Appendix A Appendix

A.1 Datasets

In this study, we used 6 datasets from 3 tissue types. The datasets included CoNSeP [21], NuCLS [22], PanNuke [54] (we refer to the breast and colon samples of this dataset as 2 different datasets), Lizard[55], and an internal dataset called Oracledataset. The details of each dataset can be found in Table A1

A.1.1 CoNSeP

consisted of 41 H&E tiles from colorectal tissues extracted from 16 whole slide images of a single patient. All tiles were in a 40× magnification scale with the size of 1,000×1,000 pixels. Cell types included 7 different categories including normal epithelial, malignant epithelial, inflammatory, endothelial, muscle, fibroblast, and miscellaneous. However, as suggested by the original paper, normal and malignant epithelial were grouped into the epithelial category, and the muscle, fibroblast, and endothelial cells were grouped into the spindle-shaped category. Therefore, the final 4 groups included epithelial, spindle-shaped, inflammatory, and miscellaneous. We followed the same train/test split of the original dataset.

A.1.2 NuCLS

included 1744 tiles of breast cancer images from the TCGA dataset collected from 18 institutions. The tiles had different sizes, but they were roughly 300×300 pixels. There were 12 different cell types available in this dataset: tumor, fibroblast, lymphocyte, plasma, macrophage, mitotic, vascular endothelium, myoepithelium, apoptotic body, neutrophil, ductal epithelium, and eosinophil. However, as suggested by the paper, these subtypes were grouped together into 5 superclasses including tumor (containing tumor and mitotic cells), stromal (containing fibroblast, vascular endothelium, and macrophage), sTILs (containing lymphocyte and plasma cells), apoptotic cells, and others. We used the first fold of this dataset for testing our model, while the rest of the dataset was used as the training data.

A.1.3 PanNuke

included tiles from 19 different tissue types among which breast and colon were the most common ones. The slides were scanned at a 40× magnification scale, and cell types included neoplastic, epithelial, inflammatory, connective, and dead cells. In this study, we only focused on the breast and colon cells as they have a larger population compared to the rest of the tissues. Although the colon portion of the dataset included all 5 cell categories, the breast tissue does not contain any dead cells, leading to 4 cell types for this tissue. This dataset contained 3 folds in total; the first fold was used for training purposes, and the third one was adopted for testing.

A.1.4 Lizard

included colon tissue tiles collected from 6 different sources across multiple sites in the world. The image tiles were scanned at the magnification scale of 20×. Cells were divided into 6 categories including epithelial, connective tissue, lymphocyte, plasma, neutrophil, and eosinophil cells. This dataset contained 3 folds, of which the first and third were used for training and testing, respectively.

A.1.5 Oracle

was an internal ovarian dataset collected in British Columbia, Canada. It included 192 TMA cores stained with H&E alongside the multi-color IHC scans of an adjacent slice for each core. IHC images captured different biomarkers including CD3 (T-cell), CD94 (Natural Killer cell), FoxP3 (T-cell), CD79a (B-cell), CD8 (killer T-cell), CD68 (macrophages), CD16a (natural killer cell), and PanCK+ (epithelial cancer cells), each of which could be used as a map to find the type of cells in the corresponding TMA core. To this end, we registered each IHC image with its corresponding core. However, due to the circular shape of the images, only 11 of them were visually matched. The cells included in these images were used as the test set.

A.1.6 SarcCell

was an internal soft tissue sarcoma dataset collected from epithelioid sarcoma cases. Whole tissue sections were obtained from formalin-fixed paraffin-embedded source blocks and stained with CD3 (T-cell, 1:100, Leica Biosystems), CD20 (B-cell, 1:200, Biocare), and CD208 (mature dendritic cells, 1:50, Novus Biologicals) by multiplex. Adjacent sections were stained by H&E. Stained sections were scanned at 40X and were co-registered together. IHC biomarkers were used as an indicator of the cell type for label generation. The distribution of extracted cells was extremely skewed towards T-cells (96% of all the cells). Although all the extracted cells were used in the training set, we selected a balanced subset of the training set as the test set (including all the B-cells and dendritic cells, but a portion of the T-cells).

Table A1 Datasets summary. The datasets are distributed across 4 tissue and 18 cell types to demonstrate the utility of the method.

