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 October 10, 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 nonprognostic 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.

Materials and methods

NMF. Let $X \in R^{p\times n}_{\geq 0}$ denote a nonnegative gene expression matrix of p features (genes) across n subjects. The goal of NMF is to approximate X as the product of two low-rank,

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

nonnegative matrices

$$X \approx WH, \qquad [1]$$

where $W \in R_{\geq 0}^{p \times k}$ contains the gene weights (or "metagenes"), and $H \in R_{\geq 0}^{k \times n}$ contains the factor scores for each subject. The number of latent factors k determines the dimensionality of the shared low-rank representation. The NMF loss is defined as the residual sum of squares.

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

Cox partial likelihood. To incorporate survival outcomes, let T_i denote the event time and C_i the censoring time for subject i. The observed time is $y_i = \min(T_i, C_i)$, and the event indicator is δ_i. Given that W is shared across datasets, we define a lower dimensional transformation of the data:

$$Z = X^T W \in R^{n \times k}, \qquad [3]$$

where each row Z_i^T represents the factor scores for subject i. These scores serve as covariates in a Cox proportional hazards model:

$$h_i(t) = h_0(t) \exp(Z_i^T \beta), \qquad [4]$$

where $h_0(t)$ is the baseline hazard and $\beta \in R^k$ are the factor specific coefficients. The Cox log partial likelihood is then

$$\ell(W, \beta) = \sum_{i=1}^{n} \delta_i \left[Z_i^T \beta - \log \left(\sum_{j: y_j \geq y_i} \exp\left(Z_j^T \beta\right) \right) \right]. \qquad [5]$$

Maximizing this quantity . . .

DeSurv. Building on the definitions above, DeSurv combines the unsupervised NMF reconstruction loss and the supervised Cox partial likelihood into a single joint objective. The combined loss function is

$$\mathcal{L}(W, H, \beta) = (1 - \alpha)\mathcal{L}(W, H)_{NMF} - \alpha\mathcal{L}(W, \beta)_{cox}, \qquad [6]$$

where $\mathcal{L}_{cox}(W, \beta)$ is the elastic net penalized log partial likelihood:

$$\mathcal{L}_{cox}(W, \beta) = \ell(W, \beta) + \lambda(\xi||\beta||_1 + \frac{(1-\xi)}{2}||\beta||_2^2), \qquad [7]$$

where λ represents the penalty weight and ξ is the balance parameter between the L1 and L2 penalty terms.

The hyperparameter $\alpha \in [0, 1]$ controls the relative contribution of each component:

- $\alpha = 0$ recovers standard NMF, focusing purely on reconstruction;
- $\alpha = 1$ corresponds to a fully supervised Cox model in the low-dimensional space $Z = X^T W$

Intermediate values of α encourage discovery of latent molecular programs that are both biologically coherent and prognostically informative.

Update Rules. DeSurv is optimized using an alternating minimization scheme (Algorithm 1) that iteratively updates W, H, and β until convergence. The sub-problems for H and β are convex in the corresponding parameter conditional on the others.

Algorithm 1 DeSurv algorithm

```
Input: $X \in \mathbb{R}_{\geq 0}^{p \times n}$, $y \in \mathbb{R}_{\geq 0}^{n}$, $\delta \in \mathbb{R}_{0,1}^{n}$, $W^{(0)}$, $H^{(0)}$, $\beta^{(0)}$, $tol$, $maxit$
$eps = \infty$
$iter = 0$
$loss = 0$
while $eps < tol$ and $iter < maxit$ do
    $W^{(iter)} = \text{argmin}_{W \geq 0} \mathcal{L}(W^{(iter-1)}, H^{(iter-1)}, \beta^{(iter-1)})$
    $H = \text{argmin}_{H \geq 0} \mathcal{L}(W^{(iter)}, H^{(iter-1)}, \beta^{(iter-1)})$
    $\beta = \text{argmin}_{\beta} \mathcal{L}(W^{(iter)}, H^{(iter)}, \beta^{(iter-1)})$
    $lossNew = \mathcal{L}(W^{(iter)}, H^{(iter)}, \beta^{(iter)})$
    $eps = |lossNew - loss|/loss$
    $loss = lossNew$
    $iter = iter + 1$
return $W$,$H$,$\beta$
```

