Development of a Bayesian Approach to Modeling Tamoxifen Resistance in Breast Cancer Cells Through Adaptive Hamiltonian Monte-Carlo Posterior Sampling

Miguel Fernández de Retana

Degree in Data Science & Artificial Intelligence

University of Deusto
www.deusto.es

Collaboration Acknowledgment

This project has been conducted as a collaborative effort between the **Basque Center for Applied Mathematics (BCAM)** & the **Cancer Heterogeneity Lab** at **CIC bioGUNE**.

Their combined support & resources have contributed significantly to the development and completion of the **project**.

Outline

Breast Cancer: A Global Health Challenge

- **Breast cancer** (BC) is the **most diagnosed** cancer among **women globally**, and accounts for almost **15%** of all female **cancer-related deaths**.
- The most widely accepted **classification** of BC subtypes consists of the following major **molecular subtypes** [1]:

Table 1. Classification of molecular subtypes of breast cancer and therapies.

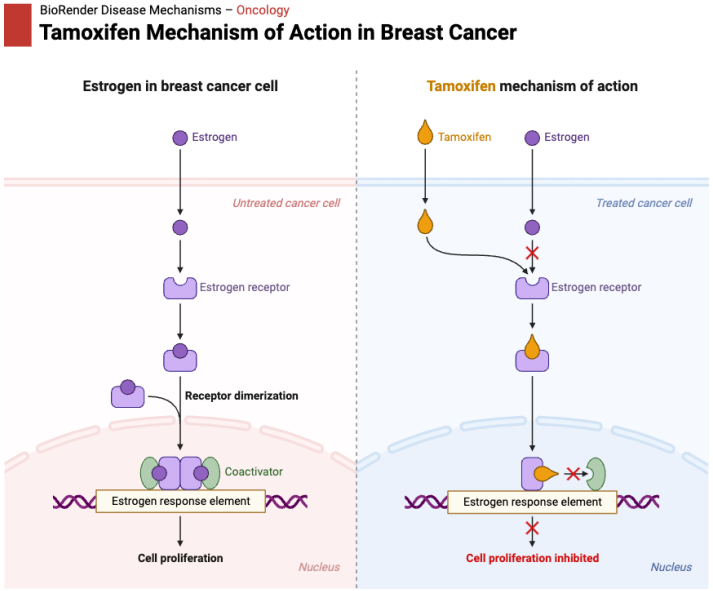
Molecular Subtypes	Luminal A	Luminal B		HER2+	TN
		(HER2-)	(HER2+)		
Biomarkers	ER+ PR+ HER2- Ki67low	ER+ PR- HER2- Ki67high	ER+ PR-/+ HER2+ Ki67low/high	ER- PR- HER2+ Ki67high	ER- PR- HER2- Ki67high
Frequency of Cases (%)	40–50	20–30		15–20	10–20
Histological Grade	Well Differentiated (Grade I)	Moderately Differentiated (Grade II)		Little Differentiated (Grade III)	Little Differentiated (Grade III)
Prognosis	Good	Intermediate		Poor	Poor
Response to Therapies	Endocrine	Endocrine Chemotherapy	Endocrine Chemotherapy Target Therapy	Target Therapy Chemotherapy	Chemotherapy PARP Inhibitors

ER: estrogen receptor; PR: progesterone receptor; HER2: human epidermal growth factor receptor 2.

Figure: Molecular Classification of Breast Cancer Cells

Breast Cancer: A Global Health Challenge

- Among these, Estrogen Receptor-Positive (**ER$^+$**) is the **most prevalent** subtype, accounting for $\approx$ **70-80% of cases**...
- ... but also offers the **best treatment path** $\rightarrow$ **Endocrine therapy**, particularly w/ **Tamoxifen**.
- By binding to ERs, **Tamoxifen** acts as an antagonist in breast tissue, effectively blocking estrogen's **proliferative effects**.



The Challenge of Tamoxifen Resistance

- The development of **tamoxifen** has shown to reduce **recurrence** by $\sim$**50%** and **mortality** by $\sim$**30%** after a **standard 5-year treatment** following surgery [2].
- However, despite the success of tamoxifen, a significant nº of patients develop **resistance** to the drug $\rightarrow$ estimated to be between **30-50%** [1]
- Early identification of resistant cases avoids wasting a **critical 5-year** therapy window.

Our Approach: A Bayesian Framework

- We propose a **Bayesian** modeling framework to analyze the complex mechanisms driving **tamoxifen resistance**...
- ... By integrating **RNA sequencing** data from both **cell-lines** and publicly available **patient** data.
- Our approach allows for the incorporation of **biologically motivated** **prior expert knowledge**, as well as SOTA developments in **computational statistics**.

Outline

Objectives

Our Goal...

To analyze this biological phenomenon responsible for the development of **resistance to tamoxifen** in ER^+ breast cancer cells by integrating **cell-patient data**.

Objectives

Our Goal...

To analyze this biological phenomenon responsible for the development of **resistance to tamoxifen** in ER^+ breast cancer cells by integrating **cell-patient data**.

More concretely:

- To identify **potential genetic biomarkers** associated with tamoxifen resistance in ER+ breast cancer cells.
- To develop **robust** and **interpretable** **prognostic models** for predicting treatment outcomes (including **integrated data**).
- To develop **pyHaiCS**, an **open-source** Python library for computational statistics based on **Hamiltonian-inspired Monte-Carlo** methods.

Available Data

The following data has been provided by the lab at **CIC bioGUNE**:

1. RNA-seq data from MCF7 **cell-lines** (both resistant & control have been independently **sequenced three times**).
2. RNA-seq data from **patients** who have been classified as **resistant** or as **responsive** to tamoxifen treatment. This data comes from **public** cancer repositories (i.e., TCGA).