Dataset	Cell Type Count	Tissue Type	Magnification Scale	Total Number of Cells	Cell Types	Number of Institutions
CoNSeP	4	Colon	40×	24,319	Spindle, epithelial, inflammatory, miscellaneous	1
NuCLS	5	Breast	20×	51,986	Tumor, stromal, sTIL, apoptotic, miscellaneous	18
PanNuke Breast	4	Breast	40×	34,806	Neoplastic, epithelial, inflammatory, and connective	> 2
PanNuke Colon	5	Colon	40×	23,831	Neoplastic, epithelial, inflammatory, connective, and dead	> 2
Lizard	6	Colon	20×	277,654	Epithelial, connective tissue, lymphocyte, plasma, neutrophil, and eosinophil	> 7
Oracle	3	Ovarian	40×	265,580	Tumor, T-cell, B-cell, and Natural Killer	1
SarcCell	3	Soft Tissue	40×	32,842	T-cell, B-cell, and dendritic	1

A.2 Evaluation Metrics

The clustering performance was measured based on multiple metrics, including Adjusted Mutual Information (AMI) [56], Adjusted Random Index (ARI) [57], and Purity of the identified cell clusters by the model and ground truth labels. In particular, AMI captures the agreement between two sets of assignments using mutual information while it is adjusted to mitigate the effect of chance on the score. On the other hand, ARI is the chance-adjusted form of the rand index [58], which calculates the quality of the clustering based on the number of matching instance pairs. Also, Purity measures how the samples within each cluster are similar to each other. In other words, it demonstrates whether each cluster is a mixture of different classes or not.

A.3 Results

A.3.1 Supervised Fine-tuning

Table A2 Fine-tuning accuracy for CoNSeP and NuCLS datasets. The supervised baselines demonstrate the performance of the HoVer-Net and NuCLS models on the CoNSeP and NuCLS datasets, respectively.

	CoNSeP	NuCLS
Supervised Baseline	79.6%	77.5%
1-Layer MLP with frozen backbone	79.3%	75.7%
1-Layer MLP with unfrozen backbone	79.6%	75.8%
2-Layer MLP with frozen backbone	76.8%	74.5%
2-Layer MLP with unfrozen backbone	**80.22%**	76.3%

Table A3 Fine-tuning of NuCLS with color normalization

	NuCLS
Fully-Supervised baseline	77.5%
1-2 1-Layer MLP with unfrozen backbone	78.1%
2-Layer MLP with unfrozen backbone	**78.2%**

A.3.2 Color Normalization

Since the CoNSeP dataset contains the images from one patient and they were scanned in one institution, we excluded it from this experiment.

Table A4 Color normalization effect on unsupervised cell clustering

	w/o Color normalization			w/ Color normalization		
	AMI	ARI	Purity	AMI	ARI	Purity
VOLTA	26.1%	**25.6%**	70.3%	**26.8%**	22.9%	**70.8%**

A.3.3 Model Size Comparison

Despite the fact that our model is capable of outperforming the supervised models, we noticed that it has 70% fewer parameters compared to HoVer-Net. However, one might relate this matter to the fact that Hover-Net performs 2 tasks of cell detection and classification, simultaneously. To address this, we only included the parameters of the common encoder and the classification head of this model to have an impartial comparison (Table A5).

Table A5 Model size comparison

	Inference Parameter Count
HoVer-Net	35.3M (common encoder: 25.6M, head: 9.7M)
Mask-RCNN in NuCLS	130M
Ours (1-Layer MLP)	**11.17M**
Ours (2-Layer MLP)	**11.43M**

A.3.4 Cancer subtype classification

To predict the subtype, the TMA core image was divided into 400×400 pixel patches, cells were divided into multiple clusters using the K-means algorithm, the number of cells within each predicted cluster was counted for each patch, and then patches were split into multiple groups based on the distributions of cell clusters. Afterward, we selected 100 patches from each group and predicted the clustering label of each original image based on the distribution of cell clusters in the selected patches. The final grouping process was conducted by using a hierarchical clustering algorithm with Ward's linkage method, preceded by dimensionality reduction using PCA.

A.4 Ablation Study

To study the effects of each component of the framework, we also conducted multiple ablation studies. In this section, we limited the training duration to 200 epochs and reduced the batch size to 512, and the studies were performed only on the CoNSeP and NuCLS datasets to decrease the time and memory requirements.

A.4.1 Memory Bank

Table A6 shows the results of the unsupervised clustering performance in the presence and absence of the negative sample memory bank. These results confirmed that the presence of the memory bank is critical to the performance of the model.

Comparing the results of this table with that of Table 1 (which was trained on 500 epochs with a batch size of 1024), we found that although the performance over the CoNSeP dataset dropped, the model produced almost similar results on the NuCLS dataset. We hypothesized that this must be related to the size of the dataset. Therefore, as the NuCLS dataset was 2× larger than the CoNSeP, the model could converge to a steady performance with even fewer training epochs.

Table A6 Memory bank ablation study

	CoNSeP			NuCLS		
	AMI	ARI	Purity	AMI	ARI	Purity
w/o Memory Bank	20.1%	**15.6%**	57.8%	22.5%	24.4%	66.6%
w/ Memory Bank	**21.3%**	15.4%	**59.4%**	**26.82%**	**28.7%**	**69.35%**

A.4.2 Masking cells in the environment patch

In this section, we studied the effects of the cell masking operation. In this regard, instead of masking all cells present in the environment patch, we only masked the target cell (the one that already exists in the associated single cell image fed into the Cell Block). The results (Table A7) showed that the masking operation increased the general performance of the model, implying that reduced bias towards other cells' types and more focus on the non-cellular environment such as tissue structure was important for environment integration.