Update for H**.** The nonnegative factor matrix H is updated using standard multiplicative updates that guarantee nonnegativity and monotonic decrease in reconstruction error as derived in (**?**):

$$H_{ij} = H_{ij} \frac{(W^T X)_{ij}}{(W^T W H)_{ij}}. \qquad [8]$$

Update for β**.** Given W, the coefficients β are updated by coordinate descent using elastic-net regularization:

$$\hat{\beta}_r = \frac{S(\frac{1}{n}\sum_{i=1}^{n} w(\tilde{\eta})_i v_{i,r} \left[z(\tilde{\eta})_i - \sum_{j \neq r} v_{ij}\beta_j, \right], \lambda\xi)}{\frac{1}{n}\sum_{i=1}^{n} w(\tilde{\eta})_i v_{i,r}^2 + \lambda(1-\xi)}, \qquad [9]$$

where $S(., \lambda\xi)$ is the soft-thresholding operator, w_i is blah, and v_{ij} . The parameters (λ, ξ) control the strength and type of regularization.

Coupled W ***update.*** The shared basis W is updated through a hybrid multiplicative rule that incorporates both NMF reconstruction gradients and Cox partial likelihood gradients. Backtracking and gradient balancing are used in this update to ensure decrease in the overall loss and avoid one component dominating the update. (Add citations for this here)

$$W^{(t+1)} = W^{(t)} \odot \max\left(\frac{\frac{(1-\alpha)}{np} X H^T + \frac{2\alpha}{N_{event}} \nabla_W \ell(W^{(t)}, \beta)}{\frac{(1-\alpha)}{np} W^{(t)} H H^T}, 0 \right) \qquad [10]$$

The quantity $\nabla_W \ell(W^{(t)}, \beta)$ denotes the gradient of the Cox partial likelihood with respect to W. This update allows the survival signal to propagate into the latent factors while preserving nonnegativity.

Publicly Available Datasets. We trained and validated DeSurv using seven publicly available pancreatic ductal adenocarcinoma (PDAC) transcriptomic datasets spanning both RNA-seq and microarray platforms. Expression data were harmonized to gene-level matrices and transformed to transcripts per million (TPM) where applicable. To reduce platform-specific bias, each dataset was rank-transformed across genes within samples prior to modeling.

Model Training. The TCGA and CPTAC datasets were used for model training. Each dataset was filtered to the top 2000 highly expressed and variable genes. These gene lists were

Table 1. Publicly available pancreatic ductal adenocarcinoma (PDAC) datasets used for model training and validation. Expression data were rank-transformed across genes within samples to mitigate platform- and scale-related effects.

Dataset	Platform	Samples (n)	Data type	Reference
TCGA-PAAD	RNA-seq (Illumina HiSeq)	∼178	Discovery / Training	(?)
CPTAC-PDAC	RNA-seq (Proteogenomic)	∼138	Validation	(?)
Dijk *et al.*	Microarray (Affymetrix)	∼80	Validation	(?)
Moffitt *et al.*	Microarray (Affymetrix)	∼84	Validation	(?)
PACA-AU (array)	Microarray (Agilent)	∼269	Validation	(?)
PACA-AU (RNA-seq)	RNA-seq (Illumina HiSeq)	∼92	Validation	(?)
Puleo *et al.*	Microarray (Affymetrix)	∼288	Validation	(?)

then intersected, resulting in XXX genes incorporated in model training.

A. Cross Validation. To identify optimal hyperparameters, we employed stratified five-fold cross-validation based on event status. An exhaustive search was performed across a grid of hyperparameters $k = 2, \dots, 12$, $\alpha \in \{0.1, 0.2, \dots, 0.9\}$, $\lambda \in 10^{\{-3,\dots,3\}}$, and $\xi \in \{0, 0.1, 0.2, \dots, 1\}$. Within each fold, models were fit on 80% of the data and evaluated on the held-out 20%. Because NMF solutions are non-unique and sensitive to initialization, each fold was repeated with 20 random seeds for W, H, and β, resulting in a total of 100 trained models per parameter configuration of k, $lambda$, and eta. Warm-start initializations were used across $alpha$ to accelerate convergence.

Final hyperparameters were selected as the combination that produced average C-index (across initializations and folds) within one standard error of the maximum (1 s.e. rule).

Comparison Methods. To benchmark the performance of DeSurv, we compared it against several established approaches for molecular subtyping and survival prediction. All comparison models were trained using the same discovery cohort, the same filtered set of 1,000 highly variable genes, and the same preprocessing steps.

