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 February 6, 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 factorization (NMF) with Cox proportional hazards modelling. In contrast to fully unsupervised approaches that evaluate survival associations only after the factorization, and to existing

Significance Statement

Tumor transcriptomes mix malignant and microenvironmental signals, making it difficult to identify programs that drive clinical outcomes. Existing deconvolution and matrix factorization methods discover latent programs but do not ensure prognostic relevance, while supervised predictors often sacrifice biological interpretability. We present DeSurv, a survival-supervised deconvolution framework that integrates nonnegative matrix factorization with Cox modeling to learn gene programs and their survival associations jointly. By embedding outcome information into discovery and using automatic model selection, DeSurv yields clinically relevant, reproducible programs across cohorts. This advances tumor deconvolution and provides a general tool for identifying actionable 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

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) together with corresponding survival times (y) and censoring indicators (δ) (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):

$$(1-\alpha)\ \mathcal{L}_{NMF}(X \approx WH) \\ -\alpha\ \mathcal{L}_{Cox}(Z^\top \beta, y, \delta) \qquad [1]$$

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 ($W^\top X$) represents patient-level factor scores (Z) - the inferred burden of each latent program across subjects. These factor scores serve as covariates in the Cox model with linear predictor $Z^\top \beta$, 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 biological interpretability, but also guides latent factors to be prognostically informative. Subsequent analyses can therefore focus on the survival-associated gene programs (Figure 1E).

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. 2A-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. 2D).

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.

DeSurv improves selection of prognostic gene signatures. To test whether supervision improves the selection of prognostic gene signatures, we used simulations with known lethal factors and tuned the number of top genes per factor (`n_top`) using BO. We compared the supervised DeSurv model to the unsupervised $\alpha = 0$ baseline across simulation regimes. Figure 3A shows the distribution of test C-index values, while Figure 3B reports the precision of recovered prognostic genes (mean across lethal factors). DeSurv consistently yields higher C-index and improved precision, indicating that survival supervision helps focus signatures on truly prognostic genes rather than reconstruction-only structure.

Survival-informed factorization reorganizes transcriptional structure in pancreatic cancer. To assess the biological structure captured by standard nonnegative matrix factorization (NMF) and DeSurv, we examined the overlap between factor-specific gene rankings and established PDAC gene programs (Fig. 4A-B). Both approaches recovered recognizable biological signals; however, the organization of these signals across factors differed in ways that reflect the objectives of each method.

In the standard NMF solution, factors largely reflected dominant sources of transcriptional variance. One factor was strongly associated with exocrine-associated expression and opposed immune and stromal signatures, consistent with a bulk composition axis separating normal or differentiated tissue from non-epithelial tumor content. A second factor aligned with classical tumor identity, while a third aggregated immune, fibroblast, and extracellular matrix–related programs into a single microenvironmental component. These patterns indicate that standard NMF captures major biological variation in bulk PDAC expression data, but merges distinct microenvironmental states and allocates substantial model capacity to differentiation-driven signals.

By contrast, DeSurv produced a factorization that emphasized axes of variation more closely aligned with disease aggressiveness. One factor corresponded to classical tumor programs opposing basal-like expression, whereas another isolated an activated microenvironmental state characterized by immune infiltration and stromal remodeling. A third factor captured aggressive tumor-intrinsic programs associated with basal-like biology. Notably, exocrine-associated expression did not dominate any DeSurv factor, suggesting that survival-informed optimization deprioritizes differentiation-related variation that is weakly associated with outcome.

These differences were reflected quantitatively by contrasting the fraction of expression variance explained by each factor with its contribution to survival (Fig. 4C). In the standard NMF solution, the factor explaining the largest proportion of transcriptional variance contributed little to survival, consistent with its enrichment for exocrine and composition-driven programs. In contrast, DeSurv concentrated survival signal into a single factor that explained substantially more variation in outcome despite accounting for a smaller fraction of expression variance. The remaining DeSurv factors contributed minimal survival signal, indicating that survival-relevant information was not diffusely distributed across components.

To further characterize how DeSurv reorganizes the transcriptional structure learned by standard NMF, we examined the correspondence between factors derived from the two methods (Fig. 4D). Tumor-intrinsic structure was largely preserved, with one DeSurv factor showing strong correspondence to the classical tumor–associated NMF factor. In contrast, the microenvironmental factor identified by standard NMF mapped primarily to a single DeSurv factor enriched for activated immune and stromal programs, indicating that DeSurv refines variance-driven microenvironmental structure into a survival-aligned axis. Notably, the NMF factor dominated by exocrine-associated expression did not correspond strongly to any single DeSurv factor, consistent with the suppression of differentiation- and composition-driven signals observed in survival-informed factorization. Together, these results indicate that DeSurv selectively preserves tumor and microenvironmental structure while reorganizing or deprioritizing axes of variation that contribute little to survival.