A.4.3 Multi-cropping

We compared the effect of multi-cropping (global-local cropping augmentations) with the conventional local-local relation where both of the augmentation pipelines in the Cell Block contained the cropping operation. Surprisingly,

Table A7 Masking cells in the environment patch ablation study

	CoNSeP			NuCLS		
	AMI	ARI	Purity	AMI	ARI	Purity
w/o Masking	18.15%	11%	57%	26.3%	26.2%	**70.2%**
w/ Masking	**21.3%**	**15.4%**	**59.4%**	**26.82%**	**28.7%**	69.35%

although multi-cropping showed a significant improvement on NuCLS dataset, it reduced the performance of the model on CoNSeP (Table A8). We hypothesized that this was due to the fact that the local augmentations obtained in this setting were not diverse enough for the model to be properly trained as the CoNSeP dataset had fewer data samples. Therefore, we increased the number of training epochs on the CoNSeP dataset and compared the effect of multi-cropping again. Our results (Table A9) demonstrated the positive effect of multi-cropping on the performance when the model was trained for a longer time.

Table A8 Multi-cropping ablation study

	CoNSeP			NuCLS		
	AMI	ARI	Purity	AMI	ARI	Purity
w/o Multi-crop	**22.7%**	**21.4%**	**60.1%**	23%	26.2%	69%
w/ Multi-crop	21.3%	15.4%	59.4%	**26.82%**	**28.7%**	**69.4%**

Table A9 Ablation study of multi-cropping on CoNSeP with longer epochs

	CoNSeP		
	AMI	ARI	Purity
w/o Multi-crop	22.7%	**20.24%**	57.5%
w/ Multi-crop	**25.5%**	19.3%	**63.2%**

A.4.4 Ensembling

We also compared the impact of using the trained momentum encoder instead of the backbone in the testing phase, and as Table A10 suggests, using the momentum encoder improved the performance of the model. This finding was expected since the momentum encoder equation forms a model similar to Polyak-Ruppert averaging [32], aggregating the the encoder network weights across all training epochs.

Table A10 Ablation study of ensembling

	CoNSeP			NuCLS		
	AMI	ARI	Purity	AMI	ARI	Purity
w/o Ensembling	20.7%	13.5%	58%	24.4%	25.2%	69.3%
w/ Ensembling	**21.3%**	**15.4%**	**59.4%**	**26.82%**	**28.7%**	**69.4%**

A.5 Cancer Subtype Clustering

fig. A1 demonstrates the flowchart of the process for cancer subtype classification on ovarian and endometrial datasets.

fig. A3 depicts the boxplot of the cell area for each cell clusters of the ovarian dataset, and fig. A4 shows the proportion of tumor 2, 4, and 5 clusters with respect to all the tumor clusters.

Table A11 Distribution of cell clusters across epithelial ovarian cancer histotypes (with standard deviation).

	Mixture of Lymphocytes and Small Stromal	Tumor	Cells with Dark Nuclei	Stromal	Lymphocytes	Spindle-shape	Tumor and Stromal
Clear Cell	3.0 ±0.4%	59.5 ± 1.0%	1.6 ± 0.3%	8.8 ± 0.6%	0.4 ± 0.1%	12.5 ± 0.8%	14 ± 0.8%
Endometrioid	3.3 ±0.3%	57.2 ± 0.9%	0.4 ± 0.1%	10.6 ± 0.6%	0.3 ± 0.1%	0.3 ± 0.6%	18.4 ± 0.8%
High grade	3.6 ±0.3%	57.2 ± 0.9%	1.9 ± 0.2%	9.9 ± 0.6%	0.5 ± 0.1%	7.0 ± 0.5%	19.6 ± 0.8%
Low grade	8.0 ±0.4%	63.8 ± 0.7%	3.5 ± 0.3%	11.9 ± 0.5%	1.6 ± 0.1%	9.7 ± 0.4%	1.0 ± 0.2%
Mucinous	1.3 ±0.2%	54.7 ± 0.9%	0.1 ± 0.07%	6.9 ± 0.5%	0.2 ± 0.1%	6.5 ± 0.4%	30 ± 0.8%

Table A12 Distribution of cell clusters across endometrial cancer molecular subtypes (with standard deviation). The non-MMRd group encompasses p53abn and POLE tumors.

