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 19, 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.

nonnegative matrix factorization | survival analysis | semi-supervised learning | tumor deconvolution | pancreatic cancer

Predicting which cancer patients will respond to therapy or progress rapidly requires identifying the transcriptional signatures that drive clinical outcomes (1). Yet in bulk RNA-sequencing, which currently provides the large clinically annotated cohorts required for survival modeling, the observed expression profile of each sample reflects contributions from malignant cells, cancer-associated fibroblasts, immune infiltrates, and other microenvironmental components (2). Nonnegative matrix factorization (NMF) has been instrumental in resolving this mixture into additive, interpretable gene programs corresponding to recognizable cell types or transcriptional states (3–5), yielding important biological insights including the identification of basal-like and classical tumor programs in pancreatic ductal adenocarcinoma (PDAC) (1, 4) and compartment-specific deconvolution across 33 cancer types (5).

However, the standard approach discovers these programs through unsupervised factorization and only then evaluates their clinical relevance retrospectively (6, 7). Because unsupervised NMF minimizes reconstruction error, its factors tend to be dominated by the highest-variance patterns in the data, which may reflect tissue composition or other outcome-neutral sources rather than prognostic biology (8, 9). In PDAC, exocrine and tissue-composition signals dominate expression variance but contribute little to survival prediction, while lower-variance basal-like and activated stromal programs carry the strongest prognostic associations (4, 10, 11); programs that are prognostically relevant but explain modest variance can be diluted across multiple factors or missed entirely. This misalignment between variance and prognosis extends beyond PDAC: tumor purity alone accounts for a large fraction of expression variation across cancer types (12), yet the objective optimized during unsupervised discovery (reconstruction error) differs fundamentally from the criterion used during evaluation (survival association) (9). Identifying the prognostically relevant subset then requires extensive downstream filtering that is ad hoc, cohort-specific, and difficult to reproduce (13). Over a decade of PDAC subtyping efforts proposed between two and six subtypes (13, 14), ultimately converging on a robust basal/classical dichotomy only after extensive retrospective evaluation across independent cohorts (1, 7, 10).

Whether incorporating survival information during factorization improves generalizability, or risks overfitting to cohort-specific outcome distributions, remains an open empirical question. Sufficient dimension reduction theory establishes that response-guided subspace estimation targets the directions most relevant to the outcome, whereas variance-maximizing projections can miss outcome-relevant structure entirely (9, 15). When the outcome depends on cell populations that are compositionally minor or contribute low variance to bulk expression, unsupervised methods systematically miss these features, while outcome-supervised methods can recover them (8, 15, 16). Supervision during factorization would filter out variance-dominant but prognostically neutral signals during learning rather than in a separate post hoc step, but whether this theoretical advantage translates to improved generalization in the NMF deconvolution setting, where nonnegative constraints, censored survival outcomes, and cohort-specific variation introduce additional challenges, has not been empirically tested.

Here we present DeSurv, a survival-supervised deconvolution framework that integrates NMF with Cox proportional

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

hazards modeling. The key architectural choice is where survival supervision enters the factorization. In DeSurv, factor scores are defined as $Z = W^\top X$, so the Cox partial likelihood is a function of the gene program matrix W and regression coefficients β, and the survival gradient acts directly on gene programs. Sample-level loadings H are updated solely through the reconstruction objective and receive no survival gradient, preserving their interpretation as mixture coefficients (3, 17) and the biological interpretability of the deconvolution. DeSurv operates as a semi-supervised method: a parameter α balances reconstruction fidelity $(1-\alpha)$ and survival prediction (α). Because α is selected via cross-validated concordance, values large enough to overfit to cohort-specific survival patterns are penalized by poor out-of-sample performance, preventing the survival term from overwhelming the factorization. Because W defines shared transcriptomic programs, new samples can be scored by projection ($Z_{\text{new}} = W^\top X_{\text{new}}$) without requiring their survival data, a property not shared by methods that route supervision through H (18, 19). Hyperparameters, including factorization rank k and α, are jointly selected via cross-validated concordance (Methods).

We evaluate DeSurv in three settings that test the predictions above. First, in simulations with known latent structure and survival associations, we show that DeSurv recovers the true factorization rank and the identity of prognostic programs more reliably than standard NMF, that its advantage scales with the degree of divergence between variance and prognosis, and that it vanishes under null conditions where no survival signal exists. Second, in PDAC, we demonstrate that survival supervision reorganizes the learned factor structure: DeSurv suppresses variance-dominant but prognostically neutral signals (e.g., exocrine content) and concentrates survival association into a smaller set of biologically interpretable factors aligned with known tumor and microenvironmental programs. These survival-aligned factors generalize across independent PDAC cohorts with consistent hazard ratios and clearer survival separation than their unsupervised counterparts. Third, we show that a PDAC-trained DeSurv factor retains prognostic signal when projected into bladder cancer samples, consistent with prior reports that basal-like transcriptional structure generalizes across epithelial cancers (20, 21).

Results

Model Overview. We developed DeSurv, a survival-supervised deconvolution framework that jointly optimizes NMF reconstruction and Cox proportional hazards likelihood (Fig. 1; Methods). The survival gradient acts on the gene program matrix W but not on the sample loadings H, directing gene programs toward outcome-relevant structure while preserving interpretability of sample loadings as mixture coefficients (3, 17). We evaluated DeSurv in three settings: controlled simulations, PDAC cohort analysis with external validation, and cross-cancer transfer.

Survival supervision clarifies NMF rank selection. Before examining DeSurv's performance across the three evaluation settings described above, we address a prerequisite question using both PDAC expression data and controlled simulations: whether incorporating clinical outcomes clarifies the well-documented instability of unsupervised rank selection heuristics (6, 22). Using gene expression data from the PDAC training cohorts (TCGA (23) and CPTAC (24); Methods), we first evaluated commonly used unsupervised criteria 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 was highest at very small ranks ($k \approx$ 2-3) and decreased monotonically thereafter, favoring low-dimensional solutions that conflicted with the other criteria. Together, these unsupervised heuristics pointed to incompatible values of k, illustrating the ambiguity of rank selection in standard NMF.

To address this ambiguity, we applied DeSurv, which incorporates survival outcomes directly into the factorization and evaluates models using cross-validated concordance index (C-index). The resulting C-index surface across the joint space of factorization rank (k) and supervision strength (α) identified a well-defined optimum, in contrast to the conflicting recommendations produced by unsupervised heuristics (Fig. 2D). Model selection followed the one-standard-error rule: we selected the smallest k whose predicted performance lay within one standard error of the maximum, yielding a parsimonious choice. Bayesian optimization selected $k = 3$ and $\alpha = 0.3339356$ (C-index 0.655; 1-SE rule; $n = 273$ patients, 139 events).

