SGPMIL: Sparse Gaussian Process Multiple Instance Learning

1. Additional Experimental Results

	ACC	AUC	ACE
ABMIL	$.964^{\star}_{.010}$	$\mathbf{.990}_{.005}$	$.032^{\star}_{.011}$
CLAM	$\underline{.978}_{.007}$	$\underline{.986}_{.007}$	$\underline{.021}_{.007}$
TransMIL	$.962^{\star}_{.009}$	$.980^{\star}_{.004}$	$.029^{\star}_{.014}$
DGRMIL	$.960^{\star}_{.012}$	$.980^{\star}_{.010}$	$.045^{\star}_{.016}$
BayesMIL	$.976_{.007}$	$.981^{\star}_{.006}$	$\mathbf{.020}_{.007}$
MixMIL	$.960^{\star}_{.009}$	$.978^{\star}_{.007}$	$.430^{\star}_{.002}$
AGP	$.883^{\star}_{.021}$	$.954^{\star}_{.019}$	$.069^{\star}_{.013}$
SGPMIL	$\mathbf{.980}_{.007}$	$\underline{.986}_{.005}$	$\underline{.021}_{.005}$

Table 1. Slide-level performance on CAMELYON16. * indicates significance based on a one-sided paired t-test across folds ($p < .05$).

	ACC (p)	AUC (p)	ACE (p)
ABMIL	$\mathbf{1.19e{-3}}$	$4.68e{-1}$	$\mathbf{2.72e{-3}}$
CLAM	$8.59e{-2}$	$2.40e{-1}$	$2.60e{-1}$
TransMIL	$\mathbf{6.40e{-5}}$	$\mathbf{5.66e{-3}}$	$\mathbf{2.11e{-2}}$
DGRMIL	$\mathbf{3.15e{-4}}$	$\mathbf{9.82e{-3}}$	$\mathbf{1.17e{-3}}$
BayesMIL	$\mathbf{4.43e{-2}}$	$\mathbf{3.07e{-5}}$	$3.79e{-1}$
MixMIL	$\mathbf{4.12e{-5}}$	$\mathbf{4.28e{-3}}$	$\mathbf{1.23e{-18}}$
AGP	$\mathbf{7.56e{-8}}$	$\mathbf{1.09e{-4}}$	$\mathbf{2.91e{-7}}$

Table 2. p-values from one-sided paired t-tests comparing SGPMIL with baseline models across folds on CAMELYON16 (slide level). Bold values indicate $p < .05$.

	ACC	AUC	ACE
ABMIL	$.953_{.003}$	$.973_{.009}$	$.039_{.008}$
CLAM	$.934^{\star}_{.014}$	$.953^{\star}_{.004}$	$.056_{.016}$
TransMIL	$.950_{.017}$	$.970_{.012}$	$.046_{.019}$
DGRMIL	$.947_{.024}$	$\underline{.974}_{.011}$	$\underline{.038}_{.022}$
BayesMIL	$\underline{.953}_{.023}$	$.973_{.021}$	$\mathbf{.033}_{.017}$
MixMIL	$.925^{\star}_{.015}$	$.963_{.014}$	$.410^{\star}_{.025}$
AGP	$.948^{\star}_{.026}$	$\mathbf{.976}_{.014}$	$.048_{.025}$
SGPMIL	$\mathbf{.955}_{.021}$	$.973_{.014}$	$.047_{.027}$

Table 3. Slide-level performance on TCGA-NSCLC. * indicates significance based on a one-sided paired t-test across folds ($p < .05$).

	ACC (p)	AUC (p)	ACE (p)
ABMIL	$2.08e{-1}$	$3.14e{-1}$	$3.66e{-1}$
CLAM	$\mathbf{3.93e{-2}}$	$\mathbf{2.18e{-2}}$	$1.50e{-1}$
TransMIL	$5.15e{-2}$	$1.26e{-1}$	$2.83e{-1}$
DGRMIL	$6.06e{-2}$	$2.59e{-1}$	$3.54e{-1}$
BayesMIL	$1.61e{-1}$	$2.89e{-1}$	$4.71e{-1}$
MixMIL	$\mathbf{1.19e{-2}}$	$5.38e{-2}$	$\mathbf{1.32e{-4}}$
AGP	$\mathbf{4.40e{-2}}$	$4.97e{-1}$	$2.22e{-1}$

Table 4. p-values from one-sided paired t-tests comparing SGPMIL with baseline models across folds on TCGA-NSCLC (slide level). Bold values indicate $p < .05$.

