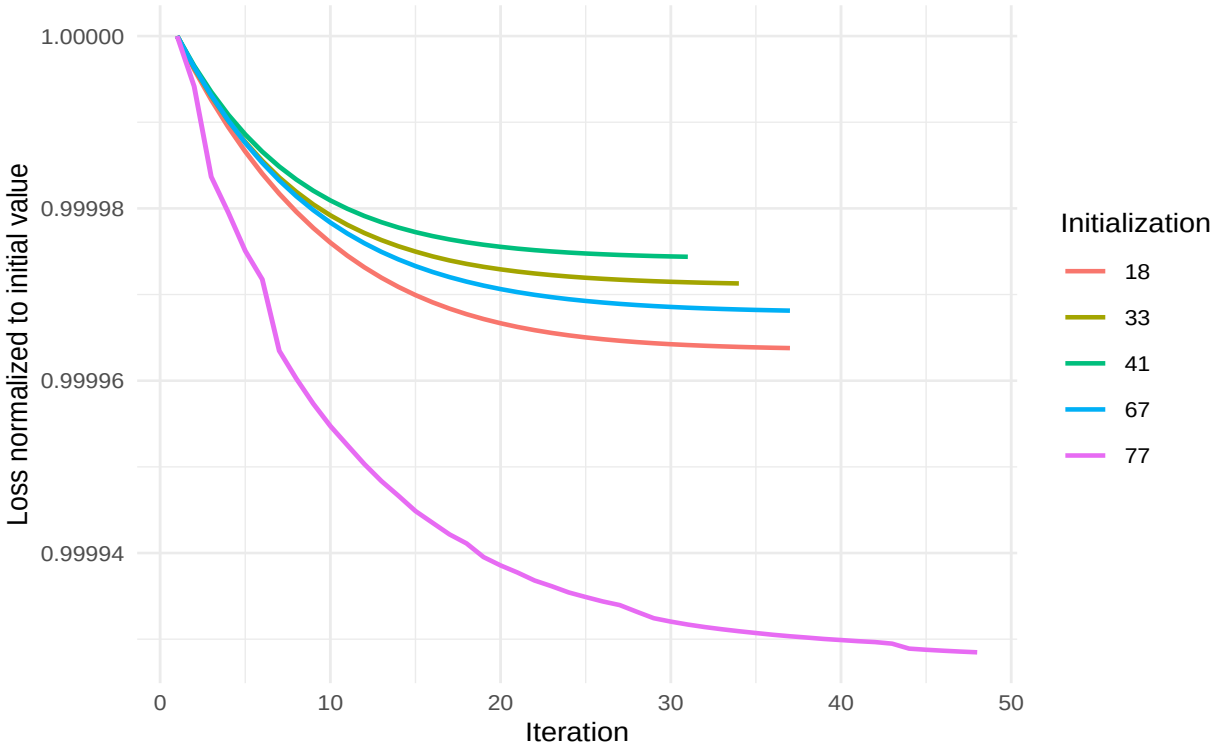

Figure 1: Convergence of model across initializations.

The NMF–Cox model shows consistent convergence and stability across restarts

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

Simulation results under null and mixed scenarios

To characterize the operating characteristics of DeSurv across different signal regimes, we evaluated performance under three simulation scenarios. In the primary scenario (R0_easy; main text Fig. 3), prognostic programs explained low variance relative to outcome-neutral background signals, and survival depended on 150 factor-specific marker genes ($\beta_1 = 2$, $\beta_2 = \beta_3 = 0$). In the null scenario (R00_null), survival times were generated independently of gene expression ($\beta = 0$), so no prognostic structure existed. In the mixed scenario (R_mixed), survival depended on 300 genes per lethal factor, of which 50% were factor-specific markers and 50% were background genes with loadings shared across all factors. All scenarios used $G = 3{,}000$ genes, $N = 200$ samples, $K = 3$ true factors, and identical gamma distribution parameters for the gene loading matrix W (see Materials and Methods). Performance was evaluated using cross-validated concordance index (C-index), precision of recovered prognostic gene sets, and the distribution of selected ranks across simulation replicates.