	Stromal and Inflammatory	Lymphocytes	Mixture of Elongated Cells	Tumor	Inflammatory and Small Stromal	Mix of Medium-Sized Cells
MMRd	6.8 ± 0.03%	9.6 ± 0.04%	10.3 ± 0.04%	54.3 ± 0.06%	0.4 ± 0.008%	18.6 ± 0.05%
non-MMRd	6.6 ± 0.03%	7.9 ± 0.03%	13.5 ± 0.04%	51.0 ± 0.06%	0.4 ± 0.007%	20.6 ± 0.04%
P53abn	6.3 ± 0.2%	7.3 ± 0.03%	12.9 ± 0.04%	53.9 ± 0.05%	0.4 ± 0.007%	19 ± 0.04%
POLE	6.8 ± 0.03%	8.2 ± 0.03%	13.8 ± 0.04%	49.1 ± 0.05%	0.3 ± 0.007%	21.56 ± 0.04%

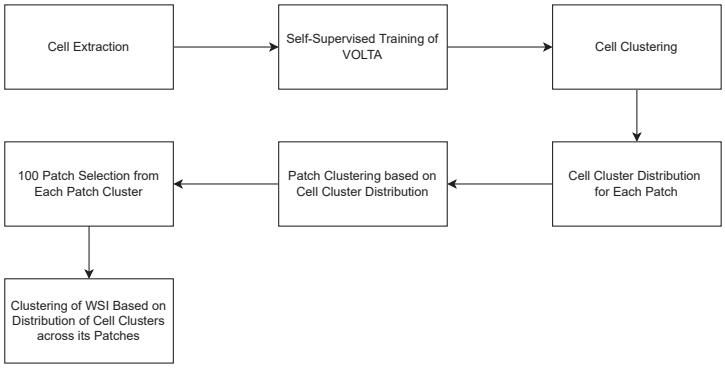

Fig. A1 Workflow of unsupervised cancer subtype classification based on the unsupervised cell representation learning

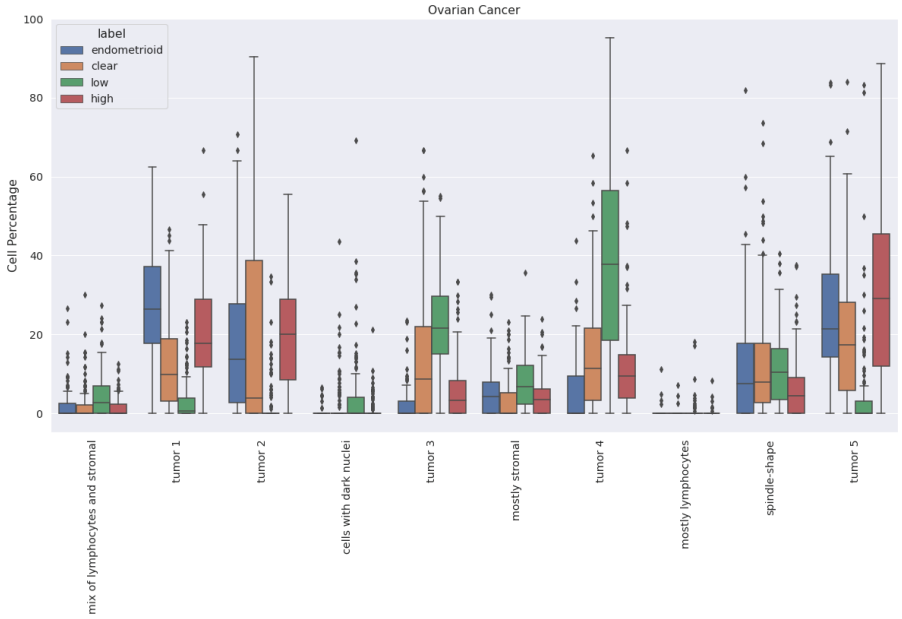

Fig. A2 Boxplot of cell cluster distribution across patches of the ovarian cancer dataset for all of the patients. The blue, orange, green, and red boxes represent the percentage of each cell cluster for endometrioid, clear-cell, low-grade serous, and high-grade serous histotypes of ovarian cancer.