	ACC	κ	ACE
ABMIL	$.834^{\star}_{.064}$	$.910^{\star}_{.028}$	$.044^{\star}_{.015}$
CLAM	$\underline{.867}^{\star}_{.061}$	$.927^{\star}_{.025}$	$.031^{\star}_{.018}$
TransMIL	$.827^{\star}_{.074}$	$.911^{\star}_{.030}$	$.043^{\star}_{.021}$
DGRMIL	$.843^{\star}_{.097}$	$\underline{.933}^{\star}_{.047}$	$.036^{\star}_{.025}$
BayesMIL	$.850^{\star}_{.060}$	$.926^{\star}_{.031}$	$.031^{\star}_{.016}$
MixMIL	$.690^{\star}_{.054}$	$.870^{\star}_{.028}$	$.180^{\star}_{.010}$
AGP	$.802^{\star}_{.086}$	$.906^{\star}_{.047}$	$\mathbf{.026}_{.013}$
SGPMIL	$\mathbf{.900}_{.065}$	$\mathbf{.955}_{.037}$	$\underline{.028}_{.022}$

Table 5. Slide-level performance on PANDA. * indicates significance based on a one-sided paired t-test across folds ($p < .05$).

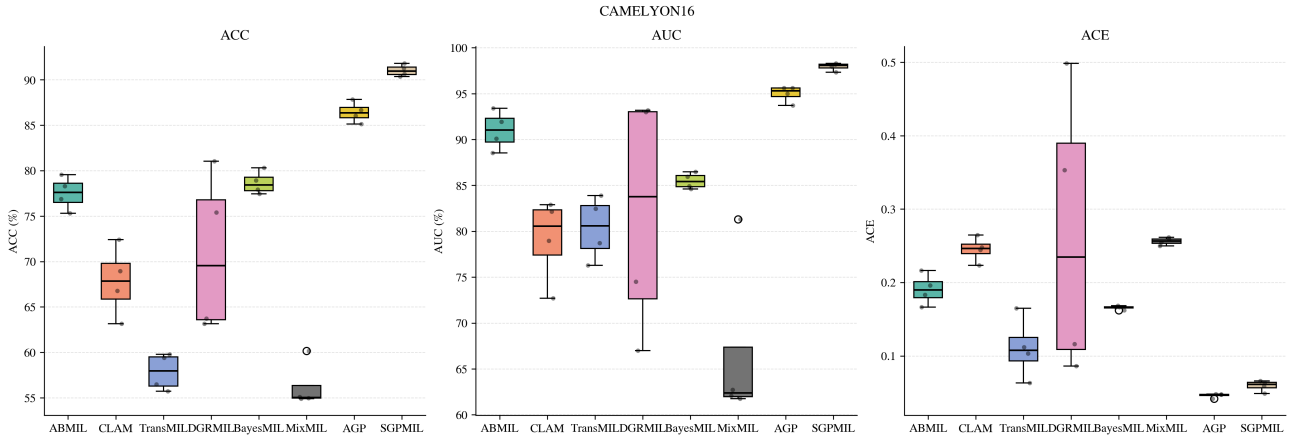

Figure 1. Instance-level performance across folds on CAMELYON16. Boxplots show the distribution of accuracy (ACC), area under the ROC curve (AUC), and adaptive expected calibration error (ACE) for each model. * denotes significance according to one-sided paired t-tests ($p < .05$).

