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

Data Science Methodology

Master Thesis Presentation

Elisa Scocco, Iñigo Exposito, Tirdod Behbehani

Barcelona School of Economics

INTRODUCTION

- Cancer is a brutally painful disease, and is one of the leading worldwide causes of death.

- In 2025, the National Cancer Institute estimates that 2,041,910 new cases of cancer will be diagnosed and 618,120 people will die from the disease in the United States alone.

- Technological advances, such as RNA sequencing (RNA-seq), have now made it possible to capture the gene expression levels of tens of thousands of genes within one snapshot.

- With these RNA-seq technologies, it is now possible to actually study the nuances and intricacies of specific cancer types.

Barcelona School of Economics

PROBLEM

- While the gene expression data lends promise, there are practical difficulties with the model implementation.
- Cohorts of gene expression data is often limited to only a few hundred patient samples, and tens of thousands of genes
- We must be wary of the curse of dimensionality and overfitting!

Barcelona School of Economics

OUR POTENTIAL SOLUTION + TRANSFER LEARNING

- Penalized regression techniques such as LASSO, Ridge regression, Elastic Net, and Bayesian inference are well-suited for high-dimensional data.

- The penalized regression sifts through the thousands of non-predictive genes, and identies a select number of potentially predictive genes.

- **General Approach:** Does gene expression data from a more common cancer (breast cancer) improve prediction performance on cancer outcomes from a less common cancer (liver cancer)?

Barcelona School of Economics

GENERAL APPROACH

- We utilize the aforementioned penalized regression techniques to predict the status (overall survival or pathological stage) of liver cancer patients using RNA-seq data.

- The gene expression data from breast cancer is utilized as prior information in the liver cancer modeling. 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 evaluate whether the external information from breast cancer improves model performance, we apply optimization techniques to maximize the Area Under the Curve (AUC) or minimize the Bayesian Information Criterion (BIC), while maintaining a balance between accuracy and retaining a suitable number of gene covariates.

Barcelona School of Economics

DATA OVERVIEW

- Liver cancer (LIHC) and Breast cancer (BRCA) datasets were collected from The Cancer Genome Atlas (TCGA) through the Genomic Data Commons (GDC):

- Data table structure:

  - Gene expression data: Matrix of genes with gene expression measurements collected via RNA-seq.

  - Survival data: Each row in the data table corresponds to an individual patient, with a binary “OS” column indicating survival or death (0 = survival, 1 = death).

  - Clinical data: Patient-level information level information such as demographic information, diagnostic information, and treatment history.

Barcelona School of Economics

DATA TRANSFORMATION AND FEATURE SELECTION

- The gene expression, survival, and clinical data were merged into one combined data frame as we began our pre-processing and initial data exploration.
- Some pathological stage values were missing, so we filled those missing values using the available information from tumor size, lymph node involvement, and metastasis components.
- A binary variable for whether the patient's cancer was advanced (i.e. in stage III or stage IV) was created.
- Sparsely expressed genes, genes with low variability, and genes with irregular expression distributions were removed from the dataset.
- Final dataset summary:
  - Liver cancer (LIHC): 25,517 genes and 341 patients
  - Breast cancer (BRCA): 33,092 genes and 1,203 patients

Barcelona School of Economics

METHODOLOGICAL FRAMEWORK

PENALIZED LIKELIHOOD METHODS

Let **S** denote a sample of **n** units from a **finite** population and $\mathbf{x_i}$ denote a vector of **p covariates**.

LOGISTIC REGRESSION

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

maximize

LIKELIHOOD

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

In high dimensions

PROBLEMS

- **Overfitting** and **inestability**
- High **variance**

SOLUTION

- Minimize **penalized loss** function
- Use **regularization** term

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

PENALIZED LIKELIHOOD METHODS

LASSO

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

ELASTIC NET

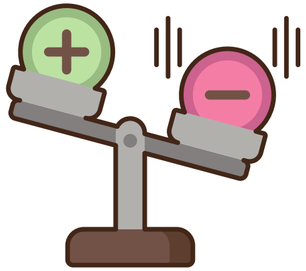
$$\hat{\boldsymbol{\beta}}^{\mathrm{EN}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \right\}$$

