

Single-Cell Multiplex Immune Biomarker Prediction from Histopathology Across Cancer Types and Clinical Endpoints

Usman Afzaal^{1*}, Usama Sajjad¹, Abdul Rehman Akbar¹, Tianyang Wang¹, Elshad Hasanov², Sachin Jhavar³, Wei Chen¹, Zaibo Li¹, Muhammad Khalid Khan Niazi¹

¹Department of Pathology, College of Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA

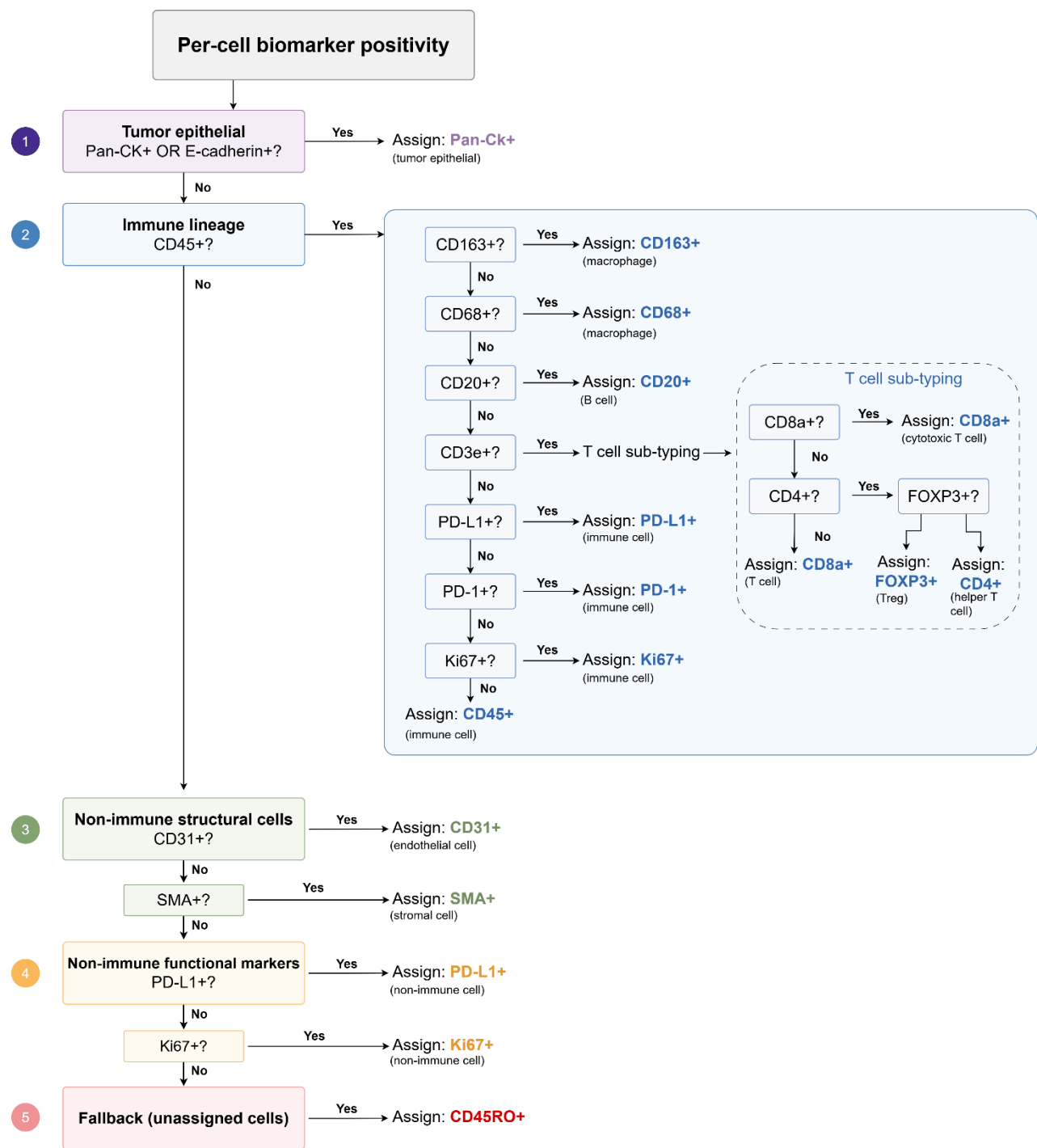
²Department of Medical Oncology, College of Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA

³Department of Radiation Oncology, College of Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA

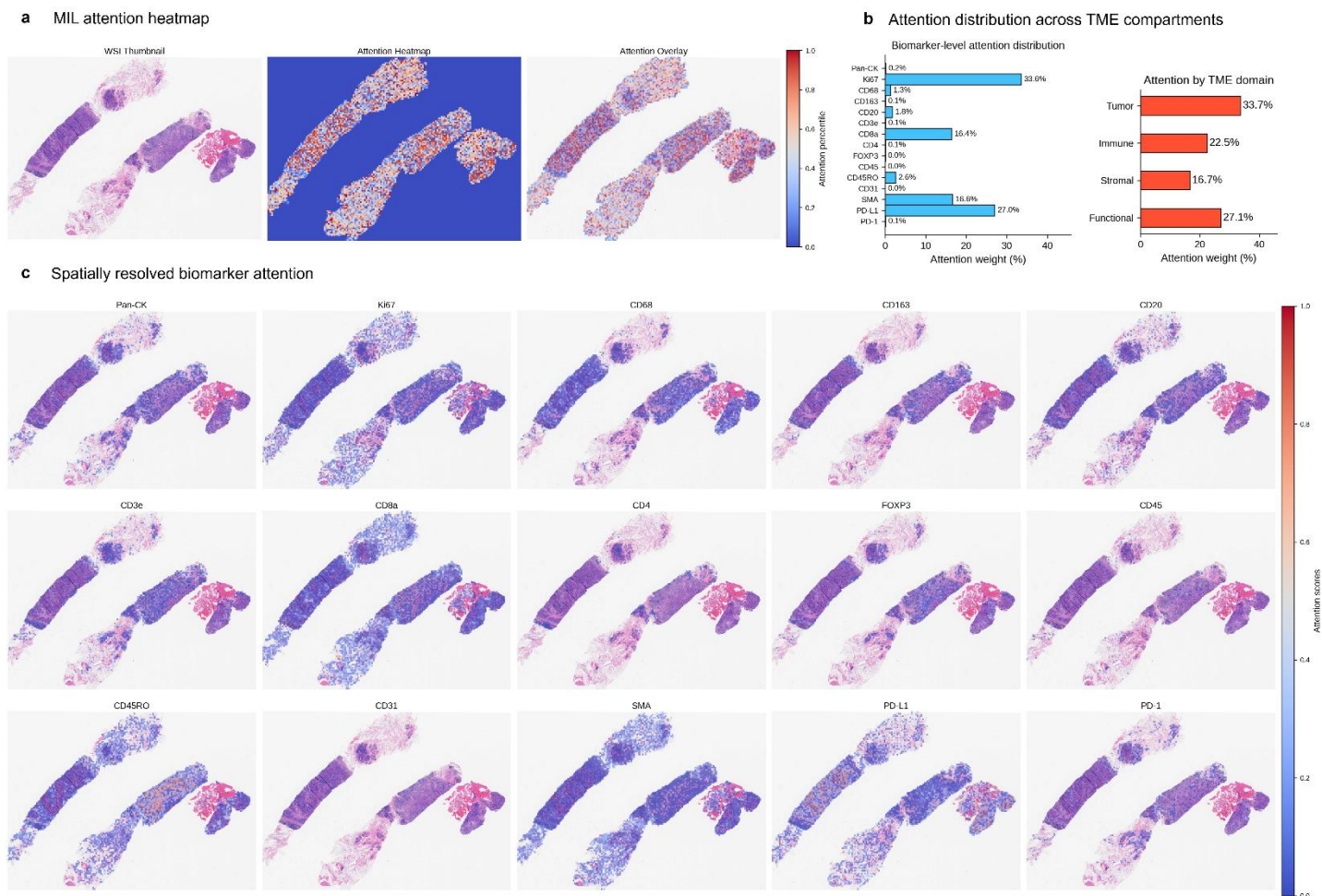
***Correspondence:** usman.afzaal@osumc.edu