To further evaluate rank recovery under controlled conditions, we conducted simulation studies in which the true underlying rank was known ($k = 3$). DeSurv consistently selected the correct rank, producing a concentrated distribution of selected k values centered at the true value (Fig. 2E). In contrast, standard NMF followed by post hoc Cox modeling ($\alpha = 0$) exhibited substantially greater variability and a systematic tendency toward under-selection. These results support the prediction that incorporating outcome information into model selection improves recovery of the true factorization rank.

DeSurv recovers prognostic gene programs when variance and prognosis diverge. Sufficient dimension reduction theory predicts that outcome-guided subspace estimation recovers directions most relevant to the response (9, 15), but this prediction has not been evaluated in the NMF deconvolution setting. To test whether it holds under nonnegative constraints with survival outcomes, we designed simulations with known ground-truth latent structure. Simulated expression matrices were generated from a nonnegative factor model in which prognostic gene programs explained low variance relative to outcome-neutral background signals, with survival times generated from marker gene expression ($p = 3{,}000$ genes, $n = 200$ samples, true $k = 3$; Methods). We compared DeSurv to standard NMF ($\alpha = 0$) followed by post hoc Cox regression, with both methods tuned via cross-validated concordance index using Bayesian optimization.

In the primary scenario, where prognostic programs explained low variance relative to outcome-neutral programs, DeSurv consistently achieved higher C-index and substantially improved precision—the fraction of genes in a learned factor

that belong to a true prognostic gene program—for recovering the true prognostic gene programs (Fig. 3A-B). The unsupervised baseline exhibited near-zero precision, indicating that variance-driven factorization failed to concentrate on the genes that actually drove survival. This result supports the prediction that outcome-guided learning recovers prognostic programs more reliably when variance and prognosis diverge. Across 100 replicates, DeSurv achieved median C-index [TODO] versus [TODO] for standard NMF (Δ = [TODO]; $p = 3{,}000$ genes, $n = 200$ samples, true $k = 3$).

To verify that these gains reflect genuine signal recovery rather than overfitting, we repeated the analysis under two additional simulation scenarios (SI Appendix, Fig. S2). In a null scenario where survival times were generated independently of gene expression ($\beta = 0$), DeSurv defaulted to standard NMF: Bayesian optimization selected low supervision strength (α), and both methods yielded C-index values near 0.5. This indicates that the semi-supervised design does not impose spurious prognostic structure when none exists. In a mixed scenario where survival depended on both factor-specific marker genes and shared background genes, creating partial overlap between variance-dominant and prognostically relevant structure, DeSurv's advantage was present but attenuated relative to the primary scenario. Together, these three scenarios establish that DeSurv's benefit scales with the degree of divergence between variance and prognosis: largest when prognostic programs explain low variance (primary scenario), moderate when variance and prognosis partially overlap (mixed), and absent when no survival signal exists (null).

Survival supervision reorganizes the learned factor structure in PDAC. The dominant source of expression variance in PDAC, exocrine content, is not the dominant source of prognostic signal; DeSurv reorganizes the learned factor structure accordingly.

To directly test whether survival supervision reorganizes the learned factor structure, as predicted by the simulation results above, we examined the overlap between factor-specific gene rankings and established PDAC gene programs (Fig. 4A-B). Bayesian optimization selected $k = 3$ and $\alpha = 0.33$ for DeSurv in the PDAC training cohorts (TCGA and CPTAC). To enable a direct comparison of how each method organizes the same number of factors, we also fit standard NMF at the same rank ($k = 3$); results for standard NMF at independently selected ranks ($k = 5$ via elbow detection, $k = 7$ via cross-validated concordance at $\alpha = 0$) are presented in the SI Appendix.

DeSurv produced a factorization in which each factor aligned with a distinct, prognostically relevant biological program (Fig. 4A). One factor was enriched for classical tumor programs relative to basal-like expression, a second isolated an activated microenvironmental state characterized by immune infiltration and stromal remodeling, and a third captured aggressive tumor-intrinsic programs associated with basal-like biology. Notably, exocrine-associated expression did not dominate any DeSurv factor, suggesting that survival supervision deprioritizes differentiation-related variation that is weakly associated with outcome.

At the same rank, standard NMF organized the three factors around dominant sources of transcriptional variance rather than prognostic biology (Fig. 4B). One factor was strongly associated with exocrine expression and negatively correlated with 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 composite microenvironmental factor. These patterns indicate that, given the same number of factors, standard NMF allocates substantial model capacity to differentiation-driven signals and merges distinct microenvironmental states that DeSurv separates.

We quantified these differences by contrasting the fraction of expression variance explained by each factor with its contribution to survival (Fig. 4C). The NMF 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 survival association despite accounting for a smaller fraction of expression variance. Specifically, the highest-variance NMF factor explained [TODO]% of expression variance but contributed $\Delta\ell$ = [TODO] to survival, whereas the DeSurv factor with the largest survival contribution explained [TODO]% of variance with $\Delta\ell$ = [TODO]. The remaining DeSurv factors contributed minimal survival signal, indicating that survival-relevant information was not diffusely distributed across factors. This result directly illustrates the misalignment between variance and prognosis anticipated by sufficient dimension reduction theory (9) and observed empirically in tumor purity analyses (12): the factors that explain the most transcriptomic variation need not be the factors most associated with patient outcomes. The observed pattern in PDAC, where the highest-variance factor shows minimal survival association while other factors show partial overlap between variance and prognosis, is consistent with the intermediate regime between the primary and mixed simulation scenarios, where DeSurv's advantage was largest.

To further characterize how the learned factor structure differs between DeSurv and 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. 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 expression did not correspond strongly to any single DeSurv factor, consistent with the suppression of differentiation- and composition-driven signals under survival supervision. Together, these results demonstrate that DeSurv selectively preserves tumor and microenvironmental structure while reorganizing or deprioritizing patterns of variation that contribute little to survival.

Survival-aligned programs generalize across independent PDAC cohorts. Having shown that DeSurv reorganizes the learned factor structure in the training data, we next tested whether this learned factor structure generalizes to independent cohorts. If outcome-guided subspaces capture biologically reproducible rather than noise-driven structure, they should transfer more readily across datasets (9, 15). We evaluated whether DeSurv factors maintain their prognostic associations in independent PDAC cohorts. We computed factor scores

in each validation dataset by projecting the gene programs learned in the training cohort ($Z = W^\top X_{\text{new}}$), yielding a continuous score for each factor per sample.

