Barcelona School of Economics

Master's Degree in Data Science:
Data Science Methodology Program

Biomarker Selection in Liver and Breast Cancer via Bayesian and Regularized Learning with Integrated Data

Authors: Elisa Scocco, Iñigo Exposito, Tirdod Behbehani

Director: David Rossell

04 June 2025

ABSTRACT IN ENGLISH:

RNA-sequencing data offers genome-wide biomarker discovery potential, but its high dimensionality poses challenges. We evaluate penalized regression and bayesian variable selection methods to identify if breast cancer gene expression data can be used to improve biomarker selection in liver cancer. Using TCGA data, we assessed LASSO, Ridge, Elastic Net, and Bayesian approaches. Cross-cancer modeling did not lead to performance improvements. However, using prior information for a within-liver cancer model did improve performance. Liver and breast cancer are too dissimilar for cross-cancer modeling, but our work offers guidance on how to conduct transfer learning on similar diseases.

ABSTRACT IN CATALAN/ SPANISH:

Los datos de secuenciación de ARN permiten la identificación de biomarcadores a escala genómica, aunque su alta dimensionalidad supone un reto. En este estudio se evalúan métodos de regresión penalizada y selección bayesiana de variables para determinar si la expresión génica en el cáncer de mama puede mejorar la selección de biomarcadores en el cáncer de hígado. Utilizando datos de TCGA, se aplicaron LASSO, Ridge, Elastic Net y enfoques bayesianos. El modelado entre estos tipos de cáncer no mejoró el rendimiento de los modelos, pero incorporar información previa en modelos intrahepáticos sí lo hizo. Los cánceres de mama e hígado presentan diferencias biológicas significativas que limitan el modelado cruzado; no obstante, este trabajo ofrece orientación para aplicar aprendizaje por transferencia en enfermedades con mayor similitud.

KEYWORDS IN ENGLISH :

Bayesian Statistics, Regularized Learning, Biomarker Selection

KEYWORDS IN CATALAN/ SPANISH:

Estadística Bayesiana, Aprendizaje Regularizado, Selección de Biomarcadores

Contents

1 Abbreviations and Symbols

We present a list of abbreviations and symbols along with their corresponding definitions and explanations, which will be continuously used throughout this thesis.

Data Sources and Projects.

Abbreviation / Symbol	Definition
UCSC Xena	University of California Santa Cruz Xena platform.
GDC	Genomic Data Commons.
TCGA	The Cancer Genome Atlas.

Cancer Types and Molecular Data.

Abbreviation / Symbol	Definition
LIHC/HCC	Liver Hepatocellular Carcinoma.
BRCA	Breast Invasive Carcinoma.
RNA-seq	Ribonucleic acid sequencing.
TPM	Transcripts Per Million.
ENSG	Ensembl Gene Identifier.

Statistical Terms and Metrics.

Abbreviation / Symbol	Definition
Q1	First quartile.
Q3	Third quartile.
IQR	Interquartile range.
SD	Standard deviation.
Cov	Covariate.

Clinical Variables and Outcomes.

Abbreviation / Symbol	Definition
OS	Overall survival status.
OS.time	Time to death or censoring.
AJCC	American Joint Committee on Cancer.
TNM	Tumor, Node, Metastasis classification system.
T	Tumor size.
N	Lymph node involvement.
M	Distant metastasis.
Is_stage_34	Binary indicator of cancer stage (0 = early, 1 = advanced).
y	Dependent or outcome variable.

Sample and Model Variables.

Abbreviation / Symbol	Definition
n	Sample size.
S	Sample of n units.
$\mathbf{x_i}$	Covariate vector for unit i.
y_i	Response variable for unit i.
$\hat{y}(c)$	Predicted binary outcome at threshold c.
D	True class label indicator.
k	Number of model parameters estimated.

Model Parameters.

Abbreviation / Symbol	Definition
$\boldsymbol{\beta}$	Vector of regression coefficients.
$\hat{\boldsymbol{\beta}}$	Maximum Likelihood Estimate (MLE) of $\boldsymbol{\beta}$.
$\hat{\boldsymbol{\beta}}^{\boldsymbol{L}}$	LASSO solution for $\boldsymbol{\beta}$.
$\hat{\boldsymbol{\beta}}^{\boldsymbol{R}}$	Ridge regression solution for $\boldsymbol{\beta}$.
$\hat{\boldsymbol{\beta}}^{\boldsymbol{EN}}$	Elastic Net solution for $\boldsymbol{\beta}$.
$\hat{\boldsymbol{\beta}}^{\boldsymbol{AL}}$	Adaptive LASSO solution for $\boldsymbol{\beta}$.

Probability and Regularized Likelihood Functions.

Abbreviation / Symbol	Definition
$p(\mathbf{x_i})$	Conditional probability of success given x_i.
$\ell(\cdot)$	Log-likelihood function.
$\ell_p(\cdot)$	Penalized log-likelihood function.
$\mathbb{E}$	Expectation operator.
$\Omega(\cdot)$	Regularization (penalty) function.
λ	Regularization parameter.
α	Elastic Net mixing parameter.
CV	Cross-validation.
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2.
Bern(p)	Bernoulli distribution with success probability p.

Classification Performance Metrics.

Abbreviation / Symbol	Definition
TP(c)	True Positives at threshold c.
FN(c)	False Negatives at threshold c.
FP(c)	False Positives at threshold c.
TN(c)	True Negatives at threshold c.
S(c)	Sensitivity at threshold c.
E(c)	Specificity at threshold c.
ROC	Receiver Operating Characteristic curve.
AUC	Area Under the ROC Curve.
J(c)	Youden index at threshold c.

Distributions and Score Variables.

Abbreviation / Symbol	Definition
$F(\cdot)$	Cumulative distribution function for negative class.
$G(\cdot)$	Cumulative distribution function for positive class.
Z_0	Classifier score for negative instances.
Z_1	Classifier score for positive instances.
$T_i^{(1)}$	Predicted score for i-th positive sample.
$T_j^{(0)}$	Predicted score for j-th negative sample.

Bayesian Optimization Notation.

Abbreviation / Symbol	Definition
η	Hyperparameter vector (η_0, η_1)
Θ.	Search space for hyperparameters.
$\mathcal{GP}(m, k)$	Gaussian Process with mean function m and covariance function k.
ϵ_i	Observation noise.
D_t	Observed data up to iteration t.
$k_t(\cdot)$	Covariance vector at iteration t.
$\mu_t(\cdot)$	Predictive posterior mean at iteration t.
$\sigma_t^2(\cdot)$	Predictive posterior variance at time t.
$a_t(\cdot)$	Acquisition function used at iteration t
UCB	Upper Confidence Bound
κ	Trade-off parameter in UCB.

Bayesian Statistical Methods.

Abbreviation / Symbol	**Definition**
$\pi_j^{(0)}$	Prior inclusion probability for gene j.
PIP	Posterior Inclusion Probability.
`varbvs`	Variational Bayesian Variable Selection algorithm.
$\sigma(\cdot)$	Sigmoid function
θ	Parameter of interest.
$p(\theta)$	Prior distribution for θ.
$\mathcal{D}$	Observed data.
$p(\mathcal{D} \mid \theta)$	Likelihood function.
$p(\theta \mid \mathcal{D})$	Posterior distribution.
$p(\mathcal{D})$	Marginal likelihood.
γ_j	Binary indicator for inclusion of variable j.
KL	Kullback–Leibler divergence.
$\mathcal{Q}$	Variational distribution family.
$q(\cdot)$	Marginal variational factor.
$\mathcal{G}$	Set of predictors.
$\mathcal{G}_{\text{screen}}$	Screening subset of predictors.
$\mathcal{G}_{\text{promising}}$	Promising predictors.
$\mathcal{G}_{\text{non}}$	Non-promising predictors.
ELBO	Evidence Lower Bound.
CIL	Confounder Importance Learning.

2 Index of Figures

3 Index of Tables

4 Introduction

Cancer is a brutally painful disease and one of the leading worldwide causes of death. While treatment techniques still have a long way to go, biostatistics and bioinformatics are making rapid advancements and lend promise to improve cancer treatment and outcomes. In particular, gene expression technologies have made it possible to capture snapshots of thousands of genes at once, which now enables researchers to study relevant biomarkers within specific cancer types. However, an inherent challenge of working with gene expression data is an enormous amount of gene covariates present in DNA, some of which including more than 70,000 genes, which creates a problem of high dimensionality. At the same time, there is often a limited amount of patient observations for specific cancer types. This creates an environment ripe for overfitting. Modeling techniques such as LASSO, Ridge, Elastic Net, and Bayesian inference are able to work with high dimensional data and narrowing down the thousands of covariates into a few, predictive genes.

In this paper, we will employ various advanced statistical techniques in order to determine the overall status of a patient with liver cancer (likelihood of death or survival) and we will predict the stages of the liver cancer based on the RNA-Seq data. In addition to standard modeling approaches, we will introduce a Bayesian variable selection framework where prior inclusion abilities depend on "external data" from a more common cancer type to assist in the predictive power for a less prominent cancer.

Our analysis focuses on two cancer types: the more common breast cancer and the less common liver cancer. Empirical evidence demonstrates that there is a link between breast cancer and liver metastasis, particularly in specific molecular subtypes [17]. We attempt to use information from the more common breast cancer as prior information on the less common liver cancer. To do so, we modify the penalized regression function to reduce the penalty on genes that are deemed as important in the breast cancer regression. To tune model hyperparameters and assess the model fit, we employ a variety of optimization strategies, such as maximizing the AUC and minimizing the BIC. Hyperparameter tuning is conducted through grid search and Bayesian optimization.

The model framework which incorporates the external information in the final model does reach a higher AUC than the baseline model. This indicates that the idea behind the framework does work, although deeper analysis revealed underlying problems. The selected genes from the breast cancer were highly uncorrelated with the relevant genes from the liver cancer. As an attempt to verify the statistical methodology in light of uninspiring results, we partitioned the liver cancer dataset into two groups: one group if the patient's liver cancer was in Stage I or Stage II, and the other group being if the patient's liver cancer was in stage III or stage IV. This transfer learning approach from within the same cancer type examined whether genes predictive in early-stage liver cancer patients serve predictive power in advanced-stage liver cancer patients.

We then implemented a two-step Bayesian variable selection framework for the liver cancer patients in Stage III/IV. We first applied a Bayesian logistic regression model assuming that all genes are equally likely to be relevant amongst the 5,000 genes with the highest variance. We then assign a higher prior inclusion probability based on the genes that were deemed to be "most promising" from the best model in the previous paragraph (Elastic Net) which assessed the predictive power of genes in stage I/II on cancer progression to stages III/IV.

In the sections that follow, we first review related studies for high-dimensional gene expression data using transfer learning, penalized regression, and Bayesian variable selection. We will then outline the data processing pipeline, statistical methodology and model implementation, and evaluation procedures. Finally, we will discuss results and discuss future implications and research ideas that could stem from our approach.

5 Literature Review

Prior to beginning our analysis, we conducted literature review to assess both the state of biomarker selection research and to gain an understanding of the statistical techniques that have been successful for this type of methodology. As mentioned in the introduction, the high-dimensional nature of gene expression data coupled with low patient sample size creates challenges for model implementation and interpretation.

Biomarker Discovery and Gene Expression Modeling

As mentioned in the introduction, the field of biomarker discovery and gene expression is making rapid advances thanks to RNA-seq technology. RNA-seq uses next-generation sequencing to assess the level of RNA molecules in a biological sample. In effect, this allows biologists to gain a novel understanding of the transcriptome levels in a given sample. *"Gene expression profiling predicts clinical outcome of breast cancer"* by Veer et al. (2002) was one of the first studies that demonstrated that gene expression data could be used to predict cancer outcomes[1]. In this study, they were able to identify a 70-gene signature which classified patients as either having a good or poor prognosis.

To focus on genes that may have some significance, the authors performed a data processing step where they removed genes whose expression levels did not exceed a minimum threshold. For each gene, they computed the correlation coefficient relative to their target variable of development of distant metastasis with 5 years. 231 genes were then selected as having strong association with the disease outcome (correlation coefficient <-0.3 or >0.3). Then, the remaining genes were put through "leave-one-out cross validation" (LOOCV) to evaluate its predictive power on correct classification. The accuracy of the classifier improved until 70 marker genes were utilized. The classifier was then able to correctly predict the actual outcome of disease for 65 of 78 (83%) of the patients.

While the study predated modern penalized regression and Bayesian frameworks, it served as a landmark study that demonstrated how it is possible to use gene expression data to predict disease outcomes. Furthermore, it demonstrated how a prognostic profile can inform treatment procedures.

Bayesian Variable Selection

Next, we will outline how Bayesian frameworks and Bayesian variable selection can assist in high-dimensional biological research. As statistical techniques advanced, the newfound potential to conduct meaningful research in biomarker selection was clear. A landmark study in this space was conducted by Li and Zhang (2010) titled *"Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces With Applications in Genomics"*[21]. Li and Zhang sought to identify a solution for high-dimensional variables selection contexts where there is known structure amongst the covariates. Li and Zhang adopt the Bayesian "spike and slab" approach to variable selection. They implemented a prior based on a graphical model, specifically an Ising-type prior. To compute the posterior probability, they used a coordinate-wise Gibbs Sampling algorithm. The posterior inclusion probability was calculated by multiplying the Ising contribution and a Bayes factor which compared the marginal likelihood of the gene predictor being included or excluded from the model. Li and Zhang then tested their framework on the modeling of transcription factor binding sites in DNA sequences. The graphical Bayesian approach lifted the AUC in every simulation, with the neighboring predictors from the graphical model uncovering relationships that are difficult to observe with more traditional statistical framework.

Transfer Learning and Incorporating External Information

Recent advances in statistical modeling have brought to light the potential for using external data as prior information in data-scarce settings. To improve variable selection within high-dimensional genomic data, we studied the Confounder Importance Learning (CLI) framework utilized by Rossell et al. (2025) in *"Inference for Multiple Treatment Effects Using Confounder Importance Learning"*[25]. Rossell et al. developed a Bayesian variable selection framework where prior inclusion probabilities depend on external data. In this case, the external data measures the difference between the treatment and the covariates. An issue in modern observtional studies stems from choosing too few covariates, which induces omitted variable bias. Conversely, keeping too many covariates renders a very high variance. To counteract this trade-off, Rossell et al. set a data-dependent prior using an empirical Bayes algorithm. The prior inclusion probability of each covariate depends on external treatment-covariation association strength. This prior is functional across multipleand treatment effects, and it performs well in high-dimensional spaces. For our purposes, it is of interest how the treatment-covariate associations can be considered as external information. Ultimately, the CIL outperformed tested model types such as the Double Lasso, ACPME, and standard Bayesian Model Averaging with regards to the root-MSE.

"A penalized regression framework for building polygenic risk models based on summary statistics from genome-wide association studies and incorporating external information" conducted by Chen et al. (2020)[9] is also relevant for our purposes. Chen et al. developed a comprehensive penalized regression framework for fitting polygenic risk score prediction models via large-scale genome-wide association study (GWA) summary statistics. The polygenic risk score prediction models were built by fitting Lasso regression models using GWA summary statistics and linkage disequilibrium based pruning. The Lasso estimates were then iteratively calculated by the coordinate descent algorithm. Then external biological information was incorporated into the model through penalization and parameter tuning. The functional annotations and pleiotropic signals enter the penalty through a given DNA variant-specific threshold. Adding the external data increased R^2 levels in all simulations. In a test case into cutaneous melanoma, the full model lifted the R^2 value from 4.9% to 5.5%. This study demonstrates how penalized regression methods can incorporate external information to boost performance in biological contexts.

6 Data Sources and Preprocessing

In this section, we provide a detailed description of the two datasets under study: liver cancer and breast cancer. We present essential information regarding the types of covariates, metadata, and clinical variables that will be utilized throughout the analysis. The section also explains the integration and preprocessing of the raw data, including the harmonization of gene expression, phenotype, and survival tables into unified datasets. Particular attention is given to ensuring data consistency, such as handling identifier formats and aligning sample structures across data types. Furthermore, we detail the processing of clinical staging variables and define the two primary binary outcomes used for predictive modeling: overall survival status and advanced cancer stage (Stage III/IV). These outcomes are critical for evaluating the prognostic value of genomic and clinical features in cancer patients.
By combining high-dimensional gene expression data with curated clinical information, we aim to build a robust foundation for the downstream tasks of biomarker selection and predictive modeling.

(a) Liver cancer

The first database used for this project was sourced from UCSC Xena, a platform developed by the University of California, Santa Cruz, specializing in genomic research. Our data is from the GDC TCGA Liver Cancer (LIHC) cohort, which was collected by The Cancer Genome Atlas (TCGA) project. It includes data from a sample of 418 individuals, each with 60,746 distinct variables, all related to patients who are currently suffering from or have previously suffered from liver cancer. The recorded data covers various aspects such as gene expression information, survival status, cancer stage, number of days to death and other relevant medical details. All of these data types were processed using official pipelines provided by the Genomic Data Commons (GDC), which helps ensure consistency and reliability across samples. The collected variables can be grouped into several categories:

- **Gene Expression Data:** The gene expression data that we received from the UCSC Xena data platform delivered a matrix with each row corresponding to a specific gene, and each column representing an individual patient sample. Genes are labeled with Ensembl gene identifiers (e.g., `ENSG00000288623.1`), while patients are labeled with unique identifiers (e.g., `TCGA-5R-AA1C-01A`). The gene expression matrix values represent normalized gene expression values, which were generated using RNA-seq and quantified as TPM.

- **Survival Data:** Patient survival data was obtained from a separate table within the UCSC Xena data platform. Each row corresponds to a unique patient, with "OS" pertaining to whether the patient survived, and "OS.time" indicating the duration from the initial diagnosis to the patient death.

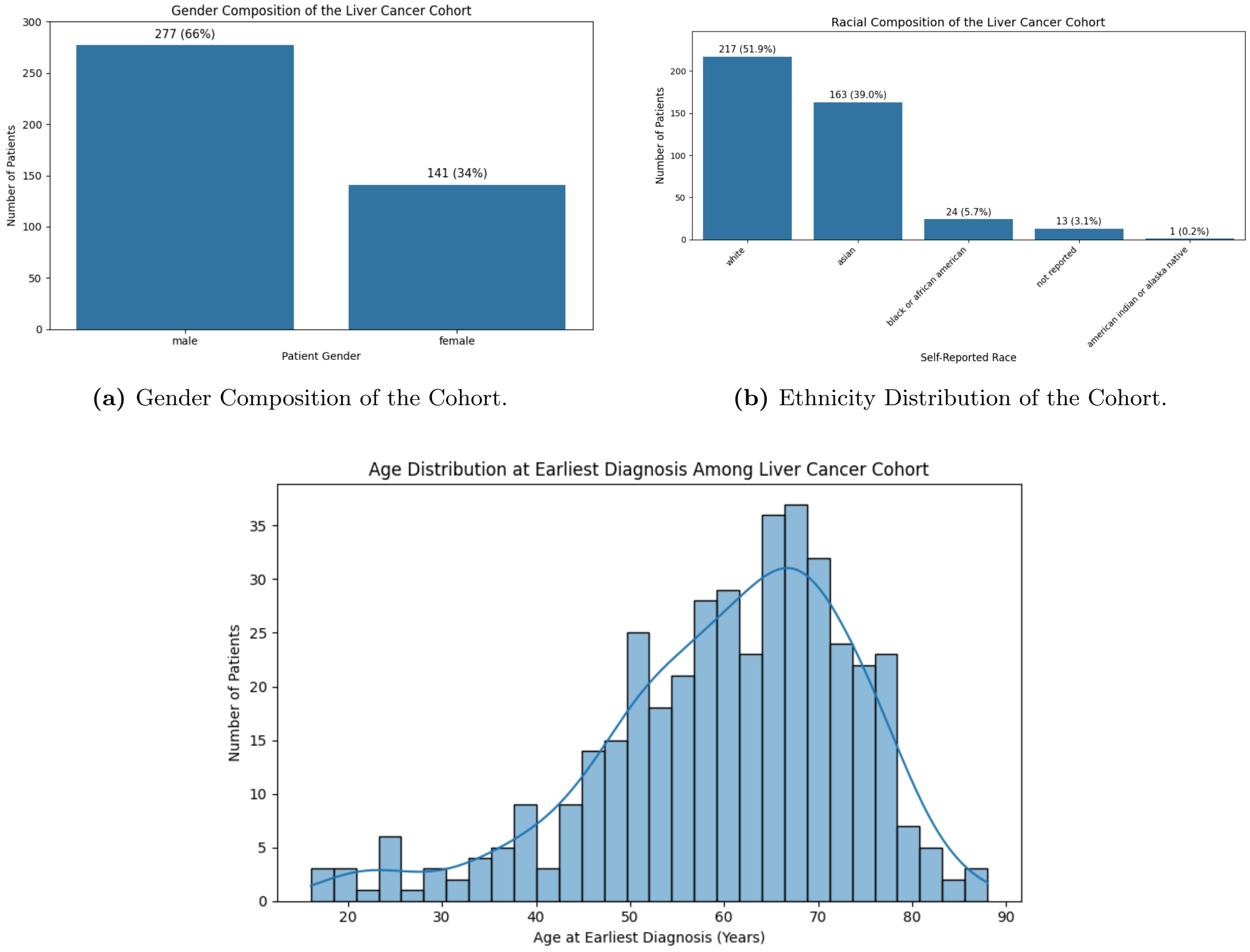
- **Clinical Data:** Patient clinical data is also separately obtained from the UCSC Xena data platform. The phenotype data contains patient-level information such as demographic information, diagnostic information, and treatment history. Key demographic variables include race, gender, ethnicity, age at index (diagnosis). Key diagnostic variables include primary diagnosis, tumor grade, and pathologic stage. Key treatment history variables include treatment type and therapeutic interventions, treatment state, and tumor description.