***Lead contact:** Usman Afzaal

Supplementary



Supplementary Figure S1 Hierarchical gating strategy for assigning 15 cell phenotypes



Supplementary Figure S2 **Multiple instance learning attention analysis for a representative pCR = 1 slide from our internal TNBC dataset.** (a) MIL attention heatmap showing the WSI thumbnail, per-patch attention heatmap, and attention overlay for the WSI. (b) Attention distribution across biomarkers and tumor microenvironment (TME) compartments. Left: percentage of attention assigned to each of the 15 profiled biomarkers. Right: attention aggregated by TME domain (Tumor, Immune, Stromal, Functional). (c) Spatially resolved, biomarker-specific attention maps for the WSI, showing the spatial distribution of model attention overlaid on the tissue for each of the 15 individual biomarkers.

Biomarker-level attention distribution

Attention percentage

Biomarker	Attention weight (%)
PD-1	20.9%
CD163	17.7%
CD8	42.4%
CD45RO	1.8%
CD45	0.0%
CD11	0.0%
PD-L1	0.3%
FOXP3	0.0%
CD14	0.0%
CD163	0.9%
CD20	0.5%
CD3e	0.1%
CD3	0.0%
CD8	0.0%
CD45RO	0.0%
CD11	0.0%
PD-L1	0.0%

Attention weight (%)

Attention by TME domain

TME Domain	Attention weight (%)
Tumor	43.5%
Immune	21.7%
Stromal	13.6%
Functional	21.1%

Attention weight (%)

Attention maps for specific biomarkers:

- CD163
- CD20
- FOXP3
- CD45
- PD-L1
- PD-1

Attention percentage

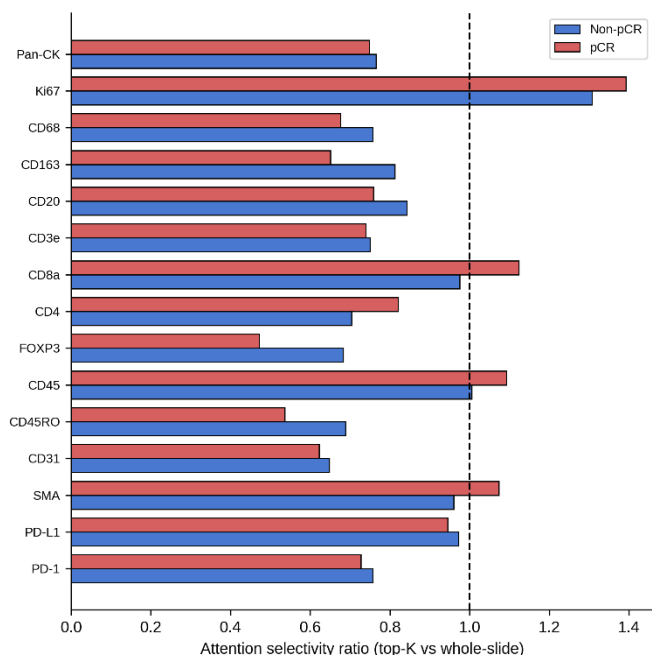
Figure 2 displays 15 immunohistochemical (IHC) images of tumor tissues, arranged in a 3x5 grid. Each image shows a histological section stained for a specific marker. The markers are labeled above each image:

- Row 1: Pan-CK, Ki67, CD68, CD163, CD20
- Row 2: CD3e, CD8a, CD4, FOXP3, CD45
- Row 3: CD45RO, CD31, SMA, PD-L1, PD-1

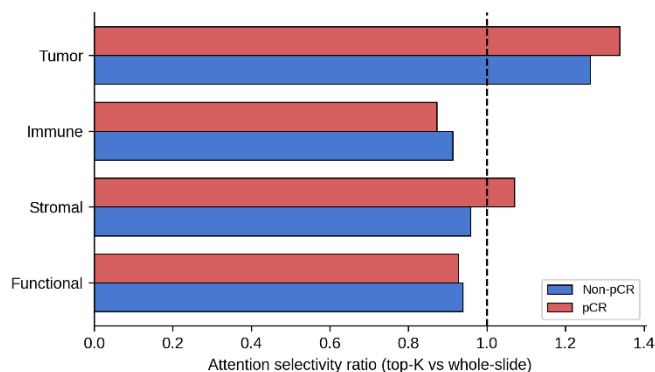
The images show various cellular structures and staining patterns, indicating the presence and distribution of these markers within the tumor tissue.

Supplementary Figure S3 **Multiple instance learning attention analysis for a representative pCR = 0 slide from our internal TNBC dataset.** (a) MIL attention heatmap showing the WSI thumbnail, per-patch attention heatmap, and attention overlay for the WSI. (b) Attention distribution across biomarkers and tumor microenvironment (TME) compartments. Left: percentage of attention assigned to each of the 15 profiled biomarkers. Right: attention aggregated by TME domain (Tumor, Immune, Stromal, Functional). (c) Spatially resolved, biomarker-specific attention maps for the WSI, showing the spatial distribution of model attention overlaid on the tissue for each of the 15 individual biomarkers.

