Z. Han, B. Helm, X. Liu, J. Zhang, T.-F. Wang, D. Ni, and K. Huang, "Correlation analysis of histopathology and proteogenomics data for breast cancer," *Molecular & Cellular Proteomics*, vol. 18, no. 8 suppl 1, S37–S51, 2019. [Online]. Available: https://doi.org/10.1074/mcp.RA118.001232.

[9] B. R. Helm, X. Zhan, P. H. Pandya, M. E. Murray, K. E. Pollok, J. L. Renbarger, M. J. Ferguson, Z. Han, D. Ni, J. Zhang, and K. Huang, "Gene co-expression networks restructured gene fusion in rhabdomyosarcoma cancers," *Genes*, vol. 10, no. 9, p. 665, 2019. [Online]. Available: https://doi.org/10.3390/genes10090665.

[10] Illumina, *A clear, more complete view of the coding transcriptome*. [Online]. Available: https://www.illumina.com/techniques/sequencing/rna-sequencing/mrna-seq.html.

[11] Z. Huang, Z. Han, T. Wang, W. Shao, S. Xiang, P. Salama, M. Rizkalla, K. Huang, and J. Zhang, "Tsunami: Translational bioinformatics tool suite for network analysis and mining," *Genomics, Proteomics, and Bioinformatics*, 2021. [Online]. Available: https://doi.org/10.1016/j.gpb.2019.05.006.

[12] T. Zahid, *An introduction to the concepts of survival analysis and its implementation in lifelines package for python*, 2019. [Online]. Available: https://towardsdatascience.com/survival-analysis-part-a-70213df21c2e.

[13] C. Davidson-Pilon, "Lifelines: Survival analysis in python," *Journal of Open Source Software*, vol. 4, no. 40, p. 1317, 2019. [Online]. Available: https://doi.org/10.21105/joss.01317.

[14] Lifelines, *Lifelines: Introduction to survival analysis*, 2014. [Online]. Available: https://lifelines.readthedocs.io/en/latest/Survival%20Analysis%20intro.html.

[15] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

[16] E. T. Lee and J. Wang, *Statistical Methods for Survival Data Analysis*. John Wiley & Sons, 2003, vol. 476. [Online]. Available: https://doi.org/10.1002/0471458546.

[17] M. Winkel, *Statistical lifetime-models*, 2017. [Online]. Available: https://www.stats.ox.ac.uk/~winkel/bs3b10_14.pdf.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539.

[19] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in Bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017. [Online]. Available: https://doi.org/10.1093/bib/bbw068.

[20] M. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey, "Deep learning of the tissue-regulated splicing code," *Bioinformatics*, vol. 30, no. 12, pp. i121–i129, 2014. [Online]. Available: https://doi.org/10.1093/bioinformatics/btu277.

[21] Y. Chen, Y. Li, R. Narayan, A. Subramanian, and X. Xie, "Gene expression inference with deep learning," *Bioinformatics*, vol. 32, no. 12, pp. 1832–1839, 2016. [Online]. Available: https://doi.org/10.1093/bioinformatics/btw074.

[22] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of dna-and rna-binding proteins by deep learning," *Nature Biotechnology*, vol. 33, no. 8, pp. 831–838, 2015. [Online]. Available: https://doi.org/10.1038/nbt.3300.

[23] W. Shao, T. Wang, Z. Huang, J. Cheng, Z. Han, D. Zhang, and K. Huang, "Diagnosis-guided multi-modal feature selection for prognosis prediction of lung squamous cell carcinoma," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 113–121. [Online]. Available: https://doi.org/10.1007/978-3-030-32251-9_13.

[24] T. Ching, X. Zhu, and L. X. Garmire, "Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data," *PLoS Computational Biology*, vol. 14, no. 4, e1006076, 2018. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1006076.

[25] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 24, 2018. [Online]. Available: https://doi.org/10.1186/s12874-018-0482-1.

[26] D. Faraggi and R. Simon, "A neural network model for survival data," *Statistics in Medicine*, vol. 14, no. 1, pp. 73–82, 1995. [Online]. Available: https://doi.org/10.1002/sim.4780140108.

[27] P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. V. Vega, D. J. Brat, and L. A. Cooper, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, E2970–E2979, 2018. [Online]. Available: https://doi.org/10.1073/pnas.1717139115.

[28] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer, *et al.*, "Random survival forests," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008. [Online]. Available: https://doi.org/10.1214/08-AOAS169.

[29] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, p. 1, 2010. [Online]. Available: https://doi.org/10.18637/jss.v033.i01.

[30] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep learning-based multi-omics integration robustly predicts survival in liver cancer," *Clinical Cancer Research*, vol. 24, no. 6, pp. 1248–1259, 2018. [Online]. Available: https://doi.org/10.1158/1078-0432.CCR-17-0853.

[31] L. Zhang, C. Lv, Y. Jin, G. Cheng, Y. Fu, D. Yuan, Y. Tao, Y. Guo, X. Ni, and T. Shi, "Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma," *Frontiers in Genetics*, vol. 9, p. 477, 2018. [Online]. Available: https://doi.org/10.3389/fgene.2018.00477.

[32] O. B. Poirion, K. Chaudhary, and L. X. Garmire, "Deep learning data integration for better risk stratification models of bladder cancer," *AMIA Summits on Translational Science Proceedings*, vol. 2018, p. 197, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961799.

[33] T.-Y. Lee, K.-Y. Huang, C.-H. Chuang, C.-Y. Lee, and T.-H. Chang, "Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication," *Computational Biology and Chemistry*, vol. 87, p. 107 277, 2020. [Online]. Available: https://doi.org/10.1016/j.compbiolchem.2020.107277.

[34] J. Zhang and K. Huang, "Normalized lmqcm: An algorithm for detecting weak quasi-cliques in weighted graph with applications in gene co-expression module discovery in cancers," *Cancer Informatics*, vol. 13, CIN–S14021, 2014. [Online]. Available: https://doi.org/10.4137/CIN.S14021.

[35] P. Langfelder and S. Horvath, "Wgcna: An r package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, pp. 1–13, 2008. [Online]. Available: https://doi.org/10.1186/1471-2105-9-559.

[36] R. De Smet and K. Marchal, "Advantages and limitations of current network inference methods," *Nature Reviews Microbiology*, vol. 8, no. 10, pp. 717–729, 2010. [Online]. Available: https://doi.org/10.1038/nrmicro2419.

[37] F.-J. Müller, L. C. Laurent, D. Kostka, I. Ulitsky, R. Williams, C. Lu, I.-H. Park, M. S. Rao, R. Shamir, P. H. Schwartz, N. O. Schmidt, and J. F. Loring, "Regulatory networks define phenotypic classes of human stem cell lines," *Nature*, vol. 455, no. 7211, pp. 401–405, 2008. [Online]. Available: https://doi.org/10.1038/nature07213.

[38] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: Back to metabolism in kegg," *Nucleic Acids Research*, vol. 42, no. D1, pp. D199–D205, 2014. [Online]. Available: https://doi.org/10.1093/nar/gkt1076.

[39] K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, *et al.*, "An empirical framework for binary interactome mapping," *Nature Methods*, vol. 6, no. 1, p. 83, 2009. [Online]. Available: https://doi.org/10.1038/nmeth.1280.

[40] J. Zhang, K. Lu, Y. Xiang, M. Islam, S. Kotian, Z. Kais, C. Lee, M. Arora, H.-w. Liu, J. D. Parvin, and K. Huang, "Weighted frequent gene co-expression network mining to identify genes involved in genome stability," *PLoS Computational Biology*, vol. 8, no. 8, e1002656, 2012. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1002656.

[41] Z. Kais, S. H. Barsky, H. Mathsyaraja, A. Zha, D. J. Ransburgh, G. He, R. T. Pilarski, C. L. Shapiro, K. Huang, and J. D. Parvin, "Kiaa0101 interacts with brca1 and regulates centrosome number," *Molecular Cancer Research*, vol. 9, no. 8, pp. 1091–1099, 2011. [Online]. Available: https://doi.org/10.1158/1541-7786.MCR-10-0503.

[42] M. A. Pujana, J.-D. J. Han, L. M. Starita, K. N. Stevens, M. Tewari, J. S. Ahn, G. Rennert, V. Moreno, T. Kirchhoff, B. Gold, *et al.*, "Network modeling links breast cancer susceptibility and centrosome dysfunction," *Nature Genetics*, vol. 39, no. 11, pp. 1338–1349, 2007. [Online]. Available: https://doi.org/10.1038/ng.2007.2.

[43] J. Zhao, X. Hu, T. He, P. Li, M. Zhang, and X. Shen, "An edge-based protein complex identification algorithm with gene co-expression data (pcia-geco)," *IEEE Transactions on Nanobioscience*, vol. 13, no. 2, pp. 80–88, 2014. [Online]. Available: https://doi.org/10.1109/TNB.2014.2317519.

[44] J. Zhang, S. Ni, Y. Xiang, J. D. Parvin, Y. Yang, Y. Zhou, and K. Huang, "Gene co-expression analysis predicts genetic aberration loci associated with colon cancer metastasis," *International Journal of Computational Biology and Drug Design*, vol. 6, no. 1-2, pp. 60–71, 2013. [Online]. Available: https://doi.org/10.1504/IJCBDD.2013.052202.

[45] Y. Ou and C. Zhang, "A new multimembership clustering method," *Journal of Industrial and Management Optimization*, vol. 3, no. 4, p. 619, 2007. [Online]. Available: https://doi.org/10.3934/jimo.2007.3.619.

[46] Z. Han, T. Johnson, J. Zhang, X. Zhang, and K. Huang, "Functional virtual flow cytometry: A visual analytic approach for characterizing single-cell gene expression patterns," *BioMed Research International*, vol. 2017, 2017. [Online]. Available: https://doi.org/10.1155/2017/3035481.

[47] P. Langfelder and S. Horvath, "Eigengene networks for studying the relationships between co-expression modules," *BMC Systems Biology*, vol. 1, no. 1, pp. 1–17, 2007. [Online]. Available: https://doi.org/10.1186/1752-0509-1-54.

[48] V. Klema and A. Laub, "The singular value decomposition: Its computation and some applications," *IEEE Transactions on Automatic Control*, vol. 25, no. 2, pp. 164–176, 1980. [Online]. Available: https://doi.org/10.1109/TAC.1980.1102314.

[49] G. Strang, *Linear Algebra and Its Applications*. Belmont, CA: Thomson, Brooks/Cole, 2006, ISBN: 0030105676 9780030105678 0534422004 9780534422004. [Online]. Available: https://www.sciencedirect.com/book/9780126736601/linear-algebra-and-its-applications.

[50] Y. Han, X. Ye, J. Cheng, S. Zhang, W. Feng, Z. Han, J. Zhang, and K. Huang, "Integrative analysis based on survival associated co-expression gene modules for predicting neuroblastoma patients' survival time," *Biology Direct*, vol. 14, no. 1, pp. 1–12, 2019. [Online]. Available: https://doi.org/10.1186/s13062-018-0229-2.

[51] J. Zhang and K. Huang, "Pan-cancer analysis of frequent dna co-methylation patterns reveals consistent epigenetic landscape changes in multiple cancers," *BMC Genomics*, vol. 18, no. 1, pp. 1–14, 2017. [Online]. Available: https://doi.org/10.1186/s12864-016-3259-0.

[52] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, 1998. [Online]. Available: https://doi.org/10.1137/1.9780898719628.

[53] R. Caruana, S. Lawrence, and C. L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," in *Advances in Neural Information Processing Systems*, 2001, pp. 402–408. [Online]. Available: https://dl.acm.org/doi/10.5555/3008751.3008807.

[54] D. M. Hawkins, "The problem of overfitting," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 1, pp. 1–12, 2004. [Online]. Available: https://doi.org/10.1021/ci0342472.

[55] E. B. Baum and D. Haussler, "What size net gives valid generalization?" *Neural Computation*, vol. 1, no. 1, pp. 151–160, 1989. [Online]. Available: https://doi.org/10.1162/neco.1989.1.1.151.

[56] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Advances in Neural Information Processing Systems*, 1992, pp. 950–957. [Online]. Available: https://dl.acm.org/doi/10.5555/2986916.2987033.

[57] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992. [Online]. Available: https://doi.org/10.1162/neco.1992.4.1.1.

[58] J. E. Moody, S. Hanson, and R. Lippmann, "The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems," *Advances in Neural Information Processing Systems*, vol. 4, pp. 847–854, 1992. [Online]. Available: https://dl.acm.org/doi/10.5555/2986916.2987020.

