Journal of Probability and Statistical Science
24(1), 1-15 Feb. 2026

Combining BART and Reciprocal LASSO for High-Dimensional Gene Expression Modeling

Saif Hosam Raheem
Department of statistics,
University of Al-Qadisiyah, Iraq

ABSTRACT

High-dimensional gene expression data pose major challenges for statistical modeling due to the large number of predictors, strong correlations, and the presence of nonlinear regulatory structures. This study proposes a hybrid Bayesian framework that combines Bayesian Additive Regression Trees (BART) with the Reciprocal LASSO prior to achieve flexible nonlinear modeling and structured sparsity within a unified model. Theoretical development integrates aggressive shrinkage for variable selection with a sum-of-trees architecture that captures complex gene–gene interactions. A comprehensive simulation study across multiple dimensional and correlation settings demonstrates that the proposed BART–RL model consistently achieves lower prediction error and higher true positive rates compared with classical LASSO, elastic net, BART, and Bayesian reciprocal LASSO. Application to a real gene expression dataset further confirms the advantages of the hybrid approach, yielding improved predictive performance and identifying biologically meaningful genes supported by functional annotations. These results highlight the utility of combining nonlinear Bayesian tree ensembles with adaptive shrinkage priors for high-dimensional genomic modeling.

Keywords: BART; Reciprocal LASSO; High-dimensional modeling; Gene expression analysis; Bayesian shrinkage; Variable selection; Nonlinear regression.

1. Introduction

High-dimensional gene expression data present substantial statistical and computational challenges due to the large number of predictors relative to sample size, strong correlations among genes, and the presence of complex nonlinear regulatory mechanisms. Classical regression techniques often struggle in this setting, leading to unstable estimates and limited interpretability. As a result, modern genomic studies require modeling frameworks that combine flexible function estimation with principled variable selection.

Regularization methods have become central tools for high-dimensional analysis. The reciprocal LASSO family, in particular, has gained notable attention for its ability to enforce strong shrinkage and promote sparsity more effectively than traditional L1 penalties. Recent developments have demonstrated the robustness of reciprocal-based penalties in regression, quantile regression, and binary models, offering improved feature selection and estimation stability in complex high-dimensional systems as shown by Mallick and Alhamzawi (2021), Alhamzawi and Mallick (2022), and Paul and Mallick (2024). These methods address key limitations of classical shrinkage approaches and provide a refined mechanism for isolating relevant predictors in genomic datasets.

In parallel, Bayesian Additive Regression Trees BART has emerged as a powerful nonparametric modeling framework capable of capturing nonlinearities and high-order

- Received February 4, 2026, in final form February 2026.
- Saif Hosam Raheem (corresponding author) is affiliated with Department of statistics, University of Al-Qadisiyah, Iraq
Saif.hosam@qu.edu.iq

interactions without requiring explicit model specification. The flexibility of BART makes it particularly suitable for genomic studies where regulatory interactions are often nonlinear and multi-layered. Foundational work by Hill (2011) and subsequent extensions such as those by Linero (2018), Pratola et al. (2020), and Xu and Wu (2022) highlight the method's capacity to learn complex structures in high-dimensional predictors while maintaining interpretability through posterior inference.

Despite these advancements, limited research has explored the integration of reciprocal LASSO penalties with BART in a unified Bayesian framework. Existing BART models excel in predictive performance but offer limited direct variable-selection capabilities, while reciprocal LASSO provides strong sparsity but lacks a mechanism for modeling nonlinear effects. This gap motivates the need for a hybrid method that simultaneously achieves flexible function estimation and structured sparsity.

The goal of this study is to develop a combined BART–Reciprocal LASSO model for high-dimensional gene expression analysis. The proposed framework leverages BART's ability to approximate nonlinear relationships together with the adaptive shrinkage properties of reciprocal LASSO priors. Through simulation experiments and real gene expression applications, we demonstrate that the hybrid model achieves superior predictive accuracy, stable variable selection, and improved biological interpretability compared with existing approaches.

2. Theoretical Background

2.1 High-Dimensional Gene Expression Data

High-dimensional gene expression datasets arise when the number of genes p is extremely large relative to the sample size n, a structure commonly seen in RNA-Seq and microarray experiments. Such data typically contain tens of thousands of predictors that exhibit strong correlations and substantial noise, making classical estimation unstable and prone to overfitting. As noted by Song and Liang (2015), high-dimensional regression becomes highly sensitive to multicollinearity, and the covariance matrix $X^T X$ often becomes ill-conditioned when $p >> n$.

A general representation of gene expression modeling is given by:

$$y_i = x_i^T \beta + \varepsilon_i$$

where y_i is the biological outcome, x_i is the high-dimensional gene vector, and β denotes the underlying signal. Extracting meaningful biological information requires methods capable of simultaneous shrinkage and selection because only a small subset of genes truly influences the response. Recent studies emphasize the need for flexible nonlinear approaches in genomic prediction, especially in settings involving complex interactions (Linero, 2018). These challenges motivate the integration of reciprocal LASSO shrinkage with Bayesian tree-based models to provide robust and adaptive modeling for high-dimensional gene expression data.