Outline

Ethical Considerations of the Project

- Our work has **direct implications** for patient care and clinical decision-making...
- This ethical responsibility is **not** limited to the **technical** aspects of our work, but also extends to the **social** and **environmental** dimensions of our research.
- We must consider the potential impact of our work on **vulnerable populations**, ensuring **equitable access** to healthcare technologies, and promoting **safety** and **sustainability**.

Ethical Considerations of the Project

This work adheres to the **ethical obligations** of service to society, health, and public welfare.

Guiding Principles

- **Beneficence:** Aim to **improve patient outcomes** by predicting resistance to **long ineffective treatments**.
- **Justice:** Promote equity through the **pyHaiCS** library, **democratizing** access to advanced methods.
- **Autonomy & Responsibility:** Promote accountability for the technical validity of our models. We respect patient autonomy by ensuring our approach supports **informed decision-making**.

Ethical Considerations of the Project

This work adheres to the **ethical obligations** of service to society, health, and public welfare.

Ethical Practice

- **Data Privacy:** All patient data is **fully anonymized** in compliance with **GDPR**, maintaining **no connection** between genetic profiles and patient identities.
- **Open Science:** Development of **pyHaiCS** promotes transparency, and **collaborative improvement**.
- **Transparency:** Maintaining **interpretable models** and **documentation** to support informed clinical decision-making.

Outline

Multi-Source Data Integration

- Our core challenge is to **bridge the gap** between controlled **in-vitro** experiments and complex **in-vivo** patient realities...

Multi-Source Data Integration

- Our core challenge is to **bridge the gap** between controlled **in-vitro** experiments and complex **in-vivo** patient realities...
- We hypothesize that genes showing **concordant expression changes** in both resistant cells and resistant patients are the most robust candidates for predicting resistance.

Multi-Source Data Integration

- Our core challenge is to **bridge the gap** between controlled **in-vitro** experiments and complex **in-vivo** patient realities...
- We hypothesize that genes showing **concordant expression changes** in both resistant cells and resistant patients are the most robust candidates for predicting resistance.

Data Integration Logic

Cell-Line Data + **Patient Data**

(MCF7 CTRL vs. TamR) (Responsive vs. Resistant)

$\Downarrow$

Refined list of high-confidence biomarkers

Multi-Source Data Integration

We follow a **multi-step pipeline** to distill the most relevant genetic biomarkers from tens of thousands of possibilities:

1. **Filter out low-expression genes** (count < 30) and apply **Relative Log Expression (RLE)** normalization to make counts comparable across all samples.

Multi-Source Data Integration

We follow a **multi-step pipeline** to distill the most relevant genetic biomarkers from tens of thousands of possibilities:

1. **Filter out low-expression genes** (count < 30) and apply **Relative Log Expression (RLE)** normalization to make counts comparable across all samples.
2. Perform **Differential Expression Analysis (DEA)** **independently** on both cell-line and patient datasets to identify statistically **differentially expressed** genes in resistant vs. sensitive groups.

Multi-Source Data Integration

We follow a **multi-step pipeline** to distill the most relevant genetic biomarkers from tens of thousands of possibilities:

1. **Filter out low-expression genes** (count < 30) and apply **Relative Log Expression (RLE)** normalization to make counts comparable across all samples.
2. Perform **Differential Expression Analysis (DEA)** **independently** on both cell-line and patient datasets to identify statistically **differentially expressed** genes in resistant vs. sensitive groups.
3. Apply a **statistical filter** to keep only the **most significant** genes: $(|\log_2 \mathrm{FC}| > 0.5)$ and $(\mathrm{FDR} < 0.1)$.

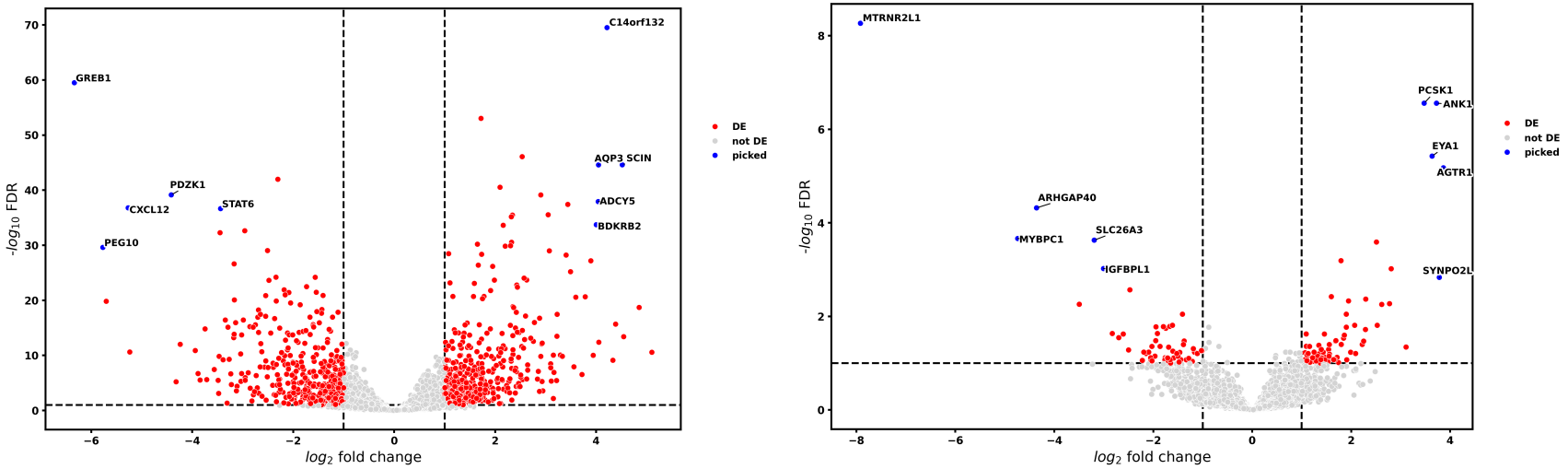
Multi-Source Data Integration

We follow a **multi-step pipeline** to distill the most relevant genetic biomarkers from tens of thousands of possibilities:

1. **Filter out low-expression genes** (count < 30) and apply **Relative Log Expression (RLE)** normalization to make counts comparable across all samples.
2. Perform **Differential Expression Analysis (DEA)** **independently** on both cell-line and patient datasets to identify statistically **differentially expressed** genes in resistant vs. sensitive groups.
3. Apply a **statistical filter** to keep only the **most significant** genes: $(|\log_2 \mathrm{FC}| > 0.5)$ and $(\mathrm{FDR} < 0.1)$.
4. **Integrate** the two filtered lists by selecting only genes that are differentially expressed in the **same direction** (i.e., over-expressed or under-expressed in both cases).

Multi-Source Data Integration

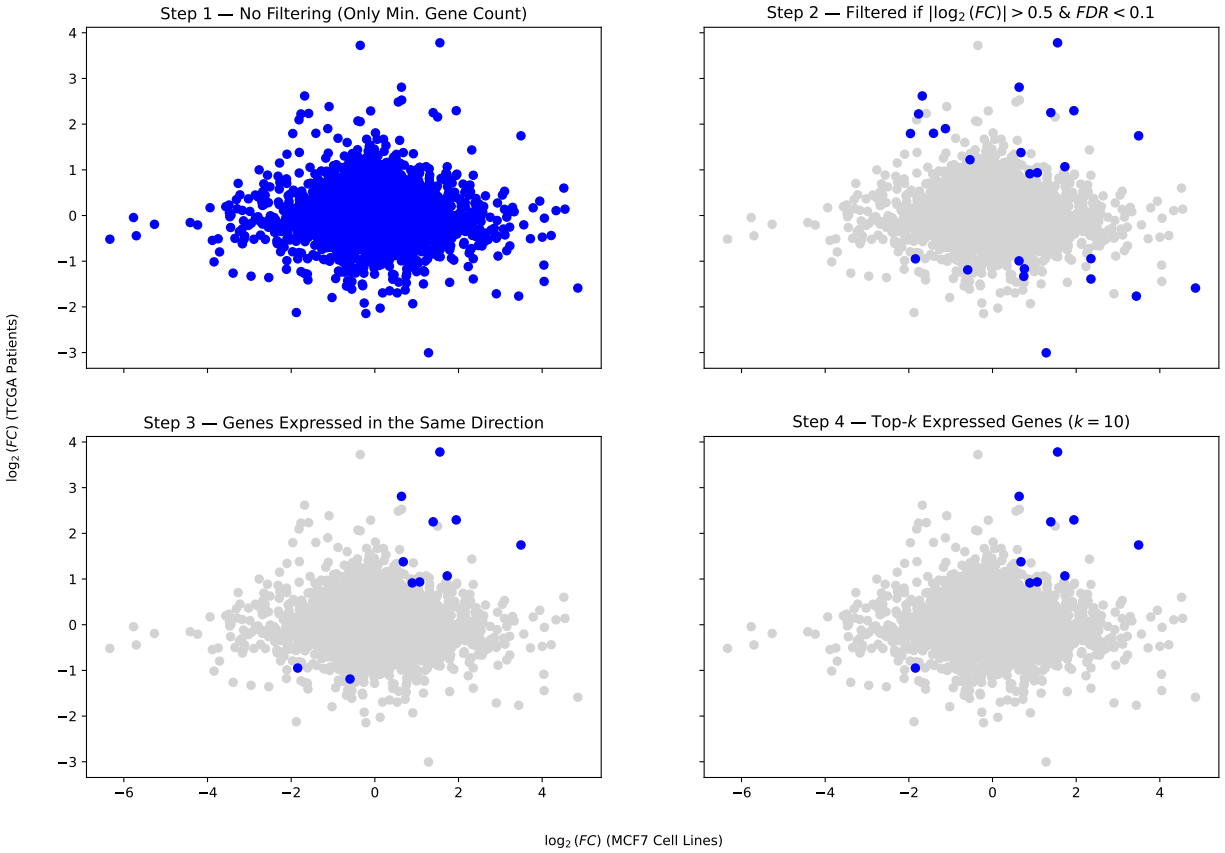
This rigorous filtering process drastically reduces the feature space from over 36,000 genes to just **10 candidate biomarkers**.



(*Left:* MCF7 Cell-Lines, *Right:* TCGA Patients)

Multi-Source Data Integration

This rigorous filtering process drastically reduces the feature space from over 36,000 genes to just **10 candidate biomarkers**.



Outline

Quick Detour: Bayesian Inference

In Bayesian inference, our goal is to understand the **posterior distribution** which describes our belief about the model parameters after seeing the data:

$$\overbrace{p(\theta|x)}^{\text{Posterior}} = \frac{\overbrace{p(\theta)}^{\text{Prior}}\ \overbrace{p(x|\theta)}^{\text{Likelihood}}}{\underbrace{p(x)}_{\text{Marginal Likelihood}}} \tag{1}$$

- This posterior is often a **complex**, high-dimensional distribution that we can't solve **analytically**.

