Survival driven deconvolution (deSurv) reveals prognostic and interpretable tumor gene signatures

Amber Young[a,1,2], **Alisa**[b], **Didong**[a], **and Naim**[a,c]

[a]University of North Carolina at Chapel Hill, Department, Street, City, State, Zip; [b]Another University Department, Street, City, State, Zip

This manuscript was compiled on September 24, 2025

Molecular subtyping in cancer is an ongoing problem that relies on the identification of robust and replicable gene signatures. While transcriptomic profiling has revealed recurrent gene expression patterns in various types of cancer, the prognostic value of these signatures is typically evaluated in retrospect. This is due to the reliance on unsupervised learning methods for identifying cell-type-specific signals and clustering patients into molecular subtypes. Here we present a Survival-driven Deconvolution tool (deSurv) that integrates bulk RNA-sequencing data with patient survival information to identify cell-type-enriched gene signatures associated with prognosis. Applying deSurv to various cohorts in pancreatic, bladder, and colorectal cancer, we uncover previously unrecognized gene signatures linked to tumor, stromal, and immune compartments, including â€¦ Several identified signatures exhibit consistent prognostic value across cohorts and cancer types and demonstrate potential as therapeutic targets or biomarkers. Our approach highlights the value of using patient outcomes during gene signature discovery.

one | two | optional | optional | optional

Molecular subtyping has become a cornerstone of precision oncology, enabling the stratification of cancer patients based on distinct gene expression patterns [Kandoth et al., 2013; Hoadley et al., 2018]. This stratification informs prognosis, guides therapeutic decisions, and enhances our understanding of tumor biology.

Nonnegative matrix factorization (NMF), first introduced by Lee and Seung for image decomposition [Lee & Seung, 1999], has become a widely used technique for dimensionality reduction and feature learning. Unlike other matrix factorization approaches, the nonnegativity constraint in NMF yields an additive, parts-based representation that facilitates interpretability of latent factors. These properties have motivated extensive methodological development, leading to extensions that incorporate domain knowledge, structural constraints, or supervision. Examples include sparsity-regularized formulations [Hoyer, 2004], graph-regularized NMF [Cai et al., 2008], and more recent supervised formulations such as NMFProfiler for multi-omics integration and clinical stratification [Mercadié et al., 2025], as well as Bayesian multi-study NMF frameworks for mutational signatures [Grabski et al., 2025]. Collectively, these frameworks highlight the flexibility of NMF as a foundation for problem-specific decompositions.

High-throughput cancer transcriptomic datasets pose unique challenges for matrix factorization: they are high-dimensional, reflect mixtures of tumor and stromal populations, and are increasingly paired with censored survival outcomes. Standard applications of NMF in this domain typically follow a two-stage procedure—first identifying latent factors in an unsupervised manner, then testing their association with overall survival [Brunet et al., 2004; Bailey et al., 2016]. This retrospective strategy can uncover biologically meaningful patterns, but it does not optimize the decomposition with respect to patient outcomes, often yielding factors dominated by non-prognostic variation such as tumor purity, stromal admixture, or batch effects [Aran et al., 2017; Thorsson et al., 2018]. Although supervised and discriminant variants of NMF have been explored [Tran et al., 2024], and some recent works have coupled factorization with survival analysis (e.g., Learning Individual Survival Models from PanCancer Whole Transcriptomes [Kumar et al., 2023]; CoxNTF [Fogel et al., 2025]), these approaches either treat survival as a downstream predictor or rely on tensor factorizations not tailored to high-dimensional gene expression data.

To address this gap, we introduce deSurv, a survival-driven deconvolution framework that integrates NMF with the Cox proportional hazards model [Cox, 1972]. deSurv directly incorporates survival information during factorization, producing interpretable, prognostic components while providing principled model selection criteria and regularization for high-dimensional stability [Tibshirani, 1997]. Implemented in a scalable pipeline for large cohorts, deSurv improves survival prediction relative to conventional unsupervised NMF while retaining interpretability. These results establish deSurv as a general framework for outcome-driven molecular subtyping across cancer types.

