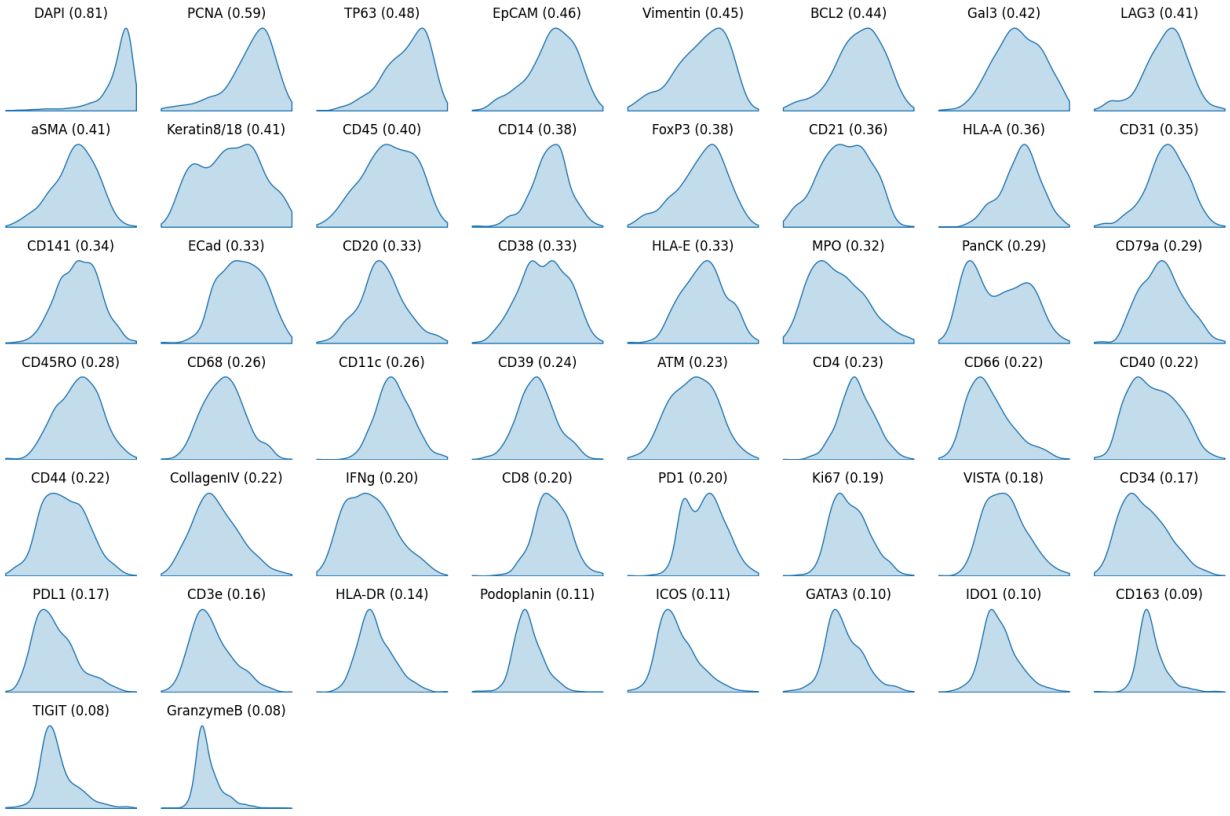
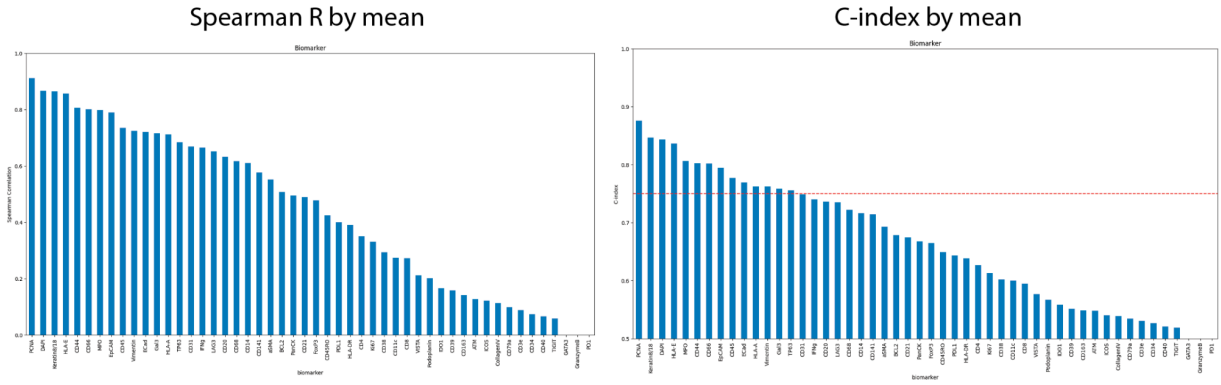
Stanford-PGC (50 biomarkers)	**Model parameters (#)**	**Pearson R**	**Spearman R**	**C-index (sample)**
- *ConvNext (Ours)*	*50.2M*	*0.319*	*0.386*	*0.694*
- UNI (VitL-16)	304.3M	0.278	0.337	0.619
- VitB-16	86.6M	0.294	0.356	0.626
- Pix2pix (GAN)	57.3M	0.034	0.091	0.559

Supplementary Table 1: Comparison of model architectures. Our CNN-based approach outperformed larger Vision Transformer models, histopathology image pre-training, and GAN-based approaches.

Supplementary Figure 1: Median samples (by Pearson R) are visualized for all 50 biomarkers from the Stanford-PGC test set.



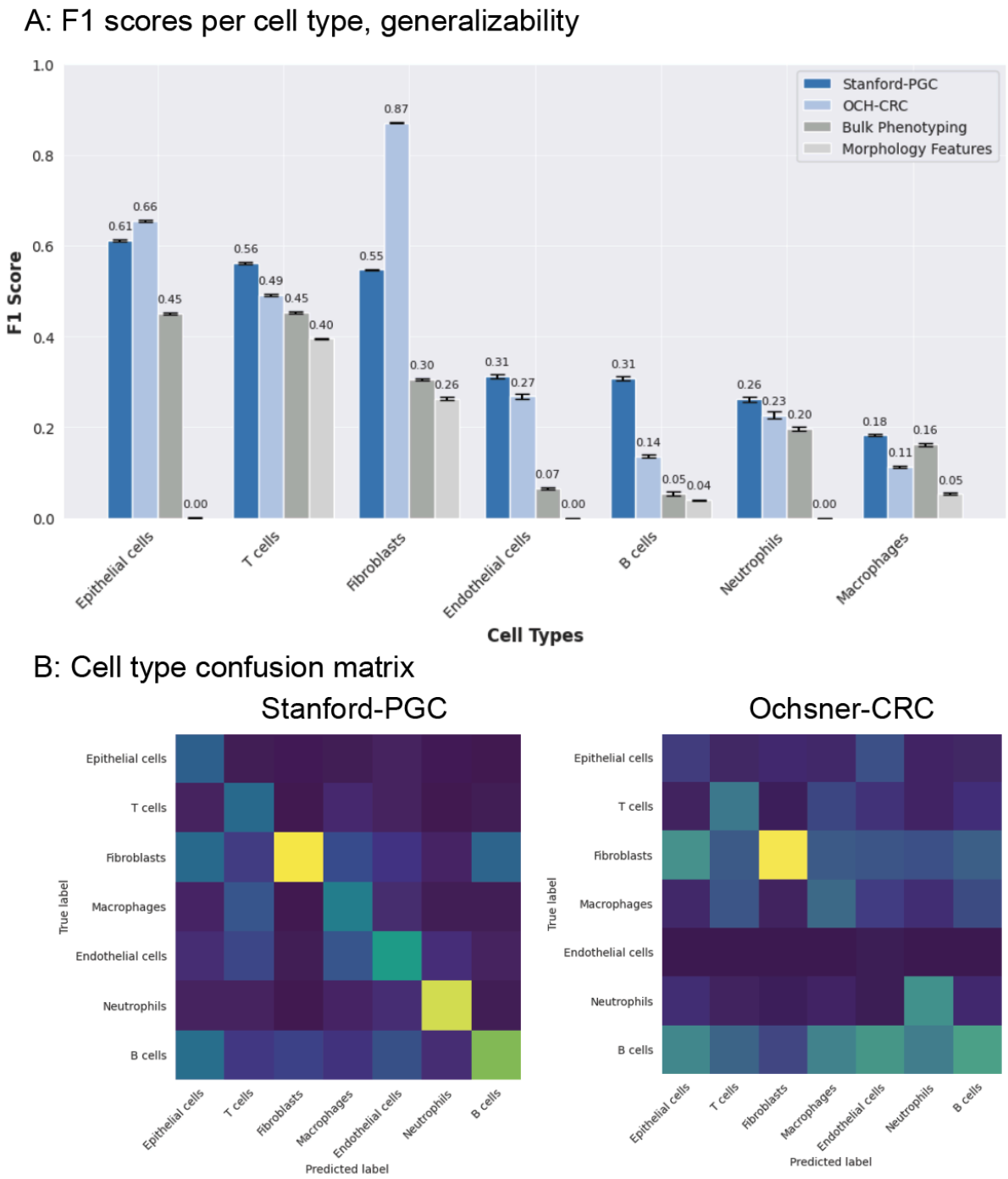
Supplementary Figure 2: Density plot of Pearson R correlation for each biomarker across the distribution of samples in the Stanford-PGC test set.



Supplementary Figure 3: Rank tests on each biomarker comparing ground truth and *ROSIE*-generated expressions in the Stanford-PGC test set (N=817,765 cells).

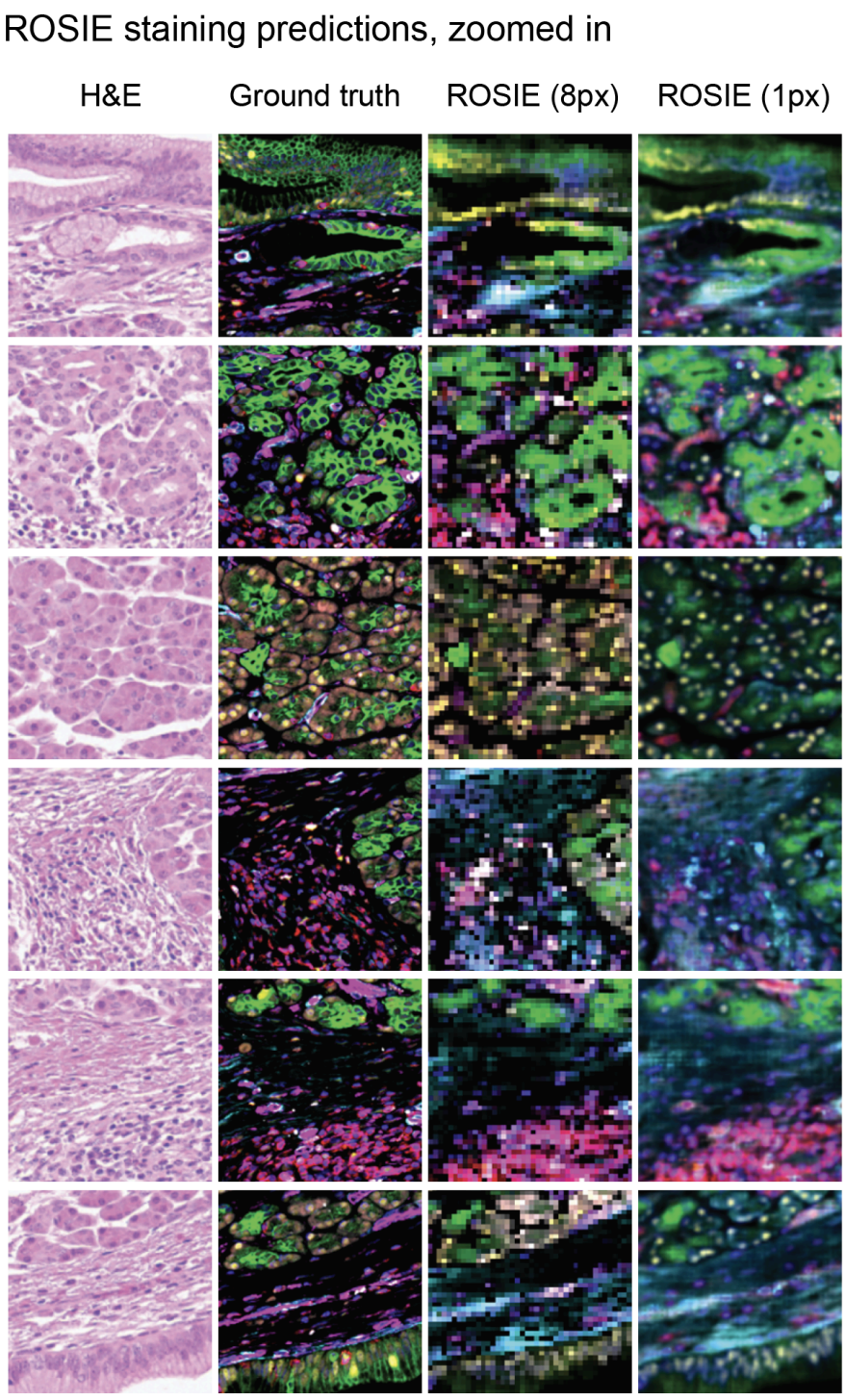
Dataset	**Stanford-PGC**			**Ochsner-CRC**
Method	*ROSIE*	*Bulk phenotyping*	*Cell morphology*	*ROSIE*
B cells	0.307 (0.304-0.312)	0.054 (0.049-0.058)	0.040 (0.039-0.041)	0.118 (0.115-0.121)
Endothelial cells	0.313 (0.308-0.317)	0.065 (0.063-0.067)	0.000 (0.000-0.000)	0.222 (0.219-0.226)
Epithelial cells	0.611 (0.609-0.613)	0.450 (0.449-0.452)	0.000 (0.000-0.001)	0.551 (0.549-0.553)
Fibroblasts	0.547 (0.545-0.548)	0.305 (0.303-0.307)	0.263 (0.261-0.265)	0.287 (0.284-0.289)
Macrophages	0.182 (0.180-0.185)	0.162 (0.159-0.164)	0.053 (0.051-0.055)	0.012 (0.000-0.027)
Neutrophils	0.262 (0.256-0.266)	0.196 (0.193-0.200)	0.000 (0.000-0.000)	0.288 (0.284-0.293)
T cells	0.561 (0.559-0.563)	0.453 (0.451-0.455)	0.395 (0.394-0.397)	0.465 (0.464-0.467)
Weighted Average	0.507	0.360	0.169	0.411

Supplementary Table 2: Cell type predictions by individual cell labels on the Stanford-PGC (N=817,765 cells) and Ochsner-CRC (N=635,649 cells) test datasets, along with two baseline methods (Bulk phenotyping and Cell morphology). We report F1 scores for each cell label and 95% bootstrapped confidence intervals in parentheses.



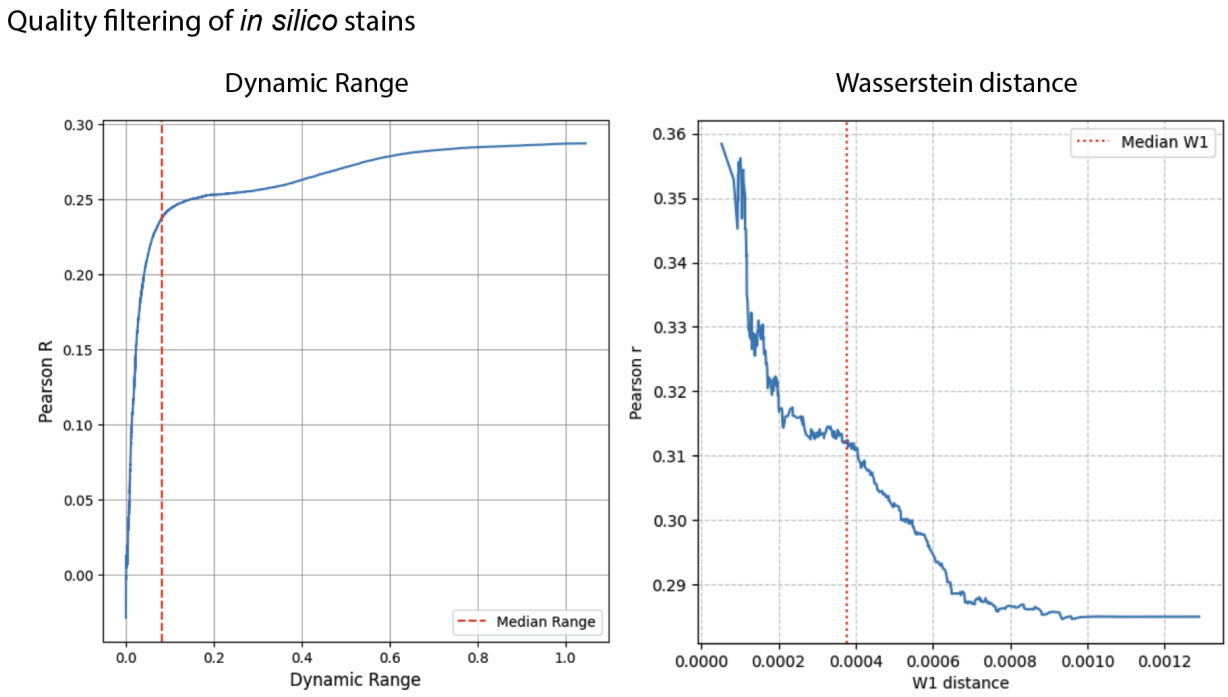
Supplementary Figure 4: *ROSIE* generalizes well in predicting cell types in a study of unseen-before colorectal cancer tumor samples (Ochsner-CRC, N=635,649 cells). **(A)** The cell type predictions on Ochsner-CRC are comparable to those in Stanford-PGC and two baseline methods, Bulk Phenotyping and Morphology Features. Data are presented as mean values with error bars as the 95% bootstrapped confidence intervals. **(B)** Cell type confusion matrices for

predictions from *ROSIE* on the Stanford-PGC and Ochsner-CRC datasets. Of note is that B cells and T cells are well-differentiated in cell predictions.

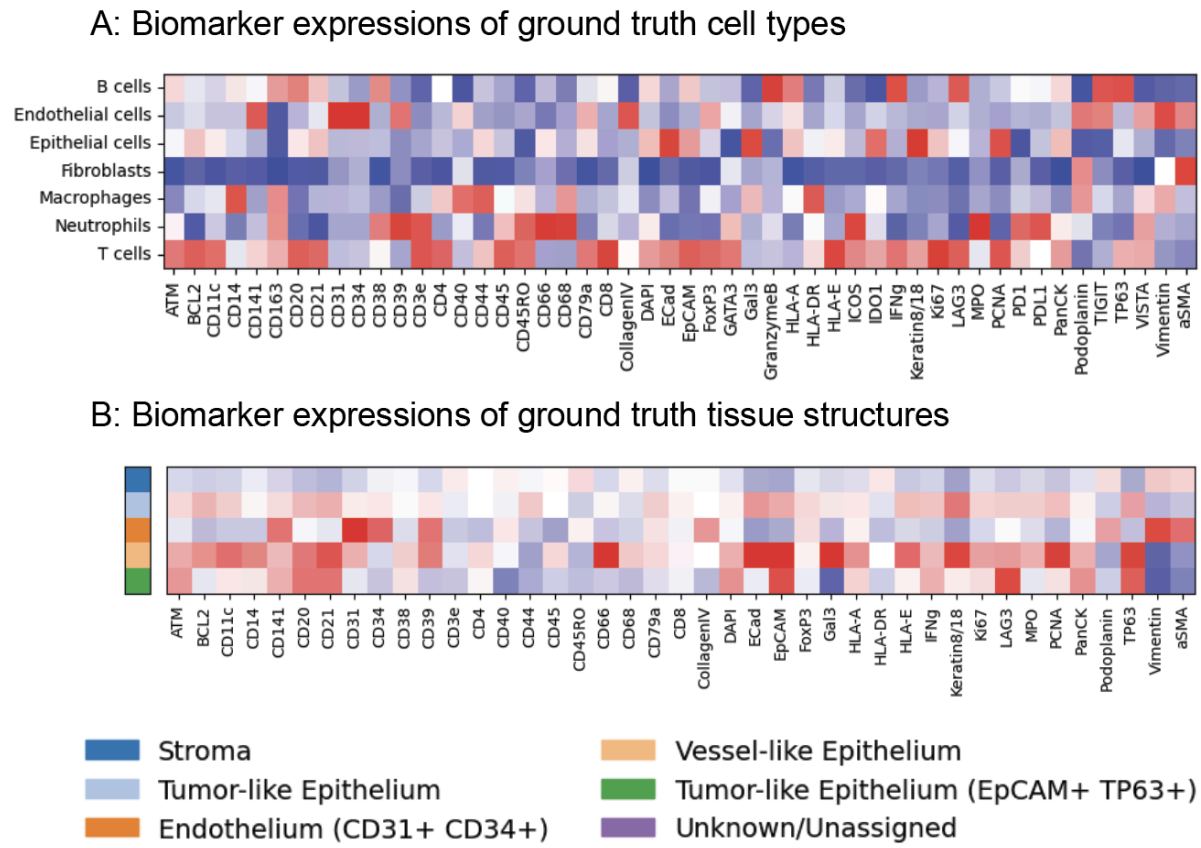


Supplementary Figure 5: Zoomed-in patches of predictions from *ROSIE*, using an 8px striding window and 1px striding window. With 1px, the resolution and detail are significantly improved;

however, generating a sample requires 64x the amount of compute resources. All analyses in this study are performed on the 8px sliding window-generated images.



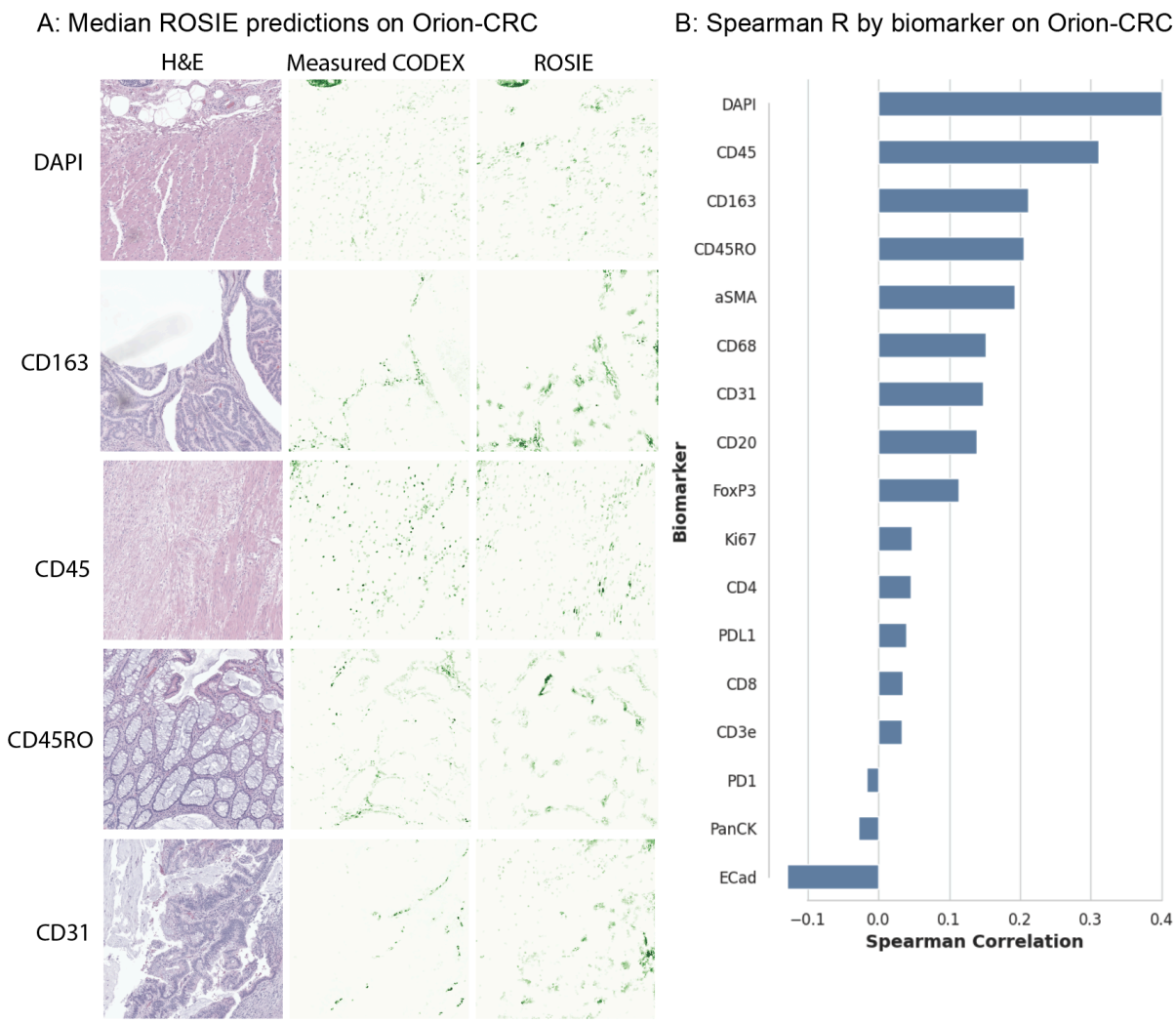
Supplementary Figure 6: Quality filtering: We show two simple methods for predicting the quality of a predicted stain using ROSIE on the Stanford-PGC dataset (N=817,765 cells). **Left**: Low-quality predictions yield samples with low dynamic range; based on this intuition, we show that the predictive accuracy (measured by average Pearson R) of a biomarker stain is strongly positively correlated with a higher dynamic range. **Right**: We also compute the Wasserstein distance between the H&E image intensity histograms of the training and test data. We also find that higher W1 distances are associated with lower Pearson R scores. In both plots, each point represents the average Pearson R of all samples with measured values less than or equal to the value on the x-axis.



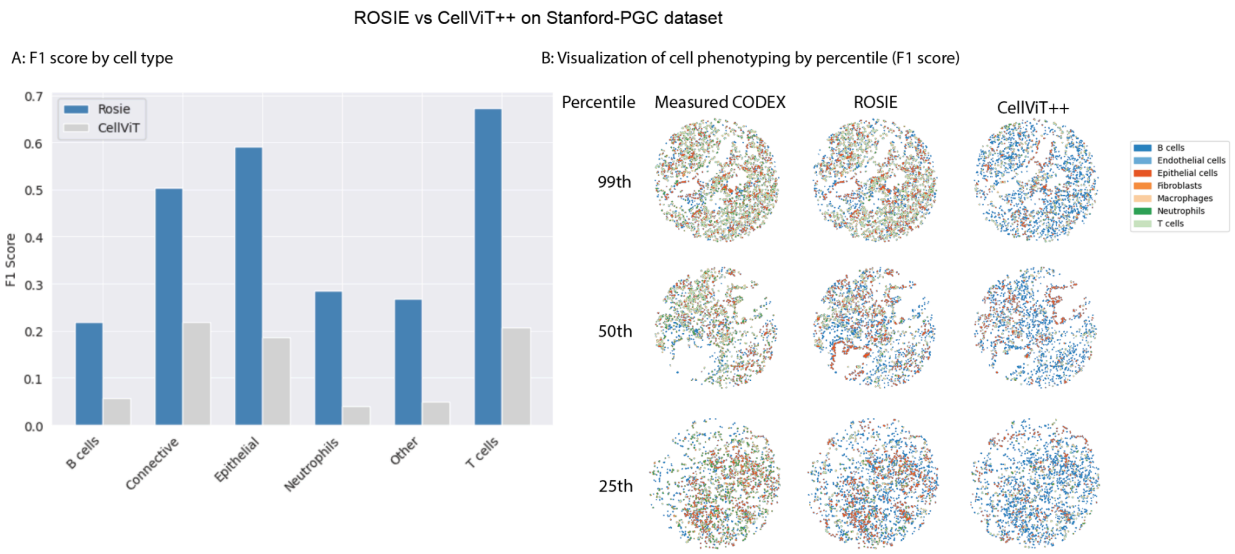
Supplementary Figure 7: (A) The mean expression profile for each ground truth cell phenotype on the Stanford-PGC test dataset (N=817,765 cells). Each biomarker expression is mean normalized across the cell types. **(B)** The mean expression profile for each ground truth tissue structure on the Stanford-PGC test dataset.

Tissue Structure	**F1 (*ROSIE*)**	**F1 (Morphology)**
Stroma	0.786 (0.784-0.788)	0.627 (0.625-0.629)
Tumor-like Epithelium	0.712 (0.710-0.714)	0.408 (0.405-0.410)
Vessel-like Epithelium	0.588 (0.583-0.593)	0.062 (0.059-0.065)
Endothelium (CD31+ CD34+)	0.416 (0.409-0.424)	0.049 (0.046-0.051)
Tumor-like Epithelium (EpCAM+ TP63+)	0.620 (0.610-0.627)	0.000 (0.000-0.000)

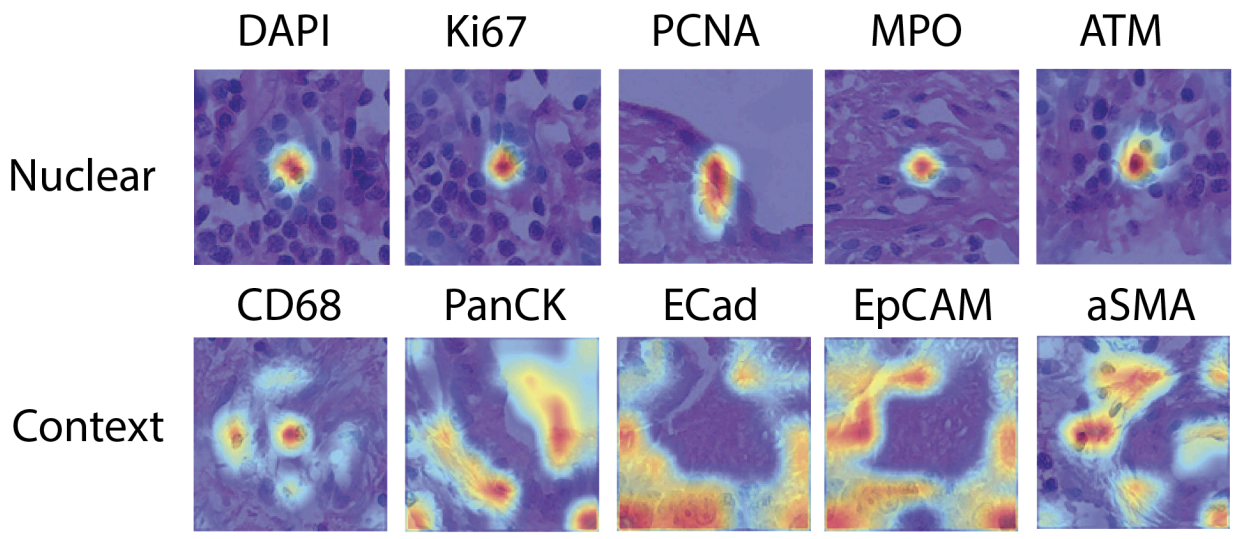
Supplementary Table 3: Tissue structure discovery performance by individual tissue structure type on the Stanford-PGC dataset (N=817,765 cells). 95% bootstrapped confidence intervals are reported in parentheses.



Supplementary Figure 8: Evaluation of ROSIE on Orion-CRC, a colorectal dataset imaged on the Orion platform. (A) visualizes the median ROSIE-predicted sample (by Spearman R) for each of the top 5 biomarkers, along with the input H&E patch and the CODEX-measured ground truth measurement. (B) shows the overall Spearman R for each of the 17 overlapping biomarkers in the Orion-CRC dataset.

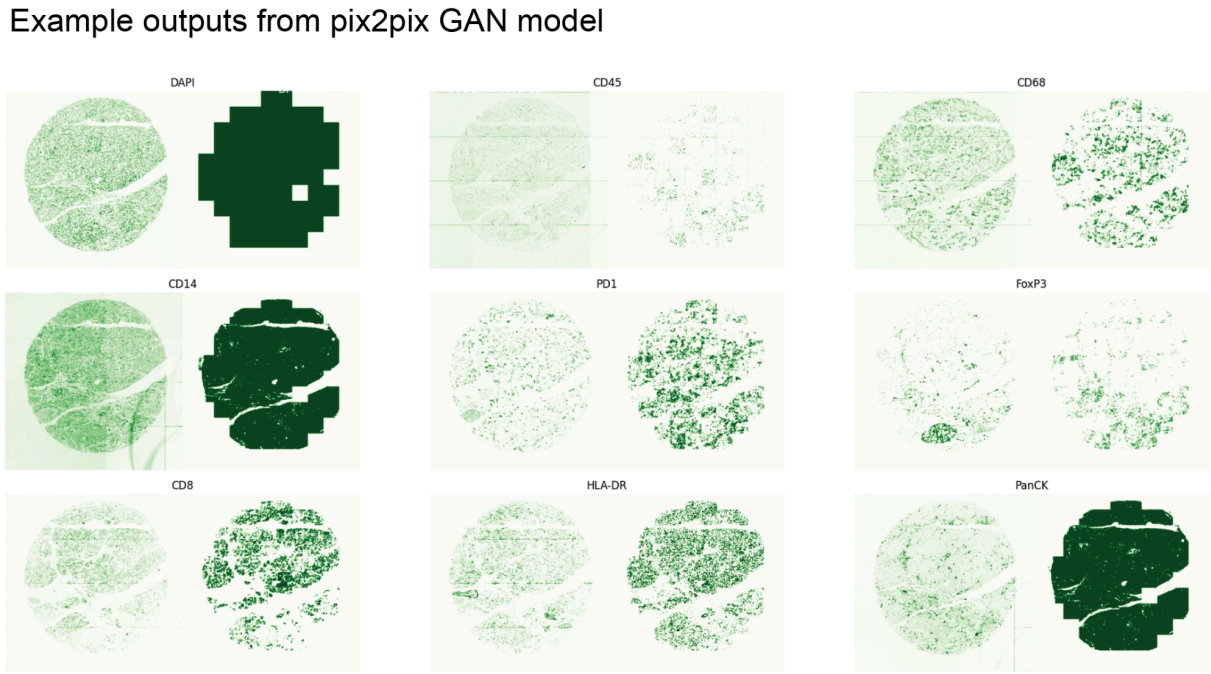


Supplementary Figure 9: Comparison of CellViT++ and ROSIE on cell phenotyping for the Stanford-PGC dataset (N=817,765 cells). A: F1 scores of the two algorithms on 220K cells for identifying six cell types. B: Visualizations of samples from Stanford-PGC, from the 99th, median, and 25th percentile scores by F1 score. The left column shows the cell types derived from the CODEX ground truth measurements, the middle column shows the cell types derived from the ROSIE-generated measurements, and the right column shows the CellViT++-predicted cell types.



Supplementary Figure 10: Gradient heatmaps created using GRAD-CAM for select biomarkers on a sample from Stanford-PGC. We visualize randomly sampled input H&E patches along with the gradient heatmaps derived from ROSIE, indicating where the model attended to in the image. The top row shows biomarkers that are associated with nuclear

morphology, whereas the bottom row shows biomarkers that are typically defined by the surrounding context.



Supplementary Figure 11: Example outputs from a GAN model architecture (pix2pix). Each pair of images shows the CODEX-measured ground truth on the left and the GAN-model-generated output on the right. This example shows that GAN model outputs can be unstable and yield degenerate images, especially when jointly training on a large biomarker panel.