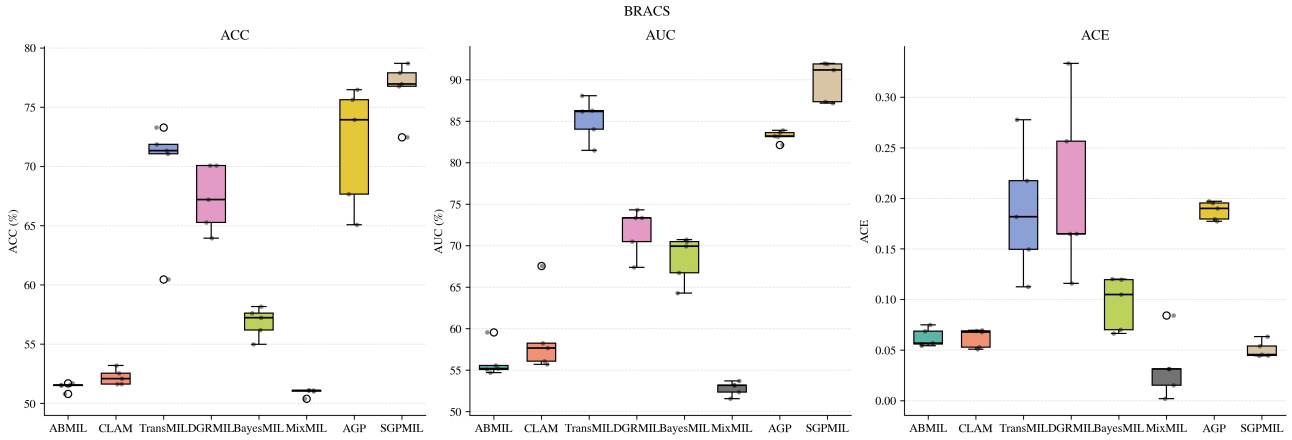

Figure 2. Instance-level performance across folds on BRACS. Boxplots show the distribution of accuracy (ACC), area under the ROC curve (AUC), and adaptive expected calibration error (ACE) for each model. * denotes significance according to one-sided paired t-tests ($p < .05$).

	ACC (p)	κ (p)	ACE (p)
ABMIL	**1.86e−3**	**8.04e−4**	**1.01e−3**
CLAM	**7.51e−3**	**2.03e−3**	**4.44e−3**
TransMIL	**4.50e−3**	**1.53e−3**	**1.56e−3**
DGRMIL	**1.51e−2**	**3.50e−2**	**2.69e−2**
BayesMIL	**1.60e−3**	**2.92e−3**	**6.92e−3**
MixMIL	**2.23e−3**	**2.95e−3**	**8.19e−5**
AGP	**1.57e−2**	**1.71e−2**	2.31e−1

Table 6. p-values from one-sided paired t-tests comparing SGPMIL with baseline models across folds on PANDA (slide level). Bold values indicate $p < .05$.

	ACC	AUC	ACE
ABMIL	$.694^{\star}_{.010}$	$.852^{\star}_{.025}$	$.175^{\star}_{.007}$
CLAM	$.699^{\star}_{.034}$	$.850^{\star}_{.021}$	$.183^{\star}_{.011}$
TransMIL	$.676^{\star}_{.005}$	$.826^{\star}_{.032}$	$.186^{\star}_{.012}$
DGRMIL	$\underline{.703}^{\star}_{.033}$	$.818^{\star}_{.035}$	$.186^{\star}_{.023}$
BayesMIL	$.648^{\star}_{.058}$	$.829^{\star}_{.022}$	$.183^{\star}_{.028}$
MixMIL	$.662^{\star}_{.031}$	$\underline{.855}^{\star}_{.010}$	$.254^{\star}_{.002}$
AGP	$.634^{\star}_{.030}$	$.830^{\star}_{.010}$	$\mathbf{.134}_{.014}$
SGPMIL	$\mathbf{.736}_{.029}$	$\mathbf{.870}_{.026}$	$\underline{.142}_{.032}$

Table 7. Slide-level performance on BRACS. * indicates significance based on a one-sided paired t-test across folds ($p < .05$).