[59] P. Burman, "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods," *Biometrika*, vol. 76, no. 3, pp. 503–514, 1989. [Online]. Available: https://doi.org/10.1093/biomet/76.3.503.

[60] T. T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognition*, vol. 48, no. 9, pp. 2839–2846, 2015. [Online]. Available: https://doi.org/10.1016/j.patcog.2015.03.009.

[61] L. Prechelt, "Automatic early stopping using cross validation: Quantifying the criteria," *Neural Networks*, vol. 11, no. 4, pp. 761–767, 1998. [Online]. Available: https://doi.org/10.1016/S0893-6080(98)00010-0.

[62] F. Santosa and W. W. Symes, "Linear inversion of band-limited reflection seismograms," *SIAM Journal on Scientific and Statistical Computing*, vol. 7, no. 4, pp. 1307–1330, 1986. [Online]. Available: https://doi.org/10.1137/0907087.

[63] C. Cortes, M. Mohri, and A. Rostamizadeh, "L2 regularization for learning kernels," *arXiv preprint arXiv:1205.2653*, 2012. [Online]. Available: https://arxiv.org/abs/1205.2653.

[64] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005. [Online]. Available: https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[65] S. Wager, S. Wang, and P. Liang, "Dropout training as adaptive regularization," *arXiv preprint arXiv:1307.1493*, 2013. [Online]. Available: https://arxiv.org/abs/1307.1493.

[66] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001. [Online]. Available: https://doi.org/10.1198/10618600152418584.

[67] H. Liu, E. R. Dougherty, J. G. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, *et al.*, "Evolving feature selection," *IEEE Intelligent Systems*, vol. 20, no. 6, pp. 64–76, 2005. [Online]. Available: https://doi.org/10.1109/MIS.2005.105.

[68] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear Algebra*, Springer, 1971, pp. 134–151. [Online]. Available: https://doi.org/10.1007/BF02163027.

[69] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987. [Online]. Available: https://doi.org/10.1016/0169-7439(87)80084-9.

[70] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proceedings of the 2005 SIAM international conference on data mining*, SIAM, 2005, pp. 606–610. [Online]. Available: https://doi.org/10.1137/1.9781611972757.70.

[71] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999. [Online]. Available: https://doi.org/10.1038/44565.

[72] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, "Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification," *IEEE Transactions on Neural Networks*, vol. 17, no. 3, pp. 683–695, 2006. [Online]. Available: https://doi.org/10.1109/TNN.2006.873291.

[73] Y. Wang, Y. Jia, C. Hu, and M. Turk, "Non-negative matrix factorization framework for face recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 19, no. 04, pp. 495–511, 2005. [Online]. Available: https://doi.org/10.1142/S0218001405004198.

[74] I. Kotsia, S. Zafeiriou, and I. Pitas, "A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 588–595, 2007. [Online]. Available: https://doi.org/10.1109/TIFS.2007.902017.

[75] S. Nikitidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Subclass discriminant nonnegative matrix factorization for facial image analysis," *Pattern Recognition*, vol. 45, no. 12, pp. 4080–4091, 2012. [Online]. Available: https://doi.org/10.1016/j.patcog.2012.04.030.

[76] A. Pascual-Montano, P. Carmona-Saez, M. Chagoyen, F. Tirado, J. M. Carazo, and R. D. Pascual-Marqui, "Bionmf: A versatile tool for non-negative matrix factorization in biology," *BMC Bioinformatics*, vol. 7, no. 1, pp. 1–9, 2006. [Online]. Available: https://doi.org/10.1186/1471-2105-7-366.

[77] R. Zhu, J. Liu, Y. Zhang, and Y. Guo, "A robust manifold graph regularized nonnegative matrix factorization algorithm for cancer gene clustering," *Molecules*, vol. 22, no. 12, p. 2131, 2017. [Online]. Available: https://doi.org/10.3390/molecules22122131.

[78] X. Zhu, T. Ching, X. Pan, S. M. Weissman, and L. Garmire, "Detecting heterogeneity in single-cell rna-seq data by non-negative matrix factorization," *PeerJ*, vol. 5, e2888, 2017. [Online]. Available: https://doi.org/10.7717/peerj.2888.

[79] X. Jiang, H. Zhang, Z. Zhang, and X. Quan, "Flexible non-negative matrix factorization to unravel disease-related genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 6, pp. 1948–1957, 2018. [Online]. Available: https://doi.org/10.1109/TCBB.2018.2823746.

[80] Y. Lai, M. Hayashida, and T. Akutsu, "Survival analysis by penalized regression and matrix factorization," *The Scientific World Journal*, vol. 2013, 2013. [Online]. Available: https://doi.org/10.1155/2013/632030.

[81] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems*, 2001, pp. 556–562. [Online]. Available: https://papers.nips.cc/paper/2000/hash/f9d1152547c0bde01830b7e8bd60024c-Abstract.html.

[82] C.-J. Hsieh and I. S. Dhillon, "Fast coordinate descent methods with variable selection for non-negative matrix factorization," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 1064–1072. [Online]. Available: https://doi.org/10.1145/2020408.2020577.

[83] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2006, pp. 126–135. [Online]. Available: https://doi.org/10.1145/1150402.1150420.

[84] S. Choi, "Algorithms for orthogonal nonnegative matrix factorization," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, IEEE, 2008, pp. 1828–1832. [Online]. Available: https://doi.org/10.1109/IJCNN.2008.4634046.

[85] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (cat. no. 98th8468)*, IEEE, 1999, pp. 41–48. [Online]. Available: https://doi.org/10.1109/NNSP.1999.788121.

[86] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 711–720, 1997. [Online]. Available: https://doi.org/10.1109/34.598228.

[87] S.-Y. Lee, H.-A. Song, and S.-i. Amari, "A new discriminant nmf algorithm and its application to the extraction of subtle emotional differences in speech," *Cognitive Neurodynamics*, vol. 6, no. 6, pp. 525–535, 2012. [Online]. Available: https://doi.org/10.1007/s11571-012-9213-1.

[88] N. Guan, X. Zhang, Z. Luo, D. Tao, and X. Yang, "Discriminant projective non-negative matrix factorization," *PloS One*, vol. 8, no. 12, e83291, 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0083291.

[89] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, and C. Yuan, "Nonnegative discriminant matrix factorization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 7, pp. 1392–1405, 2016. [Online]. Available: https://doi.org/10.1109/TCSVT.2016.2539779.

[90] Z. Jia, X. Zhang, N. Guan, X. Bo, M. R. Barnes, and Z. Luo, "Gene ranking of rna-seq data via discriminant non-negative matrix factorization," *PloS One*, vol. 10, no. 9, e0137782, 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0137782.

[91] G. Chao, C. Mao, F. Wang, Y. Zhao, and Y. Luo, "Supervised nonnegative matrix factorization to predict icu mortality risk," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2018, pp. 1189–1194. [Online]. Available: https://doi.org/10.1109/BIBM.2018.8621403.

[92] C. Boutsidis and E. Gallopoulos, "Svd based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, 2008. [Online]. Available: https://doi.org/10.1016/j.patcog.2007.09.010.

[93] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901. [Online]. Available: https://doi.org/10.1080/14786440109462720.

[94] I. T. Jolliffe, "Principal components in regression analysis," in *Principal Component Analysis*, Springer, 1986, pp. 129–155. [Online]. Available: https://doi.org/10.1007/978-1-4757-1904-8_8.

[95] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 689–696. [Online]. Available: https://doi.org/10.1145/1553374.1553463.

[96] C. M. Bishop, *Pattern Recognition and Machine Learning*. springer, 2006. [Online]. Available: https://dl.acm.org/doi/10.5555/1162264.

[97] E. Mejía-Roa, P. Carmona-Saez, R. Nogales, C. Vicente, M. Vázquez, X. Yang, C. García, F. Tirado, and A. Pascual-Montano, "Bionmf: A web-based tool for nonnegative matrix factorization in biology," *Nucleic Acids Research*, vol. 36, no. suppl_2, W523–W528, 2008. [Online]. Available: https://doi.org/10.1093/nar/gkn335.

[98] B. Zitova and J. Flusser, "Image registration methods: A survey," *Image and Vision Computing*, vol. 21, no. 11, pp. 977–1000, 2003. [Online]. Available: https://doi.org/10.1016/S0262-8856(03)00137-9.

[99] W. K. Pratt, *Introduction to Digital Image Processing*. CRC press, 2013. [Online]. Available: https://doi.org/10.1002/0470097434.

[100] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997. [Online]. Available: https://doi.org/10.1023/A:1007958904918.

[101] H. B. Curry, "The method of steepest descent for non-linear minimization problems," *Quarterly of Applied Mathematics*, vol. 2, no. 3, pp. 258–261, 1944. [Online]. Available: https://www.jstor.org/stable/43633461.

[102] E. Stiefel, "Methods of conjugate gradients for solving linear systems," *Journal of Research of the National Bureau of Standards*, vol. 49, pp. 409–435, 1952. [Online]. Available: https://doi.org/10.6028/JRES.049.044.

[103] C. T. Kelley, *Iterative Methods for Optimization*. SIAM, 1999. [Online]. Available: https://doi.org/10.1137/1.9781611970920.

[104] P. E. Gill and W. Murray, "Quasi-newton methods for unconstrained optimization," *IMA Journal of Applied Mathematics*, vol. 9, no. 1, pp. 91–108, 1972. [Online]. Available: https://doi.org/10.1093/imamat/9.1.91.

[105] K. Levenberg, "A method for the solution of certain nonlinear problems in least squares," *Quarterly of Applied Mathematics*, vol. 2, pp. 164–168, 1944. [Online]. Available: https://doi.org/10.1090/qam/10666.

[106] D. Mardquardt, "An algorithm for least square estimation of parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, pp. 431–441, 1963. [Online]. Available: https://doi.org/10.1137/0111030.

[107] A. A. Goshtasby, *2-D and 3-D Image Registration: For Medical, Remote Sensing, and Industrial Applications*. John Wiley & Sons, 2005. [Online]. Available: https://www.wiley.com/en-us/2+D+and+3+D+Image+Registration:+for+Medical,+Remote+Sensing,+and+Industrial+Applications-p-9780471649540.

[108] J.-P. Thirion, "Fast non-rigid matching of 3d medical images," Ph.D. dissertation, INRIA, 1995. [Online]. Available: https://hal.inria.fr/inria-00077268.

[109] J.-P. Thirion, "Image matching as a diffusion process: An analogy with maxwell's demons," *Medical Image Analysis*, vol. 2, no. 3, pp. 243–260, 1998. [Online]. Available: https://doi.org/10.1016/S1361-8415(98)80022-4.

[110] X. Gu, H. Pan, Y. Liang, R. Castillo, D. Yang, D. Choi, E. Castillo, A. Majumdar, T. Guerrero, and S. B. Jiang, "Implementation and evaluation of various demons deformable image registration algorithms on a gpu," *Physics in Medicine & Biology*, vol. 55, no. 1, p. 207, 2009. [Online]. Available: https://doi.org/10.1088/0031-9155/55/1/012.

[111] H. Wang, L. Dong, J. O'Daniel, R. Mohan, A. S. Garden, K. K. Ang, D. A. Kuban, M. Bonnen, J. Y. Chang, and R. Cheung, "Validation of an accelerated 'demons' algorithm for deformable image registration in radiation therapy," *Physics in Medicine & Biology*, vol. 50, no. 12, p. 2887, 2005. [Online]. Available: https://doi.org/10.1088/0031-9155/50/12/011.

[112] P. Rogelj and S. Kovačič, "Symmetric image registration," *Medical Image Analysis*, vol. 10, no. 3, pp. 484–493, 2006. [Online]. Available: https://doi.org/10.1016/j.media.2005.03.003.

[113] X. Pennec, P. Cachier, and N. Ayache, "Understanding the "demon's algorithm": 3d non-rigid registration by gradient descent," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 1999, pp. 597–605. [Online]. Available: https://doi.org/10.1007/10704282_64.

[114] M. Wodzinski and A. Skalski, "Multistep, automatic and nonrigid image registration method for histology samples acquired using multiple stains," *Physics in Medicine & Biology*, vol. 66, no. 2, p. 025 006, 2021. [Online]. Available: https://doi.org/10.1088/1361-6560/abcad7.

[115] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, M. Brady, and J. A. Schnabel, "Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Medical Image Analysis*, vol. 16, no. 7, pp. 1423–1435, 2012. [Online]. Available: https://doi.org/10.1016/j.media.2012.05.008.

[116] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, IEEE, vol. 2, 1999, pp. 1150–1157. [Online]. Available: https://doi.org/10.1109/ICCV.1999.790410.

[117] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008. [Online]. Available: https://doi.org/10.1109/ICCV.1999.790410.