The liver cancer data features gene expression measurements collected via RNA sequencing. The three data types were merged into a unified dataset using a common sample identifier. Prior to merging, the gene expression matrix was transposed so that each row corresponded to an individual patient, aligning with the structure of the survival and phenotype tables. The final dataset consists of 341 patients and 25,474 features, 25,434 of which are gene expression measurements derived from RNA sequencing, and the remaining are clinical and demographic attributes such as age at diagnosis, cancer stage, and vital status. Gene identifiers (e.g., `ENSG00000141510.15`) were truncated to remove version suffixes (e.g., `.15`), in order to ensure compatibility across datasets and retain only the core Ensembl gene ID (`ENSG00000141510`).

(i) Exploratory Data Analysis of Metadata

In addition to the gene expression data, the dataset also features survival and clinical data for each patient within the liver cancer cohort. This enriched data may provide added context and help guide understanding of the cohort and drive feature selection, modeling, and pre-processing decisions.

Patient Demographics To begin to understand the patient composition of the liver cancer cohort, exploratory data analysis is performed on the patient demographic data.



(a) Gender Composition of the Cohort.

(b) Ethnicity Distribution of the Cohort.

(c) Age at Earliest Diagnosis.

Figure 1: Demographic composition of the liver cancer cohort. These plots summarize patient gender, ethnicity, and age distribution at earliest diagnosis.

As observed in Figure 1a, the majority of the patients are male, with 66% of the cohort participants being identified as males and 34% of the cohort participants being identified as females. This gender imbalance aligns with epidemiological trends for hepatocellular carcinoma, as the National Institute of Health (NIH) states that men are two to three times more likely to develop hepatocellular carcinoma (HCC) than women [24].

Figure 1b demonstrates the racial variation amongst the liver cancer cohort. The predominantly represented racial groups are white (51.9%) and asian (39.0%) represent 90.9% of the cohort. Smaller proportions represent black or african american (5.7%), not reported (3.1%) and american indian or alaska native (0.2%). The racial group imbalance may limit the generalization power of the analysis and should be considered when interpreting results.

Figure 1c displays the age distribution among the liver cancer cohort. The age ranges from approximately 16 to 90 years, with a noticeable right skew in the data. The right skew is consistent with older adults becoming more susceptible to liver cancer than younger individuals. The mean age is approximately 60 years old with a standard deviation of approximately 14 years. The most frequently observed ages were 65 and 66.

Relevant Clinical Information In addition to gene expression data, our liver cancer data also contained clinical data which stated the specific type of liver cancer suffered by each patient. Of the 418 patients within the liver cancer cohort, 88% of the sampled tissues were tumorous, while 12% of the sampled tissue were normal. 9% of the patients had received a prior cancer diagnosis, while liver cancer was the first cancer diagnosis for 91% of the patients. The specific primary diagnoses in the liver cancer cohort are represented in Table 1.

Primary Diagnosis	Number of Diagnoses
Hepatocellular carcinoma, NOS	401
Combined hepatocellular carcinoma and cholangiocarcinoma	8
Hepatocellular carcinoma, clear cell type	4
Hepatocellular carcinoma, fibrolamellar	3
Clear cell adenocarcinoma, NOS	1
Hepatocellular carcinoma, spindle cell variant	1

Table 1: Distribution of primary diagnoses in the liver cancer cohort.

(b) Breast cancer

The breast cancer dataset comes from the same source as the liver cancer dataset: the UCSC Xena platform. Specifically, we used the GDC TCGA Breast Cancer (BRCA) cohort, which is part of The Cancer Genome Atlas (TCGA) project. This dataset includes a wide variety of molecular and clinical data collected from over 1,200 breast cancer patients. The available data types include gene expression measurements obtained via RNA sequencing, along with several other molecular features that align with those found in the liver cancer dataset. Additionally, the dataset provides phenotype information such as patient demographics, cancer stage, and survival outcomes. These features overlap makes the breast cancer data particularly suitable for integration in regression analyses.

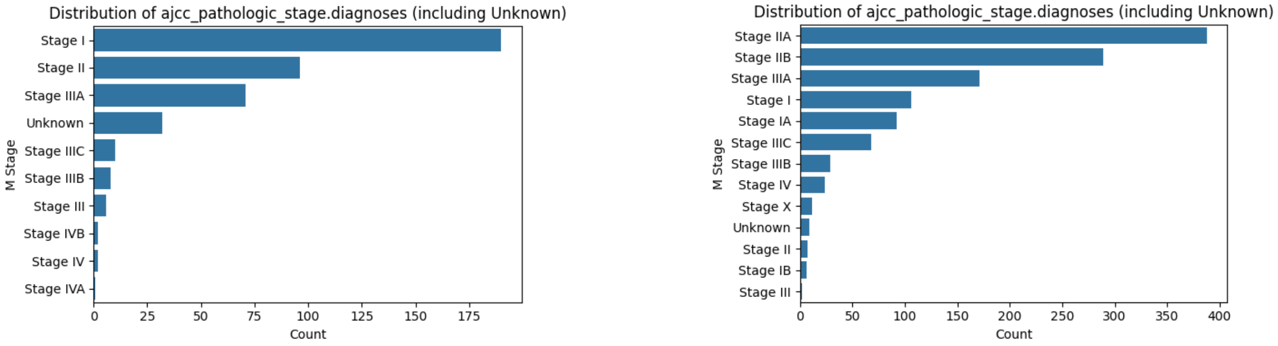
The dataset was obtained again in three separate files: the raw gene expression data that we received from the gene expression data, survival data, and clinical phenotype data. These files were merged into a unified dataset using a common sample identifier. Prior to merging, the gene expression matrix was transposed so that each row corresponded to an individual patient, aligning with the structure of the survival and phenotype tables. The final dataset consists of 1,203 patients and 60,748 features, of which 60,660 are gene expression measurements derived from RNA sequencing, and the remaining are clinical and demographic attributes such as age at diagnosis, cancer stage, and vital status. Gene identifiers (e.g., ENSG00000141510.15) were truncated to remove version suffixes (e.g., .15), in order to ensure compatibility across datasets and retain only the core Ensembl gene ID (ENSG00000141510).

(c) Clinical Stage Processing and Binary Outcome Definition.

As part of the clinical data preprocessing, we reviewed and cleaned cancer stage information provided under the AJCC pathologic staging system, which includes tumor size (T), lymph node involvement (N), and metastasis (M) components.

We inferred the stage for patients with missing entries in the pathologic stage variable (`ajcc\_pathologic\_stage.diagnoses`) by combining the available T, N, and M values: for example, if metastasis (M) was reported as `M1`, the stage was set to `Stage IV`, whereas combinations like `T1` and `N0` were classified as `Stage I`. Patients with incomplete or ambiguous information were assigned as Unknown and later on dropped.

Figure 2 shows the distribution of inferred and observed AJCC pathologic stages across both cancer types.



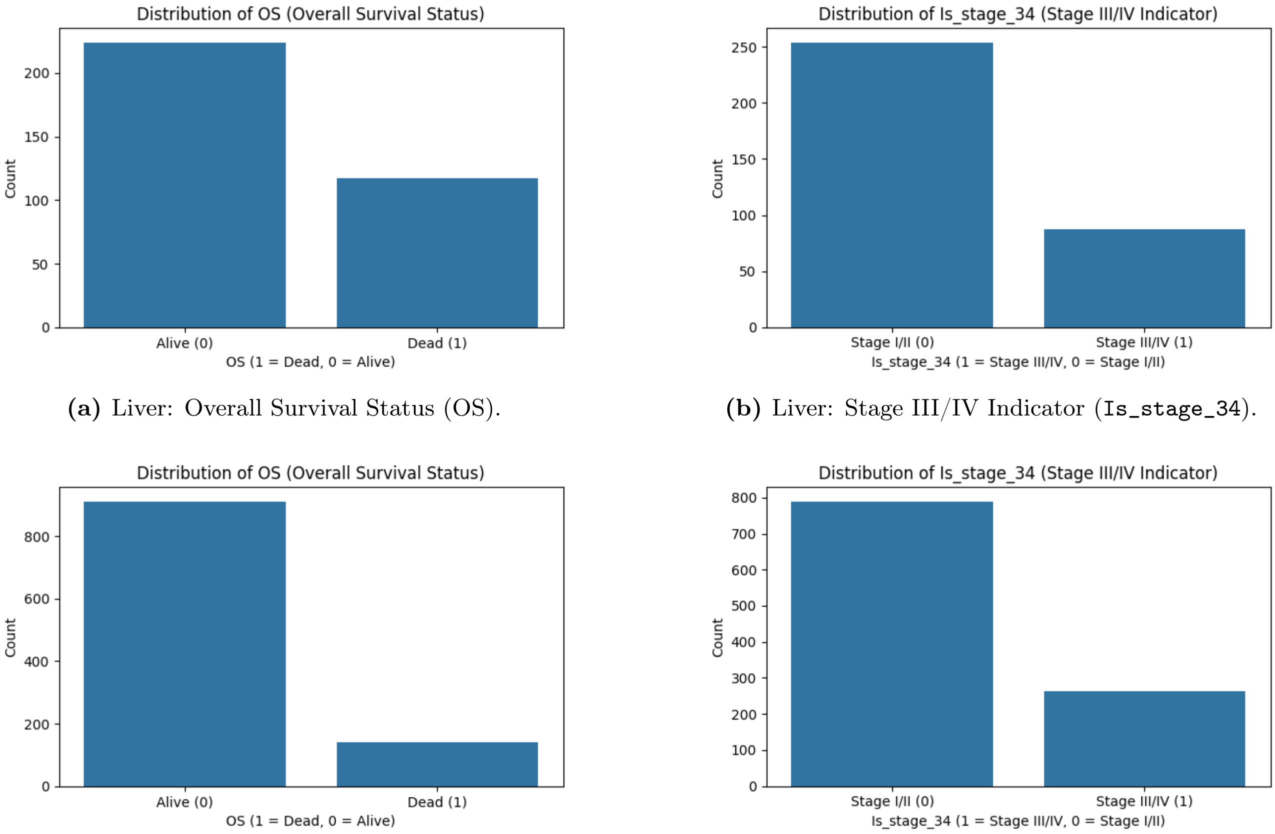
Liver dataset **Breast dataset**

Figure 2: Distribution of `ajcc_pathologic_stage.diagnoses` (including inferred values) for the liver and breast datasets.

Following this, we grouped the T, N, and M categories into simplified versions (e.g., collapsing `T1a`, `T1b`, etc., into a single `T1` group). Patients with a recorded stage of `Stage X` were excluded from the analysis. This label is used when the stage could not be assessed, often due to incomplete diagnostic information.

We transformed the cancer stage variable into a binary outcome to enable classification of patients based on disease severity. Early-stage diagnoses (Stage I, IA, IB, II, IIA, IIB) were grouped and labeled as 0, while more advanced stages (Stage III, IIIB, IIIC, IV) were labeled as 1. This new binary variable, `Is_stage_34`, serves as one of the two target variables (y) for the regression tasks based on genomic features.

In addition to `Is_stage_34`, we also use `OS`, a binary indicator of overall survival status. In this variable, 0 corresponds to patients who were alive at last follow-up and 1 indicates those who were deceased. Together, these two outcome variables enable us to evaluate the predictive value of gene expression data in relation to both disease progression and survival. Figure 3 illustrates the class distributions for both variables.



(a) Liver: Overall Survival Status (OS).

(b) Liver: Stage III/IV Indicator (`Is_stage_34`).

(c) Breast: Overall Survival Status (OS).

(d) Breast: Stage III/IV Indicator (`Is_stage_34`).

Figure 3: Distributions of the two binary outcome variables (`OS` and `Is_stage_34`) for the liver (top row) and breast (bottom row) datasets.

As shown in Figure 3, both outcome variables exhibit class imbalance across both cancer types. The `OS` variable, which captures overall survival status, shows a strong skew: the majority of breast cancer patients were alive at the last follow-up, while a smaller portion were deceased. While there is a smaller imbalance for liver cancer patients, the majority of patients were also alive at the last follow-up. Similarly, the `Is_stage_34` variable shows that most liver and breast cancer patients were diagnosed with early-stage cancer (Stage I/II), whereas only a few had more advanced disease (Stage III/IV).

This imbalance reflects real-world clinical distributions across both cancer types, where early detection is more common, and most patients survive during the study period. However, it may also pose challenges for predictive modeling, as models trained on imbalanced data can become biased toward the majority class.

7 Methodological Framework

To effectively identify genomic biomarkers capable of predicting cancer outcomes, it is essential to employ robust statistical models that can accommodate the complexity and high dimensionality inherent in gene expression data. Specifically, our dataset comprises tens of thousands of potential gene predictors, whereas the number of patient samples is relatively small. In such high-dimensional settings, classical logistic regression is prone to overfitting and numerical instability.

To address these challenges, we adopt penalized likelihood methods, which extend logistic regression by incorporating regularization terms that mitigate overfitting, improve generalization, and, in some cases, enable automatic variable selection. This section outlines the theoretical framework of penalized logistic regression, focusing on the Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, and Ridge Regression. While LASSO and Elastic Net perform variable selection by shrinking some coefficients exactly to zero, Ridge regression induces coefficient shrinkage without performing variable selection. These methods control model complexity while preserving interpretability, a crucial property in clinical applications.

To evaluate model performance, we rely on the area under the receiver operating characteristic curve (AUC-ROC) as our primary metric. This section provides a mathematically grounded explanation of AUC and its relevance in binary classification problems. To further enhance model interpretability and reliability, all predictive models were trained and evaluated using k-fold cross-validation, the details of which are discussed in subsequent sections.

Given our overarching objective, to investigate whether breast cancer gene expression data can be integrated into liver cancer analyses to improve interpretability and predictive performance, we also explore data integration strategies. These strategies were implemented using established software tools and are described in detail later. As an alternative to cross-validation, Bayesian Information Criterion (BIC) was employed to assess hyperparameter relevance. In conjunction with these criteria, Bayesian optimization was used to efficiently search for optimal hyperparameter configurations.

Finally, we introduce the foundational concepts of Bayesian statistical inference. In particular, we describe the use of the spike-and-slab prior for variable selection, and how variational inference was employed to make computation tractable in high dimensions. Our methodology and implementation details are discussed, with additional theoretical background provided in the appendix.

(a) Penalized likelihood methods for logistic regression

We begin by introducing the traditional logistic regression model and examining its limitations in high-dimensional contexts. In settings where the number of predictors substantially exceeds the number of observations, such as in gene expression data, classical logistic regression tends to exhibit instability and severe overfitting. To address these shortcomings, penalized likelihood methods have been proposed. Among the most prominent are LASSO, Ridge regression, and Elastic Net [15].

Mathematically, these techniques are formulated as optimization problems in which the log-likelihood is penalized by a norm-based constraint. The penalty term can be chosen as the ℓ_1-norm for LASSO, the ℓ_2-norm for Ridge regression, or a convex combination of both in the Elastic Net. Furthermore, each of these methods admits adaptive extensions, where data-driven weights are assigned to the penalty terms. These adaptive variants aim to improve variable selection consistency and model accuracy by allowing differential shrinkage across predictors.

In the sections that follow, we provide a detailed theoretical account of these penalized methods, including their adaptive formulations, and describe their practical implementation in the context of genomic classification.

Let $S = \{1, \ldots, n\}$ represent a sample of n units drawn from a finite population, where each unit is characterized by a vector of p covariates, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$, for $i = 1, 2, \ldots, n$. Along with the covariate information, the corresponding values of the response variable Y are available for each unit, denoted as y_i for $i \in S$. In the setting of this biomedical research, Y represents a binary outcome (such as overall survival), and under the logistic regression model, the log-odds of a positive outcome is specified as a linear function of $\mathbf{x}_i$:

$$\text{logit}(p(\mathbf{x}_i)) = \ln\left(\frac{p(\mathbf{x}_i)}{1-p(\mathbf{x}_i)}\right) = \mathbf{x}_i^T\boldsymbol{\beta}. \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients. Equivalently, the conditional probability of a positive outcome for the covariate vector $\mathbf{x}_i$, $p(\mathbf{x}_i) = P(Y = 1 \mid \mathbf{x}_i)$, is given by:

$$p(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T\boldsymbol{\beta}}}. \tag{2}$$

which ensures $p(\mathbf{x}_i) \in (0, 1)$. The maximum likelihood parameter estimator $\hat{\boldsymbol{\beta}}$ is obtained by maximizing the log-likelihood function $\ell(\boldsymbol{\beta})$, which is defined by:

$$\ell(\boldsymbol{\beta}) = \sum_{i\in S} \left[y_i \ln(p(\mathbf{x}_i)) + (1 - y_i)\ln(1 - p(\mathbf{x}_i))\right]. \tag{3}$$

In high-dimensional scenarios, particularly when the number of covariates approaches or exceeds the sample size, the classical maximum likelihood estimator often exhibits high variance and instability. To address this issue, penalized likelihood methods estimate the parameters by minimizing a penalized loss function, defined as the negative log-likelihood augmented with a regularization term:

$$\ell_p(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta}) + \lambda\,\Omega(\boldsymbol{\beta}), \tag{4}$$

where $\lambda > 0$ is a tuning parameter controlling the degree of penalization and $\Omega(\boldsymbol{\beta})$ is a penalty function that imposes structure on the coefficient estimates. By accepting a small amount of bias, these methods reduce variance, resulting in more stable estimates and improved out-of-sample performance. When $\Omega(\boldsymbol{\beta})$ promotes sparsity, the resulting model also gains interpretability, aligning clearly with the primary objective of this research. We next present a concise exposition of the penalized likelihood models that will be utilized throughout this analysis.

(i) LASSO-Regularized Logistic Regression

A widely used penalty for encouraging sparsity is the ℓ_1 norm, leading to the Least Absolute Shrinkage and Selection Operator (LASSO). In the context of logistic regression, LASSO estimates the coefficients by solving the following optimization problem:

$$\hat{\boldsymbol{\beta}}^{\text{L}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{5}$$

where λ is a regularization parameter that governs the strength of the penalty. Larger values of λ shrink more coefficients toward zero, potentially setting some exactly to zero and thus performing variable selection.

(ii) Ridge-Penalized Logistic Regression

As stated above, LASSO regression can set some coefficient estimates exactly to zero. In contrast, Ridge regression covariate estimates are non-zero with probability one. In doing so, Ridge regression uses the ℓ_2 norm and does not force any β_j^R exactly to zero. For some given $\lambda > 0$, Ridge regression solves the following optimization problem:

$$\hat{\boldsymbol{\beta}}^{R} = \arg\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{6}$$

Unlike the LASSO, Ridge regression does not set any coefficients exactly equal to zero and therefore does not perform variable selection. Instead, coefficient values can be shrunk very close to zero, thus still retaining the coefficients in the parameter vector $\hat{\boldsymbol{\beta}}$.

(iii) Elastic Net Regularization

While the LASSO is effective for producing sparse models, it tends to perform poorly when predictors are highly correlated, as it arbitrarily selects one among them [31]. The Elastic Net addresses this limitation by combining the ℓ_1 penalty of the LASSO with the ℓ_2 penalty of Ridge regression. This hybrid approach encourages both sparsity and grouping of correlated variables. For some given $\lambda > 0$, the Elastic Net estimator solves the following optimization problem:

$$\hat{\boldsymbol{\beta}}^{\text{EN}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \right\}, \tag{7}$$

where $\alpha \in [0,1]$ controls the relative balance between the LASSO and ℓ_2 penalties.

(iv) Adaptive LASSO Regularization

To improve upon the variable selection properties of the previous models, the Adaptive LASSO introduces coefficient-specific penalties based on the importance of each variable. As stated before, let y_i be our outcome variable and $\mathbf{x}_i \in \mathbb{R}^p$ the covariates for individuals $i = 1, \ldots, n$.