DeSurv-derived latent structure generalizes to independent datasets. To assess generalization of DeSurv latent factors, the gene-level signatures learned in the training cohort were used to compute factor activity scores in independent validation datasets. For each validation sample, factor activity was quantified by applying the learned gene-factor weights to the sample's gene expression profile ($W^T X$), yielding a continuous score for each factor.

For each method, we focused on the factor showing the largest increase in Cox model log partial likelihood in the training data. Across validation cohorts, this DeSurv-derived factor exhibited consistent survival effects, with hazard ratio estimates showing limited variability and predominantly protective associations (Fig. 5A). In contrast, the NMF factor identified by the same criterion showed greater heterogeneity across datasets and weaker survival associations.

When validation samples were pooled and stratified into high- and low-activity groups based on the DeSurv-derived factor, clear separation of survival trajectories was observed (Fig. 5B). Applying the same procedure to NMF resulted in weaker survival stratification (Fig. 5C).

Together, these results indicate that DeSurv identifies latent factors whose associations with survival are preserved across datasets.

Survival-informed factorization identifies prognostic structure across cancers. To assess whether the differences observed between standard nonnegative matrix factorization (NMF) and DeSurv in pancreatic ductal adenocarcinoma extend beyond a single disease context, we applied both methods to an independent bladder cancer cohort and evaluated factor structure, biological interpretation, and survival association using the same analytical framework (Fig. 6).

In bladder cancer, standard NMF again organized transcriptional structure primarily around variance-dominant axes (Fig. 6A). As in PDAC, NMF factors explained substantial fractions of expression variance but contributed little to survival, indicating that variance-driven factorization alone does not preferentially isolate prognostically relevant signals in this setting. This pattern mirrors the behavior observed in PDAC and suggests that the limitations of variance-only factorization are not cancer-type specific.

We next asked whether survival-aligned structure learned in PDAC could transfer across cancer types. Applying the DeSurv model trained in PDAC directly to bladder cancer samples, the survival-aligned factor retained prognostic relevance in the external cohort (Fig. 6B). Kaplan–Meier analysis based on a median split of projected factor scores demonstrated clear separation of survival curves, despite differences in tissue context and transcriptional background. This result indicates that DeSurv learns latent representations that capture survival-relevant structure shared across cancers.

Together, these results show that survival-informed factorization separates variance-dominant from survival-relevant structure and that prognostic signal can transfer across cancer types. By deprioritizing variance-dominant but prognostically neutral signals and concentrating outcome-relevant information into a small number of factors, DeSurv identifies latent structure that is robust across cancers.

Discussion

We present DeSurv, a survival-driven deconvolution framework that integrates nonnegative matrix factorization with Cox proportional hazards modeling to uncover latent gene-expression programs associated with patient outcomes. By coupling matrix decomposition with direct survival supervision, DeSurv targets prognostic structure during discovery rather than only in post hoc evaluation, and its Bayesian optimization strategy addresses the ambiguity of rank selection in standard NMF.

In pancreatic ductal adenocarcinoma (PDAC), survival supervision reorganized variance-dominant structure into factors more aligned with outcome. Compared with standard NMF, DeSurv suppressed exocrine- and composition-driven signals that explained substantial expression variance but contributed little to survival, and instead concentrated survival signal into a smaller set of factors. In simulations with known lethal programs, DeSurv more reliably recovered the true rank and improved both concordance and precision of prognostic gene signatures, indicating that outcome-guided learning can sharpen factor interpretability without sacrificing reconstruction.

The learned factors generalized beyond the training data. Across independent PDAC cohorts, the most prognostic DeSurv factor showed consistent hazard ratios and clearer survival separation than the corresponding NMF factor, indicating that the survival-aligned programs transfer across datasets. In bladder cancer, DeSurv again separated variance-dominant from survival-relevant structure, and a PDAC-trained factor retained prognostic signal when projected into bladder samples. Together, these results support DeSurv as a general framework for learning interpretable, outcome-aligned programs in heterogeneous tumors. Future work will test extensions to additional omics layers, alternative outcome models, and larger multi-cohort benchmarks to clarify when outcome supervision offers the greatest benefit.

Materials and methods

A. Problem formulation and notation. Let $X \in \mathbb{R}_{\geq 0}^{p \times n}$ denote the nonnegative gene expression matrix. DeSurv approximates $X \approx WH$, where $W \in \mathbb{R}_{\geq 0}^{p \times k}$ contains nonnegative gene programs and $H \in \mathbb{R}_{\geq 0}^{k \times n}$ contains sample-level activations. Additionally, let $y,\ \delta \in \mathbb{R}^n$ represent patient survival times and censoring indicators, respectively. Survival outcomes are modeled through a Cox proportional hazards model with covariates $Z = W^\top X$ and linear predictor $Z^\top \beta$.

B. The DeSurv Model. DeSurv integrates Nonnegative Matrix Factorization (NMF) with penalized Cox regression to identify gene programs associated with patient survival.

