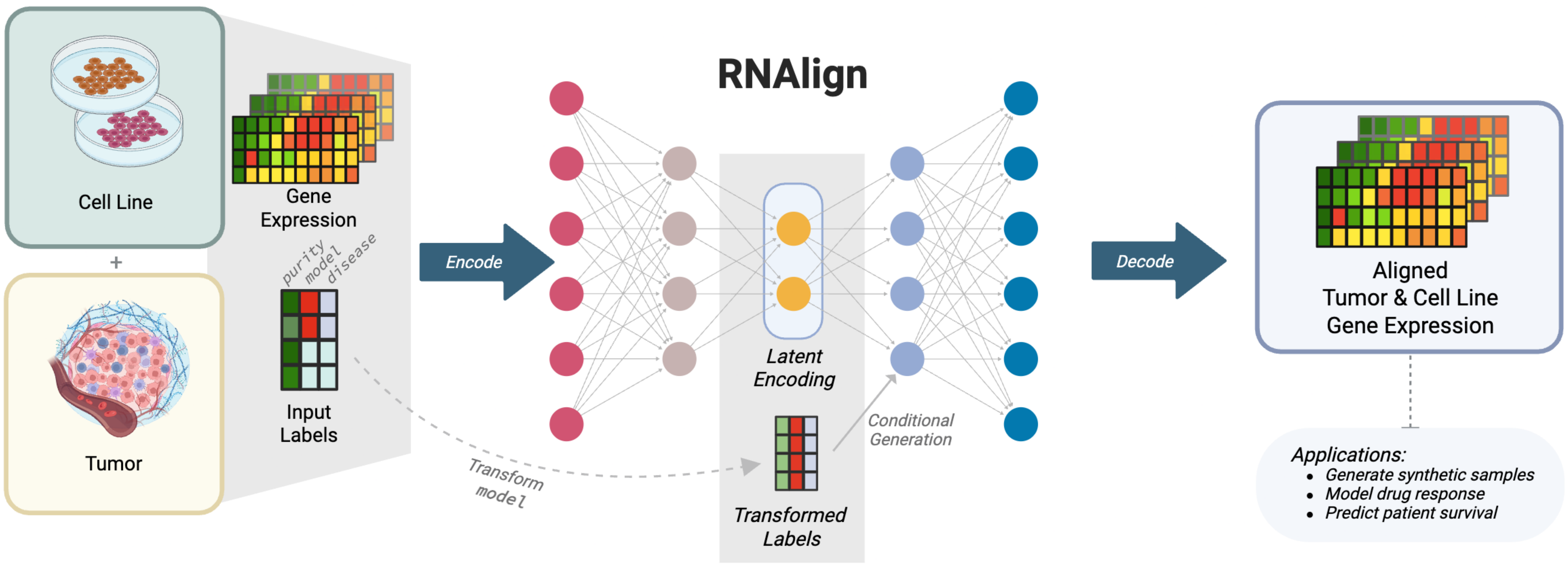
RNALIGN : ALIGNMENT OF TUMOR AND CELL LINE TRANSCRIPTOMES USING CONDITIONAL VAES

Jacob Alvarez[1,2], Kiran Krishnamachari[1], Anders Skanderup[1,2]

1. Computational Cancer Genomics, Agency for Science Technology and Research, Genome Institute of Singapore
2. School of Computing, National University of Singapore, COM1, 13, Computing Dr, 117417
jalvarez,kiran_krishnamachari,skanderupamj@gis.a-star.edu.sg