Let us consider $\hat{\boldsymbol{\beta}}^{(0)} \in \mathbb{R}^p$ as an initial parameter estimate (for instance, the estimate given by LASSO, Ridge or the Elastic Net). For some given $\lambda > 0$, the adaptive LASSO problem is defined as follows:

$$\hat{\boldsymbol{\beta}}^{\text{AL}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda \frac{|\beta_j|}{|\hat{\beta}_j^{(0)}|} \right\}, \tag{8}$$

To illustrate this advantage more formally, we present the following proposition.

Proposition 1. The solution to the adaptive LASSO problem can be written as a standard LASSO problem using only the covariates with non-zero coefficients from the initial estimates, with a common regularization parameter λ and a modified design matrix $\tilde{X}$.

Proof. In the adaptive LASSO formulation, covariate-specific penalty weights are introduced based on an initial estimate. Notably, for any covariate j such that $|\hat{\beta}_j^{(0)}| = 0$, the corresponding estimate $\hat{\beta}_j$ will necessarily be zero in the final solution. This is because the penalty term becomes infinite, effectively excluding the covariate from the model. As a result, we restrict our attention to covariates with non-zero initial coefficients. For these, we define the adjusted coefficients as:

$$\tilde{\beta}_j \;=\; \frac{\beta_j}{|\hat{\beta}_j^{(0)}|}, \qquad \tilde{x}_{ij} \;=\; |\hat{\beta}_j^{(0)}|\, x_{ij}. \tag{9}$$

Let $J = \{ j : |\hat{\beta}_j^{(0)}| > 0 \}$. The adaptive-LASSO logistic-regression objective on the full parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ is:

$$\min_{\boldsymbol{\beta}} \Big\{ -\ell(\boldsymbol{\beta}) \;+\; \lambda \sum_{j \in J} \frac{|\beta_j|}{|\hat{\beta}_j^{(0)}|} \Big\}, \tag{10}$$

where:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big[y_i \ln p(\mathbf{x}_i) + (1-y_i) \ln\big(1 - p(\mathbf{x}_i)\big) \Big], \quad p(\mathbf{x}_i) = \frac{\exp\big(\mathbf{x}_i^T \boldsymbol{\beta}\big)}{1 + \exp\big(\mathbf{x}_i^T \boldsymbol{\beta}\big)}. \tag{11}$$

Since $\beta_j = |\hat{\beta}_j^{(0)}|\, \tilde{\beta}_j$ and $x_{ij} = \tilde{x}_{ij} / |\hat{\beta}_j^{(0)}|$, the linear predictor transforms as:

$$\mathbf{x}_i^T \boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\, \beta_j = \sum_{j \in J} x_{ij} \left(|\hat{\beta}_j^{(0)}|\, \tilde{\beta}_j \right) = \sum_{j \in J} \left(|\hat{\beta}_j^{(0)}|\, x_{ij} \right) \tilde{\beta}_j = \sum_{j \in J} \tilde{x}_{ij}\, \tilde{\beta}_j. \tag{12}$$

Hence the logistic log-likelihood becomes:

$$-\ell(\boldsymbol{\beta}) = -\sum_{i=1}^{n}\Big[y_i \ln \tilde{p}(\tilde{\mathbf{x}}_i) + (1-y_i)\ln\big(1-\tilde{p}(\tilde{\mathbf{x}}_i)\big)\Big], \quad \tilde{p}(\tilde{\mathbf{x}}_i) = \frac{\exp\big(\tilde{\mathbf{x}}_i^T\,\tilde{\boldsymbol{\beta}}\big)}{1+\exp\big(\tilde{\mathbf{x}}_i^T\,\tilde{\boldsymbol{\beta}}\big)}, \tag{13}$$

where $\tilde{\mathbf{x}}_i = (\tilde{x}_{ij})_{j\in J}$. Meanwhile, the penalty term becomes:

$$\sum_{j\in J}\frac{|\beta_j|}{|\hat{\beta}_j^{(0)}|} = \sum_{j\in J}|\tilde{\beta}_j|. \tag{14}$$

Therefore, the adaptive-LASSO logistic objective:

$$\min_{\boldsymbol{\beta}}\Big\{-\ell(\boldsymbol{\beta}) + \lambda\sum_{j\in J}\frac{|\beta_j|}{|\hat{\beta}_j^{(0)}|}\Big\} \tag{15}$$

is equivalent to:

$$\min_{\tilde{\boldsymbol{\beta}}}\Big\{-\sum_{i=1}^{n}\Big[y_i \ln \tilde{p}(\tilde{\mathbf{x}}_i) + (1-y_i)\ln\big(1-\tilde{p}(\tilde{\mathbf{x}}_i)\big)\Big] \;+\; \lambda\sum_{j\in J}|\tilde{\beta}_j|\Big\}, \tag{16}$$

which is exactly the standard LASSO-penalized logistic-regression problem on the reduced design matrix $\tilde{X} = (\tilde{x}_{ij})_{i=1,\dots,n;\, j\in J}$. □

As mentioned earlier, the primary advantage of the adaptive LASSO is its ability to apply coefficient-specific penalties, which are based on the estimated importance of each covariate. By assigning smaller penalties to coefficients that are already deemed more influential by the initial LASSO or Elastic Net estimate, adaptive LASSO reduces the bias typically associated with the previous methods.

The penalty applied to each coefficient β_j is inversely proportional to the magnitude of its initial estimate. If a covariate j has little or no effect on the outcome, we expect $|\hat{\beta}_j^{(0)}|$ to be small, and the resulting larger penalty encourages its exclusion from the model. Conversely, for covariates with a strong effect, the smaller penalty helps retain their influence in the final model, thus reducing bias in the corresponding estimates $\hat{\beta}_j$.

While the adaptive LASSO does introduce a slightly higher computational cost due to the need to solve a second optimization problem, this cost is typically small. This is because the second step only involves the covariates that had non-zero coefficients in the initial estimate, thereby reducing the dimensionality and making the computation more efficient.

(b) Assessment of Classifier Performance: AUC–ROC

As widely recommended in biostatistics and bioinformatics literature [7, 19, 22], the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) will be used as the primary performance evaluation metric for this analysis. The ROC AUC has been extensively employed to assess the discriminative ability of predictive models, particularly in medical and clinical data applications, such as disease diagnosis and prognosis. To compute the ROC AUC, we must first consider the underlying classification model, where the outcome variable follows a Bernoulli distribution.

In the context of logistic regression, the model estimates the conditional probability $T = P(Y = 1 \mid X)$. Since this probability is continuous, it must be dichotomized into a binary outcome. This is achieved by applying a classification threshold. For a given threshold $c \in [0, 1]$, the classification rule is as follows:

$$\hat{y}(c) = \begin{cases} 1, & T > c, \\ 0, & T \leq c. \end{cases} \tag{17}$$

Let D be the true class label indicator variable, where $D = 1$ indicates membership in the positive class and $D = 0$ indicates membership in the negative class, and let Z_0 and Z_1 be absolutely continuous random variables representing the classifier scores for the negative and positive classes, respectively, with cumulative distribution functions F and G:

$$F(t) = P\big(T \leq t \mid D = 0\big), \qquad G(t) = P\big(T \leq t \mid D = 1\big). \tag{18}$$

In this context, the confusion matrix is defined as follows:

Obs\Pred	$\hat{y} = 1$	$\hat{y} = 0$
$D = 1$	$\mathrm{TP}(c)$ $\big(\text{True Positive}\big)$: $P(T > c \mid D = 1)$	$\mathrm{FN}(c)$ $\big(\text{False Negative}\big)$: $P(T \leq c \mid D = 1)$
$D = 0$	$\mathrm{FP}(c)$ $\big(\text{False Positive}\big)$: $P(T > c \mid D = 0)$	$\mathrm{TN}(c)$ $\big(\text{True Negative}\big)$: $P(T \leq c \mid D = 0)$

Table 2: Confusion matrix for binary classification at threshold c.

From this confusion matrix, two important performance metrics can be derived:

- The sensitivity, also called the True Positive Rate, is the probability that a positive instance is correctly classified as positive:

$$\mathrm{S}(c) = P\big(\hat{y} = 1 \mid D = 1\big) = P\big(T > c \mid D = 1\big) = 1 - G(c) = \frac{\mathrm{TP}(c)}{\mathrm{TP}(c) + \mathrm{FN}(c)}. \tag{19}$$

- The specificity, also called the True Negative Rate, is the probability that a negative instance is correctly classified as negative:

$$\mathrm{E}(c) = P\big(\hat{y} = 0 \mid D = 0\big) = P\big(T \leq c \mid D = 0\big) = F(c) = \frac{\mathrm{TN}(c)}{\mathrm{TN}(c) + \mathrm{FP}(c)}. \tag{20}$$

Based on these metrics, we can define the Receiver Operating Characteristic (ROC) curve.

Definition 1. Let F and G denote the cumulative distribution functions of the classifier scores for the negative and positive classes, respectively. For a given threshold c, let $S(c)$ denote the sensitivity and $E(c)$ denote the specificity. Then, the ROC curve is defined as:

$$\begin{aligned} \mathrm{ROC} : [0,1] &\longrightarrow [0,1] \times [0,1], \\ c &\longmapsto \big(1 - F(c),\, 1 - G(c)\big) \;=\; \big(1 - \mathrm{E}(c),\, \mathrm{S}(c)\big). \end{aligned} \tag{21}$$

To compute the area under the curve and facilitate comparisons, it is often more convenient to reparameterize the ROC by its false-positive rate $t \in [0,1]$, yielding a mapping from the unit interval into itself (Check Proposition 3 in the Appendix). This yields to the ROC function which is defined as:

$$\begin{aligned} \mathrm{ROC_f} : [0,1] &\longrightarrow [0,1], \\ t &\longmapsto \mathrm{ROC_f}(t) = 1 - G\big(F^{-1}(1-t)\big). \end{aligned} \tag{22}$$

In practice, the ROC curve is estimated using the empirical cumulative distribution functions of the model scores. Let n_1 and n_0 denote the number of positive ($D = 1$) and negative ($D = 0$) instances, respectively. Let $\{T_i^{(1)}\}_{i=1}^{n_1}$ and $\{T_j^{(0)}\}_{j=1}^{n_0}$ be the predicted scores for the positive and negative classes. The corresponding empirical distribution functions are defined as:

$$\widehat{G}(z) = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbf{1}\{T_i^{(1)} \leq z\}\,, \quad \widehat{F}(z) = \frac{1}{n_0}\sum_{j=1}^{n_0} \mathbf{1}\{T_j^{(0)} \leq z\}\,. \tag{23}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Then the empirical ROC curve is given by:

$$\widehat{\text{ROC}} \;=\; \big\{\big(1-\widehat{F}(c),\, 1-\widehat{G}(c)\big)\; : c \in [0,1]\big\} \;=\; \big\{\big(1-\widehat{E}(c),\, \widehat{S}(c)\big)\; : c \in [0,1]\big\}. \tag{24}$$

Applying the same reparametrization as in Proposition 3 (check the Appendix), the empirical ROC curve can be defined as:

$$\widehat{\text{ROC}} \;=\; \big\{\,(t,\; \widehat{\text{ROC}_\text{f}}(t)) : t \in [0,1]\big\} \;=\; \big\{\,(t,\; 1-\widehat{G}(\widehat{F}^{-1}(1-t))) : t \in [0,1]\big\}. \tag{25}$$

Observation 1. To avoid notation confusion and ensure conceptual clarity, note that the ROC function ROC_f maps a false positive rate t to the corresponding true positive rate $\text{TP}(c)$, where c is the threshold such that $\text{FP}(c) = t$. However, the notation ROC typically refers to the ROC curve itself, that is, the set of all points on the graph:

$$\big\{\, \big(t,\, \text{ROC}_f(t)\big) \;:\; t \in [0,1]\,\big\} \;\subset\; [0,1] \times [0,1]. \tag{26}$$

Once the ROC curve has been defined, it is useful to summarize its shape with a single scalar: the area under the curve. The AUC quantifies a binary classifier's discriminative ability across all decision thresholds. An AUC of 1 indicates perfect class separation, while an AUC of 0.5 corresponds to random performance.

Definition 2. Let $\text{ROC}_\text{f}(t)$ denote the receiver operating characteristic function at point t. The area under the ROC curve (AUC) is defined as the integral of the ROC function over the unit interval $[0, 1]$:

$$\text{AUC} \;=\; \int_0^1 \text{ROC}_\text{f}(t)\,\mathrm{d}t. \tag{27}$$

In practice, the AUC is often estimated using the Mann–Whitney statistic, which is equivalent to computing the probability that a randomly selected positive instance receives a higher score than a randomly selected negative instance [26].

$$\widehat{\text{AUC}} = \frac{1}{n_1 n_0}\sum_{i=1}^{n_1}\sum_{j=1}^{n_0} \mathbf{1}\left\{T_i^{(1)} > T_j^{(0)}\right\} + \frac{1}{2}\mathbf{1}\left\{T_i^{(1)} = T_j^{(0)}\right\}. \tag{28}$$

This is the form computed by the `pROC` package in R when using the default `roc()` function [27]. The estimated AUC quantifies the overall discriminative ability of the model across all possible decision thresholds. In practice, however, a binary decision must be made at a specific threshold. To select such a threshold, one often seeks to optimize classification performance using a summary statistic. A widely used criterion is the Youden index, which balances sensitivity and specificity.

$$\widehat{J(c)} = \;\widehat{\text{S}(c)} + \widehat{\text{E}(c)} \;-\; 1, \tag{29}$$

Hence, the optimal threshold c^* is chosen by maximizing the empirical Youden index over a sequence of thresholds $\{c_t\}_{t\geq 0}$.

$$c^* \;=\; \arg\max_{c_t}\; \widehat{J(c_t)}. \tag{30}$$

Values closer to zero reflect minimal class separation, whereas values approaching one indicate near perfect discrimination.

(c) Implementation: Cross-Validation for hyperparameter selection

As explained previously, various penalization methods are considered in this study. The utilized penalization methods include LASSO, adaptive LASSO initialized with previous LASSO estimates, Elastic Net and adaptive LASSO initialized with previous Elastic Net estimates. To identify the optimal value of the hyperparameters (λ and α), K-fold cross-validation is employed. In our analysis, the dataset is randomly partitioned into $K = 10$ equally sized folds. For each hyperparameter candidate within a predefined grid, the procedure detailed in Algorithm 1 is executed. The values of the hyperparameters that maximize the estimated AUC are selected as optimal. Subsequently, we re-fit the model on the full training set using these selected values; we then generate predictions on the test set and compute the corresponding AUC.

Observation 2. From this point forward, whenever a train–test split strategy is employed, the training set will comprise 80% of the observations and the test set the remaining 20%.

Algorithm 1 K-Fold Cross-Validation for Elastic Net Logistic Regression.

```
Input: Dataset $S$, training set $S_{\text{train}}$, test set $S_{\text{test}}$, number of folds $K$, grid of regularization parameters $\{\lambda_1, \dots, \lambda_L\}$, grid of mixing parameters $\{\alpha_1, \dots, \alpha_M\}$.
for $k = 1$ to $K$ do
    Split $S_{\text{train}}$ into training subset $S_{\text{train}}^{(k)}$ and validation subset $S_{\text{val}}^{(k)}$
    for each $\alpha_m$ in $\{\alpha_1, \dots, \alpha_M\}$ do
        for each $\lambda_\ell$ in $\{\lambda_1, \dots, \lambda_L\}$ do
            Fit Elastic Net logistic regression on $S_{\text{train}}^{(k)}$:
                $$\hat{\boldsymbol{\beta}}^{\text{EN}(k)} = \arg\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) \;+\; \lambda_\ell \left[ \alpha_m \sum_{j=1}^{p} |\beta_j| \;+\; (1-\alpha_m) \sum_{j=1}^{p} \beta_j^2 \right] \right\}.$$
            Predict probabilities on $S_{\text{val}}^{(k)}$: $\hat{p}_i = \frac{\exp\left(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}^{EN(k)}}\right)}{1+\exp\left(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}^{EN(k)}}\right)}$
            Compute validation $\widehat{AUC}$
        end for
    end for
end for
Compute average $\widehat{AUC}$ for each $(\alpha_m, \lambda_\ell)$: $\widehat{\text{AUC}}^{\text{CV}}(\alpha_m, \lambda_\ell) = \frac{1}{K} \sum_{k=1}^{K} \widehat{\text{AUC}}^{(k)}(\alpha_m, \lambda_\ell)$
Select $(\alpha^*, \lambda^*) = \arg\max_{\substack{m=1,\dots,M \\ \ell=1,\dots,L}} \widehat{\text{AUC}}^{\text{CV}}(\alpha_m, \lambda_\ell)$
Refit Elastic Net logistic regression on full $S_{\text{train}}$ using $(\alpha^*, \lambda^*)$
Predict on $S_{\text{test}}$ and compute test $\widehat{AUC}$:
Return: Optimal hyperparameters: $(\alpha^*, \lambda^*)$ and final estimates $\widehat{\boldsymbol{\beta}^{EN}}$
```

Observation 3. LASSO and Ridge regression are special cases of the Elastic Net, corresponding to $\alpha = 1$ and $\alpha = 0$, respectively. Thus, if either method is chosen a priori, one may fix α accordingly for all folds $m = 1, \dots, M$, and cross-validation over α becomes unnecessary.

(d) Data integration

As discussed in the Literature Review, recent studies such as *"Inference for Multiple Treatment Effects Using Confounder Importance Learning"*[25] and *"A penalized regression framework for building polygenic risk models based on summary statistics from genome-wide association studies and incorporating external information"*[9] have revealed the benefits of data integration and transfer learning procedures. In tandem, they inherently audit data quality and covariate strength, add context, reduce overfitting, and allow for cross-domain research.

Once the penalized likelihood models and the associated performance metric have been rigorously introduced, we proceed to the central objective of this study: to investigate whether breast cancer gene expression data can be leveraged to improve predictive performance in the context of liver cancer. This section outlines the methodological framework adopted to address this question, detailing the integration strategy and the procedures implemented for model training, evaluation, and interpretation.

To integrate the breast and liver cancer datasets, multiple factors must be considered. One key aspect is the trade-off between predictive performance, measured by the AUC, and the number of covariates shared between the two datasets. In addition to evaluating model performance, we also take into account the number of shared covariates, as this enhances interpretability and potential biological relevance.

To facilitate this integration, we define an auxiliary variable z_j to represent metacovariate information. For instance, for each gene j in the liver dataset, we define a binary auxiliary variable as follows:

$$z_j = \begin{cases} 1, & \text{if } \widehat{\beta}_j^{(\text{breast})} \neq 0, \\ 0, & \text{if } \widehat{\beta}_j^{(\text{breast})} = 0. \end{cases} \tag{31}$$

The rationale is that if $z_j = 1$, gene j was deemed important in the breast data by the best model selected. Consequently, when fitting the model on the liver data, we expect that the model will assign a lower penalization to that gene, thereby increasing its chances of being selected again. In other words, we expect that the regularization will favor retaining that gene rather than forcing its inclusion.

A third important consideration is the correlation between gene effects in the two datasets. For integration to be meaningful, genes with large estimated effects in the breast dataset should also tend to have large effects in the liver dataset. One way to assess this is by creating a scatterplot of estimated coefficients from the breast dataset versus those from the liver dataset, excluding genes with zero coefficients in both cases, and checking for a positive correlation. A visible trend would support the assumption that the datasets share informative features. In the case where Ridge regression is used as the modeling approach, an additional challenge arises: since Ridge does not shrink any coefficients to exactly zero, we must establish a principled method to filter out uninformative genes. To address this, we propose categorizing the genes into different groups based on the magnitude of their estimated coefficients and setting an appropriate threshold to discard covariates with negligible effect sizes.

To integrate prior knowledge from the breast cancer dataset, we assign gene-specific penalties in the Lasso model for the liver dataset. Let $z_j \in \{0, 1\}$ indicate whether gene j was selected in the breast model. We define gene-specific penalties as:

$$\log \lambda_j = \eta_0 + \eta_1 z_j, \qquad \lambda_j = \exp\big(\eta_0 + \eta_1 z_j\big). \tag{32}$$

Observation 4. From this point forward, the LASSO model that applies gene-specific penalties in the liver dataset will be referred to as the multi-penalty LASSO.

Genes for which $z_j = 0$ incur a penalty of $\exp(\eta_0)$, whereas those with $z_j = 1$ are intended to receive a smaller penalty of $\exp(\eta_0 + \eta_1)$, thereby increasing their likelihood of being selected in the liver model. To ensure that the penalty is indeed lower when $z_j = 1$, the parameter η_1 must satisfy $\eta_1 < 0$, which guarantees that $\exp(\eta_0 + \eta_1) < \exp(\eta_0)$. With this idea in mind, we then fit the multi-penalty LASSO model with the specific λ_j, compute predicted probabilities and evaluate the average AUC across k-fold cross-validation splits. Using the optimal hyperparameters $(\hat{\eta}_0, \hat{\eta}_1)$, we refit the multi-penalty LASSO on the test liver dataset to estimate the final coefficients. This is done using the optimized penalties $\hat{\lambda}_j = \exp(\hat{\eta}_0 + \hat{\eta}_1 z_j)$ by solving the penalized optimization problem:

$$\widehat{\boldsymbol{\beta}}(\eta_0, \eta_1) = \arg\min_{\boldsymbol{\beta}} \left\{ -\ell\big(\boldsymbol{\beta}; X^{(\text{liver})}, y^{(\text{liver})}\big) + \sum_{j=1}^{p} \lambda_j \, |\beta_j| \right\}. \tag{33}$$

To implement this using standard LASSO software, we reparameterize following a similar approach to that used in Adaptive LASSO in Proposition 1. To formalize this idea more rigorously, we present the following proposition.

