mathematics

MDPI

Article

A Hierarchical Bayesian Cox Model with Variational Inference for High-Dimensional Survival Analysis

Bulus I. Doroh [1] and Kazeem A. Dauda [2,*]

[1] Department of Mathematics and Statistics, Kwara State University, Malete, P.M.B. 1530, Ilorin 23431, Kwara State, Nigeria
[2] Department of Mathematics, University of Bergen, 5007 Bergen, Norway
* Correspondence: kazeem.dauda@uib.no or qdauda70@gmail.com

Abstract

Large-scale molecular datasets have become increasingly important in bioinformatics, providing a foundation for prognostic modeling and the identification of biologically relevant genes. In survival analysis, however, censored observations and the high-dimensional nature of genomic data pose significant challenges for reliable prediction and variable selection. To address these challenges, we develop a Penalized Hierarchical Bayesian Cox (PHBC) modeling framework that extends hierarchical shrinkage and global–local prior formulations using Adaptive Gaussian Priors (AGP) and a modified Global–Local Gaussian–Inverse-Gamma Hyperprior (Gaussian–IGH) structure. The framework consists of two hierarchical components: (i) Adaptive Gaussian priors assigned to regression coefficients, with coefficient-specific variances updated from the data through an empirical-Bayes rule, and (ii) Inverse-Gamma hyperpriors on variance components to provide flexible and adaptive shrinkage. Inference is based on the Cox partial likelihood, yielding a partial-likelihood (pseudo-Bayesian) posterior. We derive the corresponding posterior kernels and construct scalable variational inference approximations, resulting in two computationally efficient procedures: AGP-VI and Global–Local Gaussian–IGH-VI. The methods are evaluated through extensive simulation studies and applications to six low- and high-dimensional genomic survival datasets. Performance is assessed using the concordance index (C-index), Integrated Brier Score, time-dependent AUC, and true/false positive rates for variable selection. A dedicated simulation study further examines sensitivity to multicollinearity under weak, moderate, and high predictor correlations, with the proposed methods remaining competitive across the different correlation levels. Across the real-data applications, no single method uniformly dominated all predictive criteria. The proposed variational methods showed strong but dataset-dependent performance, with Global–Local Gaussian–IGH-VI achieving the best overall predictive performance on one dataset and the best C-index on another, while AGP-VI achieved the best discrimination on two others. Classical penalized and Bayesian Cox approaches remained competitive, particularly in calibration. These results demonstrate the complementary strengths and trade-offs of the proposed methods in high-dimensional genomic survival analysis.

Keywords: high-dimensional survival analysis; Bayesian Cox Model; hierarchical shrinkage priors; Adaptive Gaussian Prior; Inverse-Gamma hyperprior; variational inference; variable selection

MSC: 62N01

Check for updates

Academic Editor: Yanchun Bao

Received: 25 August 2026
Revised: 6 October 2026
Accepted: 14 October 2026
Published: 16 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

In genetic research, cancer is one of the most complex and heterogeneous diseases, developing through multiple molecular and genetic alterations that differ among patients and across tissue types [1,2]. Understanding these different molecular characteristics across tissue types is essential for identifying reliable prognostic biomarkers and developing personalized treatment plans [3,4]. However, there are complex hierarchical and overlapping structures among gene predictors, which are often underutilized in genetic modeling [5,6]. To improve interpretability and predictive accuracy in genomic datasets, it is essential to develop survival modeling frameworks that can accommodate complex dependence structures and high-dimensional genomic information. Survival analysis, a statistical tool used for studying time-to-event outcomes, has become increasingly relevant for integrating these molecular and genetic data to identify prognostic biomarkers and predict disease progression [7,8]. A primary challenge, however, is that the number of genes often far exceeds the number of samples, and the presence of irrelevant or redundant genes can undermine the accuracy of survival models [9,10]. The Cox proportional hazards (Cox-PH) model [11] is one of the most widely applied regression models for time-to-event data; its semi-parametric formulation allows estimation of covariate effects without the need to specify the baseline hazard function explicitly [12–15]. However, standard maximum partial-likelihood estimation becomes increasingly challenging as the number of covariates p grows relative to the sample size n, particularly in high-dimensional settings where $p \gg n$.

Specifically, when applied to modern high-dimensional or noisy genomic data with $p \gg n$, the maximum partial-likelihood estimator can become unstable, non-unique, or undefined, and multicollinearity among correlated genes can further inflate estimation variance, limiting the model's capacity to identify relevant variables reliably [9,16]. As datasets become more complex with increasing numbers of predictors and dependence structures, challenges such as overfitting, unstable parameter estimates, and multicollinearity become more pronounced. At the same time, small sample sizes and complicated models make it harder to quantify uncertainty reliably. Consequently, various penalization and Bayesian shrinkage methods have been proposed to address these challenges, particularly in variable selection, shrinkage, and uncertainty quantification. Among the early methods developed to regulate model complexity are the Lasso penalty [17], ridge regularization [18], and their further extensions [19,20]. These methods are consistently utilized in penalized modeling applications, though each has inherent drawbacks. Ridge regression cannot perform sparse variable selection, while the Lasso struggles when explanatory variables are strongly correlated. To overcome these weaknesses, the minimax concave penalty (MCP) [21] and the Elastic Net hybrid penalty introduced greater adaptability by balancing regularization and selection and have proven effective in genomic prediction research [22,23]. Hybrid strategies, such as penalized Cox models combined with boosting or ensemble methods [24], further capture nonlinear effects but often compromise interpretability under censoring. While these penalized methods control model complexity effectively, Bayesian shrinkage priors provide a probabilistic alternative that incorporates uncertainty quantification directly into the inference process.

Priors such as the Horseshoe [25] and its extensions [26], spike-and-slab [27], and global–local structures [28] enhance flexibility for sparsity and parameter shrinkage, enabling more adaptive learning in high-dimensional settings. However, these approaches can be sensitive to hyperparameter specification and computationally demanding. Building on these developments, regularized global–local variants provide additional stability in high-dimensional settings by aggressively shrinking irrelevant predictors while preserving

signals for relevant ones. Despite these advantages, careful specification of prior scales remains crucial, particularly in ultra-high-dimensional regimes.

Principled uncertainty quantification is also a central challenge in high-dimensional analysis, where complex dependence structures and limited sample sizes undermine reliable estimation of variability and uncertainty. Bayesian approaches such as Bayesian Model Averaging [29] and stochastic search variable selection [30] attempt to address model uncertainty; however, they may yield unstable inclusion probabilities in the presence of multicollinearity. Multicollinearity itself represents a critical issue in high-dimensional settings, as strong correlations among predictors can lead to unstable parameter estimates and unreliable inference [31]. Consequently, the direct application of these methods to high-dimensional survival analysis is challenged by (i) instability in variable selection, (ii) inefficient shrinkage, (iii) challenges in reliable uncertainty quantification, and (iv) limited ability to handle multicollinearity.

To address these limitations, this study builds on existing hierarchical shrinkage and global–local prior formulations [25,28,32] to construct a Penalized Hierarchical Bayesian Cox (PHBC) modeling framework tailored for high-dimensional survival data with censoring and multicollinearity. The framework combines an Adaptive Gaussian Prior (AGP) with a modified Inverse-Gamma Hyperprior (IGH) structure on the variance components, adapting these established shrinkage devices to the Cox partial-likelihood setting and to a variational (rather than MCMC-based) computational scheme. By allowing coefficient-specific regularization and adaptive control of shrinkage strength, the model aims to improve parameter stability and variable-selection sparsity in ultra-high-dimensional settings. We treat improved uncertainty quantification as a hypothesis to be examined empirically (Section 5) rather than as an established property of the method, since our variational approximation does not itself propagate uncertainty in the shrinkage-variance parameters. Posterior inference is performed using variational inference (VI) for scalable approximation, resulting in computationally efficient procedures suitable for large-scale genomic survival applications.

This study makes three main contributions. First, it adapts and combines adaptive-Gaussian and global–local Inverse-Gamma hierarchical shrinkage formulations into a unified partial-likelihood-based hierarchical Cox framework for high-dimensional variable selection. Second, it develops scalable variational approximation procedures (AGP-VI and Global–Local Gaussian–IGH-VI) that combine coordinate-ascent variational updates with empirical-Bayes point estimates for the variance components, achieving a practical balance between computational efficiency and estimation accuracy in large molecular datasets. Third, extensive simulation studies and real genomic survival applications examine the predictive performance, variable-selection behavior, and computational scalability of the proposed procedures relative to existing Bayesian and penalized Cox models, and report both the settings in which the proposed procedures perform favorably and those in which competing methods remain preferable.

The remainder of the paper is organized as follows. Section 2 presents the PHBC model formulation, the specification of the hierarchical priors, and the posterior inference procedures for the two proposed methods, AGP-VI and Global–Local Gaussian–IGH-VI, based on variational inference. Section 3 describes the simulation design under high-dimensional settings, while Section 4 presents the simulation results. Section 5 reports the applications to six real-world genomic survival datasets. Finally, Section 6 concludes the paper with a discussion of the main findings, limitations, and directions for future research.

2. Methodology

We consider a Penalized Hierarchical Bayesian Cox model with many structured predictors, where traditional Cox models often produce unstable estimates. To address this limitation, we incorporate hierarchical shrinkage priors that enable adaptive regularization of regression coefficients. In particular, we employ two related shrinkage formulations: an Adaptive Gaussian Prior (AGP) and a hierarchical Gaussian–Inverse-Gamma (Global–Local Gaussian–IGH) prior.

Furthermore, variational inference (VI) is adopted for fast and scalable posterior approximation, making the proposed approach suitable for large biomedical datasets [33,34].

2.1. Penalized Cox Proportional Hazards Model

In Bayesian modeling, analyzing survival data in high-dimensional settings requires integrating uncertainty and regularization within a unified probabilistic framework. Hierarchical priors provide a principled way to achieve this by controlling model complexity, stabilizing estimation, and promoting sparsity when the number of predictors exceeds the number of available observations. Through such priors, the Bayesian framework balances flexibility and parsimony by shrinking weak or irrelevant effects toward zero while preserving influential ones, thereby improving interpretability and predictive reliability.

Building on this, we define the observed survival data and hazard structure. Each subject i contributes an observation $Y_i = (t_i, d_i)$, where t_i is the observed time and d_i indicates whether the event occurred ($d_i = 1$) or was censored ($d_i = 0$). Let T_i and C_i denote the true survival and censoring times, such that $t_i = \min(T_i, C_i)$, with $d_i = 1$ if $T_i \leq C_i$ and $d_i = 0$ otherwise. Each individual is associated with a covariate vector $\mathbf{x}_i \in \mathbb{R}^p$.

Under this setting, the hazard function is modeled as

$$h(t \mid \mathbf{x}_i) = h_0(t) \exp(\mathbf{x}_i^\top \boldsymbol{\beta}), \tag{1}$$

where $h_0(t)$ is the baseline hazard and $\boldsymbol{\beta}$ represents regression coefficients, following the original Cox proportional hazards formulation [11,35].

Partial Likelihood

A key feature of the Cox model is the use of partial likelihood, which allows estimation of $\boldsymbol{\beta}$ without specifying $h_0(t)$. Let $R(t_i)$ denote the risk set at time t_i. The conditional probability that subject i experiences the event at t_i is

$$P(i \text{ fails at } t_i \mid R(t_i)) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\mathbf{x}_j^\top \boldsymbol{\beta})}. \tag{2}$$

The partial likelihood is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\sum_{j \in R(t_i)} \exp(\mathbf{x}_j^\top \boldsymbol{\beta})} \right]^{d_i}, \tag{3}$$

with log-partial likelihood

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} d_i \left(\mathbf{x}_i^\top \boldsymbol{\beta} - \log \sum_{j \in R(t_i)} \exp(\mathbf{x}_j^\top \boldsymbol{\beta}) \right). \tag{4}$$

Unlike a fully specified likelihood for the survival times, the Cox partial likelihood eliminates the unspecified baseline hazard $h_0(t)$ and permits inference on $\boldsymbol{\beta}$ without di-

rectly modeling $h_0(t)$ [11]. Accordingly, we combine the partial likelihood with the prior distribution on $\boldsymbol{\beta}$ to define

$$p(\boldsymbol{\beta} \mid \mathcal{D}) \propto L(\boldsymbol{\beta} \mid \mathcal{D})\, p(\boldsymbol{\beta}). \tag{5}$$

We refer to this quantity as a partial-likelihood (pseudo-Bayesian) posterior to distinguish it from a posterior obtained from a fully specified Bayesian survival model that also assigns a prior to the baseline hazard. The variational inference procedures developed in Section 2.3 approximate this partial-likelihood posterior, thereby allowing hierarchical shrinkage priors to be incorporated without explicitly specifying or estimating $h_0(t)$.

2.2. Bayesian Hierarchical Prior Specification

When the number of predictors exceeds the number of observations, regularization is essential to prevent overfitting and ensure stable estimation [17]. In the Bayesian framework, this is achieved through prior distributions that shrink coefficients toward zero while preserving important signals [32].

In this study, we adopt two complementary shrinkage formulations: the Adaptive Gaussian Prior (AGP) and the Gaussian–Inverse-Gamma (Global–Local Gaussian–IGH) prior. These priors allow flexible control of shrinkage at the coefficient level and enhance robustness in high-dimensional settings.

2.2.1. Joint Posterior Distribution via Adaptive Gaussian Prior (AGP)

Under the Adaptive Gaussian Prior (AGP) formulation, conditional on the coefficient-specific variance parameters, the joint posterior distribution of the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ given the observed data $\mathcal{D}$ is expressed as

$$p(\boldsymbol{\beta} \mid \mathcal{D}, \boldsymbol{\sigma}^2) \propto L(\boldsymbol{\beta} \mid \mathcal{D})\, p(\boldsymbol{\beta} \mid \boldsymbol{\sigma}^2), \tag{6}$$

where $L(\boldsymbol{\beta} \mid \mathcal{D})$ is the Cox partial likelihood and $p(\boldsymbol{\beta} \mid \boldsymbol{\sigma}^2)$ is the Gaussian prior with coefficient-specific variances.

The prior is defined as

$$\beta_j \mid \sigma_j^2 \sim \mathcal{N}(0, \sigma_j^2), \quad j = 1, \ldots, p, \tag{7}$$

with density

$$p(\boldsymbol{\beta} \mid \boldsymbol{\sigma}^2) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{\beta_j^2}{2\sigma_j^2}\right), \tag{8}$$

where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_p^2)$ denotes the coefficient-specific variance parameters, treated as deterministic and estimated from the data via the empirical-Bayes update given in Section 2.3.

Substituting the likelihood and prior density, the posterior kernel becomes

$$p(\boldsymbol{\beta} \mid \mathcal{D}, \boldsymbol{\sigma}^2) \propto \prod_{i \in \mathcal{E}} \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\sum_{l \in R(t_i)} \exp(\mathbf{x}_l^\top \boldsymbol{\beta})} \times \prod_{j=1}^{p} (\sigma_j^2)^{-\frac{1}{2}} \exp\left(-\frac{\beta_j^2}{2\sigma_j^2}\right). \tag{9}$$

The σ_j^2 in Equation (9) is not assigned a prior distribution but is instead a plug-in quantity re-estimated from the current fit (Equation (16)); the AGP formulation is more accurately characterized as an empirical-Bayes adaptive shrinkage procedure operating on the Cox partial likelihood, rather than as a fully specified hierarchical Bayesian model in which uncertainty about σ_j^2 is itself represented and propagated. In this respect, AGP is conceptually closer to empirical-Bayes adaptive shrinkage procedures than to fully Bayesian scale-mixture priors such as the Horseshoe [25], in which the local scale parameter is assigned its own prior distribution and estimated through the posterior. We retain

the label "AGP-VI" for continuity with the Global–Local Gaussian–IGH-VI procedure described next, but do not claim that AGP provides full uncertainty quantification for its shrinkage variances. The credible intervals reported for $\boldsymbol{\beta}$ are conditional on the plug-in estimates $\hat{\sigma}_j^2$ and therefore do not account for uncertainty associated with estimation of these coefficient-specific variances.

2.2.2. Gaussian–Inverse-Gamma (Gaussian–IGH) Prior

To further enhance shrinkage flexibility, we consider a hierarchical Gaussian prior:

$$\beta_j \mid \tau_j^2 \sim \mathcal{N}(0, \tau_j^2), \quad \tau_j^2 \sim \text{Inverse-Gamma}(a, b), \quad j = 1, \ldots, p. \tag{10}$$

The probability density of β ($p(\boldsymbol{\beta} \mid \tau^2)$) is defined as

$$p(\boldsymbol{\beta} \mid \tau^2) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\tau_j^2}} \exp\left(-\frac{\beta_j^2}{2\tau_j^2}\right). \tag{11}$$

while the density prior of τ is given by

$$p(\tau^2) = \prod_{j=1}^{p} \frac{b^a}{\Gamma(a)} (\tau_j^2)^{-(a+1)} \exp\left(-\frac{b}{\tau_j^2}\right), \tag{12}$$

where $a > 0$ and $b > 0$ are the shape and scale hyperparameters, respectively.

This formulation defines a Gaussian scale mixture, which induces a heavy-tailed marginal distribution for β_j. As a result, the model can simultaneously shrink noise variables toward zero while preserving large signals. The hierarchical structure also improves uncertainty quantification and robustness in ultra-high-dimensional settings.

2.2.3. Joint Posterior Distribution via Global–Local Gaussian–IGH Prior

Under the Global–Local Gaussian–Inverse-Gamma (Gaussian–IGH) formulation, the joint posterior distribution of the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, the local variance parameters $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_p^2)$, and the global variance parameter τ_0^2 given the observed data $\mathcal{D}$ is expressed as

$$p(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \tau_0^2 \mid \mathcal{D}) \propto L(\boldsymbol{\beta} \mid \mathcal{D})\, p(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2, \tau_0^2)\, p(\boldsymbol{\tau}^2)\, p(\tau_0^2), \tag{13}$$

where $L(\boldsymbol{\beta} \mid \mathcal{D})$ is the Cox partial likelihood, $p(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2, \tau_0^2)$ is the conditional Gaussian prior, $p(\boldsymbol{\tau}^2)$ is the local Inverse-Gamma hyperprior, and $p(\tau_0^2)$ is the global Inverse-Gamma hyperprior.

The hierarchical prior structure is specified as follows. Conditional on the local variance parameter τ_j^2 and the global variance parameter τ_0^2, the regression coefficients follow independent Gaussian distributions:

$$\beta_j \mid \tau_j^2, \tau_0^2 \sim \mathcal{N}(0, \sigma_j^2), \qquad j = 1, \ldots, p,$$

where the effective coefficient variance is defined as

$$\sigma_j^2 = \sqrt{\tau_0^2}\, \tau_j^2.$$

To avoid ambiguity, the square root applies only to τ_0^2, while τ_j^2 enters multiplicatively. Hence, the intended variance specification is $\sigma_j^2 = \sqrt{\tau_0^2}\, \tau_j^2$, not $\sigma_j^2 = \sqrt{\tau_0^2 \tau_j^2}$. All subsequent derivations (Equations (14)–(17)) are based on the former specification.