[118] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, IEEE, 2011, pp. 2564–2571. [Online]. Available: https://doi.org/10.1109/ICCV.2011.6126544.

[119] J. Borovec, J. Kybic, I. Arganda-Carreras, D. V. Sorokin, G. Bueno, A. V. Khvostikov, S. Bakas, I. Eric, C. Chang, S. Heldmann, *et al.*, "Anhir: Automatic non-rigid histological image registration challenge," *IEEE Transactions on Medical Imaging*, vol. 39, no. 10, 2020. [Online]. Available: https://doi.org/10.1109/TMI.2020.2986331.

[120] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: A review," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137–178, 2021. [Online]. Available: https://doi.org/10.1007/s10462-020-09854-1.

[121] S. Nayak, R. Gopakumar, and V. Acharya, "Nucleus segmentation from microscopic bone marrow image," in *Progress in Advanced Computing and Intelligent Engineering*, Springer, 2020, pp. 44–50. [Online]. Available: https://dx.doi.org/10.1007/978-981-15-6584-7_5.

[122] K. Leon, D. Mery, F. Pedreschi, and J. Leon, "Color measurement in l*a*b* units from rgb digital images," *Food Research International*, vol. 39, no. 10, pp. 1084–1091, 2006. [Online]. Available: https://doi.org/10.1016/j.foodres.2006.03.006.

[123] A. Ford and A. Roberts, "Colour space conversions," *Westminster University, London*, vol. 1998, pp. 1–31, 1998. [Online]. Available: https://poynton.ca/PDFs/coloureq.pdf.

[124] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440. [Online]. Available: https://arxiv.org/abs/1411.4038.

[125] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2018, pp. 1451–1460. [Online]. Available: https://doi.org/10.1109/WACV.2018.00163.

[126] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28.

[127] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017. [Online]. Available: https://arxiv.org/abs/1706.05587.

[128] S. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 1999. [Online]. Available: https://doi.org/10.1016/B978-0-12-374370-1.X0001-8.

[129] G. Papandreou, I. Kokkinos, and P.-A. Savalle, "Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 390–399. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298636.

[130] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90.

[131] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

[132] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://doi.org/10.1016/0377-0427(87)90125-7.

[133] C. Zaiontz, *Real statistics using excel: Log-rank test*, 2016. [Online]. Available: https://www.real-statistics.com/survival-analysis/kaplan-meier-procedure/log-rank-test.

[134] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemotherapy Reports*, vol. 50, pp. 163–170, 1966. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/5910392.

[135] R. Peto and J. Peto, "Asymptotically efficient rank invariant test procedures," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 2, pp. 185–198, 1972. [Online]. Available: https://doi.org/10.2307/2344317.

[136] D. Harrington, "Linear rank tests in survival analysis," *Encyclopedia of Biostatistics*, vol. 4, 2005. [Online]. Available: https://doi.org/10.1002/0470011815.b2a11047.

[137] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908. [Online]. Available: https://doi.org/10.2307/2331554.

[138] J. H. Zar, "Spearman rank correlation," *Encyclopedia of Biostatistics*, vol. 7, 2005. [Online]. Available: https://doi.org/10.1002/0470011815.b2a15150.

[139] J. A. Rice, *Mathematical Statistics and Data Analysis*. Nelson Education, 2006. [Online]. Available: https://books.google.com/books/about/Mathematical_Statistics_and_Data_Analysi.html?id=KfkYAQAAIAAJ.

[140] J. Cao and S. Zhang, "A bayesian extension of the hypergeometric test for functional enrichment analysis," *Biometrics*, vol. 70, no. 1, pp. 84–94, 2014. [Online]. Available: https://doi.org/10.1111/biom.12122.

[141] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[142] Y. Benjamini, R. Heller, and D. Yekutieli, "Selective inference in complex research," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1906, pp. 4255–4271, 2009. [Online]. Available: https://doi.org/10.1098/rsta.2009.0127.

[143] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, pp. 50–60, 1947. [Online]. Available: https://doi.org/10.1214/aoms/1177730491.

[144] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The transcriptional landscape of the yeast genome defined by rna sequencing," *Science*, vol. 320, no. 5881, pp. 1344–1349, 2008. [Online]. Available: https://doi.org/10.1126/science.1158441.

[145] B. T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers, and J. Bähler, "Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution," *Nature*, vol. 453, no. 7199, pp. 1239–1243, 2008. [Online]. Available: https://doi.org/10.1038/nature07002.

[146] Y. Chu and D. R. Corey, "Rna sequencing: Platform selection, experimental design, and data interpretation," *Nucleic Acid Therapeutics*, vol. 22, no. 4, pp. 271–274, 2012. [Online]. Available: https://doi.org/10.1089/nat.2012.0367.

[147] W. Guo, Q. Wang, Y. Zhan, X. Chen, Q. Yu, J. Zhang, Y. Wang, X.-j. Xu, and L. Zhu, "Transcriptome sequencing uncovers a three–long noncoding rna signature in predicting breast cancer survival," *Scientific Reports*, vol. 6, no. 1, pp. 1–10, 2016. [Online]. Available: https://doi.org/10.1038/srep27931.

[148] Y. Liang, J. Song, D. He, Y. Xia, Y. Wu, X. Yin, and J. Liu, "Systematic analysis of survival-associated alternative splicing signatures uncovers prognostic predictors for head and neck cancer," *Journal of Cellular Physiology*, vol. 234, no. 9, pp. 15 836–15 846, 2019. [Online]. Available: https://doi.org/10.1002/jcp.28241.

[149] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: Recent progress in multi-omics data integration methods," *Frontiers in Genetics*, vol. 8, p. 84, 2017. [Online]. Available: https://doi.org/10.3389/fgene.2017.00084.

[150] Z. Zhang, Y. Zhao, X. Liao, W. Shi, K. Li, Q. Zou, and S. Peng, "Deep learning in omics: A survey and guideline," *Briefings in Functional Genomics*, vol. 18, no. 1, pp. 41–57, 2019. [Online]. Available: https://doi.org/10.1093/bfgp/ely030.

[151] W. Shao, J. Cheng, L. Sun, Z. Han, Q. Feng, D. Zhang, and K. Huang, "Ordinal multi-modal feature selection for survival analysis of early-stage renal cancer," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2018, pp. 648–656. [Online]. Available: https://doi.org/10.1007/978-3-030-00934-2_72.

[152] D. Ramazzotti, A. Lal, B. Wang, S. Batzoglou, and A. Sidow, "Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival," *Nature Communications*, vol. 9, no. 1, pp. 1–14, 2018. [Online]. Available: https://doi.org/10.1038/s41467-018-06921-8.

[153] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 3, pp. 841–850, 2018. [Online]. Available: https://doi.org/10.1109/TCBB.2018.2806438.

[154] G. Liu, C. Dong, and L. Liu, "Integrated multiple "-omics" data reveal subtypes of hepatocellular carcinoma," *PloS One*, vol. 11, no. 11, 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0165457.

[155] S. Nagini, "Breast cancer: Current molecular therapeutic targets and new players," *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)*, vol. 17, no. 2, pp. 152–163, 2017. [Online]. Available: https://www.ingentaconnect.com/content/ben/acamc/2017/00000017/00000002/art00003.

[156] X. Wu, Y. Ye, C. H. Barcenas, W.-H. Chow, Q. H. Meng, M. Chavez-MacGregor, M. A. Hildebrandt, H. Zhao, X. Gu, Y. Deng, E. Wagar, F. J. Esteva, D. Tripathy, and G. N. Hortobagyi, "Personalized prognostic prediction models for breast cancer recurrence and survival incorporating multidimensional data," *JNCI: Journal of the National Cancer Institute*, vol. 109, no. 7, djw314, 2017. [Online]. Available: https://doi.org/10.1093/jnci/djw314.

[157] B. Pereira, S.-F. Chin, O. M. Rueda, H.-K. M. Vollan, E. Provenzano, H. A. Bardwell, M. Pugh, L. Jones, R. Russell, S.-J. Sammut, *et al.*, "The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes," *Nature Communications*, vol. 7, no. 1, pp. 1–16, 2016. [Online]. Available: https://doi.org/10.1038/ncomms11479.

[158] A. Gupta, M. Mutebi, and A. Bardia, "Gene-expression-based predictors for breast cancer," *Annals of Surgical Oncology*, vol. 22, no. 11, pp. 3418–3432, 2015. [Online]. Available: https://doi.org/10.1245/s10434-015-4703-0.

[159] F. J. Nassar, R. Nasr, and R. Talhouk, "Micrornas as biomarkers for early breast cancer diagnosis, prognosis and therapy prediction," *Pharmacology & Therapeutics*, vol. 172, pp. 34–49, 2017. [Online]. Available: https://doi.org/10.1016/j.pharmthera.2016.11.012.

[160] C. Marshall, D. Howrigan, D. Merico, B. Thiruvahindrapuram, W. Wu, D. Greer, D. Antaki, A. Shetty, P. Holmans, D. Pinto, *et al.*, "Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects," *Nature Genetics*, vol. 49, pp. 27–35, [Online]. Available: https://doi.org/10.1038/ng.3725.

[161] P. Circosta, A. R. Elia, I. Landra, R. Machiorlatti, M. Todaro, S. Aliberti, D. Brusa, S. Deaglio, S. Chiaretti, R. Bruna, *et al.*, "Tailoring cd19xcd3-dart exposure enhances t-cells to eradication of b-cell neoplasms," *Oncoimmunology*, vol. 7, no. 4, e1341032, 2018. [Online]. Available: https://doi.org/10.1080/2162402X.2017.1341032.

[162] N. J. Birkbak, B. Kochupurakkal, J. M. Izarzugaza, A. C. Eklund, Y. Li, J. Liu, Z. Szallasi, U. A. Matulonis, A. L. Richardson, J. D. Iglehart, *et al.*, "Tumor mutation burden forecasts outcome in ovarian cancer with brca1 or brca2 mutations," *PloS One*, vol. 8, no. 11, 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0080023.

[163] Z. R. Chalmers, C. F. Connelly, D. Fabrizio, L. Gay, S. M. Ali, R. Ennis, A. Schrock, B. Campbell, A. Shlien, J. Chmielecki, *et al.*, "Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden," *Genome Medicine*, vol. 9, no. 1, pp. 1–14, 2017. [Online]. Available: https://doi.org/10.1186/s13073-017-0424-2.

[164] A. M. Goodman, S. Kato, L. Bazhenova, S. P. Patel, G. M. Frampton, V. Miller, P. J. Stephens, G. A. Daniels, and R. Kurzrock, "Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers," *Molecular Cancer Therapeutics*, vol. 16, no. 11, pp. 2598–2608, 2017. [Online]. Available: https://doi.org/10.1158/1535-7163.MCT-17-0386.

[165] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz, *The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data*, 2012. [Online]. Available: https://doi.org/10.1158/2159-8290.CD-12-0095.

[166] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal," *Science Signaling*, vol. 6, no. 269, pp. 401–404, 2013. [Online]. Available: https://doi.org/10.1126/scisignal.2004088.

[167] A. S. Weigend, D. E. Rumelhart, and B. A. Huberman, "Generalization by weight-elimination with application to forecasting," in *Advances in Neural Information Processing Systems*, 1991, pp. 875–882. [Online]. Available: https://dl.acm.org/doi/10.5555/118850.119962.

[168] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015. [Online]. Available: https://doi.org/10.1016/j.neunet.2014.09.003.

[169] S. Horvath, Y. Zhang, P. Langfelder, R. S. Kahn, M. P. Boks, K. van Eijk, L. H. van den Berg, and R. A. Ophoff, "Aging effects on dna methylation modules in human brain and blood tissue," *Genome Biology*, vol. 13, no. 10, pp. 1–18, 2012. [Online]. Available: https://doi.org/10.1186/gb-2012-13-10-r97.

[170] V. Chandran, G. Coppola, H. Nawabi, T. Omura, R. Versano, E. A. Huebner, A. Zhang, M. Costigan, A. Yekkirala, L. Barrett, *et al.*, "A systems-level analysis of the peripheral nerve intrinsic axonal growth program," *Neuron*, vol. 89, no. 5, pp. 956–970, 2016. [Online]. Available: https://doi.org/10.1016/j.neuron.2016.01.034.

[171] Z. Han, J. Zhang, G. Sun, G. Liu, and K. Huang, "A matrix rank based concordance index for evaluating and detecting conditional specific co-expressed gene modules," *BMC Genomics*, vol. 17, no. 7, pp. 303–315, 2016. [Online]. Available: https://doi.org/10.1186/s12864-016-2912-y.