Standard NMF. As an unsupervised baseline, we trained conventional NMF models that minimized only the reconstruction error, corresponding to setting $alpha = 0$ in the DeSurv formulation. For each rank $k \in 2, ..., 12$, 100 random initializations were performed to address non-uniqueness. The initialization with the smallest reconstruction error for each k was selected. To select the rank, k, in the unsupervised setting, we use standard metrics including cophenetic coefficient, dispersion, explained variance, residuals, silhouette score, and sparseness. The resulting gene-factor matrix W and validation data were subsequently used as input to a Cox proportional hazards model to evaluate the prognostic value of the unsupervised factors.

Cox Elastic Net (CoxNet). We also compared DeSurv to a standard penalized Cox regression model fit directly to gene-level expression data with elastic net penalty. Five-fold internal cross-validation within the training data was used to select the penalization parameters based on the mean C-index. The final CoxNet model served as a high-dimensional supervised baseline without dimensionality reduction.

Model Validation. The remaining 6 publicly available PDAC datasets, CPTAC, Dijk, 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.

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.

Top genes. For a factor f a top gene is defined as a gene that is highly weighted in factor f while being lowly weighted in all other factors. We define s_{gf} as the difference in weights between gene g in factor f and the max weight across all other factors in gene g.

$$s_{gf} = W_{gf} - \max_{r \neq f}(W_{gr}) \tag{11}$$

Clustering.

Results

B. 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 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.

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

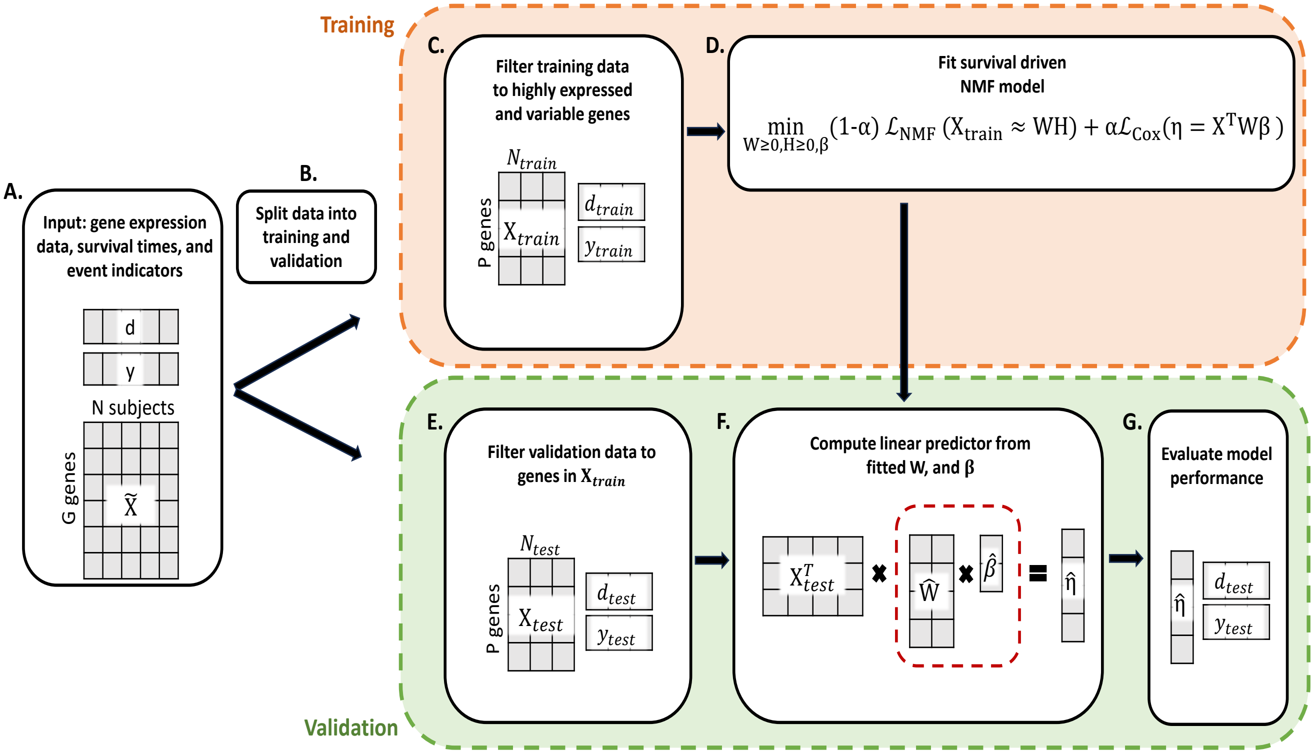
Training
A.
Input: gene expression data, survival times, and event indicators
d
y
N subjects
G genes
$\tilde{X}$
B.
Split data into training and validation
C.
Filter training data to highly expressed and variable genes
N_{train}
P genes
X_{train}
d_{train}
y_{train}
D.
Fit survival driven NMF model
$\min_{W\geq0,H\geq0,\beta}(1\text{-}\alpha)\,\mathcal{L}_{NMF}\,(X_{train}\approx WH)+\alpha\mathcal{L}_{Cox}(\eta=X^{T}W\beta\,)$
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_{test}^{T}
$\hat{W}$
$\hat{\beta}$
$\hat{\eta}$
G.
Evaluate model performance
$\hat{\eta}$
d_{test}
y_{test}
Validation