For each method, we focused on the factor showing the largest increase in Cox model log partial likelihood in the training data. Across five independent PDAC cohorts ($n =$ [TODO]), the DeSurv-derived factor exhibited consistent survival effects, with hazard ratio estimates showing limited variability and predominantly protective associations (pooled HR [TODO]; 95% CI [TODO]; log-rank $P =$ [TODO]; Fig. 5A). In contrast, the NMF factor identified by the same criterion showed greater heterogeneity across datasets and weaker survival associations, consistent with the expectation that variance-driven programs capture cohort-specific variation that does not transfer.

We pooled validation samples and stratified them into high- and low-score groups based on a median split of the DeSurv-derived factor, observing clear separation of survival trajectories (Fig. 5B). The same procedure applied to NMF yielded weaker survival stratification (Fig. 5C). These results support the prediction that outcome-aligned programs capture biology that generalizes, whereas variance-driven programs may include cohort-specific signals that attenuate across datasets.

A PDAC-trained program retains prognostic signal in bladder cancer. A PDAC-trained DeSurv program retains prognostic relevance in bladder cancer, consistent with shared basal-like transcriptional biology across epithelial cancers.

Finally, we tested whether the misalignment between variance and prognosis extends beyond PDAC and whether survival-aligned programs transfer across cancer types, a prediction motivated by prior evidence that basal-like transcriptional structure generalizes across epithelial cancers (20, 21). We first applied both standard NMF and DeSurv to an independent bladder cancer cohort and evaluated factor structure and survival association (Fig. 6A). We then asked whether the DeSurv model trained in PDAC retains prognostic relevance when projected to bladder cancer samples (Fig. 6B).

In bladder cancer, standard NMF again organized the learned factor structure primarily around variance-dominant patterns (Fig. 6A). As in PDAC, NMF factors explained substantial fractions of expression variance but contributed little to survival, replicating the misalignment between variance and prognosis observed in PDAC.

Separately, we projected the DeSurv model trained in PDAC directly onto bladder cancer samples to test cross-cancer transfer. 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 ($n =$ [TODO], [TODO] events; log-rank $P =$ [TODO]; HR [TODO]; 95% CI [TODO]). This finding suggests that survival supervision captures cross-cancer biology that unsupervised methods may distribute across multiple outcome-neutral factors.

Together, these results suggest that survival supervision separates variance-dominant from survival-relevant structure and that this separation can transfer across cancer types, providing initial evidence that outcome-guided dimension reduction targets different subspaces than variance-driven reduction.

Discussion

We have shown that incorporating survival information during NMF-based deconvolution reorganizes the learned factor structure, concentrating prognostic signal into fewer, more interpretable gene programs while suppressing variance-dominant but outcome-neutral structure. This reorganization produces a more parsimonious representation of prognostically relevant biology: in PDAC, DeSurv achieved with three factors the same cross-validated concordance that standard NMF required seven to eleven factors to match, and even at each method's unrestricted optimum, DeSurv attained a higher concordance with fewer factors. The advantage is most pronounced at low factorization ranks, where standard NMF allocates its limited representational capacity to variance-dominant signals such as exocrine content, leaving little room for prognostic programs to emerge as distinct factors; survival supervision corrects this allocation, enabling recovery of basal-like and microenvironmental programs even at $k = 3$. In simulations, DeSurv recovered the true factorization rank and the identity of prognostic programs more reliably than standard NMF, with the largest advantage occurring when prognostic programs explained modest variance relative to outcome-neutral signals. In PDAC, DeSurv isolated tumor-intrinsic and microenvironmental programs with clear survival associations, and these factors generalized across independent PDAC cohorts. For the dominant DeSurv program, prognostic signal transferred to bladder cancer, supporting the hypothesis that outcome-guided learning captures biology that generalizes more readily than variance-driven structure.

In PDAC, DeSurv recapitulated known tumor and microenvironmental programs, including the basal-like/classical distinction and activated versus normal stromal states. The methodological contribution lies not in the identity of these programs—which have been established through virtual microdissection (4), experimental microdissection (25), and unsupervised deconvolution (5)—but in how they are recovered. DeSurv discovers these programs de novo, without requiring pre-specified signatures or post hoc filtering, and directly quantifies their survival associations during the factorization. The resulting factors align with the consensus that has emerged from over a decade of PDAC subtyping efforts (1, 4, 7, 13), but are identified through a single optimization rather than iterative retrospective evaluation. The cross-cancer transfer result, in which a PDAC-trained factor stratified bladder cancer survival, is consistent with prior evidence that basal-like transcriptional programs are shared across epithelial cancers (20, 21) and suggests that survival supervision captures this shared biology more directly than variance-driven approaches.

The semi-supervised design of DeSurv reflects a deliberate balance between biological completeness and clinical relevance. The reconstruction term (weight $1 - \alpha$) penalizes deviation from the observed expression data; the survival term (weight α) directs gene programs toward outcome-relevant structure. This balance interacts with rank selection: survival supervision enables parsimonious factorizations by concentrating prognostic signal into fewer factors, so that the one-standard-error rule applied to cross-validated concordance selects smaller ranks than would be chosen by unsupervised criteria or by optimizing concordance without supervision. In PDAC, the cross-validated concordance surface was relatively flat across $k = 3$–12 for DeSurv, indicating that three supervised factors

captured nearly as much prognostic information as twelve; for standard NMF, concordance increased steadily from $k = 2$ through $k = 7$–11, reflecting the need for additional factors to distribute prognostic signal that supervision would have concentrated. Our simulations provide further guidance on when this tradeoff is most beneficial: DeSurv's advantage is greatest when prognostic programs explain modest variance relative to outcome-neutral signals and vanishes under null conditions where no survival signal exists. DeSurv is therefore not intended to replace unsupervised NMF in all settings. When the goal is exploratory discovery without clinical endpoints, when survival data is sparse or unreliable, or when the application requires comprehensive biological characterization rather than prognostic stratification, unsupervised methods remain appropriate. The α parameter, tuned via Bayesian optimization, lets the data determine the appropriate balance for a given cancer type and cohort size, and the $\alpha = 0$ endpoint recovers standard NMF as a special case.