Unlike conventional global–local shrinkage models that use the multiplicative variance specification $\tau_0^2\tau_j^2$ [25,26], we adopt a modified global–local formulation based on $\sigma_j^2 = \sqrt{\tau_0^2}\,\tau_j^2$, where the global variance enters through its square root rather than directly. This is a deliberate departure from the classical construction, introduced for numerical reasons. It reduces the influence of the global variance component during optimization and empirically improved the numerical stability of the variational algorithm and reduced variance collapse in our high-dimensional experiments. This modification changes the marginal distribution implied for β_j after integrating out τ_j^2 and τ_0^2. Therefore, β_j under this construction is not assumed to follow the same known heavy-tailed marginal family associated with classical global–local specifications. We do not derive or claim a closed-form marginal distribution for β_j under the modified variance specification, but treat it as a working prior supported by its empirical convergence behavior. For consistency, the posterior kernel in Equation (14) is derived directly from the modified variance $\sigma_j^2 = \sqrt{\tau_0^2}\,\tau_j^2$.

The local variance parameters follow independent Inverse-Gamma distributions,

$$\tau_j^2 \sim \text{Inv-Gamma}(a, b), \qquad j = 1, \ldots, p,$$

with density

$$p(\tau_j^2) = \frac{b^a}{\Gamma(a)}(\tau_j^2)^{-(a+1)} \exp\left(-\frac{b}{\tau_j^2}\right).$$

The global variance parameter follows an independent Inverse-Gamma prior,

$$\tau_0^2 \sim \text{Inv-Gamma}(a_0, b_0),$$

with density

$$p(\tau_0^2) = \frac{b_0^{a_0}}{\Gamma(a_0)}(\tau_0^2)^{-(a_0+1)} \exp\left(-\frac{b_0}{\tau_0^2}\right).$$

Substituting the prior densities, the posterior kernel becomes

$$p(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \tau_0^2 \mid \mathcal{D}) \propto L(\boldsymbol{\beta} \mid \mathcal{D}) \times \prod_{j=1}^{p} (\tau_j^2)^{-\left(a+\frac{3}{2}\right)} \exp\left(-\frac{\beta_j^2}{2\sqrt{\tau_0^2}\tau_j^2} - \frac{b}{\tau_j^2}\right) \times (\tau_0^2)^{-\left(a_0+1+\frac{p}{4}\right)} \exp\left(-\frac{b_0}{\tau_0^2}\right). \tag{14}$$

Since $\sigma_j^2 = \sqrt{\tau_0^2}\,\tau_j^2$, the Gaussian density contributes $(\sigma_j^2)^{-1/2} = (\tau_0^2)^{-1/4}(\tau_j^2)^{-1/2}$ for each coefficient. Combining $(\tau_j^2)^{-1/2}$ with the Inverse-Gamma density $(\tau_j^2)^{-(a+1)} \exp(-b/\tau_j^2)$ gives the local exponent $-(a + \frac{3}{2})$. Across all p coefficients, the global contribution is $(\tau_0^2)^{-p/4}$, which combines with $(\tau_0^2)^{-(a_0+1)} \exp(-b_0/\tau_0^2)$ to give the global exponent $-(a_0 + 1 + \frac{p}{4})$ in Equation (14). These exponents follow directly from the modified variance specification and differ from those obtained under the classical $\tau_0^2\tau_j^2$ specification.

2.3. Computational Scheme

The study estimates the parameters of the proposed posterior distributions in Equations (9) and (14) using variational inference (VI). Since the posterior distributions cannot be derived analytically due to the non-conjugacy of the Cox partial likelihood, computational approximation methods are required. Markov Chain Monte Carlo (MCMC) methods are widely recognized for producing reliable posterior samples; however, they can become computationally intensive and prohibitively slow for large-scale, high-dimensional survival datasets [36]. To address this limitation, we implement VI as a faster and more scalable alternative [33,34].

In VI, the true posterior distribution is approximated by a multivariate Gaussian variational distribution. Specifically, we approximate the posterior of the regression coefficients using

$$q(\boldsymbol{\beta}) = \mathcal{N}(\mathbf{m}, \mathbf{S}),$$

where $\mathbf{m}$ is the variational mean and $\mathbf{S}$ is the variational covariance matrix. The variance parameters, σ_j^2 in the AGP model and (τ_j^2, τ_0^2) in the Global–Local Gaussian–IGH model, are treated as deterministic quantities and updated iteratively using an empirical Bayes procedure, rather than being assigned their own variational distributions $q(\sigma_j^2)$ or $q(\tau_j^2, \tau_0^2)$.

2.3.1. Optimization Framework

We describe this explicitly as a variational expectation–maximization (variational EM)-type scheme rather than as a pure coordinate-ascent variational inference (CAVI) procedure over a single, fully specified mean-field family. In a standard CAVI derivation, every latent quantity (here, $\boldsymbol{\beta}$ and the variance components) would be assigned an explicit variational factor, and each factor's update would be obtained by maximizing the same ELBO with respect to that factor holding the others fixed, which for the Inverse-Gamma variance components would yield full Inverse-Gamma variational posteriors rather than point estimates. Our algorithm instead performs a Gaussian variational update for $\boldsymbol{\beta}$ (an "E-like" step maximizing the ELBO given the current variances) alternated with a point (M-like) update of the variance components that mixes (i) the mode of the CAVI-optimal Inverse-Gamma factor for τ_j^2 derived below, with (ii) an empirical-Bayes plug-in estimate borrowed from the AGP update. We derive the CAVI-consistent component explicitly in Equation (17) below so that the purely heuristic part of the update (the blending itself) is clearly separated from the part that does follow from maximizing the stated ELBO.

Under this approximation,

$$\mathbb{E}_q\left[\exp(\mathbf{x}_i^\top \boldsymbol{\beta})\right] = \exp\left(\mathbf{x}_i^\top \mathbf{m} + \frac{1}{2}\mathbf{x}_i^\top \mathbf{S}\mathbf{x}_i\right).$$

Thus, we approximate the expected log-partial likelihood by

$$\widetilde{\mathbb{E}}_q[\log L(\boldsymbol{\beta})] := \sum_{i:\delta_i=1}\left[\mathbf{x}_i^\top \mathbf{m} - \log \sum_{j\in R_i} \exp\left(\mathbf{x}_j^\top \mathbf{m} + \frac{1}{2}\mathbf{x}_j^\top \mathbf{S}\mathbf{x}_j\right)\right]. \tag{15}$$

The true quantity required is $\mathbb{E}_q[\log L(\boldsymbol{\beta})]$, whose risk-set term is $\mathbb{E}_q\Big[\log \sum_{j\in R_i} \exp(\mathbf{x}_j^\top \boldsymbol{\beta})\Big]$. What Equation (15) computes instead is $\log \sum_{j\in R_i} \mathbb{E}_q[\exp(\mathbf{x}_j^\top \boldsymbol{\beta})]$, i.e., the log and the expectation/sum have been interchanged. While $\mathbb{E}_q[\exp(\mathbf{x}_j^\top \boldsymbol{\beta})] = \exp(\mathbf{x}_j^\top \mathbf{m} + \frac{1}{2}\mathbf{x}_j^\top \mathbf{S}\mathbf{x}_j)$ is exact (it is the moment generating function of the Gaussian $q(\boldsymbol{\beta})$ evaluated at $\mathbf{x}_j$), the outer step is not: because $\log(\cdot)$ is concave, Jensen's inequality [37] gives

$$\mathbb{E}_q\left[\log \sum_{j\in R_i} \exp(\mathbf{x}_j^\top \boldsymbol{\beta})\right] \;\le\; \log \sum_{j\in R_i} \mathbb{E}_q\Big[\exp(\mathbf{x}_j^\top \boldsymbol{\beta})\Big],$$

so the risk-set term used in Equation (15) is an *upper bound* on the true expected log-sum-exp term. Because this term enters the log-partial likelihood with a minus sign, using the upper bound in its place means $\widetilde{\mathbb{E}}_q[\log L(\boldsymbol{\beta})] \le \mathbb{E}_q[\log L(\boldsymbol{\beta})]$: the substitution yields a further *lower bound* on the true expected log-partial likelihood, not the expectation itself. Consequently, the objective we maximize below, which we continue to call $\mathrm{ELBO}_{\mathrm{AGP}}$ and $\mathrm{ELBO}_{\mathrm{IGH}}$ for continuity with the variational-inference literature, is technically a bound on a bound: a valid (but looser than the standard mean-field bound) lower bound on the log partial-

likelihood-based marginal likelihood, analogous in spirit to the Gaussian/delta-method bounds used for expectations of log-sum-exp terms in Bayesian multinomial and Cox-type models [33]. We flag this explicitly because it is one source of the occasional non-monotone behavior of the reported ELBO trajectory (Figures A2–A7): since the blended variance update (Section 2.3) is not itself a coordinate-ascent step on this same bound, the overall iteration is not guaranteed to be monotonically non-decreasing, and small early-iteration dips are expected rather than indicative of an error in the bound itself (see the discussion in Appendix A.2).

The variational parameters are estimated by maximizing the Jensen-bounded [38,39] variational objective derived above, which is referred to as the ELBO throughout the remainder of this study for notational simplicity. On the precise relationship to KL minimization, the standard mean-field ELBO, the identity $\log p(\mathcal{D}) = \mathrm{ELBO}(q) + \mathrm{KL}[q(\boldsymbol{\beta}) \,\|\, p(\boldsymbol{\beta} \mid \mathcal{D})]$ holds exactly, because $\mathrm{KL} \geq 0$ is the *only* gap between the ELBO and the log marginal [40,41]; consequently, maximizing that ELBO over q is exactly equivalent to minimizing $\mathrm{KL}[q \,\|\, p]$. Our objective is not that ELBO: as derived above, it substitutes the further Jensen-bounded quantity $\widetilde{\mathbb{E}}_q[\log L(\boldsymbol{\beta})]$ for the true $\mathbb{E}_q[\log L(\boldsymbol{\beta})]$, so a second, non-negative gap $\Delta_{\mathrm{Jensen}}(q) \geq 0$ (the shortfall introduced by Equation (15)) sits between our objective and the log marginal, in addition to the KL term:

$$\log p(\mathcal{D}) \;=\; \underbrace{\widetilde{\mathbb{E}}_q[\log L(\boldsymbol{\beta})] + \mathbb{E}_q[\log p(\boldsymbol{\beta})] - \mathbb{E}_q[\log q(\boldsymbol{\beta})]}_{\text{our objective}} \;+\; \mathrm{KL}[q \,\|\, p] \;+\; \Delta_{\mathrm{Jensen}}(q).$$

Because $\Delta_{\mathrm{Jensen}}(q)$ is itself a function of q (through $\mathbf{m}$ and $\mathbf{S}$) and we do not have a closed form for it, maximizing our objective is only *approximately* equivalent to minimizing $\mathrm{KL}[q(\boldsymbol{\beta}) \,\|\, p(\boldsymbol{\beta} \mid \mathcal{D})]$ – exactly, it minimizes $\mathrm{KL}[q \,\|\, p] + \Delta_{\mathrm{Jensen}}(q)$ jointly. We do not attempt to bound $\Delta_{\mathrm{Jensen}}(q)$ analytically here and instead rely on the empirical convergence diagnostics reported in Appendix A.2; we make this distinction explicit so that the KL-minimization interpretation of the algorithm is not overstated relative to the standard variational-inference identity it is adapted from.

For the Global–Local Gaussian–IGH formulation, the ELBO is given by

$$\mathrm{ELBO}_{\mathrm{IGH}} = \mathbb{E}_q[\log L(\boldsymbol{\beta})] + \mathbb{E}_q[\log p(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2, \tau_0^2)] + \log p(\boldsymbol{\tau}^2) + \log p(\tau_0^2) + \mathcal{H}[q(\boldsymbol{\beta})],$$

where the entropy of the Gaussian variational distribution is

$$\mathcal{H}[q(\boldsymbol{\beta})] = \frac{1}{2} \log \det(2\pi e \mathbf{S}).$$

Explicitly, the Gaussian prior contribution is

$$\mathbb{E}_q[\log p(\boldsymbol{\beta} \mid \boldsymbol{\tau}^2, \tau_0^2)] = -\frac{1}{2} \sum_{j=1}^{p} \left[\log(2\pi\sigma_j^2) + \frac{m_j^2 + S_{jj}}{\sigma_j^2} \right],$$

where $\sigma_j^2 = \sqrt{\tau_0^2}\, \tau_j^2$. The local Inverse-Gamma prior contribution is

$$\log p(\boldsymbol{\tau}^2) = \sum_{j=1}^{p} \left[a \log b - \log \Gamma(a) - (a+1) \log \tau_j^2 - \frac{b}{\tau_j^2} \right],$$

and the global Inverse-Gamma prior contribution is

$$\log p(\tau_0^2) = a_0 \log b_0 - \log \Gamma(a_0) - (a_0 + 1) \log \tau_0^2 - \frac{b_0}{\tau_0^2}.$$

For the AGP formulation, the ELBO is

$$\mathrm{ELBO}_{\mathrm{AGP}} = \mathbb{E}_q[\log L(\boldsymbol{\beta})] + \mathbb{E}_q[\log p(\boldsymbol{\beta} \mid \sigma^2)] + \mathcal{H}[q(\boldsymbol{\beta})].$$

2.3.2. Parameter Updates

Define

$$w_j = \mathbb{E}_q[\exp(\mathbf{x}_j^\top \boldsymbol{\beta})] = \exp\left(\mathbf{x}_j^\top \mathbf{m} + \frac{1}{2}\mathbf{x}_j^\top \mathbf{S}\mathbf{x}_j\right).$$

For the Global–Local Gaussian–IGH model, the gradient of $\mathrm{ELBO}_{\mathrm{IGH}}$ with respect to $\mathbf{m}$ is

$$\frac{\partial \mathrm{ELBO}_{\mathrm{IGH}}}{\partial \mathbf{m}} = \sum_{i:\delta_i=1} \left(\mathbf{x}_i - \frac{\sum_{j\in R_i} \mathbf{x}_j w_j}{\sum_{j\in R_i} w_j}\right) - \mathrm{diag}(\boldsymbol{\sigma}^{-2})\mathbf{m},$$

where $\sigma_j^2 = \sqrt{\tau_0^2}\,\tau_j^2$.

For the AGP model, the gradient of $\mathrm{ELBO}_{\mathrm{AGP}}$ with respect to $\mathbf{m}$ is

$$\frac{\partial \mathrm{ELBO}_{\mathrm{AGP}}}{\partial \mathbf{m}} = \sum_{i:\delta_i=1} \left(\mathbf{x}_i - \frac{\sum_{j\in R_i} \mathbf{x}_j w_j}{\sum_{j\in R_i} w_j}\right) - \mathrm{diag}(\boldsymbol{\sigma}^{-2})\mathbf{m}.$$

The covariance update is obtained through the precision matrix. For the Global–Local Gaussian–IGH model,

$$\mathbf{S}^{-1} = \sum_{i:\delta_i=1} \left(\frac{\sum_{j\in R_i} w_j \mathbf{x}_j \mathbf{x}_j^\top}{\sum_{j\in R_i} w_j} - \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^\top\right) + \mathrm{diag}(\boldsymbol{\sigma}^{-2}) + \lambda_{\mathrm{ridge}}\mathbf{I},$$

and for the AGP model,

$$\mathbf{S}^{-1} = \sum_{i:\delta_i=1} \left(\frac{\sum_{j\in R_i} w_j \mathbf{x}_j \mathbf{x}_j^\top}{\sum_{j\in R_i} w_j} - \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^\top\right) + \mathrm{diag}(\boldsymbol{\sigma}^{-2}) + \lambda_{\mathrm{ridge}}\mathbf{I},$$

where

$$\bar{\mathbf{x}}_i = \frac{\sum_{j\in R_i} w_j \mathbf{x}_j}{\sum_{j\in R_i} w_j},$$

and λ_{ridge} is a small ridge penalty to ensure numerical stability. In practice, the precision matrix is inverted using a block-diagonal approximation with correlation shrinkage to improve scalability in high dimensions.

For the AGP model, the adaptive variance parameters are updated as

$$\sigma_j^2 = m_j^2 + S_{jj}, \quad j = 1, \ldots, p, \tag{16}$$

where S_{jj} denotes the jth diagonal element of $\mathbf{S}$.

2.3.3. CAVI-Derived Component

Fixing $\boldsymbol{\beta} \sim q(\boldsymbol{\beta}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$ and τ_0^2, and holding the likelihood contribution to τ_j^2 fixed through $\mathbb{E}_q[\beta_j^2] = m_j^2 + S_{jj}$, the terms in the posterior kernel (14) involving τ_j^2 alone give the conditional kernel

$$q^\star(\tau_j^2) \;\propto\; (\tau_j^2)^{-(a+\frac{3}{2})} \exp\left(-\frac{m_j^2 + S_{jj}}{2\sqrt{\tau_0^2}\,\tau_j^2} - \frac{b}{\tau_j^2}\right)$$

$$= (\tau_j^2)^{-\left(\left(a+\frac{1}{2}\right)+1\right)} \exp\left(-\frac{b+\frac{m_j^2+S_{jj}}{2\sqrt{\tau_0^2}}}{\tau_j^2}\right),$$

which is the kernel of an Inv-Gamma$\left(a+\frac{1}{2},\ b+\frac{m_j^2+S_{jj}}{2\sqrt{\tau_0^2}}\right)$ distribution.

Rather than using the mode of this distribution, the implementation uses the reciprocal of its expected precision as a point summary. For $X \sim \text{Inv-Gamma}(\alpha, \beta)$, $\mathbb{E}(1/X) = \alpha/\beta$. Therefore,

$$\tau_j^{2,\text{CAVI}} = \left[\mathbb{E}_{q^\star}\left(\frac{1}{\tau_j^2}\right)\right]^{-1} = \frac{2b + \dfrac{m_j^2 + S_{jj}}{\sqrt{\tau_0^2}}}{2a+1}. \tag{17}$$

This is exactly the second ("IGH") term used below. The quantity in Equation (17) is therefore derived from the CAVI-optimal Inverse-Gamma factor, but is carried forward as a point summary rather than as the full variational distribution $q^\star(\tau_j^2)$. Consequently, uncertainty in τ_j^2 is not propagated across iterations through a separate variational factor.