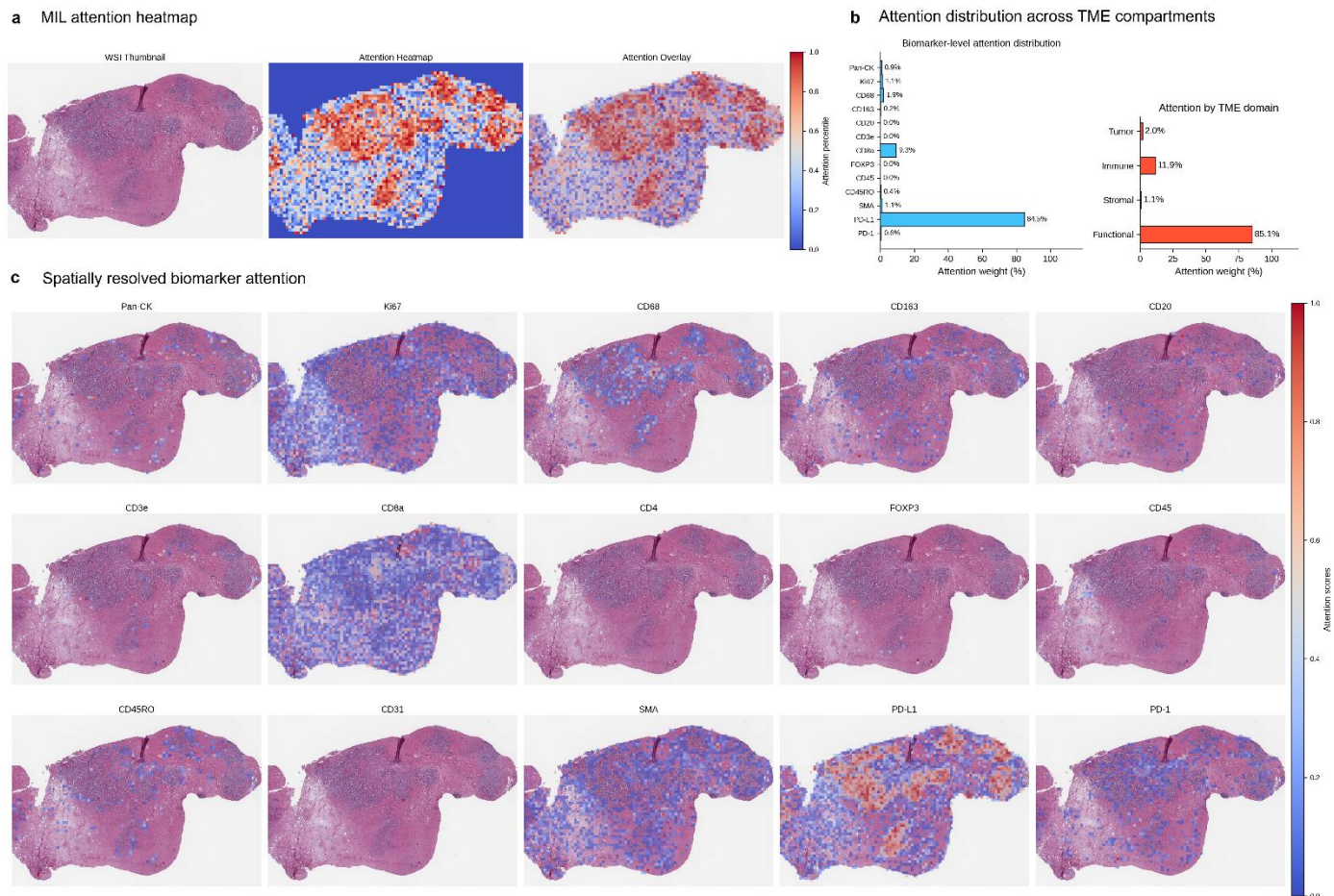
a Biomarker-level attention selectivity



b TME domain-level attention selectivity

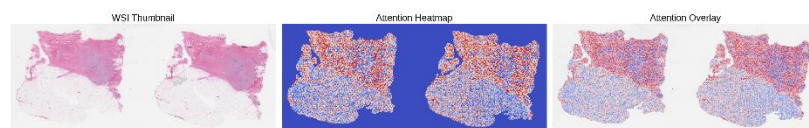


Supplementary Figure S4 **Attention selectivity analysis across pCR and Non-pCR classes for the internal TNBC dataset.** (a) Biomarker-level and (b) tumor microenvironment (TME) domain-level attention selectivity ratios, defined as the ratio between biomarker/domain composition within the top-attended patches and its composition across the whole slide, averaged across slides within each class. A ratio > 1 indicates the model attends to that biomarker/domain more than expected from its baseline abundance in the tissue, while a ratio < 1 indicates under-attention relative to baseline.

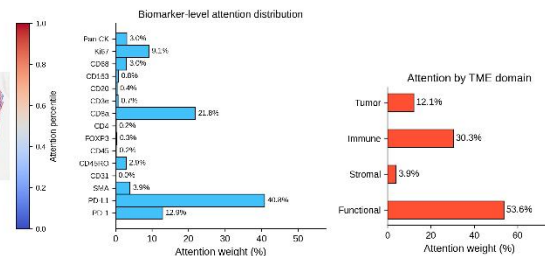


Supplementary Figure S5 **Multiple instance learning attention analysis for a representative Survival = 0 slide from TCGA-BRCA dataset.** (a) MIL attention heatmap showing the WSI thumbnail, per-patch attention heatmap, and attention overlay for the WSI. (b) Attention distribution across biomarkers and tumor microenvironment (TME) compartments. Left: percentage of attention assigned to each of the 15 profiled biomarkers. Right: attention aggregated by TME domain (Tumor, Immune, Stromal, Functional). (c) Spatially resolved, biomarker-specific attention maps for the WSI, showing the spatial distribution of model attention overlaid on the tissue for each of the 15 individual biomarkers.