Under the null scenario (Fig. S2A–B), DeSurv showed no advantage over unsupervised NMF ($\alpha = 0$). Both methods yielded C-index values near 0.5, consistent with the absence of survival signal. The distribution

of selected ranks showed similar variability across methods. These results confirm that DeSurv does not hallucinate prognostic structure when none exists: Bayesian optimization selected low values of the supervision parameter α, effectively recovering standard NMF as a special case. This specificity control is important because it establishes that the advantages observed in the primary scenario reflect genuine signal recovery rather than overfitting to noise in the survival labels.

Under the mixed scenario (Fig. S2C–E), DeSurv's advantage over unsupervised NMF was present but attenuated relative to the primary scenario. C-index values were modestly higher for DeSurv, and precision for recovering the true survival gene set—which now included both marker and background genes—showed a smaller improvement compared to the primary scenario. This attenuation is expected: when survival-relevant genes partially overlap with variance-dominant background programs, unsupervised NMF captures some of the prognostic structure incidentally, reducing the marginal benefit of supervision. The mixed scenario thus identifies the intermediate regime where DeSurv offers a moderate but not dramatic improvement over unsupervised approaches.

Together, the three scenarios provide a dose-response relationship: DeSurv's advantage is greatest when variance and prognosis diverge (primary), moderate when they partially overlap (mixed), and absent when no survival signal exists (null). This gradient is consistent with the predictions of sufficient dimension reduction theory and offers practical guidance for users: DeSurv is most beneficial in settings where the dominant transcriptional axes are expected to be prognostically neutral, as is common in cancers with high tumor purity variation or dominant tissue-composition signals.

Cross-validated C-index across factorization rank

To evaluate how the number of latent factors k affects prognostic performance, we performed an exhaustive grid search over $k \in \{2, 3, \ldots, 12\}$ and supervision strength $\alpha \in \{0, 0.05, \ldots, 0.95\}$ using 5-fold cross-validation on the TCGA+CPTAC training cohort, with all genes retained (ntop = ALL). For each k, the best-performing α was selected by maximizing the mean CV C-index. The unsupervised NMF baseline ($\alpha = 0$) was evaluated at the same k values for comparison.

Figure @ref(fig:fig-cindex-by-k) shows the resulting C-index curves for both training (CV) and external validation (pooled across held-out cohorts), with shaded ribbons indicating ± 1 standard error. In training, DeSurv consistently outperforms NMF across all k, confirming that survival supervision improves in-sample concordance. In external validation, DeSurv maintains a comparable or improved C-index relative to NMF, demonstrating that the supervised signal generalizes beyond the training data rather than reflecting overfitting to the survival labels.

Cutpoint selection and validation of dichotomized risk groups

To convert the continuous DeSurv linear predictor into a clinically interpretable binary risk stratification, we selected an optimal z-score cutpoint via cross-validation on the TCGA+CPTAC training cohort. For each candidate cutpoint in $\{-2.0, -1.8, \ldots, 2.0\}$, we computed the mean absolute log-rank z-statistic across held-out folds and selected the cutpoint maximizing this metric (Figure @ref(fig:fig-cutpoint-km)A). We then applied this cutpoint to external validation cohorts by standardizing each patient's linear predictor using the training mean and standard deviation. Figure @ref(fig:fig-cutpoint-km)B shows the resulting Kaplan–Meier curves for the pooled external validation analysis, with a stratified log-rank test accounting for dataset of origin. Panels C–F display per-cohort validation KM curves for Dijk, Moffitt, PACA-AU, and Puleo, respectively. Across all cohorts, the DeSurv-derived high-risk group showed consistently shorter survival, confirming that the training-derived cutpoint generalizes to independent datasets.

Overlap of DeSurv risk groups with known molecular subtypes

To assess whether the DeSurv-derived risk stratification captures biologically meaningful transcriptional programs, we examined the overlap between dichotomized risk groups (High vs Low) and two established

PDAC molecular classifiers: PurIST (Basal-like vs Classical) and DeCAF (proCAF vs restCAF). Figure @ref(fig:fig-subtype-overlap) shows the composition of each risk group across all pooled external validation cohorts, with Fisher's exact test p-values quantifying the association. The high-risk group was significantly enriched for Basal-like (PurIST) and proCAF (DeCAF) subtypes, consistent with the known poor prognosis of these molecular classes. This concordance confirms that DeSurv's survival-driven factorization recovers clinically relevant biology without requiring subtype labels during training.