For the Global–Local Gaussian–IGH model, we combine this CAVI-derived point estimate with the AGP-style empirical-Bayes estimate $m_j^2 + S_{jj}$ via a blended update. Let $\lambda \in [0,1]$ denote the blend weight. The local variance parameters are updated as

$$\tau_j^{2,\text{new}} = \lambda\,(m_j^2 + S_{jj}) + (1-\lambda) \underbrace{\frac{2b + \dfrac{m_j^2+S_{jj}}{\sqrt{\tau_0^2}}}{2a+1}}_{=\tau_j^{2,\text{CAVI}}\ \text{(Equation (17))}}, \quad j = 1, \ldots, p. \tag{18}$$

Unlike the second term, the first term ($m_j^2 + S_{jj}$, the AGP component) and the convex combination itself (governed by λ) do *not* arise from maximizing ELBO_{IGH} with respect to a single, coherent variational family; they are a heuristic regularization device intended to prevent the purely hierarchical update from over-shrinking variance estimates toward the prior mode in small samples. We report this explicitly as a design choice rather than as an optimization result, and we examine its sensitivity to λ in Section 3.

The global variance parameter is updated as

$$\tau_0^{2,\text{new}} = \frac{2b_0 + \sum_{j=1}^{p} \dfrac{m_j^2 + S_{jj}}{\max(\tau_j^2, \epsilon_{\min})}}{2a_0 + p}, \tag{19}$$

where $\epsilon_{\min}$ is a small positive constant to prevent division by zero. Both τ_j^2 and τ_0^2 are constrained to remain within pre-specified bounds to ensure numerical stability.

Under the modified variance specification $\sigma_j^2 = \sqrt{\tau_0^2}\,\tau_j^2$, the update in Equation (19) is used as an empirical-Bayes point update for τ_0^2 rather than as an exact CAVI update.

2.3.4. Algorithm Description

To obtain posterior estimates for the proposed models (Equations (9) and (14)), we develop scalable variational inference (VI) algorithms tailored to the Adaptive Gaussian Prior (AGP-VI) and the Global–Local Gaussian–Inverse-Gamma (Global–Local Gaussian–IGH-VI) formulations. Algorithms 1 and 2 are based on a multivariate Gaussian variational approximation with a block-diagonal covariance structure: predictors are partitioned into contiguous blocks (Section 2.3), within-block dependence is retained (with a shrinkage factor applied to off-diagonal within-block entries), and cross-block covariances are set to

zero for computational scalability at high p. The posterior distribution of the regression coefficients is approximated by a multivariate normal distribution with mean vector $\mathbf{m}$ and block-diagonal covariance matrix $\mathbf{S}$. The procedure begins by initializing the variational parameters, typically with $\mathbf{m}^{(0)} = \mathbf{0}$ and $\mathbf{S}^{(0)} = \epsilon\mathbf{I}$, where ϵ is a small positive constant to ensure numerical stability. For the Global–Local Gaussian–IGH model, the local variance parameters τ_j^2 are initialized to $\mathbf{1}$, the global variance τ_0^2 is initialized to 1, and the blend weight λ is specified (typically $\lambda = 0.90$).

Algorithm 1: AGP-VI for Bayesian Cox Model

Data: Dataset $\mathcal{D}$; initial $\mathbf{m}, \mathbf{S}, \sigma^2$
Result: Posterior estimates $\mathbf{m}, \mathbf{S}, \sigma^2$
while *not converged and iteration* $<$ `max_iter` **do**
  Compute variational Cox weights and the AGP ELBO gradient $\mathbf{g}_{\text{AGP}}$;
  Clip $\mathbf{g}_{\text{AGP}}$ elementwise and obtain the block-diagonal covariance proposal $\mathbf{S}_{\text{prop}}$;
  Use backtracking to update

$$\mathbf{m}_{\text{cand}} = \text{clip}(\mathbf{m} + \eta\mathbf{g}_{\text{AGP}})$$

  and set $\mathbf{S}_{\text{cand}} = \mathbf{S}_{\text{prop}}$;
  Update

$$\sigma^2_{j,\text{cand}} = m^2_{j,\text{cand}} + S_{jj,\text{cand}}$$

  Accept the candidate if the ELBO acceptance condition is satisfied; otherwise retain the previous state;
  Check relative ELBO convergence over consecutive accepted iterations;
end
return $\mathbf{m}, \mathbf{S}, \sigma^2$;

Algorithm 2: Global–Local Gaussian–IGH-VI for Bayesian Cox Model

Data: Dataset $\mathcal{D}$; initial $\mathbf{m}, \mathbf{S}, \boldsymbol{\tau}^2, \tau_0^2$; blend weight λ
Result: Posterior estimates $\mathbf{m}, \mathbf{S}, \boldsymbol{\tau}^2, \tau_0^2$
while *not converged* **do**
  Compute effective variances, variational Cox weights, and gradient $\mathbf{g}_{\text{IGH}}$;
  Obtain the block-diagonal covariance proposal $\mathbf{S}_{\text{prop}}$ and form

$$\mathbf{d} = \mathbf{S}_{\text{prop}}\mathbf{g}_{\text{IGH}}$$

  with a global norm safeguard;
  Use full-state backtracking to update $\mathbf{m}$ and $\mathbf{S}$;
  Compute the raw local and global variance updates using Equations (18) and (19), and apply gradual (damped) updates to $\boldsymbol{\tau}^2$ and τ_0^2 to improve numerical stability;
  Accept the full-state candidate if the ELBO criterion is satisfied; otherwise attempt safeguarded rescue updates, followed by a no-op if needed;
  Check relative ELBO convergence over consecutive accepted iterations;
end
return $\mathbf{m}, \mathbf{S}, \boldsymbol{\tau}^2, \tau_0^2$;

At each iteration, the algorithm computes the expected exponential term of the linear predictor under the variational distribution. This expectation plays a central role in approximating the Cox partial likelihood and is given by

$$w_i = \exp\left(\mathbf{x}_i^\top \mathbf{m} + \frac{1}{2}\mathbf{x}_i^\top \mathbf{S}\mathbf{x}_i\right).$$

Using these weights, the two algorithms update the variational mean m through different mechanisms. For AGP-VI, **m** is updated using an elementwise-clipped gradient-ascent step on the ELBO, with the block-diagonal covariance proposal adopted directly at each accepted step without interpolation with the previous covariance. In contrast, Global–Local Gaussian–IGH-VI updates m using a block-preconditioned approximate-Newton direction with a global norm cap and full-state backtracking, which jointly updates the mean, covariance, and variance components. Further implementation details are provided in the hyperparameter and tuning-constant specification in Section 2.3.5. In both cases the covariance matrix **S** is derived from the inverse of an approximated precision matrix that incorporates the curvature of the likelihood, the shrinkage imposed by the prior, a ridge penalty for numerical stability, and a block-diagonal approximation with correlation shrinkage for scalability.

The variance parameters are updated in a model-specific manner. For the AGP formulation, the coefficient-specific variances are updated as

$$\sigma_j^2 = m_j^2 + S_{jj},$$

reflecting an empirical Bayes estimate that adapts shrinkage at the individual coefficient level. For the Global–Local Gaussian–IGH formulation, the local variance parameters are updated using a blended rule that combines the adaptive AGP estimate with the hierarchical IGH update:

$$\tau_j^2 = \lambda\,(m_j^2 + S_{jj}) + (1-\lambda)\,\frac{2b + \frac{m_j^2 + S_{jj}}{\sqrt{\tau_0^2}}}{2a+1}.$$

The blend weight λ controls the trade-off between data-driven shrinkage (AGP component) and prior regularization (IGH component). The global variance τ_0^2 is updated as

$$\tau_0^2 = \frac{2b_0 + \sum_{j=1}^{p} \frac{m_j^2 + S_{jj}}{\max(\tau_j^2, \epsilon_{\min})}}{2a_0 + p},$$

which pools information across all coefficients to determine the overall shrinkage level. Both τ_j^2 and τ_0^2 are constrained to remain within pre-specified positive bounds to ensure valid precision matrices and numerical stability.

After updating all parameters, the ELBO is evaluated to monitor convergence. The iterative procedure continues until convergence criteria are satisfied, such as a negligible relative change in the ELBO between successive iterations. Additional stopping criteria based on the norm of parameter updates may also be used to improve numerical stability. The final output of the algorithm includes the posterior mean **m**, covariance matrix **S**, and the estimated variance parameters (σ_j^2 or τ_j^2, τ_0^2). These quantities provide point estimates and approximate uncertainty measures for the regression coefficients, conditional on the plug-in shrinkage variance estimates (σ_j^2 or τ_j^2, τ_0^2); because uncertainty in those shrinkage variances is not itself propagated (Section 2.3), this falls short of fully reliable

inference in the sense of a completely Bayesian uncertainty quantification, and should be interpreted accordingly.

2.3.5. Hyperparameter and Tuning-Constant Specification

The precision matrix is inverted using a block-diagonal approximation. Predictors are partitioned into contiguous blocks of size 10. Within each block, the precision sub-matrix is eigenvalue-floored and inverted. The off-diagonal covariance entries are then shrunk by a factor of 0.3, while the diagonal entries are left unchanged. The resulting block covariance is projected to positive-definiteness and incorporated into the block-diagonal covariance matrix $\mathbf{S}$, with cross-block covariances set to zero.

The mean updates differ between the two variational methods. For AGP-VI, the gradient is bounded elementwise as

$$g_j \leftarrow \max\{-5, \min(5, g_j)\}, \qquad j = 1, \ldots, p,$$

and the variational mean is updated as

$$m_j \leftarrow \max\{-10, \min(10, m_j + \eta g_j)\}, \qquad j = 1, \ldots, p.$$

The initial step size is $\eta = 0.001$, with up to 15 backtracking halvings. At each candidate step, the block-diagonal covariance proposal $\mathbf{S}_{\text{prop}}$ is used directly, and the coefficient-specific variances are updated as $\sigma_j^2 = m_j^2 + S_{jj}$. If no candidate satisfies the ELBO acceptance criterion, the previous variational state is retained.

For Global–Local Gaussian–IGH-VI, the mean update uses the block-preconditioned direction

$$\mathbf{d} = \mathbf{S}_{\text{prop}}\mathbf{g},$$

where $\mathbf{g}$ is the gradient and $\mathbf{S}_{\text{prop}}$ is obtained from the block-diagonal approximation to the inverse penalized precision matrix. A global norm safeguard is applied to $\mathbf{d}$. The initial step size is $\eta = 1.0$, with up to 20 backtracking halvings. Full-state backtracking is used so that the mean, covariance, and variance components are updated together. Candidate states are accepted when they satisfy the ELBO acceptance criterion described in Appendix A.2. If a full-state update is not accepted, a mean-only preconditioned update is attempted, followed, if necessary, by a norm-capped raw-gradient update. If neither update is accepted, the previous state is retained.

For both methods, the relative ELBO convergence tolerance is $\text{tol} = 10^{-4}$ and convergence requires the criterion to hold for patience $= 5$ consecutive accepted iterations. For Global–Local Gaussian–IGH-VI, rejected updates are additionally monitored through a separate stall criterion, allowing convergence, stalling, and maximum-iteration termination to be distinguished. For AGP-VI, rejected updates are recorded through the fallback count.

The variational covariance is initialized at $\mathbf{S}^{(0)} = 0.5\,\mathbf{I}$. For AGP-VI, the ridge parameter is $\lambda_{\text{ridge}} = 0.05$, and the coefficient-specific variances are restricted to $[10^{-4}, 5]$. For Global–Local Gaussian–IGH-VI, the ridge parameter is $\lambda_{\text{ridge}} = 0.04$, the local variances τ_j^2 are restricted to $[10^{-4}, 5]$, and the global variance τ_0^2 is restricted to $[0.05, 10]$. A variance-damping factor of 0.05 is used for the local and global variance updates.

Both algorithms are run for a maximum of 300 iterations in the simulation study (Section 3). In the real-data analysis (Section 5), the maximum is 1000 iterations, with earlier termination when the convergence criterion is satisfied.

For Global–Local Gaussian–IGH-VI, the real-data analysis uses an adaptive initialization procedure. Among the candidate solutions satisfying the convergence criterion, the fit with the highest final ELBO is retained. The initialization procedure uses only the training data and does not use test-set predictive performance.

The Inverse-Gamma hyperparameters for Global–Local Gaussian–IGH-VI are fixed at $a = 1$ and $b = 0.5$ for the local component and $a_0 = 1$ and $b_0 = 2$ for the global component. The blend weight is fixed at $\lambda = 0.90$. These values are specified a priori and are not selected by cross-validation or other data-driven tuning procedures.

2.3.6. Convergence and Stability

The proposed variational inference (VI) algorithms, namely the Adaptive Gaussian Prior (AGP-VI) and Global–Local Gaussian–Inverse-Gamma (Global–Local Gaussian–IGH-VI) procedures (Algorithms 1 and 2), are optimized by maximizing the Evidence Lower Bound (ELBO), which provides a tractable objective for variational posterior approximation. Each iteration updates the variational parameters through gradient-based optimization of the ELBO, together with the corresponding variance-parameter updates.

Because the computational scheme uses an approximation to the expected Cox partial log-likelihood, a block-diagonal covariance approximation, bounded variance parameters, and, for Global–Local Gaussian–IGH-VI, a blended Global–Local variance update, convergence is assessed numerically rather than interpreted as convergence to a stationary point of a single exact ELBO. In particular, convergence is evaluated from the stabilization of the variational objective over successive accepted iterations (Appendix A.2).

Convergence is assessed using the relative change in the ELBO between successive iterations. Specifically,

$$\frac{|\mathrm{ELBO}^{(t)} - \mathrm{ELBO}^{(t-1)}|}{|\mathrm{ELBO}^{(t-1)}| + \epsilon_{\mathrm{tol}}} < \epsilon,$$

where $\epsilon > 0$ is a predefined tolerance and ϵ_{tol} is a small positive constant that prevents division by zero.

In the reported analyses, $\epsilon = 10^{-4}$, and this criterion is required to hold for five consecutive accepted iterations. The maximum number of iterations is 300 in the simulation study and 1000 in the real-data analysis.

To enhance numerical stability, the variational covariance is initialized as $\mathbf{S}^{(0)} = 0.5\mathbf{I}$, and the variance parameters (σ_j^2, τ_j^2, and τ_0^2) are constrained to positive, pre-specified ranges throughout the optimization.

The two algorithms use different safeguards for the mean update. AGP-VI uses an elementwise-bounded gradient update with backtracking. Global–Local Gaussian–IGH-VI instead uses a block-preconditioned approximate-Newton direction with a global norm safeguard, full-state backtracking, and gradual updates of the local and global variance components. When the full-state proposal is not accepted, the safeguarded rescue updates described in Section 2.3.5 are considered.

The multivariate Gaussian variational approximation uses a block-diagonal covariance structure, which retains dependencies among predictors within each block while setting cross-block covariances to zero. This provides a computationally tractable alternative to a fully dense covariance matrix in high-dimensional settings.

In high-dimensional settings, the prior precision terms, $\mathrm{diag}(\sigma^{-2})$ and $\mathrm{diag}(\boldsymbol{\sigma}^{-2})$, provide regularization of the regression coefficients, similar in spirit to shrinkage-based penalized Cox regression methods [16,42].

For Global–Local Gaussian–IGH-VI, the blended local and global variance updates are applied gradually through under-relaxation to reduce large changes in the variance components between successive iterations.

The code for reproducing the results of the proposed AGP-VI and Global–Local Gaussian–IGH-VI methods is available at https://doi.org/10.5281/zenodo.23558120 (accessed on 29 May 2026).

2.4. Model Evaluation Criteria

This study considers two key evaluation metrics for assessing predictive performance in survival analysis: the concordance index (C-index) and the Brier Score. These metrics are particularly suitable in high-dimensional Bayesian survival modeling, where heavy censoring and complex covariate structures can obscure true predictive performance.

The C-index evaluates the discriminative ability of a model, whereas the Brier Score measures calibration and overall predictive accuracy. Together, they provide a comprehensive assessment of model performance by quantifying both the ability to rank individuals by risk and the agreement between predicted and observed outcomes.

2.4.1. C-Index Criterion

The concordance index (C-index) measures the ability of a model to correctly order survival times based on predicted risk scores. Originally proposed by [43], it is widely used in survival analysis. Its classical form is given by:

$$C = \frac{\sum_{i<j} \mathbb{I}(\hat{r}_i > \hat{r}_j) \cdot \mathbb{I}(T_i < T_j,\, \delta_i = 1)}{\sum_{i<j} \mathbb{I}(T_i < T_j,\, \delta_i = 1)}, \tag{20}$$

where $\hat{r}_i = \mathbf{x}_i^\top \hat{\beta}$ denotes the predicted risk score, T_i is the observed survival time, and δ_i indicates event occurrence.

The C-index takes values in the range $[0.5, 1]$, where 0.5 corresponds to random prediction and 1 indicates perfect concordance. In the presence of censoring, modifications such as the inverse probability of censoring weighting (IPCW) estimator proposed by [44] provide more reliable estimates, particularly in heavily censored datasets.

2.4.2. Brier Score Procedure

While the C-index assesses discriminative performance, the Brier Score evaluates calibration by comparing predicted survival probabilities with observed outcomes at a given time point t. Predicted survival probabilities are obtained as $\hat{S}(t \mid \mathbf{x}_i) = \hat{S}_0(t)^{\exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})}$, where $\hat{\boldsymbol{\beta}}$ is the posterior mean $\mathbf{m}$ (for AGP-VI and Global–Local Gaussian–IGH-VI) or the corresponding point estimate for the competing methods, and $\hat{S}_0(t) = \exp(-\hat{H}_0(t))$ is obtained from the Breslow estimator of the cumulative baseline hazard, $\hat{H}_0(t) = \sum_{t_i \le t} \delta_i / \sum_{j \in R(t_i)} \exp(\mathbf{x}_j^\top \hat{\boldsymbol{\beta}})$, evaluated on the training data and applied to the test data at evaluation. This is a standard post hoc step for any Cox-type model (proportional-hazards-based or hierarchical-Bayesian) whose likelihood is a partial likelihood: fitting $\boldsymbol{\beta}$ via the partial likelihood does not by itself estimate $\hat{S}_0(t)$, so a baseline-hazard estimator is needed separately, after model fitting, to convert the fitted linear predictor into a survival probability for the Brier Score and time-dependent AUC below. Introduced by [45], it is defined as:

$$\mathrm{BS}(t) = \frac{1}{N} \sum_{i=1}^{N} \big(\hat{S}(t \mid \mathbf{x}_i) - \mathbb{I}(T_i > t)\big)^2, \tag{21}$$

where $\hat{S}(t \mid \mathbf{x}_i)$ denotes the predicted survival probability.

Because censoring complicates direct evaluation, the IPCW-adjusted Brier Score is commonly used:

$$\mathrm{BS}_{\mathrm{IPCW}}(t) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\delta_i \cdot \mathbb{I}(T_i \le t)}{\hat{G}(T_i)} \big(\hat{S}(t \mid \mathbf{x}_i)\big)^2 + \frac{\mathbb{I}(T_i > t)}{\hat{G}(t)} \big(1 - \hat{S}(t \mid \mathbf{x}_i)\big)^2 \right], \tag{22}$$

where $\hat{G}(\cdot)$ is the Kaplan–Meier estimator of the censoring distribution.