Proposition 2. Let $X \in \mathbb{R}^{n \times p}$ be the design matrix and $y \in \mathbb{R}^n$ the response vector from the liver dataset. Suppose we assign gene-specific penalties $\lambda_j = \exp(\eta_0 + \eta_1 z_j)$, where $z_j \in \{0, 1\}$ indicates whether gene j was selected in a prior breast cancer model. Then the following penalized likelihood problem:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ -\ell(\beta; X, y) + \sum_{j=1}^{p} \lambda_j |\beta_j| \right\}, \tag{34}$$

can be reparameterized as a standard LASSO problem via a transformation of the design matrix and coefficients.

Proof. We define the transformation:

$$\tilde{\beta}_j = \lambda_j \beta_j, \quad w_{ij} = \frac{x_{ij}}{\lambda_j}, \tag{35}$$

so that the transformed design matrix is:

$$W = XD^{-1}, \quad \text{where } D = \operatorname{diag}(\lambda_1, \ldots, \lambda_p). \tag{36}$$

Substituting $\beta_j = \tilde{\beta}_j / \lambda_j$ into the objective (34), we get:

$$\hat{\tilde{\beta}} = \arg\min_{\tilde{\beta} \in \mathbb{R}^p} \left\{ -\ell\left(D^{-1}\tilde{\beta}; X, y\right) + \sum_{j=1}^{p} \left|\tilde{\beta}_j\right| \right\}. \tag{37}$$

Noting that:

$$X\beta = XD^{-1}\tilde{\beta} = W\tilde{\beta}, \tag{38}$$

we rewrite the log-likelihood as:

$$\ell(\beta; X, y) = \ell(\tilde{\beta}; W, y). \tag{39}$$

Thus, the optimization problem becomes:

$$\hat{\tilde{\beta}} = \arg\min_{\tilde{\beta} \in \mathbb{R}^p} \left\{ -\ell(\tilde{\beta}; W, y) + \sum_{j=1}^{p} |\tilde{\beta}_j| \right\}, \tag{40}$$

This reformulation yields a standard LASSO problem, where the design matrix is given by W, the penalty parameter is uniform across all covariates with $\lambda = 1$, and the response vector remains y. □

After solving (40), we recover the original coefficients using:

$$\hat{\beta}_j = \frac{\hat{\tilde{\beta}}_j}{\lambda_j}, \quad \forall j \in \{1, \ldots, p\}. \tag{41}$$

In what follows, we consider two alternative strategies for selecting the penalty parameters (η_0, η_1). The first approach selects (η_0, η_1) by maximizing the cross-validated AUC, directly targeting predictive performance. The second approach chooses (η_0, η_1) by minimizing the Bayesian Information Criterion (BIC), thereby balancing model fit against complexity.

(i) Grid Search and Bayesian optimization

To tune the hyperparameters $\eta = (\eta_0, \eta_1)$ in the integrative penalty function, we explore both grid search and Bayesian optimization strategies. Initially, we perform a standard grid search over a predefined grid of size $N \times M$ on the domain $\Theta = [\eta_0^{\min}, \eta_0^{\max}] \times [\eta_1^{\min}, \eta_1^{\max}]$, selecting the values that maximize the cross-validated AUC:

$$\hat{\eta} \;=\; \arg\max_{\eta \in \Theta} \text{CV-AUC}(\eta) \;=\; \arg\max_{\eta \in \Theta} \frac{1}{k} \sum_{m=1}^{k} \text{AUC}^{(m)}(\eta), \tag{42}$$

where $\text{AUC}^{(m)}(\eta)$ denotes the area under the ROC curve on the mth fold, based on fitting the multi-penalty LASSO with per-feature penalties $\lambda_j(\eta) = \exp(\eta_0 + \eta_1 z_j)$. However, grid search is computationally expensive, as each of the $N \cdot M$ evaluations requires a full k-fold cross-validation, leading to a higher total cost. To implement this, we defined a grid of candidate values for η_0 and η_1: η_0 was chosen from a dense sequence between -5 and -2 (incremented by 0.25), and η_1 from -2 to 2 (in steps of 0.25). This grid search was applied to the liver dataset using prior-selected genes from breast cancer (obtained via Adaptive Ridge). To mitigate this, we also apply Bayesian optimization as an efficient alternative. Bayesian optimization models the objective function as a draw from a Gaussian process prior:

$$f \;\sim\; \mathcal{GP}\big(m(\eta),\, k(\eta, \eta')\big), \tag{43}$$

where $m : \Theta \to \mathbb{R}$ denotes the prior mean and $k : \Theta \times \Theta \to \mathbb{R}$ the covariance kernel. Hence, for any finite collection of parameters $\{\eta^{(i)}\}_{i=1}^{t} = (\eta_0^{(i)}, \eta_1^{(i)}) \subset \Theta$, where the vector of latent values follows a multivariate normal distribution

$$f_{1:t} := \big(f(\eta^{(1)}), \ldots, f(\eta^{(t)})\big)^\top \sim \mathcal{N}\big(m_{1:t}, K_t\big), \tag{44}$$

with

$$m_{1:t} = \big(m(\eta^{(1)}), \ldots, m(\eta^{(t)})\big)^\top, \qquad K_t = \big[k(\eta^{(i)}, \eta^{(j)})\big]_{i,j=1}^{t}. \tag{45}$$

In our setting, $f : \Theta \to \mathbb{R}$ denotes the latent noise-free AUC function and for each candidate η_i in the parameter space, we perform cross-validation and observe

$$y_i \;=\; \text{CV}-\text{AUC}(\eta^{(i)}) \;+\; \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_n^2), \tag{46}$$

so that

$$y_i \sim \mathcal{N}\big(f(\eta^{(i)}),\, \sigma_n^2\big). \tag{47}$$

Further implementation details are provided in the Appendix. The Bayesian optimization procedure was implemented in `RStudio` using the `BayesOpt` package. A full description of the algorithmic steps can be found in Algorithm 4, along with technical specifications and parameter settings.

(ii) Bayesian Information Criterion (BIC)

As a complementary approach, fitting η_0 and η_1 is attempted by minimizing the BIC. A key benefit of the BIC relative to cross-validation is how the BIC asks whether the added model complexity introduced by each additional parameter is warranted based on the updated model's performance.

Definition 3. Given a statistical model $\mathcal{M}$ with likelihood function $L(\theta \mid \mathcal{D})$, where $\theta \in \mathbb{R}^k$ is the parameter vector and $\mathcal{D} = \{x_1, \ldots, x_n\}$ is the observed dataset, the Bayesian Information Criterion (BIC) is defined as:

$$\mathrm{BIC} = -2 \cdot \log L(\hat{\theta} \mid \mathcal{D}) + k \cdot \log n, \tag{48}$$

where $\hat{\theta}$ denotes the maximum likelihood estimator of the parameter vector, k is the number of free parameters, and n is the number of observations.

Observation 5. In our setting, the model complexity component (k) of the Bayesian Information Criterion accounts for the number of non-zero β_j coefficients, as well as the number of non-zero (η) parameters in equation (49). Consequently, assigning a non-zero value to η_1 increases the effective model dimensionality and is penalized by the BIC. This induces a preference for excluding the external covariate z_j when it does not substantially improve model fit, thereby promoting model parsimony.

The BIC provides a principled framework for model selection by explicitly balancing model fit and complexity. As the sample size n increases, the penalty term $k \log n$ exerts a progressively stronger influence, thereby reducing the risk of overfitting. This asymptotic behavior ensures that the BIC is a consistent model selection criterion, in the sense that it selects the true data-generating model with probability approaching one as $n \to \infty$. Therefore, given two statistical models M_1 and M_2 with corresponding BIC values BIC_1 and BIC_2, the model with the smaller BIC value is preferred, reflecting a more favorable trade-off between goodness of fit and model complexity [15].

To incorporate external information into the liver cancer modeling, we incorporate prior information from the Adaptive Ridge survival and stage modeling from breast cancer. The prior information represents the gene-importance within the breast cancer setting.

Before tuning η_0 and η_1, recall that if Ridge regression or Adaptive Ridge is selected as the primary penalized-likelihood method, we first remove the 25% of genes with the smallest absolute coefficients from the original breast-cancer model. This step removes dimensionality without substantially worsening performance, as mentioned before. The remaining 75% of the genes will be then converted into a binary indicator indicating whether the gene is considered "promising" or not. Let $z_j \in \{0, 1\}$ indicate whether gene j is considered "promising". We define gene-specific penalties as:

$$\log \lambda_j = \eta_0 + \eta_1 \, z_j, \qquad \lambda_j = \exp\bigl(\eta_0 + \eta_1 \, z_j\bigr). \tag{49}$$

η_0 is the baseline level of shrinkage applied to all genes, and η_1 adjusts the gene-specific penalty based on the whether the indicator $z_j = 1$. A large η_1 represents heavy penalties, a negative η_1 reverses the penalty, and a η_1 value close to 0 represents a uniform penalty across all genes.

(e) Bayesian Statistical Inference

High-dimensional classification tasks, especially those encountered in genomic data analysis, present unique challenges that require a careful trade-off between predictive performance and model interpretability. To address these challenges, we employ a Bayesian logistic regression framework with spike-and-slab priors, which enables principled variable selection by promoting sparsity in the model coefficients. Given the computational complexity typically associated with fully Bayesian methods in high dimensions, we utilize variational inference to approximate the posterior distribution in an efficient manner [4].

In the context described above, let $\mathbf{y} \in \{0,1\}^n$ denote the binary response vector, and let $\mathbf{X} \in \mathbb{R}^{n\times p}$ represent the design matrix containing p predictors (gene expression levels), where each row $\mathbf{x}_i^\top$ corresponds to the predictor values associated with the i-th observation. The binary outcome is modeled using logistic regression, defined as follows:

$$y_i \mid \mathbf{x}_i, \boldsymbol{\beta} \sim \mathrm{Bern}(\sigma(\mathbf{x}_i^\top \boldsymbol{\beta})), \quad \text{for } i = 1, \dots, n, \tag{50}$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$ denotes the sigmoid function. The corresponding log-likelihood function is provided in Equation (3).

In the previous setting, parameter estimation was carried out using maximum likelihood, supplemented with a regularization penalty to mitigate overfitting by reducing the variance of the coefficient estimates, albeit at the cost of introducing some bias. In this section, we will adopt an alternative framework of Bayesian inference. Unlike the frequentist approach, which treats model parameters as fixed but unknown quantities, Bayesian inference considers parameters as random variables characterized by probability distributions.

Bayesian inference provides a rigorous probabilistic framework for parameter estimation by integrating prior knowledge with observed data. Uncertainty about the parameters θ is initially expressed through a prior distribution $p(\theta)$, reflecting any existing information or assumptions. The likelihood function $p(\mathcal{D} \mid \theta)$ quantifies the plausibility of the observed data $\mathcal{D}$ under different parameter values. Bayes' Theorem then combines the prior and likelihood to yield the posterior distribution $p(\theta \mid \mathcal{D})$, which represents the updated uncertainty about θ after observing the data. This posterior distribution serves as the foundation for inference and decision-making, offering a coherent and comprehensive quantification of uncertainty.

Theorem 1. [Bayes' Theorem] Let $\theta \in \Theta \subseteq \mathbb{R}^p$ denote the vector of model parameters, and let $\mathcal{D}$ represent the observed data. Then, the posterior distribution of θ conditional on $\mathcal{D}$ is given by:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})}, \tag{51}$$

where the marginal likelihood $p(\mathcal{D})$, is defined as:

$$p(\mathcal{D}) = \int_\Theta p(\mathcal{D} \mid \theta)\, p(\theta)\, d\theta. \tag{52}$$

Observation 6. In the context of this study, the parameter vector is denoted by $\theta = \boldsymbol{\beta}$, and the observed data is given by $\mathcal{D} = (\mathbf{X}, \mathbf{y})$. The posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})$, however, does not admit a closed-form expression, necessitating the use of approximation methods for its evaluation. In this work, we employ variational Bayesian inference as a computationally efficient approach to approximate the posterior distribution.

(i) Bayesian model variable selection

Bayesian variable selection methods often employ spike-and-slab priors to induce sparsity in the coefficient vector $\boldsymbol{\beta}$. This prior formulation explicitly models the inclusion or exclusion of each predictor via latent indicator variables, allowing for flexible and interpretable feature selection within a probabilistic framework.

Definition 4. Let $\beta = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^p$ denote the vector of regression coefficients. The spike-and-slab prior introduces a collection of latent binary indicator variables $\gamma = (\gamma_1, \dots, \gamma_p)^\top \in \{0,1\}^p$, where each γ_j encodes the inclusion ($\gamma_j = 1$) or exclusion ($\gamma_j = 0$) of the j-th predictor.

$$\mathbb{P}(\gamma_j = 1) = \pi, \quad \mathbb{P}(\gamma_j = 0) = 1 - \pi, \quad \text{for } j = 1, \dots, p, \tag{53}$$

where $\pi \in [0, 1]$ denotes the prior inclusion probability for any given predictor.

Conditional on γ_j, the prior for the coefficient β_j is given by a mixture model:

$$p(\beta_j \mid \gamma_j) = \begin{cases} \mathcal{N}(\beta_j; 0, \sigma^2), & \text{if } \gamma_j = 1, \\ \delta_0(\beta_j), & \text{if } \gamma_j = 0, \end{cases} \tag{54}$$

where $\delta_0(\cdot)$ denotes the Dirac delta measure centered at zero, that is, a point mass assigning probability one to the value zero. This setup defines a hierarchical prior over the regression coefficients β in two stages.

The first stage specifies a prior distribution over the inclusion indicators γ as:

$$p(\gamma) = \prod_{j=1}^{p} \text{Bern}(\gamma_j; \pi) = \prod_{j=1}^{p} \pi^{\gamma_j} (1 - \pi)^{1 - \gamma_j}, \tag{55}$$

Conditioned on the inclusion indicators $\boldsymbol{\gamma}$, the prior distribution of the regression coefficients corresponding to the active predictors is given by:

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}) = \prod_{j:\gamma_j = 1} \mathcal{N}\big(\beta_j; 0, \sigma^2\big) = \mathcal{N}\big(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; \mathbf{0}, \sigma^2 \mathbf{I}_{|\boldsymbol{\gamma}|}\big), \tag{56}$$

where $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the subvector of $\boldsymbol{\beta}$ for which $\gamma_j = 1$, and $\mathbf{I}_{|\boldsymbol{\gamma}|}$ is the identity matrix of size equal to the number of active predictors.

Building upon the hierarchical spike-and-slab prior specification for the regression coefficients $\boldsymbol{\beta}$ and inclusion indicators $\boldsymbol{\gamma}$, and given the observed data $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, the posterior distribution over the latent inclusion indicators can be derived using Bayes' theorem as:

$$p(\boldsymbol{\gamma} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\gamma})\, p(\boldsymbol{\gamma})}{p(\mathcal{D})} \propto p(\mathcal{D} \mid \boldsymbol{\gamma})\, p(\boldsymbol{\gamma}), \tag{18}$$

where the marginal likelihood is defined by the integral:

$$p(\mathcal{D} \mid \boldsymbol{\gamma}) \;=\; \int p\big(\mathcal{D} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}\big)\, p\big(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}\big)\, d\boldsymbol{\beta}_{\boldsymbol{\gamma}}. \tag{57}$$

Observation 7. For notational simplicity, from this point forward we denote the posterior inclusion probability (PIP) of the j-th variable as:

$$\text{PIP}_j = P(\gamma_j = 1 \mid \mathcal{D}), \quad \text{for all } j = 1, \dots, p. \tag{58}$$

Due to the computational intractability of the integral in high-dimensional parameter spaces, exact evaluation of the posterior distribution and related quantities is generally infeasible. In the present context, Coordinate Ascent Variational Inference (CAVI) will be employed as an approximate inference framework to obtain computationally tractable estimates of the posterior distribution.

(ii) Variational Inference for Approximation of Posterior Inclusion Probabilities

For the purpose of inference, we utilize the `R` programming language. In particular, the `varbvs` package is employed, which implements a Coordinate Ascent Variational Inference (CAVI) algorithm tailored for sparse Bayesian logistic regression models incorporating continuous spike-and-slab priors. This approach is well-suited for scalable variable selection in high-dimensional settings and has been widely applied in fields such as genetics and biostatistics.

Probabilistic Model.

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote the matrix of predictor variables and $\mathbf{y} \in \{0,1\}^n$ the corresponding binary response vector. The model is specified as follows:

$$y_i \mid \mathbf{x}_i, \boldsymbol{\beta} \sim \text{Bern}(\sigma(\mathbf{x}_i^\top \boldsymbol{\beta})), \quad \text{for } i = 1, \ldots, n, \tag{59}$$

$$\sigma(t) = \frac{1}{1 + e^{-t}}, \tag{60}$$

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j) \cdot \delta_0 + \gamma_j \cdot \mathcal{N}(0, \sigma^2), \quad \text{for } j = 1, \ldots, p, \tag{61}$$

$$\gamma_j \sim \text{Bern}(\pi), \quad \text{for } j = 1, \ldots, p \tag{62}$$

Variational Approximation.

Variational inference seeks to approximate the intractable posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$ by selecting a distribution from a tractable variational family $\mathcal{Q}$ that is closest in Kullback–Leibler (KL) divergence (see Definition 5 in Appendix). Formally, the optimal approximation is obtained by solving:

$$q^* = \arg\min_{q \in \mathcal{Q}} \text{KL}\left(q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \,\|\, p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})\right), \tag{63}$$

To approximate the true posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y}, \mathbf{X})$, the `varbvs` algorithm adopts a mean-field variational approximation. Specifically, it assumes a fully factorized variational family of the form:

$$q(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{j=1}^{p} q(\beta_j, \gamma_j), \tag{64}$$

where each marginal variational factor $q(\beta_j, \gamma_j)$ is defined as a spike-and-slab distribution:

$$q_j(\gamma_j, \beta_j) = PIP_j \, \mathbf{1}_{\{\gamma_j = 1\}} \mathcal{N}(\beta_j; m_j, s_j^2) + (1 - PIP_j) \, \mathbf{1}_{\{\gamma_j = 0\}} \delta_0(\beta_j). \tag{65}$$

where PIP_j, m_j and s_j^2 denote the posterior inclusion probability, posterior mean and posterior variance, respectively, of the coefficient β_j.This factorization implies a strong independence assumption: under q, all pairs (β_j, γ_j) are mutually independent, and within each pair, β_j is conditionally independent of other variables given γ_j. This is a standard assumption in mean-field variational inference, which greatly simplifies the optimization of the evidence lower bound by decoupling the parameters across coordinates.

Coordinate Ascent Updates.

The coordinate ascent variational inference procedure updates variational parameters sequentially, optimizing each while holding others fixed. Each variational factor is updated as:

$$\log q_j^*(\theta_j) = \mathbb{E}_{q_{-j}}\left[\log p(\theta_j, \theta_{-j}, y, X)\right] + \text{const}, \tag{66}$$

where the expectation is taken with respect to the variational distribution over all variables except θ_j.

The resulting closed-form update equations are applied iteratively until convergence. The following algorithm outlines this procedure as applied to Bayesian logistic regression with continuous spike-and-slab priors. Further details and derivations of this coordinate ascent variational inference scheme can be found in [16, 8] and the Appendix.