RIDGE

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

ADAPTIVE METHODS

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



ADVANTAGES

- Coefficient **specific penalization**
- Reduce **bias**

DISADVANTAGES

- Computationally **costlier:** two optimization problems

Barcelona School of Economics

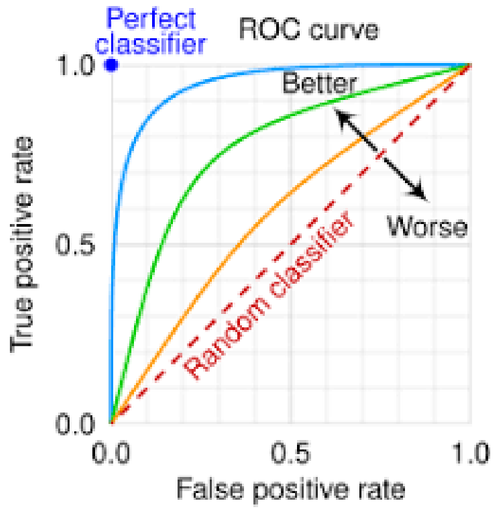
BINARY CLASSIFICATION

CONFUSION MATRIX

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

ROC CURVE

Plots the **True Positive Rate** against the **False Positive Rate** at various threshold settings.



AUC

- Measures the ability of a binary classification model to **distinguish** between the two **classes**
- Perfect classification: AUC= 1
- Random guessing: AUC= 0.5

CROSS VALIDATION

- Split into **train** and **test** set
- Divide the train set in **10 folds**
- Select **hyperparameters** that **maximize** average **AUC** in the train set
- **Predict** on **test** set

Barcelona School of Economics

MULTI PENALTY LASSO

INITIAL CONSIDERATIONS

- Trade off between **predictive performance** and number of **shared covariates**
- **Correlation** between genes (large effects in both cancers)
- In **Ridge** regression: set an appropiate **threshold** to **discard covariates**

IMPLEMENTATION

-Define binary auxiliary variable

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

-Gene specific penalty

$$\log \lambda_j = \eta_0 + \eta_1 \, z_j, \qquad \lambda_j = \exp\left(\eta_0 + \eta_1 \, z_j\right).$$

OPTIMIZATION PROBLEM

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

How can we **choose** the **hyperparameters**?

HYPERPARAMETER TUNNING

1. **MAXIMIZE CV AUC**

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

- GRID SEARCH
- BAYESIAN OPTIMIZATION (Gaussian processes)

2. **MINIMIZE BIC**

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

- k accounts for the number of **non-zero coefficients**
- Trade off between **model fit** and **complexity**
- assigning a non-zero value to η_1 is penalized by the BIC.

BAYESIAN STATISTICAL INFERENCE

Previous information about the parameters expressed through a **prior distribution**

SPIKE AND SLAB PRIOR MODEL

$$
\begin{aligned}
y_i \mid \mathbf{x}_i, \boldsymbol{\beta} &\sim \text{Bern}(\sigma(\mathbf{x}_i^\top \boldsymbol{\beta})), \quad \text{for } i = 1, \ldots, n, \\
\sigma(t) &= \frac{1}{1 + e^{-t}}, \\
\beta_j \mid \gamma_j &\sim (1 - \gamma_j) \cdot \delta_0 + \gamma_j \cdot \mathcal{N}(0, \sigma^2), \quad \text{for } j = 1, \ldots, p, \\
\gamma_j &\sim \text{Bern}(\pi), \quad \text{for } j = 1, \ldots, p
\end{aligned}
$$

Posterior inclusion probability

$$\text{PIP}_j = P(\gamma_j = 1 \mid \mathcal{D})$$

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

Computation of posterior inclusion probability is **intractable**

Need **variational inference**

RESULTS

REGRESSIONS ANALYSIS FINDINGS

The methods used

- LASSO
- Adaptive LASSO with initial LASSO estimates
- Ridge
- Adaptive LASSO with initial Ridge estimates
- Elastic Net
- Adaptive LASSO with initial Elastic Net estimates

The target outcome

- Survival status
- Survival status (with pathological cancer stage as covariate)
- Survival status (with patient metadata information, liver only)
- Pathological cancer stage

Barcelona School of Economics

BREAST CANCER

Survival Status (Without Knowledge of Pathological Stage)

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.

Pathological Stage

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.

Survival Status (With Knowledge of 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.