[172] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems*, vol. 14, pp. 849–856, 2002. [Online]. Available: https://dl.acm.org/doi/10.5555/2980539.2980649.

[173] Y. Xiang, C.-Q. Zhang, and K. Huang, "Predicting glioblastoma prognosis networks using weighted gene co-expression network analysis on tcga data," in *BMC Bioinformatics*, Springer, vol. 13, 2012, pp. 1–8. [Online]. Available: https://doi.org/10.1186/1471-2105-13-S2-S12.

[174] J. Zhang, Z. Abrams, J. D. Parvin, and K. Huang, "Integrative analysis of somatic mutations and transcriptomic data to functionally stratify breast cancer patients," *BMC Genomics*, vol. 17, no. 7, pp. 183–193, 2016. [Online]. Available: https://doi.org/10.1186/s12864-016-2902-0.

[175] S. Shroff, J. Zhang, and K. Huang, "Gene co-expression analysis predicts genetic variants associated with drug responsiveness in lung cancer," *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 32, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5001757.

[176] J. Cheng, J. Zhang, Y. Han, X. Wang, X. Ye, Y. Meng, A. Parwani, Z. Han, Q. Feng, and K. Huang, "Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis," *Cancer Research*, vol. 77, no. 21, e91–e100, 2017. [Online]. Available: https://doi.org/10.1158/0008-5472.CAN-17-0313.

[177] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Medical Journal*, vol. 24, no. 3, pp. 69–71, 2012. [Online]. Available: https://www.ajol.info/index.php/mmj/article/view/81576.

[178] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980.

[179] T. Ma and A. Zhang, "Multi-view factorization autoencoder with network constraints for multi-omic integrative analysis," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2018, pp. 702–707. [Online]. Available: https://doi.org/10.1109/BIBM.2018.8621379.

[180] H. Steck, B. Krishnapuram, C. Dehing-Oberije, P. Lambin, and V. C. Raykar, "On ranking in survival analysis: Bounds on the concordance index," in *Advances in Neural Information Processing Systems*, 2008, pp. 1209–1216. [Online]. Available: https://dl.acm.org/doi/10.5555/2981562.2981714.

[181] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997. [Online]. Available: https://doi.org/10.1016/S0031-3203(96)00142-2.

[182] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, "Toppgene suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic Acids Research*, vol. 37, no. suppl_2, W305–W311, 2009. [Online]. Available: https://doi.org/10.1093/nar/gkp427.

[183] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan, "Enrichr: A comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Research*, vol. 44, no. W1, W90–W97, 2016. [Online]. Available: https://doi.org/10.1093/nar/gkw377.

[184] R. Setiono and H. Liu, "Neural-network feature selector," *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp. 654–662, 1997. [Online]. Available: https://doi.org/10.1109/72.572104.

[185] G. P. Zhang, "Neural networks for classification: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451–462, 2000. [Online]. Available: https://doi.org/10.1109/5326.897072.

[186] A. H. Sung and S. Mukkamala, "Identifying important features for intrusion detection using support vector machines and neural networks," in *2003 Symposium on Applications and the Internet, 2003. Proceedings.*, IEEE, 2003, pp. 209–216. [Online]. Available: https://doi.org/10.1109/SAINT.2003.1183050.

[187] H.-O. Adami, B. Malker, L. Holmberg, I. Persson, and B. Stone, "The relation between survival and age at diagnosis in breast cancer," *New England Journal of Medicine*, vol. 315, no. 9, pp. 559–563, 1986. [Online]. Available: https://doi.org/10.1056/NEJM198608283150906.

[188] B. Burwinkel, M. Wirtenberger, R. Klaes, R. K. Schmutzler, E. Grzybowska, A. Försti, B. Frank, J. L. Bermejo, P. Bugert, B. Wappenschmidt, *et al.*, "Association of ncoa3 polymorphisms with breast cancer risk," *Clinical Cancer Research*, vol. 11, no. 6, pp. 2169–2174, 2005. [Online]. Available: https://doi.org/10.1158/1078-0432.CCR-04-1621.

[189] K. B. Meyer and J. S. Carroll, "Foxa1 and breast cancer risk," *Nature Genetics*, vol. 44, no. 11, pp. 1176–1177, 2012. [Online]. Available: https://doi.org/10.1038/ng.2449.

[190] N. Rangel, N. Fortunati, S. Osella-Abate, L. Annaratone, C. Isella, M. G. Catalano, L. Rinella, J. Metovic, R. Boldorini, D. Balmativola, *et al.*, "Foxa1 and ar in invasive breast cancer: New findings on their co-expression and impact on prognosis in er-positive patients," *BMC Cancer*, vol. 18, no. 1, pp. 1–9, 2018. [Online]. Available: https://doi.org/10.1186/s12885-018-4624-y.

[191] L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu, and J. Bergh, "An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival," *Proceedings of the National Academy of Sciences*, vol. 102, no. 38, pp. 13 550–13 555, 2005. [Online]. Available: https://doi.org/10.1073/pnas.0506230102.

[192] W. T. Barry, D. N. Kernagis, H. K. Dressman, R. J. Griffis, J. D. Hunter, J. A. Olson, J. R. Marks, G. S. Ginsburg, P. K. Marcom, J. R. Nevins, *et al.*, "Intratumor heterogeneity and precision of microarray-based predictors of breast cancer biology and clinical outcome," *Journal of Clinical Oncology*, vol. 28, no. 13, p. 2198, 2010. [Online]. Available: https://doi.org/10.1200/JCO.2009.26.7245.

[193] B. Győrffy, A. Lanczky, A. C. Eklund, C. Denkert, J. Budczies, Q. Li, and Z. Szallasi, "An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients," *Breast Cancer Research and Treatment*, vol. 123, no. 3, pp. 725–731, 2010. [Online]. Available: https://doi.org/10.1007/s10549-009-0674-9.

[194] D. G. Cox, S. E. Hankinson, and D. J. Hunter, "Polymorphisms of the aurka (stk15/aurora kinase) gene and breast cancer risk (united states)," *Cancer Causes & Control*, vol. 17, no. 1, pp. 81–83, 2006. [Online]. Available: https://doi.org/10.1007/s10552-005-0429-9.

[195] G. O. Consortium, "The gene ontology project in 2008," *Nucleic Acids Research*, vol. 36, no. suppl_1, pp. D440–D444, 2008. [Online]. Available: https://doi.org/10.1093/nar/gkm883.

[196] N. Boyd, J. Byng, R. Jong, E. Fishell, L. Little, A. Miller, G. Lockwood, D. Tritchler, and M. J. Yaffe, "Quantitative classification of mammographic densities and breast cancer risk: Results from the canadian national breast screening study," *JNCI: Journal of the National Cancer Institute*, vol. 87, no. 9, pp. 670–675, 1995. [Online]. Available: https://doi.org/10.1093/jnci/87.9.670.

[197] W. Huang, B. Newman, R. C. Millikan, M. J. Schell, B. S. Hulka, and P. G. Moorman, "Hormone-related factors and risk of breast cancer in relation to estrogen receptor and progesterone receptor status," *American Journal of Epidemiology*, vol. 151, no. 7, pp. 703–714, 2000. [Online]. Available: https://doi.org/10.1093/oxfordjournals.aje.a010265.

[198] K. R. Bauer, M. Brown, R. D. Cress, C. A. Parise, and V. Caggiano, "Descriptive analysis of estrogen receptor (er)-negative, progesterone receptor (pr)-negative, and her2-negative invasive breast cancer, the so-called triple-negative phenotype: A population-based study from the california cancer registry," *Cancer*, vol. 109, no. 9, pp. 1721–1728, 2007. [Online]. Available: https://doi.org/10.1002/cncr.22618.

[199] K. Hung, R. Hayashi, A. Lafond-Walker, C. Lowenstein, D. Pardoll, and H. Levitsky, "The central role of cd4+ t cells in the antitumor immune response," *The Journal of Experimental Medicine*, vol. 188, no. 12, pp. 2357–2368, 1998. [Online]. Available: https://doi.org/10.1084/jem.188.12.2357.

[200] S. Hadrup, M. Donia, and P. Thor Straten, "Effector cd4 and cd8 t cells and their role in the tumor microenvironment," *Cancer Microenvironment*, vol. 6, no. 2, pp. 123–133, 2013. [Online]. Available: https://doi.org/10.1007/s12307-012-0127-6.

[201] E. H. Arash, A. Shiban, S. Song, and L. Attisano, "Mark4 inhibits hippo signaling to promote proliferation and migration of breast cancer cells," *EMBO Reports*, vol. 18, no. 3, pp. 420–436, 2017. [Online]. Available: https://doi.org/10.15252/embr.201642455.

[202] C. Ercolani, A. Di Benedetto, I. Terrenato, L. Pizzuti, L. Di Lauro, D. Sergi, F. Sperati, S. Buglioni, M. T. Ramieri, L. Mentuccia, *et al.*, "Expression of phosphorylated hippo pathway kinases (mst1/2 and lats1/2) in her2-positive and triple-negative breast cancer patients treated with neoadjuvant therapy," *Cancer Biology & Therapy*, vol. 18, no. 5, pp. 339–346, 2017. [Online]. Available: https://doi.org/10.1080/15384047.2017.1312230.

[203] T. Wang, J. F. Fahrmann, H. Lee, Y.-J. Li, S. C. Tripathi, C. Yue, C. Zhang, V. Lifshitz, J. Song, Y. Yuan, *et al.*, "Jak/stat3-regulated fatty acid β-oxidation is critical for breast cancer stem cell self-renewal and chemoresistance," *Cell Metabolism*, vol. 27, no. 1, pp. 136–150, 2018. [Online]. Available: https://doi.org/10.1016/j.cmet.2017.11.001.

[204] A. Lachmann, D. Torre, A. B. Keenan, K. M. Jagodnik, H. J. Lee, L. Wang, M. C. Silverstein, and A. Ma'ayan, "Massive mining of publicly available rna-seq data from human and mouse," *Nature Communications*, vol. 9, no. 1, pp. 1–10, 2018. [Online]. Available: https://doi.org/10.1038/s41467-018-03751-6.

[205] J. Totzke, D. Gurbani, R. Raphemot, P. F. Hughes, K. Bodoor, D. A. Carlson, D. R. Loiselle, A. K. Bera, L. S. Eibschutz, M. M. Perkins, *et al.*, "Takinib, a selective tak1 inhibitor, broadens the therapeutic efficacy of tnf-α inhibition for cancer and autoimmune disease," *Cell Chemical Biology*, vol. 24, no. 8, pp. 1029–1039, 2017. [Online]. Available: https://doi.org/10.1016/j.chembiol.2017.07.011.

[206] E. V Broude, B. Gyorffy, A. A Chumanevich, M. Chen, M. SJ McDermott, M. Shtutman, J. F Catroppo, and I. B Roninson, "Expression of cdk8 and cdk8-interacting genes as potential biomarkers in breast cancer," *Current Cancer Drug Targets*, vol. 15, no. 8, pp. 739–749, 2015. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26452386.

[207] T. Ren, W. Zhang, X. Liu, H. Zhao, J. Zhang, J. Zhang, X. Li, Y. Zhang, X. Bu, M. Shi, *et al.*, "Discoidin domain receptor 2 (ddr2) promotes breast cancer cell metastasis and the mechanism implicates epithelial–mesenchymal transition programme under hypoxia," *The Journal of Pathology*, vol. 234, no. 4, pp. 526–537, 2014. [Online]. Available: https://doi.org/10.1002/path.4415.

[208] Y. Feng, M. H. Chen, I. P. Moskowitz, A. M. Mendonza, L. Vidali, F. Nakamura, D. J. Kwiatkowski, and C. A. Walsh, "Filamin a (flna) is required for cell–cell contact in vascular development and cardiac morphogenesis," *Proceedings of the National Academy of Sciences*, vol. 103, no. 52, pp. 19 836–19 841, 2006. [Online]. Available: https://doi.org/10.1073/pnas.0609628104.

[209] H.-M. Tian, X.-H. Liu, W. Han, L.-L. Zhao, B. Yuan, and C.-J. Yuan, "Differential expression of filamin a and its clinical significance in breast cancer," *Oncology Letters*, vol. 6, no. 3, pp. 681–686, 2013. [Online]. Available: https://doi.org/10.3892/ol.2013.1454.

[210] G. R. de Garibay, F. Mateo, A. Stradella, R. Valdés-Mas, L. Palomero, J. Serra-Musach, D. A. Puente, A. Díaz-Navarro, G. Vargas-Parra, E. Tornero, *et al.*, "Tumor xenograft modeling identifies an association between tcf4 loss and breast cancer chemoresistance," *Disease Models & Mechanisms*, vol. 11, no. 5, 2018. [Online]. Available: https://doi.org/10.1242/dmm.032292.