Algorithm 2 Coordinate Ascent Variational Inference for Spike-and-Slab Logistic Regression.

```
Input: Data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, response vector $\mathbf{y} \in \{0,1\}^n$, prior inclusion probability $\pi$, prior
slab variance $\sigma^2$, initial variational parameters: $m_j^{(0)}, s_j^{2(0)}, PIP_j^{(0)}$, and auxiliary variables $\xi_i^{(0)}$.
repeat
    Compute $\lambda(\xi_i) = \frac{1}{2\xi_i}\tanh(\xi_i/2)$ for all $i = 1, \ldots, n$.
    Construct diagonal matrix $\Lambda = \mathrm{diag}(\lambda(\xi_1), \ldots, \lambda(\xi_n))$.
    for each variable $j = 1, \ldots, p$ do
        Update posterior variance: $s_j^2 \leftarrow \left(\mathbf{x}_j^\top \Lambda \mathbf{x}_j + \frac{1}{\sigma^2}\right)^{-1}$
        Update posterior mean: $m_j \leftarrow s_j^2 \cdot \mathbf{x}_j^\top \left(\mathbf{y} - \frac{1}{2} - 2\Lambda(\sum_{k\neq j} \mathbf{x}_k PIP_k m_k\right))$
        Update logit of inclusion probability: $\mathrm{logit}(PIP_j) \leftarrow \log\left(\frac{\pi}{1-\pi}\right) + \frac{1}{2}\log\left(\frac{s_j^2}{\sigma^2}\right) + \frac{m_j^2}{2s_j^2}$
        Compute updated inclusion probability: $PIP_j \leftarrow \frac{1}{1+\exp(-\mathrm{logit}(PIP_j))}$
    end for
    Update auxiliary variables: $\xi_i \leftarrow \sqrt{\sum_{j=1}^p x_{ij}^2 PIP_j (m_j^2 + s_j^2)}$
until convergence
Output: Posterior inclusion probabilities $PIP_j$, posterior means $m_j$, and variances $s_j^2$.
```

(iii) Implementation

In this section, we present a two-step empirical Bayes methodology designed to incorporate structured prior information into a variational Bayesian variable selection model. The goal is to reflect prior biological beliefs that genes identified through prior knowledge or external models may have a higher likelihood of being associated with the outcome. Our strategy estimates these prior inclusion probabilities in a data-driven yet principled way, without excluding any variable from the final analysis. The development of the methodology in this work is primarily informed by the foundational contributions in the literature, particularly the works of [14], [3], and [6], which provide essential theoretical and methodological underpinnings for Bayesian variable selection.

Let $\mathcal{G} = \{1, \ldots, p\}$ denote the complete set of gene predictors, where $\mathbf{X} \in \mathbb{R}^{n\times p}$ represents the standardized design matrix and $\mathbf{y} \in \{0,1\}^n$ is the binary response vector. To reduce computational complexity at the initial stage, we define a screening subset $\mathcal{G}_{\text{screen}} \subset \mathcal{G}$ of size k, which is composed of two disjoint groups:

$$\mathcal{G}_{\text{screen}} = \mathcal{G}_{\text{prom}} \cup \mathcal{G}_{\text{non}}, \tag{67}$$

where $\mathcal{G}_{\text{prom}}$ comprises k_1 promising genes selected based on prior biological evidence or their inclusion in previous models. The complementary set, $\mathcal{G}_{\text{non}}$, consisting of $k_2 = k - k_1$ genes, is defined as

$$\mathcal{G}_{\text{non}} = \underset{\substack{\mathcal{S}\subseteq\mathcal{G}\setminus\mathcal{G}_{\text{prom}} \\ |\mathcal{S}|=k_2}}{\arg\max} \sum_{j\in\mathcal{S}} \mathrm{var}(X_j), \tag{68}$$

selecting genes with the highest empirical variance across samples.

Next, we apply a variational Bayesian variable selection model (implemented via `varbvs`) to this screening subset, assuming a uniform prior inclusion probability:

$$\pi_j^{(0)} = \pi_0, \quad \forall j \in \mathcal{G}_{\text{screen}}. \tag{69}$$

This procedure yields posterior inclusion probabilities (PIPs), denoted by $\widehat{\mathrm{PIP}}_j \in [0,1]$, for each $j \in \mathcal{G}_{\text{screen}}$:

$$\widehat{\mathrm{PIP}}_j = \mathbb{P}(\beta_j \neq 0 \mid \mathbf{y}, \mathbf{X}_{\mathcal{G}_{\text{screen}}}). \tag{70}$$

We then calculate group-specific empirical prior inclusion probabilities by averaging the PIPs within each group:

$$\bar{\pi}_{\text{prom}} = \frac{1}{|\mathcal{G}_{\text{prom}}|} \sum_{j \in \mathcal{G}_{\text{prom}}} \widehat{\mathrm{PIP}}_j, \quad \bar{\pi}_{\text{non}} = \frac{1}{|\mathcal{G}_{\text{non}}|} \sum_{j \in \mathcal{G}_{\text{non}}} \widehat{\mathrm{PIP}}_j. \tag{71}$$

These averages are used to define a structured prior inclusion probability $\pi_j^{(1)}$ for each gene, depending on its membership in the promising or non-promising groups:

$$\pi_j^{(1)} = \begin{cases} \bar{\pi}_{\text{prom}}, & \text{if } j \in \mathcal{G}_{\text{prom}}, \\ \bar{\pi}_{\text{non}}, & \text{if } j \in \mathcal{G}_{\text{non}}. \end{cases} \tag{72}$$

Accordingly, each gene $j \in \mathcal{G}$ is assigned a prior inclusion probability that reflects both empirical evidence and prior biological knowledge. Finally, these prior probabilities are converted into prior log-odds for use in the subsequent variational inference step:

$$\text{logodds}_j = \log\left(\frac{\pi_j^{(1)}}{1-\pi_j^{(1)}}\right), \quad \forall j \in \mathcal{G}. \tag{73}$$

Subsequently, predictors with posterior inclusion probabilities exceeding 0.5 are selected. This cutoff corresponds to the median probability model, a widely accepted criterion for variable selection [13].

The entire methodology is formalized in Algorithm 3. This algorithm encapsulates the essential steps of the empirical Bayes framework that underpins the variable selection process. Recall that the estimation of posterior inclusion probabilities is detailed in Algorithm 2. As mentioned before, we split the data into training and testing sets to evaluate predictive performance and enable comparison with prior models. The model is trained on the training set using the two-stage prior updating procedure described below, and performance metrics are computed on the held-out test set.

Algorithm 3 Empirical Bayes Variational Variable Selection.

```
Input: Data $\mathbf{X} \in \mathbb{R}^{n \times p}$, response $\mathbf{y} \in \{0,1\}^n$, screening size $k$, uniform prior inclusion probability $\pi_0 \in (0,1)$
Split data into training and testing sets: $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}), (\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$
Define screening subset on training data: $\mathcal{G}_{\text{screen}} = \mathcal{G}_{\text{prom}} \cup \mathcal{G}_{\text{non}}$ where $\mathcal{G}_{\text{prom}}$ are genes selected based on prior biological evidence or previous models, and
    $$\mathcal{G}_{\text{non}} = \underset{\substack{\mathcal{S} \subseteq \mathcal{G} \setminus \mathcal{G}_{\text{prom}} \\ |\mathcal{S}| = k - |\mathcal{G}_{\text{prom}}|}}{\arg\max} \sum_{j \in \mathcal{S}} \text{var}((X_{\text{train}})_j)$$
Extract $\mathbf{X}_{\text{train,screen}}$ and $\mathbf{X}_{\text{test,screen}}$ corresponding to $\mathcal{G}_{\text{screen}}$
Fit variational Bayesian logistic regression on $(\mathbf{X}_{\text{train,screen}}, \mathbf{y}_{\text{train}})$ with prior $\pi_j^{(0)} = \pi_0$
Compute posterior inclusion probabilities (PIPs):
    $$\widehat{\text{PIP}}_j = \mathbb{P}(\beta_j \neq 0 \mid \mathbf{y}_{\text{train}}, \mathbf{X}_{\text{train,screen}}), \quad \forall j \in \mathcal{G}_{\text{screen}}$$
Calculate group-wise average PIPs:
    $$\bar{\pi}_{\text{prom}} = \frac{1}{|\mathcal{G}_{\text{prom}}|} \sum_{j \in \mathcal{G}_{\text{prom}}} \widehat{\text{PIP}}_j, \quad \bar{\pi}_{\text{non}} = \frac{1}{|\mathcal{G}_{\text{non}}|} \sum_{j \in \mathcal{G}_{\text{non}}} \widehat{\text{PIP}}_j$$
Update prior inclusion probabilities based on group membership:
    $$\pi_j^{(1)} = \begin{cases} \bar{\pi}_{\text{prom}}, & \text{if } j \in \mathcal{G}_{\text{prom}}, \\ \bar{\pi}_{\text{non}}, & \text{if } j \in \mathcal{G}_{\text{non}} \end{cases}$$
Fit variational Bayesian logistic regression again on $(\mathbf{X}_{\text{train,screen}}, \mathbf{y}_{\text{train}})$ with updated priors $\pi_j^{(1)}$
Select predictors with $\widehat{\text{PIP}}_j > 0.5$:
    $$\widehat{\mathcal{G}} = \{ j \in \mathcal{G}_{\text{screen}} : \widehat{\text{PIP}}_j > 0.5 \}.$$
Predict on test set using $\mathbf{X}_{\text{test},\widehat{\mathcal{G}}}$.
Evaluate AUC on $(\mathbf{X}_{\text{test},\widehat{\mathcal{G}}}, \mathbf{y}_{\text{test}})$.
Output: $\widehat{\mathcal{G}}$ and $\{\widehat{\text{PIP}}_j\}_{j \in \widehat{\mathcal{G}}}$.
```

8 Validation and Results

In this section, we apply the statistical methodologies described earlier to the biomedical datasets under consideration. We begin with an exploratory descriptive analysis, assessing both sample-level and gene-level summary statistics in order to understand the underlying structure and variability within the data.

Subsequently, we implement the regression models introduced in previous sections, applying them to different target variables. These models form the basis for the subsequent data integration procedures. Initially, the obtained results did not meet expectations, prompting the inclusion of a validation study solely within liver cancer stage progression. This additional step demonstrated the robustness of our methodology, although the predictive power of the selected within-cancer features is limited.

We then proceed to the Bayesian statistical inference component, where the results obtained using Bayesian variable selection are analyzed. This section also includes a biological interpretation of the findings. Some modifications to the initial approach were necessary due to computational limitations encountered during the inference process.

(a) Descriptive Statistical Analysis of Data

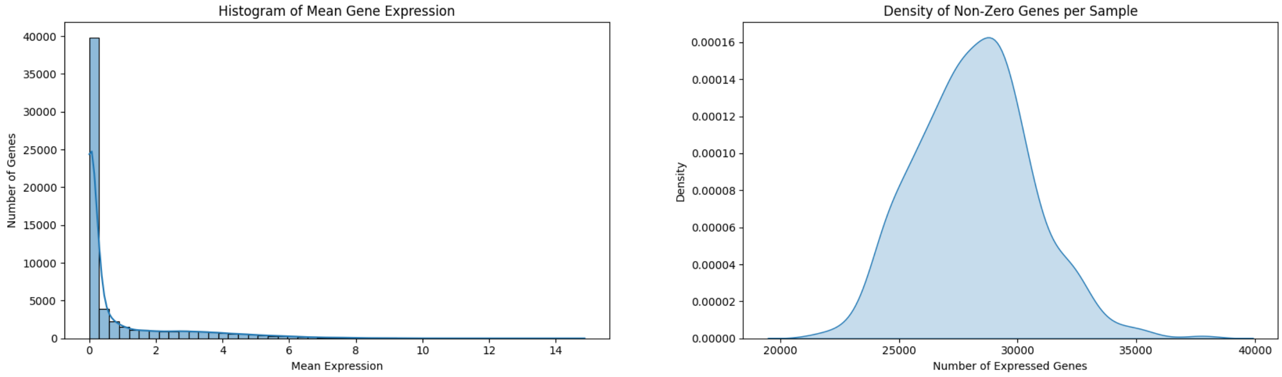
In this phase, we summarize key features of the datasets by calculating central tendency and dispersion measures such as means, medians, and standard deviations at both the sample and gene levels. We also assess data quality by identifying missing values and outliers, as well as examining distribution patterns. This initial exploration provides critical insights that guide the selection and implementation of subsequent regression models.

(i) Descriptive Analysis of Liver Gene Expression

Summary statistics were computed at the sample and gene level to explore the structure and distribution of the liver cancer gene expression data. For each sample, total gene expression, average expression, median expression, number of expressed genes (non-zero entries), and the fraction of missing values were calculated. At the gene level, the mean and standard deviation of expression across all patients were computed.

Sample-Level Statistics. Figure 4a, which represents the histogram for mean gene expression, clearly demonstrates a positively skewed distribution. The distribution of mean gene expression is heavily concentrated around 0, with only a select few genes demonstrating high expression values. This is not uncommon in gene expression datasets.

Figure 4b shows the density distribution of non-zero genes per patient samples. On average, each patient sample has 28,200 active genes, with a standard deviation of 2,405. We observe a slightly right-skewed distribution. While most samples have similar density, we do observe a fair degree of variation across patient samples.



(a) Histogram of Mean Gene Expression.

(b) Density of Non-Zero Genes per Sample.

Figure 4: Summary of Sample-Level Gene Expression Statistics. These plots provide an overview of expression magnitude and transcriptome coverage across all samples.

Gene-Level Statistics. Figure 5 shows a scatter plot of the mean and standard deviation of gene expression across all patient samples. We observe a positive relationship between the covariates, as a higher mean gene expression tends to be associated with higher variance across patients.

However, there is clearly a non-linear and heteroskedastic relationship between the covariates. Genes with low mean expression tend to have small variance and center around the origin. As the mean expression increases, the standard deviation increases, albeit with varying spread across the plot. These differing variations with increased expression could potentially be associated with factors such as different cancer sub-types or tumor malignancies.

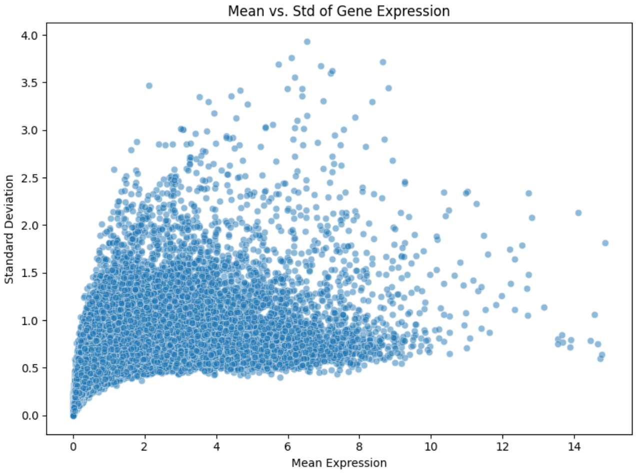
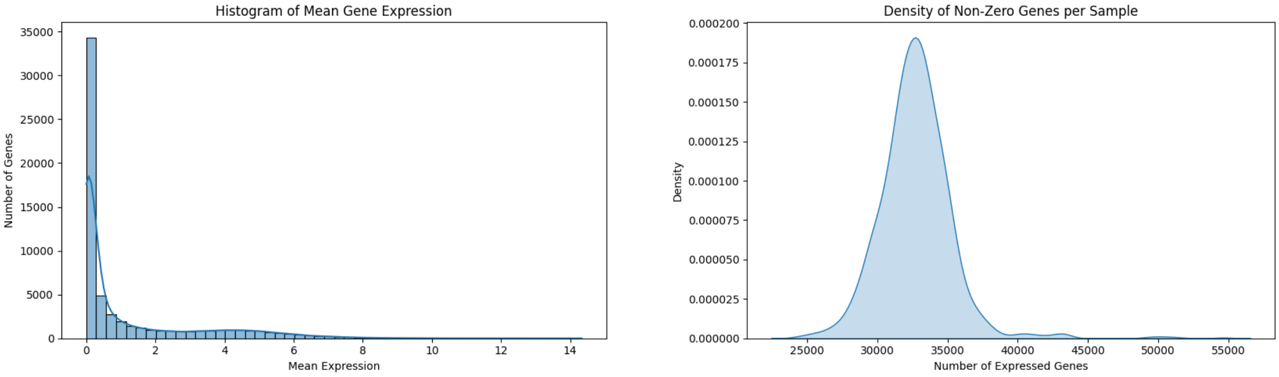


Figure 5: Mean vs. Standard Deviation of Gene Expression. Each point represents a gene. Higher expression is often associated with higher variability.

(ii) Descriptive Analysis of Breast Gene Expression

The same summary statistics were computed both at the sample level and the gene level for the breast dataset.

Sample-Level Statistics The histogram of mean gene expression (Figure 6a) reveals a highly right-skewed distribution, indicating that the majority of genes are expressed at very low levels across patients. This sparsity is a common feature in genomic datasets and motivates the decision to filter out lowly expressed genes in later preprocessing steps. The density plot of the number of non-zero genes per sample (Figure 6b) shows that samples typically express between 30,000 and 35,000 genes, indicating broad coverage.



(a) Histogram of Mean Gene Expression.

(b) Density of Non-Zero Genes per Sample.

Figure 6: Summary of Sample-Level Gene Expression Statistics. These plots provide an overview of expression magnitude and transcriptome coverage across all samples.

Gene-Level Statistics At the gene level, a scatterplot of mean versus standard deviation of expression (Figure 7) reveals a clear positive relationship between average expression and variability. Genes with higher mean expression tend to also show greater dispersion across patients, which may indicate biologically meaningful variability. For example, genes that are more active in some patients than others may be involved in disease subtypes or outcomes, making them useful for further analysis.

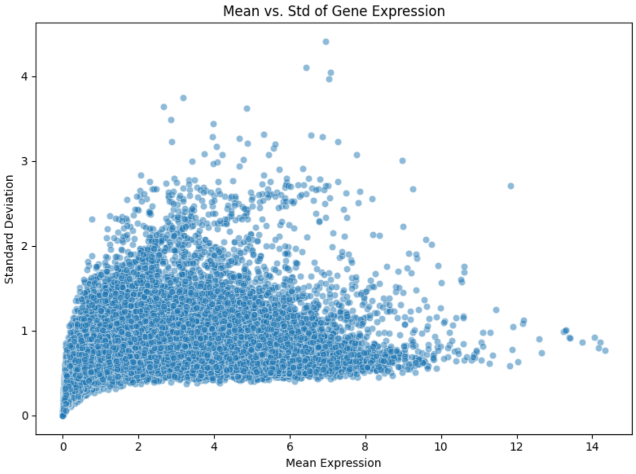


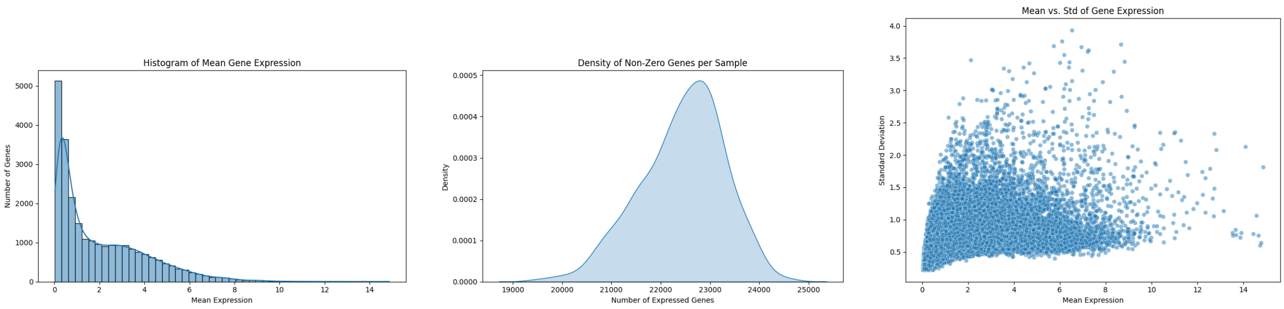
Figure 7: Mean vs. Standard Deviation of Gene Expression. Each point represents a gene. Higher expression is often associated with higher variability.

(b) Gene Feature Selection

In this section, we overview the feature selection process that was informed from our descriptive statistical analysis of the data. The exploratory analysis confirms a broad dynamic range in expression levels across the gene and samples for both liver and breast cancer, with most genes expressed at low levels and a smaller subset exhibiting high expression. We then applied several filtering steps to reduce the number of uninformative or low-quality gene features.

First, we removed genes that were sparsely expressed across the cohort. For the liver dataset, we dropped 4,265 genes that had zero expression in all samples, and an additional 26,321 genes that were expressed (non-zero) in less than 20% of patients. This step reduced the number of gene features from 60,660 to 34,455. For the breast dataset, we dropped 2,660 genes that had zero expression in all samples, and an additional 21,146 genes that were expressed (non-zero) in less than 20% of patients. This step reduced the number of gene features from 60,660 to 39,624.

Next, we filtered out genes with very low variability or irregular expression distributions. Genes that are expressed in only a small fraction of samples, or that show almost no variation across patients, are often considered “asymmetrical” or biologically uninformative. They can also introduce noise or bias in statistical models. After removing these low-variance or highly skewed genes, the liver dataset was reduced to 25,517 gene expression features and the breast dataset was reduced to 33,092 gene expression features. These filtering steps ensure that the remaining genes are more likely to be informative for distinguishing patient subgroups and predicting clinical outcomes.