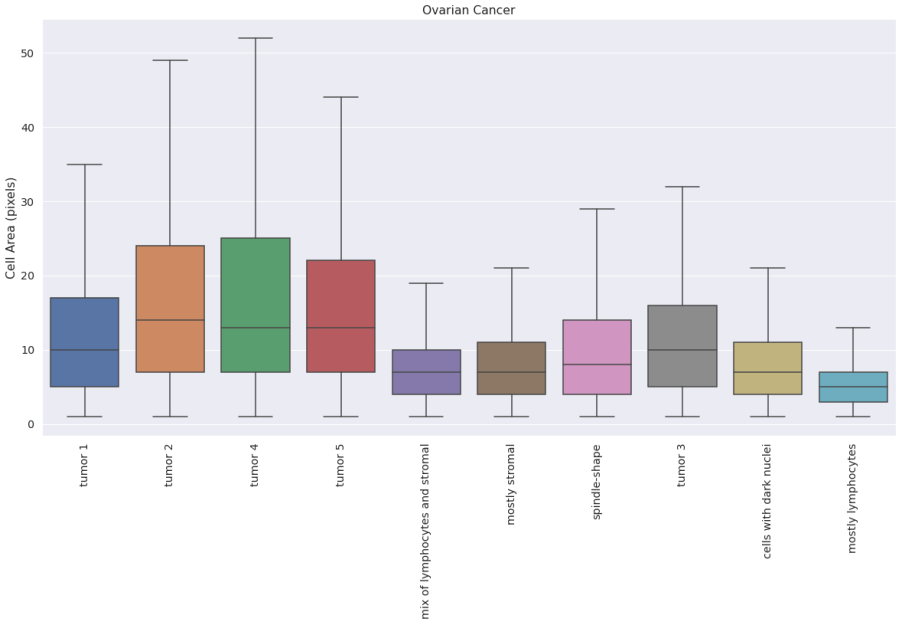

Fig. A3 Cell area boxplot for cell clusters of the ovarian dataset

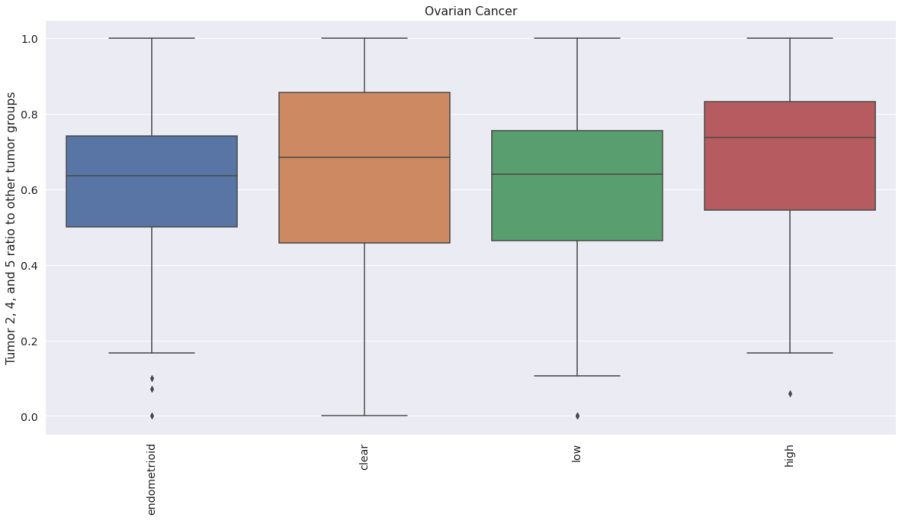

Fig. A4 Tumor 2, 4, and 5 clusters proportion ratio with respect to all the tumor clusters for each histotype

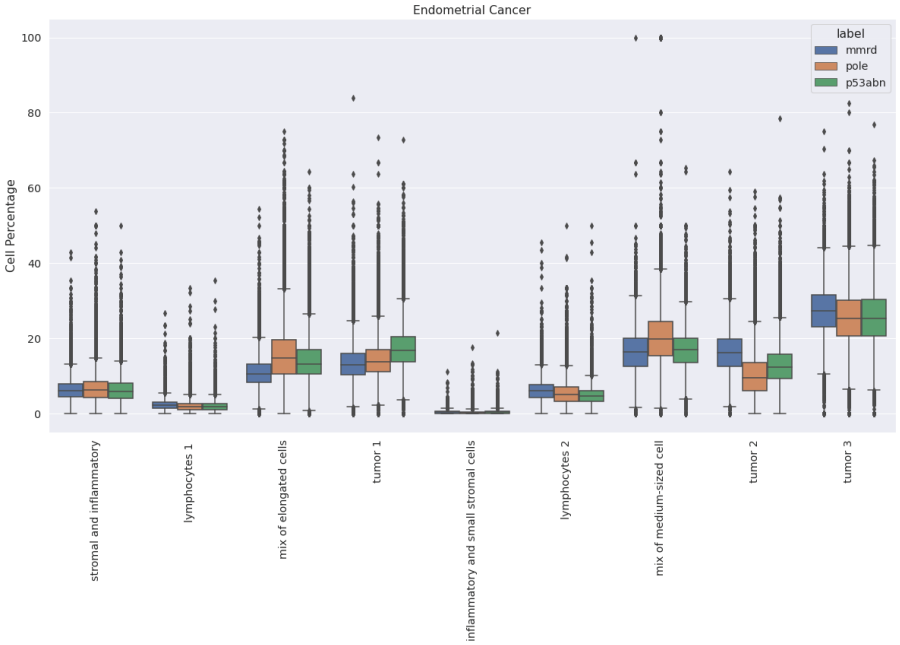

Fig. A5 Boxplot of cell cluster distribution across patches of the Endometrial Cancer for all of the patients