[211] N. Sato, M. Maeda, M. Sugiyama, S. Ito, T. Hyodo, A. Masuda, N. Tsunoda, T. Kokuryo, M. Hamaguchi, M. Nagino, *et al.*, "Inhibition of snw 1 association with spliceosomal proteins promotes apoptosis in breast cancer cells," *Cancer Medicine*, vol. 4, no. 2, pp. 268–277, 2015. [Online]. Available: https://doi.org/10.1002/cam4.366.

[212] M. B. Yaffe, G. W. Farr, D. Miklos, A. L. Horwich, M. L. Sternlicht, and H. Sternlicht, "Tcp1 complex is a molecular chaperone in tubulin biogenesis," *Nature*, vol. 358, no. 6383, pp. 245–248, 1992. [Online]. Available: https://doi.org/10.1038/358245a0.

[213] R. Bassiouni, K. N. Nemec, A. Iketani, O. Flores, A. Showalter, A. S. Khaled, P. Vishnubhotla, R. W. Sprung, C. Kaittanis, J. M. Perez, *et al.*, "Chaperonin containing tcp-1 protein level in breast cancer cells predicts therapeutic application of a cytotoxic peptide," *Clinical Cancer Research*, vol. 22, no. 17, pp. 4366–4379, 2016. [Online]. Available: https://doi.org/10.1158/1078-0432.CCR-15-2502.

[214] W. Shan, Y. Jiang, H. Yu, Q. Huang, L. Liu, X. Guo, L. Li, Q. Mi, K. Zhang, and Z. Yang, "Hdac2 overexpression correlates with aggressive clinicopathological features and dna-damage response pathway of breast cancer," *American Journal of Cancer Research*, vol. 7, no. 5, p. 1213, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5446485.

[215] A. Shadeo and W. L. Lam, "Comprehensive copy number profiles of breast cancer cell model genomes," *Breast Cancer Research*, vol. 8, no. 1, pp. 1–14, 2006. [Online]. Available: https://doi.org/10.1186/bcr1370.

[216] S. F. Chin, A. E. Teschendorff, J. C. Marioni, Y. Wang, N. L. Barbosa-Morais, N. P. Thorne, J. L. Costa, S. E. Pinder, M. A. Van de Wiel, A. R. Green, *et al.*, "High-resolution acgh and expression profiling identifies a novel genomic subtype of er negative breast cancer," *Genome Biology*, vol. 8, no. 10, pp. 1–17, 2007. [Online]. Available: https://doi.org/10.1186/gb-2007-8-10-r215.

[217] A. Rouault, G. Banneau, G. MacGrogan, N. Jones, N. Elarouci, E. Barouk-Simonet, L. Venat, I. Coupier, E. Letouzé, A. de Reyniès, *et al.*, "Deletion of chromosomes 13q and 14q is a common feature of tumors with brca2 mutations," *PloS One*, vol. 7, no. 12, e52079, 2012. [Online]. Available: https://doi.org/10.1371/journal.pone.0052079.

[218] D. F. Quail and J. A. Joyce, "Microenvironmental regulation of tumor progression and metastasis," *Nature Medicine*, vol. 19, no. 11, p. 1423, 2013. [Online]. Available: https://doi.org/10.1038/nm.3394.

[219] C. Robertson, "The extracellular matrix in breast cancer predicts prognosis through composition, splicing, and crosslinking," *Experimental Cell Research*, vol. 343, no. 1, pp. 73–81, 2016. [Online]. Available: https://doi.org/10.1016/j.yexcr.2015.11.009.

[220] M. C. Moh and S. Shen, "The roles of cell adhesion molecules in tumor suppression and cell migration: A new paradox," *Cell Adhesion & Migration*, vol. 3, no. 4, pp. 334–336, 2009. [Online]. Available: https://doi.org/10.4161/cam.3.4.9246.

[221] S. Saadatmand, E. de Kruijf, A. Sajet, N. Dekker-Ensink, J. van Nes, H. Putter, V. Smit, C. van de Velde, G. Liefers, and P. Kuppen, "Expression of cell adhesion molecules and prognosis in breast cancer," *British Journal of Surgery*, vol. 100, no. 2, pp. 252–260, 2013. [Online]. Available: https://doi.org/10.1002/bjs.8980.

[222] A. Vincent-Salomon and J. P. Thiery, "Host microenvironment in breast cancer development: Epithelial–mesenchymal transition in breast cancer development," *Breast Cancer Research*, vol. 5, no. 2, pp. 1–6, 2003. [Online]. Available: https://doi.org/10.1186/bcr578.

[223] M. Muleris, A. Almeida, M. Gerbault-Seureau, B. Malfoy, and B. Dutrillaux, "Identification of amplified dna sequences in breast cancer and their organization within homogeneously staining regions," *Genes, Chromosomes and Cancer*, vol. 14, no. 3, pp. 155–163, 1995. [Online]. Available: https://doi.org/10.1002/gcc.2870140302.

[224] M. B. Kastan and J. Bartek, "Cell-cycle checkpoints and cancer," *Nature*, vol. 432, no. 7015, pp. 316–323, 2004. [Online]. Available: https://doi.org/10.1038/nature03097.

[225] W. Han, M.-R. Han, J. J. Kang, J.-Y. Bae, J. H. Lee, Y. J. Bae, J. E. Lee, H.-J. Shin, K.-T. Hwang, S.-E. Hwang, *et al.*, "Genomic alterations identified by array comparative genomic hybridization as prognostic markers in tamoxifen-treated estrogen receptor-positive breast cancer," *BMC Cancer*, vol. 6, no. 1, pp. 1–13, 2006. [Online]. Available: https://doi.org/10.1186/1471-2407-6-92.

[226] C. Ton, J. Guenthoer, and P. L. Porter, "Somatic alterations and implications in breast cancer," in *The Role of Genetics in Breast and Reproductive Cancers*, Springer, 2009, pp. 183–213. [Online]. Available: https://doi.org/10.1007/978-1-4419-0477-5_9.

[227] R. H. Bartels and G. H. Golub, "The simplex method of linear programming using lu decomposition," *Communications of the ACM*, vol. 12, no. 5, pp. 266–268, 1969. [Online]. Available: https://doi.org/10.1145/362946.362974.

[228] J. Ye and Q. Li, "A two-stage linear discriminant analysis via qr-decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 929–941, 2005. [Online]. Available: https://doi.org/10.1109/TPAMI.2005.110.

[229] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences*, vol. 101, no. 12, pp. 4164–4169, 2004. [Online]. Available: https://doi.org/10.1073/pnas.0308531101.

[230] W. Liu, K. Yuan, and D. Ye, "Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis," *Journal of Biomedical Informatics*, vol. 41, no. 4, pp. 602–606, 2008. [Online]. Available: https://doi.org/10.1016/j.jbi.2007.12.003.

[231] C.-H. Zheng, D.-S. Huang, L. Zhang, and X.-Z. Kong, "Tumor clustering using nonnegative matrix factorization with gene selection," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 4, pp. 599–607, 2009. [Online]. Available: https://doi.org/10.1109/TITB.2009.2018115.

[232] J. J.-Y. Wang, X. Wang, and X. Gao, "Non-negative matrix factorization by maximizing correntropy for cancer clustering," *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–11, 2013. [Online]. Available: https://doi.org/10.1186/1471-2105-14-107.

[233] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, no. 21, pp. 3970–3975, 2005. [Online]. Available: https://doi.org/10.1093/bioinformatics/bti653.

[234] G. Wang, A. V. Kossenkov, and M. F. Ochs, "Ls-nmf: A modified non-negative matrix factorization algorithm utilizing uncertainty estimates," *BMC Bioinformatics*, vol. 7, no. 1, pp. 1–10, 2006. [Online]. Available: https://doi.org/10.1186/1471-2105-7-175.

[235] M. Schmidt, G. Fung, and R. Rosales, "Optimization methods for l1-regularization," *University of British Columbia, Technical Report TR-2009*, vol. 19, 2009. [Online]. Available: https://www.cs.ubc.ca/sites/default/files/tr/2009/TR-2009-19_0.pdf.

[236] Y.-J. Lee and O. L. Mangasarian, "Ssvm: A smooth support vector machine for classification," *Computational Optimization and Applications*, vol. 20, no. 1, pp. 5–22, 2001. [Online]. Available: https://doi.org/10.1023/A:1011215321374.

[237] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756–2779, 2007. [Online]. Available: https://doi.org/10.1162/neco.2007.19.10.2756.

[238] D. González, M. Piña, and L. Torres, "Estimation of parameters in cox's proportional hazard model: Comparisons between evolutionary algorithms and the newton-raphson approach," in *Mexican International Conference on Artificial Intelligence*, Springer, 2008, pp. 513–523. [Online]. Available: https://doi.org/10.1007/978-3-540-88636-5_49.

[239] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011. [Online]. Available: https://doi.org/10.1137/090771806.

[240] P.-G. Martinsson, V. Rokhlin, and M. Tygert, "A randomized algorithm for the decomposition of matrices," *Applied and Computational Harmonic Analysis*, vol. 30, no. 1, pp. 47–68, 2011. [Online]. Available: https://doi.org/10.1016/j.acha.2010.02.003.

[241] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet, "A direct formulation for sparse pca using semidefinite programming," 3, vol. 49, SIAM, 2007, pp. 434–448. [Online]. Available: https://doi.org/10.1137/050645506.

[242] B. Everett, *An Introduction to Latent Variable Models*. Springer Science & Business Media, 2013. [Online]. Available: https://www.springer.com/gp/book/9789401089548.

[243] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 366–373. [Online]. Available: https://proceedings.mlr.press/v9/jenatton10a.

[244] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963. [Online]. Available: https://doi.org/10.1080/01621459.1963.10500845.

[245] R. Bender, T. Augustin, and M. Blettner, "Generating survival times to simulate cox proportional hazards models," *Statistics in Medicine*, vol. 24, no. 11, pp. 1713–1723, 2005. [Online]. Available: https://doi.org/10.1002/sim.2059.

[246] G. P. Wagner, K. Kin, and V. J. Lynch, "A model based criterion for gene expression calls using rna-seq data," *Theory in Biosciences*, vol. 132, no. 3, pp. 159–164, 2013. [Online]. Available: https://doi.org/10.1007/s12064-013-0178-3.

[247] M. Thattai and A. Van Oudenaarden, "Intrinsic noise in gene regulatory networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 15, pp. 8614–8619, 2001. [Online]. Available: https://doi.org/10.1073/pnas.151588598.

[248] B. Munsky, G. Neuert, and A. Van Oudenaarden, "Using gene expression noise to understand gene regulation," *Science*, vol. 336, no. 6078, pp. 183–187, 2012. [Online]. Available: https://doi.org/10.1126/science.1216379.

[249] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, vol. 18, no. 9, pp. 1509–1517, 2008. [Online]. Available: https://doi.org/10.1101/gr.079558.108.

[250] B. Li and C. N. Dewey, "Rsem: Accurate transcript quantification from rna-seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, no. 1, pp. 1–16, 2011. [Online]. Available: https://doi.org/10.1186/1471-2105-12-323.

[251] Y. Kawaguchi, T. Endo, K. Ichige, and K. Hamada, "Non-negative novelty extraction: A new non-negativity constraint for nmf," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2018, pp. 256–260. [Online]. Available: https://doi.org/10.1109/IWAENC.2018.8521320.

[252] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945. [Online]. Available: https://doi.org/10.2307/1932409.

[253] E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock, "Go::termfinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes," *Bioinformatics*, vol. 20, no. 18, pp. 3710–3715, 2004. [Online]. Available: https://doi.org/10.1093/bioinformatics/bth456.

[254] A. P. Møller and N. Saino, "Immune response and survival," *Oikos*, vol. 104, no. 2, pp. 299–304, 2004. [Online]. Available: https://doi.org/10.1111/j.0030-1299.2004.12844.x.

[255] A. House and A. Watt, "Survival and the immune response in patients with carcinoma of the colorectum," *Gut*, vol. 20, no. 10, pp. 868–874, 1979. [Online]. Available: https://doi.org/10.1136/gut.20.10.868.

[256] S. Tang, W. K. Wu, X. Li, S. H. Wong, N. Wong, M. T. Chan, J. J. Sung, and J. Yu, "Stratification of digestive cancers with different pathological features and survival outcomes by microrna expression," *Scientific Reports*, vol. 6, no. 1, pp. 1–10, 2016. [Online]. Available: https://doi.org/10.1038/srep24466.