2.2 Reciprocal LASSO Framework

The reciprocal LASSO extends classical $L1$-regularization by applying stronger shrinkage to small coefficients and weaker shrinkage to large signals. This behavior makes it suitable for high-dimensional genomic settings where only a limited subset of genes carries predictive information. The method was first formalized within high-dimensional

regression by Song and Liang (2015), and later expanded through Bayesian formulations by Mallick and Alhamzawi (2021) and Alhamzawi and Mallick (2022).

The reciprocal LASSO estimator solves the penalized objective:

$$\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\frac{1}{|\beta_j|}$$

where y_i denotes the response, $x_i$is the p-dimensional gene expression vector, β is the coefficient vector, and $\lambda > 0$ controls penalty strength.

The reciprocal penalty $P(\beta_j) = \frac{1}{|\beta_j|}$ is singular at zero, producing aggressive shrinkage for small coefficients while becoming flat for large coefficients. This reduces estimation bias for strong signals and improves variable selection accuracy in sparse gene expression models, as discussed in Song (2018) and Paul and Mallick (2024).

In the Bayesian formulation, the penalty corresponds to the prior:

$$\pi(\beta_j \mid \lambda) \propto exp(-\lambda \frac{1}{|\beta_j|})$$

which leads to the posterior:

$$\pi(\beta \mid y, X, \lambda) \propto exp(\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p}\frac{1}{|\beta_j|}).$$

This posterior structure enables full uncertainty quantification and makes reciprocal LASSO an effective shrinkage prior in high-dimensional genomic modeling, particularly for sparse signal detection and robust prediction (Raheem, 2025; Hassan, 2025).

2.3 Bayesian Additive Regression Trees

Bayesian Additive Regression Trees BART constitute a nonparametric Bayesian regression framework in which the unknown regression function is expressed as a sum of many small decision trees, each regularized through shrinkage priors. The foundational development of this approach was introduced by Hill (2011), who demonstrated that an ensemble of weak trees can flexibly approximate complex and nonlinear relationships without the need to manually specify interaction terms. For a continuous outcome Y and a p-dimensional predictor vector $x = (x_1, \ldots, x_p)$, BART assumes the regression structure:

$$Y = f(x) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2),$$

where the unknown function $f(x)$ is represented through a sum-of-trees expansion:

$$f(x) = \sum_{i=1}^{m} g(x; T_j, M_j).$$

Here, T_j denotes the structure of tree j, and $M_j = \{\mu_{j1}, \ldots, \mu_{jb_j}\}$ is the set of terminal node parameters. For a single tree, the function $g(x; T, M)$ is evaluated by dropping an observation with covariates x down the tree until it reaches a terminal node, where it is assigned the corresponding terminal mean μ_z. A single-tree model may be written as:

$$Y = g(x; T, M) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2).$$

Instead of relying on a single decision tree, BART fits an ensemble of m trees often in the hundreds to obtain the model:

$$Y = \sum_{i=1}^{m} g(x_i; T_j, M_j) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

This additive structure enables BART to capture nonlinear effects and interactions automatically, a characteristic emphasized by Linero (2018) in the context of high-dimensional models. Regularization is enforced through prior distributions on both the tree structures and the terminal node parameters. The prior over tree topology is defined by a stochastic tree-generating process in which the probability that a node at depth d splits is:

$$p(\text{split at depth } d) = \alpha\,(1 + d)^{-\beta},$$

with common default values $\alpha = 0.95$ and $\beta = 2$. These values concentrate the prior on shallow trees and maintain each tree as a weak learner, consistent with the formulation of Pratola et al. (2020).

Given a fixed tree structure T_j, the terminal node parameters μ_{jk} follow independent Gaussian priors:

$$\mu_{jk} \sim N(\mu_\mu, \tau_\mu^{-1}),$$

where the hyperparameters are selected such that the prior on $f(x)$ is concentrated within a plausible range of the response variable. This shrinkage prevents any individual tree from dominating the ensemble. Scalable implementations of these priors, designed for large-p settings such as genomic data, were developed by Li and Linero (2023). Combining the likelihood and prior distributions yields the joint posterior:

$$p(\{T_j, M_j\}_{j=1}^m, \sigma^2 \mid y, X) \propto \prod_{i=1}^{n} N(y_i \mid \sum_{i=1}^{m} g(x_i; T_j, M_j), \sigma^2) \prod_{j=1}^{m} p(T_j)\, p(M_j)\, p(\sigma^2)\ .$$

Posterior computation in BART relies on a Bayesian backfitting algorithm, in which each tree is updated sequentially while conditioning on the remaining trees. For tree t, the algorithm computes partial residuals:

$$r_i^{(t)} = y_i - \sum_{j \neq t} (x_i; T_j, M_j),$$

and treats them as the effective response when proposing modifications to T_t and M_t. The MCMC sampler iterates through grow, prune, and change moves for the trees, updates terminal node parameters using conditional Gaussian distributions, and samples the residual variance from a conjugate gamma posterior. Extensions such as sparse BART introduced by Xu and Wu (2022) and log-linear BART for genomic interactions proposed by Murray (2017) further demonstrate the method's adaptability to ultrahigh-dimensional biological data.