Results

A. The NMF–Cox framework provides an end-to-end workflow for prognostic modeling. We developed an integrated framework that combines nonnegative matrix factorization (NMF) with Cox proportional hazards regression to identify latent gene expression factors associated with survival. As illustrated

Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the Significance Statement is to explain the relevance of the work in broad context to a broad readership. The Significance Statement appears in the paper itself and is required for all research papers.

Please provide details of author contributions here.

Please declare any conflict of interest here.

[2] To whom correspondence should be addressed. E-mail: bob@email.com

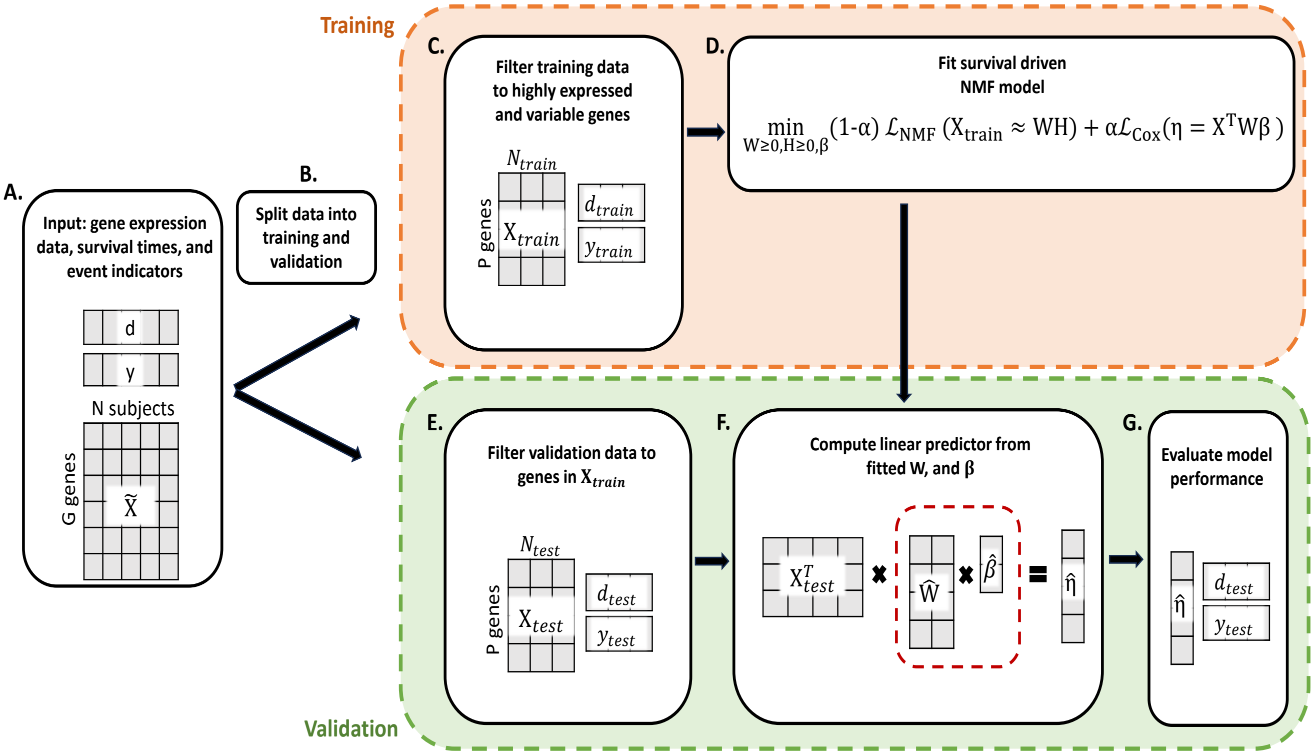
Training
A.
Input: gene expression data, survival times, and event indicators
d
y
N subjects
G genes
$\tilde{X}$
B.
Split data into training and validation
C.
Filter training data to highly expressed and variable genes
N_{train}
P genes
X_{train}
d_{train}
y_{train}
D.
Fit survival driven NMF model
$\min_{W\geq 0, H\geq 0, \beta} (1\text{-}\alpha)\, \mathcal{L}_{NMF}\, (X_{train} \approx WH) + \alpha \mathcal{L}_{Cox}(\eta = X^T W\beta)$
E.
Filter validation data to genes in X_{train}
N_{test}
P genes
X_{test}
d_{test}
y_{test}
F.
Compute linear predictor from fitted W, and β
X^T_{test}
$\hat{W}$
$\hat{\beta}$
$\hat{\eta}$
G.
Evaluate model performance
$\hat{\eta}$
d_{test}
y_{test}
Validation