[257] X. Li, P. Larsson, I. Ljuslinder, D. Öhlund, R. Myte, A. Löfgren-Burström, C. Zingmark, A. Ling, S. Edin, and R. Palmqvist, "Ex vivo organoid cultures reveal the importance of the tumor microenvironment for maintenance of colorectal cancer stem cells," *Cancers*, vol. 12, no. 4, p. 923, 2020. [Online]. Available: https://doi.org/10.3390/cancers12040923.

[258] J. K. Maesaka and S. Fishbane, "Regulation of renal urate excretion: A critical review," *American Journal of Kidney Diseases*, vol. 32, no. 6, pp. 917–933, 1998. [Online]. Available: https://doi.org/10.1016/S0272-6386(98)70067-8.

[259] Q. Ling and H. Kalthoff, "Transportome malfunctions and the hallmarks of pancreatic cancer," pp. 1–23, 2020. [Online]. Available: https://doi.org/10.1007/112_2020_20.

[260] D. Dou, S. Yang, Y. Lin, and J. Zhang, "An eight-mirna signature expression-based risk scoring system for prediction of survival in pancreatic adenocarcinoma," *Cancer Biomarkers*, vol. 23, no. 1, pp. 79–93, 2018. [Online]. Available: https://doi.org/10.3233/CBM-181420.

[261] L. Gullo, D. Ancona, R. Pezzilli, R. Casadei, and O. Campione, "Glucose tolerance and insulin secretion in pancreatic cancer," *The Italian Journal of Gastroenterology*, vol. 25, no. 9, pp. 487–489, 1993. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/8123896.

[262] J. Munkley, "The glycosylation landscape of pancreatic cancer," *Oncology Letters*, vol. 17, no. 3, pp. 2569–2575, 2019. [Online]. Available: https://doi.org/10.3892/ol.2019.9885.

[263] G. Ferone, J.-Y. Song, K. D. Sutherland, R. Bhaskaran, K. Monkhorst, J.-P. Lambooij, N. Proost, G. Gargiulo, and A. Berns, "Sox2 is the determining oncogenic switch in promoting lung squamous cell carcinoma from different cells of origin," *Cancer Cell*, vol. 30, no. 4, pp. 519–532, 2016. [Online]. Available: https://doi.org/10.1016/j.ccell.2016.09.001.

[264] H. J. Park, Y.-J. Cha, S. H. Kim, A. Kim, E. Y. Kim, and Y. S. Chang, "Keratinization of lung squamous cell carcinoma is associated with poor clinical outcome," *Tuberculosis and Respiratory Diseases*, vol. 80, no. 2, p. 179, 2017. [Online]. Available: https://doi.org/10.4046/trd.2017.80.2.179.

[265] M. A. Qadir, B. Kwok, W. H. Dragowska, K. H. To, D. Le, M. Bally, and S. M. Gorski, "Macroautophagy inhibition sensitizes tamoxifen-resistant breast cancer cells and enhances mitochondrial depolarization," *Breast Cancer Research and Treatment*, vol. 112, no. 3, pp. 389–403, 2008. [Online]. Available: https://doi.org/10.1007/s10549-007-9873-4.

[266] E. H. Baehrecke, "Autophagy: Dual roles in life and death?" *Nature Reviews Molecular Cell Biology*, vol. 6, no. 6, pp. 505–510, 2005. [Online]. Available: https://doi.org/10.1038/nrm1666.

[267] I. H. McKillop and L. W. Schrum, "Alcohol and liver cancer," *Alcohol*, vol. 35, no. 3, pp. 195–203, 2005. [Online]. Available: https://doi.org/10.1016/j.alcohol.2005.04.004.

[268] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 36, no. 2, pp. 59–80, 2019. [Online]. Available: https://doi.org/10.1109/MSP.2018.2877582.

[269] N. Gillis, "The why and how of nonnegative matrix factorization," *arXiv preprint arXiv:1401.5226*, 2014. [Online]. Available: https://arxiv.org/abs/1401.5226.

[270] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 211–224, 2013. [Online]. Available: https://doi.org/10.1109/TSP.2013.2285514.

[271] D. Kitamura and N. Ono, "Efficient initialization for nonnegative matrix factorization based on nonnegative independent component analysis," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2016, pp. 1–5. [Online]. Available: https://doi.org/10.1109/IWAENC.2016.7602947.

[272] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, no. 8, pp. 1301–1309, 2019. [Online]. Available: https://doi.org/10.1038/s41591-019-0508-1.

[273] E. Abels, L. Pantanowitz, F. Aeffner, M. D. Zarella, J. van der Laak, M. M. Bui, V. N. Vemuri, A. V. Parwani, J. Gibbs, E. Agosto-Arroyo, *et al.*, "Computational pathology definitions, best practices, and recommendations for regulatory guidance: A white paper from the digital pathology association," *The Journal of Pathology*, vol. 249, no. 3, pp. 286–294, 2019. [Online]. Available: https://doi.org/10.1002/path.5331.

[274] J. N. Kather, A. T. Pearson, N. Halama, D. Jäger, J. Krause, S. H. Loosen, A. Marx, P. Boor, F. Tacke, U. P. Neumann, *et al.*, "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," *Nature Medicine*, vol. 25, no. 7, pp. 1054–1056, 2019. [Online]. Available: https://doi.org/10.1038/s41591-019-0462-y.

[275] J. Cheng, Z. Han, R. Mehra, W. Shao, M. Cheng, Q. Feng, D. Ni, K. Huang, L. Cheng, and J. Zhang, "Computational analysis of pathological images enables a better diagnosis of tfe3 xp11.2 translocation renal cell carcinoma," *Nature Communications*, vol. 11, no. 1, pp. 1–9, 2020. [Online]. Available: https://doi.org/10.1038/s41467-020-15671-5.

[276] W. Shao, Z. Han, J. Cheng, L. Cheng, T. Wang, L. Sun, Z. Lu, J. Zhang, D. Zhang, and K. Huang, "Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis," *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 99–110, 2019. [Online]. Available: https://doi.org/10.1109/TMI.2019.2920608.

[277] T. Whiteside, "The tumor microenvironment and its role in promoting tumor growth," *Oncogene*, vol. 27, no. 45, pp. 5904–5912, 2008. [Online]. Available: https://doi.org/10.1038/onc.2008.271.

[278] P. Cortazar, L. Zhang, M. Untch, K. Mehta, J. P. Costantino, N. Wolmark, H. Bonnefoi, D. Cameron, L. Gianni, P. Valagussa, *et al.*, "Pathological complete response and long-term clinical benefit in breast cancer: The ctneobc pooled analysis," *The Lancet*, vol. 384, no. 9938, pp. 164–172, 2014. [Online]. Available: https://doi.org/10.1016/S0140-6736(13)62422-8.

[279] A. Berruti, V. Amoroso, F. Gallo, V. Bertaglia, E. Simoncini, R. Pedersini, L. Ferrari, A. Bottini, P. Bruzzi, and M. P. Sormani, "Pathologic complete response as a potential surrogate for the clinical outcome in patients with breast cancer after neoadjuvant therapy: A meta-regression of 29 randomized prospective studies," *Journal of Clinical Oncology*, vol. 32, no. 34, pp. 3883–3891, 2014. [Online]. Available: https://doi.org/10.1200/JCO.2014.55.2836.

[280] J. Baselga, I. Bradbury, H. Eidtmann, S. Di Cosimo, E. De Azambuja, C. Aura, H. Gómez, P. Dinh, K. Fauria, V. Van Dooren, *et al.*, "Lapatinib with trastuzumab for her2-positive early breast cancer (neoaltto): A randomised, open-label, multicentre, phase 3 trial," *The Lancet*, vol. 379, no. 9816, pp. 633–640, 2012. [Online]. Available: https://doi.org/10.1016/S0140-6736(11)61847-3.

[281] M. Untch, P. A. Fasching, G. E. Konecny, S. Hasmüller, A. Lebeau, R. Kreienberg, O. Camara, V. Müller, A. du Bois, T. Kühn, *et al.*, "Pathologic complete response after neoadjuvant chemotherapy plus trastuzumab predicts favorable survival in human epidermal growth factor receptor 2–overexpressing breast cancer: Results from the techno trial of the ago and gbg study groups," *Journal of Clinical Oncology*, vol. 29, no. 25, pp. 3351–3357, 2011. [Online]. Available: https://doi.org/10.1200/JCO.2010.31.4930.

[282] V. Guarneri, A. Frassoldati, A. Bottini, K. Cagossi, G. Bisagni, S. Sarti, A. Ravaioli, L. Cavanna, G. Giardina, A. Musolino, *et al.*, "Preoperative chemotherapy plus trastuzumab, lapatinib, or both in human epidermal growth factor receptor 2–positive operable breast cancer: Results of the randomized phase ii cher-lob study," *Journal of Clinical Oncology*, vol. 30, no. 16, pp. 1989–1995, 2012. [Online]. Available: https://doi.org/10.1200/JCO.2011.39.0823.

[283] W. Y. Sun, Y. K. Lee, and J. S. Koo, "Expression of pd-l1 in triple-negative breast cancer based on different immunohistochemical antibodies," *Journal of Translational Medicine*, vol. 14, no. 1, pp. 1–12, 2016. [Online]. Available: https://doi.org/10.1186/s12967-016-0925-6.

[284] A. Sánchez-Muñoz, V. Navarro-Perez, Y. Plata-Fernández, A. Santonja, I. Moreno, N. Ribelles, and E. Alba, "Proliferation determined by ki-67 defines different pathologic response to neoadjuvant trastuzumab-based chemotherapy in her2-positive breast cancer," *Clinical Breast Cancer*, vol. 15, no. 5, pp. 343–347, 2015. [Online]. Available: https://doi.org/10.1016/j.clbc.2015.01.005.

[285] A. Seo, H. Lee, E. Kim, H. Kim, M. Jang, H. Lee, Y. J. Kim, J. H. Kim, and S. Y. Park, "Tumour-infiltrating cd8+ lymphocytes as an independent predictive factor for pathological complete response to primary systemic therapy in breast cancer," *British Journal of Cancer*, vol. 109, no. 10, pp. 2705–2713, 2013. [Online]. Available: https://doi.org/10.1038/bjc.2013.634.

[286] H. Hornychova, B. Melichar, M. Tomšová, J. Mergancová, H. Urminska, and A. Ryška, "Tumor-infiltrating lymphocytes predict response to neoadjuvant chemotherapy in patients with breast carcinoma," *Cancer Investigation*, vol. 26, no. 10, pp. 1024–1031, 2008. [Online]. Available: https://doi.org/10.1080/07357900802098165.

[287] R. Yamaguchi, M. Tanaka, A. Yano, M. T. Gary, M. Yamaguchi, K. Koura, N. Kanomata, A. Kawaguchi, J. Akiba, Y. Naito, *et al.*, "Tumor-infiltrating lymphocytes are important pathologic predictors for neoadjuvant chemotherapy in patients with breast cancer," *Human Pathology*, vol. 43, no. 10, pp. 1688–1694, 2012. [Online]. Available: https://doi.org/10.1016/j.humpath.2011.12.013.

[288] M. Ono, H. Tsuda, C. Shimizu, S. Yamamoto, T. Shibata, H. Yamamoto, T. Hirata, K. Yonemori, M. Ando, K. Tamura, *et al.*, "Tumor-infiltrating lymphocytes are correlated with response to neoadjuvant chemotherapy in triple-negative breast cancer," *Breast Cancer Research and Treatment*, vol. 132, no. 3, pp. 793–805, 2012. [Online]. Available: https://doi.org/10.1007/s10549-011-1554-7.

[289] S. M. Mahmoud, E. C. Paish, D. G. Powe, R. D. Macmillan, M. J. Grainge, A. H. Lee, I. O. Ellis, and A. R. Green, "Tumor-infiltrating cd8+ lymphocytes predict clinical outcome in breast cancer," *Journal of Clinical Oncology*, vol. 29, no. 15, pp. 1949–1955, 2011. [Online]. Available: https://doi.org/10.1200/JCO.2010.30.5037.

[290] H. J. Lee, J.-Y. Seo, J.-H. Ahn, S.-H. Ahn, and G. Gong, "Tumor-associated lymphocytes predict response to neoadjuvant chemotherapy in breast cancer patients," *Journal of Breast Cancer*, vol. 16, no. 1, pp. 32–39, 2013. [Online]. Available: https://doi.org/10.4048/jbc.2013.16.1.32.

[291] C. Denkert, S. Loibl, A. Noske, M. Roller, B. Muller, M. Komor, J. Budczies, S. Darb-Esfahani, R. Kronenwett, C. Hanusch, *et al.*, "Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer," *Journal of Clinical Oncology*, vol. 28, no. 1, pp. 105–113, 2010. [Online]. Available: https://doi.org/10.1200/JCO.2009.23.7370.