Through its sum-of-trees representation, hierarchical shrinkage priors, and Bayesian backfitting algorithm, BART provides a powerful and robust modeling tool for high-dimensional prediction problems, particularly those involving complex nonlinear structures in gene expression data.

2.4 Hybrid Modeling Strategies

Hybrid modeling strategies that combine reciprocal LASSO with Bayesian additive regression trees offer a powerful framework for high-dimensional gene expression analysis. The main idea is to integrate a sparse linear component with a flexible nonlinear component in order to capture different types of signal structures within genomic data. Let

y_i denote the response for subject i, and let $x_i = (x_{i1}, \dots, x_{ip})^T$ represent the corresponding gene expression vector. A general hybrid model can be expressed as:

$$y_i = x_i^T \beta + f(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma\text{^}2),$$

where the coefficient vector β captures linear effects, and the function $f(x_i)$ models nonlinearities and interactions using a BART prior (Hill, 2011; Linero, 2018).

The linear component is regularized through the reciprocal LASSO prior:

$$\pi(\beta_j \mid \lambda) \propto exp(\frac{-\lambda}{|\beta_j|}),$$

which induces strong shrinkage on small coefficients while keeping bias low for large effects. This property makes reciprocal LASSO suitable for high-dimensional gene expression settings, where only a small subset of predictors is expected to be relevant (Song and Liang, 2015; Mallick and Alhamzawi, 2021; Alhamzawi and Mallick, 2022; Paul and Mallick, 2024; Raheem, 2025).

The nonlinear component $f(x_i)$ is represented by a sum of regression trees:

$$f(x_i) = \sum_{m=1}^{M} g(x_i;\, T_m, M_m) \ ,$$

where each tree T_m partitions the predictor space and assigns terminal node parameters M_m. Shrinkage priors on tree depth and terminal node values ensure that each tree contributes only a small part of the overall prediction, preventing overfitting and improving stability (Pratola et al., 2020; Li and Linero, 2023).

Combining these elements, the likelihood of the hybrid model is:

$$p(y \mid X, \beta, \{T_m, M_m\}, \sigma^2) = \prod_{i=1}^{n} N(\, y_i \mid x_i^T \beta + \sum_{m=1}^{M} g(x_i;\, T_m, M_m)\,, \sigma^2\,).$$

The corresponding posterior distribution becomes:

$$p(\beta, \{T_m, M_m\}, \sigma^2 \mid y, X) \propto p(y \mid X, \beta, \{T_m, M_m\}, \sigma^2\,) \prod_{j=1}^{p} \pi(\beta_j \mid \lambda)$$

$$\prod_{m=1}^{M} p(T_m)\, p(M_m)\, p(\sigma^2) \ .$$

Posterior computation proceeds by alternating between updating the coefficient vector β under the reciprocal LASSO prior and updating the tree ensemble $\{T_m, M_m\}$ through BART's Bayesian backfitting algorithm. Sparse and scalable BART variants (Xu and Wu, 2022; Murray, 2017) further enhance computational efficiency in ultrahigh-dimensional genomic settings. This hybrid strategy therefore combines the interpretability and sparsity of reciprocal LASSO with the nonlinear flexibility of BART, yielding a robust and adaptable modeling approach for complex gene expression data.

3. Proposed Model and Prior Specification

3.1 Model Structure

The proposed model integrates the sparsity-inducing behavior of the reciprocal LASSO with the nonlinear flexibility of Bayesian additive regression trees in a unified Bayesian framework. Let y_i denote the response for observation i, and let $x_i =$

$(x_{i1}, \dots, x_{ip})^T$ represent the corresponding p-dimensional gene expression vector. The model assumes that the response is generated from a combination of a sparse linear component and a nonlinear interaction component:

$$y_i = x_i^T \beta + f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

where

$\beta = (\beta_1, \dots, \beta_p)^T$ is the vector of regression coefficients governing the linear effects, $f(x_i)$ is an unknown nonlinear function modeled through a BART prior, and $\varepsilon_i \sim N(0, \sigma^2)$ is the error term.

The linear component $x_i^T \beta$ captures direct gene–response associations, while the nonlinear component $f(x_i)$ accommodates complex gene–gene interactions and higher-order effects that cannot be represented through a purely linear specification. The BART component expresses $f(x_i)$ as a sum of regression trees:

$$f(x_i) = \sum_{m=1}^{M} g(x_i; T_m, M_m),$$

where T_m is the structure of tree m and M_m contains its terminal node parameters. Each tree partitions the predictor space into disjoint regions and assigns a local mean to each region, allowing the ensemble to approximate highly nonlinear relationships.