The joint objective is

$$\mathcal{L}(W, H, \beta) = (1 - \alpha)\, \mathcal{L}_{\text{NMF}}(W, H) - \alpha\, \mathcal{L}_{\text{Cox}}(W, \beta), \quad [2]$$

where $\mathcal{L}_{\text{NMF}}(W, H)$ is the NMF reconstruction error and $\mathcal{L}_{\text{Cox}}(W, \beta)$ is the elastic-net penalized partial log-likelihood. Optimization proceeds by alternating updates for H, W, and β, using multiplicative rules for H (13), projected gradients for W, and coordinate descent for β. Although non-convex, these updates are shown to converge to a stationary point under mild conditions (SI Appendix). Complete derivations and algorithmic details are provided in the SI Appendix.

C. Hyperparameter selection and cross-validation. Hyperparameters $(k, \alpha, \lambda_H, \lambda, \xi)$ were selected by maximizing the cross-validated C-index using Bayesian optimization. Each fold was trained using multiple random initializations, and fold-level performance was defined as the average C-index across initializations. For stability, we used a consensus-based initialization for the final model, aggregating multiple DeSurv runs into a gene–gene co-occurrence matrix and constructing an initialization W_0 from the resulting clusters (SI Appendix). Before validation, each column of W was truncated to its BO-selected number of top genes (details in SI Appendix), denoted $\tilde{W}$. External validation was performed by first projecting new datasets onto the learned programs via $Z = \tilde{W}^\top X_{\text{new}}$ and evaluating survival associations using C-index and log-rank statistics. To evaluate the quality of the DeSurv derived gene signatures for subtyping, the new datasets X_new were clustered on genes in $\tilde{W}$, and survival differences were analyzed for the derived clusters. Further training, validation, and runtime details appear in the SI Appendix.

D. Simulation studies. Simulation studies were conducted to assess recovery of prognostic latent structure and survival prediction. Gene expression data were generated from a nonnegative factor model $X = WH$, where gene loadings W comprised three gene classes: marker genes, background genes, and noise genes. Marker genes were simulated to load strongly on a single factor and weakly on others, background genes to load strongly across all factors, and noise genes to have uniformly low loadings; each class was generated from a distinct gamma distribution. Sample-level factor activities H were generated from a gamma distribution.

Survival times were generated from an exponential distribution in which risk depended on marker gene structure through $X^T\tilde{W}$, where $\tilde{W}$ retained marker gene loadings for their corresponding factors and was zero otherwise; censoring times were generated independently from an exponential distribution. Each dataset was analyzed using DeSurv and standard NMF followed by Cox regression on inferred factors, both tuned using cross-validated concordance index. Performance was summarized across repeated simulation replicates.

E. Real-world datasets. We analyzed publicly available RNA-seq and microarray cohorts of pancreatic ductal adenocarcinoma (PDAC) and bladder cancer with corresponding overall survival outcomes. Gene expression matrices were converted to TPM, log-transformed, and filtered to remove low-expression genes. Survival times and censoring indicators were taken from the associated clinical annotations. Of the seven PDAC cohorts we considered, two were used for training (TCGA and CPTAC) and the rest were used for external validation (Dijk, Moffitt, PACA, Puleo). The bladder cohort was split into training and validation cohorts via a 70/30 split. To harmonize differences in scale across cohorts, filtered gene expression data was within-subject rank transformed before model training. More details about the datasets can be found in the SI Appendix.

F. Simulations, Benchmarking, and Availability. An R package has been developed for DeSurv at github.com/ayoung31/DeSurv. Code and processed data used in this study are available at github.com/ayoung31/DeSurv-paper.

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.

Versioning

```
## DeSurv package version: HEAD
## DeSurv git branch: installed
## DeSurv git commit: unknown
## Paper git branch: naimedits0125
## Paper git commit: f3f9ee43804fc505d7415a16db3b53c577cf2b23
```


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*.