[292] Y. Issa-Nummer, S. Darb-Esfahani, S. Loibl, G. Kunz, V. Nekljudova, I. Schrader, B. V. Sinn, H.-U. Ulmer, R. Kronenwett, M. Just, *et al.*, "Prospective validation of immunological infiltrate for prediction of response to neoadjuvant chemotherapy in her2-negative breast cancer–a substudy of the neoadjuvant geparquinto trial," *PloS One*, vol. 8, no. 12, e79775, 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0079775.

[293] S. Loi, N. Sirtaine, F. Piette, R. Salgado, G. Viale, F. Van Eenoo, G. Rouas, P. Francis, J. Crown, E. Hitre, *et al.*, "Prognostic and predictive value of tumor-infiltrating lymphocytes in a phase iii randomized adjuvant breast cancer trial in node-positive breast cancer comparing the addition of docetaxel to doxorubicin with doxorubicin-based chemotherapy: Big 02-98," *Journal of Clinical Oncology*, vol. 31, no. 7, pp. 860–867, 2013. [Online]. Available: https://doi.org/10.1200/JCO.2011.41.0902.

[294] L. Gianni, M. Zambetti, K. Clark, J. Baker, M. Cronin, J. Wu, G. Mariani, J. Rodriguez, M. Carcangiu, D. Watson, *et al.*, "Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer," *Journal of Clinical Oncology*, vol. 23, no. 29, pp. 7265–7277, 2005. [Online]. Available: https://doi.org/10.1200/JCO.2005.02.0818.

[295] H. Wimberly, J. R. Brown, K. Schalper, H. Haack, M. R. Silver, C. Nixon, V. Bossuyt, L. Pusztai, D. R. Lannin, and D. L. Rimm, "Pd-l1 expression correlates with tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy in breast cancer," *Cancer Immunology Research*, vol. 3, no. 4, pp. 326–332, 2015. [Online]. Available: https://doi.org/10.1158/2326-6066.CIR-14-0133.

[296] H.-M. Baek, J.-H. Chen, K. Nie, H. J. Yu, S. Bahri, R. S. Mehta, O. Nalcioglu, and M.-Y. Su, "Predicting pathologic response to neoadjuvant chemotherapy in breast cancer by using mr imaging and quantitative 1h mr spectroscopy," *Radiology*, vol. 251, no. 3, pp. 653–662, 2009. [Online]. Available: https://doi.org/10.1148/radiol.2512080553.

[297] N. Michoux, S. Van den Broeck, L. Lacoste, L. Fellah, C. Galant, M. Berlière, and I. Leconte, "Texture analysis on mr images helps predicting non-response to nac in breast cancer," *BMC Cancer*, vol. 15, no. 1, p. 574, 2015. [Online]. Available: https://doi.org/10.1186/s12885-015-1563-8.

[298] J. Lee, S. H. Kim, and B. J. Kang, "Pretreatment prediction of pathologic complete response to neoadjuvant chemotherapy in breast cancer: Perfusion metrics of dynamic contrast enhanced mri," *Scientific Reports*, vol. 8, no. 1, pp. 1–8, 2018. [Online]. Available: https://doi.org/10.1038/s41598-018-27764-9.

[299] Y.-H. Qu, H.-T. Zhu, K. Cao, X.-T. Li, M. Ye, and Y.-S. Sun, "Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using a deep learning (dl) method," *Thoracic Cancer*, vol. 11, no. 3, pp. 651–658, 2020. [Online]. Available: https://doi.org/10.1111/1759-7714.13309.

[300] H. R. Ali, A. Dariush, E. Provenzano, H. Bardwell, J. E. Abraham, M. Iddawela, A.-L. Vallier, L. Hiller, J. A. Dunn, S. J. Bowden, *et al.*, "Computational pathology of pre-treatment biopsies identifies lymphocyte density as a predictor of response to neoadjuvant chemotherapy in breast cancer," *Breast Cancer Research*, vol. 18, no. 1, pp. 1–11, 2016. [Online]. Available: https://doi.org/10.1186/s13058-016-0682-8.

[301] H. W. Hwang, H. Jung, J. Hyeon, Y. H. Park, J. S. Ahn, Y.-H. Im, S. J. Nam, S. W. Kim, J. E. Lee, J.-H. Yu, *et al.*, "A nomogram to predict pathologic complete response (pcr) and the value of tumor-infiltrating lymphocytes (tils) for prediction of response to neoadjuvant chemotherapy (nac) in breast cancer patients," *Breast Cancer Research and Treatment*, vol. 173, no. 2, pp. 255–266, 2019. [Online]. Available: https://doi.org/10.1007/s10549-018-4981-x.

[302] C. Denkert, G. Von Minckwitz, J. C. Brase, B. V. Sinn, S. Gade, R. Kronenwett, B. M. Pfitzner, C. Salat, S. Loi, W. D. Schmitt, *et al.*, "Tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy with or without carboplatin in human epidermal growth factor receptor 2-positive and triple-negative primary breast cancers," *Journal of Clinical Oncology*, vol. 33, no. 9, pp. 983–991, 2015. [Online]. Available: https://doi.org/10.1200/JCO.2014.58.1967.

[303] S. Adams, R. J. Gray, S. Demaria, L. Goldstein, E. A. Perez, L. N. Shulman, S. Martino, M. Wang, V. E. Jones, T. J. Saphner, *et al.*, "Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase iii randomized adjuvant breast cancer trials: Ecog 2197 and ecog 1199," *Journal of Clinical Oncology*, vol. 32, no. 27, p. 2959, 2014. [Online]. Available: https://doi.org/10.1200/JCO.2013.55.0491.

[304] P. G. Tsoutsou, J. Bourhis, and G. Coukos, "Tumor-infiltrating lymphocytes in triple-negative breast cancer: A biomarker for use beyond prognosis?" *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, vol. 33, no. 11, pp. 1297–1298, 2015. [Online]. Available: https://doi.org/10.1200/jco.2014.59.2808.

[305] R. Salgado, C. Denkert, S. Demaria, N. Sirtaine, F. Klauschen, G. Pruneri, S. Wienert, G. Van den Eynden, F. L. Baehner, F. Pénault-Llorca, *et al.*, "The evaluation of tumor-infiltrating lymphocytes (tils) in breast cancer: Recommendations by an international tils working group 2014," *Annals of Oncology*, vol. 26, no. 2, pp. 259–271, 2015. [Online]. Available: https://doi.org/10.1093/annonc/mdu450.

[306] C. Denkert, G. von Minckwitz, S. Darb-Esfahani, B. Lederer, B. I. Heppner, K. E. Weber, J. Budczies, J. Huober, F. Klauschen, J. Furlanetto, *et al.*, "Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: A pooled analysis of 3771 patients treated with neoadjuvant therapy," *The Lancet Oncology*, vol. 19, no. 1, pp. 40–50, 2018. [Online]. Available: https://doi.org/10.1016/S1470-2045(17)30904-X.

[307] F. Zhang, M. Huang, H. Zhou, K. Chen, J. Jin, Y. Wu, L. Ying, X. Ding, D. Su, and D. Zou, "A nomogram to predict the pathologic complete response of neoadjuvant chemotherapy in triple-negative breast cancer based on simple laboratory indicators," *Annals of Surgical Oncology*, vol. 26, no. 12, pp. 3912–3919, 2019. [Online]. Available: https://doi.org/10.1245/s10434-019-07655-7.

[308] S. Blom, L. Paavolainen, D. Bychkov, R. Turkki, P. Mäki-Teeri, A. Hemmes, K. Välimäki, J. Lundin, O. Kallioniemi, and T. Pellinen, "Systems pathology by multiplexed immunohistochemistry and whole-slide digital image analysis," *Scientific Reports*, vol. 7, no. 1, pp. 1–13, 2017. [Online]. Available: https://doi.org/10.1038/s41598-017-15798-4.

[309] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, "Openslide: A vendor-neutral software foundation for digital pathology," *Journal of Pathology Informatics*, vol. 4, 2013. [Online]. Available: https://doi.org/10.4103/2153-3539.119005.

[310] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html.

[311] I. Culjak, D. Abram, T. Pribanic, H. Dzapo, and M. Cifrek, "A brief introduction to opencv," in *2012 Proceedings of the 35th International Convention MIPRO*, IEEE, 2012, pp. 1725–1730. [Online]. Available: https://ieeexplore.ieee.org/document/6240859.

[312] A. C. Wolff, M. E. H. Hammond, D. G. Hicks, M. Dowsett, L. M. McShane, K. H. Allison, D. C. Allred, J. M. Bartlett, M. Bilous, P. Fitzgibbons, *et al.*, "Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline update," *Archives of Pathology and Laboratory Medicine*, vol. 138, no. 2, pp. 241–256, 2014. [Online]. Available: https://doi.org/10.5858/arpa.2013-0953-SA.

[313] W. F. Symmans, F. Peintinger, C. Hatzis, R. Rajan, H. Kuerer, V. Valero, L. Assad, A. Poniecka, B. Hennessy, M. Green, *et al.*, "Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy," *Journal of Clinical Oncology*, vol. 25, no. 28, pp. 4414–4422, 2007. [Online]. Available: https://doi.org/10.1200/JCO.2007.10.6823.

[314] Y. Hou, H. Nitta, L. Wei, P. M. Banks, M. Lustberg, R. Wesolowski, B. Ramaswamy, A. V. Parwani, and Z. Li, "Pd-l1 expression and cd8-positive t cells are associated with favorable survival in her2-positive invasive breast cancer," *The Breast Journal*, vol. 24, no. 6, pp. 911–919, 2018. [Online]. Available: https://doi.org/10.1111/tbj.13112.

[315] Y. Hou, H. Nitta, L. Wei, P. M. Banks, A. V. Parwani, and Z. Li, "Evaluation of immune reaction and pd-l1 expression using multiplex immunohistochemistry in her2-positive breast cancer: The association with response to anti-her2 neoadjuvant therapy," *Clinical Breast Cancer*, vol. 18, no. 2, e237–e244, 2018. [Online]. Available: https://doi.org/10.1016/j.clbc.2017.11.001.

[316] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The cancer genome atlas (tcga): An immeasurable source of knowledge," *Contemporary Oncology*, vol. 19, no. 1A, A68, 2015. [Online]. Available: https://doi.org/10.5114/wo.2014.47136.

[317] M. Amgad, H. Elfandy, H. Hussein, L. A. Atteya, M. A. Elsebaie, L. S. Abo Elnasr, R. A. Sakr, H. S. Salem, A. F. Ismail, A. M. Saad, *et al.*, "Structured crowdsourcing enables convolutional segmentation of histology images," *Bioinformatics*, vol. 35, no. 18, pp. 3461–3467, 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz083.

[318] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90.

[319] S. Wang, R. Rong, D. M. Yang, J. Fujimoto, S. Yan, L. Cai, L. Yang, D. Luo, C. Behrens, E. R. Parra, *et al.*, "Computational staining of pathology images to study the tumor microenvironment in lung cancer," *Cancer Research*, vol. 80, no. 10, pp. 2056–2066, 2020. [Online]. Available: https://doi.org/10.1158/0008-5472.CAN-19-1629.

[320] M.-N. Wu, C.-C. Lin, and C.-C. Chang, "Brain tumor detection using color-based k-means clustering segmentation," in *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007)*, IEEE, vol. 2, 2007, pp. 245–250. [Online]. Available: https://doi.org/10.1109/IIHMSP.2007.4457697.

[321] Z. Huang, Z. Han, A. Parwani, K. Huang, and Z. Li, "Predicting response to neoadjuvant chemotherapy in her2-positive breast cancer using machine learning models with combined tissue imaging and clinical features," in *Laboratory Investigation*, NATURE PUBLISHING GROUP 75 VARICK ST, 9TH FLR, NEW YORK, NY 10013-1917 USA, vol. 99, 2019.

[322] Z. Huang, Z. Han, A. V. Parwani, K. Huang, and Z. Li, "Artificial intelligence driven neoadjuvant chemotherapy response prediction in triple negative breast cancer (tnbc) unveils non-linear feature interactions," in *Laboratory investigation*, NATURE PUBLISHING GROUP 75 VARICK ST, 9TH FLR, NEW YORK, NY 10013-1917 USA, vol. 100, 2020, pp. 1459–1461.

[323] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *European Conference on Information Retrieval*, Springer, 2005, pp. 345–359. [Online]. Available: https://doi.org/10.1007/978-3-540-31865-1_25.

[324] C. Spearman, "The proof and measurement of association between two things," 1961. [Online]. Available: https://doi.org/10.1037/11491-005.