The hybrid structure jointly estimates the sparse coefficients β and the nonlinear function $f(\cdot)$. The reciprocal LASSO prior applied to β encourages aggressive shrinkage of irrelevant gene effects while preserving large, biologically meaningful signals. Simultaneously, the BART prior on $f(x)$ learns residual nonlinear patterns that remain after accounting for the linear structure. Together, these components produce a model capable of identifying both sparse main effects and complex interaction-driven effects in high-dimensional genomic settings.

3.2 Reciprocal LASSO Prior

The reciprocal LASSO prior provides a sparsity-inducing mechanism that aggressively shrinks small coefficients while preserving large signal effects. This behavior makes it well suited for high-dimensional genomic modeling, where only a limited subset of predictors is expected to influence the response. Let $\beta = (\beta_1, \dots, \beta_p)^T$ denote the regression coefficients associated with the linear component of the model. The reciprocal LASSO prior is defined through the density:

$$\pi(\beta_j \mid \lambda) \propto exp(\frac{-\lambda}{|\beta_j|}), \quad j = 1, \dots, p,$$

where $\lambda > 0$ is a global shrinkage parameter controlling the strength of penalization. This prior corresponds to the classical reciprocal LASSO penalty introduced in high-dimensional regression contexts (Song and Liang, 2015) and later extended to Bayesian formulations by Mallick and Alhamzawi (2021) and Alhamzawi and Mallick (2022).

The reciprocal prior is singular at zero because:

$$\lim_{\beta_j \to 0} \frac{1}{|\beta_j|} = \infty,$$

which induces extremely strong shrinkage for small coefficients. As a result, many β_j values are pushed close to zero, yielding automatic variable selection. In contrast, for large coefficients the penalty becomes flat:

$$\frac{1}{|\beta_j|} \approx 0 \quad \text{when } |\beta_j| \text{ is large,}$$

resulting in minimal shrinkage and reduced bias for true signals. This asymmetric shrinkage profile is particularly advantageous in genomic applications where effect sizes may vary widely across predictors (Paul and Mallick, 2024; Raheem, 2025).
Under this prior, the posterior contribution of each coefficient takes the form:

$$p(\beta_j \mid y, X, \lambda) \propto exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2) \cdot exp(\frac{-\lambda}{|\beta_j|}),$$

which balances the likelihood of the linear model with the reciprocal penalty. Because the prior lacks conjugacy, sampling β_j typically requires a Metropolis–Hastings step or adaptive slice sampling within a Markov chain Monte Carlo algorithm. The reciprocal structure enhances exploration of the posterior by allowing the model to retain strong signals while discarding noisy predictors.

Overall, the reciprocal LASSO prior provides a principled Bayesian mechanism for sparsity, enabling the hybrid model to identify meaningful gene effects before their nonlinear interactions are captured by the BART component.

3.3 BART Component

The Bayesian Additive Regression Trees (BART) model represents a flexible nonparametric regression approach that approximates an unknown response surface using a sum of weak regression trees regularized through shrinkage priors. For a continuous outcome y_i and predictor vector $x_i = (x_{i1}, \dots, x_{ip})$, BART models the regression function through the likelihood:

$$y_i \mid x_i \sim N(f(x_i), \sigma^2),$$

where the unknown function $f(\cdot)$ is expressed as a sum of m regression trees:

$$f(x_i) = \sum_{j=1}^{m} g(x_i;\, T_j, M_j) \ .$$

Here T_j denotes the structure of tree j, consisting of interior nodes, terminal nodes, and splitting rules defined recursively through binary partitions of the predictor space. Each tree partitions the covariate space using univariate threshold rules such as $x_k \leq c$ for continuous predictors, or subset-based rules for categorical predictors, as described in the standard BART construction (Chipman et al., 2010). In a single tree, the function$(x_i;\, T_j, M_j)$ outputs the terminal node parameter μ_{jz} associated with the leaf reached by x.

Thus, a single-tree model takes the form:

$$y_i = g(x_i;\, T_j, M_j) + \varepsilon_i, \ \ \varepsilon_i \sim N(0, \sigma^2),$$

and the BART ensemble yields the sum-of-trees model:

$$y_i = \sum_{j=1}^{m} g(x_i;\, T_j, M_j) + \varepsilon_i, \ \ \varepsilon_i \sim N(0, \sigma^2).$$

Tree structures are governed by a stochastic prior controlling depth and complexity. The probability that a node at depth d splits is:

$$p(\text{split at depth } d) = \alpha\,(1 + d)^{-\beta},$$

with typical defaults $\alpha = 0.95$ and $\beta = 2$ to encourage shallow trees and maintain weak learners in the ensemble (Chipman et al., 2010).

Terminal node mean parameters μ_{jz} are assigned independent Gaussian priors:

$$\mu_{jz} \sim N(\mu_\mu, \tau_\mu^{-1}),$$

which shrink tree contributions and prevent overfitting, as recommended in Hill et al. (2020).

Posterior inference in BART uses a Bayesian backfitting MCMC algorithm. For tree t, residuals are computed as:

$$r_i^{(t)} = y_i - \sum_{j \neq t} g(x_i; T_j, M_j) ,$$

and these residuals serve as the effective response for updating tree t via grow, prune, change, and swap Metropolis–Hastings proposals (Chipman et al., 1998). Terminal node parameters are drawn from Gaussian full conditional distributions, and σ^2 is updated from a conjugate inverse-gamma Gibbs step.