Fig. 1. DeSurv overview

in Figure 1, the workflow begins with preprocessing and normalization of RNA-seq data, followed by NMF decomposition into patient factor loadings (W) and gene weightings (H). A Cox model is then fit using projected covariates derived from W. The framework incorporates a balancing parameter α to control the relative influence of reconstruction error versus survival likelihood. Model selection is performed via cross-validation across k, penalty parameters, and α, with downstream evaluation focusing on both predictive performance and biological interpretability.

B. The NMF–Cox model shows consistent convergence and stability across restarts. Across datasets and initialization schemes, the NMF–Cox algorithm consistently converged to numerically stable solutions within the designated iteration budget. Warm-start strategies, in which solutions at $\alpha = 0$ initialized supervised runs, substantially reduced variability across restarts and improved reproducibility. Figure 2 shows representative loss trajectories demonstrating monotone decreases until convergence, while Figure 3 summarizes performance variability across restarts. Compared with naïve random initialization, warm-starts produced tighter distributions of cross-validated C-index and partial likelihood, confirming stability of the optimization procedure.

C. Cross-validation of NMF–Cox identifies parameter settings that balance prediction and reconstruction. We evaluated performance across a grid of factor ranks (k), penalties, and values of α. Cross-validated C-index varied modestly across conditions, with no consistent improvement for $\alpha > 0$. Instead, supervised extensions altered the orientation of latent factors while maintaining comparable discrimination. Figure @ref(fig::fig-cv)A shows a heatmap of mean C-index across k and α, and @ref(fig::fig-cv)B illustrates C-index trends across α stratified by rank.

D. NMF–Cox uncovers biologically interpretable latent factors associated with clinical outcomes. Despite limited performance gains from supervision, the latent factors identified by NMF–Cox exhibited strong biological interpretability. The projected covariates, $W^T X$, aligned with known clinical and molecular subtypes, including basal-like versus classical subgroups in pancreatic cancer (Figure 7). Kaplan–Meier curves stratified by factor exposures revealed significant survival differences (Figure 8), supporting the prognostic relevance of the factors. At the gene level, W highlighted pathway-level enrichment for immune signaling, stromal activity, and hallmark oncogenic processes. Overlap analysis (Figure 9) demonstrated consistency with external signatures, confirming that NMF–Cox produces reproducible biological features.

E. NMF–Cox factors generalize to independent cohorts in external validation. To assess generalizability, models trained on TCGA-PAAD and CPTAC were applied to external cohorts including PACA, Moffitt, and Puleo. Factor exposures in validation datasets recapitulated subgroup structures identified in training and stratified patients into groups with distinct survival outcomes (Figure 10). Factor correlation analyses (Figure 11) confirmed reproducibility of core latent dimensions, particularly those separating basal-like and classical subtypes. Predictive accuracy in external cohorts was comparable to cross-validation results, with simpler models ($k \leq 5$) showing greater reproducibility. These findings indicate that NMF–Cox captures transferable biological signals across studies.