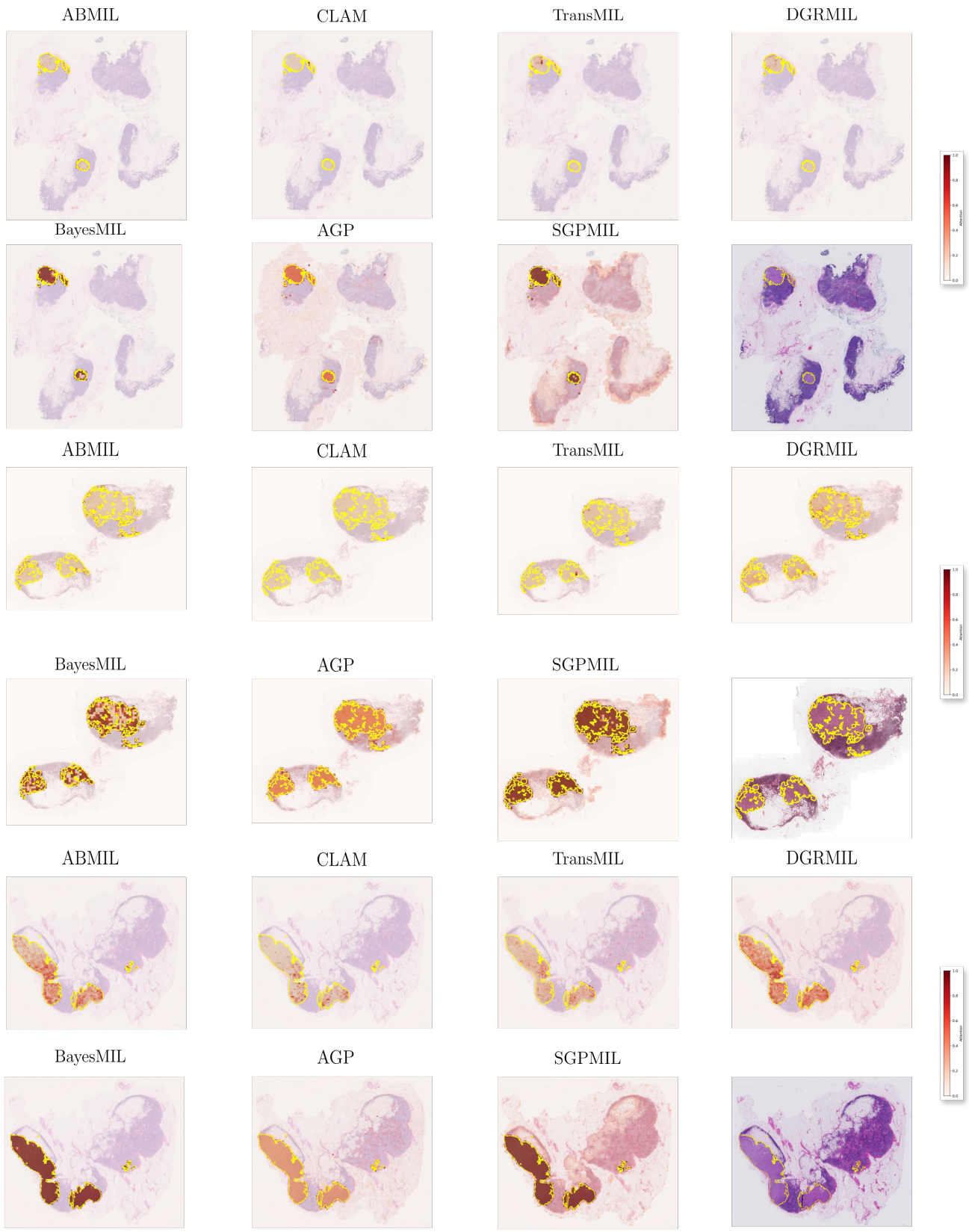

Figure 3. Additional attention heatmaps on CAMELYON16. Ground-truth tumor annotations are overlaid as yellow contours.