Analysis of scRNA-seq

Next, we verify our findings at the cellular level using the Elyada scRNA-seq data [@elyada2019cross]. We found that our DeSurv factor 1 signature was expressed primarily in iCAF and B cells, factor 2 was expressed in Acinar and Classical PDAC 2, and factor 3 was expressed primarily in Basal-like PDAC cells (Figure @ref(fig:fig-sc)B-E). This is consistent with our ORA analysis which found that factor 1 was mostly immune and factor 3 was basal-like, and with the overlap we saw with published gene lists that identified factor 1 as iCAF/classical/immune and factor 3 as basal-like. Interestingly, we do not see a large overlap of factor 1 with the classical PDAC cell types despite some overlap with classical gene lists. Need more here...

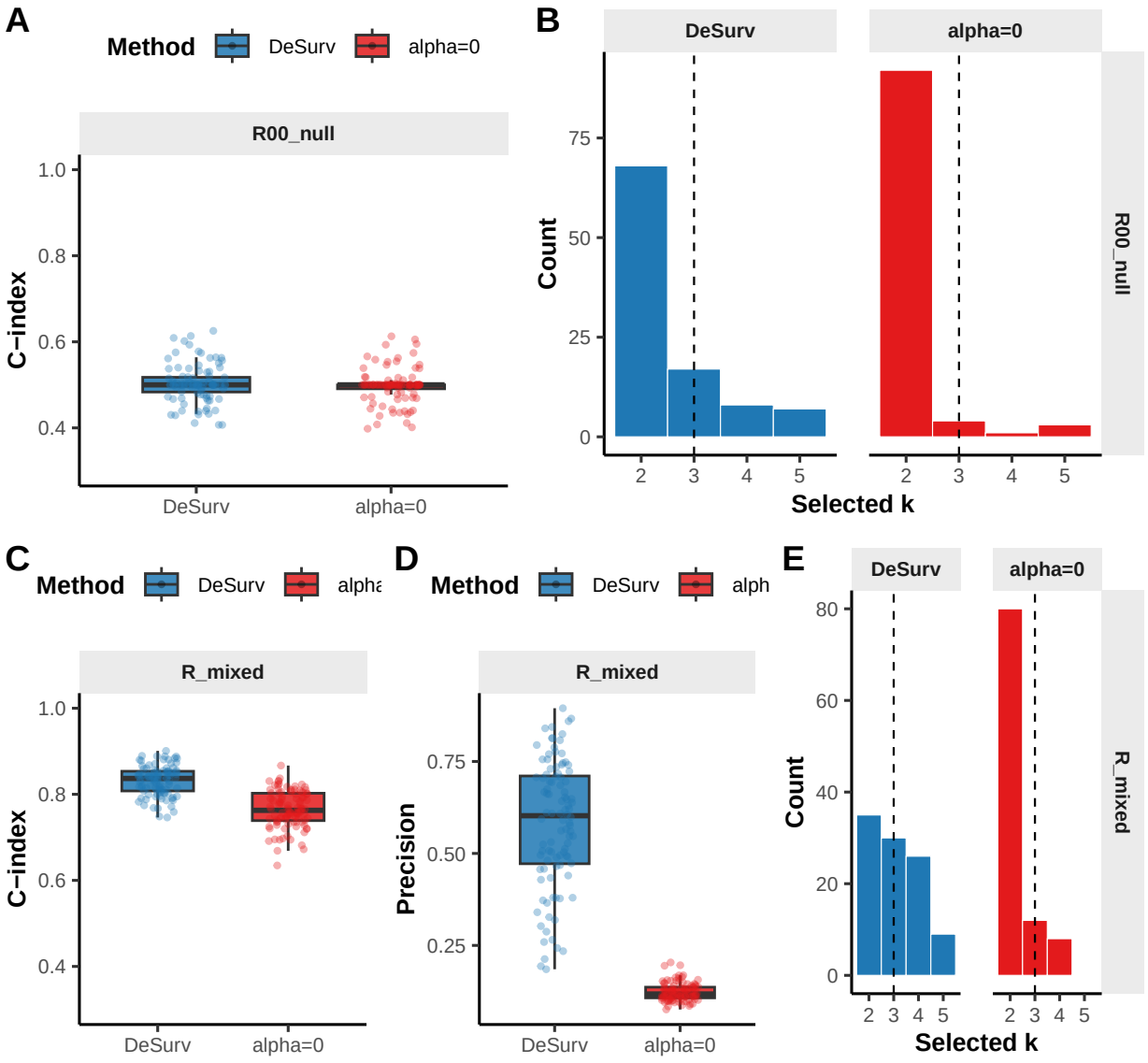

Figure 2: DeSurv performance under null and mixed simulation scenarios. (A–B) Null scenario: survival times are independent of gene expression ($\beta = 0$). DeSurv shows no advantage over unsupervised NMF ($\alpha = 0$), confirming that the method does not hallucinate prognostic structure when none exists. (A) C-index distributions are comparable and centered near 0.5. (B) Distribution of selected