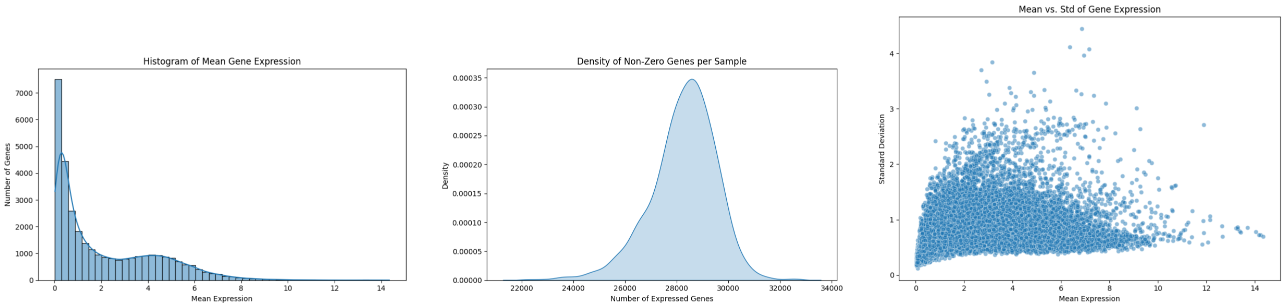


(a) Histogram of Mean Expression. **(b)** Density of Non-Zero Genes. **(c)** Mean vs. Std of Expression.

Figure 8: Summary plots for liver gene expression dataset.

The histogram of mean gene expression for both liver cancer (Figure 8a) and breast cancer (Figure 9a) reveal that while most genes still exhibit low average gene expression, the long tail of highly expressed genes is more distinct, reflecting the removal of genes with minimal or no activity across samples. The density plot of non-zero gene counts per sample for liver cancer (Figure 8b) and for breast cancer (Figure 9b) confirm that each sample retains expression for a large number of genes, indicating good coverage and minimal sparsity post-filtering. However, we do observe higher variance of the non-zero genes in the liver cancer dataset.

Finally, the scatterplot of mean versus standard deviation for the liver cancer (Figure 8c) and breast cancer datasets (Figure 9c) demonstrate a clear relationship between average expression and variability, with low-variance genes largely removed. These distributions confirm that the filtering steps effectively eliminated uninformative features while preserving biologically relevant variability across the patient cohort.



(a) Histogram of Mean Expression. **(b)** Density of Non-Zero Genes. **(c)** Mean vs. Std of Expression.

Figure 9: Summary plots for breast gene expression dataset.

We chose not to log-transform the data because our analysis focuses on modeling absolute gene expression magnitudes, where preserving the original scale is essential for interpretability, particularly for identifying genes that exceed meaningful expression thresholds relevant to marker detection. Observing if log-transormation leads to a meaningful difference to performance, albeit at the cost of interpretability, is an idea for future research.

(c) Regressions Analysis Findings

Now that the requisite data cleaning and pre-processing steps have been undertaken, we proceed to the regression analysis. Six regularized logistic regression methods (LASSO, Adaptive LASSO with initial LASSO estimates, Ridge, Adaptive LASSO with initial Ridge estimates, Elastic Net, and Adaptive LASSO with initial Elastic Net estimates) were applied to both the liver and breast cancer datasets to predict binary outcomes. These binary outcomes were survival status and whether the pathological cancer stage was in stage I/II or in stage III/IV. For the survival status, there were two different models: one that included pathological cancer stage as a covariate, and one that excluded it. Finally, the effect of including metadata was also analyzed.

Observation 8. From now on, we will refer to the adaptive LASSO model that utilizes initial estimates obtained from Ridge regression as Adaptive Ridge, and the version that employs initial estimates from the Elastic Net as Adaptive Elastic Net. These terms will be used throughout the remainder of this thesis for clarity and brevity.

As mentioned in Section 3(b), the ROC-AUC was used as the primary evaluation metric for each classifier. As mentioned in Section 3(c), 10-fold cross-validation was employed to identify the optimal λ and α. Once models were fitted separately for both datasets, data integration was performed to assign differential penalties to the coefficients depending on whether they originated from liver or breast cancer data. The optimal penalization parameters were determined using Bayesian optimization, and model selection was guided by the Bayesian Information Criterion (BIC). Finally, we assessed whether incorporating this differential penalty improved model performance compared to using liver cancer data alone.

(i) Breast cancer

Survival Status (Without Knowledge of Pathological Stage) Now, we will discuss the results for the various regression variants on the different target variables of interest. First, we will assess the overall survival status without knowing the pathological stage of breast cancer. See Table 3 for the preliminary results.

Method	**AUC**	**Non-Zero Covariates**
Adaptive Elastic Net	0.636	148
Adaptive Ridge	0.585	33,004
Ridge	0.583	33,004
Adaptive LASSO	0.569	39
Elastic Net	0.568	444
LASSO	0.514	39

Table 3: Preliminary model results for predicting overall survival (`y_os`) in the breast cancer dataset, sorted by descending AUC.

As shown in Table 3, the Adaptive Elastic Net model achieved the highest AUC (0.636) for predicting overall survival, followed by Adaptive Ridge and Ridge regression. Both Adaptive Ridge and Ridge regression had an AUC slightly above 0.580. Recall that Ridge regression does not set covariate values to zero, so all genes (33,004) remain in the model. In contrast, the Adaptive Elastic Net achieved better performance with substantially fewer gene covariates, with only 148 non-zero gene covariates being selected.

The Adaptive LASSO, Elastic Net, and LASSO models achieved worse AUC values than the aforementioned models. The standard LASSO only preserved 39 gene covariates, which could have led to an overly simplistic model that struggled with unseen data.

As explained in Section *(i)*, the LASSO tends to perform poorly when covariates are highly correlated.

Thus, it is not surprising that the model struggled, given that the gene covariates in question are highly correlated. Recall that the Elastic Net addresses this limitation by combining the ℓ_1 penalty of the LASSO with the ℓ_2 penalty of Ridge regression. The Adaptive Elastic Net goes a step further by assigning different penalty weights to different coefficients. In doing so, the Adaptive Elastic Net improved model performance by simultaneously preserving important genes and heavily penalizing non-informative genes.

Pathological Stage See Table 4 for the preliminary results for predicting pathological stage (i.e. whether the cancer is in stage III/IV or not).

Method	**AUC**	**Non-Zero Covariates**
Ridge	0.633	33,004
Adaptive Ridge	0.627	33,004
Elastic Net	0.578	1,490
LASSO	0.573	241
Adaptive Elastic Net	0.561	283
Adaptive LASSO	0.555	161

Table 4: Preliminary model results for predicting pathological stage (stage III/IV indicator, `is_stage34`) in the breast cancer dataset, sorted by descending AUC.

As shown in Table 4, Ridge regression achieved the highest AUC (0.633) for predicting pathological stage, followed closely by Adaptive Ridge (0.627). Recall that these two models used all available gene covariates (33,004), shrinking coefficients towards zero but not removing them altogether as defined within the ℓ_2 penalty. This indicates that retaining all gene covariates is the best strategy. Ridge regression shrinks all coefficients uniformly, whereas Adaptive Ridge regression has a unique data-driven weight for each covariate. The standard Ridge performs slightly better than the Adaptive Ridge, which indicates that there may not be specific, dominant genes that are predictive of pathological stage. Instead, applying a balanced weighting across many genes is more effective.

Despite selecting far fewer features, the sparsity-inducing methods such as LASSO and Elastic Net performed much worse than the nonsparsity Ridge regression methods. The results indicate that the biologically relevant information lies in the summation of all gene covariates as opposed to a small subset of genes.

Survival Status (With Knowledge of Pathological Stage) See Table 5 for the preliminary results for predicting overall survival, while now knowing the patient's pathological stage.

Method	AUC	Non-Zero Covariates
Adaptive Elastic Net	0.644	146
Adaptive LASSO	0.615	37
Elastic Net	0.595	455
Adaptive Ridge	0.588	33,005
Ridge	0.586	33,005
LASSO	0.574	45

Table 5: Preliminary model results for predicting overall survival with access to pathological stage information in the breast cancer dataset, sorted by descending AUC.

By adding in knowledge of the pathological stage, each model's AUC improves relative to the same model type without knowledge of the pathological stage. As one would expect, knowledge of the pathological stage is predictive of breast cancer survival status.

The Adaptive Elastic Net model achieves the highest AUC (0.644) by a fair margin, while selecting two fewer covariates than the version which did not include pathological stage as a covariate. The Adaptive LASSO was the second-best model with an AUC of 0.615, while selecting only 37 gene covariates. The substantial improvement in the Adaptive LASSO's performance following the knowledge of the pathological stage likely stems from the strong predictive power of the pathological stage. LASSO models strive to shrink covariates and isolate the most important features, so adding in a covariate with strong predictive power boosts the performance of the Adaptive LASSO. The flexibility of the adaptive models, given the weak signal of the individual genes coupled with the powerful binary predictor of cancer stage, proved effective in this use case.

In contrast, the Ridge and Adaptive Ridge regression models struggled in this context. This likely originated from the pathological stage covariate coefficient being suppressed by over 30,000 gene features. The LASSO model had the lowest AUC (0.574). We observe that the Adaptive LASSO actually selected less coefficients than the standard LASSO, while yielding a higher AUC. A possible explanation for this is that the LASSO retained some non-zero coefficients which were not particularly predictive, and these coefficients were eliminated in the Adaptive LASSO and led to improved model performance.

(ii) Liver cancer

For the liver cancer data, model performance showed a clear trade-off between accuracy and the number of selected covariates. LASSO achieved the best results for predicting survival, both when used alone (AUC = 0.637) and in the combined predictions without metadata (AUC = 0.651), while selecting a relatively small number of covariates (124). On the other hand, Adaptive Ridge and Ridge regression performed best for predicting pathological stage (AUC = 0.692 and 0.684), which indicates that keeping all genes in the model was more effective for in the liver dataset than it was for the breast dataset. When metadata was included, the Adaptive Elastic Net achieved the highest AUC (0.736) while keeping the number of covariates low (104), showing that this method could efficiently combine gene information with clinical variables.

(a) Survival Status (y_os).

Method	AUC	Non-Zero Covariates
LASSO	0.637	125
Adaptive LASSO	0.628	63
Adaptive Elastic Net	0.606	7
Elastic Net	0.597	8
Adaptive Ridge	0.596	25,431
Ridge	0.583	25,431

(b) Pathological Stage ($is\_stage34$).

Method	AUC	Non-Zero Covariates
Adaptive Ridge	0.692	25,431
Ridge	0.684	25,431
LASSO	0.663	73
Adaptive Elastic Net	0.643	19
Elastic Net	0.633	19
Adaptive LASSO	0.590	50

(c) Combined Predictions (No Metadata).

Method	AUC	Non-Zero Covariates
LASSO	0.651	124
Adaptive LASSO	0.620	63
Adaptive Elastic Net	0.605	7
Elastic Net	0.604	8
Adaptive Ridge	0.596	25,432
Ridge	0.585	25,432

(d) Combined Predictions (With Metadata).

Method	AUC	Non-Zero Covariates
Adaptive Elastic Net	0.736	104
LASSO	0.719	165
Adaptive LASSO	0.711	73
Elastic Net	0.694	535
Adaptive Ridge	0.604	25,446
Ridge	0.589	25,466

Table 6: Preliminary model results for liver cancer predictions, organized by task.

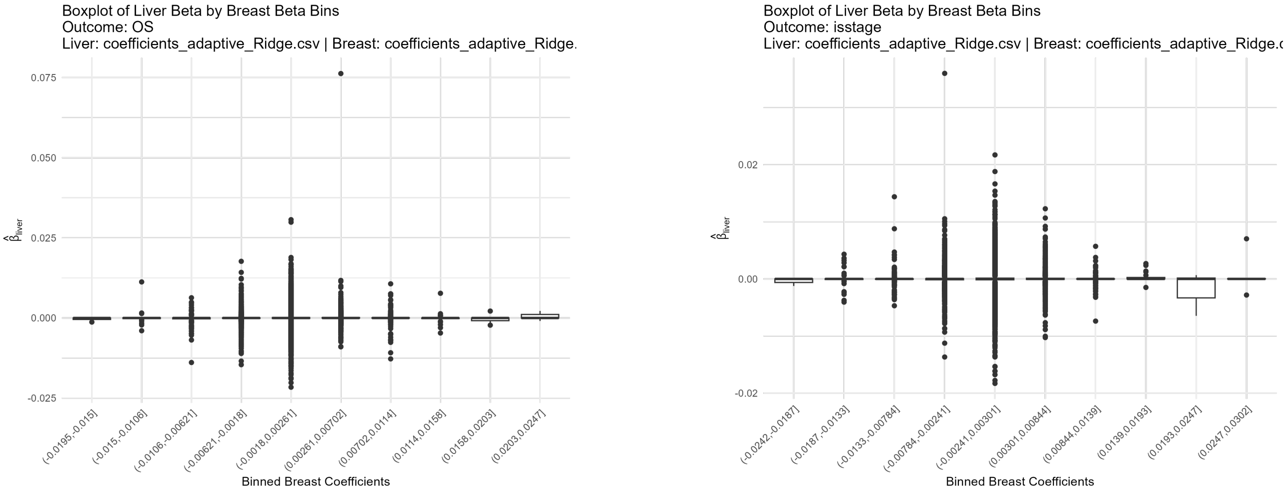
The liver cancer results have some interesting differences relative to the breast cancer results. In both datasets, Ridge-based models were best for predicting pathological stage, suggesting that spreading weights across many genes is important for stage classification. However, for survival prediction, breast cancer models performed best with Adaptive Elastic Net, while liver cancer survival was better predicted with the stronger sparsity employed in the LASSO. This suggests that, for liver cancer, a smaller group of genes carries more useful information for survival prediction, especially when combined with metadata. In contrast, breast cancer survival seems to require more flexible methods that can handle many correlated genes. Overall, these results highlight that the best modeling approach depends not only on the type of prediction task but also on the biological characteristics of the cancer type.

Across the four metadata-inclusive models that incorporated variable selection, the most predictive feature as to whether the patient would die of liver cancer was whether the sampled tumor was "Normal" as opposed to "Tumorous". This is counterintuitive and it is unclear why this is the case, but it may have to do with post-mortem tumor sampling. Age, whether the pathological stage was in stage III or stage IV, and whether the primary diagnosis was HCC or NOS were all predictive of patient death across the different models which performed feature selection. Given the lack of improvement in performance and the counterintuitive findings, we decided to not incorporate metadata covariates within the subsequent data integration steps. Instead, we only proceed with gene covariates.

In this following sections, we focused only on model variants predicting survival status and pathological stage. The combined model, which would include cancer stage as an additional covariate, and the model with all the metadata were not considered for further analysis. This decision was based on the fact that models using only gene expression data did not achieve sufficiently strong results to justify a more in-depth investigation. Furthermore, including metadata in the data integration methods could have downweighted the gene expression data to an undesirable level.

(d) Data integration

In this section, we present the results from our data integration methods. We evaluate predictive performance using the ROC-AUC, and assess model complexity and fit using the BIC. Furthermore, we attempt to validate the underlying statistical methodology by performing a validation test on the within-liver cancer progression from stages I/II to stages III/IV. Prior to beginning the data integration, we first assess the correlation between the gene effects in breast and liver cancer.



(a) Survival status correlation.

(b) Cancer state correlation.

Figure 10: Correlation between liver and breast genes for different outcomes.

The plots in Figure 10 clearly demonstrate that there is a low correlation amongst the genes effects in breast and liver cancer.

(i) ROC-AUC

Informed by the previous results, we now proceed to integrate the data. The first step in this process is to select the most appropriate models for each type of cancer. The aim was to balance high predictive performance (AUC) with model sparsity (number of selected covariates). Our primary goal was to choose the models with the highest AUC. However, this often resulted in very few genes being shared between the two cancers, which limited the potential for a meaningful joint analysis. For this reason, we decided to also consider the number of common covariates between the models as a secondary criterion for model selection.

As described in the methodology, the initial step for data integration involves analyzing the correlation between gene effects across the two cancer datasets. To facilitate this assessment, we employed the following plots in Figure 10. For the survival status covariate, Adaptive Lasso with initial Ridge estimates was applied to both cancer datasets. In the case of breast cancer, this model was ranked second by AUC and selected all covariates. In comparison, the standard LASSO model retained only 148 covariates for breast cancer, none of which overlapped with those selected for liver cancer.

The above results in Table 6 indicate that the genes identified by LASSO in the breast cancer model were not sufficient or informative when integrating with the liver cancer data. To enable a broader and more robust gene selection for the joint analysis, we opted for the Adaptive Lasso model (using initial Ridge estimates) for breast cancer. The coefficients selected by this model will serve as prior information to guide the regression analysis for liver cancer.

Since Adaptive Ridge regression was selected as the modeling approach, and no coefficients were shrunk exactly to zero, a post hoc variable selection procedure was implemented to identify the most relevant genes. Specifically, the regression coefficients were ranked in descending order based on their absolute values, and the 25% of genes corresponding to the smallest coefficients were subsequently removed from the model.

Informed by the lack of correlation amongst the gene effects in breast and liver cancer (See Figure 10), we do not expect a significant improvement in AUC by applying the methodology adapted penalty. The plots indicate that the gene effects are generally weak and do not reveal a strong or consistent shared signal between breast and liver cancer. To confirm this, we applied the transfer learning method using both grid search and Bayesian-optimized adaptive LASSO.

The optimal parameter combination with the grid search was found at $\theta_0 = -2.75$ and $\theta_1 = 0.25$. To assess the practical impact of learning θ_1, we compared the test AUC when using a global penalty ($\theta_1 = 0$) versus the learned value ($\theta_1 = 0.25$). The test AUC improved from 0.647 to 0.6538, resulting in a modest gain of 0.0068. This suggests that incorporating prior information from the breast cancer gene selection provided a small, but measurable, improvement in out-of-sample performance. The learned θ_1 being close to zero indicates that the model did not heavily differentiate between prior-selected and other genes, which is consistent with the earlier observation (shown in Figure 10) that the shared signal between breast and liver cancer appears weak.

Compared to the standard LASSO implementation using `cv.glmnet`, which achieved a test AUC of 0.637 (see Table 6a), our multi-penalty LASSO led to a modest improvement in performance, reaching an AUC of 0.6538. In addition to the performance gain, our method resulted in a more parsimonious model, selecting only 21 genes compared to 125 genes selected by the standard LASSO. Among these, the following genes in Table 7 were common to both models:

Gene ID	Coefficient
ENSG00000136371	-0.00512
ENSG00000174326	-0.00399
ENSG00000184350	0.00189
ENSG00000197245	-0.00391
ENSG00000221598	0.00273
ENSG00000226372	0.00008
ENSG00000233337	-0.00368
ENSG00000237453	-0.00138
ENSG00000243694	0.00191
ENSG00000251015	-0.00054
ENSG00000266373	0.00277
ENSG00000268883	0.00175
ENSG00000271715	0.00819
ENSG00000279307	0.01497

Table 7: Common genes between multi-penalty LASSO and standard LASSO, with coefficients from the multi-penalty LASSO model.

Similar results are obtained when regressing on the cancer stages, and are summarized as follows:

Outcome	η_0^*	η_1^*	**AUC**	**Non-Zero Cov.**	**Common Cov.**
Survival status	-2.75	0.25	0.6538	21	14
Cancer state	-4.75	1.25	0.7342	131	25

Table 8: Performance and selection metrics for multi-penalty LASSO models (with grid search).

The optimal parameters were found at $\theta_0 = -4.75$ and $\theta_1 = 1.25$, indicating a stronger weighting for genes previously selected in the breast cancer model. However, unlike the OS outcome, incorporating this prior information did not lead to an AUC improvement. In fact, performance slightly decreased from 0.7398 (global penalty) to 0.7342. The multi-penalty model selected more covariates than the standard LASSO implementation via `cv.glmnet` (131 vs. 73), but it also achieved a higher AUC (0.7342 vs. 0.663, as shown in Table 6b).

The list of common genes and their corresponding coefficients can be found in the Appendix (see Table 11).

Outcome	η_0^*	η_1^*	**AUC**	**Non-Zero Cov.**	**Common Cov.**
Survival status	-4.36	1.65	0.62	139	0
Cancer state	-3.37	-0.72	0.63	131	131

Table 9: Performance and selection metrics for multi-penalty LASSO (with Bayesian optimization).

As an alternative to the grid search strategy, we also applied a Bayesian optimization approach to search for the optimal values of the regularization parameters (η_0^*, η_1^*) in the multi-penalty LASSO framework. Table 9 summarizes the optimal hyperparameters (η_0^*, η_1^*), corresponding AUC values, and the number of selected genes for the classification tasks of survival status and cancer state, as determined using Bayesian-optimized multi-penalty LASSO. For survival status, the optimal configuration $(\eta_0^* = -4.36, \eta_1^* = 1.65)$ resulted in an AUC of 0.62, with 139 genes selected. None of those 139 selected genes were common to both breast and liver cancer. In contrast, for cancer state, the method achieved a slightly higher AUC of 0.63 using $(\eta_0^* = -3.37, \eta_1^* = -0.72)$, selecting 131 genes that were entirely shared across the two cancer types. Despite these results, the penalty-based integration strategy did not provide meaningful improvements in predictive performance compared to baseline methods such as LASSO or adaptive Ridge (see Table 6), which achieved better AUC values without incorporating such penalties.