Barcelona School of Economics

LIVER CANCER

Survival Status (Without Knowledge of Pathological Stage)

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

Pathological Stage

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

Survival Status (With Knowledge of Pathological Stage)

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

Survival Status (With Information on 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

SUMMARY OF REGRESSION RESULTS

For survival prediction, breast cancer models performed best with Adaptive Elastic Net, while liver cancer survival was better predicted with LASSO

In both datasets, Ridge-based models were best for predicting pathological stage

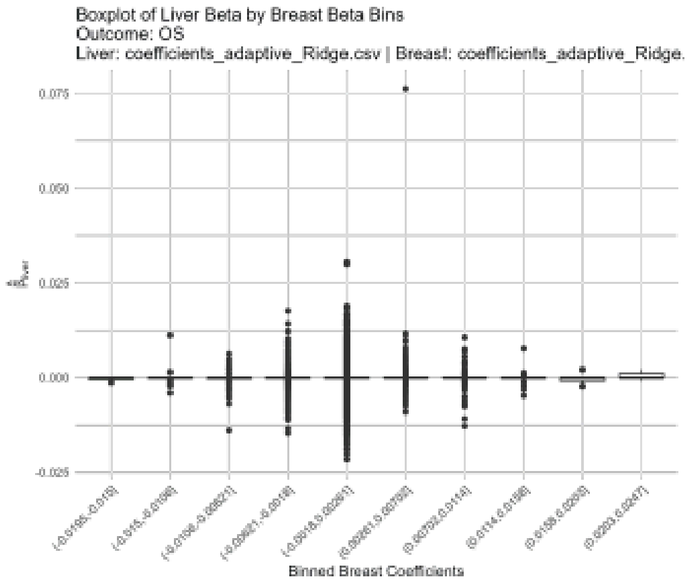
Given the marginal of improvement in performance, we decided to not incorporate metadata covariates with in the subsequent data integration steps

We only proceed with gene covariates. We aim to balance high predictive performance (AUC) with model sparsity (number of selected covariates)

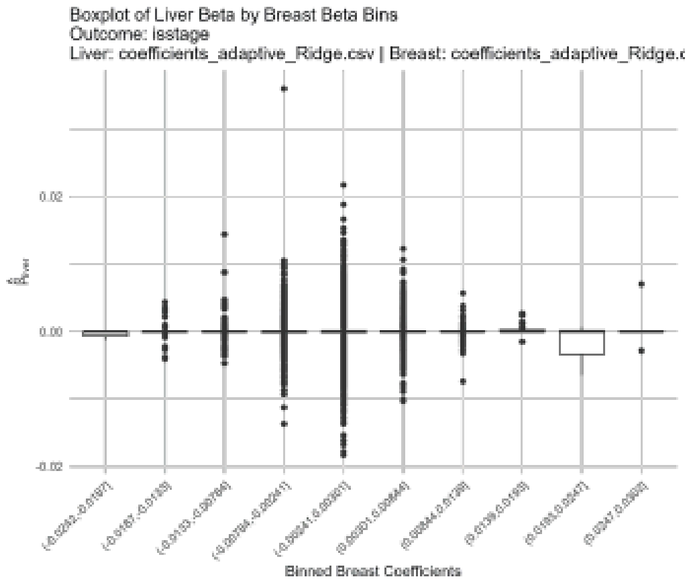
Barcelona School of Economics

DATA INTEGRATION

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.

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

Barcelona School of Economics

MULTI-PENALTY LASSO FOR LIVER CANCER PREDICTION-GRID SEARCH

Approach:

- Used Multi-Penalty LASSO with initial Ridge estimates on breast cancer data.
- Coefficients from breast model used as prior information for liver cancer regression.
- Removed 25% of genes with smallest coefficients in absolute value.

Results:

- AUC improved from 0.637 from standard LASSO to 0.6538 for Survival Status regression.
- AUC increased from 0.663 from standard LASSO to 0.7342 for Cancer Stage regression.
- Common covariates (transfering learning) are 14 for Survival Status and 25 for Stage Prediction

Conclusion:

- Prior-guided selection yields modest yet measurable gain.
- Learned $\eta_1 \approx 0$ suggests weak shared signal between breast and liver cancer.

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

Barcelona School of Economics

MULTI-PENALTY LASSO FOR LIVER CANCER PREDICTION-BIC

Approach:

- Used Multi-Penalty LASSO with initial Ridge estimates on breast cancer data.
- Coefficients from breast model used as prior information for liver cancer regression.
- Removed 25% of genes with smallest coefficients in absolute value.

Results:

- η_0 is heavily penalized in both the survival status and cancer state models.
- η_1 is equal to zero in both the survival status and cancer state models.

Conclusion:

- The BIC dictates that the best fitting models are sparse with very few gene covariates.
- The gene-specific penalities are discarded from the final model, and only the global η_0 is applied.

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

Barcelona School of Economics

VALIDATION CASE STUDY: USING STAGE I/II TO PREDICT STAGE III/IV IN LIVER CANCER

We then test if the method is valid by applying it in a setting where true correlation is expected.