a MIL attention heatmap



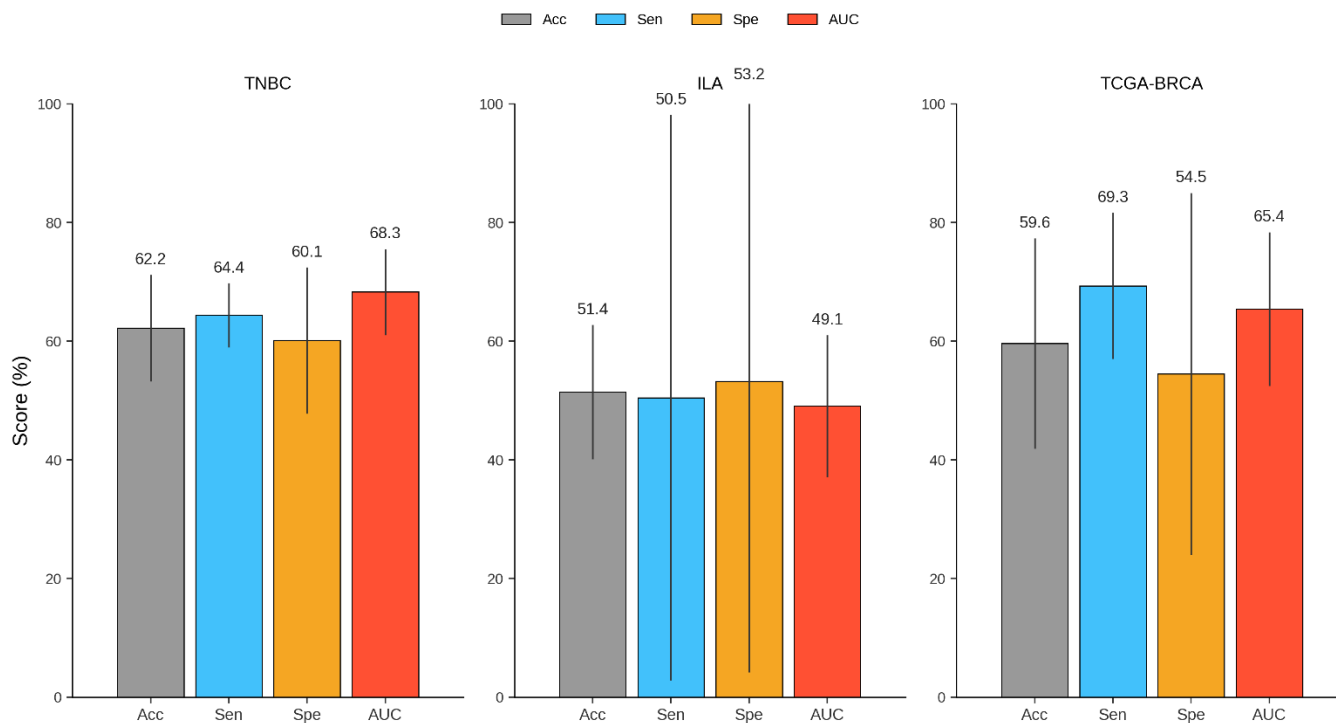
b Attention distribution across TME compartments



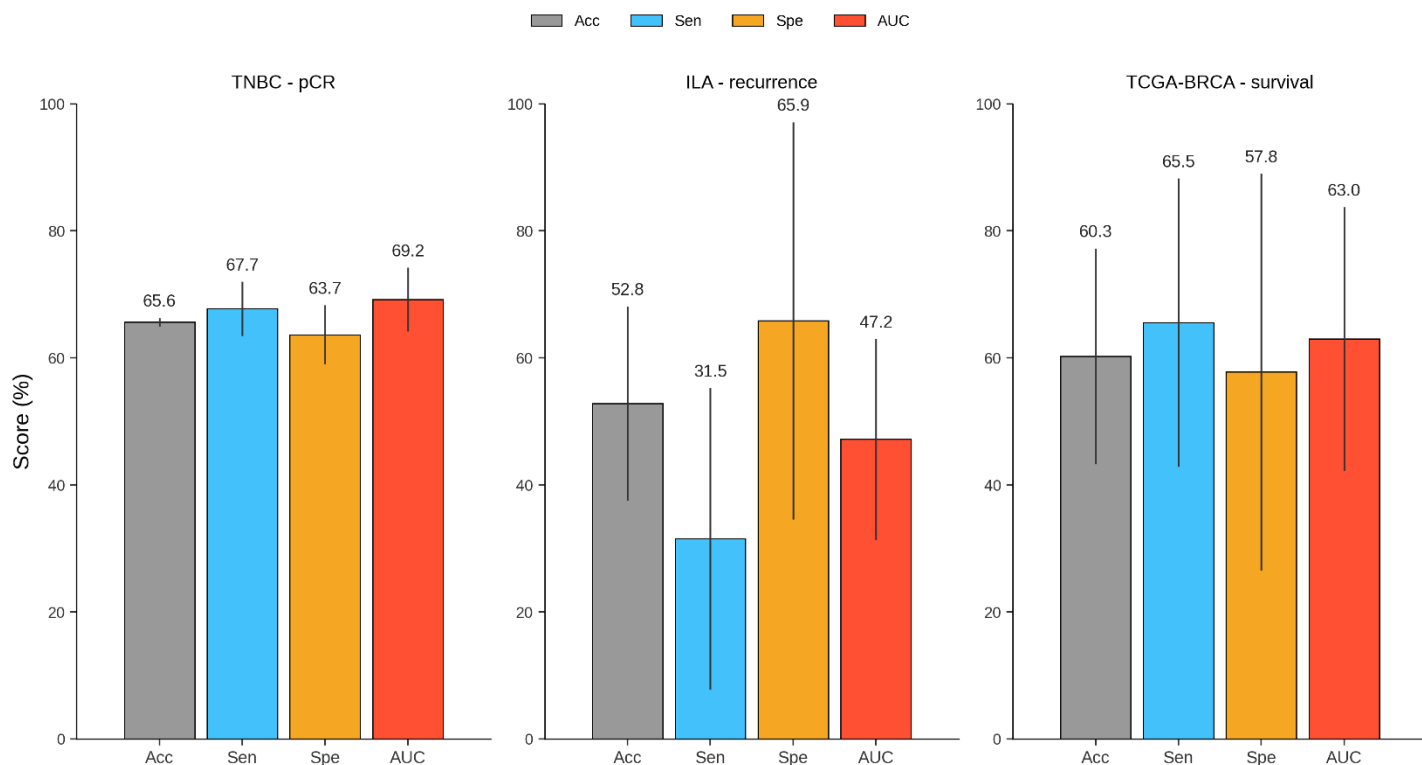
c Spatially resolved biomarker attention