Quick Detour: Bayesian Inference

In Bayesian inference, our goal is to understand the **posterior distribution** which describes our belief about the model parameters after seeing the data:

$$\overbrace{p(\theta|x)}^{\text{Posterior}} = \frac{\overbrace{p(\theta)}^{\text{Prior}}\ \overbrace{p(x|\theta)}^{\text{Likelihood}}}{\underbrace{p(x)}_{\text{Marginal Likelihood}}} \tag{2}$$

- Actually, given a **new observation** $\tilde{x}$, predictions can be made by using the *posterior predictive distribution* as:

$$p(\tilde{x}|x) = \int p(\tilde{x}, \theta|x)\, d\theta = \int p(\tilde{x}|\theta) p(\theta|x)\, d\theta \tag{3}$$

Quick Detour: Bayesian Inference

Since the integral has no analytical solution, we must **draw samples** to approximate it.

- Standard MCMC methods use inefficient **random walks** to explore the probability space, which converge poorly in high dimensions.
- **Hamiltonian Monte-Carlo (HMC)** uses a much smarter approach inspired by **classical mechanics** to propose new samples.
- The core idea is to augment our parameters, or the **position** (θ), with an auxiliary **momentum** variable (p). This creates a physical system whose total energy is described by the **Hamiltonian**:

$$H(\theta, p) = K(p) + U(\theta) = \frac{1}{2} p^T M^{-1} p + U(\theta) \tag{4}$$

Quick Detour: Bayesian Inference

- Instead of a random step, we simulate this system's evolution through the **numerical integration** of the **Hamiltonian dynamics**:

$$\dot{\theta} = H_p(\theta, p) = M^{-1}p, \qquad \dot{p} = -H_\theta(\theta, p) = -U_\theta(\theta) \tag{5}$$

- Where the **Hamiltonian potential** $U(\theta)$ is related to the **target distribution** by: $U(\theta) = -\log \pi(\theta) + \mathcal{C}$.
- In practice, the integration of the target dynamics is carried out by combining the following **solution flows**:

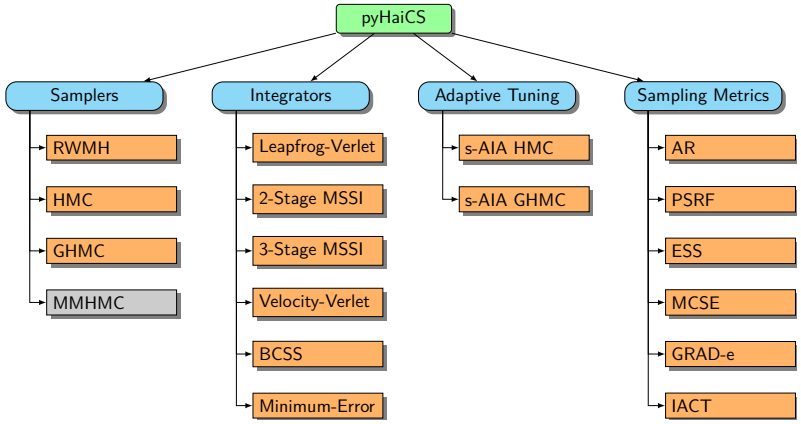
$$\varphi_t^A(\theta, p) = (\theta + tM^{-1}p, p), \qquad \varphi_t^B(\theta, p) = (\theta, p - tU_\theta(\theta)) \tag{6}$$

Combining these flows gives rise to various **numerical integrators**, and further enhancements lead to different **sampling algorithms** like **HMC**, **GHMC**, etc.

pyHaiCS: A New Tool for Bayesian Inference

- Introducing **pyHaiCS**: a new Python library for **Hamiltonian Monte-Carlo (HMC)** methods.
- Built on **JAX** for high performance: Just-In-Time (JIT) compilation, **automatic differentiation**, and **hardware acceleration** (CPU/GPU/TPU).
- Implements a wide range of SOTA **samplers**, numerical **integrators**, and **adaptive tuning** algorithms.
- Designed to be **user-friendly** and easily integrable with existing scientific Python workflows.

pyHaiCS



The Challenge of Data Imbalance

- The patient dataset is highly **imbalanced**: only $\sim$**30%** of patients are resistant to tamoxifen.
- In this clinical context, **failing to identify a resistant patient** is far more critical than misclassifying a sensitive one.
- Therefore, standard metrics like accuracy are misleading. We focus on **Recall** (Sensitivity) and the **Matthews Corr. Coeff. (MCC)**.

The Challenge of Data Imbalance

- The patient dataset is highly **imbalanced**: only **$\sim$30%** of patients are resistant to tamoxifen.
- In this clinical context, **failing to identify a resistant patient** is far more critical than misclassifying a sensitive one.
- Therefore, standard metrics like accuracy are misleading. We focus on **Recall** (Sensitivity) and the **Matthews Corr. Coeff. (MCC)**.

Solution: Data Augmentation with SMOTE

We use the **Synthetic Minority Oversampling Technique (SMOTE)** to balance the dataset by generating synthetic samples for the minority (resistant) class. All models were trained and evaluated on both the original and augmented datasets to measure the impact.