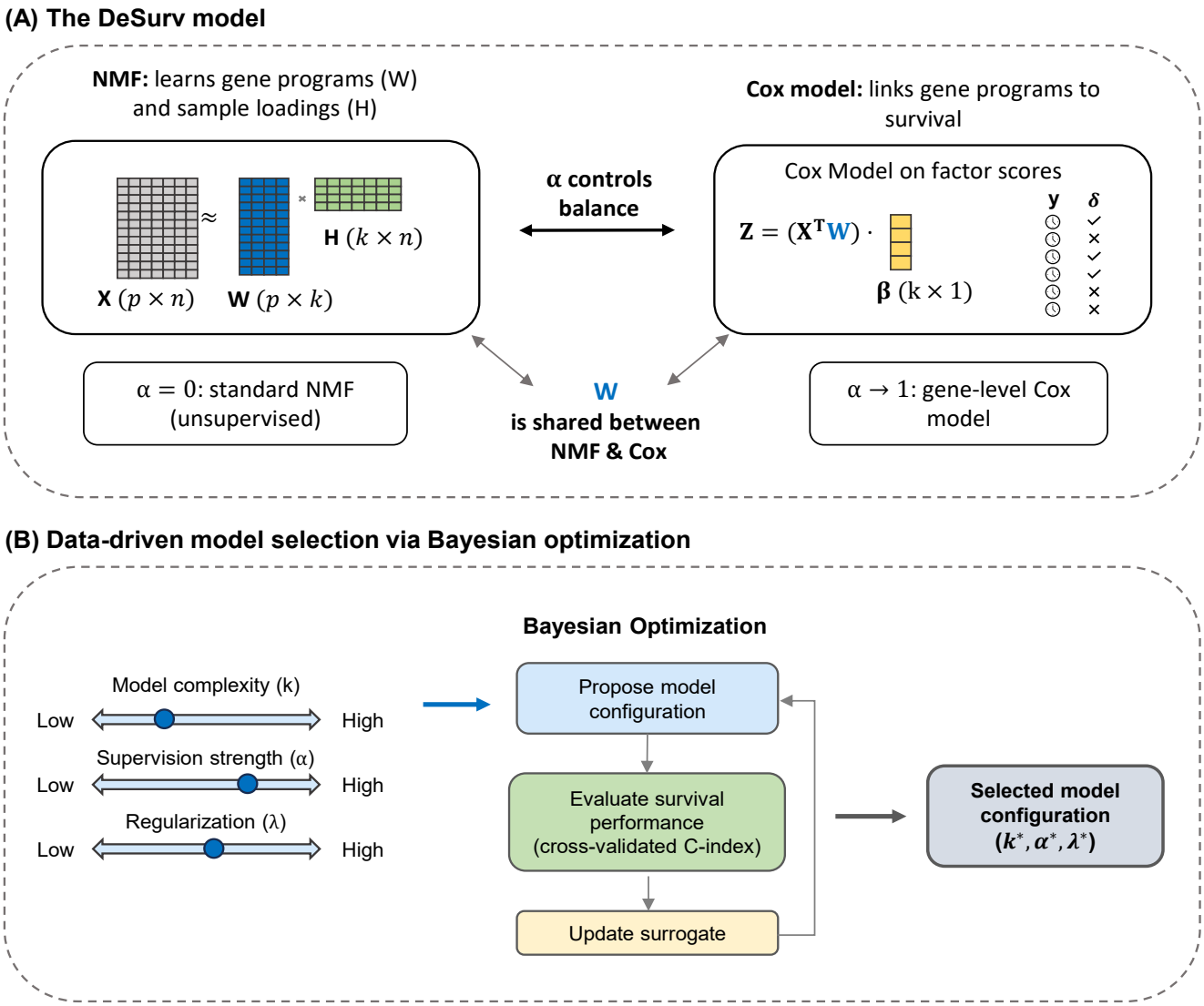

Fig. 1. Overview of the DeSurv framework and data-driven model selection. (A) DeSurv integrates nonnegative matrix factorization (NMF) with survival modeling to learn prognostic gene programs from a gene expression matrix X. NMF decomposes $X \approx WH$, where W represents gene programs and H sample loadings; the learned programs W are shared with a Cox proportional hazards model that links factor-derived scores $Z = W^\top X$ to survival outcomes via linear predictor $Z^\top \beta$. 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.