ranks shows comparable variability; both methods favor parsimonious solutions. Precision is not shown because no true survival genes exist under the null ($\beta = 0$). (C–E) Mixed scenario: survival depends on a mixture of factor-specific marker genes (50%) and shared background genes (50%), creating partial overlap between variance-dominant and prognostically relevant structure. DeSurv's advantage is present but attenuated relative to the primary scenario (main text Fig. 3). (C) C-index is modestly higher for DeSurv. (D) Precision for recovering the true survival gene set (now including both marker and background genes) remains substantially higher for DeSurv. (E) Selected-rank distributions. In all panels, DeSurv (blue) denotes the supervised model and NMF (red) denotes the unsupervised baseline ($\alpha = 0$), both tuned via Bayesian optimization.

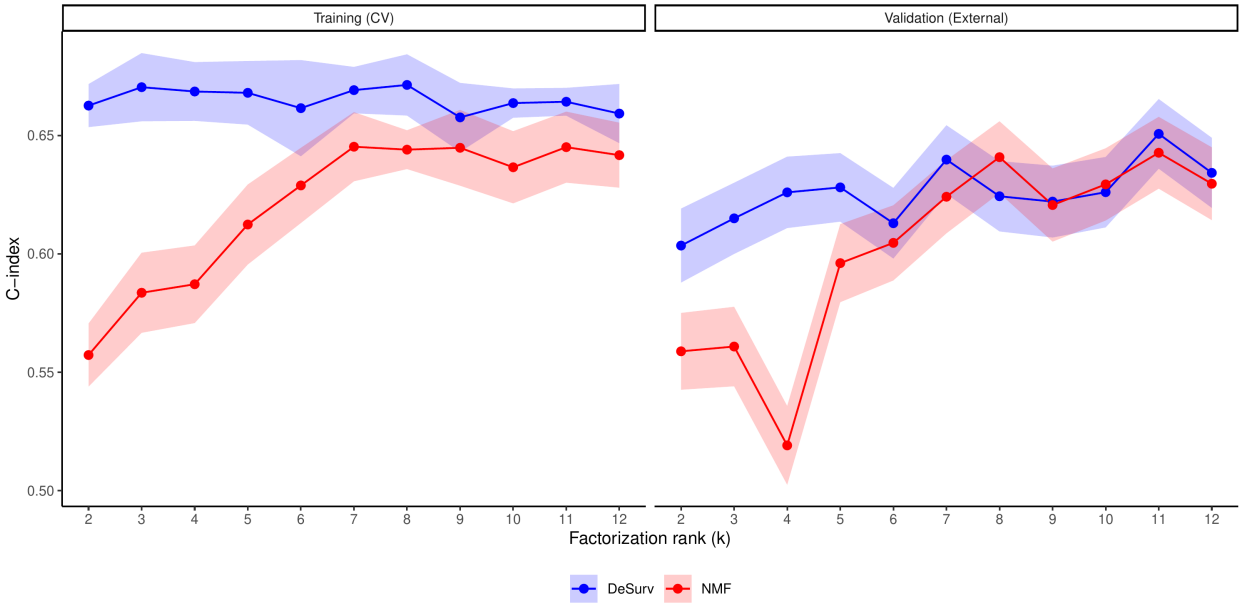

Figure 3: Training (CV) and external validation C-index as a function of factorization rank k for DeSurv (blue) and standard NMF (red, $\alpha = 0$). Shaded ribbons show ± 1 SE. For each k, the best α was selected by maximizing the mean CV C-index. Validation C-index is pooled across all external cohorts.