(ii) BIC

Next, we present and analyze the results from the BIC-based modeling. The η_0 and η_1 values are selected via a grid search of candidate values. η_0 values were chosen from a dense sequence between -10 and -2, incremented by 0.25, while η_1 ranged from -2 to 2 in steps of 0.25.

Table 10 displays the results attempting to fit η_0 and η_1 by minimizing the BIC.

Outcome	η_0^*	η_1^*	**BIC**
Survival status	−8.00	0.00	148316.4
Cancer state	−8.25	0.00	148316.4

Table 10: Optimal hyperparameters (η_0, η_1) and corresponding BIC for liver cancer survival status and cancer state.

We observe strong η_0 baseline penalization in both models. In effect, the majority of the coefficients are being heavily shrunk towards 0. The BIC dictates that the best fitting models are sparse with very few covariates. At the same time, the optimal η_1 values for both models was set at 0. In effect, the gene-specific penalities are discarded from the final model and only the global η_0 penalty is applied. The transfer learning from the promising genes in breast cancer had no additional explanatory power in the liver cancer context. In line with our previous data integration results, minimizing the BIC suggests that genes that may be important to breast cancer do not generalize well to liver cancer.

Interestingly enough, we observe the exact same BIC values for the two model variants within the same dataset. In other words, after accounting for the gene priors and the penalized prior covariates, the two model variants perform equally well. It demonstrates that survival is highly correlated with stage across datasets. Generally speaking, the η_1 values across both the survival status and cancer state models indicate that weighting for specific genes did not identify any signal for survival or pathological stage across liver or breast cancer.

(iii) Validation Case Study: Using Stage I/II to Predict Stage III/IV in Liver Cancer

We did not find any meaningful correlation between gene effects in breast and liver cancer. As a result, when trying to fit a LASSO model that would penalize genes shared by both cancers less, the model performed poorly. Penalization either became larger or was set to zero, which reflects the lack of an explanatory relationship and therefore no transferable information from one cancer type to the other.

To verify that the method and the reasoning behind it are valid, we applied the same approach to a different study where we expected better results. Specifically, we split the liver cancer dataset into two groups: patients in stage I/II and patients in stage III/IV. External information from the breast cancer dataset is not considered. We then used the coefficients selected by the best liver cancer model for stage I/II to predict outcomes in the stage III/IV group. This setup allows us to check whether our method can successfully transfer information when a real underlying correlation exists.

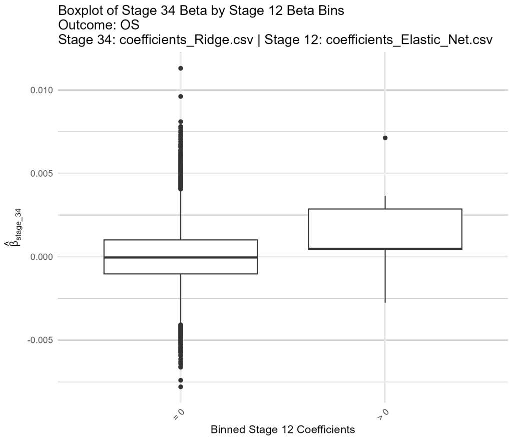


Figure 11: Survival Status correlation between Stage III/IV and Stage I/II coefficient for liver cancer.

The boxplot shows the correlation between the coefficients of genes selected by Elastic Net in Stage I/II and their corresponding coefficients obtained from Ridge regression in Stage III/IV. Although Elastic Net selected only six genes with positive coefficients in Stage I/II, the same genes also exhibit predominantly positive coefficients in Stage III/IV according to Ridge regression. This positive correlation suggests that these genes may play a relevant role in both early and advanced stages of liver cancer, supporting the biological plausibility of the selected features and the potential transferability of information between these patient groups.

To further investigate the correlation between the models, we applied the multi-penalty LASSO regression model, designed to prioritize features previously selected in stage I/II liver cancer. We found the optimal parameter combination at $\theta_0 = -2.75$ and $\theta_1 = -0.75$. To assess whether learning θ_1 meaningfully improves predictive performance, we compared the test AUC using the learned value ($\theta_1 = -0.75$) versus a global penalty approach ($\theta_1 = 0$). The test AUC decreased slightly from 0.5844 to 0.5801 (AUC improvement: –0.0043), suggesting that incorporating prior coefficients from early-stage liver cancer did not lead to measurable gains in out-of-sample performance. However, when compared to a standard LASSO model implemented with cv.glmnet, which yielded a lower AUC of 0.556, our multi-penalty method still provided a modest improvement. Additionally, our model was more sparse, selecting 49 coefficients compared to the 80 selected by standard LASSO. A list of all six model performances and number of covariates selected can be found in Appendix Table 12. Notably, of the six genes previously identified by the Elastic Net model in stage I/II (and which showed a consistent directional signal in the Ridge regression of stage III/IV, as illustrated in Figure 11), only one gene, ENSG00000250969, was retained in the final adaptive model. As of now, this gene is not recognized as a well-established cancer-related gene, and it does not appear among classical oncogenes or tumor suppressors, indicating the potential for novel biomarker discovery but also highlighting the need for further validation.

These findings suggest that while the methodology and its implementation were correct, the predictive power of the selected features was limited. This could be due to a weak signal in the data, insufficient overlap in gene relevance between stages, or the need for additional informative covariates to improve performance. Given these constraints, we opted not to pursue further optimization using Bayesian methods. Instead, we proceed to the next chapter, where we explore alternative strategies for improving model generalization and interpretability.

(e) Bayesian Statistical Inference

During the initial stages of this research, we explored penalized regression methods using integrated gene expression data from multiple platforms. While data integration increased the overall sample size, it did not yield substantial improvements in predictive performance, as measured by the AUC. The modest gains observed were insufficient to justify the additional complexity introduced by gene penalization. Consequently, we chose to replicate the procedure, following the approach used in the previous validation case study, within a Bayesian framework, focusing exclusively on liver cancer data in relation to cancer stage.

Given the high dimensionality of the transcriptomic data, we applied an initial filtering step to retain a subset of 5,000 genes. To incorporate prior knowledge, 95 of these genes (promising genes) were selected from the Elastic Net model on predicting overall survival for liver cancer patients in stages III and IV (see Table 12). While the Adaptive Ridge (AUC = 0.560) had a slightly higher AUC than the Elastic Net (AUC = 0.558), we elected to use the Elastic Net model in order to perform variable selection and extract genes which were more likely to have higher predictive power. The remaining 4,905 genes (non-promising genes) were selected based on their highest variance across samples. Including all genes in the analysis led to computational challenges; therefore, this commonly adopted approach mitigates computational burden while concentrating the analysis on the most variable and potentially informative features.

We then applied a variational Bayesian logistic regression model using the `varbvs` algorithm, with uniform prior inclusion probabilities set at $\pi_j^{(0)} = 0.5$ for all genes. This initial model produced posterior inclusion probabilities (PIPs) for each gene, which represent the inferred probability that a given gene is associated with the outcome.

Based on that posterior inclusion probabilities, the average PIP within each group was computed and then used as an empirical estimate of the group-specific prior inclusion probability, yielding:

$$\text{Avg PIP}_{\text{promising}} = 0.534, \quad \text{Avg PIP}_{\text{non-promising}} = 0.484. \tag{74}$$

We observe that the "promising gene" PIP is approximately 0.052 greater than that of the "non-promising gene". In other words, the "promising" genes have a 5.2% greater chance of being selected in the final model than the "non-promising" genes. These updated priors were incorporated into a second round of variational inference. Specifically, each gene was assigned a prior inclusion probability based on its group membership, and the model was re-estimated under this structured prior scheme. To enhance interpretability and prioritize candidates for downstream biological validation, we selected genes with second-step posterior inclusion probabilities greater than 0.5, following the two-step Bayesian variable selection model. These genes are considered to have strong evidence for association with our outcome. This two-stage procedure enables principled incorporation of empirical evidence into prior beliefs, enhancing model adaptivity while maintaining computational efficiency.

This approach led to a much higher AUC than we had observed previously, with an AUC of 0.7439 on the test set with 80 selected gene covariates. For reference, the AUC of the baseline Elastic Net model on predicting liver cancer survival was 0.558 with 95 covariates. The Bayesian approach renders an increase of 0.1859 to the AUC, which validates our Bayesian variable selection framework. 80 genes had a PIP greater than 0.5, with no gene having a PIP higher than 0.65. In the appendix, Table 13 lists the subset of genes exceeding this 0.5 threshold.

To contextualize these findings biologically, we reviewed the functional annotations of the high-PIP genes. Several genes appear to be involved in key processes related to ischemic stroke progression, including inflammation, immune signaling, and vascular remodeling, suggesting that the structured

Bayesian approach not only enhances statistical interpretability but also yields biologically meaningful results.

Of the 80 genes with PIP values greater than 0.5, we discuss 5 genes which we identified as having strong empirical association or functional relevance with liver cancer below:

- ENSG00000148082 (SHC3): SHC3 had a PIP of 0.62. SHC3 is a protein coding gene which is responsible for enabling phosphotyrosine residue binding activity [11]. It acts as a bridge between growth factor receptors and the Ras/MAPK pathway, which helps cells respond to external signals. The Ras/MAPK is an important signaling pathway that regulates cell proliferation, survival, and differentiation. SHC3 has been as a contributor to tumor progression in hepatocellular carcinoma (HCC). Liu et al. 2021 [30] observed significant SHC3 overexpression in HCC tissues and cell lines. At the same time, therapetic techniques with SHC3 cell inhibition restored HCC cell sensitivity to normal levels. The study outlines the potential of SHC3 therapeutic techniques to aid in the clinical treatment of HCC.

- ENSG00000004948 (CALCR): CALCR had a PIP of 0.61. CALCR is a G protein-coupled Calcitonin receptor that is involved in maintaining calcium homeostasis and in regulating bone resorption [12]. CALCR has been identified has a significant prognostic marker for LIHC. Wang et al. (2024) identified that patients with higher CALCR expression exhibited shorter overall survival (OS) and disease-specific survival (DSS) [29]. Furthermore, cells with suppressed CALCR expression were less aggressive and less likely to lead to die. CALCR is a dual-use biomarker, both as a therapeutic target and a predictor of how patients will benefit from immunotherapy.

- ENSG00000108433 (GOSR2): GOSR2 had a PIP of 0.57. GOSR2 encodes a part of the Golgi SNAP receptor complex, which is essential for transporting vesicles between the endoplasmic reticulum and the Golgi apparatus. This transport system is important for processing and secreting proteins—functions often altered in cancer cells to support their uncontrolled growth. In hepatocellular carcinoma, higher levels of GOSR2 have been linked to worse patient outcomes [28]. More recent research has identified GOSR2 as part of a gene signature related to ferroptosis, a type of iron-dependent cell death. This signature predicts lower overall survival and resistance to sorafenib, a common drug used to treat advanced HCC [20]. Since ferroptosis plays a role in HCC progression and drug resistance, the involvement of GOSR2 suggests it may help cancer cells avoid this form of cell death, making it a promising marker for prognosis and a potential target for therapy.

- ENSG00000214711 (CAPN14): CAPN14 had a PIP of 0.56. CAPN14 is predicted to enable calcium-dependent cysteine-type endopeptidase activity[2]. Dai et al. (2023) conducted a bioinformatics study where they studied key factors in the progression and prognosis of HCC[10]. CAPN14 was found to be up-regulated in tumor tissues, and that it was an independent risk factor for poorer overall survival prognosis. It was identified to belong to a "high-risk" cluster of genes, which collectively showed enriched B-cell infiltration and demonstrated evidence of suppressing immune system processes. Further research into CAPN14's relevance into liver cancer can build upon these findings.

- ENSG00000104835 (MMP14) MMP14 had a PIP of 0.55. MMP14 is involved in the deconstruction of extracellular matrix in physiological processes such as embryonic development and reproduction. It also plays a role in disease processes such as artritis and metastisis [23]. Jin et al. (2020) [18] identified that elevated MMP14 expression levels were highly predictive of poor prognosis in HCC. Furthermore, MMP14 expression was positively correlated with tumor size. MMP14 was negatively associated with both overall and disease-free survival amongst the tested cohort. By having relevance in both tumor size and survival, MMP14 is clearly highly prognostic in liver cancer cases, and further research should be conducted to see how downregulating MMP14 expression levels can offer potential treatment solutions.

Generally speaking, there is a lack of understanding of the true role these genes play in liver cancer development, and regarding how therapeutic techniques can be developed to curb tumor growth. Understanding how to use expression patterns into predictive or therapeutic strategies remains limited. While there seems to be some altered expression or weak associations with prognosis, there is a lack of clinical review and robust studies. Further experimentation and research should be conducted to assess the true biological relevance of these genes in liver cancer, and to grow in the understanding of how to translate that knowledge into actionable strategies that can improve prognosis.

9 Conclusion

This paper explored how advanced statistical techniques can be used to identify important genes that could help predict survival or cancer stage in liver and breast patients. Within that framework, we explored how Bayesian methods can assign statistical priors from breast cancer to see if it improves performance in liver cancer models.

We employed model techniques such as Ridge, LASSO, Elastic Net, other Adaptive frameworks, and Bayesian variable selection using variational inference. There was no single model type that performed best in all test cases. Within our model selection framework, we balanced the objective of balancing the AUC, while also ensuring that we selected models with picked a sufficient number of non-zero gene covariates. Model selection depended heavily on the cancer type, classification task, and feature selection.

- For breast cancer, the Adaptive Elastic Net provided the best results for predicting patient survival (AUC = 0.636 with 148 covariates). Knowledge of the pathological stage led to a marginal improvement in the AUC (AUC = 0.644 with 146 covariates). For predicting pathological stage, the best models were the Ridge and Adaptive Ridge models, respectively (AUC = 0.633 for Ridge, AUC = 0.627 for Adaptive Ridge).
- For liver cancer, the LASSO and Adaptive LASSO models had the strongest performance for predicting patient survival (AUC = 0.637 with 125 covariates). Knowledge of the pathological stage slightly boosted the AUC (AUC = 0.651 with 124 covariates). As with the breast cancer models, the Ridge and Adaptive Ridge model variants were the best at predicting pathological stage (AUC = 0.692 for Adaptive Ridge, AUC = 0.684 for Ridge). We made the decision to discard the models with patient metadeta due to the models downweighting gene significance too much in favor of metadata.

Both the breast and liver cancer datasets favored the Ridge-based models for predicting pathological stage. The Ridge models performing the best indicates that there may not be specific, dominant genes that are predictive of pathological stage. Meanwhile, the differing model selection for survival status among breast and liver cancer hints at a key discrepancy amongst the gene significance between liver and breast cancer. There is no clear and obvious modeling strategy for accurate prognostics in overall survival and stage classification across cancers.

Next, we attempted to use the breast cancer results to set a prior as to which genes should be prioritized in the liver cancer models. We used the coefficients from the Adaptive Lasso model (with Ridge initialization) applied to breast cancer, ranked them by absolute value, and removed the bottom 25% to form a more informative and manageable set of prior genes. However, we observed a very low correlation amongst the gene affects in breast and liver cancer. Our initial aim was to balance maximizing predictive performance (AUC) with selecting a feasible number of gene covariates. Unfortunately, this exercise revealed a very weak correlation amongst the gene effects in liver and breast cancer on both survival status and cancer state.

Both the grid search and Bayesian optimization methods confirmed that our data integration method was performed correctly, as η_1 gene coefficient values were close to 0. Furthermore, the breast cancer priors did not consistently lead to meaningful improvements in the liver cancer AUC. In some cases, the AUC declined or there were no shared gene covariates between the breast cancer prior and the liver cancer model. Learning η_0 and η_1 through minimizing the BIC told the same story. η_1, which dictates the specific weight applied to each selected breast cancer gene, was equal to 0 in both the survival status and stage classification models. In other words, the breast cancer prior information was completely ignored in the liver cancer model.

To verify whether our statistical methodology was functional, we elected to partition the liver cancer set into two categories of stage I/II and stage III/IV. Instead of using prior knowledge from one cancer to predict outcomes in a separate cancer, we used gene data from early-stage patients (stage I/II) to predict outcomes for late-stage patients (stage III/IV) in liver cancer. The multi-penalty LASSO reached a higher AUC than the standard LASSO model, and selected 49 genes compared to the 80 genes selected in the standard LASSO model. These findings demonstrate that with sufficient biological overlap, the transfer learning frameworks we have developed can be successful.

We then set out to determine whether a two-step Bayesian variable selection framework could outperform the traditional penalized regression methods for liver cancer overall survival outcomes. We shrunk the gene sample to 5,000 genes, then implemented a uniform prior across all 5,000 genes. Then, we implemented an empirical prior based on the prior inclusion probability based on whether the gene coefficient was "promising" based on the Elastic Net model from the validation case study of using stage I/II to predict stage III/IV in liver cancer. The two step variational approach lifted the AUC up to 0.7439. 80 genes were found to have a PIP greater than 0.5. SHC3, CALCR, GOSR2, CAPN14, and MMP14 were among those genes, all of which have empirical evidence linking them to cancer significance. The within-liver cancer improvement in model performance indicated that Bayesian variable selection can successfully capture biological trends. However, our analysis is limited due to only being able to test on a 5,000 gene sample due to computational cost concerns.

Despite our statistical methodology, the biological nature of liver and breast cancer are seemingly too uncorrelated for us to have uncovered any novel ideas within our transfer learning approach. No matter how we tweaked with different model variants and different statistical techniques, there is only so much that can be done in the absence of a shared signal between two datasets. The gene effects in breast and liver cancer were simply too different for transfer learning to be successful. However, we were able to report a high AUC for the two-step Bayesian model. This indicates that the Bayesian framework was sound, but the lack of connection between the liver and breast cancer data inhibited the performance of the Bayesian modeling.

Future research maybe more effective to use this type of data integrative modeling between cancers that are more biologically similar or correlated. Adding other data types such as DNA methlyation or protein expression may also help uncover deeper connections across different diseases. While we demonstrated that our statistical methodology could work within a single cancer type, we learned the hard way that inputing gene-level prior knowledge from one disease to another is easier said than done. Successful modeling is highly dependent on the shared biology between diseases.

10 Appendix

This appendix presents the mathematical foundations and key derivations underlying the variational inference techniques, Bayesian optimization and the Receiver Operating Characteristic curve utilized in this work, along with relevant empirical results. We begin by reparameterizing the Receiver Operating Characteristic curve to improve interpretability and derive a suitable expression for computing the Area Under the Curve as a function defined over the unit interval. We then analyze Bayesian optimization, which will be used for hyperparameter tuning of $\eta = (\eta_0, \eta_1)$.

Next, we review the Kullback–Leibler (KL) divergence, a measure of discrepancy between probability distributions that forms the basis for variational approximations. We then introduce the Evidence Lower Bound (ELBO), a tractable objective function whose maximization is equivalent to minimizing the KL divergence. Furthermore, we discuss the use of a quadratic lower bound on the logistic likelihood, as proposed by Jaakkola and Jordan [16], which enables efficient variational optimization in Bayesian logistic regression models with spike-and-slab priors.

Finally, we present empirical results, including tables that summarize the genes selected by both the within-liver cancer model and the two-step Bayesian variable selection method, to support and illustrate the proposed methodology.

(a) Alternative Parametrization of the ROC Curve

We propose an alternative reparameterization of the ROC curve based on [26] and [7] that allows us to define a mapping from the unit interval to itself, rather than the original mapping in Definition 1.

Proposition 3. Let $D \in \{0,1\}$ denote the true class label, where $D = 1$ corresponds to the positive class and $D = 0$ to the negative class. Let Z_1 and Z_0 be absolutely continuous random variables representing the classifier scores for observations from the positive and negative classes, respectively, with cumulative distribution functions G and F. For a given threshold c, let $S(c)$ denote the sensitivity and $E(c)$ denote the specificity. Then, the ROC funtion can be defined as:

$$\begin{aligned} \mathrm{ROC_f} : [0,1] &\longrightarrow [0,1], \\ t &\longmapsto \mathrm{ROC_f}(t) = 1 - G\big(F^{-1}(1-t)\big). \end{aligned} \tag{75}$$

where $F^{-1}(t) = \inf\big\{u \in \mathbb{R} \mid F(u) \geq t\big\}$.

Proof. Starting from the parametric form of the ROC curve in Definition 1 and setting $t = 1 - F(c)$, then $c = F^{-1}(1-t)$, since Z_0 is absolutely continuous. Substituting into the second component of the ROC parametric mapping (21) yields in

$$1 - G(c) \;=\; 1 - G\big(F^{-1}(1-t)\big). \tag{76}$$

□

(b) Foundations of Bayesian Optimization

In this section, we present the mathematical formulation of Bayesian optimization, aimed at selecting the optimal hyperparameters $\eta = (\eta_0, \eta_1)$. The central objective is to model an unknown objective function f, which in our context corresponds to the cross-validated AUC, as a realization from a Gaussian process. This probabilistic framework enables efficient exploration and exploitation of the hyperparameter space to identify configurations that yield superior predictive performance.