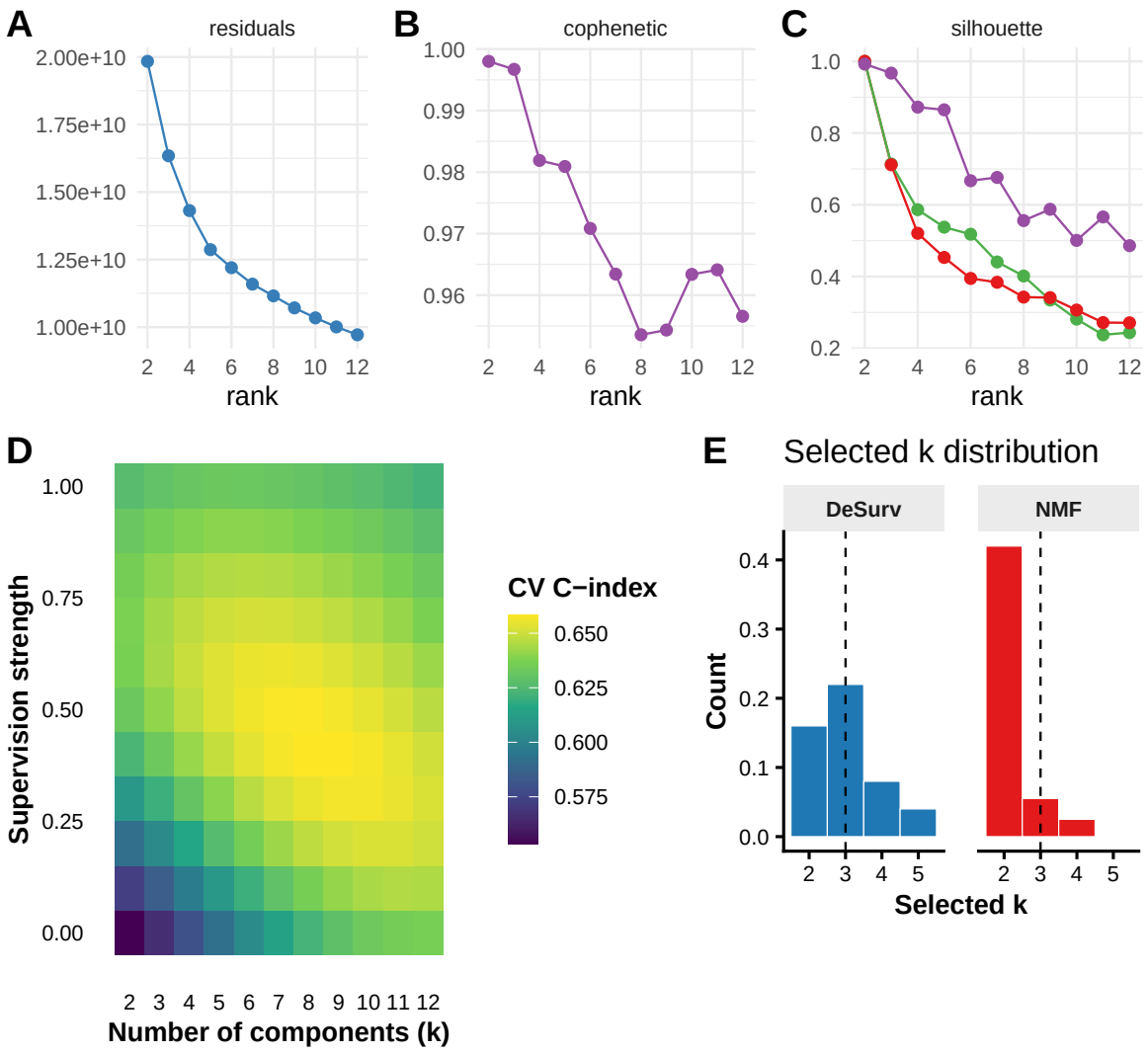

Fig. 2. (A-D) Analyses based on real pancreatic ductal adenocarcinoma (PDAC) gene expression data illustrate the ambiguity of rank selection in standard nonnegative matrix factorization (NMF) and the use of outcome supervision in DeSurv. (A-C) Commonly used unsupervised heuristics for selecting the number 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 ≈ 3-4) or larger (k ≈ 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 ≈ 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 ≈ 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 (α), 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.

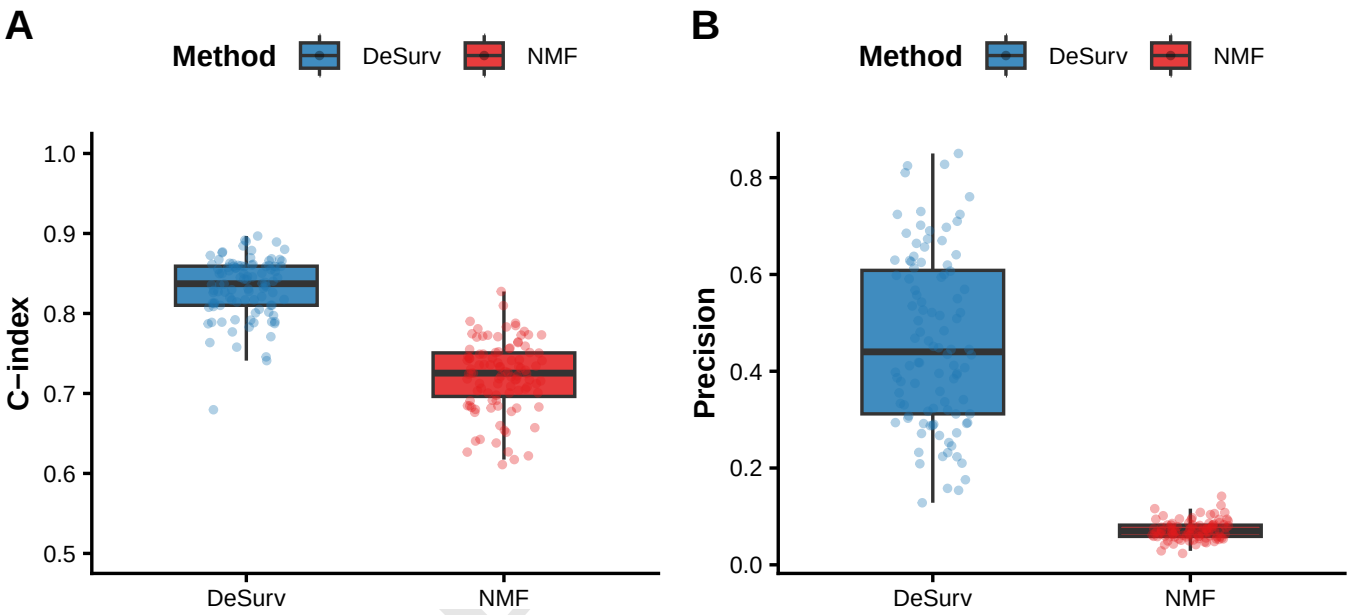

Fig. 3. Performance comparison between DeSurv and unsupervised NMF ($\alpha = 0$) in simulation. (A) Distribution of cross-validated concordance index shows that DeSurv achieves consistently higher survival prediction performance than the unsupervised baseline. (B) Precision for recovering true underlying gene programs is substantially higher for DeSurv, whereas the unsupervised approach exhibits near-zero precision, indicating poor alignment with the ground-truth factors.