Several limitations should be noted. First, DeSurv assumes Cox proportional hazards, which may not hold in settings where treatment effects are delayed or non-proportional, such as immunotherapy response; extensions to more flexible survival models (e.g., accelerated failure time models or neural network-based hazard functions) are a natural direction. Second, the computational cost of Bayesian optimization over cross-validated concordance exceeds that of standard NMF and may limit scalability to very large cohorts or rapid iterative analyses; however, hyperparameter selection is performed once per dataset and the final model can then be applied to arbitrarily many validation cohorts by projection. Third, while the cross-cancer transfer result is encouraging, we tested only one transfer pair (PDAC to bladder), and broader benchmarking across cancer types is needed to establish generality. Fourth, DeSurv's convergence guarantee (SI Appendix) applies to an idealized algorithm; the implementation includes practical modifications (backtracking, gradient clamping) that depart from the theoretical analysis, though empirically these do not affect convergence behavior. Fifth, the supervised factorization is optimized with respect to a specific clinical endpoint, and the resulting molecular programs may differ under alternative outcome definitions (e.g., progression-free vs. overall survival). Moreover, because supervision incorporates observed survival, which reflects both tumor biology and treatment received, the learned factors may implicitly encode treatment-response associations present in the training cohort. The Cox component can incorporate additional clinical covariates to adjust for known confounders during optimization, partially mitigating this concern, though residual confounding from unmeasured factors remains possible. This limits transportability to settings with substantially different treatment landscapes and means the biological interpretation of supervised factors should be understood as conditional on the therapeutic context in which they were derived. Finally, as with any supervised method, DeSurv's value depends on the quality of the outcome data: heavily censored, short follow-up, or misannotated survival information will degrade the learned programs. The nested cross-validation framework mitigates overfitting to training labels, but cannot compensate for systematic outcome misspecification.

More broadly, the principle instantiated by DeSurv—that outcome-guided dimensionality reduction targets different subspaces than variance-driven reduction—extends beyond cancer genomics. Sufficient dimension reduction theory (9) and the information bottleneck framework (26) both predict that supervised compression retains outcome-relevant structure while discarding nuisance variation. DeSurv realizes this principle within the specific constraints of NMF deconvolution, where nonnegativity preserves biological interpretability and the factorization structure enables single-sample scoring. The success of this approach in PDAC and bladder cancer suggests that analogous frameworks may be valuable in other settings where high-dimensional measurements contain both signal and nuisance variation—including multi-omics integration, spatial transcriptomics, and electronic health records. Extending DeSurv to these domains, and to alternative outcome models beyond Cox regression, is a natural next step. As single-cell cohorts with clinical annotation grow in size, direct survival modeling at cellular resolution may complement deconvolution-based approaches, though the cohort sizes required for stable survival analysis remain available primarily in bulk expression data.

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 loadings. Additionally, let $y \in \mathbb{R}_{>0}^{n}$ denote patient survival times, $\delta \in \{0, 1\}^n$ the censoring indicators ($\delta_i = 1$ if the event is observed), and $\beta \in \mathbb{R}^k$ the Cox regression coefficients. Survival outcomes are modeled through a Cox proportional hazards model with factor scores $Z = W^\top X \in \mathbb{R}^{k \times n}$ 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 method operates in a semi-supervised framework: the reconstruction loss penalizes deviation from the observed expression data, while the survival term directs gene programs toward outcome-relevant structure. The parameter $\alpha \in [0, 1]$ controls the relative contribution of each objective.

The joint objective is

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

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. A critical architectural choice is where survival supervision enters the factorization. Because factor scores are defined as $Z = W^\top X$, the Cox partial likelihood $\mathcal{L}_{\text{Cox}}$ is a direct function of W and β, and the survival gradient acts explicitly on the gene program matrix W. The sample-level loadings H enter only through the reconstruction term and receive no survival gradient, preserving their interpretation as mixture coefficients (3, 17). When $\alpha = 0$, DeSurv reduces to standard unsupervised NMF. The Cox component also accommodates additional sample-level covariates (e.g., tumor stage or grade) alongside Z, enabling adjustment for known prognostic factors during optimization.

Optimization proceeds by alternating updates for H, W, and β, using multiplicative rules for H (3), 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, with final rank k chosen by the one-standard-error rule (the smallest k whose predicted performance lay within one standard error of the maximum). All model selection was performed entirely within training data using nested cross-validation; no validation cohort data were used during any stage of tuning. Because α is selected via cross-validated concordance, values large enough to overfit to cohort-specific survival patterns are penalized by poor out-of-sample performance, preventing the survival term from overwhelming the factorization. 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 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. Because the learned W is fixed at training time and validation samples are scored by simple projection, no retraining or access to validation survival data is required. Reported external hazard ratios and log-rank statistics therefore reflect purely out-of-sample generalization. To evaluate the quality of DeSurv-derived gene signatures for subtyping, 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 under controlled conditions. 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 expression 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 via Bayesian optimization. Performance was summarized across repeated simulation replicates. We considered three simulation scenarios: a primary scenario where prognostic programs explain low variance relative to outcome-neutral signals, a null scenario with no survival signal, and a mixed scenario where prognostic and variance-dominant programs partially overlap.