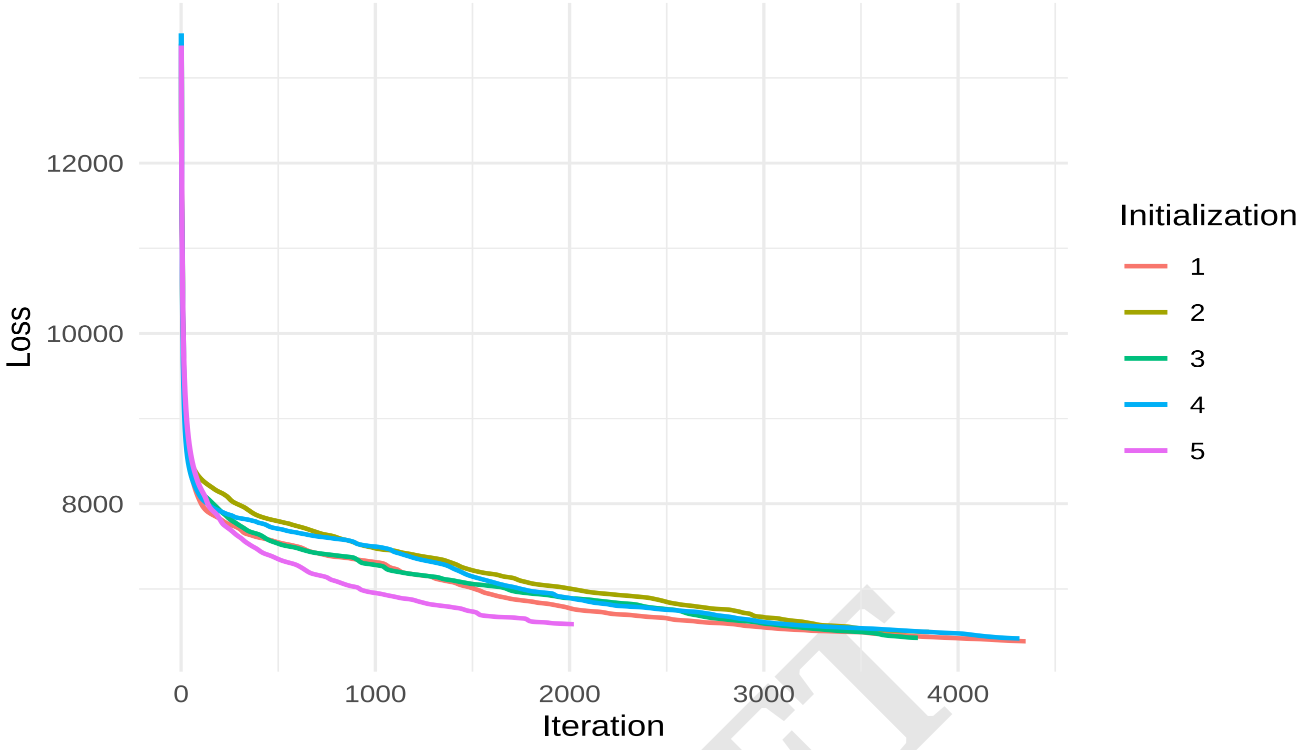

Fig. 2. Convergence of model for k=8 alpha=.5

Discussion

Materials and methods

F. Joint loss function. DeSurv is constructs a joint loss function of NMF reconstruction error and the cox partial likelihood. The contribution of each component to the overall loss is weighted by hyperparameter α.

$$\mathcal{L}(W,H,\beta) = (1-\alpha)\mathcal{L}(W,H)_{NMF} - \alpha\mathcal{L}(W,\beta)_{cox} \qquad [1]$$

F.1. Reconstruction error. Let X be a matrix of p features and n subjects. Then the NMF portion of the loss is

$$\mathcal{L}(W,H)_{NMF} = ||X - WH||_F^2 \qquad [2]$$

where $W \in R^{p\times k}$ is the matrix of feature weights and $H \in R^{k\times n}$ is the matrix of subject weights, where k represents the number of latent factors.