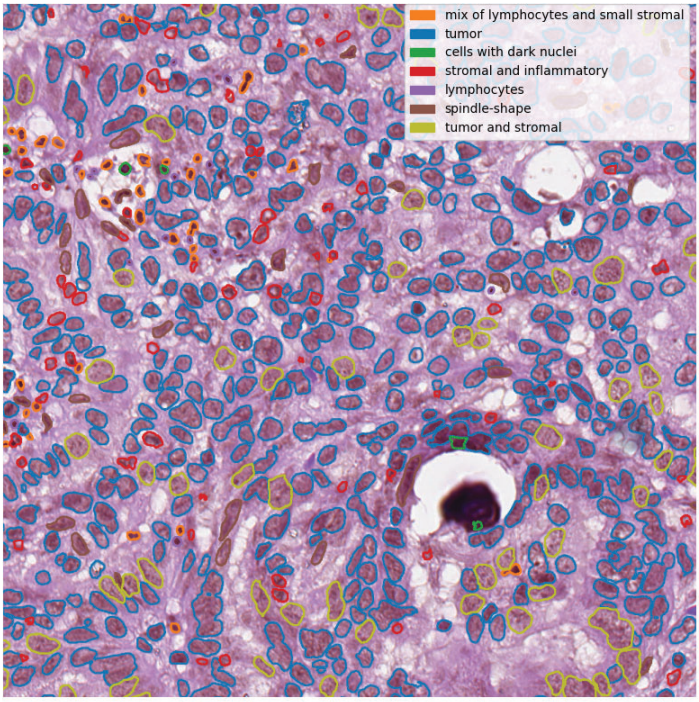

Fig. A6 Inferred cell cluster labels in a representative example of a low-grade serous ovarian carcinoma with psammomatous calcification. Cell cluster labels are depicted as coloured nuclear outlines, and represent morphologically-distinct clusters that are associated with cell types. Green captures cells with darkly staining nuclei, and brown labels spindled cells that are mostly stromal.

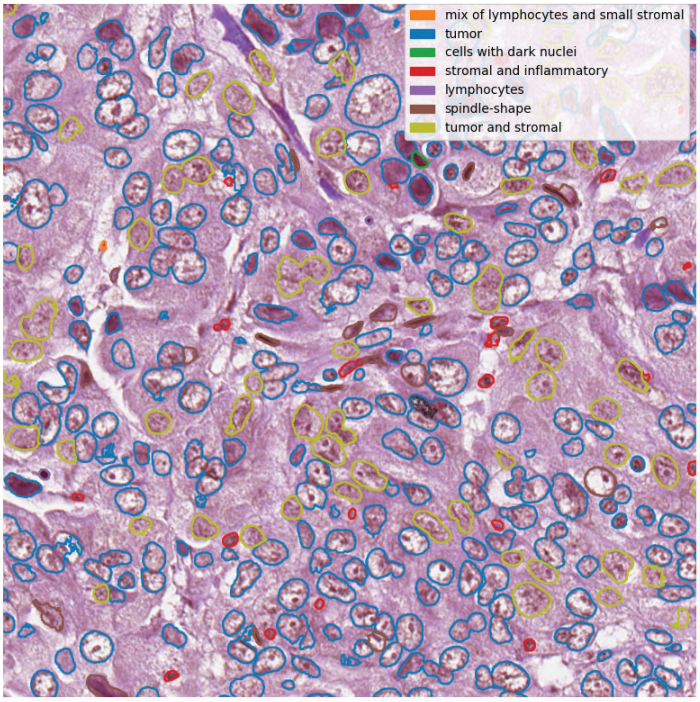

Fig. A7 Inferred cell cluster labels in a representative example of a high-grade serous ovarian carcinoma. Cell cluster labels are depicted as coloured nuclear outlines, and represent morphologically-distinct clusters that are associated with cell types. Green captures cells with darkly staining nuclei, and brown labels spindled cells that are mostly stromal.