Development of Prognostic Models

We developed a wide range of models to predict **tamoxifen resistance**:

Bayesian Models:

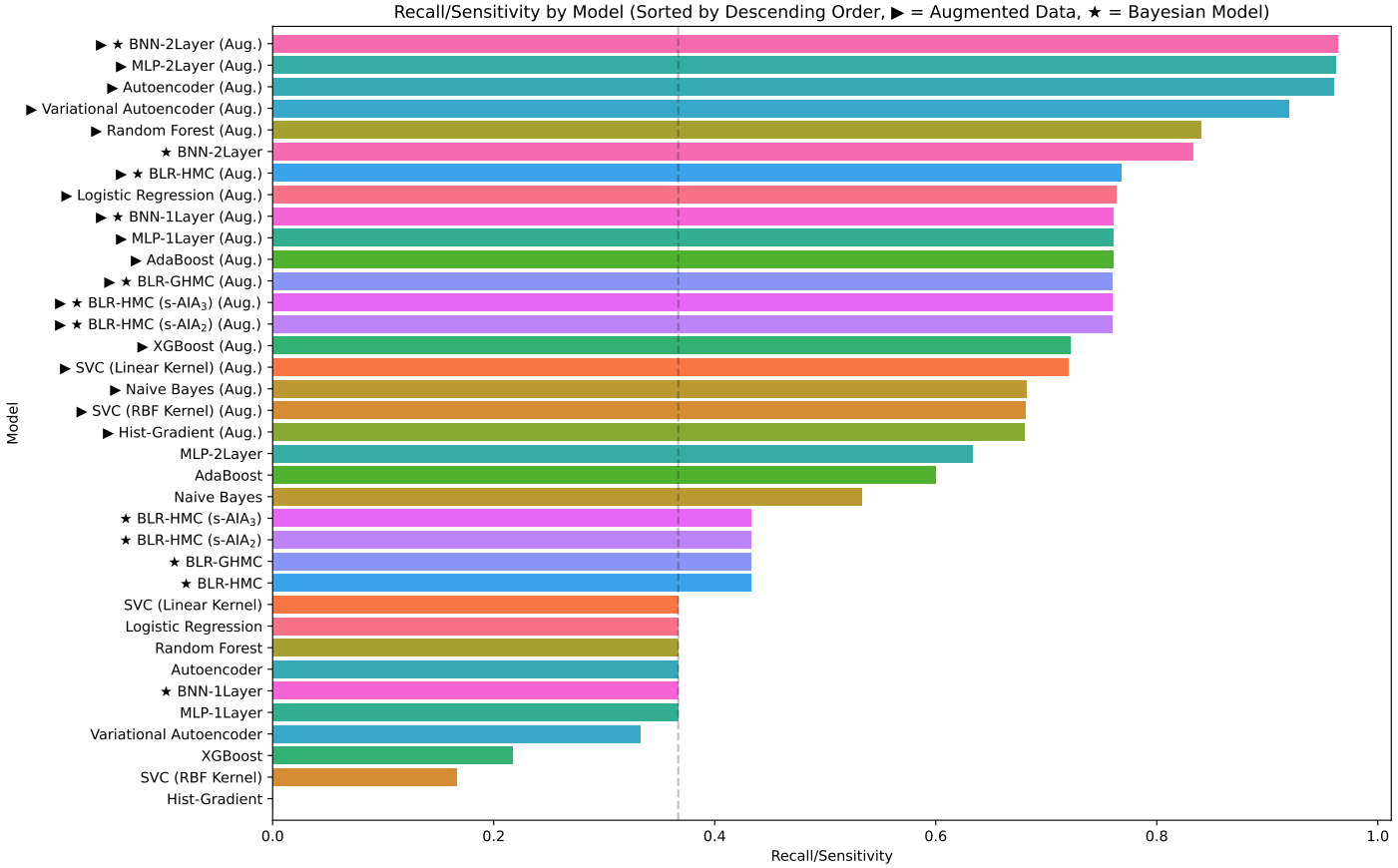
- **Bayesian Logistic Regression (BLR) w/ HMC:** Priors for gene coefficients were set using **cell-line data**: $\theta_i \sim \mathcal{N}(\log_2 FC_i, 2.5^2)$.
- **Bayesian Neural Networks (BNN):** Advanced NNs that treat weights as probability distributions.

Baseline Models:

- **Shallow Models:** Logistic Regression, Support Vector Classifiers.
- **Ensemble Methods:** Random Forest, XGBoost, AdaBoost.
- **Neural Networks:** MLPs, Autoencoders, Variational Autoencoders.