Extensions developed by [46] ensure consistency of the Brier Score under right censoring. Lower values indicate better calibration and predictive accuracy.

The C-index and Brier Score provide complementary insights into model performance. The former evaluates the ability to correctly rank individuals by risk, while the latter assesses how well predicted survival probabilities align with observed outcomes. Their combined use is therefore essential for evaluating Bayesian survival models in high-dimensional settings, where censoring and model complexity pose significant challenges to reliable predictive assessment.

2.4.3. Time-Dependent AUC

In addition to the C-index and Brier Score, we report the time-dependent Area Under the Receiver Operating Characteristic Curve (AUC). At a fixed evaluation time t, subjects are classified as cases if they experience the event by time t ($T_i \leq t, \delta_i = 1$) and as controls if they are event-free and known to survive beyond t ($T_i > t$). The time-dependent AUC is defined as

$$\mathrm{AUC}(t) = P\big(\hat{r}_i > \hat{r}_j \mid T_i \leq t < T_j\big), \tag{23}$$

and is estimated using inverse probability of censoring weighting (IPCW) to account for censoring before time t. Higher AUC values indicate better discrimination between subjects who experience the event by time t and those who remain event-free beyond t.

2.4.4. Variable Selection Rule, True Positive Rate, and False Positive Rate

Because the AGP and Global–Local Gaussian–IGH priors are continuous shrinkage priors, their variational coefficient estimates are generally not exactly zero. A selection rule is therefore required to classify individual coefficients as selected or not selected.

For AGP-VI and Global–Local Gaussian–IGH-VI, coefficient j is declared selected when its approximate 95% credible interval excludes zero:

$$\mathrm{Selected}_j = \mathbb{1}\Big[0 \notin \Big(\hat{\beta}_j - 1.96\sqrt{S_{jj}},\ \hat{\beta}_j + 1.96\sqrt{S_{jj}}\Big)\Big]. \tag{24}$$

The same posterior-interval principle is used for Bayesian-Lasso-Cox when posterior draws are available. For Cox-Lasso and Cox-ElasticNet, selection is determined by non-zero estimated coefficients at the cross-validated regularization parameter. No additional coefficient-magnitude threshold is imposed on these penalized Cox models.

Given the selected variables and the known true support $\mathcal{S} = \{j : \beta_j^{\mathrm{true}} \neq 0\}$ in the simulation study, the True Positive Rate (TPR) and False Positive Rate (FPR) are defined as

$$\mathrm{TPR} = \frac{|\{j \in \mathcal{S} : j \text{ is selected}\}|}{|\mathcal{S}|}, \tag{25}$$

$$\mathrm{FPR} = \frac{|\{j \notin \mathcal{S} : j \text{ is selected}\}|}{|\{1, \ldots, p\} \setminus \mathcal{S}|}. \tag{26}$$

TPR measures the proportion of truly non-zero coefficients correctly identified, whereas FPR measures the proportion of truly zero coefficients incorrectly selected. An ideal variable-selection procedure has TPR $= 1$ and FPR $= 0$. The mean and standard deviation of these quantities across simulation replications are reported in Section 4.

A small non-zero effect may nevertheless have a credible interval that includes zero, particularly when correlated genomic predictors share the predictive signal. Consequently, the number of individually selected variables should be interpreted together with the predictive performance measures reported in Section 5.

2.4.5. Random Survival Forests for Feature Screening

For the ultra-high-dimensional real datasets (Section 5), Random Survival Forests (RSF) [47–49] are used as a pre-processing feature-screening procedure. RSF constructs

an ensemble of survival trees from bootstrap samples and uses survival-specific splitting rules to accommodate censored outcomes. Variable-importance measures obtained from the fitted forest are used to rank predictors before the competing survival models are fitted.

The RSF screening is performed using 500 trees. Predictors with permutation importance above the 95th percentile of the importance-score distribution are retained, subject to a minimum of 20 and a maximum of 1000 predictors. Screening is applied only to datasets containing more than 100 predictors (DBCD, CHOP, VDX, and TCGA Ovarian). PBC and NKI70, for which $p \leq 100$, are analyzed without RSF screening.

3. Simulation Procedure

In this study, we develop a comprehensive simulation framework to evaluate the performance of the proposed variational inference-based Bayesian Cox models, namely the Adaptive Gaussian Prior Variational Inference (AGP-VI) and the Gaussian–Inverse-Gamma Hierarchical Variational Inference (Global–Local Gaussian–IGH-VI). The simulation setting is designed to reflect realistic challenges commonly encountered in high-dimensional survival analysis, including sparse signals, correlated predictors, fixed sample size, and a particular level of censoring.

To provide a meaningful benchmark, the proposed methods are compared with three widely used competing approaches: Cox-Lasso [42], Cox-ElasticNet [22], and Bayesian-Lasso-Cox [50]. These methods represent commonly adopted penalized and Bayesian regularization techniques for high-dimensional survival analysis and provide useful benchmarks for assessing predictive performance, variable selection, and computational efficiency. The comparison therefore highlights the trade-off between computational efficiency and statistical performance, contrasting scalable variational inference methods with traditional MCMC-based approaches.

In high-dimensional settings, strong regularization and scalable inference techniques are essential for achieving reliable estimation and stable model performance, especially when predictors are noisy or highly correlated [9,16]. To mimic such conditions, datasets were generated with varying dimensionality, where the number of predictors was set to $p \in \{50, 100, 200, 300, 400, 500\}$, while the sample size $n = 100$ was kept fixed. This reflects common scenarios in modern biomedical and genomic studies, where the number of predictors can greatly exceed the number of observations [16].

3.1. Data Generation Mechanism

To ensure a realistic and unbiased evaluation, the true regression coefficients were generated under a sparse structure, where only a small subset of predictors had non-zero effects on the hazard function, while the remaining coefficients were set to zero.

Specifically, $|\mathcal{S}| = n_{\text{signal}} = 10$ active predictors were assigned non-zero coefficients, held fixed across all $p \in \{50, 100, 200, 300, 400, 500\}$, with the remaining $p - 10$ coefficients set to zero. Thus, the proportion of active predictors decreased as p increased. The signal strength was fixed at $|\beta_j^{\text{true}}| = 1.20$ for all active predictors, with alternating signs $(+1.20, -1.20, +1.20, \ldots)$ across the 10 active predictors, corresponding to hazard ratios of $\exp(\pm 1.20) \approx \{0.30, 3.32\}$.

This setup allows for a clear assessment of each model's ability to identify relevant variables while shrinking irrelevant ones.

The covariates were generated from a multivariate normal distribution $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with an AR(1)-type covariance structure, $\Sigma_{jk} = \rho^{|j-k|}$ with $\rho = 0.15$, 0.50, and 0.90 in the three simulation settings (weak, moderate, and high correlation, respectively), to induce correlation among predictors that decreases with the distance between predictor indices. The covariates were standardized to have mean zero and unit variance after simulation.

This simulation design allows the assessment of model performance under multicollinearity, a common characteristic of high-dimensional datasets that can adversely affect estimation stability and variable selection accuracy [9,16].

Survival times were generated according to the Cox proportional hazards model in Equation (1), using a constant exponential baseline hazard rate of $h_0 = 0.05$. Specifically,

$$T_i \sim \text{Exponential}\Big(\text{rate} = h_0 \exp(\mathbf{x}_i^\top \boldsymbol{\beta}^{\text{true}})\Big),$$

where the linear predictor was restricted to $[-10, 10]$ for numerical stability.

This ensures that the simulated data satisfy the proportional hazards assumption [11].

Independent censoring times $C_i \sim \text{Exponential}(\text{rate} = c)$ were generated.

The rate c was calibrated numerically using root-finding on a pilot sample to obtain an expected censoring proportion of 50%. The observed survival time and event indicator were defined as

$$(t_i, d_i) = (\min(T_i, C_i), \mathbb{1}\{T_i \leq C_i\}),$$

and the realized censoring proportion was recorded for each simulated dataset.

3.1.1. Summary of Simulation Settings

For ease of reference, the design quantities (see Table 1) used throughout the simulation study are summarized below. The sample size, number of active predictors, signal strength, and censoring proportion were held fixed in every simulated dataset; only the number of predictors p (across rows within each table) and the correlation parameter ρ (across the three tables) were varied.

Table 1. Summary of simulation design settings.

Design Quantity	Value Used	Fixed or Varied
Sample size n	100	Fixed
Number of active predictors	10 (all other $p - 10$ coefficients are zero)	Fixed for every p
Signal strength	$\lvert\beta_j^{\text{true}}\rvert = 1.20$, alternating signs	Fixed
Censoring proportion	50% (expected; censoring rate calibrated numerically)	Fixed
Number of predictors p	50, 100, 200, 300, 400, 500	Varied (rows of each table)
Predictor correlation	AR(1), $\Sigma_{jk} = \rho^{\lvert j-k\rvert}$, $\rho \in \{0.15, 0.50, 0.90\}$	Varied

The first four quantities were held fixed; therefore, the present study does not assess sensitivity to sample size, censoring level, sparsity level, or signal strength.

3.1.2. Validation Protocol

For each of $n_{\text{rep}} = 20$ replications at each value of p, a single random 70%/30% training/test split was generated. All five competing methods were fitted using the training data, and predictive performance was evaluated using the held-out 30% test data. Centering and scaling parameters applied to the test data were estimated from the corresponding training data only.

This procedure represents repeated holdout rather than k-fold cross-validation. The Mean and SD values reported are the mean and standard deviation of the corresponding measures across the twenty independent replications at each value of p.

3.2. Evaluation Metrics

Model performance was assessed using multiple criteria. Predictive accuracy was evaluated using the concordance index (C-index) and the Brier Score, which measure discrimination and calibration, respectively [43,45].

Variable selection performance was assessed using the True Positive Rate (TPR) and False Positive Rate (FPR), computed per the selection rule and definitions given in Section 2.4.4, together with their standard deviation across simulation replications. This provides insight into each method's ability to correctly identify relevant predictors while controlling false discoveries.

In general, the simulation framework provides a realistic setting for evaluating the proposed AGP-VI and Global–Local Gaussian–IGH-VI methods. The results provide insight into the trade-off between estimation accuracy, prediction error as measured by the Brier Score, and variable selection performance, and illustrate the settings in which the proposed approaches offer advantages relative to competing penalized and Bayesian methods in high-dimensional survival analysis, as well as settings in which they do not.

4. Simulation Results

The simulation results presented in this section provide a comprehensive assessment of the proposed Bayesian variational inference framework across a range of high-dimensional survival settings. By varying the number of predictors while maintaining a sparse underlying signal structure, the study evaluates the competing methods in terms of predictive accuracy and variable selection performance. Specifically, performance is assessed using the concordance index (C-index), Brier Score, True Positive Rate (TPR), and False Positive Rate (FPR).

The results provide insight into how the methods behave as model dimensionality increases and the variable selection task becomes more challenging. Overall, the findings highlight the limitations of conventional penalized and Bayesian shrinkage approaches in ultra-high-dimensional settings while demonstrating the effectiveness of the proposed AGP-VI and Global–Local Gaussian–IGH-VI methods. In particular, the proposed variational inference procedures exhibit competitive predictive performance, improved control of false positive selections, and greater robustness as the number of predictors increases. These characteristics make the proposed framework a scalable and reliable alternative for high-dimensional survival analysis.

Table 2 summarizes the performance of the five competing methods across increasing predictor dimensions ($p = 50, \ldots, 500$). Performance is evaluated using the concordance index (C-index) and Brier Score for predictive accuracy, together with the mean and standard deviation of the True Positive Rate (TPR) and False Positive Rate (FPR) to assess variable selection performance.

For the low-dimensional setting ($p = 50$), Cox-ElasticNet achieved the highest C-index (0.8824), indicating the strongest discriminative ability, whereas AGP-VI attained the lowest Brier Score (0.0908), reflecting the best prediction accuracy. Bayesian-Lasso-Cox achieved perfect sensitivity with a TPR of 1.0000 and zero variability across replications, successfully identifying all active predictors. However, this gain in sensitivity was accompanied by a substantially larger FPR (0.7050), indicating a tendency to select many irrelevant variables.

As the dimensionality increased to $p = 100$ and $p = 200$, the proposed variational inference methods remained highly competitive. AGP-VI achieved the highest C-index at $p = 100$ (0.7205), while Global–Local Gaussian–IGH-VI delivered the strongest predictive performance at $p = 200$, obtaining both the highest C-index (0.6911) and the lowest Brier Score (0.1457). Although Bayesian-Lasso-Cox continued to exhibit the highest TPR

values, its substantially larger FPR values suggest reduced sparsity and a higher risk of over-selection.

Table 2. Performance metrics for Cox-Lasso, Cox-ElasticNet, Bayesian-Lasso-Cox, AGP-VI, and Global–Local Gaussian–IGH-VI at varying predictor dimensions ($\rho = 0.15$, weak correlation).

p	Method	C-Index	Brier Score	Mean TPR	SD TPR	Mean FPR	SD FPR
50	Cox-Lasso	0.8480	0.107933	0.6800	0.3701	**0.0200**	0.0326
	Cox-ElasticNet	**0.8824**	0.096017	0.9200	0.1789	0.0600	0.0548
	Bayesian-Lasso-Cox	0.8316	0.121067	**1.0000**	**0.0000**	0.7050	0.0671
	AGP-VI	0.8683	**0.090765**	0.8800	0.1304	0.0650	**0.0224**
	Global–Local Gaussian–IGH-VI	0.8366	0.116572	0.7800	0.4382	0.1450	0.0837
100	Cox-Lasso	0.7012	**0.153298**	0.4000	0.3082	0.0111	0.0192
	Cox-ElasticNet	0.7185	0.158260	0.5000	0.2739	**0.0089**	0.0199
	Bayesian-Lasso-Cox	0.6984	0.212182	**0.9600**	**0.0548**	0.7511	0.0435
	AGP-VI	**0.7205**	0.165089	0.7600	0.1140	0.0800	0.0253
	Global–Local Gaussian–IGH-VI	0.7063	0.167074	0.8000	0.1225	0.1333	0.0333
200	Cox-Lasso	0.6114	0.154959	0.3000	0.2550	**0.0042**	**0.0044**
	Cox-ElasticNet	0.6229	0.161352	0.3200	0.2588	0.0074	0.0088
	Bayesian-Lasso-Cox	0.6647	0.186858	**0.9600**	0.0548	0.6905	0.0425
	AGP-VI	0.6705	0.159540	0.6400	0.0548	0.0516	0.0164
	Global–Local Gaussian–IGH-VI	**0.6911**	**0.145723**	0.6800	**0.0447**	0.0737	0.0211
300	Cox-Lasso	0.6245	**0.160079**	0.1600	0.1517	**0.0028**	**0.0029**
	Cox-ElasticNet	0.6259	0.161819	0.1600	0.1517	0.0041	0.0057
	Bayesian-Lasso-Cox	0.6306	0.221743	**0.8200**	**0.0837**	0.6828	0.1150
	AGP-VI	0.6544	0.172993	0.5000	0.1225	0.0421	**0.0029**
	Global–Local Gaussian–IGH-VI	**0.6786**	0.161081	0.4600	0.1517	0.0393	0.0058
400	Cox-Lasso	0.5741	0.207310	0.0400	0.0548	0.0015	0.0034
	Cox-ElasticNet	0.5569	**0.204535**	0.0200	**0.0447**	**0.0005**	**0.0011**
	Bayesian-Lasso-Cox	0.4717	0.385817	**0.9000**	0.1000	0.8364	0.1220
	AGP-VI	**0.6147**	0.220444	0.4600	0.2191	0.0369	0.0067
	Global–Local Gaussian–IGH-VI	0.5491	0.235773	0.2200	0.0837	0.0159	0.0049
500	Cox-Lasso	0.5450	0.187278	0.0800	**0.0837**	**0.0012**	**0.0027**
	Cox-ElasticNet	0.5854	0.182554	0.0800	**0.0837**	0.0016	0.0037
	Bayesian-Lasso-Cox	0.5146	0.313940	**0.8800**	**0.0837**	0.8355	0.1398
	AGP-VI	**0.6464**	**0.177752**	0.4800	0.1924	0.0278	0.0040
	Global–Local Gaussian–IGH-VI	0.5977	0.199637	0.2600	0.0894	0.0082	0.0050

Note: Bold values indicate the best performance for each metric within each predictor dimension (p). Higher C-index and mean TPR values indicate better performance, whereas lower Brier scores and mean FPR values are preferred. Lower standard deviations indicate greater consistency across simulation runs. Tied values are all shown in bold.

The differences between the methods became more pronounced in the high-dimensional settings ($p = 300$, 400, and 500). The performance of Cox-Lasso and Cox-ElasticNet deteriorated considerably in terms of discrimination and sensitivity, with very low TPR values and declining C-indices as dimensionality increased. In contrast, the proposed variational inference approaches exhibited greater robustness in discrimination. Global–Local Gaussian–IGH-VI achieved the highest C-index at $p = 300$ (0.6786), while AGP-VI provided the strongest C-index at $p = 400$ and $p = 500$ (0.6147 and 0.6464, respectively). The Brier Score picture is more mixed: at $p = 500$, AGP-VI also attained the lowest (best) Brier Score (0.177752) among the competing methods; however, at $p = 400$, Cox-ElasticNet attained the lowest Brier Score (0.204535), lower than AGP-VI's (0.220444), despite Cox-ElasticNet's much poorer C-index (0.5569) at that dimension. This illustrates that strong discrimination (C-index) and strong calibration (Brier Score) do not always coincide in the same method at a given dimension.

Across all dimensions, Bayesian-Lasso-Cox consistently achieved the highest TPR values, demonstrating strong sensitivity for detecting relevant predictors. Nevertheless,

this advantage came at the cost of substantially larger FPR values, indicating a tendency to include a large number of irrelevant variables. By comparison, AGP-VI and Global–Local Gaussian–IGH-VI achieved a more favorable balance between prediction accuracy and variable selection, maintaining competitive C-index and Brier Score values while controlling false positive selections more effectively. Overall, the results suggest that the proposed variational inference methods provide a robust and scalable alternative for high-dimensional survival analysis, particularly in settings where both predictive performance and sparse variable selection are important.

The simulation design above uses a single AR(1) correlation setting, $\rho = 0.15$, corresponding to weak dependence among predictors. Therefore, we repeated the full simulation study, using the same data-generating mechanism, predictor grid, and $n_{\text{rep}} = 20$ replications per p, at two additional, substantially stronger correlation levels: $\rho = 0.50$ (moderate correlation) and $\rho = 0.90$ (high correlation, representative of the strong co-expression or linkage patterns seen in real genomic data). Tables 3 and 4 report the corresponding results, using the same metrics and format as Table 2.