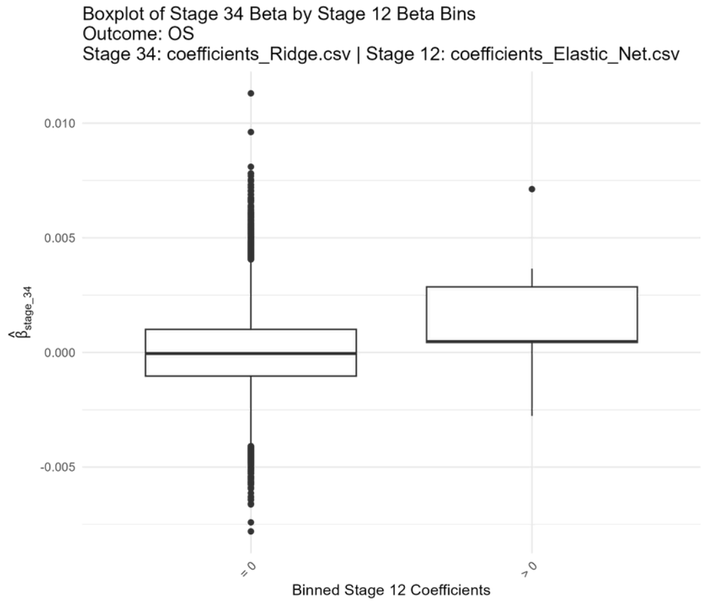
Design:

Liver cancer data split into:

- Stage I/II patients
- Stage III/IV patients

Approach:

- Selected coefficients from the best model trained on Stage I/II
- Applied these to predict outcomes in Stage III/IV patients



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.

Barcelona School of Economics

VALIDATION CASE STUDY: USING STAGE I/II TO PREDICT STAGE III/IV IN LIVER CANCER

Results:

- Compared to standard LASSO (AUC = 0.556), the multi-penalty method still performed better (0.5801)
- Model was more sparse: 49 coefficients selected vs. 80 in standard LASSO
- Of the 6 genes with consistent directional signal, only 1 gene (ENSG00000250969) was retained (not a known oncogene or tumor suppressor)

While the implementation was sound, the predictive power of selected features was limited.
This may reflect weak signals, low overlap in gene relevance, or missing covariates.

Barcelona School of Economics

BAYESIAN STATISTICAL INFERENCE

1. **Initial filtering** (5000 genes):

-**95 promising** genes (from Elastic Net).

-**4905 additional** genes with highest variance

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

2. Initial uninformative **prior** probability of **0.5** for all genes.

3. After running **variational inference**:

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

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

4. We run again the **variational inference** with the group **specific** priors.

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

Model	Number of Covariates	AUC	AUC Improvement
Elastic Net	95	0.558	—
Bayesian Approach	80	0.7439	0.1859

Barcelona School of Economics

NOTABLE GENES FROM BAYESIAN STATISTICAL INFERENCE

- ENSG00000148082 (SHC3):
  - SHC3 had a PIP of 0.62.
  - Liu et al. (2021) found that SHC3 is a contributor to tumor progression in hepatocellular carcinoma (HCC)
    - SHC3 was significantly overexpressed in HCC tissues and cell lines
    - Therapetic techniques which targeted SHC3 restoreed HCC cell sensitivity to normal levels

- ENSG00000004948 (CALCR):
  - CALCR had a PIP of 0.61.
  - Wang et al. (2024) identified that patients with higher CALCR expression exhibited shorter overall survival (OS) and disease-specific survival (DSS)
  - Cells with suppressed CALCR expression were less aggressive and less likely to lead to die.

- ENSG00000104835 (MMP14):
  - MMP14 had a PIP of 0.55.
  - Jin et al. (2020) identifed that elevated MMP14 expression levels were highly predictive of poor prognosis in HCC.
  - MMP14 expression was positively correlated with tumor size and negatively associated with survival.

Barcelona School of Economics

CONCLUSION

CONCLUSION

- Our liver cancer survival and pathological stage models with prior breast cancer information struggled. Prior breast cancer information had either minimal impact on model performance or was completely omitted from the final modeling.

- Our validation study on within-liver cancer progression also indicated that the predictive power of selected features was limited. However, the two-step within-liver cancer Bayesian model did report a higher AUC.

- 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.

- While we only consider gene expression in our analysis, adding other data types such as DNA methlyation or protein expression may also help uncover deeper connections across different diseases.

- For transfer learning approaches to be successful, there must be sufficient biological and gene expression overlap between the two diseases.

Barcelona School of Economics

Thank you!

Barcelona School of Economics