AI for Science (AI 4 X) Conference 2025



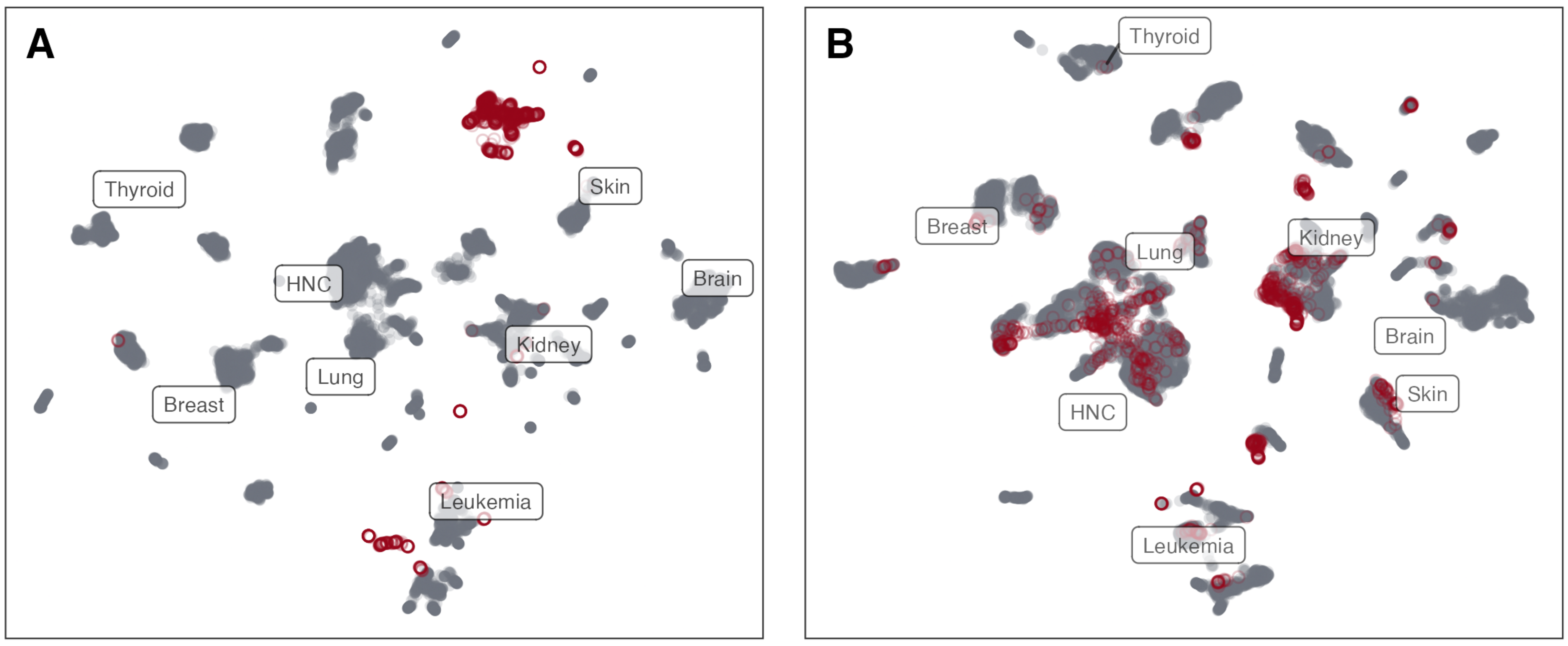
Applications:
- *Generate synthetic samples*
- *Model drug response*
- *Predict patient survival*

To harmonize innately discordant cell line and tumor gene expression data, RNAlign is a CVAE trained on cell line and tumor data, conditioned with class labels. Two novel regularisation terms are introduced - distance correlation loss $\mathbf{L_{cor}}$ regularizes the latent space to be independent from class labels, and a gradient-based loss $\mathbf{L_{grad}}$ on the ELBO with respect to class labels, to increase decoder sensitivity to class labels.

RESULTS

Data Integration: RNAlign was trained with samples from TCGA and CCLE, using purity, sample type, cancer type as class labels.

Enhanced Alignment: Transformed data enhances clustering of tumors and CLs by cancer type (B). To align, we decode with homogenized class labels – model set to 'CL' and purity set to 1.