Table 3. Performance metrics for Cox-Lasso, Cox-ElasticNet, Bayesian-Lasso-Cox, AGP-VI, and Global–Local Gaussian–IGH-VI at varying predictor dimensions ($\rho = 0.50$, moderate correlation)

p	Method	C-Index	Brier Score	Mean TPR	SD TPR	Mean FPR	SD FPR
50	Cox-Lasso	0.6897	0.213635	0.6750	0.3796	0.4200	0.3788
	Cox-ElasticNet	0.6805	0.183460	0.5550	0.3677	0.2387	0.2403
	Bayesian-Lasso-Cox	0.7403	0.187001	0.6350	0.1565	**0.0263**	**0.0263**
	AGP-VI	0.6318	0.273007	0.8550	0.1877	0.6850	0.1947
	Global–Local Gaussian–IGH-VI	**0.7541**	**0.176975**	**0.9100**	**0.1119**	0.1437	0.0493
100	Cox-Lasso	0.5590	0.186418	0.2400	0.2780	0.0811	0.1549
	Cox-ElasticNet	0.5759	**0.171283**	0.2400	0.1957	0.0567	0.0660
	Bayesian-Lasso-Cox	0.6298	0.265340	0.0800	**0.0834**	**0.0017**	**0.0041**
	AGP-VI	0.5633	0.277724	**0.9100**	0.1252	0.7894	0.1389
	Global–Local Gaussian–IGH-VI	**0.6516**	0.238329	0.6550	0.1638	0.1278	0.0349
200	Cox-Lasso	0.5276	0.188948	0.0600	0.1142	0.0063	0.0117
	Cox-ElasticNet	0.5329	**0.188928**	0.1000	0.1654	0.0158	0.0210
	Bayesian-Lasso-Cox	0.5553	0.259626	0.0000	**0.0000**	**0.0000**	**0.0000**
	AGP-VI	0.5347	0.273094	**0.9250**	0.0786	0.9203	0.0251
	Global–Local Gaussian–IGH-VI	**0.5651**	0.265584	0.3600	0.1501	0.0379	0.0147
300	Cox-Lasso	0.5512	**0.204317**	0.0700	0.1031	0.0167	0.0434
	Cox-ElasticNet	0.5445	0.206718	0.1100	0.1410	0.0140	0.0214
	Bayesian-Lasso-Cox	0.5099	0.338580	0.0050	**0.0224**	**0.0097**	0.0432
	AGP-VI	0.5222	0.313096	**0.9750**	0.0550	0.9095	0.0215
	Global–Local Gaussian–IGH-VI	**0.5694**	0.277070	0.1700	0.1174	0.0140	**0.0068**
400	Cox-Lasso	0.5286	0.282419	0.1150	0.0933	**0.0083**	**0.0059**
	Cox-ElasticNet	0.5400	0.192618	0.0900	0.1210	0.0119	0.0164
	Bayesian-Lasso-Cox	0.4961	0.332989	0.0950	0.1538	0.0873	0.1266
	AGP-VI	0.5119	0.289805	**0.9450**	**0.0759**	0.9110	0.0205
	Global–Local Gaussian–IGH-VI	**0.5435**	**0.190416**	0.1150	0.1226	0.0090	0.0095
500	Cox-Lasso	0.5276	0.327768	0.2100	0.1483	0.0398	0.0444
	Cox-ElasticNet	0.5266	0.196084	0.1250	0.1482	0.0137	0.0183
	Bayesian-Lasso-Cox	0.5051	0.349176	0.1550	0.1504	0.1893	0.1424
	AGP-VI	0.5276	0.326247	**0.9400**	**0.0598**	0.9106	0.0170
	Global–Local Gaussian–IGH-VI	**0.5310**	**0.194973**	0.0950	0.1234	**0.0108**	**0.0148**

Note: Bold values indicate the best performance for each metric within each predictor dimension (p). Higher C-index and mean TPR values indicate better performance, whereas lower Brier scores and mean FPR values are preferred. Lower standard deviations indicate greater consistency across simulation runs. Tied values are all shown in bold.

At $\rho = 0.50$, Global–Local Gaussian–IGH-VI shows the strongest overall discrimination, achieving the highest C-index at all six predictor dimensions ($p = 50, 100, 200, 300, 400, 500$: 0.7541, 0.6516, 0.5651, 0.5694, 0.5435, and 0.5310, respectively). It also achieves the best Brier Score at $p = 50$ (0.176975), $p = 400$ (0.190416), and $p = 500$ (0.194973), while Cox-ElasticNet achieves the lowest Brier Score at $p = 100$ (0.171283) and $p = 200$ (0.188928), and Cox-Lasso at $p = 300$ (0.204317). These results indicate that the discrimination advantage of Global–Local Gaussian–IGH-VI persists across the examined dimensions under moderate correlation, although this advantage becomes smaller at higher dimensions and does not extend uniformly to prediction error as measured by the Brier Score. A striking pattern under moderate correlation is AGP-VI's variable-selection behavior: its Mean TPR remains very high throughout (0.855 at $p = 50$, rising to 0.975 at $p = 300$), but this is accompanied by a high Mean FPR (0.6850 at $p = 50$, 0.7894 at $p = 100$, and above 0.90 at $p = 200$–500), indicating that AGP-VI's selection rule becomes markedly over-inclusive under moderate predictor correlation, selecting a large proportion of both true and null coefficients rather than achieving sparse variable selection. Global–Local Gaussian–IGH-VI retains comparatively better-controlled false positive rates across all dimensions (0.0090–0.1437), making it the more balanced of the two proposed methods under this correlation setting. Bayesian-Lasso-Cox, by contrast, achieves very low Mean FPR at the lower dimensions but at the cost of low sensitivity, particularly at $p = 200$ and $p = 300$, where its Mean TPR is 0.0000 and 0.0050, respectively.

At $\rho = 0.90$, no single method dominates discrimination: the highest C-index is split across different methods depending on p (Bayesian-Lasso-Cox at $p = 50, 100$: 0.6371 and 0.6059; AGP-VI at $p = 200$: 0.5333; Cox-ElasticNet at $p = 300$: 0.5549; and Global–Local Gaussian–IGH-VI at $p = 400, 500$: 0.5488 and 0.5348). Thus, despite the severe predictor correlation, the proposed methods remain competitive in terms of discrimination, with AGP-VI achieving the highest C-index at $p = 200$ and Global–Local Gaussian–IGH-VI achieving the highest C-index at the two largest predictor dimensions. The Brier Score shows a different pattern: the classical penalized methods, Cox-Lasso and Cox-ElasticNet, achieve the lowest Brier Score at all six predictor dimensions, with neither proposed variational method achieving the lowest Brier Score under high correlation. AGP-VI's over-inclusive selection behavior also persists under high correlation, with Mean TPR ranging from 0.835 to 0.920 alongside Mean FPR ranging from 0.7825 to 0.8672. In contrast, Global–Local Gaussian–IGH-VI maintains substantially lower false positive rates while remaining competitive in discrimination, suggesting a more balanced variable-selection behavior under severe predictor correlation.

Taken together, Tables 2–4 show that the performance of the proposed variational methods is correlation-dependent, but remains competitive as predictor correlation increases. At moderate correlation ($\rho = 0.50$), Global–Local Gaussian–IGH-VI achieves the highest C-index across all six predictor dimensions, although its discrimination advantage generally narrows as dimensionality increases. Under high correlation ($\rho = 0.90$), no single method consistently dominates discrimination. Bayesian-Lasso-Cox achieves the highest C-index at $p = 50$ and $p = 100$, AGP-VI at $p = 200$, and Cox-ElasticNet at $p = 300$. Notably, at the two highest dimensions ($p = 400$ and $p = 500$), Global–Local Gaussian–IGH-VI achieves the highest C-index (0.5488 and 0.5348, respectively), indicating that it remains competitive even under the combined challenge of severe predictor correlation and high dimensionality. The classical penalized methods, however, show an advantage in terms of Brier Score under severe correlation, achieving the lowest values across all six dimensions. For variable selection, AGP-VI maintains high sensitivity but becomes markedly over-inclusive, as reflected by its high false positive rates, whereas Global–Local Gaussian–IGH-VI provides substantially better control of false positive selections. Overall,

these findings indicate that the proposed methods remain competitive under moderate-to-high multicollinearity, while also revealing important trade-offs in prediction error and variable-selection behavior as predictor correlation and dimensionality increase.

Table 4. Performance metrics for Cox-Lasso, Cox-ElasticNet, Bayesian-Lasso-Cox, AGP-VI, and Global–Local Gaussian–IGH-VI at varying predictor dimensions ($\rho = 0.90$, high correlation)

p	Method	C-Index	Brier Score	Mean TPR	SD TPR	Mean FPR	SD FPR
50	Cox-Lasso	0.5801	0.242551	0.3000	0.3095	0.1787	0.3140
	Cox-ElasticNet	0.5876	**0.217042**	0.3000	0.3277	0.2325	0.3404
	Bayesian-Lasso-Cox	**0.6371**	0.243170	0.1100	**0.1071**	**0.0050**	**0.0103**
	AGP-VI	0.5438	0.285968	**0.8350**	0.2477	0.7825	0.3393
	Global–Local Gaussian–IGH-VI	0.6041	0.222400	0.2500	0.1573	0.1062	0.0663
100	Cox-Lasso	0.5697	**0.168921**	0.1250	0.1482	0.0189	0.0247
	Cox-ElasticNet	0.5755	0.170445	0.1850	0.1814	0.0444	0.0470
	Bayesian-Lasso-Cox	**0.6059**	0.256069	0.0300	**0.0470**	**0.0006**	**0.0025**
	AGP-VI	0.5525	0.232590	**0.8950**	0.2089	0.8494	0.1451
	Global–Local Gaussian–IGH-VI	0.5755	0.242752	0.2250	0.1333	0.0789	0.0342
200	Cox-Lasso	0.5225	0.196223	0.0600	0.1095	0.0134	0.0193
	Cox-ElasticNet	0.5256	**0.192758**	0.0500	0.0761	0.0145	0.0191
	Bayesian-Lasso-Cox	0.5233	0.321194	0.0000	**0.0000**	**0.0000**	**0.0000**
	AGP-VI	**0.5333**	0.274401	**0.9200**	0.0894	0.8482	0.0539
	Global–Local Gaussian–IGH-VI	0.4989	0.305947	0.1050	0.0887	0.0297	0.0211
300	Cox-Lasso	0.5542	**0.212119**	0.0450	0.0686	0.0116	0.0115
	Cox-ElasticNet	**0.5549**	0.213623	0.0550	0.0686	0.0148	0.0187
	Bayesian-Lasso-Cox	0.5259	0.332127	0.0050	**0.0224**	0.0110	0.0493
	AGP-VI	0.5310	0.288974	**0.9100**	0.1210	0.8571	0.0363
	Global–Local Gaussian–IGH-VI	0.5355	0.349078	0.0200	0.0410	**0.0040**	**0.0049**
400	Cox-Lasso	0.5185	0.194244	0.0350	0.0671	0.0067	**0.0086**
	Cox-ElasticNet	0.5131	**0.193783**	0.0250	**0.0639**	**0.0059**	0.0125
	Bayesian-Lasso-Cox	0.5058	0.351523	0.0900	0.1518	0.0991	0.1343
	AGP-VI	0.5447	0.264053	**0.9100**	0.0852	0.8437	0.0480
	Global–Local Gaussian–IGH-VI	**0.5488**	0.269618	0.0350	0.0813	0.0159	0.0533
500	Cox-Lasso	0.5311	**0.199585**	0.0400	**0.0598**	**0.0035**	**0.0059**
	Cox-ElasticNet	0.5334	0.202180	0.0650	0.0745	0.0128	0.0161
	Bayesian-Lasso-Cox	0.5039	0.369849	0.1200	0.1609	0.1286	0.1302
	AGP-VI	0.5332	0.283218	**0.9000**	0.1076	0.8672	0.0291
	Global–Local Gaussian–IGH-VI	**0.5348**	0.300183	0.1300	0.1809	0.0520	0.0645

Note: Bold values indicate the best performance for each metric within each predictor dimension (p). Higher C-index and mean TPR values indicate better performance, whereas lower Brier scores and mean FPR values are preferred. Lower standard deviations indicate greater consistency across simulation runs. Tied values are all shown in bold.

Furthermore, Figure 1 demonstrates that the proposed variational inference approaches exhibit greater robustness and stability as the dimensionality of the predictor space increases. Unlike the competing methods, whose predictive performance deteriorates noticeably in higher-dimensional settings, AGP-VI and Global–Local Gaussian–IGH-VI maintain relatively stable C-index and Brier Score trajectories across a wide range of predictor dimensions. These results provide strong evidence that the proposed hierarchical shrinkage framework effectively addresses the challenges of high-dimensional survival analysis by simultaneously improving predictive performance, enhancing variable selection accuracy, and maintaining scalability in increasingly complex modeling scenarios.

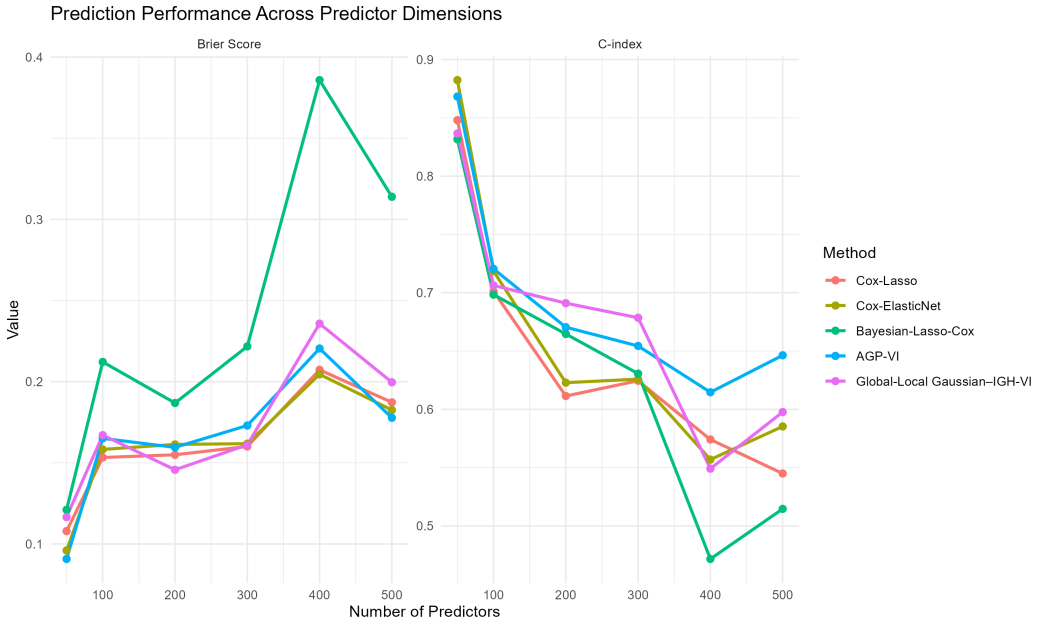


Figure 1. Prediction performance of the competing methods across increasing predictor dimensions. The left panel reports the Brier Score, where smaller values indicate better predictive accuracy, while the right panel reports the concordance index (C-index), where larger values indicate superior discriminative performance.

Figure 2 compares the variable selection performance of the competing methods across increasing predictor dimensions. The left panel presents the False Positive Rate (FPR), while the right panel shows the corresponding True Positive Rate (TPR). Together, these metrics provide insight into each method's ability to correctly identify relevant predictors while avoiding the inclusion of irrelevant variables.

Overall, Figure 2 demonstrates that the proposed variational inference methods provide a favorable compromise between variable selection sensitivity and specificity. By balancing the identification of relevant predictors with effective control of false positive selections, AGP-VI and Global–Local Gaussian–IGH-VI offer a robust and scalable framework for variable selection in high-dimensional survival analysis.

Additional insight into the variable selection behavior of the competing methods is provided through coefficient recovery plots presented in Figure A1. The supplementary figure compares the estimated regression coefficients with the true underlying coefficients for each method and illustrates the differing shrinkage characteristics of the competing approaches. Consistent with the results reported in Table 2 and Figures 1 and 2, the proposed AGP-VI and Global–Local Gaussian–IGH-VI methods achieve a favorable balance between signal recovery and shrinkage, whereas Bayesian-Lasso-Cox tends to over-select variables and the penalized Cox methods exhibit stronger coefficient shrinkage. A detailed discussion is provided in Appendix A.1.

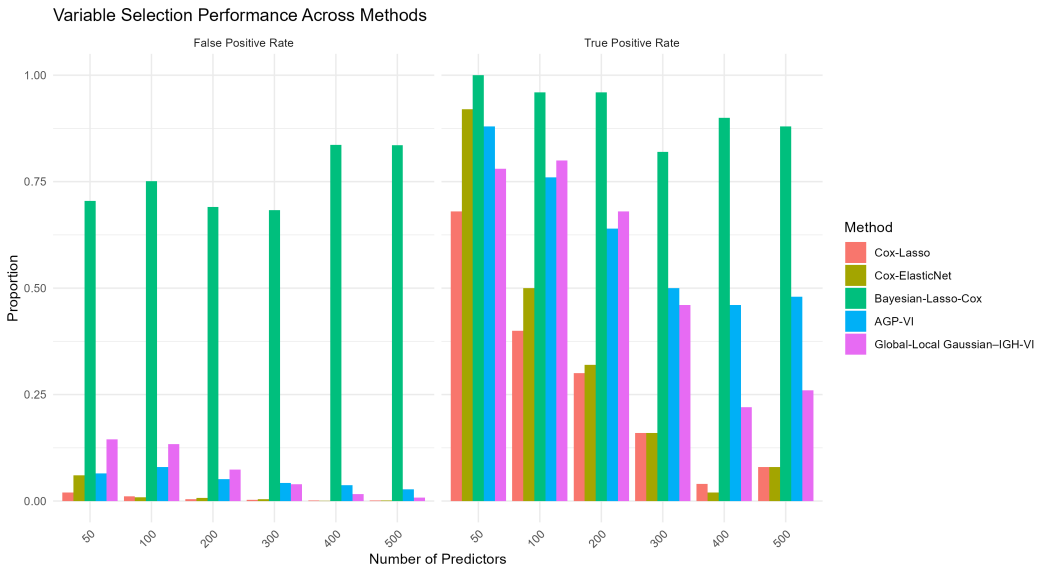


Figure 2. Variable selection performance of the competing methods across increasing predictor dimensions. The left panel reports the False Positive Rate (FPR), where smaller values indicate better control of irrelevant variable selection, while the right panel reports the True Positive Rate (TPR), where larger values indicate greater sensitivity in identifying relevant predictors.