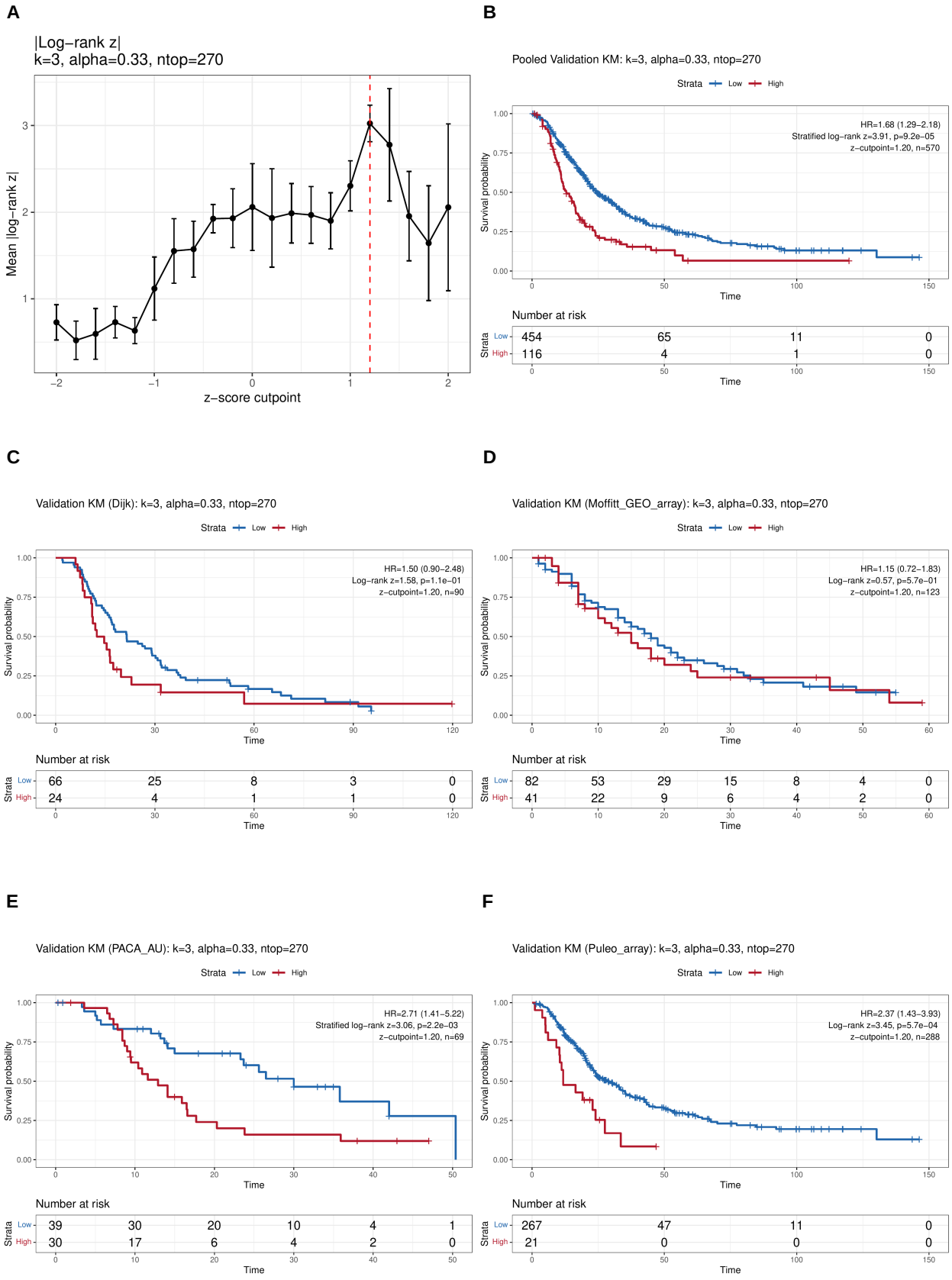

Figure 4: Cutpoint selection and external validation of dichotomized risk groups. (A) Cross-validated mean absolute log-rank z-statistic as a function of z-score cutpoint; red dashed line indicates the selected optimum. (B) Pooled validation Kaplan–Meier curves (stratified log-rank test). (C–D) Per-cohort validation KM curves: (C) Dijk, (D) Moffitt. (E–F) Per-cohort validation KM curves: (E) PACA-AU, (F) Puleo.