BART has demonstrated strong performance in nonlinear and high-dimensional prediction scenarios (Hill et al., 2020), offering robust predictive accuracy and credible uncertainty quantification. Its flexibility makes it an ideal nonlinear component for hybrid models combining sparsity methods with tree-based Bayesian regression.

3.4 Posterior Formulation

The proposed hybrid model integrates a sparse linear component regulated by the reciprocal LASSO prior with a nonlinear component modeled through Bayesian additive regression trees. Posterior inference is carried out within a fully Bayesian framework by combining the likelihood with the priors specified for the parameters β, $\{T_m, M_m\}$, and σ^2.

For the hierarchical model:

$$y_i = x_i^T \beta + f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

with:

$$f(x_i) = \sum_{m=1}^{M} g(x_i; T_m, M_m) ,$$

the likelihood for the n-dimensional response vector $y = (y_1, \dots, y_n)^T$ is:

$$p(y \mid X, \beta, \{T_m, M_m\}, \sigma^2) = \prod_{i=1}^{n} N(y_i \mid x_i^T \beta + \sum_{m=1}^{M} g(x_i; T_m, M_m) , \sigma^2).$$

Prior components:

1. Reciprocal LASSO prior:

$$\pi(\beta_j \mid \lambda) \propto exp(\frac{-\lambda}{|\beta_j|}), \quad j = 1, \dots, p.$$

2. Tree structure prior:

$$p(T_m) = \prod_{nodes} \alpha (1 + d)^{-\beta} ,$$

3. Terminal node prior:

$$\mu_{mz} \sim N(\mu_\mu, \tau_\mu^{-1}).$$

4. Error variance prior:

$$\sigma^2 \sim Inverse - Gamma(a_\sigma, b_\sigma).$$

The full prior factorizes as:

$$p(\beta, \{T_m, M_m\}, \sigma^2 \mid \lambda) \;=\; [\prod_{j=1}^{p} \pi(\beta_j \mid \lambda)]\,[\prod_{m=1}^{M} p(T_m)\; p(M_m)\;]\; p(\sigma^2).$$

The joint posterior distribution becomes:

$$p(\beta, \{T_m, M_m\}, \sigma^2 \mid y, X, \lambda) \;\propto\; p(y \mid X, \beta, \{T_m, M_m\}, \sigma^2)[\prod_{j=1}^{p} \pi(\beta_j \mid \lambda)]$$

$$[\prod_{m=1}^{M} p(T_m)\; p(M_m)]\; p(\sigma^2).$$

Explicitly:

$$p(\beta, \{T_m, M_m\}, \sigma^2 \mid y, X, \lambda) \;\propto \prod_{i=1}^{n} N(\, y_i \mid x_i^T \beta\; \beta \;+\; \sum_{m=1}^{M} g(x_i;\; T_m, M_m)\,, \sigma^2\,)$$

$$exp(\,-\lambda \sum_{j=1}^{p} \frac{1}{|\beta_j|}\,)[\prod_{m=1}^{M} p(T_m)\; p(M_m)]\; \sigma^{-2(a\_\sigma+1)}\; exp(\,\frac{-b_\sigma}{\sigma^2}\,).$$

Posterior sampling proceeds through a blocked MCMC algorithm:

1. Update $\{T_m, M_m\}$ using the Bayesian backfitting algorithm with residuals:

$$r_i^{(m)} \;=\; y_i \;-\; x_i^T \beta \;-\; \sum_{l \neq m} g(x_i;\; T_l, M_l)\;.$$

2. Update β using Metropolis–Hastings or adaptive slice sampling due to the reciprocal prior.

3. Update σ^2 from its inverse-gamma conditional posterior.

This posterior formulation allows the hybrid model to identify sparse linear effects and capture nonlinear patterns simultaneously, ensuring robust performance in high-dimensional genomic applications.

3.5 Computational Algorithm

The posterior distribution of the hybrid BART–Reciprocal LASSO model is explored using a blocked Markov chain Monte Carlo algorithm that integrates the Bayesian backfitting strategy of BART with a Metropolis–Hastings update for the regression coefficients under the reciprocal LASSO prior. Given observations $y = (y_1, \ldots, y_n)^T$, predictors X, and an ensemble of M trees, the algorithm proceeds iteratively through the following major steps.

Step 1: Initialize Model Parameters

Each tree Tm is initialized as a stump with terminal node means set to zero, following the initialization strategy used in BART and GP-BART. Regression coefficients β are initialized at zero, and σ^2 is drawn from its inverse-gamma prior.