5. Application to Real Data

To further evaluate the practical utility of the proposed methods, we applied them to six real-world survival datasets that vary substantially in both dimensionality and clinical context. These datasets comprise one low-dimensional clinical dataset (PBC) and five high-dimensional genomic datasets (DBCD, CHOP, NKI70, VDX, and TCGA Ovarian), thereby providing a comprehensive framework for assessing predictive performance and variable selection in diverse survival modeling scenarios. A summary of the six datasets used in this study is presented in Table 5.

Table 5. Description of datasets used in the study.

S/N	Dataset	Description	Subjects (n)	Predictors (p)
1	PBC	Primary Biliary Cirrhosis dataset representing patients with chronic liver disease obtained from the `survival` package in R [51].	276	20
2	DBCD	Dutch Breast Cancer Dataset containing high-dimensional gene expression profiles of breast cancer patients from the Netherlands Cancer Institute [52].	295	4921
3	CHOP	Diffuse Large B-cell Lymphoma dataset consisting of gene expression measurements for patients treated with R-CHOP chemotherapy [53].	181	3835
4	NKI70	Netherlands Cancer Institute 70-gene breast cancer dataset widely used for prognostic survival modelling and genomic risk prediction [54].	144	71
5	VDX	High-dimensional breast cancer microarray dataset from the VDX cohort available in the `breastCancerVDX` Bioconductor package [55].	344	22,283
6	TCGA Ovarian	The Cancer Genome Atlas (TCGA) ovarian cancer dataset containing genomic and survival information for ovarian serous cystadenocarcinoma patients [56].	557	13,104

For the ultra-high-dimensional genomic datasets, an initial feature screening step was performed using Random Survival Forests (RSFs; see Section 2.4.4 for a description) [47–49] prior to model fitting. The primary objective of this screening procedure was to reduce the dimensionality of the predictor space by removing noise variables and alleviating the effects of extreme multicollinearity, which is commonly encountered in large-scale genomic expression data. Variable importance measures were obtained from the fitted RSF model, and predictors whose importance exceeded the 95th percentile of the importance-score distribution for that dataset were retained, subject to retaining at least 20 and at most 1000 predictors. Screening was applied only to the four datasets with more than 100 original predictors (DBCD, CHOP, VDX, and TCGA Ovarian), while PBC and NKI70 ($p \leq 100$) were used without screening.

The 70/30 train/test split was performed before RSF screening. RSF was fitted using the training partition only, and the resulting selected predictor set was then applied to both the training and held-out test partitions. Covariate standardization statistics (mean and standard deviation for each predictor) were also estimated from the training partition only and subsequently applied to the test partition. All five competing methods (Cox-Lasso, Cox-ElasticNet, Bayesian-Lasso-Cox, AGP-VI, and Global–Local Gaussian–IGH-VI) were therefore fitted and evaluated using the same training-derived predictor set and standardization.

The RSF screening procedure was applied consistently to all five competing methods within each dataset, ensuring that the methods were compared using the same set of screened predictors. This strategy improves computational feasibility in ultra-high-dimensional settings and focuses the analysis on the most informative predictors while preserving potentially relevant biological signals. We note that the reported variable-selection counts and TPR/FPR for the four screened datasets are therefore conditional on this screening step, not unconditional properties of each method alone: a predictor removed by RSF screening before model fitting cannot subsequently be selected by any of the five methods, regardless of its true relevance. This is an inherent consequence of screening at an ultra-high dimension and should be kept in mind when interpreting selection-based results for these datasets.

Prior to the train/test split, RSF screening, and model fitting, each dataset's predictor matrix was restricted to numeric columns with non-zero variance, and rows with missing outcome or predictor values were removed. RSF screening uses internal imputation for any remaining missingness encountered during tree growing. A summary of the RSF screening results for the four high-dimensional datasets, including the dataset-specific importance thresholds and the number of predictors retained, is presented in Table 6.

Table 6. Random Survival Forest feature-screening results for the high-dimensional real datasets.

Dataset	Original p	RSF Importance Threshold	Predictors Retained
DBCD	4919	0.00143	246
CHOP	3833	0.00094	192
VDX	22,283	0.00055	1000
TCGA Ovarian	13,104	0.00049	656

Table 7 reports point estimates of the C-index, Brier Score, and time-dependent AUC without accompanying variability estimates. Because the real-data evaluation uses a single 70/30 train/test split for each dataset rather than repeated splits or cross-validation, the reported results reflect one particular partition of each dataset and do not provide an empirical estimate of variability across data partitions. Consequently, differences in

predictive performance between methods should be interpreted with appropriate caution. This limitation is discussed further in Section 6.

In addition to the performance measures described earlier, we also consider the time-dependent Area Under the Curve (AUC) as an additional measure of predictive performance. Together, these metrics provide complementary assessments of predictive discrimination, prediction accuracy, and model sparsity. By evaluating the methods across datasets with varying sample sizes, predictor dimensions, and clinical characteristics, the real-data analysis provides a rigorous assessment of the proposed hierarchical Bayesian framework under realistic survival modeling conditions.

The results of the Random Survival Forest (RSF) feature screening for the four high-dimensional real datasets are presented in Table 6. As described in Section 2.4.5, predictors were ranked according to their RSF permutation importance, and predictors with importance values above the 95th percentile of the importance-score distribution were retained, subject to a minimum of 20 and a maximum of 1000 predictors. The screening procedure retained 246 predictors for DBCD, 192 for CHOP, 1000 for VDX, and 656 for TCGA Ovarian. The maximum screening limit of 1000 predictors was reached for VDX. PBC and NKI70 were not subjected to RSF screening because their original predictor dimensions did not exceed the screening criterion of $p > 100$.

The screening results demonstrate the substantial reduction in predictor dimensionality achieved before fitting the competing survival models. The retained predictor sets were subsequently used consistently across all competing methods within each dataset. Importantly, RSF screening was performed using the training partition only, and the resulting selected predictor set was subsequently applied to the held-out test partition, thereby preventing test-set outcomes from contributing to the feature-screening procedure.

In Table 7, the best value for each performance metric is bolded separately within each dataset: the highest C-index, the lowest Brier Score, and the highest mean time-dependent AUC. Consequently, different methods may be highlighted for different performance metrics within the same dataset. For example, for DBCD, Global–Local Gaussian–IGH-VI is bolded for the C-index, whereas Cox-ElasticNet is bolded for the Brier Score and mean time-dependent AUC. Table 7 presents the predictive performance and variable-selection characteristics of the competing methods across six real-world survival datasets. Performance was assessed using the concordance index (C-index), Brier Score, mean time-dependent Area Under the Curve (AUC), and the number of selected variables under the variable-selection rule described in Section 2.4.4.

Across the six real-data applications, no single method uniformly dominated all three predictive criteria; each of the five methods achieved the best performance on at least one dataset and one predictive metric. On the CHOP dataset, Cox-Lasso achieved the best C-index, Brier Score, and mean time-dependent AUC among all five methods (0.6068, 0.2171, 0.6729), while Global–Local Gaussian–IGH-VI (0.5524, 0.3960, 0.5917) and AGP-VI (0.4793, 0.3189, 0.5291) showed lower predictive performance on this dataset. On DBCD, Global–Local Gaussian–IGH-VI achieved the highest C-index (0.7160), while Cox-ElasticNet achieved the best Brier Score and mean time-dependent AUC (0.1033, 0.7588). Global–Local Gaussian–IGH-VI also selected the fewest variables among the methods with non-zero selections on this dataset (11).

NKI70 was the strongest dataset for the proposed variational methods. Global–Local Gaussian–IGH-VI achieved the best C-index, Brier Score, and mean time-dependent AUC simultaneously (0.7918, 0.0455, 0.9653). On the low-dimensional PBC dataset, Cox-ElasticNet achieved the best C-index and mean time-dependent AUC (0.9004, 0.9460), while Bayesian-Lasso-Cox achieved the best Brier Score (0.0764). Global–Local Gaussian–IGH-VI (0.8775,

0.0786, 0.9281) and AGP-VI (0.8967, 0.0766, 0.9440) remained competitive, although neither achieved the best value for an individual predictive metric.

Table 7. Real-life survival analysis results across six datasets.

Dataset	Method	C-Index	Brier Score	Mean Time-Dependent AUC	No. Selected
CHOP	AGP-VI	0.4793	0.3189	0.5291	20
	Bayesian-Lasso-Cox	0.5500	0.3386	0.5897	0
	Cox-ElasticNet	0.5987	0.2174	0.6610	14
	Cox-Lasso	**0.6068**	**0.2171**	**0.6729**	14
	Global–Local Gaussian–IGH-VI	0.5524	0.3960	0.5917	70
DBCD	AGP-VI	0.7043	0.1122	0.7435	7
	Bayesian-Lasso-Cox	0.5160	0.1638	0.5371	52
	Cox-ElasticNet	0.7049	**0.1033**	**0.7588**	26
	Cox-Lasso	0.6981	0.1048	0.7545	22
	Global–Local Gaussian–IGH-VI	**0.7160**	0.1129	0.7434	11
NKI70	AGP-VI	0.6921	0.0769	0.8413	6
	Bayesian-Lasso-Cox	0.7654	0.0458	0.9522	2
	Cox-ElasticNet	0.6891	0.0726	0.8560	25
	Cox-Lasso	0.6921	0.0751	0.8578	16
	Global–Local Gaussian–IGH-VI	**0.7918**	**0.0455**	**0.9653**	22
PBC	AGP-VI	0.8967	0.0766	0.9440	5
	Bayesian-Lasso-Cox	0.8929	**0.0764**	0.9391	4
	Cox-ElasticNet	**0.9004**	0.0779	**0.9460**	11
	Cox-Lasso	0.8892	0.0848	0.9382	7
	Global–Local Gaussian–IGH-VI	0.8775	0.0786	0.9281	5
TCGA Ovarian	AGP-VI	**0.5962**	0.2416	**0.5989**	16
	Bayesian-Lasso-Cox	0.4364	0.4500	0.3993	188
	Cox-ElasticNet	0.5567	**0.2268**	0.5681	31
	Cox-Lasso	0.5119	0.2305	0.5094	7
	Global–Local Gaussian–IGH-VI	0.5869	0.3772	0.5976	281
VDX	AGP-VI	**0.6777**	0.1241	0.6820	18
	Bayesian-Lasso-Cox	0.6148	**0.1083**	0.6755	285
	Cox-ElasticNet	0.6438	0.1093	0.6515	54
	Cox-Lasso	0.6254	0.1121	0.6127	43
	Global–Local Gaussian–IGH-VI	0.6548	0.1858	**0.7017**	59

Note: Bold values indicate the best performance for each metric within each dataset. Higher C-index and mean time-dependent AUC values indicate better performance, whereas lower Brier scores are preferred. No. Selected is the number of predictors selected by each method and is not ranked.

On TCGA Ovarian, AGP-VI achieved the best C-index and mean time-dependent AUC (0.5962, 0.5989), while Cox-ElasticNet achieved the best Brier Score (0.2268). Global–Local Gaussian–IGH-VI showed comparable discrimination to AGP-VI (C-index = 0.5869; mean time-dependent AUC = 0.5976), but had a higher Brier Score (0.3772) and selected substantially more variables (281) than AGP-VI (16). This indicates differences in the variable-selection and predictive behavior of the global–local formulation on this dataset.

On VDX, AGP-VI achieved the best C-index (0.6777), Bayesian-Lasso-Cox achieved the best Brier Score (0.1083), and Global–Local Gaussian–IGH-VI achieved the best mean time-dependent AUC (0.7017). Bayesian-Lasso-Cox selected 285 variables on this dataset, substantially more than the other four methods.

Taken together, the real-data results show that the relative performance of the competing methods varied across datasets and evaluation criteria, reflecting differences in discrimination, calibration, and variable-selection sparsity.

Figure 3a,b present the time-dependent AUC profiles of the competing methods across the six real-world survival datasets. The time-dependent AUC evaluates the abil-

ity of a model to discriminate between individuals who experience the event and those who remain event-free at different follow-up times. Higher values indicate superior discriminative performance.

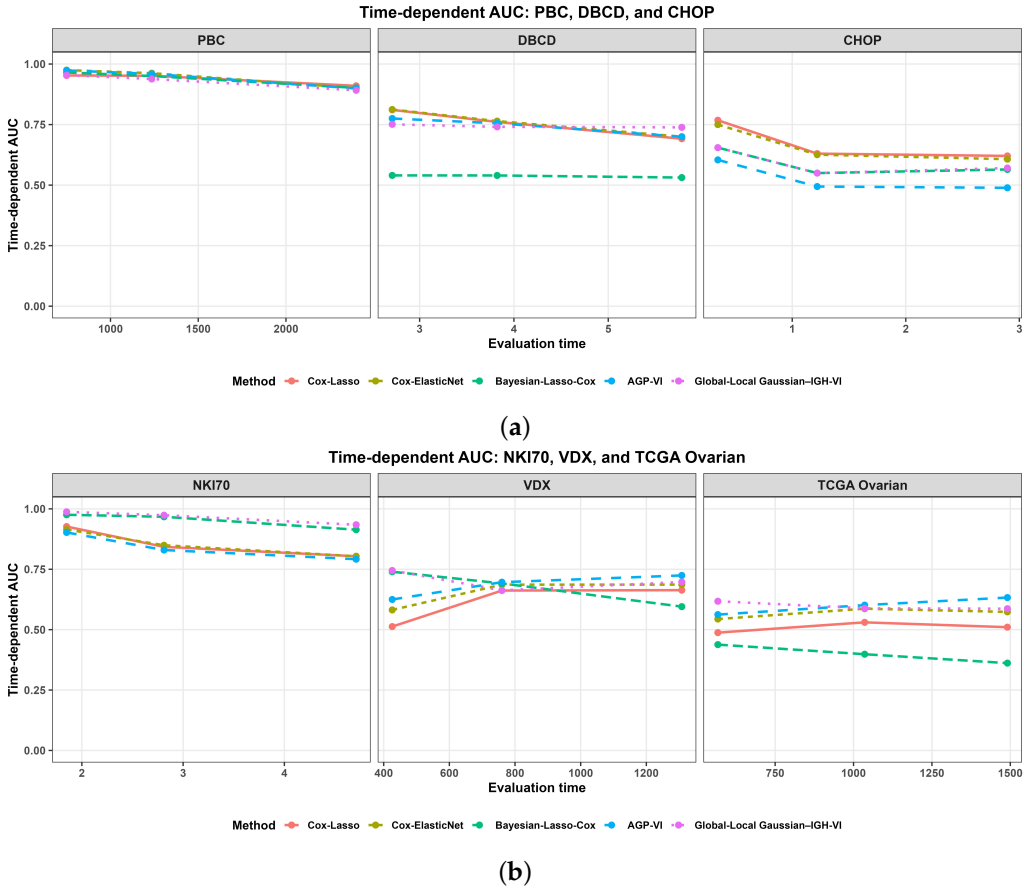


Figure 3. Time-dependent AUC profiles for the six real-world survival datasets. Higher AUC values indicate superior discriminative performance over time. Across most datasets, AGP-VI and Global–Local Gaussian–IGH-VI exhibit competitive and stable predictive discrimination throughout the follow-up period. (**a**) Time-dependent AUC curves for PBC, DBCD, and CHOP datasets. (**b**) Time-dependent AUC curves for NKI70, VDX, and TCGA Ovarian datasets.

For the PBC dataset, all methods achieved consistently high time-dependent AUC values throughout the evaluation period, with only minor differences observed among the competing approaches. AGP-VI maintained among the highest AUC values across all time points (mean AUC 0.9440, second only to Cox-ElasticNet's 0.9460). Global–Local Gaussian–IGH-VI recorded the lowest mean time-dependent AUC of the five methods on this dataset (0.9281), with its curve lying at or near the bottom of the PBC panel across the evaluation times. Similarly, for the DBCD dataset, the proposed variational inference methods did not consistently outperform the competing approaches across the full evaluation horizon: Cox-Lasso and Cox-ElasticNet visibly lead both AGP-VI and Global–Local Gaussian–IGH-VI at the first two evaluation times, and Cox-ElasticNet attains the highest mean time-dependent AUC on this dataset (0.7588), above both AGP-VI (0.7435) and Global–Local Gaussian–IGH-VI (0.7434); Global–Local Gaussian–IGH-VI overtakes the penalized Cox methods only at the final evaluation time. Although a gradual decline in AUC was observed for all methods as follow-up time increased, AGP-VI and Global–Local Gaussian–IGH-VI clearly retained superior discrimination relative to Bayesian-Lasso-Cox, whose AUC remains flat and substantially lower (around 0.53–0.54) throughout the DBCD panel; the comparison with the penalized Cox models is more mixed, as just described.

For the CHOP dataset, Cox-Lasso achieved the highest AUC values across the evaluation times, with a mean AUC of 0.6729, while AGP-VI showed the lowest AUC values in the CHOP panel, with the lowest mean AUC among the five methods (0.5291). Global–Local Gaussian–IGH-VI (mean AUC 0.5917) and Bayesian-Lasso-Cox (mean AUC 0.5897) showed intermediate performance, both below the penalized Cox methods.

The results for the NKI70 dataset reveal a gradual decline in predictive discrimination for all methods as follow-up time increased. Nevertheless, Bayesian-Lasso-Cox and Global–Local Gaussian–IGH-VI maintained the highest AUC values throughout the evaluation period, with AGP-VI performing comparably to (marginally below) the penalized Cox approaches (mean AUC 0.8413 for AGP-VI versus 0.8560 and 0.8578 for Cox-ElasticNet and Cox-Lasso, respectively). For the VDX dataset, Global–Local Gaussian–IGH-VI achieved the highest mean time-dependent AUC (0.7017, versus AGP-VI's 0.6820), but not at every individual evaluation time: it leads only at the first evaluation time, after which AGP-VI's curve is visibly the highest of all five methods at the two later evaluation times. Bayesian-Lasso-Cox does not exhibit substantially lower AUC values throughout the VDX panel—it is essentially tied for highest at the first evaluation time—but its discrimination deteriorates sharply by the final evaluation time, ending as the clearly lowest-performing method there, despite selecting a large number of variables.

For the TCGA Ovarian dataset, the proposed variational inference methods again demonstrated competitive discrimination throughout the follow-up period. Global–Local Gaussian–IGH-VI maintained the highest or near-highest AUC values across most evaluation times, while AGP-VI consistently outperformed Bayesian-Lasso-Cox; in fact, AGP-VI's curve is the highest of all five methods at the two later evaluation times and attains the best mean time-dependent AUC on this dataset (0.5989, marginally above Global–Local Gaussian–IGH-VI's 0.5976), consistent with the AGP-VI result already reported for TCGA Ovarian in Section 5. Although all methods experienced some reduction in discrimination at later time points, the decline was less pronounced for the proposed hierarchical Bayesian approaches.