F.2. Cox partial likelihood. Let $y_i = \min(T_i, C_i)$ where T_i is the event time and C_i is the censoring time for the ith subject; let δ_i represent the indicator that the event time for the ith subject is observed. Since W is not dataset dependent, we take $Z = X^TW \in R^{n\times k}$ to be the covariates passed to the proportional hazards model. This can be interpreted as a transformation of the data matrix X into the lower dimensional space. The log partial likelihood is

$$\ell(W,\beta) = \sum_{i=1}^{n} \delta_i \left[Z_i^T\beta - \log\left(\sum_{j=1}^{n} \exp\left(Z_j^T\beta\right) \mathbb{1}(y_j \geq y_i) \right) \right] \qquad [3]$$

Update Rules. The joint loss function in (1) is nonconvex. Algorithm (1) provides an update alternates between updating W, H, and β until convergence.

```
Algorithm 1 DeSurv algorithm
Input: $X \in \mathbb{R}^{p\times n}_{\geq 0}$, $y \in \mathbb{R}^{n}_{\geq 0}$, $\delta \in \mathbb{R}^{n}_{0,1}$
$eps \leftarrow \infty$
$iter \leftarrow 0$
$W_{jr} \sim Unif(0, \max(X))$ for $j = 1,\dots,p$ and $r = 1,\dots,k$
$H_{ri} \sim Unif(0, \max(X))$ for $r = 1,\dots,k$ and $i = 1,\dots,n$
while $eps < tol$ and $iter < maxit$ do
    $W \leftarrow \text{argmin}_{W\geq 0} \mathcal{L}(W,H,\beta)$
    $H \leftarrow \text{argmin}_{H\geq 0} \mathcal{L}(W,H,\beta)$
    $\beta \leftarrow \text{argmin}_{\beta} \mathcal{L}(W,H,\beta)$
    $errNew \leftarrow \mathcal{L}(W,H,\beta)$
    $relErr \leftarrow |errNew - err|/err$
    $err \leftarrow errNew$
    $iter \leftarrow iter + 1$
return $W, H, \beta$
```

F.3. H update. Since the H update does not depend on β, a standard multiplicative update can be used

$$H_{ij} = H_{ij}\frac{(W^TX)_{ij}}{(W^TWH)_{ij}} \qquad [4]$$

F.4. β update. Holding W fixed, the *beta* update for covariates Z solves the standard convex penalized weighted least squares problem. The update as derived in (**?**) is