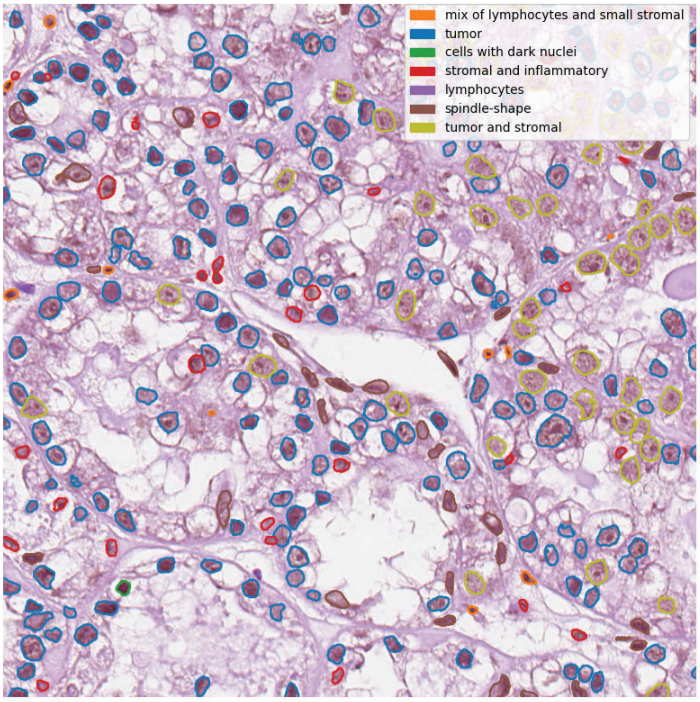

Fig. A8 Ovarian - clear cell overlay visualization

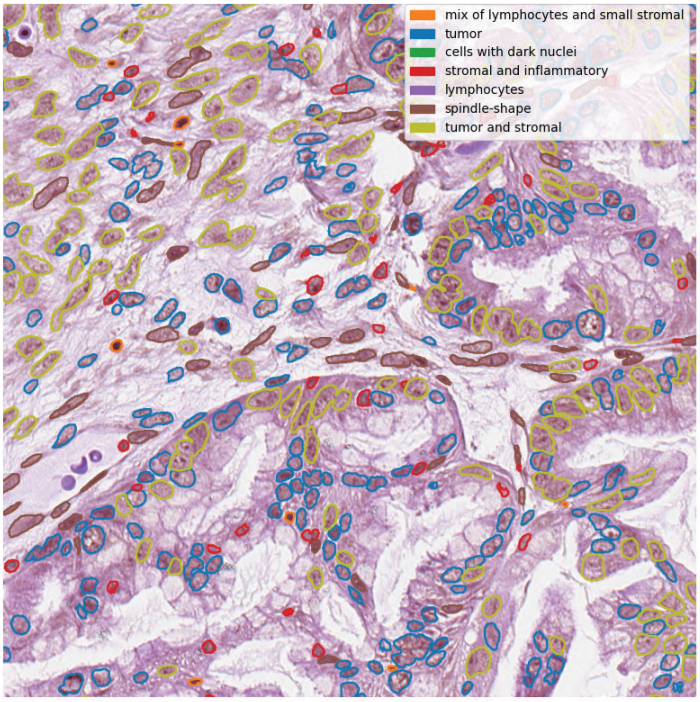

Fig. A9 Inferred cell cluster labels in a representative example of a mucinous ovarian carcinoma. Cell cluster labels are depicted as coloured nuclear outlines, and represent morphologically-distinct clusters that are associated with cell types. Green captures cells with darkly staining nuclei, and brown labels spindled cells that are mostly stromal. In this example, blue also captures reactive stromal cells with plump nuclei.

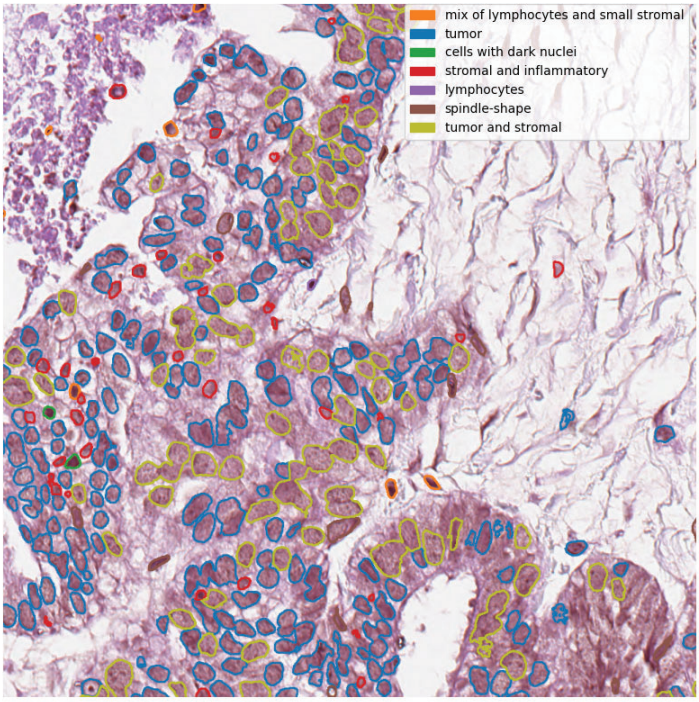

Fig. A10 Inferred cell cluster labels in a representative example of an endometrioid ovarian carcinoma. Cell cluster labels are depicted as coloured nuclear outlines, and represent morphologically-distinct clusters that are associated with cell types. Green captures cells with darkly staining nuclei, and brown labels spindled cells that are mostly stromal.