Fig. 1. DeSurv overview

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 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 X). Factor correlation analyses (Figure X) 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.

E. 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 X). Kaplan–Meier curves stratified by factor exposures revealed significant survival differences (Figure X), 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 X) demonstrated consistency with external signatures, confirming that NMF–Cox produces reproducible biological features.

Discussion

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.

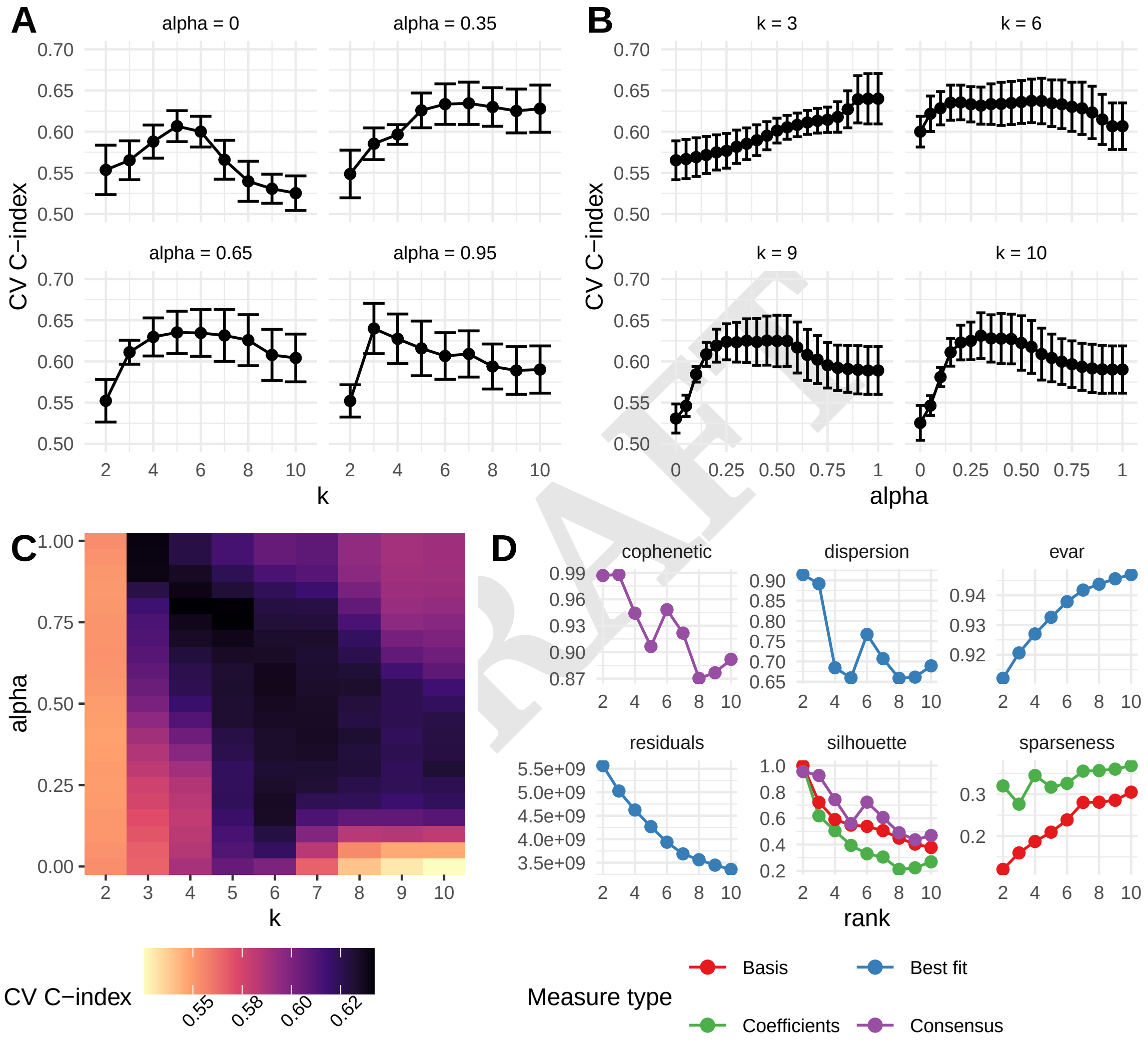

Fig. 2. A. Cross validated c-index over rank (k) with panels for various values of the trade-off parameter (alpha). B. Cross validated c-index over alpha with panels for various values of k. C. Heatmap of cross-validated c-index with columns k and rows alpha. D. Plots of various metrics used to choose k in the standard NMF setting.

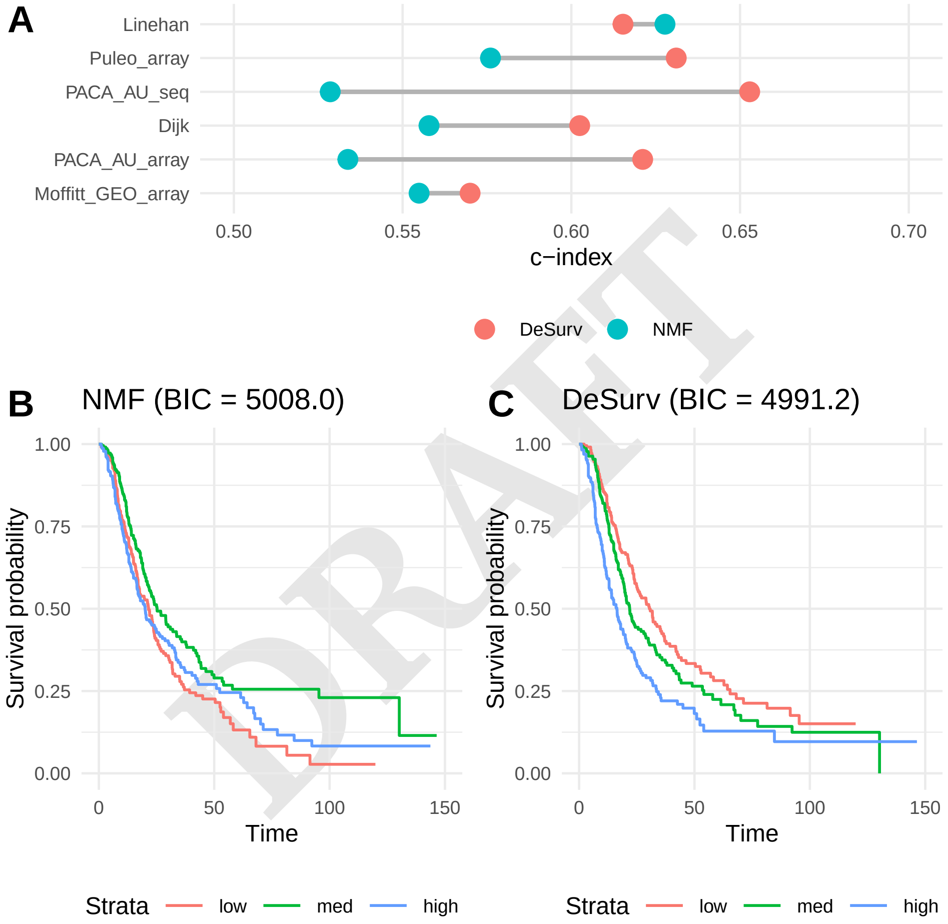

Fig. 3. Model performance in valiation data. A. C-index for standard NMF vs DeSurv. B. Kaplan Meier curves by quantile of the linear predictor for standard NMF. C. Kaplan Meier curves by quantile of the linear predictor for DeSurv