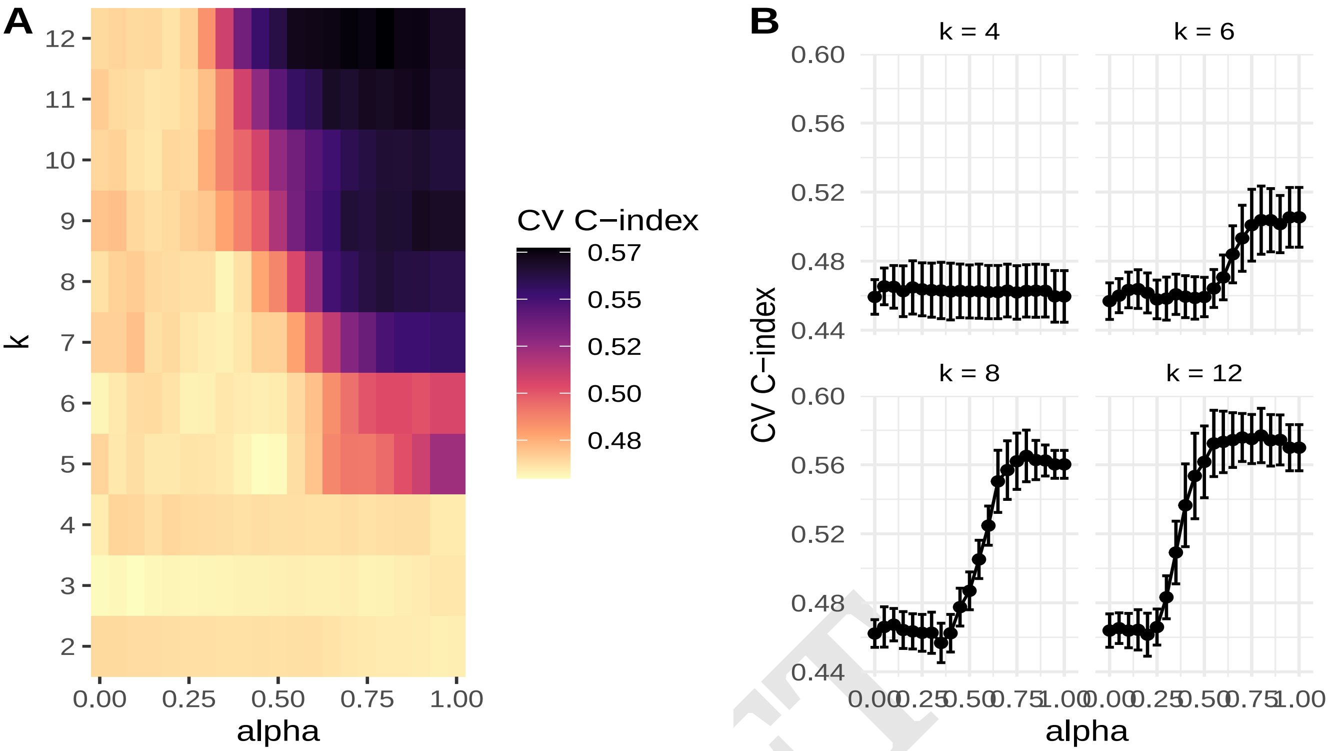

Fig. 3. A. Heatmap of cross-validated C-index across key parameters k and alpha. B. Cross validated C-index by alpha with standard error bars, stratified by k.

$$\hat{\beta}_r = \frac{S(\frac{1}{n}\sum_{i=1}^{n} w(\tilde{\eta})_i v_{i,r}\left[z(\tilde{\eta})_i - \sum_{j\neq r} v_{ij}\beta_j,\right], \lambda\xi)}{\frac{1}{n}\sum_{i=1}^{n} w(\tilde{\eta})_i v_{i,r}^2 + \lambda(1-\xi)} \qquad [5]$$

where describe wtilde and ztilde

G. Publicly Available Datasets. To train and validate our model we used seven publicly available PDAC datasets. RNAseq datasets were in TPM units. Each dataset was log2 + 1 transformed for variance stabilization. Additionally each dataset was rank transformed to mitigate scale differences between datasets and platforms. For each include sample size, citation, platform

- TCGA
- CPTAC
- Dijk
- Moffitt
- PACA array
- PACA seq
- Puleo

H. Model Training. The TCGA dataset was used for model training. The data was filtered to the top 1000 highly expressed and variable genes. Models were trained across a grid of hyperparameters $\alpha \in \{0, .95\}$, $\lambda \in$, $\xi \in$, $\gamma \in$, $\nu \in$, and $k = 2, \ldots, 15$

H.1. Hyperparameter selection. The hyperparameters α, λ, ξ, γ, and ν were selected to adequately balance the supervised and unsupervised portions of the model using a metric we defined as the c-index of the proportional hazards model divided by the reconstruction error. The parameters were chosen to maximize this metric. Since the reconstruction error exclusively decreases as the dimension k increases, this metric was not adequate to choose k.

Cross-validation was used to select the optimal hyperparameters. To account for lack of uniqueness of NMF solutions, we used 50 random initializations of W,H, and beta per fold. Validation metrics were averaged across fold and initialization.

I. Model Validation. The remaining 7 publicly available PDAC datasets, CPTAC, Dijk, Linehan, Moffitt, PACA microarray, PACA RNAseq, and Puleo, were used to validate our models. The datasets were restricted to the same highly variable genes in the TCGA dataset, and the fitted model was applied to each dataset individually. The partial likelihood and c-index were calculated for each dataset. Hazard ratios were reported for each factor.