Overall, the time-dependent AUC analysis is broadly, but not uniformly, consistent with the findings obtained from the C-index and Brier Score results. By mean time-dependent AUC, the proposed methods lead on three of the six datasets (NKI70, VDX, and TCGA Ovarian) but trail the penalized Cox methods on the other three (PBC, DBCD, and CHOP), so this should not be read as showing the proposed methods superior "across the majority of datasets." AGP-VI and Global–Local Gaussian–IGH-VI each exhibited their own strongest discriminative performance on a dataset-dependent subset of datasets rather than uniformly relative to the competing methods, mirroring the dataset-dependent trade-offs already reported for the C-index and Brier Score in Section 5. The results do, however, demonstrate that the proposed hierarchical shrinkage framework can provide stable predictive performance over time on the datasets where it is competitive, while the framework's robustness across datasets with substantially different dimensional characteristics and clinical contexts is dataset-dependent rather than uniform, as detailed above.

To further assess the numerical stability of the proposed variational inference algorithms, ELBO convergence plots for AGP-VI and Global–Local Gaussian–IGH-VI are provided in Figures A2–A7. Across the six real-data applications, the ELBO values generally increased and stabilized over iterations, indicating that the proposed optimization procedures achieved stable convergence. These supplementary results support the computational reliability of the proposed variational inference framework.

6. Conclusions

This study developed a hierarchical Bayesian Cox modeling framework for high-dimensional survival analysis, combining an Adaptive Gaussian Prior (AGP) and a modified Global–Local Gaussian–Inverse-Gamma Hierarchical (GL-IGH) shrinkage prior with the Cox partial likelihood. To enable scalable inference in high-dimensional settings, variational inference (VI) was employed to derive the AGP-VI and Global–Local Gaussian–IGH-VI algorithms; as discussed in Section 2.3, these algorithms target a partial-likelihood (pseudo-Bayesian) posterior and combine coordinate-ascent variational updates with empirical-Bayes point estimates for the shrinkage-variance components, rather than performing fully Bayesian uncertainty propagation over those components.

We conducted extensive simulation studies and analyzed six real-world survival datasets. On the real-data applications (Section 5), no single method uniformly dominated all three predictive criteria across the six datasets. Global–Local Gaussian–IGH-VI achieved the best C-index, Brier Score, and mean time-dependent AUC simultaneously on NKI70 and the best C-index on DBCD, while also attaining the highest mean AUC on VDX. AGP-VI achieved the highest C-index and mean time-dependent AUC on TCGA Ovarian and the highest C-index on VDX. Classical penalized and Bayesian Cox approaches remained competitive, achieving the best Brier Score on four of the six datasets (CHOP, DBCD, PBC, and VDX) and the best C-index and mean time-dependent AUC on CHOP and PBC. Variable-selection sparsity for the proposed methods was similarly dataset-dependent, ranging from relatively sparse solutions on some datasets to substantially less sparse solutions on others. In particular, on TCGA Ovarian, Global–Local Gaussian–IGH-VI selected 281 variables compared with 16 for AGP-VI, despite comparable discrimination. Overall, the results indicate that the relative performance of the proposed and competing methods depends on the dataset and the performance criterion considered.

In addition, the proposed variational inference algorithms demonstrated stable convergence behavior once the optimization procedure described in Section 2.3 was met and substantially lower computational cost than MCMC-based Bayesian sampling approaches, making the framework attractive for large-scale biomedical studies involving thousands of molecular features.

Limitations

The following methodological limitations and implementation challenges were identified during the development and evaluation of the proposed models. First, because inference targets the Cox partial-likelihood posterior rather than a fully generative survival model, the credible intervals and shrinkage-variance estimates reported here do not carry the same interpretation as those from a fully Bayesian generative model (Section 2.3). Second, the variance/hyperprior updates in both AGP-VI and Global–Local Gaussian–IGH-VI combine a coordinate-ascent-derived component with an empirical-Bayes component (Equation (17) and surrounding discussion). Uncertainty in these variance components is not propagated across iterations, so the framework provides only partial uncertainty quantification for these components.

Third, Random Survival Forest feature screening and covariate standardization are performed using the training partition only and are subsequently applied to the held-out test partition (Section 5). However, a single 70/30 train/test split per dataset is used rather than repeated splits or cross-validation. Consequently, the reported real-data performance measures reflect one particular partition of each dataset and do not provide an empirical assessment of variability across data partitions. Future evaluation using repeated train/test splits or cross-validation would provide a more comprehensive assessment of the stability of these performance estimates.

Fourth, Global–Local Gaussian–IGH-VI required additional numerical safeguards to achieve stable convergence in some of the higher-dimensional real-data settings. Although the adaptive initialization procedure described in Section 2.3 provided stable convergence for the reported analyses, the coupled global–local variance update remains more numerically demanding than the simpler per-coordinate update used by AGP-VI. Further investigation of alternative damped or trust-region updates for the global variance component may therefore be useful. In addition, on TCGA Ovarian, Global–Local Gaussian–IGH-VI selected substantially more variables (281) than AGP-VI (16), despite comparable discrimination and a higher Brier Score (Section 5). This suggests that the variable-selection and predictive behavior of the global–local formulation may vary across datasets and warrants further investigation.

The hierarchical structure of the proposed models nonetheless provides flexibility for adaptive coefficient-specific shrinkage while maintaining computational tractability. Future research directions include: extending the framework to accommodate nonlinear covariate effects, time-varying coefficients, competing risks, and multi-omics data integration; replacing the empirical-Bayes/CAVI blend with a fully conjugate variational family over the shrinkage variances (e.g., an explicit Inverse-Gamma variational factor with no plug-in blending) so that uncertainty in the shrinkage variances is fully propagated; conducting formal sensitivity analyses over the blend weight λ and the Inverse-Gamma hyperparameters (a, b, a_0, b_0); and evaluating calibration (credible-interval coverage) directly, rather than inferring it indirectly from the Brier Score. These developments have the potential to address the limitations above and further strengthen the framework's predictive performance and uncertainty quantification in complex survival modeling settings.

Overall, the proposed hierarchical Bayesian Cox framework offers a computationally scalable approach to high-dimensional survival analysis that most reliably improves discrimination and variable-selection sparsity, with calibration and formal uncertainty quantification identified above as areas requiring further development before the framework's advantages can be considered fully established.

Author Contributions: Conceptualization, K.A.D.; Methodology, B.I.D. and K.A.D.; Software, K.A.D.; Validation, B.I.D.; Formal analysis, B.I.D.; Investigation, B.I.D. and K.A.D.; Resources, K.A.D.; Data curation, B.I.D.; Writing—original draft, B.I.D. and K.A.D.; Writing—review & editing, K.A.D.; Visualization, B.I.D.; Supervision, K.A.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets analyzed in this study are all previously published and are publicly available as follows. The Primary Biliary Cirrhosis (PBC) dataset is included in the survival R package (Therneau, 2024 [51]), available through the Comprehensive R Archive Network (CRAN). The CHOP diffuse large B-cell lymphoma dataset (Lenz et al., 2008 [53]) is available from the Gene Expression Omnibus (GEO) under accession number GSE10846. The NKI70 dataset (van 't Veer et al., 2002 [54]) is included as nki70 in the penalized R package, available through CRAN. The VDX dataset is available through the breastCancerVDX Bioconductor package. The TCGA Ovarian dataset is available from the Genomic Data Commons (GDC) Data Portal under project TCGA-OV. The Dutch Breast Cancer Dataset (DBCD) (van Houwelingen et al., 2006 [52]) derives from the 295-patient Netherlands Cancer Institute breast cancer cohort of van de Vijver et al. (2002) [54]. [https://CRAN.R-project.org/package=survival, accessed on 10 June 2026 [48,49,51], GSE10846].

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AGP	Adaptive Gaussian Prior
AGP-VI	Adaptive Gaussian Prior Variational Inference
AUC	Area Under the Curve
C-index	Concordance Index
Cox-PH	Cox Proportional Hazards
DBCD	Dutch Breast Cancer Dataset
CHOP	B-Cell Lymphoma dataset treated with R-CHOP chemotherapy
ELBO	Evidence Lower Bound
FP	False Positive
FPR	False Positive Rate
GL-IGH	Global–Local Gaussian–Inverse-Gamma Hierarchical Prior
Global–Local Gaussian–IGH-VI	Global–Local G-Inverse-Gamma Hierarchical Variational Inference
IBS	Integrated Brier Score
IGH	Inverse-Gamma Hyperprior
KL	Kullback–Leibler
MCMC	Markov Chain Monte Carlo
NKI70	Netherlands Cancer Institute 70-Gene Dataset
PBC	Primary Biliary Cirrhosis Dataset
PHBC	Penalized Hierarchical Bayesian Cox Model
RSF	Random Survival Forest
TCGA	The Cancer Genome Atlas
TP	True Positive
TPR	True Positive Rate
VDX	van de Vijver Breast Cancer Dataset
VI	Variational Inference

Appendix A. Supplementary Analyses

This appendix contains supplementary results supporting the simulation and real-data analyses presented in the main manuscript. Specifically, we provide coefficient recovery plots and Evidence Lower Bound (ELBO) convergence diagnostics for the proposed Adaptive Gaussian Prior Variational Inference (AGP-VI) and Global–Local Gaussian–IGH-VI methods.

Appendix A.1. Coefficient Recovery Analysis

Figure A1 presents the coefficient recovery performance of the competing methods in a representative high-dimensional simulation setting. The figure provides additional insight into the shrinkage behaviour underlying the prediction and variable selection results reported in the main manuscript.

Cox-Lasso and Cox-ElasticNet exhibit strong shrinkage toward zero, resulting in sparse coefficient estimates but also substantial attenuation of the true signal magnitudes. Bayesian-Lasso-Cox recovers a larger proportion of the true signals but produces many sizeable non-zero coefficients among irrelevant predictors, explaining its consistently high true positive rates and false positive rates.

In contrast, AGP-VI and Global–Local Gaussian–IGH-VI provide a more balanced shrinkage pattern. Both methods effectively shrink irrelevant coefficients toward zero while preserving meaningful signal information for important predictors. The resulting coefficient estimates are more stable and interpretable, supporting the improved balance between predictive accuracy and variable selection observed in the simulation study. These findings

further illustrate the effectiveness of the proposed hierarchical shrinkage framework for high-dimensional survival analysis.

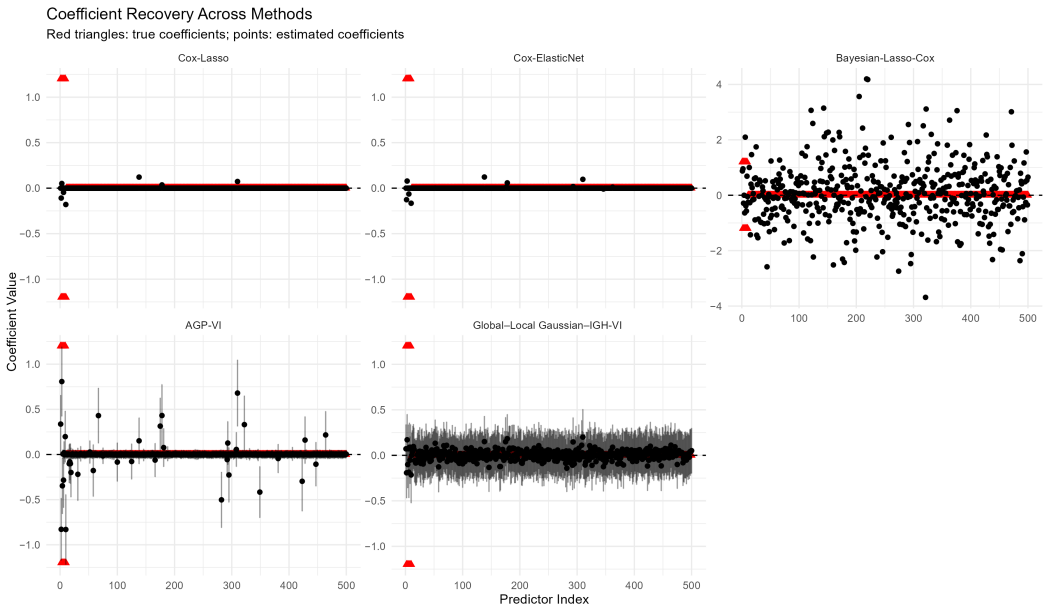
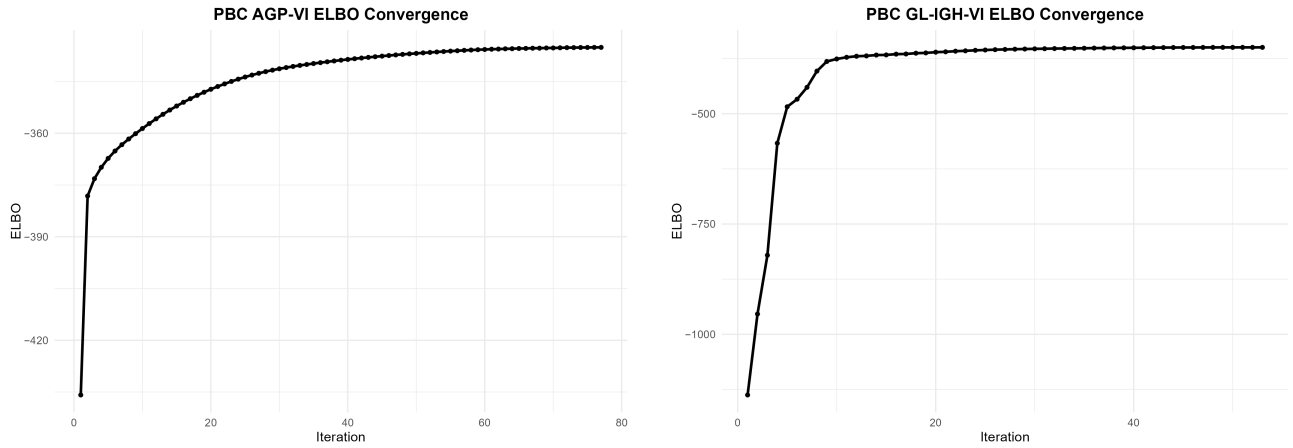


Figure A1. Coefficient recovery performance across competing methods. Red triangles denote the true regression coefficients; the non-zero coefficients appear at ± 1.2, and the overlapping triangles for the true zero coefficients form the red band along zero. Black points represent the estimated coefficients, and the black dashed horizontal line marks zero. For the variational inference methods (AGP-VI and Global–Local Gaussian–IGH-VI), grey vertical lines indicate approximate 95% posterior credible intervals. Note that the Bayesian-Lasso-Cox panel uses a different vertical scale.

Appendix A.2. ELBO Convergence Diagnostics

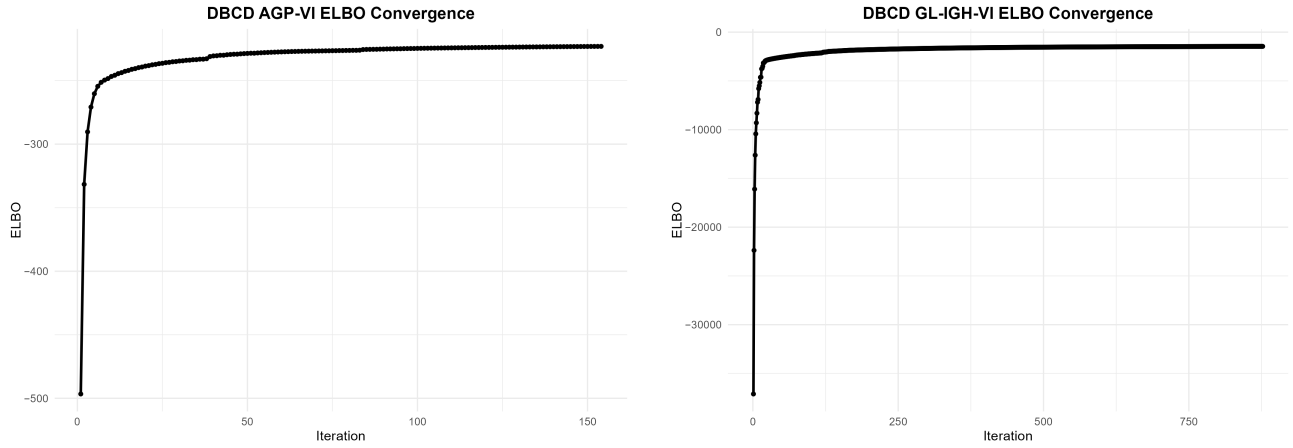
The ELBO convergence plots for the two proposed variational inference methods are presented in Figures A2–A7. These plots provide diagnostic evidence on the numerical stability of the AGP-VI and Global–Local Gaussian–IGH-VI algorithms across the six real-world datasets. The ELBO trajectories show stable convergence behavior for both AGP-VI and Global–Local Gaussian–IGH-VI across the six real-world datasets. In general, the ELBO increases rapidly during the initial iterations and subsequently improves more gradually toward a stable plateau. AGP-VI typically reaches this stable region in fewer iterations, whereas Global–Local Gaussian–IGH-VI requires more iterations for some datasets, particularly in the higher-dimensional settings.

The number of iterations plotted differs considerably across dataset–method combinations, ranging from about 48 iterations for PBC under Global–Local Gaussian–IGH-VI to about 880 iterations for DBCD under Global–Local Gaussian–IGH-VI. The TCGA Ovarian panel for Global–Local Gaussian–IGH-VI (Figure A7) shows a distinct two-phase pattern, with a rapid initial rise followed by a longer period of gradual improvement before approaching the final plateau. For AGP-VI, all six datasets approach a stable plateau within fewer than 350 displayed iterations, with the low-dimensional PBC dataset requiring fewer than 80 iterations. These results provide empirical evidence of numerical convergence of the implemented algorithms across the datasets considered.



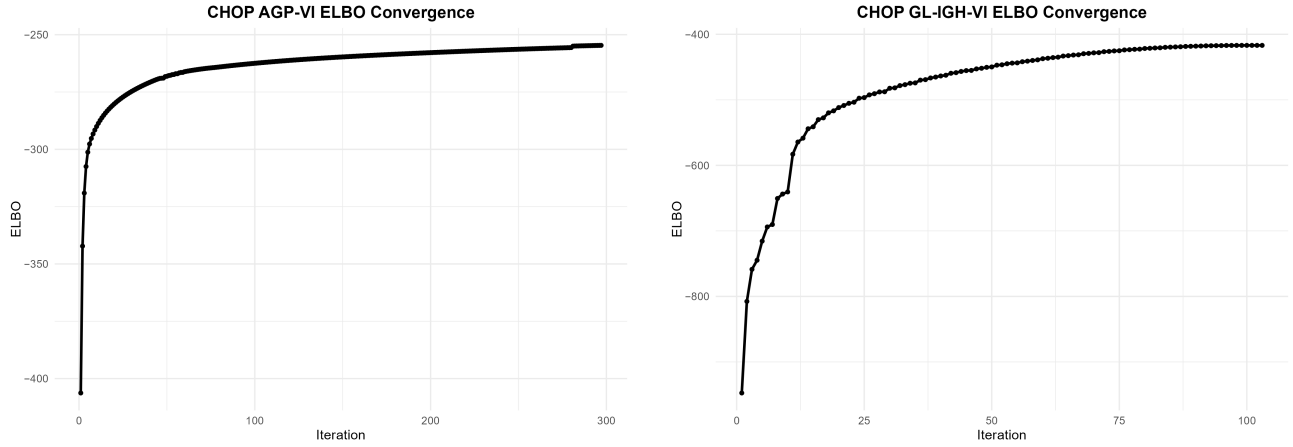
(**a**) (**b**)

Figure A2. ELBO convergence for the PBC dataset. (**a**) PBC: AGP-VI; (**b**) PBC: GL-IGH-VI.