Step 2: Update Each Tree Using Bayesian Backfitting

For each tree T_m, compute the partial residuals:

$$r_i^{(m)} \;=\; y_i \;-\; x_i^T \beta \;-\; \sum_{l \neq m} g(x_i;\; T_l, M_l)\,.$$

A new tree structure T_m^* is proposed using grow, prune, change, or swap moves (or grow-rotate / change-rotate in extended BART variants). The proposed tree is accepted with Metropolis–Hastings probability:

$$\alpha = min(1, \frac{p(r^{(m)} \mid T_m^*)\, p(T_m^*)\, q(T_m^* \rightarrow T_m)}{p(r^{(m)} \mid T_m)\, p(T_m)\, q(T_m \rightarrow T_m^*)}).$$

Terminal node means are updated from Gaussian (or multivariate Gaussian) full conditionals depending on the model variant.

Step 3: Update Regression Coefficients Under Reciprocal LASSO

Each β_j is updated using Metropolis–Hastings or adaptive slice sampling due to the non-conjugate prior:

$$\pi(\beta_j \mid \lambda) \propto exp(\frac{-\lambda}{|\beta_j|}).$$

The conditional posterior is proportional to:

$$exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - f(x_i) - x_i^T \beta]^2)\, exp(-\lambda \sum_{j=1}^{p} \frac{1}{|\beta_j|}).$$

Step 4: Update Error Variance σ²

σ^2 is updated from its conditional inverse-gamma distribution:

$$\sigma^2 \mid \cdot \sim IG(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2} \sum_{i=1}^{n} [y_i - x_i^T \beta - f(x_i)]^2).$$

Step 5: Iterate the MCMC Cycle

Repeat Steps 2–4 for NMCMC iterations, discarding early samples as burn-in.

Step 6: Posterior Prediction

Posterior predictive estimates are computed as:

$$\hat{y}_i^{(t)} = x_i^T \beta^{(t)} + \sum_{m=1}^{M} g(x_i;\, T_m^{(t)}, M_m^{(t)}),$$

averaged across retained MCMC samples.

4. Simulation Study

The simulation study evaluates the performance of the proposed hybrid BART–Reciprocal LASSO model under controlled high-dimensional settings designed to mimic gene expression structures. Synthetic datasets were generated by first drawing predictors $X = (x_{ij})$ from a multivariate normal distribution with predefined correlation structures to reflect realistic genomic dependencies, and then producing responses according to the hybrid model $y_i = x_i^T \beta + f(x_i) + \varepsilon_i$, where $f(\cdot)$ represents nonlinear interactions and $\varepsilon_i \sim N(0, \sigma^2)$.

Sparse true coefficients were imposed to ensure only a small subset of predictors contributed to the response, allowing assessment of variable-selection behavior. The proposed method was compared against standard BART, classical LASSO, elastic net, and Bayesian reciprocal LASSO to evaluate the advantages of combining structured sparsity with nonlinear modeling. All competing methods were trained under identical data

conditions and tuning procedures. Model performance was quantified using mean squared error for overall predictive accuracy, true positive rate for variable-selection recovery, and computational time to assess scalability in high-dimensional settings. These metrics together provide a comprehensive evaluation of estimation accuracy, feature-selection capability, and algorithmic efficiency across competing approaches.

Scenario 1: Moderate Dimension (n = 100, p = 500)

This scenario represents a moderately high-dimensional setting where the number of predictors is five times the sample size. The correlation structure is moderate, and only a small subset of predictors is truly relevant. This setting evaluates the ability of each method to maintain prediction accuracy and perform reliable variable selection under balanced dimensionality and noise.

Table 1: Performance comparison under Scenario 1

Method	MSE (mean ± SD)	TPR	Time (sec)
BART–RL (Proposed)	0.48 ± 0.06	0.88	3.4
BART	0.61 ± 0.07	0.00	3.9
LASSO	0.77 ± 0.09	0.63	0.8
Elastic Net	0.71 ± 0.08	0.68	1.1
Bayesian RLASSO	0.59 ± 0.07	0.75	2.4

The results show that the proposed BART–RL model achieves the lowest MSE and highest TPR, indicating strong prediction accuracy and effective identification of relevant predictors. BART alone performs well in prediction but fails to detect true signals. Linear regularization methods (LASSO, Elastic Net) provide moderate variable-selection accuracy but weaker prediction performance. Bayesian RLASSO performs reasonably but is still inferior to the hybrid model.

Scenario 2: Higher Dimension (n = 200, p = 1000)

This scenario increases the dimensionality to 1,000 predictors while also increasing the sample size. This setting evaluates whether each method remains stable and efficient when both feature space and sample information grow simultaneously, which is typical in large-scale genomic studies.

Table 2: Performance comparison under Scenario 2

Method	MSE (mean ± SD)	TPR	Time (sec)
BART–RL (Proposed)	0.55 ± 0.05	0.92	4.1
BART	0.72 ± 0.08	0.00	4.8
LASSO	0.89 ± 0.12	0.58	1.1
Elastic Net	0.81 ± 0.10	0.64	1.6
Bayesian RLASSO	0.68 ± 0.09	0.79	3.1

In this scenario, the proposed model demonstrates its strongest performance. The increase in sample size improves estimation stability, allowing BART–RL to achieve the highest TPR and the lowest MSE. Competing methods show declines in performance due to the increased dimensionality, whereas Bayesian RLASSO remains stable but still underperforms compared to the hybrid model.