Given initial data $\mathcal{D}_t = \{(\eta^{(i)}, y_i)\}_{i=1}^t$ and by defining $k_t(\eta) = \big[k(\eta, \eta^{(i)})\big]_{i=1}^t$, the posterior predictive distribution at any $\eta \in \Theta$ is

$$f(\eta) \mid \mathcal{D}_t \ \sim\ \mathcal{N}\big(\mu_t(\eta),\, \sigma_t^2(\eta)\big), \tag{77}$$

with

$$\begin{aligned} \mu_t(\eta) &= m(\eta) + k_t(\eta)^\top \big(K_t + \sigma_n^2 I\big)^{-1} \big(y_{1:t} - m(\eta_{1:t})\big), \\ \sigma_t^2(\eta) &= k(\eta, \eta) - k_t(\eta)^\top \big(K_t + \sigma_n^2 I\big)^{-1} k_t(\eta). \end{aligned} \tag{78}$$

where $K_t = [k(\eta^{(i)}, \eta^{(j)})]_{i,j=1}^t$ is the kernel matrix, $m_{1:t} = \big[m(\eta^{(i)})\big]_{i=1}^t$ is the vector of prior means, and σ_n^2 denotes the variance of the observation noise.

We then define an acquisition function $a_t(\eta)$, that guides the selection of the next hyperparameter by balancing exploitation, which targets points where the surrogate's predictive mean $\mu_t(\eta)$ is high and thus likely to improve the objective, and exploration, which targets points where the surrogate's predictive standard deviation $\sigma_t(\eta)$ is large and may uncover better optima. We adopt the Upper Confidence Bound (UCB) acquisition function in its heuristic form. At iteration t, UCB is defined by:

$$a_t(\eta) = \mu_t(\eta) + \kappa \sigma_t(\eta), \tag{79}$$

where $\kappa > 0$ balances exploration against exploitation. In our case, we set $\kappa = 2.576$, corresponding to the 99% quantile of the standard normal distribution, exploring points whose predicted value might lie near the top 1% of the uncertainty interval. At each iteration we select the next hyperparameter by maximizing the acquisition function:

$$\eta^{(t+1)} \ = \ \arg\max_{\eta \in \Theta} a_t(\eta). \tag{80}$$

Repeating this process until convergence or running the loop for N total evaluations yields the final estimate, so that:

$$\hat{\eta} \ = \ \arg\max_{\eta \in \Theta} f(\eta). \tag{81}$$

The hyperparameter optimization procedure detailed above is formalized in Algorithm 4. In our experiments, we set the UCB parameter to $\kappa = 2.576$ (the 99% quantile of the standard normal) and initialize the Bayesian optimizer with $n_0 = 25$ randomly sampled evaluations of f.

Algorithm 4 Bayesian Optimization for Hyperparameters $\eta = (\eta_0, \eta_1)$.

```
Input: $\Theta \subset \mathbb{R}^2$ (search domain), objective $f(\eta) = \mathrm{CV{-}AUC}(\eta)$, initial sample size $n_0$, total iterations $N$, exploration parameter $\kappa$.
$\mathcal{D}_{n_0} \leftarrow \{$evaluate $f$ at $n_0$ points$\}$
for $t = n_0, \dots, N-1$ do
    Fit GP surrogate to $\mathcal{D}_t$, obtaining $\mu_t(\eta), \sigma_t(\eta)$.
    Define UCB acquisition $a_t(\eta) = \mu_t(\eta) + \kappa\, \sigma_t(\eta)$.
    Select next hyperparameter
        $\eta^{(t+1)} = \arg\max_{\eta \in \Theta} a_t(\eta)$.
    Evaluate $y_{t+1} = f(\eta^{(t+1)})$.
    Update dataset: $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{(\eta^{(t+1)}, y_{t+1})\}$.
end for
Return: $\hat{\eta} \ = \ \arg\max_{(\eta,y) \in \mathcal{D}_N} y$.
```

(c) Foundations of Variational Inference: Kullback-Leibler Divergence and Evidence Lower Bound

This subsection delineates the mathematical framework foundational to variational inference. It commences with a formal definition of the Kullback–Leibler divergence, elucidates its theoretical interpretation, and subsequently establishes its connection to the Evidence Lower Bound, following the exposition in [5].

Definition 5. Let P and Q be two probability distributions defined on the same probability space $\mathcal{X}$, with density functions $p(x)$ and $q(x)$, respectively. The Kullback-Leibler Divergence from Q to P, denoted $\mathrm{KL}(P \,\|\, Q)$, is defined as:

$$\mathrm{KL}(P \,\|\, Q) = \int_{\mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx. \tag{82}$$

Observation 9. KL divergence quantifies the dissimilarity between two probability distributions. It can be interpreted as the amount of information lost when the distribution Q is used to approximate P. A divergence of zero implies that the two distributions are equal almost everywhere, while larger values indicate greater discrepancy.

To implement the variational algorithm and derive its updates, it is essential to first define a fundamental concept: the Evidence Lower Bound (ELBO).

Definition 6. Let $p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{y})$ denote the joint distribution over the parameters and initial data, and let $q(\boldsymbol{\beta}, \boldsymbol{\gamma})$ be a variational approximation to the posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y})$. The ELBO is defined as:

$$\mathrm{ELBO}(q) = \mathbb{E}_q \left[\log p(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \right] - \mathbb{E}_q \left[\log q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \right]. \tag{83}$$

Maximizing the Evidence Lower Bound over a family of variational distributions q is mathematically equivalent to minimizing the Kullback–Leibler (KL) divergence between q and the true posterior distribution as shown in Proposition 4.

Proposition 4. Let $p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathcal{D})$ denote the true posterior distribution over latent variables given data $\mathcal{D}$, and let $q(\boldsymbol{\beta}, \boldsymbol{\gamma})$ be a variational approximation. Then minimizing the Kullback–Leibler divergence from q to the true posterior is equivalent to maximizing the Evidence Lower Bound.

Proof. The KL divergence between the variational distribution $q(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and the true posterior $p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathcal{D})$ is defined as:

$$\mathrm{KL}(q \,\|\, p) = \int q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \log \frac{q(\boldsymbol{\beta}, \boldsymbol{\gamma})}{p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathcal{D})} \, d\boldsymbol{\beta} \, d\boldsymbol{\gamma}. \tag{84}$$

Using Bayes' theorem, we can write:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathcal{D}) = \frac{p(\mathcal{D}, \boldsymbol{\beta}, \boldsymbol{\gamma})}{p(\mathcal{D})}. \tag{85}$$

Substituting this expression into the KL divergence yields:

$$\mathrm{KL}(q \,\|\, p) = \int q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \left[\log q(\boldsymbol{\beta}, \boldsymbol{\gamma}) - \log p(\mathcal{D}, \boldsymbol{\beta}, \boldsymbol{\gamma}) + \log p(\mathcal{D}) \right] \, d\boldsymbol{\beta} \, d\boldsymbol{\gamma}. \tag{86}$$

Since $\log p(\mathcal{D})$ does not depend on γ or β, we can rewriting the expression as:

$$\mathrm{KL}(q \,\|\, p) = \mathbb{E}_q[\log q(\boldsymbol{\beta}, \boldsymbol{\gamma})] - \mathbb{E}_q[\log p(\mathcal{D}, \boldsymbol{\beta}, \boldsymbol{\gamma})] + \log p(\mathcal{D}). \tag{87}$$

Rearranging terms, we find:

$$\log p(\mathcal{D}) = \mathrm{KL}(q \,\|\, p) + \mathrm{ELBO}(q). \tag{88}$$

Since $\log p(\mathcal{D})$ does not depend on q, minimizing the KL divergence $\mathrm{KL}(q \,\|\, p)$ is equivalent to maximizing the ELBO:

$$q^* = \arg\min_q \mathrm{KL}(q \,\|\, p) \quad \Longleftrightarrow \quad q^* = \arg\max_q \mathrm{ELBO}(q). \tag{89}$$

□

(d) Jaakkola and Jordan's Bound for Variational Optimization

In variational inference for Bayesian logistic regression, a central difficulty lies in computing the expected log-likelihood under the variational distribution. Specifically, for a logistic likelihood of the form:

$$\log p(\mathbf{y} \mid \boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \log \sigma(\mathbf{x}_i^\top \boldsymbol{\beta}) + (1 - y_i) \log(1 - \sigma(\mathbf{x}_i^\top \boldsymbol{\beta})) \right], \tag{90}$$

the expectation $\mathbb{E}_q[\log p(\mathbf{y} \mid \boldsymbol{\beta})]$ does not admit a closed-form expression when the variational distribution $q(\boldsymbol{\beta})$ is a mixture with Gaussian components, as in spike-and-slab models. To address this intractability, `varbvs` employs the bound introduced by Jaakkola and Jordan [16], which provides a quadratic lower bound on the logistic log-likelihood:

$$\log \sigma(t) \;\geq\; -\lambda(\xi)\, t^2 \;+\; \frac{t}{2} \;-\; \psi(\xi), \tag{91}$$

where $\lambda(\xi) = \frac{1}{2\xi} \tanh(\xi/2)$ and $\psi(\xi) = -\frac{\xi}{2} + \log(1 + e^{\xi}) - \lambda(\xi)\xi^2$. The variational parameter ξ_i is introduced for each observation i and is optimized alongside other variational parameters . This bound replaces each $\log \sigma(\mathbf{x}_i^\top \boldsymbol{\beta})$ term in the likelihood with a concave quadratic approximation in $\boldsymbol{\beta}$, allowing the entire likelihood contribution to the ELBO to be approximated in closed form under $q(\boldsymbol{\beta})$. This formulation renders the evidence lower bound both tractable and amenable to optimization.

(e) Cancer-stage common genes table

The table lists the genes selected for liver cancer (predicting cancer stage) by both standard LASSO and by the adapted LASSO method.

Gene ID	Coefficient
ENSG00000104490	-2.07e-04
ENSG00000135702	3.58e-03
ENSG00000137869	1.15e-03
ENSG00000155428	-1.36e-03
ENSG00000160870	-1.42e-04
ENSG00000181511	2.24e-03
ENSG00000188508	3.48e-03
ENSG00000198216	-8.98e-04
ENSG00000215186	3.73e-03
ENSG00000229941	8.07e-04
ENSG00000230074	3.41e-03
ENSG00000230290	-3.58e-03
ENSG00000232876	-2.44e-03
ENSG00000233005	1.81e-03
ENSG00000234639	1.16e-03
ENSG00000251299	-8.73e-04
ENSG00000252929	1.14e-03
ENSG00000253190	-1.42e-03
ENSG00000255883	3.07e-03
ENSG00000259868	-3.24e-03
ENSG00000264940	3.33e-03
ENSG00000273444	2.27e-03
ENSG00000282301	-1.85e-03
ENSG00000287189	-2.61e-03
ENSG00000287928	-4.07e-03

Table 11: Common genes between multi-penalty LASSO method and Standard LASSO with their coefficients.

(f) Model results or predicting overall survival in the liver cancer dataset stages III and IV

Method	AUC	Non-Zero Covariates
Adaptive Ridge	0.560	33,004
Elastic Net	0.558	95
Lasso	0.556	80
Ridge	0.498	33,004
Adaptive Elastic Net	0.453	46
Adaptive Lasso	0.447	43

Table 12: Model results for predicting overall survival in the liver cancer dataset stages III and IV, sorted by descending AUC.

(g) Bayesian-selected genes and their Posterior Inclusion Probabilities

The table lists the genes selected for predicting overall survival in liver cancer using a Bayesian variable selection model with a two-step prior. The first prior step is a uniform prior fit using the `varbvs` package, and the resulting posterior inclusion probabilities (PIPs) informed the priors for a second model. Genes with a PIP value greater than 0.5 (i.e. stronger evidence of gene relevance within Bayesian framework) are shown below:

Gene ID	PIP	Gene ID	PIP	Gene ID	PIP
ENSG00000260990	0.65	ENSG00000147206	0.53	ENSG00000208037	0.52
ENSG00000250483	0.65	ENSG00000135100	0.53	ENSG00000260047	0.51
ENSG00000287446	0.63	ENSG00000169040	0.53	ENSG00000100344	0.51
ENSG00000148082	0.62	ENSG00000255062	0.53	ENSG00000162039	0.51
ENSG00000004948	0.61	ENSG00000177675	0.53	ENSG00000254934	0.51
ENSG00000089169	0.61	ENSG00000232352	0.53	ENSG00000159723	0.51
ENSG00000202229	0.60	ENSG00000170423	0.53	ENSG00000242737	0.51
ENSG00000182586	0.59	ENSG00000108298	0.53	ENSG00000263955	0.51
ENSG00000278383	0.58	ENSG00000221882	0.53	ENSG00000253429	0.51
ENSG00000250387	0.57	ENSG00000237921	0.53	ENSG00000268278	0.51
ENSG00000108433	0.57	ENSG00000221265	0.53	ENSG00000261594	0.51
ENSG00000224473	0.56	ENSG00000253971	0.52	ENSG00000253622	0.51
ENSG00000253954	0.56	ENSG00000261377	0.52	ENSG00000213512	0.51
ENSG00000283959	0.56	ENSG00000255622	0.52	ENSG00000276545	0.51
ENSG00000214711	0.56	ENSG00000225912	0.52	ENSG00000161270	0.51
ENSG00000189325	0.55	ENSG00000079689	0.52	ENSG00000234880	0.51
ENSG00000104835	0.55	ENSG00000169894	0.52	ENSG00000241877	0.51
ENSG00000212332	0.55	ENSG00000237758	0.52	ENSG00000257927	0.51
ENSG00000267496	0.55	ENSG00000168334	0.52	ENSG00000242020	0.51
ENSG00000198910	0.55	ENSG00000183134	0.52	ENSG00000183586	0.51
ENSG00000230362	0.54	ENSG00000248859	0.52	ENSG00000231185	0.50
ENSG00000254972	0.54	ENSG00000219642	0.52	ENSG00000222465	0.50
ENSG00000121871	0.54	ENSG00000010219	0.52	ENSG00000221947	0.50
ENSG00000185674	0.54	ENSG00000227748	0.52	ENSG00000131737	0.50
ENSG00000227550	0.53	ENSG00000140443	0.52	ENSG00000095932	0.50
ENSG00000228286	0.53	ENSG00000255474	0.52	ENSG00000279570	0.50
ENSG00000211945	0.53	ENSG00000251972	0.52		

Table 13: Bayesian-selected genes with Posterior Inclusion Probabilities (PIPs) > 0.5 (ordered column-wise).

References

[1] Laura J. van 't Veer et al. "Gene expression profiling predicts clinical outcome of breast cancer". In: *Nature* 415.6871 (2002), pp. 530–536. DOI: 10.1038/415530a. URL: https://doi.org/10.1038/415530a.

[2] AllianceGenome Consortium. *CAPN14 (HGNC 16664) Gene Detail Page*. Accessed 2 June 2025. 2025. URL: https://www.alliancegenome.org/gene/HGNC:16664.

[3] Marcia M. Barbieri and James O. Berger. "Optimal Predictive Model Selection". In: *The Annals of Statistics* 32.3 (2004), pp. 870–897. DOI: 10.1214/009053604000000238.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Chapter 1.2. New York: Springer, 2006.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. 1st. New York: Springer, 2006. ISBN: 978-0-387-31073-2.

[6] Philip J. Brown, Marina Vannucci, and T. Fearn. "Multivariate Bayesian Variable Selection and Prediction". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.3 (1998), pp. 627–641.

[7] Camilla Calı‘ and Maria Longobardi. "Some mathematical properties of the ROC curve and their applications". In: *Ricerche di Matematica* 64 (2015), pp. 391–402.

[8] Peter Carbonetto and Matthew Stephens. "Scalable variational inference for variable selection in regression, and its accuracy in genetic association studies". In: *Bayesian Analysis* 7.1 (2012), pp. 73–108.

[9] Ting-Huei Chen et al. "A Penalized Regression Framework for Building Polygenic Risk Models Based on Summary Statistics from Genome-Wide Association Studies and Incorporating External Information". In: *Journal of the American Statistical Association* 116.533 (2021). Epub 2020 Oct 12, pp. 133–143. DOI: 10.1080/01621459.2020.1764849.

[10] Dongjun Dai et al. "Novel insights into the progression and prognosis of the calpain family members in hepatocellular carcinoma: a comprehensive integrated analysis". In: *Frontiers in Molecular Biosciences* 10 (July 2023), p. 1162409. ISSN: 2296-889X. DOI: 10.3389/fmolb.2023.1162409. URL: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1162409.

[11] GeneCards Human Gene Database. *SHC3 Gene - SHC Adaptor Protein 3*. Accessed: 2025-05-31. 2024. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=SHC3.

[12] GeneCards, The Human Gene Database. *CALCR Gene - Calcitonin Receptor*. https://www.genecards.org/cgi-bin/carddisp.pl?gene=CALCR. Accessed: 2025-05-26. 2025.

[13] Jayanta K. Ghosh. "Bayesian model selection using the median probability model". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 7.2 (2015), pp. 124–134.

[14] Yongtao Guan and Matthew Stephens. "Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems". In: *The Annals of Applied Statistics* 5.3 (2011), pp. 1780–1815. DOI: 10.1214/11-AOAS455.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Accessed: 2025-05-26. Springer, 2009. URL: https://web.stanford.edu/~hastie/ElemStatLearn/.

[16] T. S. Jaakkola and M. I. Jordan. "Bayesian parameter estimation via variational methods". In: *Statistics and Computing* 10 (2000), pp. 25–37.

[17] Lin Ji et al. "Risk and prognostic factors of breast cancer with liver metastases". In: *BMC Cancer* 21 (2021), p. 238. DOI: 10.1186/s12885-021-07968-5.

[18] Ye Jin et al. "High MMP14 Expression Is Predictive of Poor Prognosis in Resectable Hepatocellular Carcinoma". In: *Pathology* 52.3 (Apr. 2020), pp. 359–365. ISSN: 0031-3025. DOI: 10.1016/j.pathol.2020.01.436. URL: https://pubmed.ncbi.nlm.nih.gov/32122646/.

[19] T. A. Lasko et al. "The use of receiver operating characteristic curves in biomedical informatics". In: *Journal of Biomedical Informatics* 38.5 (2005), pp. 404–415. DOI: 10.1016/j.jbi.2005.02.001.

[20] C. Li et al. "Identification of ferroptosis-related gene signature for predicting prognosis and immune response in hepatocellular carcinoma". In: *Frontiers in Pharmacology* 14 (2023), p. 1212719. DOI: 10.3389/fphar.2023.1212719. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10348016/.

[21] Fan Li and Nancy R. Zhang. "Bayesian Variable Selection in Structured High-Dimensional Covariate Spaces with Applications in Genomics". In: *Journal of the American Statistical Association* 105.491 (2010), pp. 1202–1214. ISSN: 0162-1459. DOI: 10.1198/jasa.2010.tm08177. URL: https://doi.org/10.1198/jasa.2010.tm08177.

[22] Ariel Linden. "Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis". In: *Journal of Evaluation in Clinical Practice* 12.2 (2006), pp. 132–139.

[23] National Center for Biotechnology Information. *MMP1 – Matrix Metallopeptidase 1 (Homo sapiens)*. Accessed 2025-06-02. 2025. URL: https://www.ncbi.nlm.nih.gov/gene/4323.

[24] R. Nevola et al. "Gender Differences in the Pathogenesis and Risk Factors of Hepatocellular Carcinoma". In: *Biology (Basel)* 12.7 (July 2023), p. 984. DOI: 10.3390/biology12070984.

[25] Omiros Papaspiliopoulos, David Rossell, and Miquel Torrens-i-Dinarès. *Inference for Multiple Treatment Effects Using Confounder Importance Learning*. 2025. arXiv: 2110.00314 [stat.ME]. URL: https://arxiv.org/abs/2110.00314.

[26] Margaret Sullivan Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.

[27] Xavier Robin et al. "Package 'pROC'". In: *pROC Package* (2021).

[28] H. Wang et al. "Golgi-related gene signature predicts prognosis in hepatocellular carcinoma". In: *BMC Cancer* 21 (2021), p. 1204. DOI: 10.1186/s12885-021-08939-3. URL: https://doi.org/10.1186/s12885-021-08939-3.

[29] Wenlong Wang et al. "CALCR is a potential prognostic biomarker and modulator of immune infiltration in liver hepatocellular carcinoma". In: *Molecular Medicine Reports* 29.6 (2024), p. 13406. DOI: 10.3892/mmr.2024.13406.

[30] Xianbin Zhang et al. "SHC3 promotes drug resistance and cancer stemness by enhancing -catenin signaling in hepatocellular carcinoma". In: *Cell Death & Disease* 15.5 (2024), p. 253. DOI: 10.1038/s41419-024-06420-3.

[31] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.