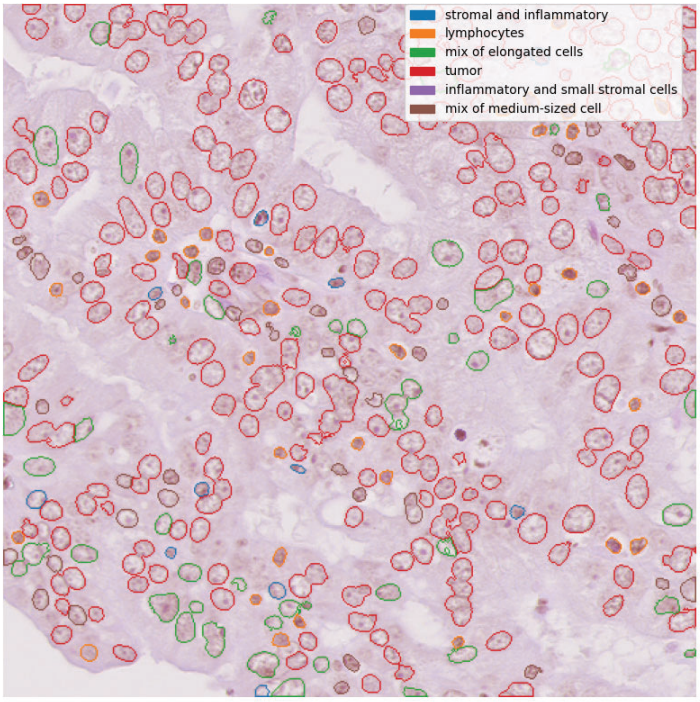

Fig. A11 Inferred cell cluster labels in a representative example of a POLE-mutated endometrial carcinoma. Cell cluster labels are depicted as colored nuclear outlines, and represent morphologically-distinct clusters that are associated with cell types. Green cells comprise a mixture of stromal and cancer cells with elongated nuclei. Brown denotes cells with intermediate-sized nuclei, comprising a mixture of tumor, inflammatory, and stromal cells. In this example, some cancer cells are labeled green, highlighting the uncertainty of cell type classification in morphologically pleomorphic POLE tumors.

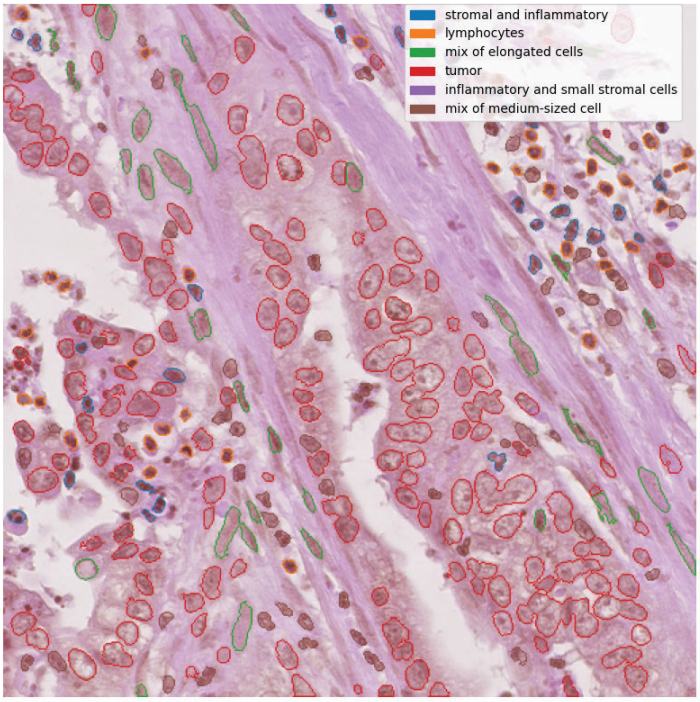

Fig. A12 Endometrial - MMRd overlay visualization

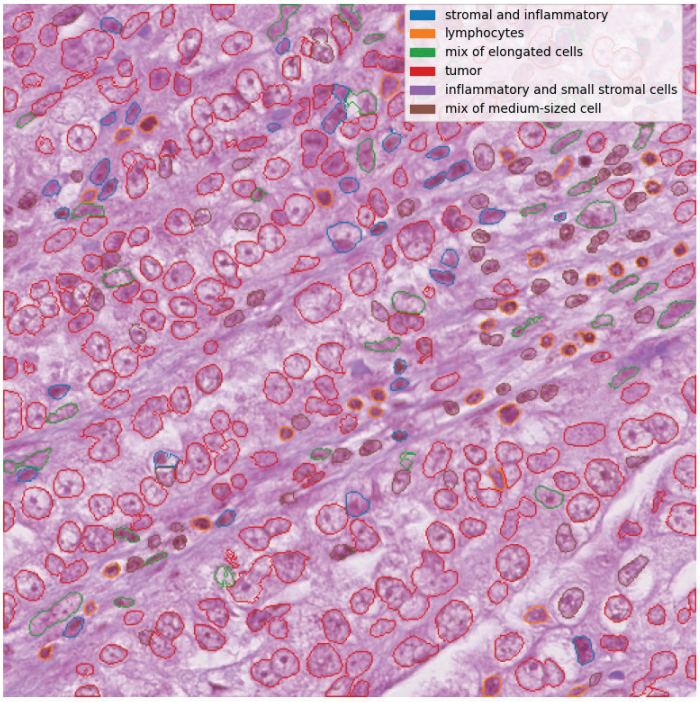

Fig. A13 Inferred cell cluster labels in a representative example of a p53-mutated endometrial carcinoma. Cell cluster labels are depicted as coloured nuclear outlines, and represent morphologically-distinct clusters that are associated with cell types. Green cells comprise mixture of stromal and tumour cells with elongated nuclei. Brown denotes cells with intermediate sized nuclei, comprising a mixture of tumour, inflammatory, and stromal cells.