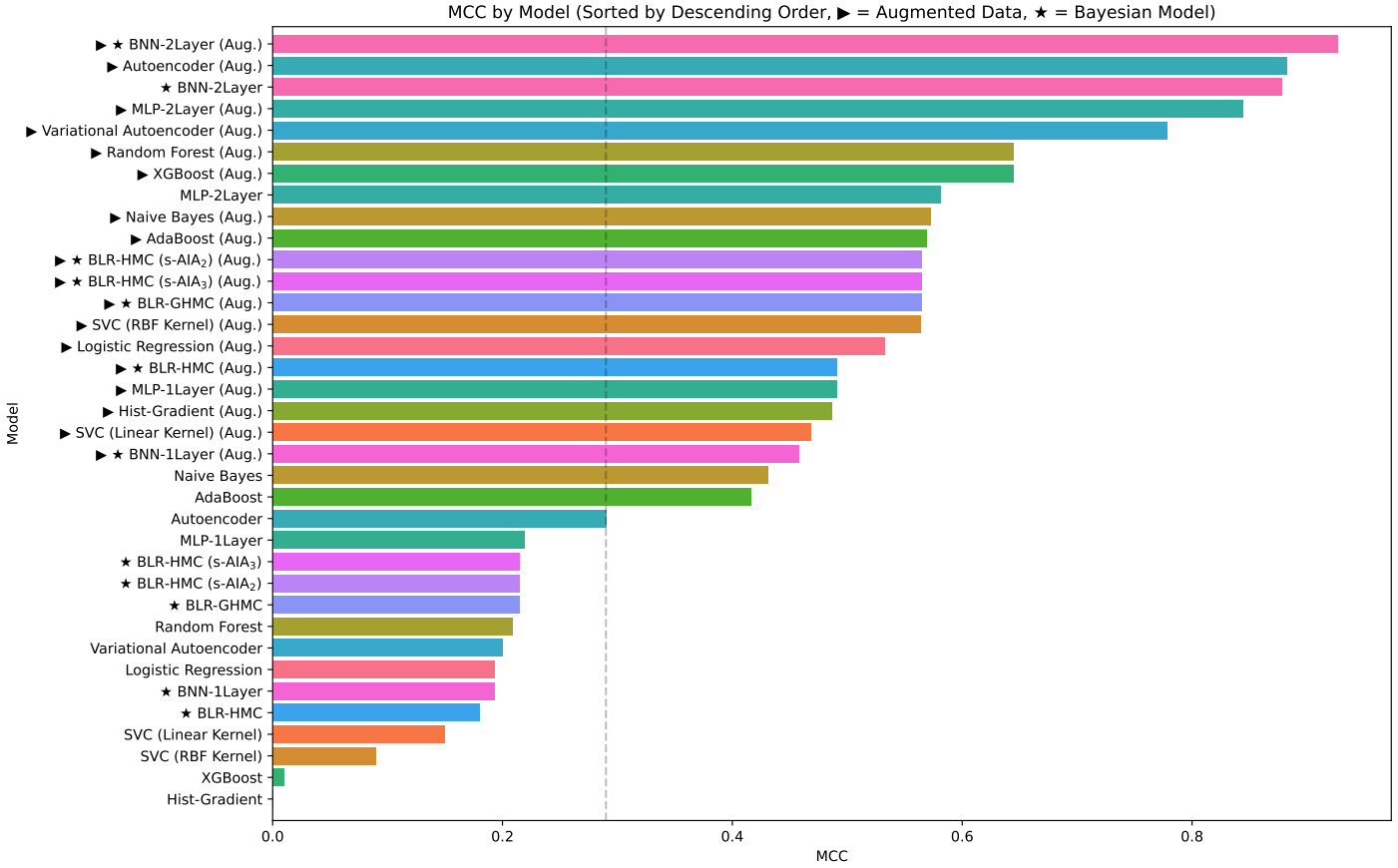
Development of Prognostic Models

Data augmentation with SMOTE **drastically improved** the models' ability to identify resistant patients...



Development of Prognostic Models

Bayesian and **deep learning** models showed the best overall performance, especially on the augmented dataset.



Development of Prognostic Models

Key Takeaways:

- **Data augmentation** (SMOTE) improved all models' performance
  Synthetic samples may not fully represent the underlying **biological variability**...
- On the **sensitivity** of models to **class imbalance**...
- **(Bayesian) Neural Networks** achieved best performance
  (BNN-2Layer: Recall=0.964, MCC=0.927)
- Still, simple **shallow** models (Logistic Regression/RF) remained **competitive** when dealing with the **augmented dataset**.
- Significant improvements over **previous work**... (Recall of > 0.9 vs. 0.367)

Limitations:

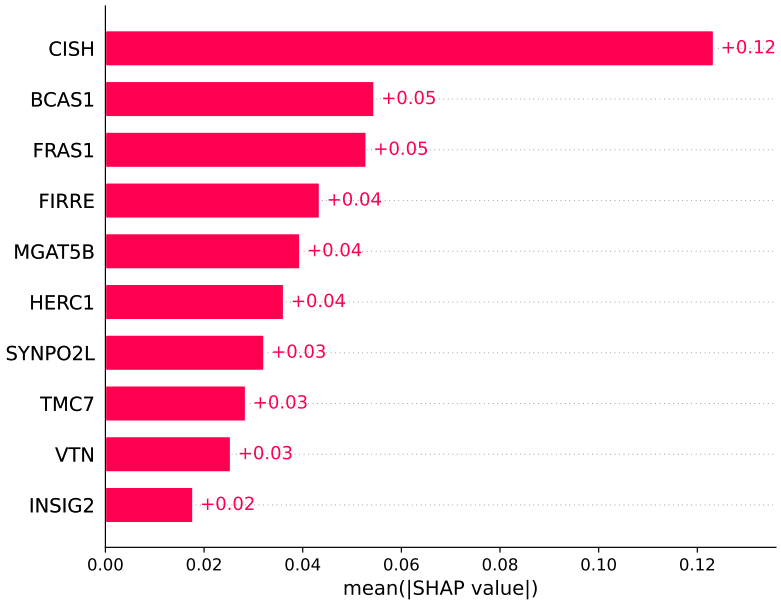
- Small cohort size ($n = 37$ patients), even after augmentation
- RNA-seq integration showed limited impact
  (BLR-HMC MCC=0.565 vs. BNN-2Layer MCC=0.927, BLR-HMC vs *Vanilla* LR)