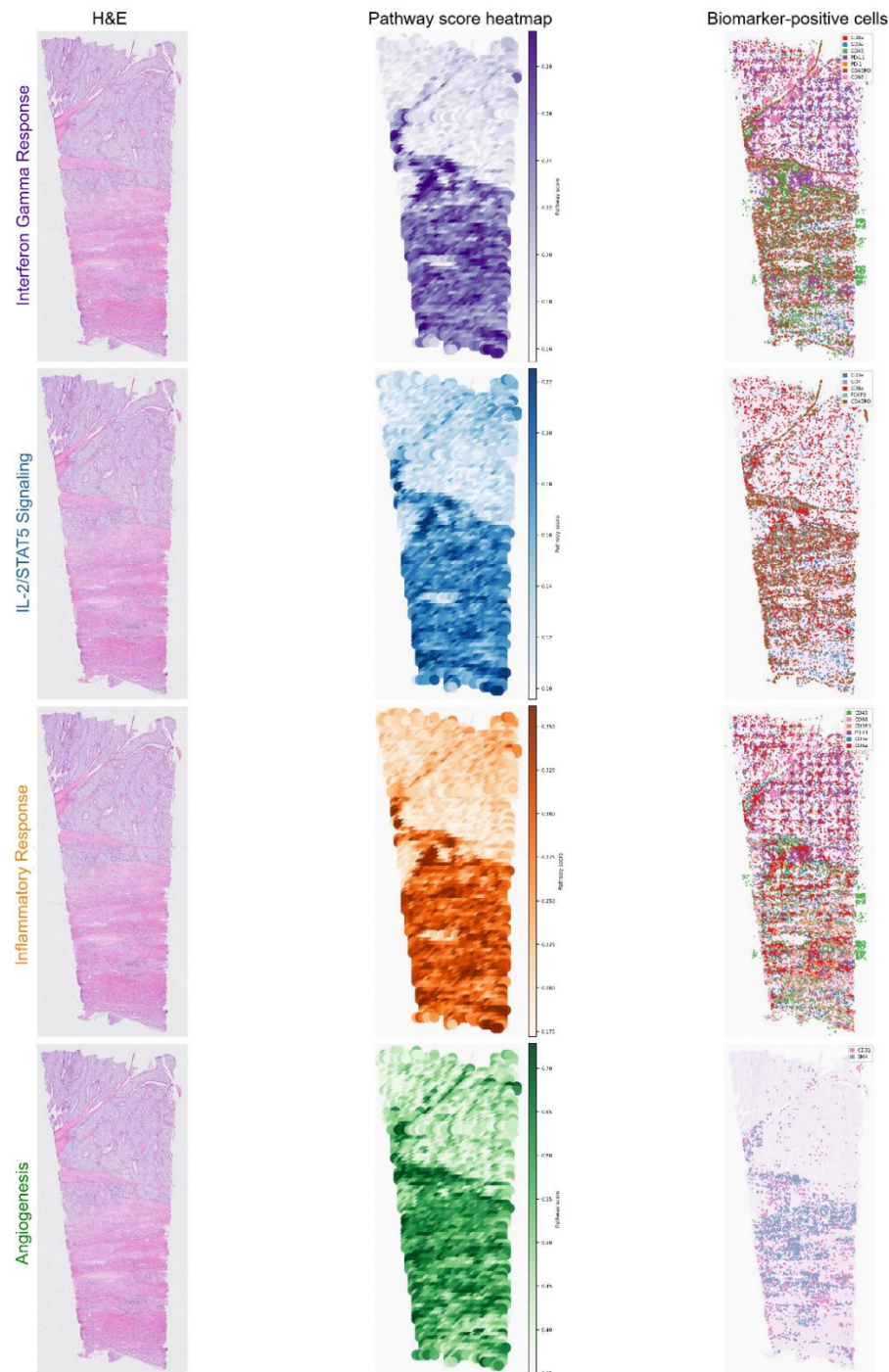
Supplementary Figure S6 **Multiple instance learning attention analysis for a representative Survival = 1 slide from TCGA-BRCA dataset.** (a) MIL attention heatmap showing the WSI thumbnail, per-patch attention heatmap, and attention overlay for the WSI. (b) Attention distribution across biomarkers and tumor microenvironment (TME) compartments. Left: percentage of attention assigned to each of the 15 profiled biomarkers. Right: attention aggregated by TME domain (Tumor, Immune, Stromal, Functional). (c) Spatially resolved, biomarker-specific attention maps for the WSI, showing the spatial distribution of model attention overlaid on the tissue for each of the 15 individual biomarkers.



Supplementary Figure S7 CellPlexer performance using only the top 5% most-attended patches per slide, across multiple prediction tasks. For each task (TNBC pathological complete response (pCR) prediction, ILA recurrence prediction and TCGA-BRCA survival) the trained model was evaluated using only the top 5% of patches ranked by attention weight per slide, with no retraining or fine-tuning. Bars show accuracy (Acc), sensitivity (Sen), specificity (Spe), and area under the ROC curve (AUC), all expressed as percentages and averaged across cross-validation folds, where error bars denote standard deviation across folds.



Supplementary Figure S8 CellPlexer performance using only random 5% patches per slide, across multiple prediction tasks. For each task (TNBC pathological complete response (pCR) prediction, ILA recurrence prediction and TCGA-BRCA survival) the trained model was evaluated using only random 5% of patches per slide, with no retraining or fine-tuning. Bars show accuracy (Acc), sensitivity (Sen), specificity (Spe), and area under the ROC curve (AUC), all expressed as percentages and averaged across cross-validation folds, where error bars denote standard deviation across folds.