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. Software and availability. An R package implementing DeSurv is available 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: 1.0.1
## DeSurv git branch: main
## DeSurv git commit: 370c88aa9a89e0c71507fbb161d226c7e2e4c61f
## Paper git branch: main
## Paper git commit: 662e85c253084798310f3fa1b42b7665233512c6
```

1. Collisson EA, et al. (2011) Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature medicine* 17(4):500–503.

2. 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.

3. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *nature* 401(6755):788–791.

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

5. 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.

6. Brunet J-P, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences* 101(12):4164–4169.

7. Bailey P, Chang DK, et al. (2016) Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 531(7592):47–52.

8. Bair E, Tibshirani R (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS biology* 2(4):e108.

9. Cook RD (2007) Fisher lecture: Dimension reduction in regression. *Statistical Science* 22(1):1–26.

10. Rashid NU, et al. (2020) Purity independent subtyping of tumors (PurIST), a clinically robust, single-sample classifier for tumor subtyping in pancreatic cancer. *Clinical Cancer Research* 26(1):82–92.

11. Peng XL, et al. (2024) Determination of permissive and restraining cancer-associated fibroblast (DeCAF) subtypes. *bioRxiv*.

12. Aran D, Sirota M, Butte AJ (2015) Systematic pan-cancer analysis of tumour purity. *Nature Communications* 6:8971.

13. Collisson EA, Bailey P, Chang DK, Biankin AV (2019) Molecular subtypes of pancreatic cancer. *Nature reviews Gastroenterology & hepatology* 16(4):207–220.

14. Schwarzová L, Bouchal P, Brychtová S, Hrstka R (2023) Stroma-rich bladder cancers: Biological features, clinical relevance, and therapeutic targeting. *Cancers* 15(5):1503.