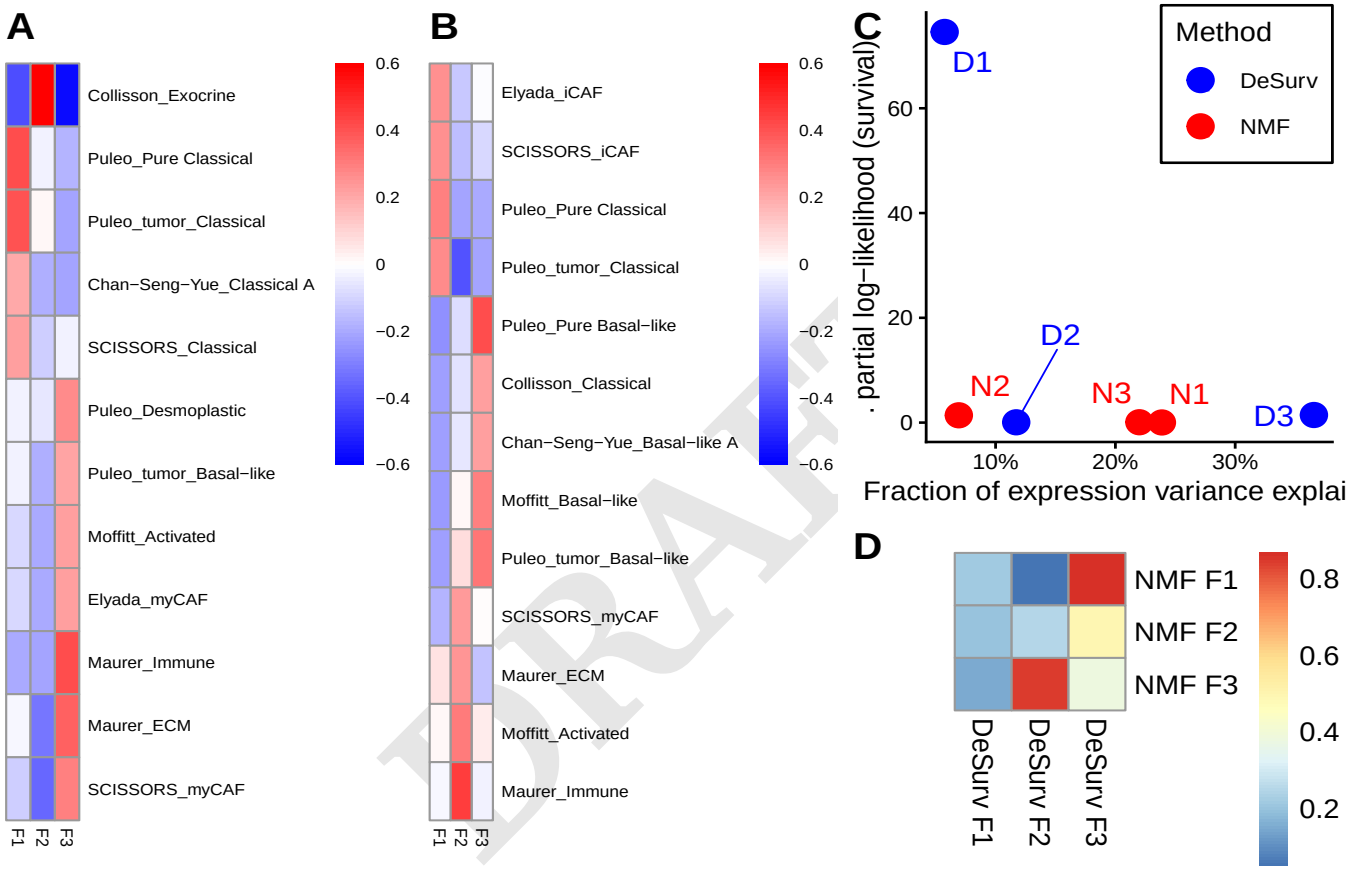

Fig. 4. Survival-informed factorization reorganizes transcriptional structure relative to variance-driven NMF. (A) Correlation of standard NMF factor gene rankings with established pancreatic ductal adenocarcinoma (PDAC) gene programs. NMF recovers variance-dominant structure, including an exocrine-associated factor opposing immune and stromal signatures, a classical tumor factor, and a composite microenvironmental factor. (B) Corresponding correlations for DeSurv. DeSurv isolates tumor-intrinsic classical and basal-like programs and resolves an activated immune–stromal microenvironment, while exocrine-associated expression does not dominate any factor. Asterisks indicate significant correlations after multiple testing correction. Only gene lists with correlation > 0.2 included in A and B. (C) Fraction of expression variance explained versus survival contribution for each factor, quantified by the change in partial log-likelihood from univariate Cox models. Standard NMF factors explain substantial variance with minimal survival relevance, whereas DeSurv concentrates survival signal into a single factor despite