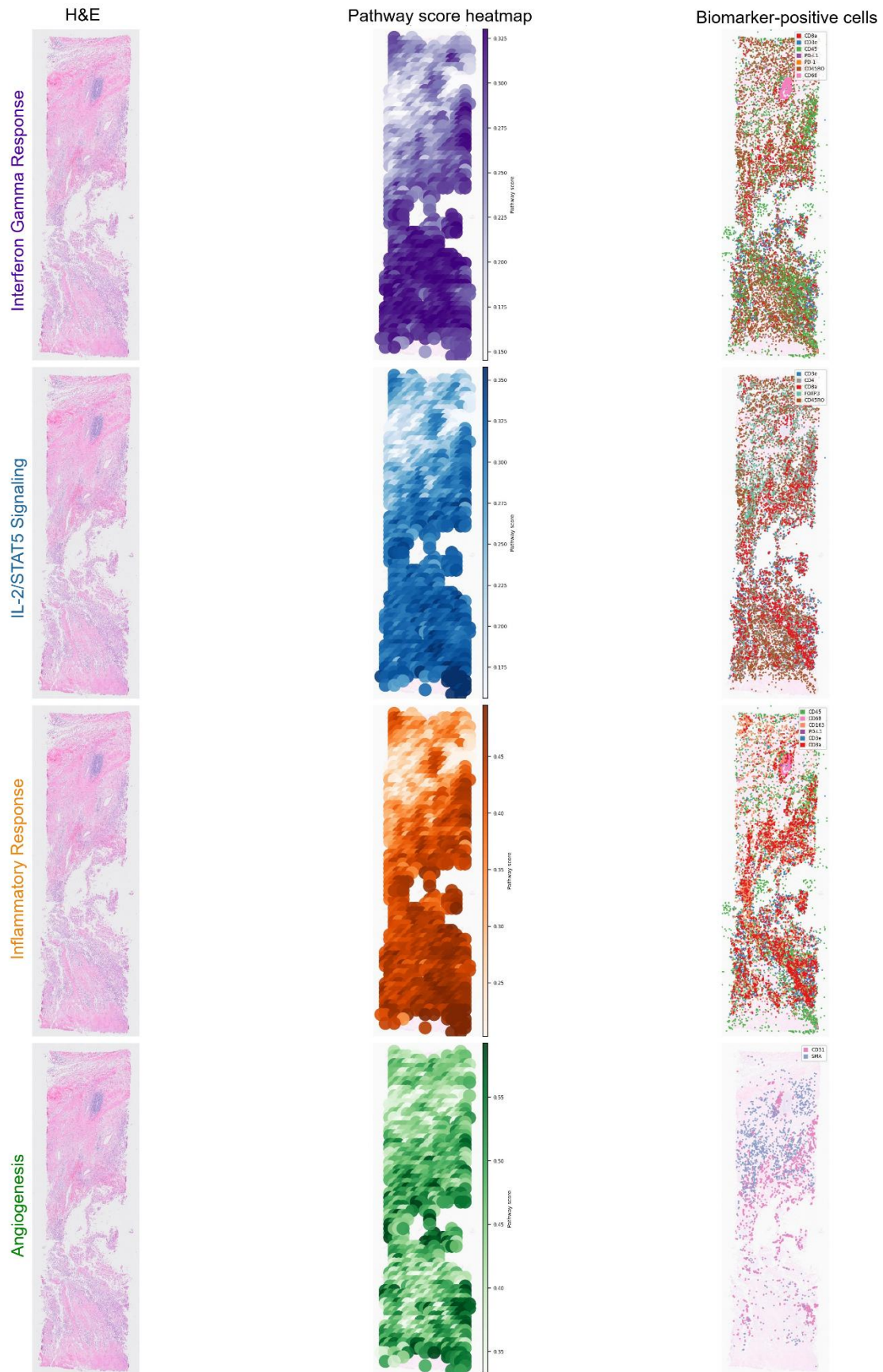


Supplementary Figure S9 **Spatial concordance between Hallmark pathway activity and CellPlexer-detected biomarker-positive cells in a representative tissue section.** Serial visualization of a single HEST spatial transcriptomics tissue section, shown across four Molecular Signatures Database (MSigDB) Hallmark pathways (rows, top to bottom: Interferon Gamma Response, IL-2/STAT5 Signaling, Inflammatory Response, Angiogenesis). Left column (H&E): reference hematoxylin and eosin whole-slide image thumbnail, identical across rows. Middle column (Pathway score heatmap): per-spot Hallmark pathway score (mean log-normalized expression of pathway member genes), overlaid on a H&E background. Each circle represents one spatial transcriptomics capture spot, enlarged relative to its true physical diameter to visualize spatial continuity between neighboring high-scoring spots. Right column (Biomarker-positive cells); individual cell centroids classified as positive for that pathway's hypothesized primary and secondary biomarkers (see Supplementary Figure S11 legend for the full biomarker-to-pathway assignment), detected using a CellPlexer model trained for 15-biomarker multiplex classification, color-coded by biomarker identity and overlaid on the same H&E

background. Only spots and cells falling within the annotated tissue region are shown. This representative section illustrates the spatial basis for the cohort-wide enrichment statistics reported in Supplementary Figure S11.

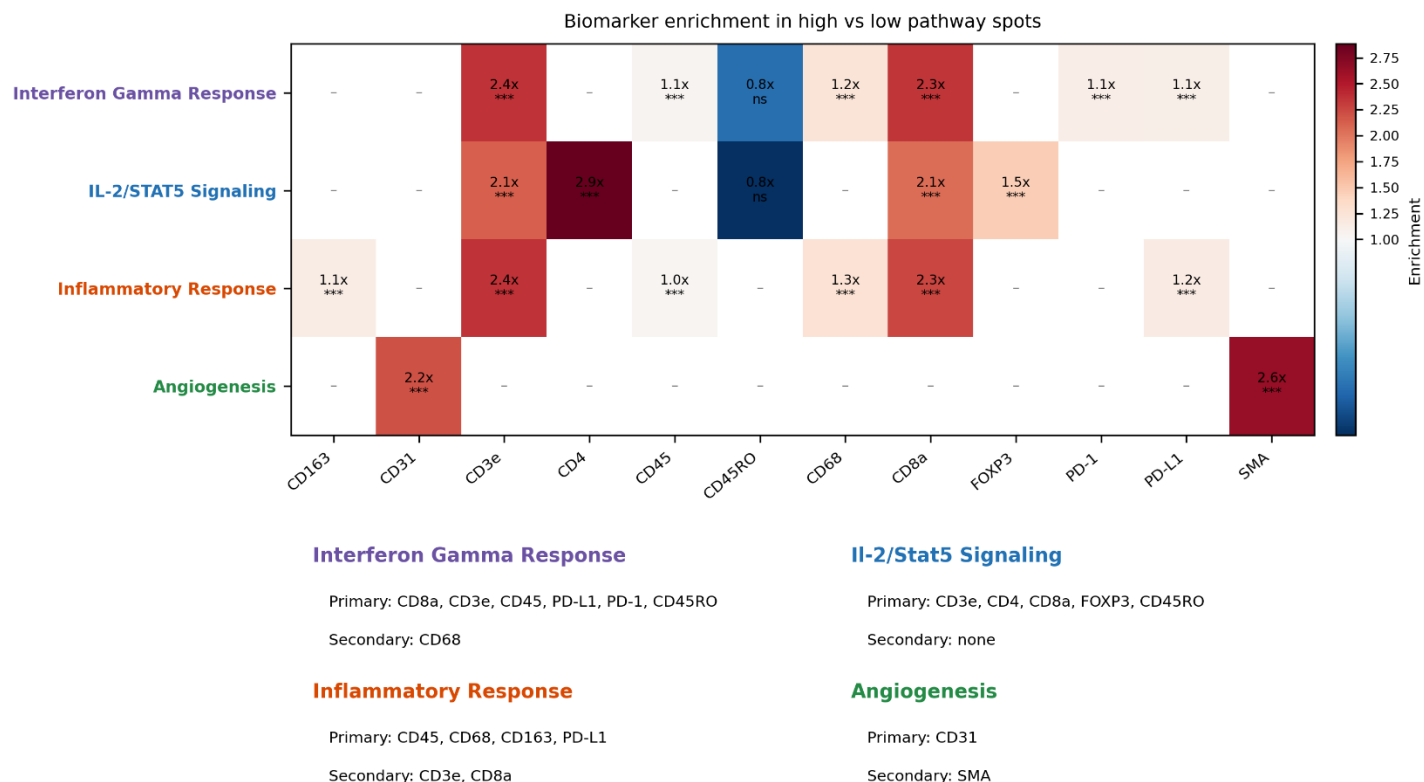


Supplementary Figure S10 **Spatial concordance between Hallmark pathway activity and CellPlexer-detected biomarker-positive cells in a second representative tissue section.** As in Figure Supplementary Figure S9, shown for an independent tissue section to illustrate consistency of the spatial pattern across samples.



Supplementary Figure S11 **Spatial concordance between Hallmark pathway activity and CellPlexer-detected biomarker-positive cells in a second representative tissue section.** As in

Figure Supplementary Figure S9, shown for an independent tissue section to illustrate consistency of the spatial pattern across samples.