15. Bair E, Hastie T, Paul D, Tibshirani R (2006) Prediction by supervised principal components. *Journal of the American Statistical Association* 101(473):119–137.

16. Arora A, Olshen AB, Seshan VE, Shen R (2020) SurvClust: An integrative survival-weighted clustering method for multi-omic data. *bioRxiv*. doi:10.1101/2020.09.04.283838.

17. Gaujoux R, Seoighe C (2010) A flexible r package for nonnegative matrix factorization. *BMC bioinformatics* 11(1):367.

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*.

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

20. Damrauer JS, et al. (2014) Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. *Proceedings of the National Academy of Sciences* 111(8):3110–3115.

21. Hoadley KA, Yau C, et al. (2018) Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173(2):291–304.

22. Frigyesi A, Höglund M (2008) Non-negative matrix factorization for the analysis of complex gene expression data: Identification of clinically relevant tumor subtypes. *Cancer Informatics* 6:275–292.

23. 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.

24. Ellis M, et al. (2013) Clinical proteomic tumor analysis consortium (CPTAC): Connecting genomic alterations to cancer biology with proteomics: The NCI clinical proteomic tumor analysis consortium. *Cancer Discov* 3:1108–1112.

25. Maurer C, et al. (2019) Experimental microdissection enables functional harmonisation of pancreatic cancer subtypes. *Gut* 68(6):1034–1043.

26. Tishby N, Pereira FC, Bialek W (1999) The information bottleneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*:368–377.

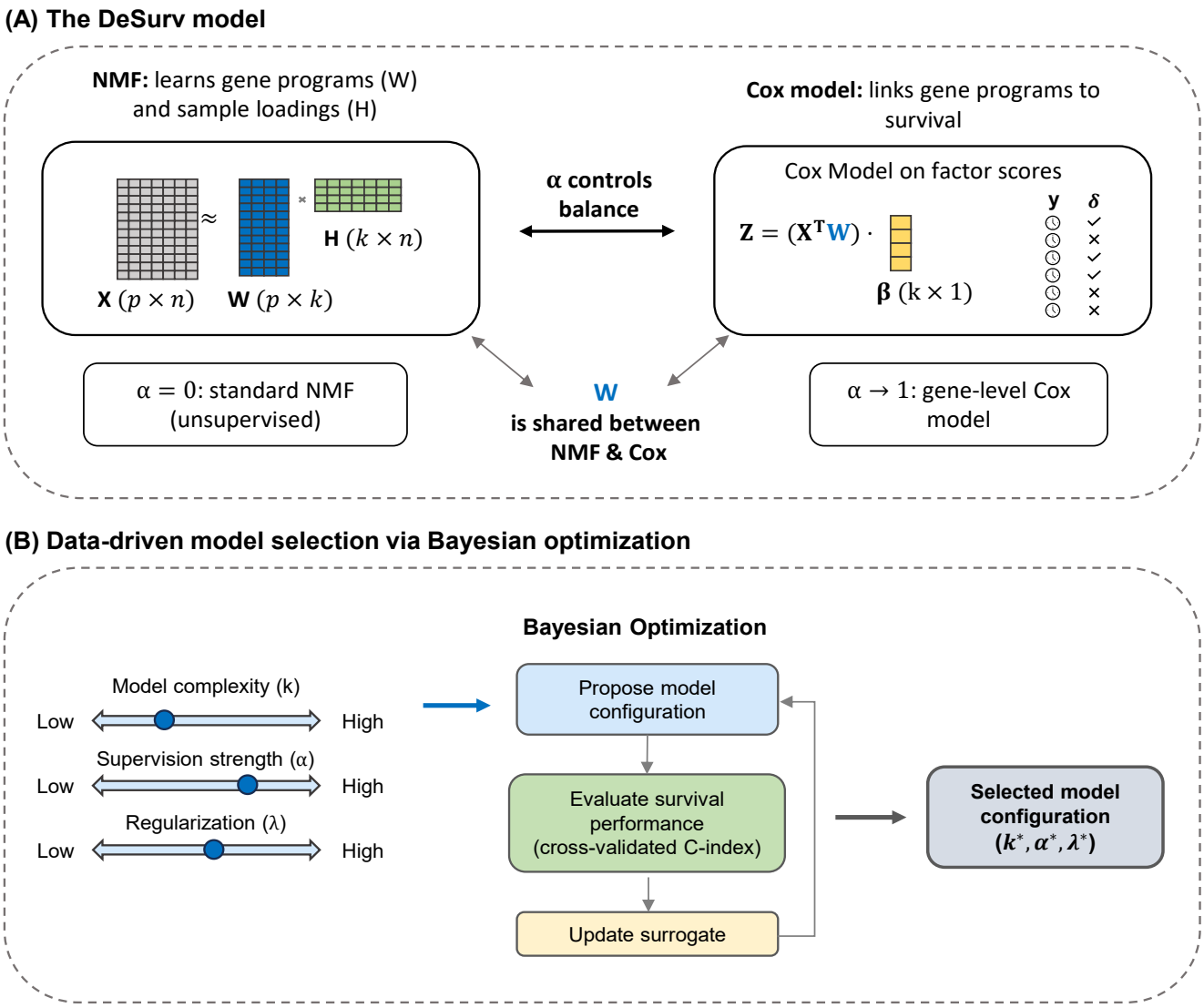