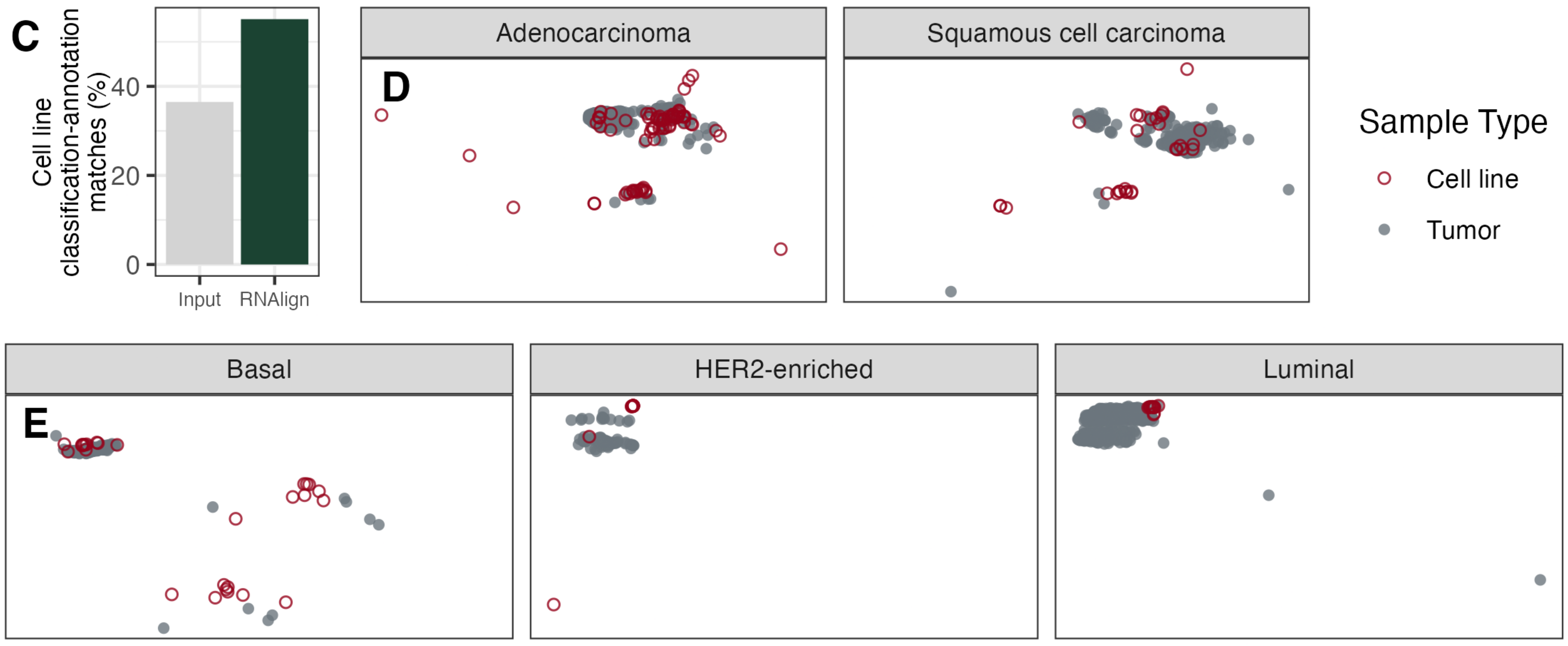


*2D UMAP representation of 12,236 tumors and 1,249 CL pan-cancer samples used for training RNAlign before **(A)** and after **(B)** RNAlign transformation.*

Quantified Improvement: CL-tumor match increases from 36.5% to 55.1% (C). RNAlign is effective at removing sample type related variation (Table 1).

Subtype Preservation: Retains intra-disease heterogeneity in NSCLC (D) and BRCA (E), resulting in subtype alignment of unsupervised factors.

Novel regularization terms: $\mathbf{L_{cor}}$ and $\mathbf{L_{grad}}$ are key to disentanglement and improves conditional generation (Table 2)



*Median percentage of CL samples clustering around tumors of same cancer type **(C)**. 2D UMAP representation of transformed non-small cell lung cancer **(D)** and breast cancer **(E)** samples across **unsupervised** subtype information*

Method	ΔD	PVCA	ΔkBET
Input	22.56	0.27	0.90
Linear Projection	10.57	0.16	0.91
Celligner	8.59	**0.10**	0.86
RNAlign	**4.75**	0.13	**0.65**

***Table 1** : RNAlign tops cancer-type batch effect removal metrics.*

Method	ΔD
Input data	22.56
RNAlign	**4.75**
RNAlign (*no* L_{grad})	6.12
RNAlign (*no* L_{cor})	10.70
RNAlign (*no purity labels*)	17.34

***Table 2** : Model ablation impairs batch effect removal performance.*