explaining less expression variance. (D) Pairwise correspondence between NMF and DeSurv factors. Tumor-intrinsic structure is preserved across methods, while variance-driven microenvironmental structure is reorganized into a survival-aligned axis; the exocrine-dominated NMF factor shows no strong one-to-one correspondence.

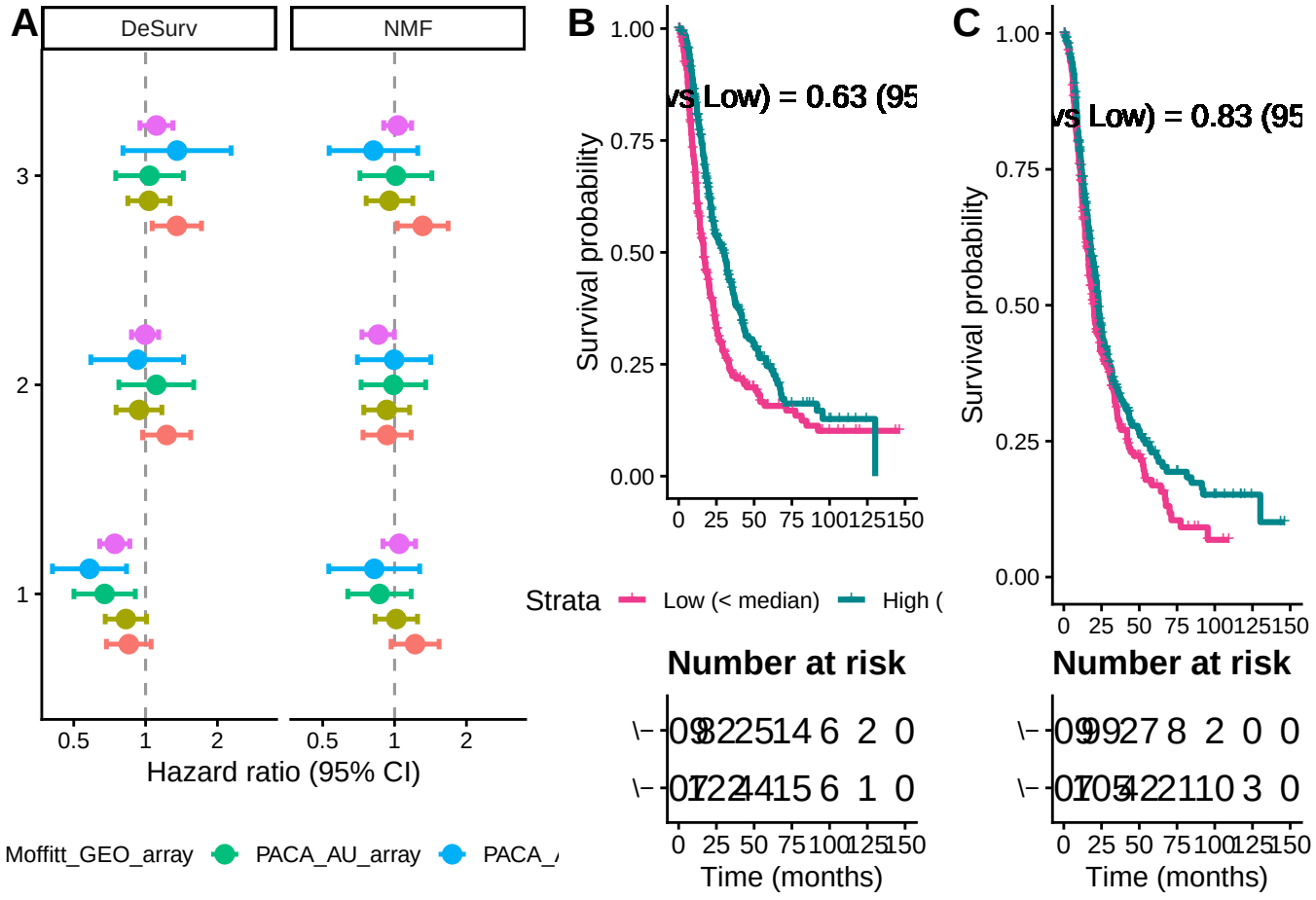

Fig. 5. DeSurv learns prognostic structure that generalizes across independent cohorts. (A) Forest plot summarizing hazard ratios (HRs; 95% CIs) for the latent factor most predictive of survival in held-out validation datasets. Estimates from DeSurv (blue) are more stable across cohorts than those obtained with standard NMF (red). (B) Kaplan–Meier curves for pooled validation samples stratified into high and low groups based on values of the DeSurv-derived factor providing the strongest survival signal. Median-based stratification yields clear separation of survival trajectories, indicating robust out-of-sample prognostic performance. (C) Corresponding analysis for the NMF-derived factor selected using the same survival-based criterion, showing weaker discrimination between risk groups.