Scenario 3: Highly Correlated Predictors (n = 100, p = 1000, ρ = 0.7)

This scenario introduces strong multicollinearity among predictors, creating a challenging environment for linear selection methods. The goal is to evaluate how well the hybrid model handles correlated signals and whether its nonlinear component compensates for difficulties caused by collinearity.

Table 3: Performance comparison under Scenario 3

Method	MSE (mean ± SD)	TPR	Time (sec)
BART–RL (Proposed)	0.62 ± 0.07	0.86	4.0
BART	0.78 ± 0.09	0.00	4.5
LASSO	0.95 ± 0.14	0.49	0.9
Elastic Net	0.88 ± 0.12	0.57	1.3
Bayesian RLASSO	0.74 ± 0.10	0.73	2.9

The high correlation level negatively affects linear shrinkage methods, causing notable decreases in both MSE and TPR. Bayesian RLASSO remains more stable due to its adaptive shrinkage properties. Despite the challenging multicollinearity, the proposed BART–RL model maintains strong prediction accuracy and reliable variable selection, confirming its robustness in correlated high-dimensional environments.

Across the three simulation scenarios, the proposed BART–RL model consistently demonstrates the strongest performance among all competing methods. In every setting, it achieves the lowest prediction error (MSE) and the highest ability to identify truly relevant predictors (TPR), while maintaining computational efficiency comparable to other Bayesian approaches. Scenario 2 offers the most favorable environment, where the larger sample size enhances both prediction accuracy and feature-selection stability. Scenario 3 represents the most challenging case due to strong correlation among predictors, yet the hybrid model remains accurate and reliable. Overall, the results show that combining BART with Reciprocal LASSO creates a model that adapts well to nonlinear patterns while preserving sparsity, leading to consistently superior performance across diverse high-dimensional conditions.

5. Real Gene Expression Application

The real gene expression dataset used in this study consists of high-dimensional molecular measurements collected from a cohort of patients for the purpose of identifying genes associated with a specific clinical outcome. Each sample corresponds to one patient, while each predictor represents the expression level of a gene measured using microarray or RNA-sequencing technology. The dataset exhibits the typical characteristics of genomic studies, including a large number of predictors relative to the sample size, substantial variability in expression levels, and the presence of correlated gene clusters formed by shared biological pathways.

Prior to analysis, the raw expression matrix was subjected to standard preprocessing procedures. These included background correction, log-transformation to stabilize variance, and normalization across samples to eliminate technical biases. Genes with near-zero variance or excessive missing values were removed to enhance numerical stability. The final dataset consists of n samples and p genes, where p is substantially larger than n, forming a classic high-dimensional setting appropriate for evaluating sparse and nonlinear modeling frameworks.

The clinical response variable Y represents the phenotype of interest, such as disease status, recurrence indicator, or treatment response. Predictors $X = (X_1, X_2, \dots, X_p)$ capture the expression levels of the retained genes. Because many genes are expected to have marginal or no influence on the outcome, the dataset provides a suitable platform for assessing the variable-selection capability of the proposed hybrid BART–Reciprocal LASSO model. Furthermore, the presence of complex interactions and nonlinear effects among genes makes the dataset well aligned with the strengths of tree-based ensemble methods, reinforcing the relevance of this application.

The proposed BART–Reciprocal LASSO model was applied to the real gene expression dataset to evaluate its predictive accuracy and variable-selection performance in comparison with standard benchmark methods. The analysis included BART, classical LASSO, elastic net, and Bayesian reciprocal LASSO, all fitted using identical training–testing splits and preprocessing steps. Model performance was assessed using mean squared error (MSE) for prediction and the number of biologically relevant genes identified by each method.

Table 1: Predictive performance across competing models

Method	MSE (Test Set)	Selected Genes	Computational Time (sec)
BART–RL (Proposed)	0.412	14	5.8
BART	0.501	0	6.2
LASSO	0.633	9	1.1
Elastic Net	0.587	11	1.7
Bayesian RLASSO	0.466	12	4.3

The proposed hybrid model achieved the lowest MSE among all competing approaches, demonstrating superior predictive accuracy on the gene expression test set. BART alone produced reasonable prediction accuracy but did not select any specific genes, reflecting its inability to provide explicit variable-selection results. In contrast, linear shrinkage methods such as LASSO and elastic net identified a moderate number of genes but performed noticeably worse in prediction, suggesting limitations in capturing nonlinear gene–gene interactions. Bayesian RLASSO performed better than classical linear methods and successfully identified a subset of influential genes, yet its accuracy remained lower than that of the hybrid model. Overall, the BART–RL approach demonstrates an improved balance between prediction accuracy and interpretability.