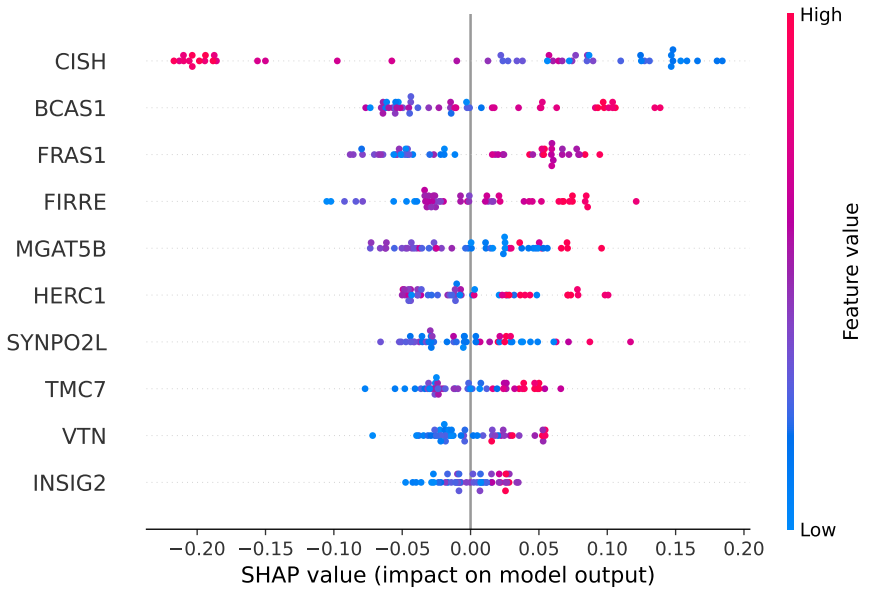
Outline

Explaining the “Black Box”: SHAP Values

- To understand **why** our model makes certain predictions, we use **SHAP (SHapley Additive exPlanations)** values.
- SHAP is a method from cooperative game theory that explains the output of any machine learning model by computing the **contribution of each feature** (gene) to the prediction.
- It allows us to identify the most important genes that our models use to predict tamoxifen resistance.
- **Important Note:** SHAP values show **correlation**, not causation. They reveal which genes are most predictive for the model, but do not prove a direct causal role in resistance.

Potential Genetic Biomarkers (Global)





- **Low expression** of **CISH** (blue dots on the right) is strongly associated with a higher probability of resistance.
- **CISH** has the greatest discerning power in the model's predictions.
- Conversely, **high expression** of genes like **BCAS1** and **FRAS1** (red dots on the right) is associated with higher resistance.

Potential Genetic Biomarkers (Local)

We can also break down the predictions for a **single patient**...

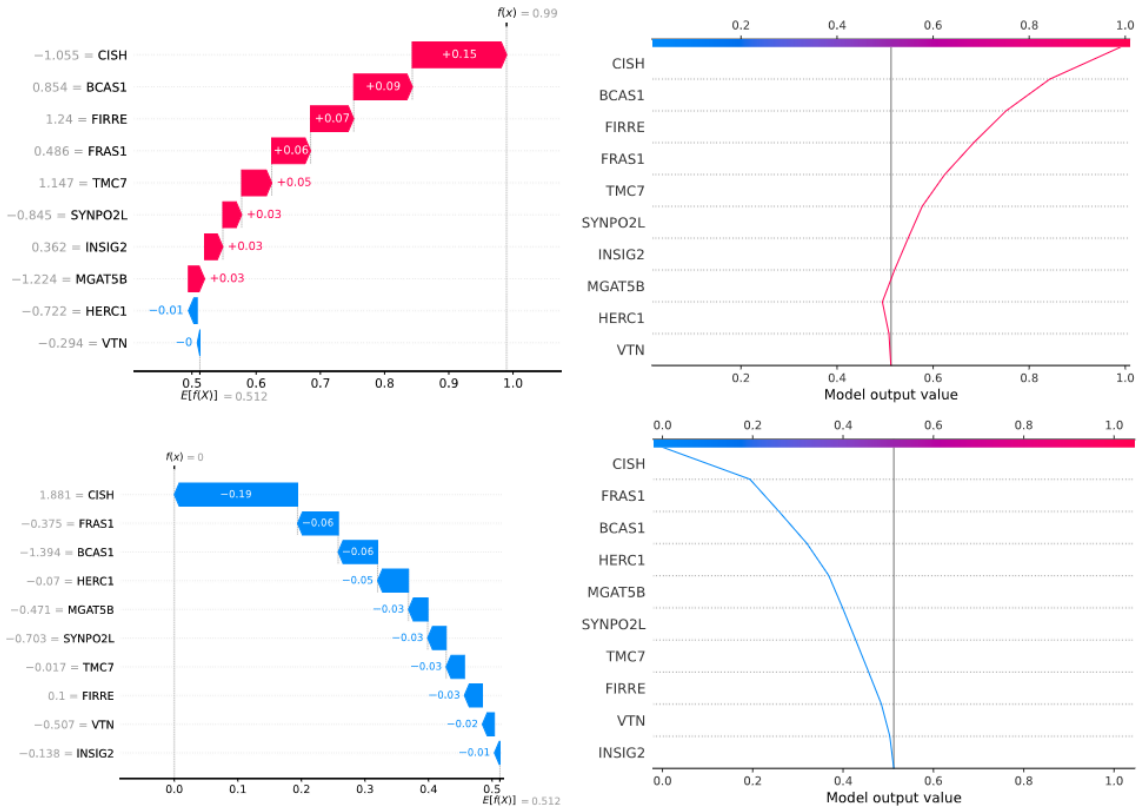


Figure 5.6: Local Gene Contribution – Per-Gene Contribution for Two Sample Patients (*Above:* Resistant, *Below:* Favorable Treatment) (*Left:* Waterfall Contribution Plots, *Right:* Decision Contribution Plots)

External Validation with Survival Analysis

To validate our findings, we performed a **Kaplan-Meier Survival Analysis** on an **independent** cohort of **178** ER$^+$ breast cancer patients treated with tamoxifen...

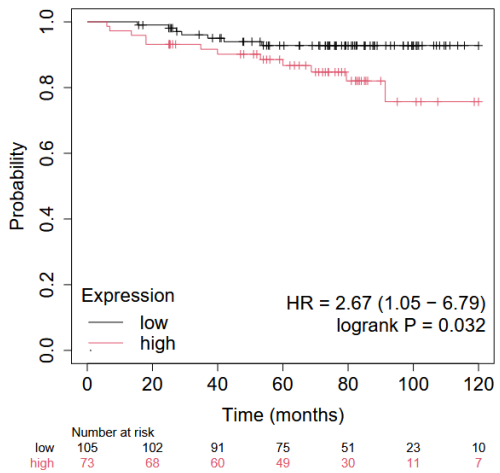


Figure 5.7: Kaplan-Meier Survival Analysis for the Identified Genes in ER+ Breast Cancer Patients – Relapse-Free Survival (RFS) Probability Over Time

- The analysis shows a **stat. significant** difference in RFS between patients with ↑ vs. ↓ expression of our signature (p-value = 0.032).
- Patients with high expression of the signature had a **2.67 times greater risk** of relapse (Hazard Ratio = 2.67).