J. Score based approach. In a clinical setting, it may not be feasible to sequence all of the genes required to port the full W matrix to future test sets. For this purpose we also propose a score based method, where only the identity of top genes for each factor must be ported to future datasets. To construct the scores we define Wtilde as a binary matrix... To predict patient outcomes in future datasets, we restrict those datasets to these g x k genes, and compute XtWtilde as linear predictor in the survival model.

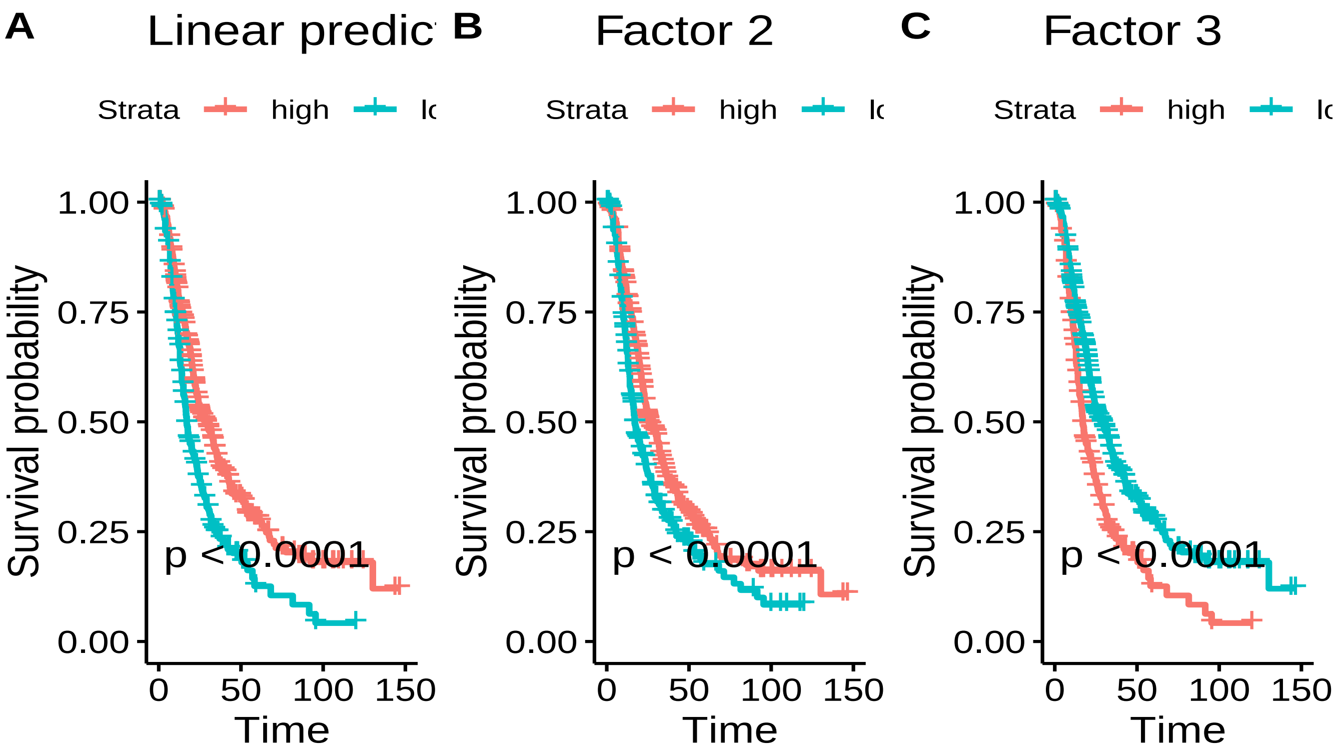

Fig. 4. Kaplan Meier curves for a median split on A. linear predictor, B. Factor 2, C. Factor 3

J.1. Top genes. The top genes were extracted from each factor of W in the selected model at each value of k. A top gene was defined as . . .

ACKNOWLEDGMENTS. Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript.