Survival driven deconvolution (DeSurv) reveals prognostic and interpretable cancer subtypes

Amber M. Young[a,1,2], **Alisa Yurovsky**[b], **Didong Li**[a], **and Naim U. Rashid**[a,c]

[a]University of North Carolina at Chapel Hill, Biostatistics, Street, City, State, Zip; [b]Stony Brook University, Street, City, State, Zip

This manuscript was compiled on January 22, 2026

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 cancer, we uncover prognostic and biologically interpretable subtypes that reflect the complex interactions between stroma, tumor, and immune cells in the tumor microenvironment. Our approach highlights the value of using patient outcomes during gene signature discovery.

one | two | optional | optional | optional

Molecular subtyping has transformed precision oncology by stratifying patients into biologically and clinically meaningful groups that inform prognosis and guide therapy (1–5). Subtyping relies on the identification of robust biological signals that define subtypes such as transcriptomic signatures. However, the tumor microenvironment (TME) contains mixtures of diverse cell types such as malignant, stromal, immune, and endothelial cells, and disentangling tumor specific signals from this mixture can be challenging. As such, subtyping pipelines typically rely on the deconvolution of bulk transcriptomic data or single-cell analysis to discover distinct cell types and their corresponding signatures. Downstream, the signatures are evaluated for clinical relevance such as overall survival or response to treatment.

Separating discovery from validation can risk overfitting and limits biological and clinical generalizability. Identified cell types may capture dataset-specific noise rather than reproducible biological signals, undermining their utility in downstream analyses or therapeutic targeting (6, 7). Moreover, even when discovered cell types are biologically valid and reproducible, they may not correspond to the cellular programs most relevant for predicting or influencing clinical outcomes (8, 9). Therefore, there is a clear need for integrative methods that jointly uncover biologically meaningful programs while directly incorporating clinical endpoints to ensure prognostic relevance.

However, integrating patient outcomes into the discovery phase is not straightforward with current technology and methodology. Single-cell transcriptomics can resolve programs at the cellular level, but cohort sizes are often too small to support survival analyses. In contrast, large bulk transcriptomic cohorts with clinical annotations are well-suited for outcome modeling (10, 11), yet deconvolution is needed to disentangle overlapping cellular signals. Reference-based deconvolution methods focus on estimating cell-type proportions from predefined signatures, which limits the utility of these methods for discovery of novel programs (12).

Nonnegative matrix factorization (NMF) is widely used in cancer genomics because its nonnegativity constraints produce biologically interpretable, additive molecular programs (13–16). Although recent extensions have incorporated supervision into the factorization, most target regression or classification rather than time-to-event outcomes. Two studies have proposed survival-aware NMF formulations (17, 18), but both integrate the survival objective through the sample-specific loadings rather than the gene-level programs. This design emphasizes prediction accuracy but limits the model's ability to restructure or refine the underlying molecular programs, reducing its value for biological interpretation and subtype discovery, which are core objectives in cancer transcriptomics. In addition, neither study provides a principled approach for hyperparameter selection or model assessment, and convergence properties are only briefly addressed in one manuscript. Both works remain unpublished and unreviewed, leaving their methodological robustness and reproducibility uncertain. These gaps highlight the need for a rigorously formulated, survival-aware deconvolution method that jointly estimates interpretable molecular programs and their prognostic relevance.

Here we present DeSurv, a Survival-supervised Deconvolution framework that integrates non-negative matrix factor-

Significance Statement

Tumor transcriptomes reflect mixtures of malignant and microenvironmental cell populations, making it challenging to identify the molecular programs that truly drive clinical outcomes. Existing deconvolution and matrix factorization methods discover latent transcriptional programs but do not ensure that these programs are prognostic, while supervised extensions optimized for prediction offer limited biological interpretability. We present DeSurv, a survival-supervised deconvolution framework that integrates nonnegative matrix factorization with Cox modeling to jointly learn biologically coherent gene programs and their associations with patient survival. By embedding outcome information directly into the discovery process and performing automatic model selection, DeSurv reveals clinically relevant transcriptional programs that are reproducible across cohorts. This approach advances the statistical foundations of tumor deconvolution and provides a general tool for identifying actionable molecular drivers of disease progression.

Please provide details of author contributions here.

Please declare any conflict of interest here.

[2] To whom correspondence should be addressed. E-mail: ayoung31@live.unc.edu

ization (NMF) with Cox proportional hazards modelling. In contrast to fully unsupervised approaches that evaluate survival associations only after the factorization, and to existing supervised NMF models that link outcomes to the subject-level factor loadings, DeSurv integrates survival information directly into the gene signature matrix. This design ensures that the discovered transcriptional programs are not only biologically interpretable but also intrinsically aligned with patient outcomes. To enhance robustness and reproducibility, DeSurv performs automatic parameter selection via Bayesian optimization, addressing the quintessential challenge of rank determination in matrix factorization.

By coupling latent program discovery with direct survival supervision, DeSurv resolves longstanding challenges in disentangling tumor–microenvironment interactions and aligns molecular heterogeneity with clinical outcomes. This unified approach represents a methodological advance in translational cancer genomics and provides a general framework for deriving actionable insights from high-dimensional transcriptomic data.

Results

Model Overview. We have developed an integrated framework, DeSurv, that couples Nonnegative Matrix Factorization (NMF) with Cox proportional hazards regression to identify latent gene-expression programs associated with patient survival (Figure @ref(fig:fig-schema)). The model takes as input a bulk expression matrix of p genes by n patients (X_{Train}) together with corresponding survival times (y_{Train}) and censoring indicators (δ_{Train}) (Figure 1A).

DeSurv optimizes a joint objective combining the NMF reconstruction loss and the Cox model's log-partial likelihood, weighted by a supervision parameter (α) that determines the relative contribution of each term (fig. 1B):

$$(1-\alpha)\ \mathcal{L}_{NMF}(X_{Train} \approx WH) \\ -\alpha\ \mathcal{L}_{Cox}(X_{Train}^T W\beta, y_{Train}, \delta_{Train}) \quad [1]$$

When $\alpha = 0$, the method reduces to standard unsupervised NMF; when $\alpha > 0$, survival information directly guides the learned factors toward prognostic structure.

Within this framework, the product ($X_{Train}^T W$) represents patient-level factor scores - the inferred burden of each latent program across subjects. These factor scores serve as covariates in the Cox model, and their regression coefficients (β) indicate whether higher activity of a given program corresponds to improved or reduced survival.

Model training yields gene weights ($\hat{W}$), factor loadings ($\hat{H}$), and Cox coefficients ($\hat{\beta}$) (Figure 1C), where the inner dimension (k) specifies the number of latent factors. Genes with high gene weights in one factor and low gene weights in all others define the factor-specific signature genes (Figure 1D). By integrating survival supervision into the factorization, DeSurv not only reconstructs the underlying expression structure, preserving biologial interpretability, but also guides latent factors to be prognostically informative. Subsequent analyses can therefore focus on the survival-associated gene programs (Figure 1E).

Bayesian optimisation selects the DeSurv hyperparameters (k, α).

A. Outcome-guided model selection resolves ambiguity in NMF rank choice. We examined the problem of selecting the number of latent components (k) in nonnegative matrix factorization (NMF) using gene expression data from pancreatic ductal adenocarcinoma (PDAC) cohorts. These heterogeneous tumor transcriptomes provide a representative setting in which to evaluate how commonly used unsupervised rank-selection heuristics behave in practice.

Across a range of candidate ranks, standard NMF diagnostics yielded inconsistent guidance (Fig. **??**A-C). Reconstruction residuals decreased smoothly with increasing k and did not exhibit a clear elbow, a pattern consistent with both relatively small solutions ($k \approx$ 3-4) and substantially larger ranks ($k \approx$ 6-8). The cophenetic correlation coefficient began to decline at low ranks ($k \approx$ 3-4) but continued to fluctuate at higher values without a distinct transition point. In contrast, mean silhouette width, evaluated across multiple distance metrics, was highest at very small ranks ($k \approx$ 2-3) and decreased monotonically thereafter, favoring low-dimensional solutions that conflicted with recommendations based on reconstruction error or cophenetic correlation. Together, these unsupervised criteria pointed to different and incompatible values of k, highlighting the ambiguity of rank selection in standard NMF when applied to PDAC data.

To resolve this ambiguity, we applied DeSurv, which incorporates survival outcomes directly into the factorization process and evaluates models using survival-based predictive performance. Using the same PDAC gene expression data, we assessed model performance across the joint space of the number of components (k) and supervision strength (α) using cross-validated concordance index (C-index). The resulting C-index surface summarizes expected predictive performance across candidate models and enables direct comparison of solutions that differ in both model complexity and degree of supervision (Fig. **??**D).

Model selection was based on standard cross-validation principles. Rather than selecting the single parameter combination with the highest predicted C-index, we selected the smallest value of k whose predicted performance lay within one standard error of the maximum. This criterion yielded a stable and parsimonious choice of model rank in the PDAC data, in contrast to the conflicting recommendations produced by unsupervised NMF heuristics.

To further evaluate rank recovery under controlled conditions, we conducted simulation studies in which the true underlying rank was known ($k = 3$). Across repeated simulation replicates, DeSurv consistently selected the correct rank, producing a concentrated distribution of selected k values centered at the true value. In contrast, standard NMF followed by post hoc Cox modeling ($\alpha = 0$) exhibited substantially greater variability and a systematic tendency toward under-selection. Together, these results indicate that incorporating outcome information during model fitting improves the reliability of rank selection in settings where unsupervised criteria yield conflicting conclusions.

"'{r fig-bo, fig.width = 6, fig.height= 5.5, fig.cap = "(A-D) Analyses based on real pancreatic ductal adenocarcinoma (PDAC) gene expression data illustrate the ambiguity of rank selection in standard nonnegative matrix factorization (NMF) and the use of outcome supervision in DeSurv. (A-C) Commonly used unsupervised heuristics for selecting the number

of components (k) yield inconsistent conclusions. (A) Reconstruction residuals decrease smoothly with increasing k and do not exhibit a clear elbow; diminishing returns could be inferred at intermediate (k
approx 3-4) or larger (k
approx 6-8) ranks. (B) The cophenetic correlation coefficient, often used to select the largest k prior to a marked loss of clustering stability, begins to decline at low ranks (k
approx 3-4) but continues to fluctuate thereafter, providing no unambiguous selection criterion. (C) Mean silhouette width across multiple distance metrics is highest at small ranks (k
approx 2-3) and decreases monotonically with increasing k, favoring lower-dimensional solutions that conflict with the other criteria. (D) Heatmap of the Gaussian process predicted mean cross-validated concordance index (C-index) from Bayesian optimization over the joint space of the number of components (k) and supervision strength (
alpha), computed on the same PDAC data. The predicted performance surface summarizes survival prediction accuracy across parameter settings and illustrates how DeSurv uses outcome information to inform model selection. (E) Results from simulation studies with a known underlying rank (k = 3) showing the distribution of selected k values across repeated replicates. DeSurv more consistently recovers the true rank, yielding a concentrated distribution centered at k = 3, whereas standard NMF with post hoc Cox modeling (
alpha = 0) exhibits greater variability and a tendency toward under-selection. \label{fig:bo}", fig.env='figure*', fig.pos='t', out.height= "5.5in", out.width="6in"}
tar_load(fig_bo_heat_tcgacptac)
tar_load(fig_residuals_tcgacptac)
tar_load(fig_cophenetic_tcgacptac)
tar_load(fig_silhouette_tcgacptac)

1. upper = plot_grid(fig_bo_cvk_tcgacptac,fig_bo_cvalpha_tcgacptac,labels=c("A","B"))

upper = plot_grid(fig_residuals_tcgacptac, fig_cophenetic_tcgacptac, fig_silhouette_tcgacptac, ncol = 3, labels = c("A","B","C"))
k_hist = plot_grid(alt_plots$k_hist,NULL,nrow=2,rel_heights = c(4,1))
lower = plot_grid(fig_bo_heat_tcgacptac, k_hist,ncol=2,labels=c("D","E"),rel_widths = c(3,2))
plot_grid(upper,lower,nrow=2,rel_heights = c(2,3))

DeSurv improves selection of prognostic gene signatures {.unnumbered}
To test whether supervision improves the selection of prognostic gene signatures, we used simulations with known lethal
\begin{figure*}[t]

{\centering \includegraphics[width=\textwidth]{/work/users/a/y/ayucong23/DeSurv_paper/paper/paper_files/figure-latex/fig

}

\caption{Performance comparison between DeSurv and unsupervised NMF ($\alpha = 0$) in simulation. (A) Distribution of c
\end{figure*}

Survival-informed factorization reorganizes transcriptional structure in pancreatic cancer {.unnumbered}
To assess the biological structure captured by standard nonnegative matrix factorization (NMF) and DeSurv, we examined

In the standard NMF solution, factors largely reflected dominant

By contrast, DeSurv produced a factorization that emphasized axe

These differences were reflected quantitatively by contrasting t

To further characterize how DeSurv reorganizes the transcription
\begin{figure*}[t]

{\centering \includegraphics[width=\textwidth,height=4.5in]{/wor

}

\caption{Survival-informed factorization reorganizes transcripti
\end{figure*}

DeSurv-derived latent structure generalizes to independent da
To assess generalization of DeSurv latent factors, the gene-leve

For each method, we focused on the factor showing the largest in

When validation samples were pooled and stratified into high- an

Together, these results indicate that DeSurv identifies latent f
\begin{figure*}[t]

{\centering \includegraphics[width=\textwidth]{/work/users/a/y/a

}

\caption{Figure X. DeSurv learns prognostic structure that gener
\end{figure*}

Bladder cancer analysis {.unnumbered}
We will report a focused bladder cancer analysis here, including

Discussion {.unnumbered}

We present DeSurv, a survival-driven deconvolution framework tha

In pancreatic ductal adenocarcinoma (PDAC), DeSurv identified th
Importantly, DeSurv outperformed unsupervised NMF in cross-valid
Beyond PDAC, DeSurv provides a generalizable framework for disc

Materials and methods {.unnumbered}

Problem formulation and notation
Let $X \in \mathbb{R}^{p \times n}_{\ge 0}$ denote the nonnegati

The DeSurv Model
DeSurv integrates Nonnegative Matrix Factorization (NMF) with pe

The joint objective is
\begin{equation}
\label{eqn:desurv}
\mathcal{L}(W,H,\beta) =
(1-\alpha)\,\mathcal{L}_{\mathrm{NMF}}(W,H)
- \alpha\,\mathcal{L}_{\mathrm{Cox}}(W,\beta),
\end{equation}
where $\mathcal{L}_{\mathrm{NMF}}(W,H)$ is the NMF reconstruction error and $\mathcal{L}_{\mathrm{Cox}}(W,\beta)$ is the negative Cox partial log-likelihood.

Hyperparameter selection and cross-validation
Hyperparameters $(k,\alpha,\lambda_H,\lambda,\xi)$ were selected by maximizing the cross-validated C-index.

Simulation studies
Simulation studies were conducted to assess recovery of prognostic latent structure and survival prediction.

Survival times were generated from an exponential proportional hazards model in which risk depended on marker gene str

Evaluation metrics

Real-world datasets
We analyzed publicly available RNA-seq and microarray cohorts of pancreatic ductal adenocarcinoma (PDAC) and bladder c

Simulations, Benchmarking, and Availability
Code and processed data used in this study are available at [repository link].

<!-- Leave these lines as they are at the end of your .Rmd file to ensure placement of methods & acknowledgements sect
\showmatmethods
\showacknow

\subsection{Model details}

Let $X \in \mathbb{R}_{\ge 0}^{p \times n}$ denote the nonnegat
expression matrix with p genes and n subjects. DeSurv appro
$X \approx WH$ with $W \in \mathbb{R}_{\ge 0}^{p \times k}$ and
$H \in \mathbb{R}_{\ge 0}^{k \times n}$, and links the shared p
W to survival via an elastic-net-penalized Cox model.

Let $y \in \mathbb{R}^n_{\ge 0}$ be the observed survival times
$\delta \in \{0,1\}^n$ the event indicators, and
$Z = W^\top \in \mathbb{R}^{n \times k}$ the sample-wise prog
loadings. We write $Z_i^\top$ for the ith row of Z, and
$n_{\text{event}} = \sum_{i=1}^n \delta_i$. For convenience we
Subtracting the constant $\frac{1}{2n}\lVert X \rVert_F^2$ and
\begin{align}
  \mathcal{L}(W,H,\beta) &=\nonumber \\
  &(1-\alpha)\left(\frac{1}{2np}\lVert X - WH \rVert_F^2 + \fra
  &\alpha\left(\frac{2}{n_{\text{event}}}\ell(W,\beta) + \lambd
\end{align}
where $\lVert \cdot \rVert_F$ is the Frobenius norm and
$\beta \in \mathbb{R}^k$ are the Cox coefficients. The Cox log-
likelihood $\ell(W,\beta)$ is
\begin{equation}
\ell(W,\beta)
=
\sum_{i=1}^n
\delta_i
\left[
  Z_i^\top \beta
  -
  \log \left(
    \sum_{j
    \exp(Z_j^\top \beta)
  \right)
\right].
\end{equation}
Hyperparameters satisfy $\alpha \in [0,1)$,
$\lambda_H \ge 0$, $\lambda \ge 0$, and $\xi \in [0,1]$.
The constants $1/(2np)$, $2/n_{\text{event}}$, and $1/(2nk)$ ar
for numerical convenience and do not affect the minimizer.

\subsection{Optimization Algorithm}

\subsubsection{Block coordinate descent scheme}
Algorithm~\ref{alg:desurv} summarizes the block coordinate desc
scheme used to minimize $\mathcal{L}(W,H,\beta)$. At each outer
we update H, then W, then β while holding the other b
fixed.

\begin{algorithm}[H]
\caption{DeSurv block coordinate descent}
\label{alg:desurv}
    \begin{algorithmic}[1]
        \Require $X \in \mathbb{R}_{\geq 0}^{p \times n}$, surv
        \Ensure Fitted DeSurv parameters (W,H,β)

        \State Initialize $W^{(0)}, H^{(0)}$ with positive entr
        \State Initialize $\beta^{(0)}$ (e.g., $\beta^{(0)} = \
        \State $loss = \mathcal{L}(W^{(0)},H^{(0)},\beta^{(0)})
        \State $eps = \infty$, $t = 0$
        \While{$eps \geq tol$ \textbf{and} $t < maxit$}

```
            \State \textbf{(H-update)}
            \State \hspace{0.5cm} Update $H^{(t+1)}$ from $H^{(t)}$ using the multiplicative rule in Eq.~\ref{eqn:Hupda
            \State \hspace{0.5cm} holding $W^{(t)}$ and $\beta^{(t)}$ fixed.
            \State \textbf{(W-update)}
            \State \hspace{0.5cm} Update $W^{(t+1)}$ from $W^{(t)}$ using the hybrid multiplicative rule in Eq.~\ref{eq
            \State \hspace{0.5cm} holding $H^{(t+1)}$ and $\beta^{(t)}$ fixed, and perform backtracking to ensure
            \State \hspace{0.5cm} $\mathcal{L}$ does not increase.
            \State \textbf{($\beta$-update)}
            \State \hspace{0.5cm} Update $\beta^{(t+1)}$ from $\beta^{(t)}$ using a Newton-like step for the Cox loss
            \State \hspace{0.5cm} (Eq.~\ref{eqn:beta_update}), holding $W^{(t+1)}$ and $H^{(t+1)}$ fixed.
            \State $lossNew = \mathcal{L}(W^{(t+1)},H^{(t+1)},\beta^{(t+1)})$
            \State $eps = |lossNew - loss| / |loss|$
            \State $loss = lossNew$
            \State $t = t + 1$
        \EndWhile
        \State \Return $\hat{W} = W^{(t+1)}, \hat{H} = H^{(t+1)}, \hat{\beta}=\beta^{(t+1)}$
    \end{algorithmic}
\end{algorithm}

\subsubsection{Update for $H$}
Conditional on $W$, the loss $\mathcal{L}(W,H,\beta)$ reduces to a strictly
convex quadratic function of $H$. We adopt the standard NMF multiplicative
update with $\ell_2$ penalty:
\begin{equation}
\label{eqn:Hupdate}
H \leftarrow
\max\!\left(
  H \odot \frac{W^\top X}{
    W^\top W H + \lambda_H H + \varepsilon_H
  },
  \varepsilon_H
\right),
\end{equation}
where $\odot$ denotes elementwise multiplication, all divisions are
elementwise, and $\varepsilon_H > 0$ is a small floor to prevent entries from
becoming exactly zero. This update is equivalent to a majorization-minimization
step and guarantees a nonincreasing reconstruction term conditional on $W$
[@seung2001algorithms; @pascualmontano2006nonsmooth]. The floor $\varepsilon_H$
ensures that iterates remain in the interior of the nonnegative orthant, which
simplifies the convergence analysis.


\subsubsection{Update for $W$}
For $W$, we construct a hybrid multiplicative update that balances the
contributions of the NMF and Cox gradients. Let
\[
\nabla_W \mathcal{L}_{\mathrm{NMF}}(W,H)
=
\frac{1}{np}(W H - X)H^\top,
\]
and let
\[
\nabla_W \mathcal{L}_{\mathrm{Cox}}(W,\beta)
=
\frac{2}{n_{event}}\nabla_W \ell(W,\beta).
\]
where $\nabla_W \ell(W,\beta)$ denotes the Cox gradient with respect to $W$. A closed-form expression for $\nabla_W \ell$ is
\begin{equation}
\nabla_W \ell(W,\beta)
=
\sum_{i=1}^n
```

```
\delta_i
\left(
  x_i
  -\frac{
    \sum_{j : y_j \ge y_i}
    x_j \exp(x_j^\top W \beta)
  }{
    \sum_{j : y_j \ge y_i}
    \exp(x_j^\top W \beta)
  }
\right)\beta^\top,
\end{equation}
where $x_i$ denotes the $i$th column of $X$.

We define the gradient-balancing factor
\begin{equation}
\label{eq:supp-delta}
\delta^{(t)}
=
\frac{
  \left\lVert (1-\alpha) \nabla_W \mathcal{L}_{\mathrm{NMF}}(W^{
}{
  \left\lVert \alpha \nabla_W \mathcal{L}_{Cox}(W^{(t)},\beta^{(
},
\end{equation}
and clip $\delta^{(t)}$ to avoid extreme values:
$\delta^{(t)} \leftarrow \min(\delta^{(t)}, \delta_{\max})$
with $\delta_{\max} = 10^6$ in all experiments.

The multiplicative factor for $W$ is then
\begin{equation}
R^{(t)}
=
\frac{
  (1-\alpha) X H^{(t+1)\top}
  +\frac{2\alpha}{n_{\mathrm{event}}}\,\delta^{(t)}\,\nabla_W \e
}{
  (1-\alpha) W^{(t)} H^{(t+1)}H^{(t+1)\top}
  + \varepsilon_W
},
\end{equation}
and the proposed update is
\begin{equation}
\label{eqn:Wupdate}
W^{(t+1)}
=
\max\!\left(
  W^{(t)} \odot R^{(t)},
  \varepsilon_W
\right),
\end{equation}
with a small floor $\varepsilon_W > 0$. When $\alpha=0$, this re
to the standard multiplicative update for NMF [@seung2001algorit
for $\alpha>0$, the Cox gradient perturbs the update in a direct
decreases the supervised loss.

Because the combined multiplicative step does not, by itself, gu
nonincrease in the full loss $\mathcal{L}$, we embed it in a bac
line search.

\begin{algorithm}[H]
```

```
\caption{$W$ update with backtracking}
\label{alg:backtrack}
    \begin{algorithmic}[1]
        \Require Value of $W$ and the previous iteration $t
        \Ensure $W^{(t+1)}$
        \State $\theta =1$
        \State b = 1
        \State $flag\_accept = FALSE$
        \While{$b \leq max\_bt$}
            \State $W^{(t+1)}_{cand} = W^{(t)} \odot [(\theta
            \State \textbf{(Column normalization)}
            \State \hspace{0.5cm} Compute $D = \mathrm{
            \State \hspace{0.5cm} and set $W^{(t+1)}_{c
            \If{$\mathcal{L}(W^{(t+1)}_{cand},H^{(t+1)}
              \State $W^{(t+1)} = W^{(t+1)}_{cand}$
              \State $H^{(t+1)} = H^{(t+1)}_{cand}$
              \State $\beta^{(t)} = \beta^{(t)}_{cand}$
              \State $flag\_accept = TRUE$
              \State break
            \EndIf
            \State $\theta = \theta*\rho$
            \State $b = b + 1$
        \EndWhile
        \If{$flag\_accept = FALSE$}
            \State $W^{(t+1)}=W^{(t)}$
        \EndIf
    \end{algorithmic}
\end{algorithm}

The column normalization preserves both $WH$ and $W\beta$, and therefore
leaves the loss $\mathcal{L}$ invariant up to numerical
guarantees that the accepted $W$ update does not increa

\subsubsection{Update for $\beta$}

Conditional on $(W,H)$, the loss in $\beta$ reduces to a co
elastic-net-penalized Cox problem:
\[
\min_{\beta \in \mathbb{R}^k}
\left\{
  -\frac{2\alpha}{n_{\mathrm{event}}}\,\ell(W,\beta)
  +
  \lambda\left(
    \xi \lVert \beta \rVert_1
    +\frac{1-\xi}{2} \lVert \beta \rVert_2^2
  \right)
\right\}.
\]
We solve this subproblem by cyclic coordinate descent fo
[@simon2011regularization]. Writing $\ell(\beta) = \ell
$\tilde{\eta}_i = Z_i^\top \beta$, the update for coordi
closed form
\begin{equation}
\label{eqn:beta_update}
\hat{\beta}_r
=
\frac{
  S\!\left(
    \dfrac{1}{n}
    \sum_{i=1}^n w(\tilde{\eta})_i
    z_{i,r}
```

```
    \left[
      v(\tilde{\eta})_i - \sum_{j \ne r} z_{i,j} \beta_j
    \right],
  \lambda(t)}$, the multiplicative update term at the current
  \right)
}{
  \frac{1}{n} \sum_{i=1}^n w(\tilde{\eta})_i z_{i,r}^2 + \lambd
},
\end{equation}
where $S(a,\tau) = \mathrm{sign}(a)\max(|a|-\tau,0)$ is the sof
operator, and $w(\tilde{\eta})$, $v(\tilde{\eta})$ are standard
quadratic approximation of the Cox log-partial likelihood
We iterate coordinate updates until the subproblem converges
tolerance, yielding a minimizer in $\beta$ for the current $W$}
\subsubsection{Normalization of W}

After each accepted $W$ update, we normalize the columns of $W$
$W \leftarrow W D^{-1}$, and adjust $H$ and $\beta$ as
\[
H \leftarrow D H, \qquad \beta \leftarrow D \beta,
\]
where $D$ is the diagonal matrix of column $\ell_2$-norms of $W
This preserves the reconstruction $WH$ and the linear predictor
leaving both the NMF and Cox terms in the loss unchanged. Norma
prevents degeneracy in the scale-nonidentifiable factorization
columns of $W$ comparable in magnitude and interpretable as gen

\subsubsection{Derivation of W update from projected coordinate
The $W$ update can be derived directly from the projected coordin

Recall that the overall loss function is
\begin{align}
  \mathcal{L}(W,H,\beta) = \frac{(1-\alpha)}{2np} ||X - WH||^
\end{align}
where $\ell(W,\beta)$ is the log-partial likelihood for the Cox

\begin{equation}
  \frac{\partial L}{\partial W} = \frac{(1-\alpha)}{np}\left( W
\end{equation}

Then the gradient descent update rule at iteration $t$ is
\begin{align}
  W^{(t)} &= W^{(t-1)} - \gamma \left( \frac{\partial L}{\parti
          &= W^{(t-1)} - \gamma \left( \frac{(1-\alpha)}{np}\le
\end{align}
Let the step size $\gamma$ be defined as in CITE:
\begin{equation}
  \gamma = \frac{np}{(1-\alpha)}\frac{W}{WHH^T}
\end{equation}
Then the update becomes
\begin{align}
  W^{(t)} &= W^{(t-1)} -\frac{np}{(1-\alpha)}\frac{W}{WHH^T} \l
          &= \frac{W}{WHH^T} XH^T + \frac{2\alpha np}{n_{event}
          &= W \odot \frac{XH^T + \frac{2\alpha np}{n_{event}(1
          &= W \odot \frac{\frac{(1-\alpha)}{np} XH^T + \frac{2
\end{align}
Finally, projected coordinate descent projects the $W$ update i
\begin{equation}
```

```
  W^{(t)} =  \max \left(W \odot \frac{\frac{(1-\alpha)}{np} XH^T]+ \frac{2\alpha}{n_{event}} \nabla_W \ell}{\frac{(1-\a
\end{equation}                                                   If no candidate satisfies the Armijo-type condition with
Since $W \geq 0$ this is equivalent to                           number of backtracking steps, the algorithm sets
\begin{equation}                                                 $W^{(t+1)} := W^{(t)}$.
  W^{(t)} =  W \odot \max \left( \frac{\frac{(1-\alpha)}{np}XH\emph{$\beta$-update.} For fixed $(W^{(t+1)},H^{(t+1)})$
\end{equation}                                                   $\beta^{(t)} \mapsto \beta^{(t+1)}$ is obtained by runni
which matches the multiplicative form used in the software implementation.descent method of [@simon2011regularization]
                                                                 subproblem until convergence. Thus
                                                                 \[
\subsection{Convergence proof}                                     \beta^{(t+1)} \in
                                                                   \arg\min_{\beta\in\mathbb{R}^k}
We show that, under mild regularity conditions, the block coordinate descent(W^{(t+1)},H^{(t+1)},\beta),
(BCD) Algorithm~S\ref{alg:desurv} converges to a stationary point of the DeSurv
loss function                                                    and
\[                                                               \[
\mathcal{L}(W,H,\beta).                                            \mathcal{L}(W^{(t+1)},H^{(t+1)},\beta^{(t+1)})
\]                                                                 \le
For clarity, we first analyze the algorithm \emph{without} the column\mathcal{L}(W^{(t+1)},H^{(t+1)},\beta^{(t)}).
normalization of $W$ inside the loop, and then argue in          \]
Section~\ref{subsec:conv_normalization} that the normalization step preserves
stationarity of limit points.                                \end{enumerate}
                                                         \vspace{0.5em}
Throughout, let $\theta = (W,H,\beta)$ and denote        \textbf{Lemma 1 (Continuity and bounded level sets).}
\(\mathcal{L}(\theta) = \mathcal{L}(W,H,\beta)\).        Under Assumption 1, $\mathcal{L}:\Theta\to\mathbb{R}$ is continu
The feasible set is                                      initial sublevel set
\[                                                       \[
\Theta = \{(W,H,\beta):                                    \mathcal{S}_0 := \{\theta \in \Theta:
  W \in \mathbb{R}_{\ge 0}^{p\times k},\;                    \mathcal{L}(\theta) \le \mathcal{L}(\theta^{(0)})\}
  H \in \mathbb{R}_{\ge 0}^{k\times n},\;                \]
  \beta \in \mathbb{R}^k\}.                              is nonempty, closed, and bounded.
\]
                                                         \emph{Proof.}
\textbf{Assumption 1 (Regularity and parameter space).}The NMF term $\|X-WH\|_F^2$ is a polynomial in the entries of $W
We assume:                                               is continuously differentiable. The penalty $\|H\|_F^2$ is conti
\begin{enumerate}[label=(\roman*)]                       The Cox partial log-likelihood is a smooth function of the linea
  \item The data $X \in \mathbb{R}_{\ge 0}^{p\times n}$$\y_i\in\mathbb{R}^n$, and $\delta_i\in\{0,1\}^n$ are fixed and
  \item The hyperparameters satisfy $\lambda_H > 0$, $\lambda\ge0$, $\alpha\in[0,1)$, and $\xi\in [0,1]$. The $\ell_2$ penalty
  \item The initial iterate $\theta^{(0)} = (W^{(0)},H^{(0)},\beta^{(0)})$ lies in $\Theta$ and ... $W^{(0)},H^{(
\end{enumerate}                                          $\mathcal{L}$ is continuous on $\Theta$.

\textbf{Assumption 2 (Block updates).}                   The constraints $W,H\ge 0$ define closed convex cones, and
At each outer iteration $t$, the three block updates in$\beta\in\mathbb{R}^k$ is unconstrained. Because $\lambda_H>0$ a
Algorithm~S\ref{alg:desurv} satisfy:                     the terms $\tfrac{\lambda_H}{nk}\|H\|_F^2$ and
\begin{enumerate}[label=(\roman*)]                       $\lambda\tfrac{1-\xi}{2}\|\beta\|_2^2$ dominate the objective as
  \item \emph{$H$-update.} For fixed $(W^{(t)},\beta^{(t)})$, the update $\|H\|_F\to\infty$ or $\|\beta\|_2\to\infty$, respectively. Thus
        $H^{(t)} \mapsto H^{(t+1)}$ is given by the multiplicative$\mathcal{S}_0$ is bounded in $(H,\beta)$. Since $W\ge 0$ an
        \[                                               fixed, the reconstruction term and Cox term being bounded preven
        H \leftarrow H \odot \frac{W^\top X}{W^\top W H + \lambda_H H},$\mathcal{L}$ remains below $\mathcal{L}(\theta^
        \]                                               $W$ is bounded on $\mathcal{S}_0$. Because $\Theta$ is closed an
        applied at $(W^{(t)},H^{(t)})$. This rule preserves nonnegativity and continuous, $\mathcal{S}_0$ is closed. Nonempti
        yields a nonincreasing value of the reconstruction term (conditional on from $\theta^{(0)}\in\mathcal{S}_0$. \hfill$\square$
        $W^{(t)}$ and $\beta^{(t)}$; see e.g.
        [@seung2001algorithms; @pascualmontano2006nonsmooth; @lin2007convergence].\vspace{0.5em}
  \item \emph{$W$-update.} For fixed $(H^{(t+1)},\beta^{(t)})$, the update\textbf{Lemma 2 (Monotone descent and existence of limit points)
        $W^{(t)} \mapsto W^{(t+1)}$ is the hybrid multiplicative step Under Assumptions 1 and 2, Algorithm~S\ref{alg:desurv} (without
        followed by backtracking as in Algorithm~S\ref{alg:backtrack}, producing normalization) generates a sequence $\{\theta^{(t)}\}_{t\ge 0}\s
        $W^{(t+1)}$ such that                            such that
        \[                                               \[
          \mathcal{L}(W^{(t+1)},H^{(t+1)},\beta^{(t)})     \mathcal{L}(\theta^{(t+1)}) \le \mathcal{L}(\theta^{(t)})\quad
          \le                                            \]
          \mathcal{L}(W^{(t)},H^{(t+1)},\beta^{(t)}).    and $\{\mathcal{L}(\theta^{(t)})\}$ converges to a finite limit
```

Moreover, $\{\theta^{(t)}\}$ is bounded and therefore admits at least one limit point.

\emph{Proof.}
By Assumption 2(i),
\[
  \mathcal{L}(W^{(t)},H^{(t+1)},\beta^{(t)})
  \le
  \mathcal{L}(W^{(t)},H^{(t)},\beta^{(t)}).
\]
By Assumption 2(ii),
\[
  \mathcal{L}(W^{(t+1)},H^{(t+1)},\beta^{(t)})
  \le
  \mathcal{L}(W^{(t)},H^{(t+1)},\beta^{(t)}).
\]
By Assumption 2(iii),
\[
  \mathcal{L}(W^{(t+1)},H^{(t+1)},\beta^{(t+1)})
  \le
  \mathcal{L}(W^{(t+1)},H^{(t+1)},\beta^{(t)}).
\]
Combining these inequalities gives
\[
  \mathcal{L}(\theta^{(t+1)}) \le \mathcal{L}(\theta^{(t)}) \quad \text{for all } t.
\]
so $\{\mathcal{L}(\theta^{(t)})\}$ is monotonically nonincreasing. Since $\mathcal{L}$ is bounded below on Θ, the sequence converges to some finite $\mathcal{L}^*$.

Each iterate lies in $\mathcal{S}_0$ by construction, and $\mathcal{S}_0$ is bounded by Lemma 1. Therefore $\{\theta^{(t)}\}$ is bounded and has at least one limit point by the Bolzano--Weierstrass theorem. \hfill\square

\vspace{0.5em}
\textbf{Lemma 3 (Blockwise optimality of limit points).}
Let $\theta^*=(W^*,H^*,\beta^*)$ be any limit point of $\{\theta^{(t)}\}$ under Assumptions~1--2. Then:
\begin{enumerate}[label=(\roman*)]
  \item H^* satisfies the KKT conditions for
        \[
          \min_{H\ge 0} \mathcal{L}(W^*,H,\beta^*).
        \]
  \item W^* satisfies the KKT conditions for
        \[
          \min_{W\ge 0} \mathcal{L}(W,H^*,\beta^*).
        \]
  \item β^* is the unique minimizer of
        \[
          \min_{\beta\in\mathbb{R}^k} \mathcal{L}(W^*,H^*,\beta),
        \]
        and satisfies the KKT conditions for the elastic-net subproblem.
\end{enumerate}

\emph{Proof.}
Let $\{t_j\}$ be a subsequence such that $\theta^{(t_j)}\to\theta^*=(W^*,H^*,\beta^*)$. By continuity of $\mathcal{L}$, $\mathcal{L}(\theta^{(t_j)})\to\mathcal{L}^*$.

(i) Conditional on (W,β), the H-subproblem is
\begin{equation}
\min_{H\ge 0} \frac{1}{np}(\lVert X-WH \rVert_F^2 +
\end{equation}
a strictly convex quadratic program. The multiplicative update
to be a descent method whose fixed points coincide with the KKT
constrained subproblem [@seung2001algorithms; @lin2007convergen
objective $\mathcal{L}(\theta^{(t)})$ converges, the per-iterat
in $\mathcal{L}$ due to the H-update must vanish along $\{t_j
If H^* were not KKT for the H-subproblem given $(W^*,\beta^
exist a feasible descent direction for H at (W^*,β^*),
for sufficiently large j the H-update would still produce a
in $\mathcal{L}$, contradicting convergence. Hence H^* satisf
conditions.

(ii) Conditional on (H,β), the W-subproblem is differen
convex in W (sum of a quadratic term and a negative Cox parti
log-likelihood in a linear predictor). The hybrid multiplicativ
backtracking can be viewed as a projected gradient-like update
nonnegative orthant. Under mild conditions on the search direct
sizes, any limit point of such a projected gradient scheme is a
the constrained subproblem; see, for example, [@tseng2001conver
If W^* did not satisfy the KKT conditions, there would be a f
descent direction at (W^*,H^*,β^*) and a sufficiently sma
reduces $\mathcal{L}$, which the line search would eventually a
would contradict the fact that $\mathcal{L}(\theta^{(t_j)})\to\
Therefore W^* satisfies KKT for the W-subproblem.

(iii) For fixed (W,H), the β-subproblem is the convex e
penalized Cox log-partial
β-block given (W^*,H^*) and hence satisfies its KKT con
\hfill$\square$

\vspace{0.5em}
\textbf{Theorem 1 (Convergence to a stationary point).}
Under Assumptions 1 and 2, the sequence $\{\theta^{(t)}\}$ prod
Algorithm~S\ref{alg:desurv} (without W normalization) satisfi
\begin{enumerate}[label=(\alph*)]
  \item The objective values $\{\mathcal{L}(\theta^{(t)})\}$ ar
        nonincreasing and converge to a finite limit $\mathcal{
  \item Every limit point $\theta^* = (W^*,H^*,\beta^*)$ of $\{
        is a stationary point of $\mathcal{L}$ on Θ, i.e
        the first-order KKT conditions for
        \[
          \min_{\theta\in\Theta} \mathcal{L}(\theta).
        \]
\end{enumerate}

\emph{Proof.}
Part (a) is Lemma 2. For part (b), Lemma 3 shows that any limit
is blockwise optimal: each block is optimal (in the KKT sense)
are held fixed.

The Cox loss can be written as
\[
\mathcal{L}(\theta) = f(W,H,\beta) + g(\beta)
                    + I_{\{W\ge 0\}}(W) + I_{\{H\ge 0\}}(H),
\]
where f is continuously differentiable, $g(\beta)=\lambda\xi\
a separable convex (possibly nondifferentiable) penalty, and $I
indicator of a closed convex set C. This is the smooth+nonsmo
form considered in block coordinate descent analyses such as [@
In this setting, any point that is optimal with respect to each

individually (a block coordinatewise minimizer) is a stationary point of the full problem: equivalently,

$$
0 \in \nabla_{W,H} f(W^*,H^*,\beta^*) + \partial I_{\{W\ge 0\}}(W^*) + \partial I_{\{H\ge 0\}}(H^*),
$$

$$
0 \in \nabla_\beta f(W^*,H^*,\beta^*) + \partial g(\beta^*),
$$

which are precisely the KKT conditions for $\min_{\theta\in\Theta}\mathcal{L}(\theta)$. Hence every limit point of Algorithm~S\ref{alg:desurv} is stationary. \hfill\square$

\subsubsection{Effect of column normalization of W}
\label{subsec:conv_normalization}

The above analysis omits the column-normalization step in Algorithm~S\ref{alg:backtrack}. We briefly argue that this normalization does not affect stationarity of limit points.

Let D be a diagonal matrix with strictly positive diagonal entries. The transformation

$$
(W,H,\beta) \mapsto (W',H',\beta') = (W D^{-1},\, D H,\, D\beta)
$$

preserves both the product WH and the linear predictor $WH\beta$, and hence leaves $\mathcal{L}$ invariant. The column-normalization step is exactly such a transformation, with D chosen from the column norms of W.

Let $\{\theta^{(t)}\}$ be the sequence generated by the algorithm without normalization, and let $\{\tilde{\theta}^{(t)}\}$ be the sequence with normalization. For each t there exists a diagonal $D^{(t)}$ with strictly positive entries such that

$$
\tilde{\theta}^{(t)} = (W^{(t)} D^{(t),-1},\, D^{(t)} H^{(t)},\, D^{(t)}\beta^{(t)}),
$$

and $\mathcal{L}(\tilde{\theta}^{(t)}) = \mathcal{L}(\theta^{(t)})$. Thus limit points of $\{\tilde{\theta}^{(t)}\}$ are obtained from limit points of $\{\theta^{(t)}\}$ by such invertible diagonal scalings.

Since the KKT conditions are expressed in terms of the gradients with respect to WH and $W\beta$, and these quantities are invariant under such scaling, stationarity is preserved under the transformation. Therefore Theorem~1 implies that every limit point of the \emph{normalized} algorithm is also a stationary point of $\mathcal{L}$ (up to this scaling ambiguity), which establishes convergence of the implementation used in practice.

\subsubsection{Remark (Coxnet implementation)}
Our implementation updates the β block using a Coxnet-style coordinate descent that relies on an approximate Hess

\subsubsection{Supervision on \(W\) versus \(H\)}

Classical nonnegative matrix factorization (NMF) is non-identifiable: for any invertible matrix \(R\) with nonnegative entries, the factorization \((W, H)\) can be transformed to \((W R, R^{-1} H)\) without changing the reconstruction error, yielding multiple valid solutions [@donoho2004does; @laurberg2008uniqueness; @gillis2014nmf]. Altho columns of \(W\) or rows of \(H\) resolves the trivial scaling a nonnegative mixing transformations persist, and empirical studie show that the sample-loading matrix \(H\) is more sensitive to i and local minima than the program matrix \(W\) [@brunet2004metagenes; @kim2007sparse; @gillis2012accelerated].

In DeSurv, survival supervision enters through the projection $Z

$$
\nabla \mathcal{L} = (1-\alpha) \nabla_W \mathcal{L}_{\mathrm{Co}
$$

and therefore the update rule for W as in Equation~\ref{eqn:W_ genes aligned with the hazard gradient and suppresses those that prognostic structure, ensuring that the learned programs encode biological variation.

By contrast, if supervision were applied to the sample loadings could reduce the Cox loss by redistributing patient-specific coe leaving the gene-level programs \(W\) nearly unchanged, a behavi for W in supervised variants of NMF and supervised topic models, supervision applied only to the coefficient matrix modifies \(H\ basis \(W\) [@cai2011graph; @wang2014supervised; @blei2007slda].

Supervising \(W\) also provides a practical advantage for **port because \(W\) defines gene-level programs, new samples can be em closed-form projection \(Z_{\text{new}} = X_{\text{new}}^\top W\ Supervising \(H\) would not yield such a mapping; instead, obtai loadings for external datasets would require solving a nonnegati (NNLS) problem for each new sample, introducing optimization noi reproducibility.

Together, these considerations motivate supervising through \(W\ aligns programs toward prognostic directions, stabilizes solut interpretation, and yields biologically interpretable signature generalize across cohorts.

\subsubsection{Cross-validation procedure}
Algorithm~S\ref{alg:cv} describes the cross validation procedure
We define the c-index using comparable pairs
$\mathcal{P} = \{(i,j) : y_i < y_j,\ \delta_i = 1\}$ and linear
$\hat{\eta}_i$. The concordance index is

$$
\begin{equation}
\label{eqn:cindex}
\mathrm{c\text{-}index} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathbb{1}(\hat{\eta}_i > \hat{\eta}_j) + \frac{1}{2}\mathbb{1}(\hat{\eta}_i = \hat{\eta}_j) \Big],
\end{equation}
$$

\begin{algorithm}[H]
\caption{Cross-validation for DeSurv pipeline}
\label{alg:cv}
\begin{algorithmic}[1]
\Require
\State Number of folds F
\State Number of initializations R
\State Observed data (X, y, δ)

```
\Statex Hyperparameters $k$, $\alpha$, $\lambda$, $\xi$, and $\lambda_H$
\Ensure The cross-validated c-index

\State Divide subjects into $F$ folds.
\For{$f = 1$ to $F$}
    \State Split data into training and validation $(X_{(-f)}, y_{(-f)}, \delta_{(-f)})$ and $(X_{(f)}, y_{(f)}, \delta
    \For{$r = 1$ to $R$}
        \State set.seed$(r)$
        \State Apply Algorithm~\ref{alg:desurv} with inputs $(X_{(-f)},y_{(-f)}, \delta_{(-f)})$, and $(k, \alpha, \la
        \State Obtain $\hat{W}$ and $\hat{\beta}$ as output from Algorithm~\ref{alg:desurv}
        \State Compute the estimated linear predictor:
        $
        \hat{\eta} = X_{(f)}^T\hat{W}\hat{\beta}
        $
        \State Compute c-index, denoted $\hat{c}_{(f)r}$, according to Equation~\ref{eqn:cindex}
    \EndFor
    \State \textbf{end loop over $r$}
    \State Compute average c-index across initializations:
    $
    \hat{c}_{(f)} = \frac{1}{r}\sum_{r=1}^R\hat{c}_{(f)r}
    $
\EndFor
\State \textbf{end loop over $f$}
\State Compute final cross-validated c-index:
$
\hat{c} = \frac{1}{F}\sum_{f=1}^F \hat{c}_{(f)}
$
\State \Return Final estimate $\hat{c}$

\end{algorithmic}
\end{algorithm}


\subsection{Bayesian Optimization}

\subsection{PDAC Datasets}

\subsection{Bladder Datasets}

\subsection{Consensus Clustering}

\subsection{Survival Analysis}

\pnasbreak

# Versioning {.unnumbered}
```

A. DeSurv package version: 1.0.1.

B. DeSurv git branch: 20260107bugfix.

C. DeSurv git commit: fea641a96e4743315a35b9b6addcc1a1ff53da9d.

D. Paper git branch: main.

E. Paper git commit: 524e354dd777e74a98130a69b6296cca4eefbd51.

“‘

1. Pareja F, et al. (2016) Triple-negative breast cancer: The importance of molecular and histologic subtyping, and recognition of low-grade variants. *NPJ breast cancer* 2(1):1–11.
2. Dienstmann R, Salazar R, Tabernero J (2018) Molecular subtypes and the evolution of treatment decisions in metastatic colorectal cancer. *Am Soc Clin Oncol Educ Book* 38(38):231–8.
3. Zhou X, et al. (2021) Clinical impact of molecular subtyping of pancreatic cancer. *Frontiers in cell and developmental biology* 9:743908.
4. Seiler R, et al. (2017) Impact of molecular subtypes in muscle-invasive bladder cancer on predicting response and survival after neoadjuvant chemotherapy. *European urology* 72(4):544–554.
5. Prat A, et al. (2015) Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast* 24:S26–S35.
6. Ou F, Michiels S, Shyr Y, Adjei AA, Oberg AL (2021) Biomarker discovery and validation: Statistical considerations. *Journal of Thoracic Oncology* 16(Suppl 15):S539–S547.
7. Planey Catherine R, Gevaert O (2016) CoINcIDE: A framework for discovery of patient subtypes across multiple datasets. *Genome Medicine* 8(1):27.
8. Prat A, Pineda E, Adamo B, et al. (2014) Molecular features and survival outcomes of the intrinsic subtypes in the international breast cancer study group trial 10-93. *Journal of the National Cancer Institute* 106(8):dju152.
9. Ellrott K, et al. (2025) Classification of non-TCGA cancer samples to TCGA molecular subtypes using compact feature sets. *Cancer cell* 43(2):195–212.
10. Tomczak K, Czerwińska P, Wiznerowicz M (2015) Review the cancer genome atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia* 2015(1):68–77.
11. Zhang J, et al. (2019) The international cancer genome consortium data portal. *Nature biotechnology* 37(4):367–369.
12. Nguyen H, Nguyen H, Tran D, Draghici S, Nguyen T (2024) Fourteen years of cellular deconvolution: Methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Research* 52(9):4761–4783.
13. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *nature* 401(6755):788–791.
14. Bailey P, Chang DK, et al. (2016) Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 531(7592):47–52.

15. Moffitt RA, et al. (2015) Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature genetics* 47(10):1168–1178.

16. Peng XL, Moffitt RA, Torphy RJ, Volmar KE, Yeh JJ (2019) De novo compartment deconvolution and weight estimation of tumor samples using DECODER. *Nature communications* 10(1):4729.

17. Le Goff V, et al. (2025) SurvNMF: Non-negative matrix factorization supervised for survival data analysis. PhD thesis (Institut Pasteur Paris; CEA).

18. Huang Z, Salama P, Shao W, Zhang J, Huang K (2020) Low-rank reorganization via proportional hazards non-negative matrix factorization unveils survival associated gene clusters. *arXiv preprint arXiv:200803776*.

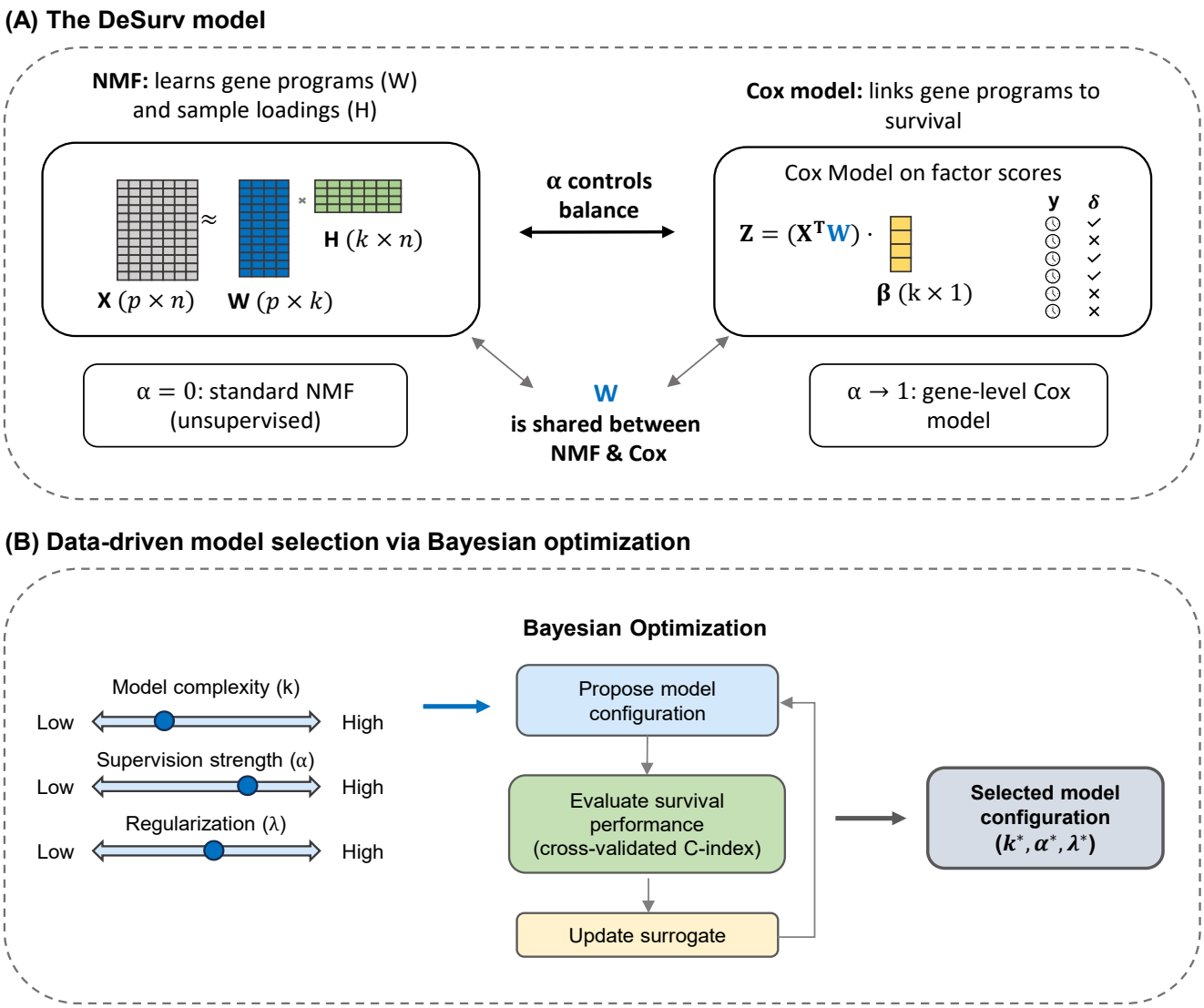

Fig. 1. Overview of the DeSurv framework and data-driven model selection. (A) DeSurv integrates nonnegative matrix factorization (NMF) with survival modeling to learn prognostic gene programs from a gene expression matrix X. NMF decomposes $X \approx WH$, where W represents gene programs and H sample loadings; the learned programs W are shared with a Cox proportional hazards model that links factor-derived scores $Z = X^\top W$ to survival outcomes via regression coefficients β. A tuning parameter α controls the balance between unsupervised structure learning ($\alpha = 0$) and supervised survival association ($\alpha = 1$). (B) Model complexity (k), supervision strength (α), and regularization (λ) are selected via Bayesian optimization using cross-validated concordance index.