Associated Biological Pathways

An **enrichment analysis** suggests the identified genes are involved in biological pathways known to be critical in breast cancer progression.

- The **Human ECM-receptor interaction pathway**, which plays a critical role in cancer progression and survival.
- The **Interleukin-7 (IL-7) signaling pathway**, which is known to be involved in promoting breast cancer cell proliferation.

Limitation: Due to the small size of our gene signature, it is difficult to extract definitive conclusions from pathway analysis. Further biological investigation is required...

Outline

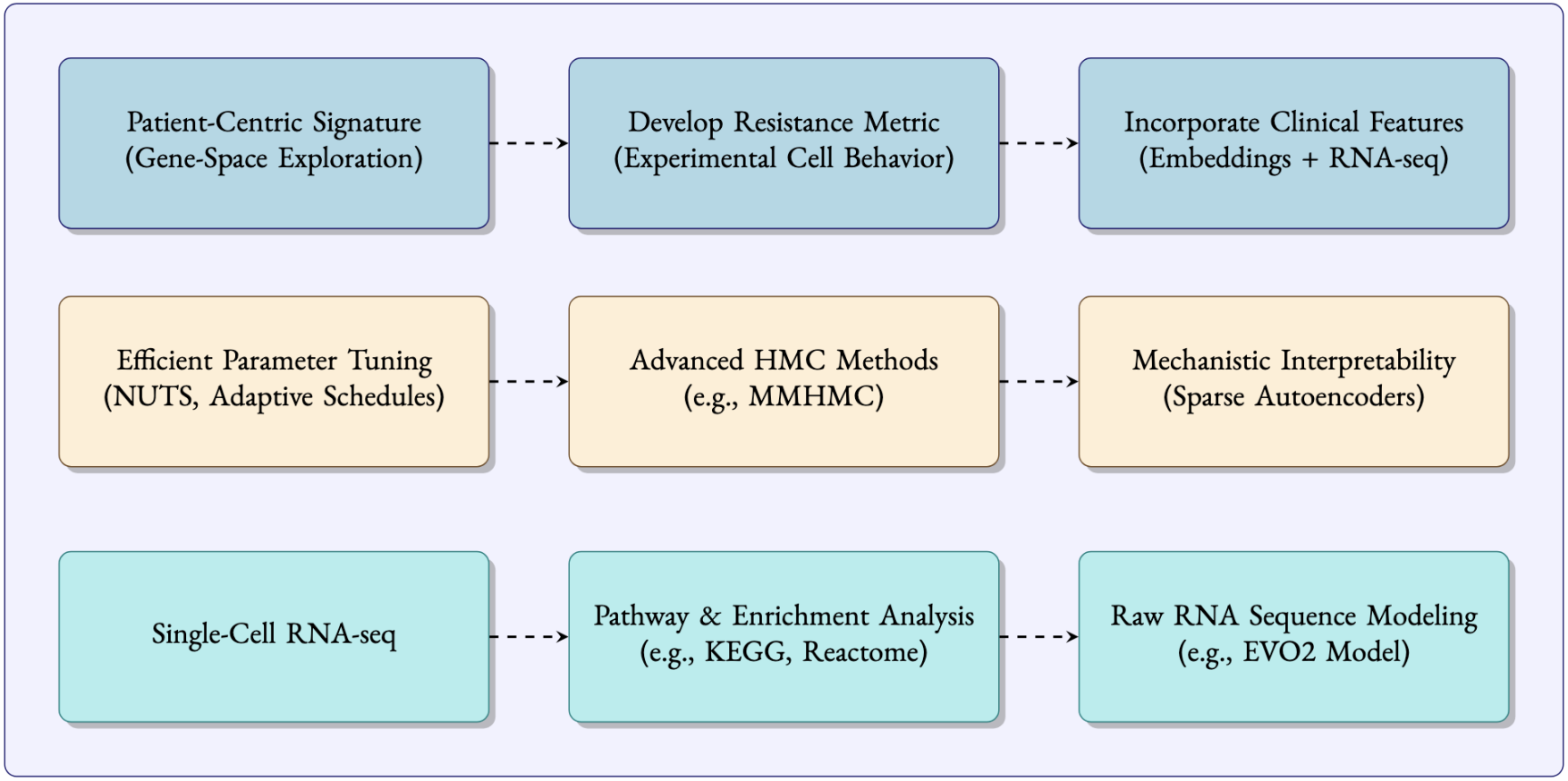
Conclusions & Future Work

Key Findings & Contributions:

1. Data augmentation significantly improved performance across all models.
2. BNNs achieved near-optimal results.
3. Identified 10 key biomarkers contributing to tamoxifen resistance.
   (CISH, BCAS1, FRAS1, FIRRE, MGAT5B, HERC1, SYNPO2L, TMC7, VTN, INSIG2)
4. Cell-line priors did not yield expected improvements.
5. **pyHaiCS** library for HMC-based Bayesian inference.

Conclusions & Future Work

Future Research Directions



Questions?

Bibliography I

Therese Sørlie, Charles M Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, et al.

Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.

Proceedings of the National Academy of Sciences, 98(19):10869–10874, 2001.

Early Breast Cancer Trialists' Collaborative Group et al.

Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials.

The lancet, 378(9793):771–784, 2011.