Fig. 1. Overview of the DeSurv framework. (A) DeSurv jointly optimizes NMF reconstruction ($X \approx WH$) and Cox proportional hazards likelihood. The gene program matrix W is shared between objectives: factor scores $Z = W^\top X$ serve as covariates in the Cox model with coefficients β. A supervision parameter α controls the balance between reconstruction fidelity ($\alpha = 0$, standard NMF) and survival prediction ($\alpha > 0$). (B) The factorization rank (k), supervision strength (α), and regularization (λ) are selected via Bayesian optimization over cross-validated concordance index.

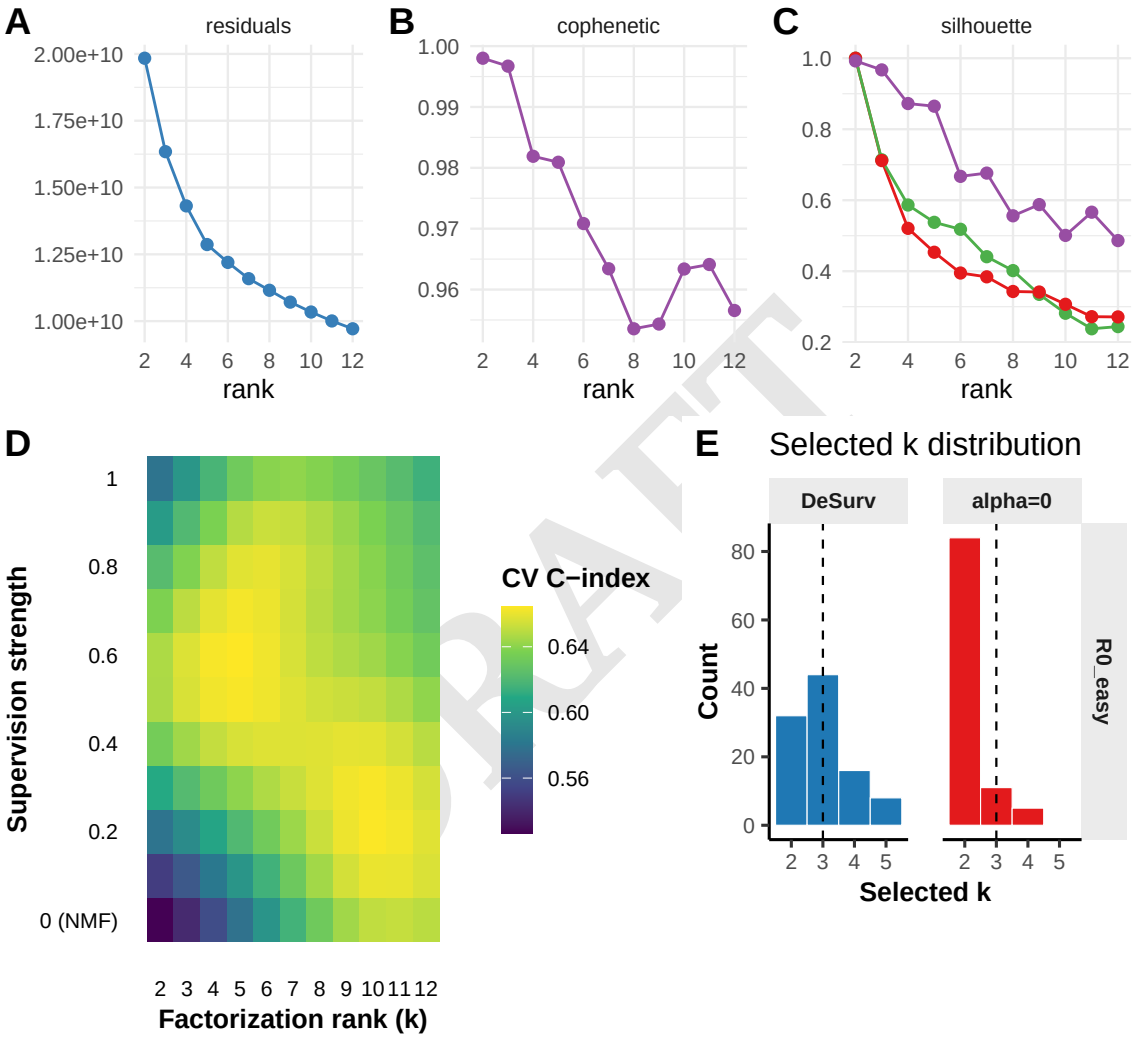

Fig. 2. (A–D) Analyses based on pancreatic ductal adenocarcinoma (PDAC) gene expression data from TCGA and CPTAC cohorts. (A–C) Standard unsupervised rank selection heuristics yield inconsistent guidance for selecting the factorization rank k. (A) Reconstruction residuals as a function of k. (B) Cophenetic correlation coefficient as a function of k. (C) Mean silhouette width across multiple distance metrics as a function of k. (D) Gaussian process predicted mean cross-validated concordance index (C-index) from Bayesian optimization over the joint space of factorization rank (k) and supervision strength (α). (E) Simulation studies with known rank ($k = 3$): distribution of selected k values across repeated replicates for DeSurv versus standard NMF with post hoc Cox modeling ($\alpha = 0$).

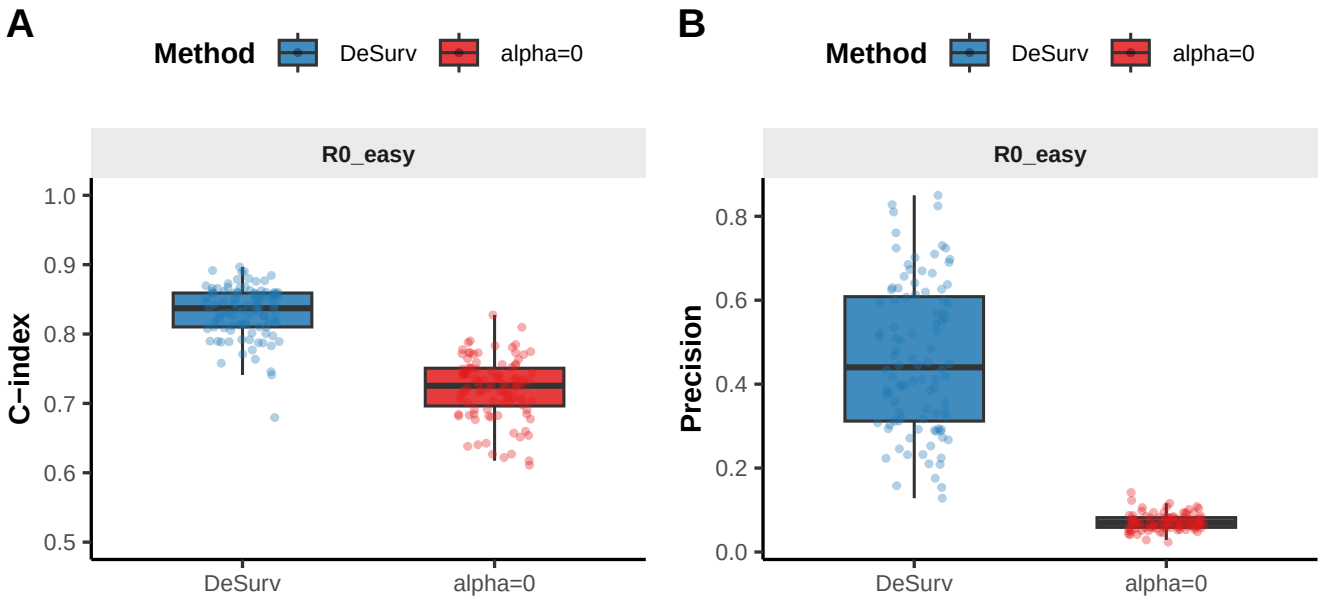

Fig. 3. Performance comparison between DeSurv and standard NMF ($\alpha = 0$) in the primary simulation scenario ($p = 3{,}000$ genes, $n = 200$ samples, true $k = 3$, 100 replicates), where prognostic programs explain low variance relative to outcome-neutral programs. Both methods were tuned via Bayesian optimization over cross-validated concordance index. (A) Test-set concordance index across simulation replicates. (B) Precision (fraction of genes in a learned factor belonging to a true prognostic gene program) across simulation replicates. Each point represents one replicate.

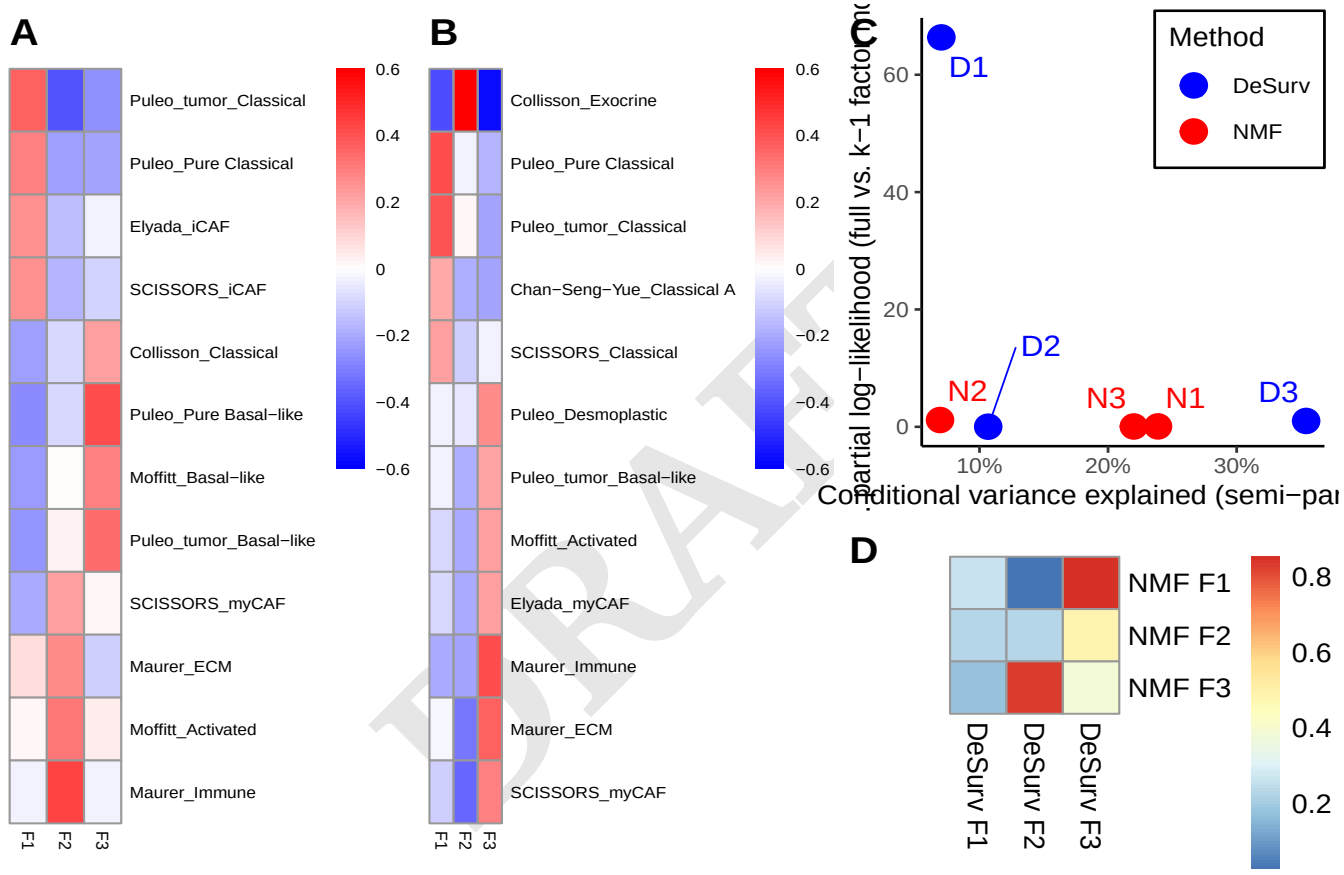