Supplementary Figure S12 **Cellplexer-detected immune and vascular biomarkers are enriched in spatial transcriptomic spots with high activity of biologically matched Hallmark pathways.** Heatmap showing fold-enrichment of 12 Cellplexer-detected biomarkers (columns) across four Molecular Signatures Database (MSigDB) Hallmark pathways (rows: Interferon Gamma Response, IL-2/STAT5 Signaling, Inflammatory Response, Angiogenesis) in HEST spatial transcriptomics spots. For each pathway, spots were ranked by Hallmark pathway score (mean log-normalized expression of member genes) within each tissue section and stratified into high-scoring (top quartile) and low-scoring (bottom quartile) groups. A biomarker was scored as "present" in a spot if at least three positively classified cells of that type were detected within the spot's cell-assignment radius; spots with fewer than five total detected cells were excluded from analysis. Cell values indicate the fold-enrichment in the percentage of high-scoring spots with the biomarker present relative to low-scoring spots (color scale, right; white = no enrichment, dark red = strongest enrichment, blue = depletion). Statistical significance was assessed using a one-sided Fisher's exact test per biomarker-pathway pair, testing specifically for enrichment (not depletion) in high-scoring spots, with Benjamini-Hochberg false-discovery-rate correction applied across all tests (* $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$; ns, not significant). Dashes indicate biomarker-pathway combinations that were not tested, as the biomarker was not part of the a priori hypothesized biomarker set for that pathway. Each pathway's hypothesized primary (strong prior biological association) and secondary (partial association) biomarkers, defined prior to analysis, are listed at the bottom of the figure.

Supplementary Table S1 CellPlexer cell-level training dataset statistics

Cell Type	Train	Val	Test	Total
CD31	460992	115064	148640	724696
CD45	1963260	491637	389799	2844696
CD68	1213748	163866	387463	1765077
CD4	983900	173775	273941	1431616
FOXP3	253692	70284	79232	403208
CD8a	1242988	133349	248816	1625153
CD45RO	2593496	717995	625167	3936658
CD20	2075173	102222	319085	2496480
PD-L1	381171	20475	244251	645897
CD3e	138554	20683	20586	179823
CD163	1953755	213136	600504	2767395
PD-1	170278	25324	27168	222770
Ki67	93487	55039	13276	161802
Pan-CK	9216973	2894053	2808944	14919970
SMA	1365996	358493	551880	2276369
			Total Cells	36401610

Supplementary Table S2 CellPlexer patch-level training dataset statistics

Split	No. of Patches
Train	660409
Validation	138031
Test	171450

Supplementary Table S3 CRC-SurGen Cohort Dataset Splits for Colorectal Cancer Survival Prediction

Split	No. of WSIs
Train	296
Validation	43
Test	85
Total	424

Supplementary Table S4 TCGA-BRCA Cohort Dataset Splits for Breast Cancer Survival Prediction

Split	No. of WSIs
Train	87
Validation	29
Test	30
Total	146

Supplementary Table S5 Internal TNBC Cohort Dataset Splits for Pathological Complete Response (pCR) Prediction

Split	No. of WSIs
Train	118
Validation	14
Test	34
Total	166

Supplementary Table S6 Internal ILA Cohort Dataset Splits for Recurrence Prediction

Split	No. of WSIs
Train	297
Validation	67
Test	92
Total	456

Supplementary Table S7 CRC-SurGen Cohort Dataset Splits for KRAS mutation Prediction

Split	No. of WSIs
Train	291
Validation	42
Test	83
Total	416

Supplementary Table S8 CRC-SurGen Cohort Dataset Splits for BRAF mutation Prediction

Split	No. of WSIs
Train	235
Validation	34
Test	67
Total	336

Supplementary Table S9 TCGA-BRCA performance comparison of state-of-the-art MIL methods with our best performing model. Metrics include accuracy, sensitivity, specificity, and AUC (mean $\pm$ standard deviation across test folds).

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
ABMIL	63.80 $\pm$ 5.18	59.56 $\pm$ 17.40	66.26 $\pm$ 14.80	66.04 $\pm$ 5.92
CLAM	62.39 $\pm$ 16.90	48.89 $\pm$ 16.89	69.32 $\pm$ 16.90	65.27 $\pm$ 6.01
TransMIL	62.34 $\pm$ 6.04	46.89 $\pm$ 18.34	70.26 $\pm$ 16.17	58.52 $\pm$ 5.88
CellPlexer	67.84 $\pm$ 3.14	59.56 $\pm$ 22.43	72.47 $\pm$ 13.67	71.43 $\pm$ 5.09

Supplementary Table S10 Internal TNBC dataset performance comparison of state-of-the-art MIL methods with our best performing model. Metrics include accuracy, sensitivity, specificity, and AUC (mean $\pm$ standard deviation across test folds).

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
ABMIL	64.25 $\pm$ 2.89	58.40 $\pm$ 27.54	66.73 $\pm$ 15.35	68.65 $\pm$ 5.94
CLAM	66.30 $\pm$ 2.69	63.61 $\pm$ 14.66	66.80 $\pm$ 14.08	67.97 $\pm$ 4.97
TransMIL	66.84 $\pm$ 1.22	66.84 $\pm$ 13.73	65.69 $\pm$ 10.37	69.53 $\pm$ 5.12
GigaTIME	58.23 $\pm$ 5.31	30.65 $\pm$ 22.67	79.67 $\pm$ 16.20	58.08 $\pm$ 7.18
PathRosetta	60.59 $\pm$ 6.17	56.14 $\pm$ 19.51	62.29 $\pm$ 13.69	64.77 $\pm$ 6.18
CellPlexer	66.24 $\pm$ 1.79	69.06 $\pm$ 4.35	63.66 $\pm$ 2.20	72.19 $\pm$ 4.48
CellPlexer (with clinical features)	68.76 $\pm$ 2.80	70.31 $\pm$ 6.32	67.12 $\pm$ 6.19	75.06 $\pm$ 2.10

Supplementary Table S11 Internal ILA dataset performance comparison of state-of-the-art MIL methods with our best performing model. Metrics include accuracy, sensitivity, specificity, and AUC (mean $\pm$ standard deviation across test folds).

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
ABMIL	61.55 $\pm$ 8.23	55.77 $\pm$ 12.04	65.39 $\pm$ 5.55	66.57 $\pm$ 5.55
CLAM	61.13 $\pm$ 7.52	56.52 $\pm$ 12.89	64.56 $\pm$ 12.18	66.13 $\pm$ 5.90
TransMIL	58.62 $\pm$ 10.63	65.54 $\pm$ 12.44	54.19 $\pm$ 10.63	64.93 $\pm$ 6.25
GigaTIME	60.26 $\pm$ 10.47	54.17 $\pm$ 18.61	63.35 $\pm$ 25	62.66 $\pm$ 5.92
CellPlexer	63.22 $\pm$ 4.66	59.18 $\pm$ 6.33	65.85 $\pm$ 5.42	68.89 $\pm$ 5.35

Supplementary Table S12 CRC SurGen dataset Survival performance comparison of state-of-the-art MIL methods with our best performing model. Metrics include accuracy, sensitivity, specificity, and AUC (mean $\pm$ standard deviation across test folds).