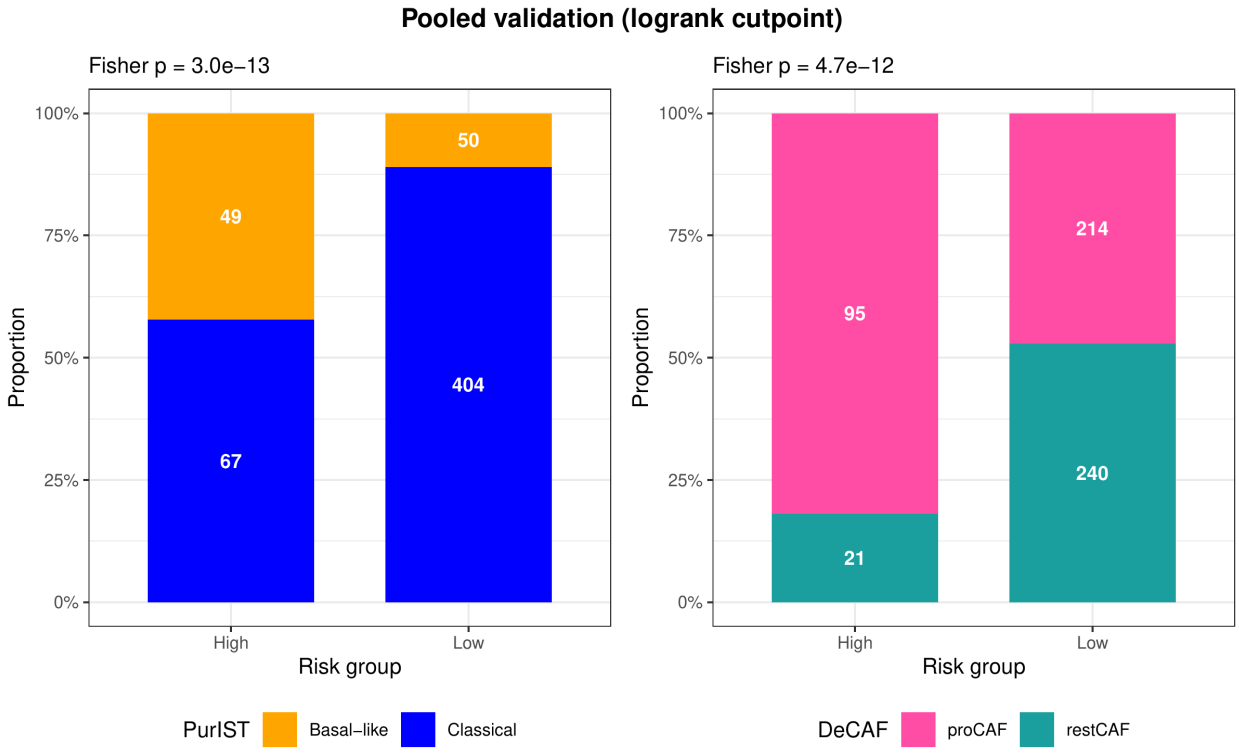

Figure 5: Subtype composition of DeSurv risk groups in pooled external validation cohorts. Stacked bar charts show the proportion of PurIST (left) and DeCAF (right) subtypes within Low and High risk groups defined by the log-rank-optimized cutpoint. Numbers indicate sample counts; Fisher's exact test p-values are shown above each panel.