[325] J. L. Meisel, J. Zhao, A. Suo, C. Zhang, Z. Wei, C. Taylor, R. Aneja, U. Krishnamurti, Z. Li, R. Nahta, *et al.*, "Clinicopathologic factors associated with response to neoadjuvant anti-her2–directed chemotherapy in her2-positive breast cancer," *Clinical Breast Cancer*, vol. 20, no. 1, pp. 19–24, 2020. [Online]. Available: https://doi.org/10.1016/j.clbc.2019.09.003.

[326] M. Miyashita, H. Sasano, K. Tamaki, H. Hirakawa, Y. Takahashi, S. Nakagawa, G. Watanabe, H. Tada, A. Suzuki, N. Ohuchi, *et al.*, "Prognostic significance of tumor-infiltrating cd8+ and foxp3+ lymphocytes in residual tumors and alterations in these parameters after neoadjuvant chemotherapy in triple-negative breast cancer: A retrospective multicenter study," *Breast Cancer Research*, vol. 17, no. 1, p. 124, 2015. [Online]. Available: https://doi.org/10.1186/s13058-015-0632-x.

[327] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Computational and Structural Biotechnology Journal*, vol. 16, pp. 34–42, 2018. [Online]. Available: https://doi.org/10.1016/j.csbj.2018.01.001.

[328] I. B. Buchwalow and W. Böcker, "Multiple multicolor immunoenzyme staining," in *Immunohistochemistry: Basics and Methods*, Springer, 2010, pp. 61–67. [Online]. Available: http://doi.org/10.1007/978-3-642-04609-4_7.

[329] A. Petrovic, M. Abramovic, D. Mihailovic, J. Gligorijevic, V. Zivkovic, M. Mojsilovic, and I. Ilic, "Multicolor counterstaining for immunohistochemistry–a modified movat's pentachrome," *Biotechnic & Histochemistry*, vol. 86, no. 6, pp. 429–435, 2011. [Online]. Available: https://doi.org/10.1016/j.ymeth.2014.08.016.

[330] S. Vranic, F. S. Cyprian, Z. Gatalica, and J. Palazzo, "Pd-l1 status in breast cancer: Current view and perspectives," in *Seminars in Cancer Biology*, Elsevier, 2019. [Online]. Available: https://doi.org/10.1016/j.semcancer.2019.12.003.

[331] C. Sun, R. Mezzadra, and T. N. Schumacher, "Regulation and function of the pd-l1 checkpoint," *Immunity*, vol. 48, no. 3, pp. 434–452, 2018. [Online]. Available: https://doi.org/10.1016/j.immuni.2018.03.014.

[332] P. Dong, Y. Xiong, J. Yue, S. J. Hanley, and H. Watari, "Tumor-intrinsic pd-l1 signaling in cancer initiation, development and treatment: Beyond immune evasion," *Frontiers in Oncology*, vol. 8, p. 386, 2018. [Online]. Available: https://doi.org/10.3389/fonc.2018.00386.

[333] M. Z. Baptista, L. O. Sarian, S. F. Derchain, G. A. Pinto, and J. Vassallo, "Prognostic significance of pd-l1 and pd-l2 in breast cancer," *Human pathology*, vol. 47, no. 1, pp. 78–84, 2016. [Online]. Available: https://doi.org/10.1016/j.humpath.2015.09.006.

[334] R. Sabatier, P. Finetti, E. Mamessier, J. Adelaide, M. Chaffanet, H. R. Ali, P. Viens, C. Caldas, D. Birnbaum, and F. Bertucci, "Prognostic and predictive value of pdl1 expression in breast cancer," *Oncotarget*, vol. 6, no. 7, p. 5449, 2015. [Online]. Available: https://doi.org/10.18632/oncotarget.3216.

[335] R. K. Beckers, C. I. Selinger, R. Vilain, J. Madore, J. S. Wilmott, K. Harvey, A. Holliday, C. L. Cooper, E. Robbins, D. Gillett, *et al.*, "Programmed death ligand 1 expression in triple-negative breast cancer is associated with tumour-infiltrating lymphocytes and improved outcome," *Histopathology*, vol. 69, no. 1, pp. 25–34, 2016. [Online]. Available: https://doi.org/10.1111/his.12904.

[336] E. A. Mittendorf, A. V. Philips, F. Meric-Bernstam, N. Qiao, Y. Wu, S. Harrington, X. Su, Y. Wang, A. M. Gonzalez-Angulo, A. Akcakanat, *et al.*, "Pd-l1 expression in triple-negative breast cancer," *Cancer Immunology Research*, vol. 2, no. 4, pp. 361–370, 2014. [Online]. Available: https://doi.org/10.1158/2326-6066.CIR-13-0127.

[337] E. A. Dill, A. A. Gru, K. A. Atkins, L. A. Friedman, M. E. Moore, T. N. Bullock, J. V. Cross, P. M. Dillon, and A. M. Mills, "Pd-l1 expression and intratumoral heterogeneity across breast cancer subtypes and stages," *The American Journal of Surgical Pathology*, vol. 41, no. 3, pp. 334–342, 2017. [Online]. Available: https://doi.org/10.1097/PAS.0000000000000780.

[338] S. B. Bae, H. D. Cho, M.-H. Oh, J.-H. Lee, S.-H. Jang, S. A. Hong, J. Cho, S. Y. Kim, S. W. Han, J. E. Lee, *et al.*, "Expression of programmed death receptor ligand 1 with high tumor-infiltrating lymphocytes is associated with better prognosis in breast cancer," *Journal of Breast Cancer*, vol. 19, no. 3, pp. 242–251, 2016. [Online]. Available: https://doi.org/10.4048/jbc.2016.19.3.242.

[339] H. Mori, M. Kubo, R. Yamaguchi, R. Nishimura, T. Osako, N. Arima, Y. Okumura, M. Okido, M. Yamada, M. Kai, *et al.*, "The combination of pd-l1 expression and decreased tumor-infiltrating lymphocytes is associated with a poor prognosis in triple-negative breast cancer," *Oncotarget*, vol. 8, no. 9, p. 15 584, 2017. [Online]. Available: https://doi.org/10.18632/oncotarget.14698.

[340] A. Cimino-Mathews, E. Thompson, J. M. Taube, X. Ye, Y. Lu, A. Meeker, H. Xu, R. Sharma, K. Lecksell, T. C. Cornish, *et al.*, "Pd-l1 (b7-h1) expression and the immune tumor microenvironment in primary and metastatic breast carcinomas," *Human Pathology*, vol. 47, no. 1, pp. 52–63, 2016. [Online]. Available: https://doi.org/10.1016/j.humpath.2015.09.003.

[341] P. Kong, J. Wang, Z. Song, S. Liu, W. He, C. Jiang, Q. Xie, L. Yang, X. Xia, and L. Xia, "Circulating lymphocytes, pd-l1 expression on tumor-infiltrating lymphocytes, and survival of colorectal cancer patients with different mismatch repair gene status," *Journal of Cancer*, vol. 10, no. 7, p. 1745, 2019. [Online]. Available: https://doi.org/10.7150/jca.25187.

[342] J. Zhang, T. Knobloch, J. Parvin, C. Weghorst, and K. Huang, "Identifying smoking associated gene co-expression networks related to oral cancer initiation," in *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, IEEE, 2011, pp. 1039–1041. [Online]. Available: https://doi.org/10.1109/BIBMW.2011.6112553.

[343] J. Zhang, Y. Xiang, L. Ding, T. B. Borlawsky, H. G. Ozer, R. Jin, P. Payne, and K. Huang, "Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia," in *BMC Bioinformatics*, Springer, vol. 11, 2010, pp. 1–9. [Online]. Available: https://doi.org/10.1186/1471-2105-11-S9-S5.

[344] Y. Cao, J. Zhu, P. Jia, and Z. Zhao, "Scrnaseqdb: A database for gene expression profiling in human single cell by rna-seq," *Genes*, vol. 8, no. 12, pp. 1–10, 2017. [Online]. Available: https://doi.org/10.3390/genes8120368.

[345] L. Collado-Torres, A. Nellore, K. Kammers, S. E. Ellis, M. A. Taub, K. D. Hansen, A. E. Jaffe, B. Langmead, and J. T. Leek, "Reproducible rna-seq analysis using recount2," *Nature Biotechnology*, vol. 35, no. 4, pp. 319–321, 2017. [Online]. Available: https://doi.org/10.1038/nbt.3838.

[346] T. J. Freeman, J. J. Smith, X. Chen, M. K. Washington, J. T. Roland, A. L. Means, S. A. Eschrich, T. J. Yeatman, N. G. Deane, and R. D. Beauchamp, "Smad4-mediated signaling inhibits intestinal neoplasia by inhibiting expression of β-catenin," *Gastroenterology*, vol. 142, no. 3, pp. 562–571, 2012. [Online]. Available: https://doi.org/10.1053/j.gastro.2011.11.026.

[347] J. J. Smith, N. G. Deane, F. Wu, N. B. Merchant, B. Zhang, A. Jiang, P. Lu, J. C. Johnson, C. Schmidt, C. E. Bailey, *et al.*, "Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer," *Gastroenterology*, vol. 138, no. 3, pp. 958–968, 2010. [Online]. Available: https://doi.org/10.1053/j.gastro.2009.11.005.

[348] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, *et al.*, "Orchestrating high-throughput genomic analysis with bioconductor," *Nature Methods*, vol. 12, no. 2, p. 115, 2015. [Online]. Available: https://doi.org/10.1038/nmeth.3252.

[349] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan, "Enrichr: Interactive and collaborative html5 gene list enrichment analysis tool," *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–14, 2013. [Online]. Available: https://doi.org/10.1186/1471-2105-14-128.

[350] Z. Gu, L. Gu, R. Eils, M. Schlesner, and B. Brors, "Circlize implements and enhances circular visualization in r," *Bioinformatics*, vol. 30, no. 19, pp. 2811–2812, 2014. [Online]. Available: https://doi.org/10.1093/bioinformatics/btu393.

[351] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, "The human genome browser at ucsc," *Genome Research*, vol. 12, no. 6, pp. 996–1006, 2002. [Online]. Available: https://doi.org/10.1101/gr.229102.

[352] J. M. Bland and D. G. Altman, "Survival probabilities (the kaplan-meier method)," *BMJ*, vol. 317, no. 7172, pp. 1572–1580, 1998. [Online]. Available: https://doi.org/10.1136/bmj.317.7172.1572.

[353] D. G. Kleinbaum and M. Klein, "Kaplan-meier survival curves and the log-rank test," in *Survival Analysis*, Springer, 2012, ch. 2, pp. 55–96. [Online]. Available: https://doi.org/10.1007/978-1-4419-6646-9_2.

[354] S. Suciu, A. M. Eggermont, P. Lorigan, J. M. Kirkwood, S. N. Markovic, C. Garbe, D. Cameron, S. Kotapati, T.-T. Chen, K. Wheatley, *et al.*, "Relapse-free survival as a surrogate for overall survival in the evaluation of stage ii–iii melanoma adjuvant therapy," *JNCI: Journal of the National Cancer Institute*, vol. 110, no. 1, pp. 87–96, 2018. [Online]. Available: https://doi.org/10.1093/jnci/djx133.

VITA

Zhi Huang was born in Hefei, China in 1994. In June 2015, he received his Bachelor of Science degree in Automation (BS–MS straight entrance class) from Xi'an Jiaotong University School of Electronic and Information Engineering. He was an exchange student in the University of Michigan – Dearborn School of Electrical and Computer Engineering in his senior year (August 2014 – May 2015).

In December 2016, he received his Master of Science degree in Computer Engineering from the Purdue School of Electrical and Computer Engineering, Indiana University – Purdue University Indianapolis, Indiana, USA.

His Ph.D. program is at Purdue University, West Lafayette, Indiana, USA, with the major of Electrical and Computer Engineering, and the area of Communications, Networking, Signal and Image Processing. His current research interests include machine and deep learning, matrix factorization, survival analysis, bioinformatics, computational biology, and biomedical image analysis. He is a research assistant at Indiana University School of Medicine. He is a student member of the IEEE and the IEEE Young Professionals since 2016. He has served as a reviewer since 2017 for Scientific Report (2020), Medical Science Monitor (2019), the International Conference on Intelligent Biology and Medicine (ICIBM) (2019), the Medical Image Computing and Computer Assisted Intervention (MICCAI) (2017, 2018, 2020), the AMIA Annual Symposium (2020), and ACM SIGKDD Conference on Knowledge Discovery and Data Mining (2020).

From May 2019 to August 2019, he was at Philips Research North America, Cambridge, Massachusetts, USA as a Research Intern.