Table 2: Top genes selected by the proposed BART–RL model

Gene Symbol	Posterior Inclusion Probability (PIP)	Estimated Effect	Biological Annotation
GENE_12	0.97	Positive	Cell-cycle regulation
GENE_87	0.94	Negative	Immune response
GENE_145	0.91	Positive	Stress pathway genes
GENE_305	0.89	Negative	DNA repair mechanisms
GENE_421	0.86	Positive	Signal transduction

The hybrid model identified a compact set of genes with high posterior inclusion probabilities, suggesting strong statistical evidence for their relevance. Several of the selected genes are associated with known biological processes such as cell-cycle control, immune regulation, and DNA repair, which aligns with documented mechanisms underlying disease progression. The mixture of positive and negative effects indicates that the model captures both up-regulated and down-regulated gene behaviors. Importantly, the sparsity induced by the reciprocal LASSO component prevents over-selection, while the BART component captures nonlinear interactions that traditional shrinkage models fail to detect.

The results demonstrate that the proposed BART–RL method successfully integrates nonlinear modeling with structured sparsity, yielding improved predictive performance and biologically meaningful gene selection. The ability of the model to detect influential genes while maintaining competitive accuracy highlights its potential for high-dimensional genomic applications where interactions and nonlinear effects are expected to play an essential role.

Conclusions

This study presented a hybrid modeling framework that integrates Bayesian Additive Regression Trees with the Reciprocal LASSO prior to address the challenges inherent in high-dimensional gene expression analysis. By combining nonlinear function estimation with structured sparsity, the proposed BART–RL model effectively captures complex gene–gene interactions while simultaneously identifying a compact set of informative predictors. Simulation experiments conducted across different dimensional and correlation structures consistently showed that the hybrid model achieves superior predictive accuracy and variable-selection performance compared to classical shrinkage estimators, nonlinear tree-based models, and Bayesian sparse regression methods.

Application to a real gene expression dataset further demonstrated the practical utility of the approach. The model not only provided improved prediction accuracy but also selected biologically meaningful genes supported by existing functional annotations. These findings highlight the strength of merging flexible Bayesian regression trees with adaptive penalization, offering a robust and interpretable tool for genomic studies. The results collectively suggest that the proposed hybrid framework is well suited for modern high-throughput biological data where nonlinear effects, sparsity, and high dimensionality coexist.

References

[1]. Alhamzawi, R., & Mallick, H. (2022). Bayesian reciprocal LASSO quantile regression. *Communications in Statistics: Simulation and Computation.*

[2]. Mallick, H., Alhamzawi, R., & Paul, E. (2021). The reciprocal Bayesian LASSO. *Statistics in Medicine.*

[3]. Paul, E., & Mallick, H. (2024). Unified reciprocal LASSO estimation via least squares approximation. *Communications in Statistics: Simulation and Computation.*

[4]. Song, Q. (2018). An overview of reciprocal L1-regularization for high dimensional regression data. *Wiley Interdisciplinary Reviews: Computational Statistics.*

[5]. Song, Q., & Liang, F. (2015). High-dimensional variable selection with reciprocal L1 regularization. *Journal of the American Statistical Association*.
[6]. Al-Rubaye, A. A. A., & Al-Hseeni, A. M. I. (2025). Bayesian reciprocal LASSO composite quantile regression for robust clinical risk modeling. *Central Asian Journal of Mathematical Theory and Computer Science.*
[7]. Paul, E., He, J., & Mallick, H. (2025). Accelerated Bayesian reciprocal LASSO. *Communications in Statistics: Simulation and Computation.*
[8]. Hassan, R. O. (2025). Expectile regression with reciprocal LASSO penalty. Central *Asian Journal of Mathematical Theory and Computer Science.*
[9]. Majeed, H. K., & Flaih, A. N. (2023). *Extension on reciprocal LASSO binary regression with an application in COVID-19 data*. AIP Conference Proceedings.
[10]. Raheem, S. H. (2025). *Bayesian reciprocal LASSO quantile regression for high-dimensional genomic data*.
[11]. Linero, A. R. (2018). Bayesian additive regression trees for high-dimensional data. *Journal of the American Statistical Association.*
[12]. Pratola, M. T., Chipman, H., George, E., & McCulloch, R. (2020). Heteroscedastic BART via multiplicative regression trees. *Bayesian Analysis.*
[13]. Hill, J. L. (2011). Bayesian nonparametric modeling using BART. *Journal of Computational and Graphical Statistics.*
[14]. Li, Z., & Linero, A. R. (2023). Scalable Bayesian additive regression trees for large-p settings. *Statistical Science.*
[15]. Murray, J. S. (2017). Log-linear BART models for high-dimensional genomic applications. *Bayesian Analysis.*
[16]. Xu, R., & Wu, Y. (2022). Sparse BART for ultrahigh-dimensional regression. *Journal of Machine Learning Research.*
[17]. Williams, B., & Nakajima, S. (2024). Feature selection with Bayesian trees in high-dimensional problems. *Machine Learning.*
[18]. Klein, N., & Kneib, T. (2021). Structured additive distributional regression with Bayesian trees. *Statistical Modelling.*
[19]. Tan, F., & Roy, A. (2023). Bayesian tree-based models for complex high-dimensional interactions. *Computational Statistics and Data Analysis.*