Fig. 4. Survival supervision reorganizes the learned factor structure relative to standard NMF in PDAC training data (TCGA and CPTAC). Both methods are compared at factorization rank $k = 3$, selected by DeSurv's Bayesian optimization; standard NMF at independently selected ranks is shown in SI Appendix. (A) Correlation of DeSurv factor gene rankings with established PDAC gene programs. (B) Corresponding correlations for standard NMF factors. Asterisks indicate significant correlations after multiple testing correction; only gene programs with $|r| > 0.2$ shown. (C) Fraction of expression variance explained versus survival contribution for each factor, quantified by the change in partial log-likelihood from univariate Cox models. (D) Pairwise correlation between NMF and DeSurv factor gene rankings, showing which biological programs are preserved, reorganized, or suppressed under survival supervision.

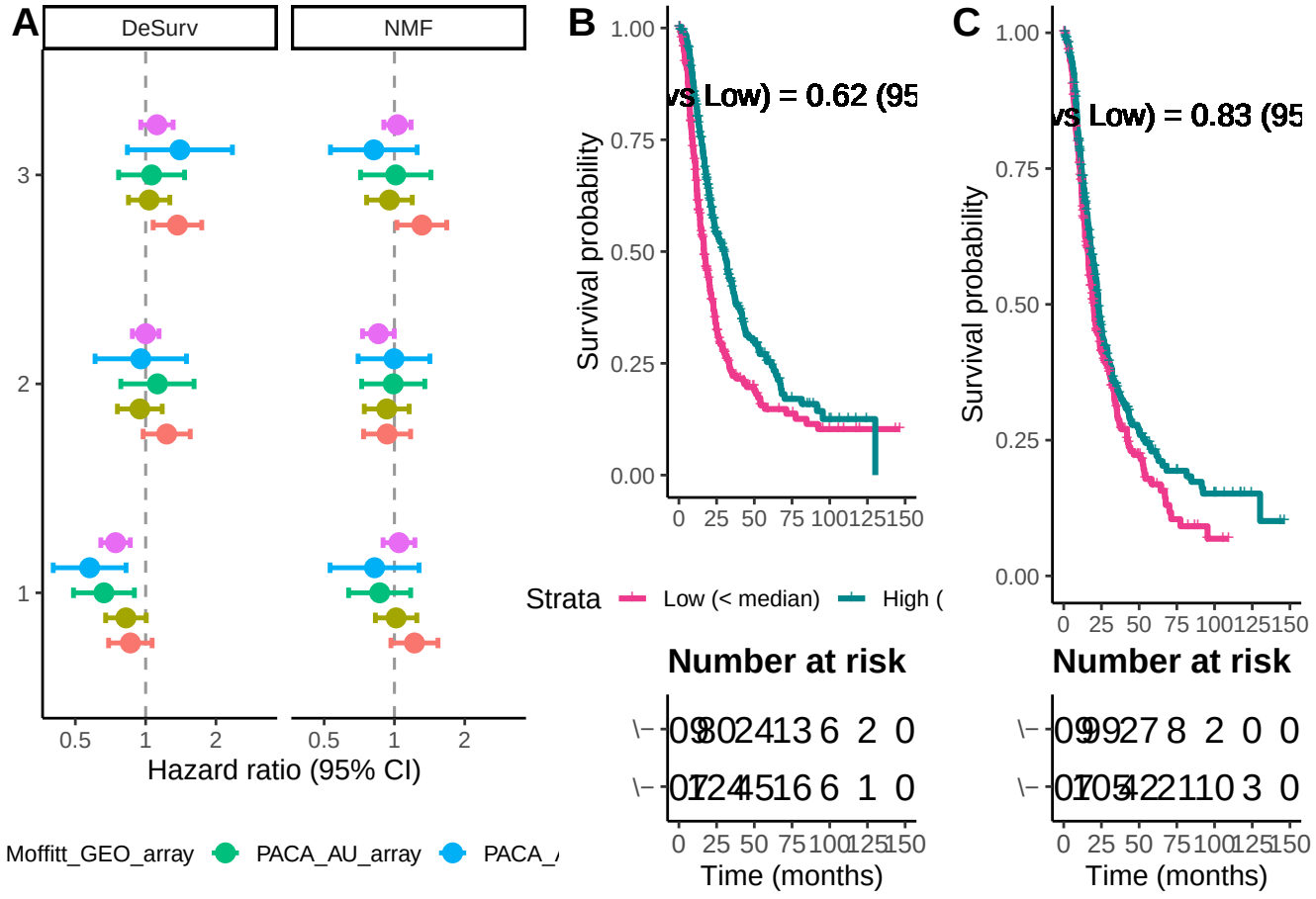

Fig. 5. External validation of DeSurv and standard NMF prognostic factors in independent PDAC cohorts, both at factorization rank $k = 3$. (A) Forest plot of hazard ratios (HRs; 95% CIs) for the factor with the largest training-set Cox partial log-likelihood contribution, evaluated in five held-out PDAC cohorts. DeSurv (blue) and standard NMF (red). (B) Kaplan–Meier curves for pooled validation samples stratified by median split of the DeSurv-derived factor score. (C) Corresponding median-split stratification for the standard NMF-derived factor.

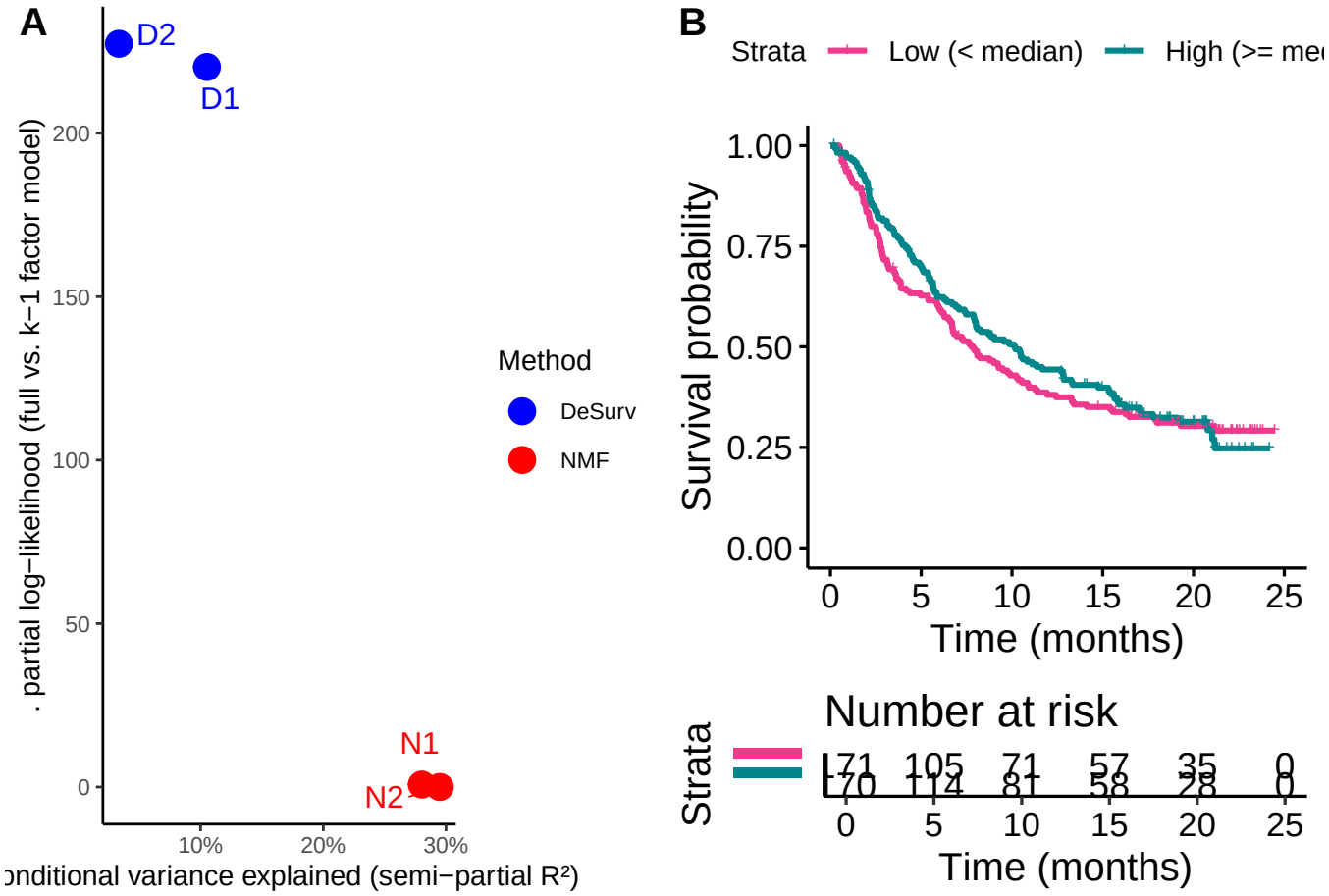

Fig. 6. Survival-associated factor structure in bladder cancer and cross-cancer transfer from PDAC. (A) Fraction of expression variance explained versus survival contribution (change in Cox partial log-likelihood) for each factor in bladder cancer, comparing DeSurv and standard NMF. (B) Kaplan–Meier curves for bladder cancer samples stratified by median split of projected factor scores from a PDAC-trained DeSurv model, with numbers at risk shown below.