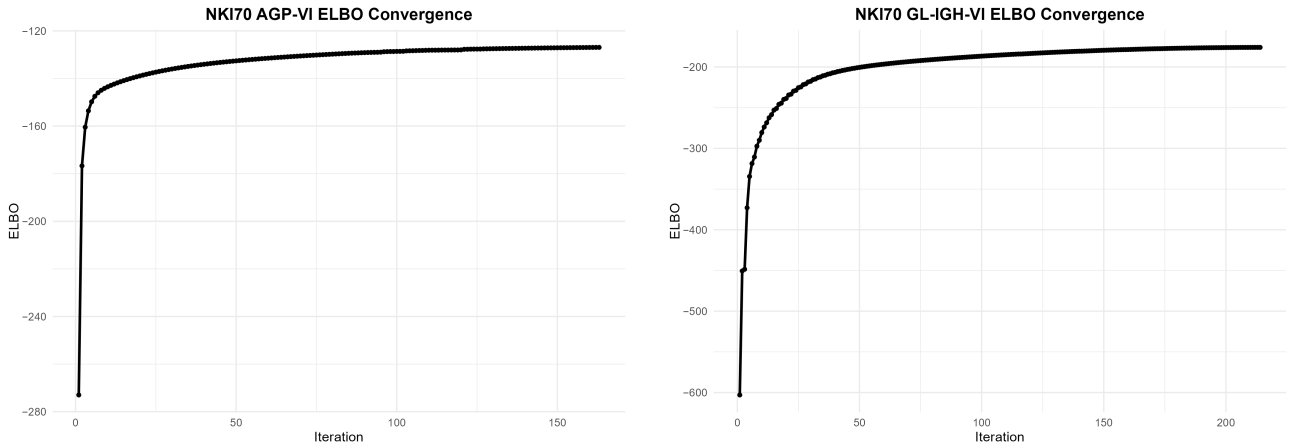
(**a**) (**b**)

Figure A3. ELBO convergence for the DBCD dataset. (**a**) DBCD: AGP-VI; (**b**) DBCD: GL-IGH-VI.



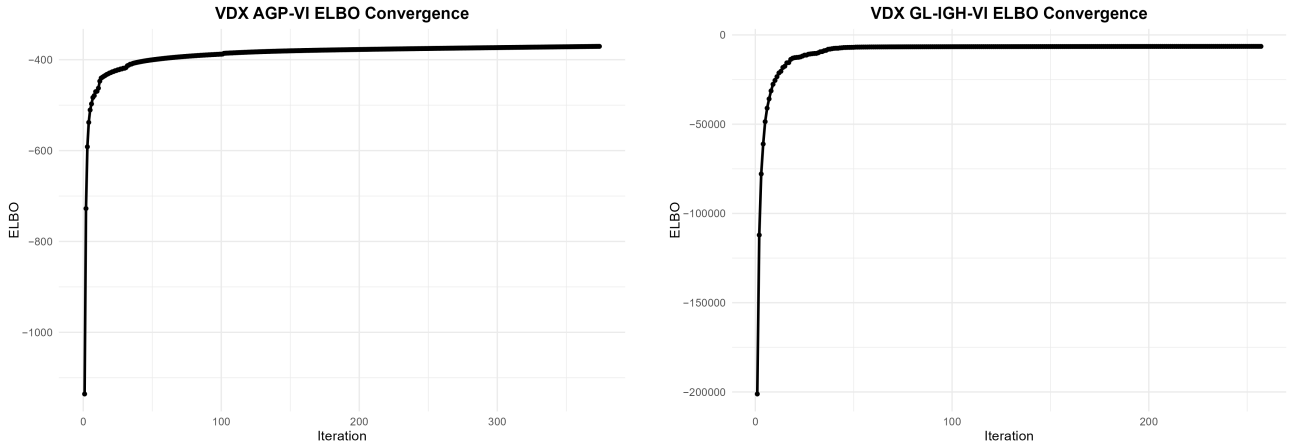
(**a**) (**b**)

Figure A4. ELBO convergence for the CHOP dataset. (**a**) CHOP: AGP-VI; (**b**) CHOP: GL-IGH-VI.



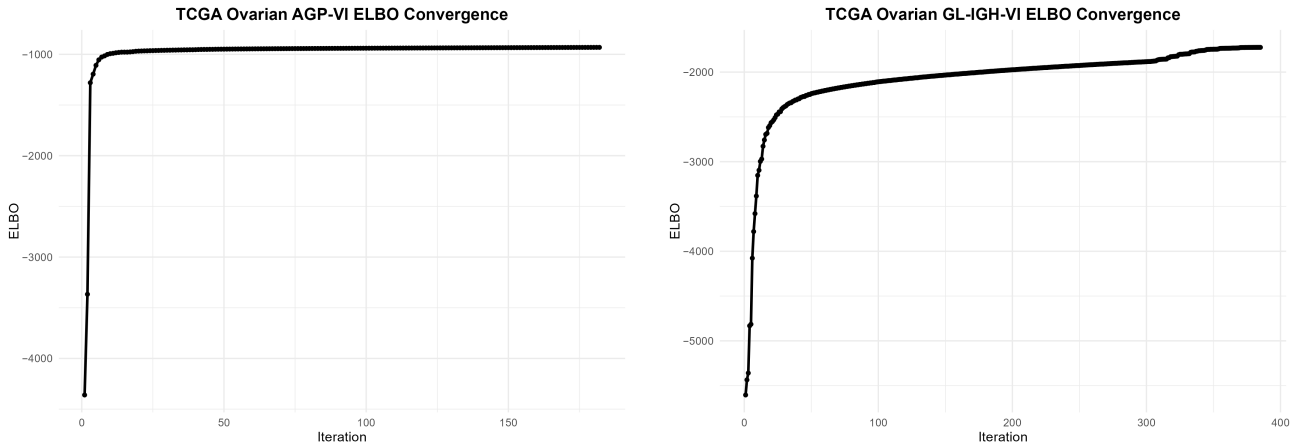
(**a**) (**b**)

Figure A5. ELBO convergence for the NKI70 dataset. (**a**) NKI70: AGP-VI; (**b**) NKI70: GL-IGH-VI.



(**a**) (**b**)

Figure A6. ELBO convergence for the VDX dataset. (**a**) VDX: AGP-VI; (**b**) VDX: GL-IGH-VI.



(**a**) (**b**)

Figure A7. ELBO convergence for the TCGA Ovarian dataset. (**a**) TCGA Ovarian: AGP-VI; (**b**) TCGA Ovarian: GL-IGH-VI.

https://doi.org/10.3390/math14203763

References

1. MacDonald, W.J.; Purcell, C.; Pinho-Schwermann, M.; Stubbs, N.M.; Srinivasan, P.R.; El-Deiry, W.S. Heterogeneity in Cancer. *Cancers* **2025**, *17*, 441. [CrossRef] [PubMed]
2. Fu, Y.C.; Liang, S.B.; Luo, M.; Wang, X.P. Intratumoral heterogeneity and drug resistance in cancer. *Cancer Cell Int.* **2025**, *25*, 103. [CrossRef] [PubMed]
3. Li, A.Z.; Du, Y.; Liu, Y.; Chen, L.; Liu, R. Benchmarking computational methods for multi-omics biomarker discovery in cancer. *bioRxiv* **2025**. [CrossRef]
4. Deng, C.; Xie, C.; Li, Z.; Mei, J.; Wang, K. Multi-omics analysis identifies diagnostic circulating biomarkers and potential therapeutic targets, revealing IQGAP1 as an oncogene in gastric cancer. *npj Precis. Oncol.* **2025**, *9*, 105. [CrossRef] [PubMed]
5. Mok, L.; Kim, Y.; Lee, S.; Choi, S.; Lee, S.; Jang, J.Y.; Park, T. HisCoM-PAGE: Hierarchical Structural Component Models for Pathway Analysis of Gene Expression Data. *Genes* **2019**, *10*, 931. [CrossRef] [PubMed]
6. Dauda, K.A.; Olorede, K.O.; Banjoko, A.W.; Yahya, W.B.; Ayipo, Y.O. Genetic Diagnosis, Classification, and Risk Prediction in Cancer Using Next-Generation Sequencing in Oncology. In *Computational Approaches in Biomaterials and Biomedical Engineering Applications*; CRC Press: Boca Raton, FL, USA, 2024; pp. 107–122.
7. Li, H.; Luan, Y. Kernel Cox Regression Models for Linking Gene Expression Profiles to Censored Survival Data. *Pac. Symp. Biocomput.* **2003**, *8*, 65–76. [CrossRef]
8. Dauda, K.A.; Lamidi, R.K. Consistency and Stability in Feature Selection for High-Dimensional Microarray Survival Data in Diffuse Large B-Cell Lymphoma Cancer. *Data* **2025**, *10*, 26. [CrossRef]
9. Gui, J.; Li, H. Penalized Cox Regression Analysis in the High-Dimensional and Low-Sample Size Settings, with Applications to Microarray Gene Expression Data. *Bioinformatics* **2005**, *21*, 3001–3008. [CrossRef] [PubMed]
10. Dauda, K.A.; Adeniyi, E.J.; Lamidi, R.K.; Wahab, O.T. Exploring Flexible Penalization of Bayesian Survival Analysis Using Beta Process Prior for Baseline Hazard. *Computation* **2025**, *13*, 21. [CrossRef]
11. Cox, D.R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B Methodol.* **1972**, *34*, 187–220. [CrossRef]
12. Devarajan, K.; Ebrahimi, N. A Semi-Parametric Generalization of the Cox Proportional Hazards Regression Model: Inference and Applications. *Commun. Stat.—Theory Methods* **2011**, *40*, 3256–3275. [CrossRef]
13. Dauda, K.A.; Pradhan, B.; Shankar, B.U.; Mitra, S. Decision Tree for Modeling Survival Data with Competing Risks. *Biocybern. Biomed. Eng.* **2019**, *39*, 697–708. [CrossRef]
14. Abadi, A.; Yavari, P.; Dehghani-Arani, M.; Alavi-Majd, H.; Ghasemi, E.; Amanpour, F.; Bajdik, C. Cox Models Survival Analysis Based on Breast Cancer Treatments. *Iran. J. Cancer Prev.* **2014**, *7*, 124–129. Available online: https://pmc.ncbi.nlm.nih.gov/articles/PMC4209689/ (accessed on 10 June 2026). [PubMed]
15. Dauda, K.A. Optimal Tuning of Random Survival Forest Hyperparameter with an Application to Liver Disease. *Malays. J. Med. Sci.* **2022**, *29*, 67. [CrossRef] [PubMed]
16. Bradic, J.; Fan, J.; Jiang, J. Regularization for Cox's Proportional Hazards Model with NP-Dimensionality. *Ann. Stat.* **2011**, *39*, 3092–3120. [CrossRef] [PubMed]
17. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 267–288. [CrossRef]
18. Hoerl, A.E.; Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
19. Fan, J.; Li, R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
20. Hans, C. Bayesian Lasso Regression. *Biometrika* **2009**, *96*, 835–845. [CrossRef]
21. Zhang, C.-H. Nearly Unbiased Variable Selection under Minimax Concave Penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef] [PubMed]
22. Zou, H.; Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
23. Doroh, B.I.; Dauda, K.A.; Lamidi, R.K. A Bayesian Hierarchical Cox Model with Elastic Net Regularization for Improved Survival Prediction and Feature Selection. *Mathematics* **2026**, *14*, 767. [CrossRef]
24. Bühlmann, P. Boosting for High-Dimensional Linear Models. *Ann. Stat.* **2006**, *34*, 559–583. [CrossRef]
25. Carvalho, C.M.; Polson, N.G.; Scott, J.G. The Horseshoe Estimator for Sparse Signals. *Biometrika* **2010**, *97*, 465–480. [CrossRef]
26. Bhadra, A.; Datta, J.; Polson, N.G.; Willard, B.T. The Horseshoe+ Estimator of Ultra-Sparse Signals. *Bayesian Anal.* **2017**, *12*, 1105–1131. [CrossRef]
27. Ishwaran, H.; Rao, J.S. Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *Ann. Stat.* **2005**, *33*, 730–773. [CrossRef]
28. Polson, N.G.; Scott, J.G. Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction. In *Bayesian Statistics 9*; Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M., Eds.; Oxford University Press: Oxford, UK, 2011; pp. 501–538. [CrossRef]

29. Raftery, A.E.; Madigan, D.; Hoeting, J.A. Bayesian Model Averaging for Linear Regression Models. *J. Am. Stat. Assoc.* **1997**, *92*, 179–191. [CrossRef]
30. George, E.I.; McCulloch, R.E. Variable Selection via Gibbs Sampling. *J. Am. Stat. Assoc.* **1993**, *88*, 881–889. [CrossRef]
31. Pesaran, M.H.; Smith, R.P. A Bayesian Analysis of Linear Regression Models with Highly Collinear Regressors. *Econom. Stat.* **2019**, *11*, 1–21. [CrossRef]
32. Gelman, A. Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Anal.* **2006**, *1*, 515–533. [CrossRef]
33. Kucukelbir, A.; Tran, D.; Ranganath, R.; Gelman, A.; Blei, D.M. Automatic Differentiation Variational Inference. *J. Mach. Learn. Res.* **2017**, *18*, 1–45. Available online: https://jmlr.org/papers/v18/16-107.html (accessed on 10 June 2026).
34. Carbonetto, P.; Stephens, M. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Anal.* **2012**, *7*, 73–108. [CrossRef]
35. Haredasht, F.N.; Dauda, K.A.; Vens, C. Exploiting Censored Information in Self-Training for Time-to-Event Prediction. *IEEE Access* **2023**, *11*, 96831–96840. [CrossRef]
36. Guo, R.; Wang, X.; Fan, K.; Broderick, T. Boosting Variational Inference. *arXiv* **2016**, arXiv:1611.05559. [CrossRef]
37. Jensen, J.L.W.V. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Math.* **1906**, *30*, 175–193. [CrossRef]
38. Struski, Ł.; Mazur, M.; Batorski, P.; Spurek, P.; Tabor, J. Bounding Evidence and Estimating Log-Likelihood in VAE. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*; PMLR: Breckenridge, CO, USA, 2023; Volume 206, pp. 5036–5051. Available online: https://proceedings.mlr.press/v206/struski23a.html (accessed on 10 June 2026).
39. Mazur, M.; Kościelniak, P.; Struski, Ł. Tight Bounds on Jensen's Gap: Novel Approach with Applications in Generative Modeling. *arXiv* **2025**, arXiv:2502.03988. [CrossRef]
40. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
41. Challis, E.; Barber, D. Gaussian Kullback–Leibler Approximate Inference. *J. Mach. Learn. Res.* **2013**, *14*, 2239–2286. Available online: https://jmlr.org/papers/v14/challis13a.html (accessed on 10 June 2026).
42. Tibshirani, R. The Lasso Method for Variable Selection in the Cox Model. *Stat. Med.* **1997**, *16*, 385–395. [CrossRef]
43. Harrell, F.E.; Califf, R.M.; Pryor, D.B.; Lee, K.L.; Rosati, R.A. Evaluating the Yield of Medical Tests. *JAMA* **1982**, *247*, 2543–2546. [CrossRef]
44. Uno, H.; Cai, T.; Pencina, M.J.; D'Agostino, R.B.; Wei, L.-J. On the C-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Stat. Med.* **2011**, *30*, 1105–1117. [CrossRef] [PubMed]
45. Graf, E.; Schmoor, C.; Sauerbrei, W.; Schumacher, M. Assessment and Comparison of Prognostic Classification Schemes for Survival Data. *Stat. Med.* **1999**, *18*, 2529–2545. [CrossRef]
46. Gerds, T.A.; Schumacher, M. Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biom. J.* **2006**, *48*, 1029–1040. [CrossRef] [PubMed]
47. Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; Lauer, M.S. Random Survival Forests. *Ann. Appl. Stat.* **2008**, *2*, 841–860. [CrossRef]
48. Ishwaran, H.; Kogalur, U.B.; Gorodeski, E.Z.; Minn, A.J.; Lauer, M.S. High-Dimensional Variable Selection for Survival Data. *J. Am. Stat. Assoc.* **2010**, *105*, 205–217. [CrossRef]
49. Ishwaran, H.; Kogalur, U.B. randomForestSRC: Fast Unified Random Forests for Survival, Regression, and Classification. R Package Version 3.3.3. 2024. Available online: https://cran.r-project.org/package=randomForestSRC (accessed on 10 June 2026).
50. Park, T.; Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686. [CrossRef]
51. Therneau, T.M. A Package for Survival Analysis in R. R Package Version 3.5-7. 2024. Available online: https://CRAN.R-project.org/package=survival (accessed on 10 June 2026).
52. van Houwelingen, H.C.; Bruinsma, T.; Hart, A.A.M.; van't Veer, L.J.; Wessels, L.F.A. Cross-Validated Cox Regression on Microarray Gene Expression Data. *Stat. Med.* **2006**, *25*, 3201–3216. [CrossRef] [PubMed]
53. Lenz, G.; Wright, G.; Dave, S.S.; Xiao, W.; Powell, J.; Zhao, H.; Xu, W.; Tan, B.; Goldschmidt, N.; Iqbal, J.; et al. Stromal Gene Signatures in Large-B-Cell Lymphomas. *N. Engl. J. Med.* **2008**, *359*, 2313–2323. [CrossRef] [PubMed]
54. van 't Veer, L.J.; Dai, H.; van de Vijver, M.J.; He, Y.D.; Hart, A.A.; Mao, M.; Peterse, H.L.; van der Kooy, K.; Marton, M.J.; Witteveen, A.T.; et al. Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature* **2002**, *415*, 530–536. [CrossRef] [PubMed]
55. Schroeder, M.; Haibe-Kains, B.; Culhane, A.; Sotiriou, C.; Bontempi, G.; Quackenbush, J. breastCancerVDX: Gene Expression Datasets for Breast Cancer. Bioconductor Package. 2024. Available online: https://bioconductor.org/packages/breastCancerVDX (accessed on 10 June 2026).
56. The Cancer Genome Atlas Research Network. Integrated Genomic Analyses of Ovarian Carcinoma. *Nature* **2011**, *474*, 609–615. [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.