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
ABMIL	59.80 $\pm$ 5.02	54.05 $\pm$ 12.94	65.55 $\pm$ 8.64	65.46 $\pm$ 7.12
CLAM	60.87 $\pm$ 5.42	45.58 $\pm$ 16.75	73.16 $\pm$ 10.02	66.41 $\pm$ 5.03
TransMIL	52.87 $\pm$ 3.82	5.51 $\pm$ 8.22	98.71 $\pm$ 2.42	66.82 $\pm$ 5.12
CellPlexer	65.11 $\pm$ 4.68	54.68 $\pm$ 7.13	71.49 $\pm$ 3.71	68.28 $\pm$ 5.31

Supplementary Table S13 Biomarker-specific ablation analysis on the TNBC cohort. Metrics include accuracy, sensitivity, specificity, and AUC (mean $\pm$ standard deviation across test folds).

Cell-type	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
CD31	63.10 $\pm$ 1.82	64.83 $\pm$ 10.05	61.18 $\pm$ 7.59	69.82 $\pm$ 3.95
CD45	67.36 $\pm$ 4.92	71.64 $\pm$ 10.41	63.66 $\pm$ 2.20	71.93 $\pm$ 6.98
CD68	59.49 $\pm$ 6.79	64.94 $\pm$ 19.30	53.47 $\pm$ 13.69	64.60 $\pm$ 6.96
CD4	62.05 $\pm$ 4.37	63.44 $\pm$ 19.38	59.08 $\pm$ 18.41	68.87 $\pm$ 7.31
FOXP3	61.33 $\pm$ 6.55	64.94 $\pm$ 15.74	56.86 $\pm$ 15.40	67.08 $\pm$ 8.57
CD8a	63.97 $\pm$ 9.08	62.27 $\pm$ 15.79	63.79 $\pm$ 19.22	69.16 $\pm$ 9.68
CD45RO	61.28 $\pm$ 5.52	68.49 $\pm$ 8.80	54.71 $\pm$ 14.36	66.80 $\pm$ 3.79
CD20	58.79 $\pm$ 5.23	65.99 $\pm$ 10.48	52.09 $\pm$ 13.74	68.21 $\pm$ 6.59
PD-L1	61.27 $\pm$ 4.33	55.29 $\pm$ 14.83	64.77 $\pm$ 9.59	69.33 $\pm$ 4.92
CD3e	59.40 $\pm$ 2.19	64.01 $\pm$ 14.33	54.51 $\pm$ 12.79	67.03 $\pm$ 6.61
CD163	63.71 $\pm$ 1.33	59.92 $\pm$ 8.31	66.93 $\pm$ 5.03	68.53 $\pm$ 5.09
PD-1	62.04 $\pm$ 5.47	67.24 $\pm$ 9.35	56.80 $\pm$ 13.45	69.00 $\pm$ 7.28
Ki67	63.77 $\pm$ 7.85	60.62 $\pm$ 21.33	64.77 $\pm$ 8.84	67.69 $\pm$ 11.66
Pan-CK	65.65 $\pm$ 1.71	58.84 $\pm$ 8.61	70.46 $\pm$ 6.53	70.09 $\pm$ 4.63
SMA	62.52 $\pm$ 6.61	64.20 $\pm$ 22.86	59.21 $\pm$ 19.45	68.64 $\pm$ 9.14

Supplementary Table S14 KRAS dataset performance comparison of state-of-the-art MIL methods with our best performing model. Metrics include accuracy, sensitivity, specificity, and AUC (mean $\pm$ standard deviation across test folds).

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	Balanced Accuracy (%)
ABMIL	56.72 $\pm$ 2.86	48.08 $\pm$ 5.54	63.17 $\pm$ 6.70	57.40 $\pm$ 3.82	56.89 $\pm$ 2.46
CLAM	58.41 $\pm$ 2.98	47.27 $\pm$ 9.02	65.67 $\pm$ 8.61	60.28 $\pm$ 2.86	56.19 $\pm$ 4.23
TransMIL	55.76 $\pm$ 3.98	61.56 $\pm$ 11.03	51.47 $\pm$ 2.92	58.81 $\pm$ 5.74	52.29 $\pm$ 3.56
CellPlexer	60.24 $\pm$ 6.41	59 $\pm$ 9.94	61.06 $\pm$ 13.46	62.56 $\pm$ 4.77	59.49 $\pm$ 6.80

Supplementary Table S15 BRAF dataset performance comparison of state-of-the-art MIL methods with our best performing model. Metrics include accuracy, sensitivity, specificity, and AUC (mean $\pm$ standard deviation across test folds).

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	Balanced Accuracy (%)
ABMIL	80.66 $\pm$ 5.96	63.49 $\pm$ 20.19	84.18 $\pm$ 8.14	81.90 $\pm$ 6.26	73.83 $\pm$ 9.93
CLAM	81.86 $\pm$ 2.80	46.36 $\pm$ 24.08	89.28 $\pm$ 5.41	83.03 $\pm$ 5.65	67.82 $\pm$ 11.27
TransMIL	78.88 $\pm$ 5.80	51.36 $\pm$ 18.56	84.62 $\pm$ 7.52	80.34 $\pm$ 6.08	67.99 $\pm$ 9.66
CellPlexer	75.24 $\pm$ 5.33	75.76 $\pm$ 10.77	75.17 $\pm$ 7.16	83.09 $\pm$ 6.59	75.46 $\pm$ 5.36

Supplementary Table S16 CellPlexer performance metrics using only the top 5% most-attended patches per slide, for TNBC pCR prediction, ILA recurrence and TCGA-BRCA survival prediction. Metrics include accuracy, sensitivity, specificity, and AUC (mean $\pm$ standard deviation across test folds).

Task	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
TNBC – pCR	62.22 $\pm$ 8.99	64.38 $\pm$ 5.40	60.13 $\pm$ 12.32	68.27 $\pm$ 7.26
ILA – recurrence	51.43 $\pm$ 11.31	50.48 $\pm$ 47.67	53.24 $\pm$ 49.06	49.08 $\pm$ 11.96
TCGA BRCA – survival	59.61 $\pm$ 17.74	69.33 $\pm$ 12.33	54.47 $\pm$ 30.52	65.42 $\pm$ 12.95

Supplementary Table S17 CellPlexer performance metrics using only random 5% patches per slide, for TNBC pCR prediction, ILA recurrence and TCGA-BRCA survival prediction. Metrics include accuracy, sensitivity, specificity, and AUC (mean $\pm$ standard deviation across test folds).

Task	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
TNBC – pCR	65.63 $\pm$ 0.67	67.73 $\pm$ 4.27	63.66 $\pm$ 4.64	69.19 $\pm$ 5.04
ILA – recurrence	52.84 $\pm$ 15.28	31.51 $\pm$ 23.74	65.86 $\pm$ 31.27	47.18 $\pm$ 15.81
TCGA BRCA – survival	60.26 $\pm$ 16.98	65.55 $\pm$ 22.70	57.79 $\pm$ 31.23	63 $\pm$ 20.75