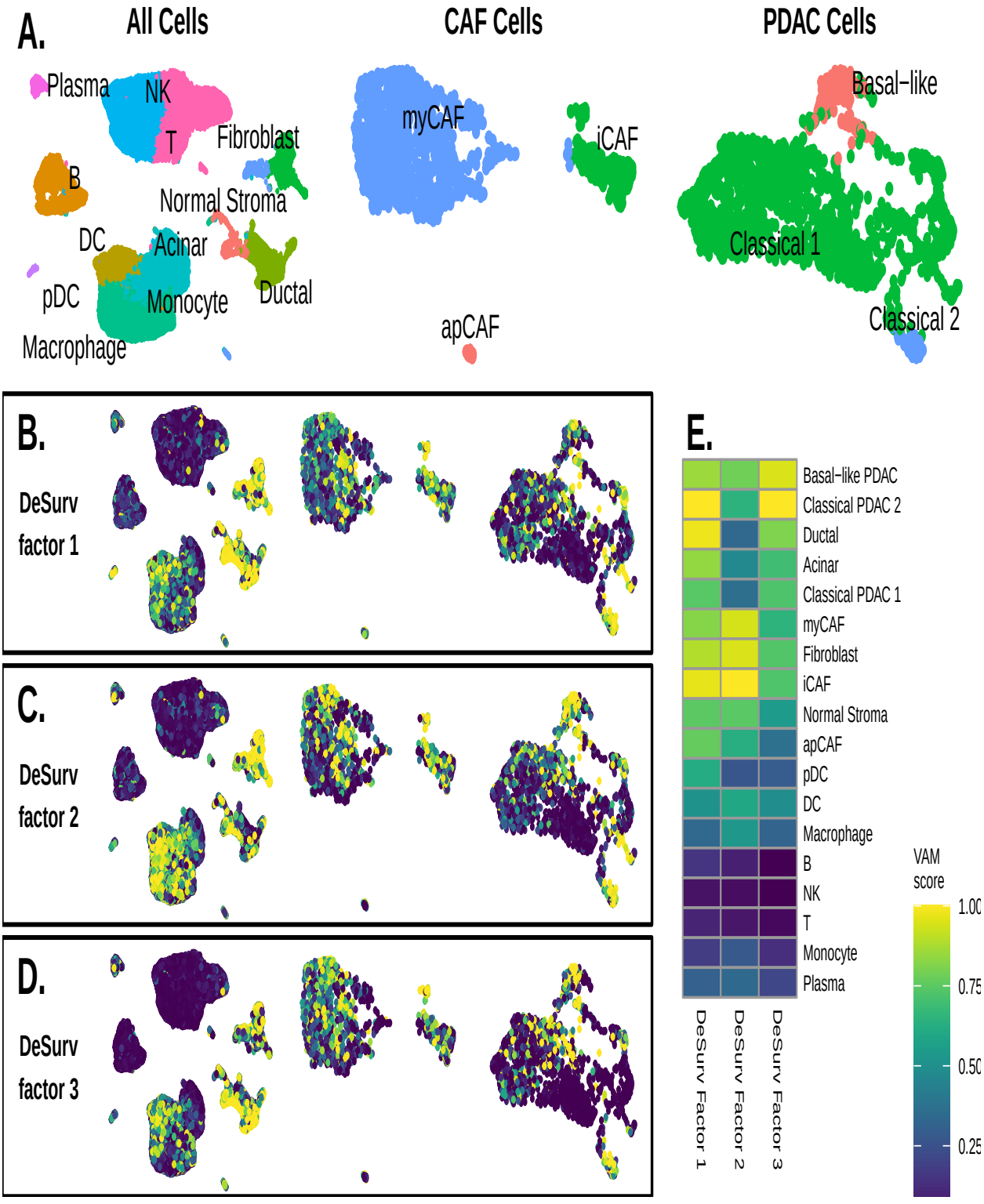

Figure 6: A. Cell type clusters from the Elyada scRNA-seq data. B-D. VAM scores for the DeSurv factor signatures in each cell shown on the UMAP of cells in the Elyada-sc data. E. Heatmap of average VAM scores by cell type and factor signature in the Elyada-sc data.