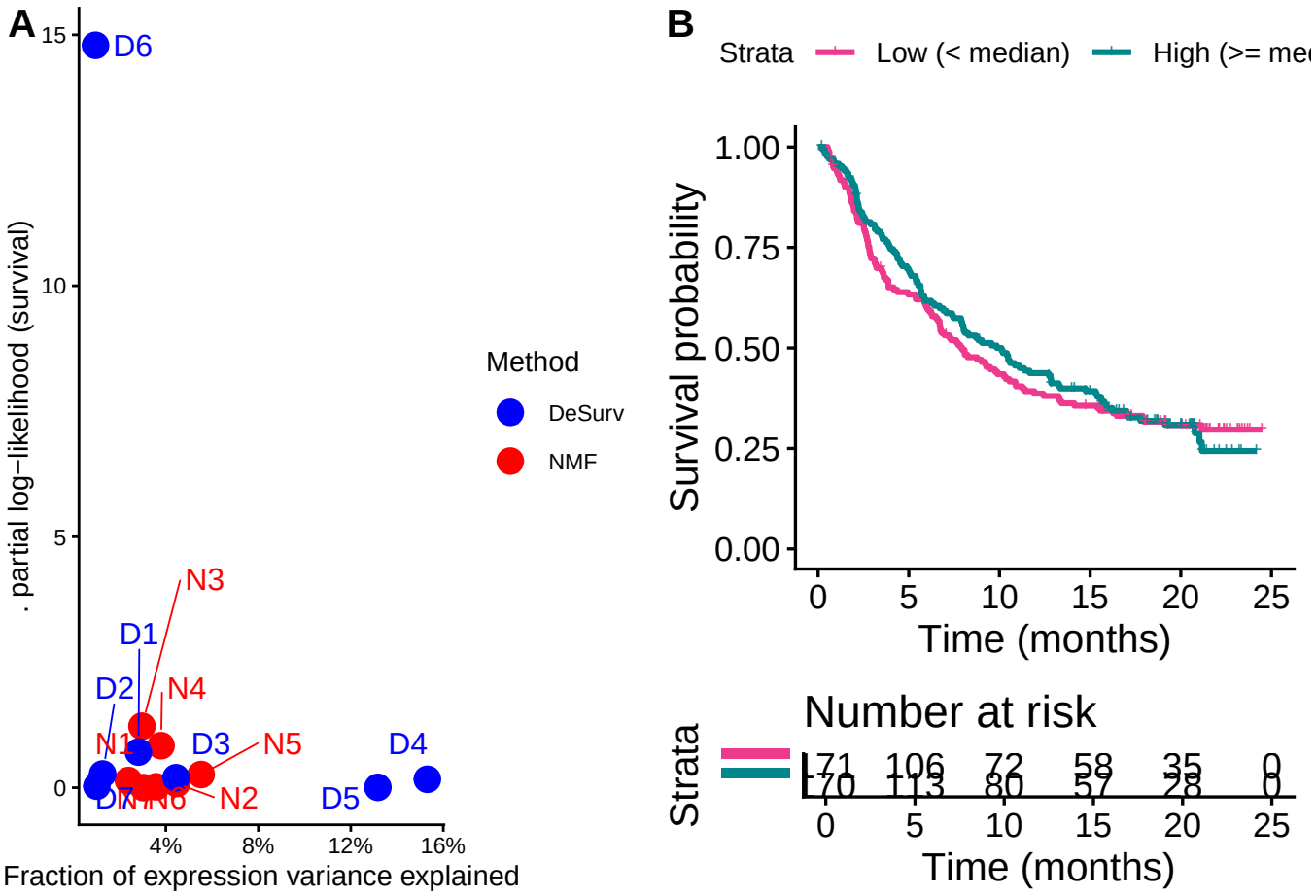

Fig. 6. Survival-associated structure in bladder cancer is recovered by DeSurv and transfers across tumor types.(A) In bladder cancer, DeSurv identifies latent factors (F1, F2) with strong associations to survival, as measured by Cox partial log-likelihood, despite explaining only a small fraction of total expression variance. In contrast, factors derived from standard NMF explain substantially more variance but exhibit weak survival association. (B) Applying gene-level factor definitions learned in pancreatic ductal adenocarcinoma (PDAC) to bladder cancer samples yields clear separation of survival outcomes when patients are stratified by factor score (median split), with numbers at risk shown below. These results indicate that prognostically relevant structure may be weakly aligned with dominant transcriptional variation and can generalize across cancer types.