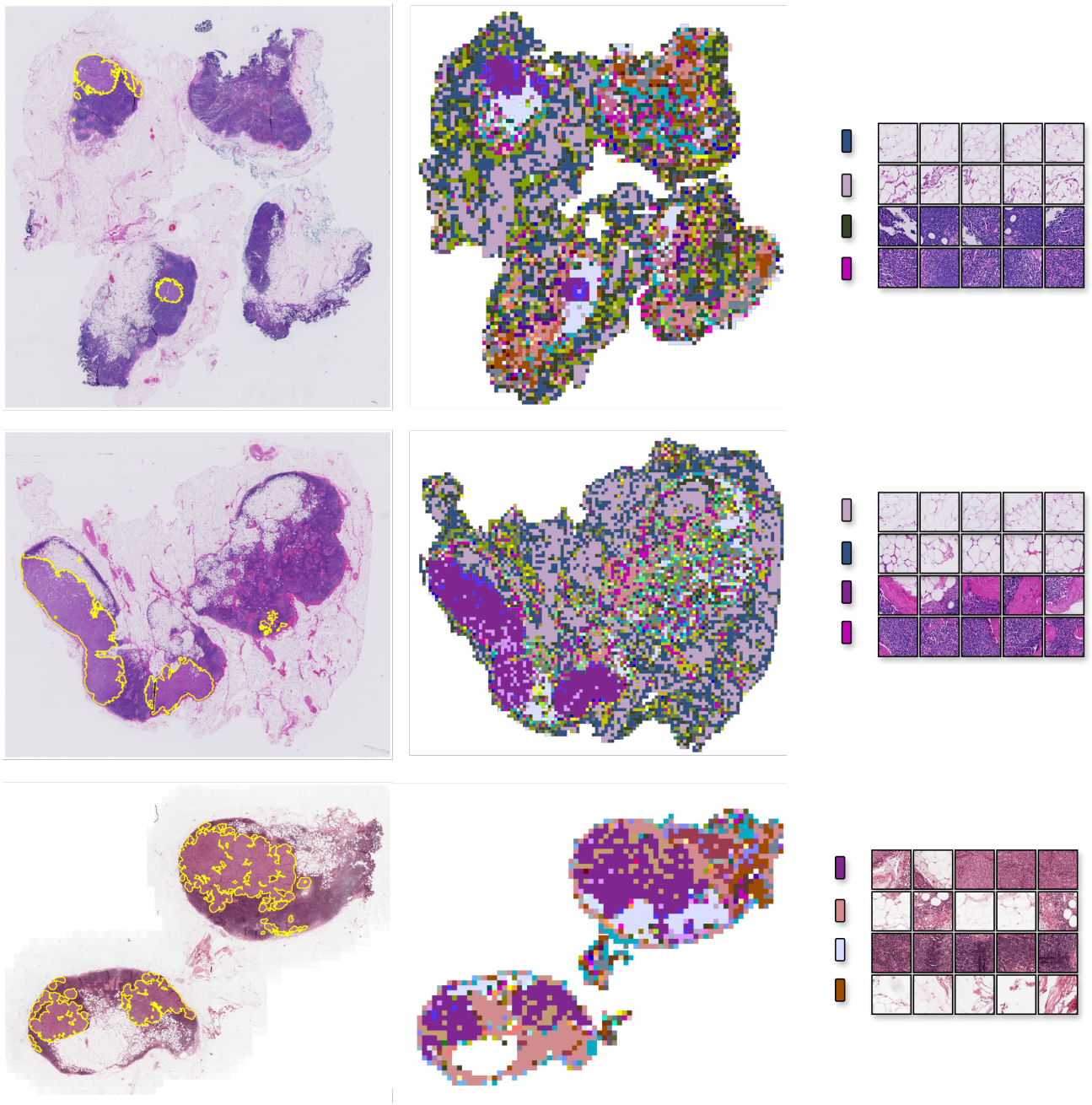

Figure 4. Additional inducing point label maps. Left: CAMELYON16 test set WSIs with ground-truth tumor annotations (yellow contours). Middle: label maps where each patch embedding is assigned to its most similar inducing point based on cosine similarity. Right: top-5 most similar patches for selected inducing points. These visualizations highlight how inducing points specialize in different tissue morphologies, including tumor, stroma, and interface regions.

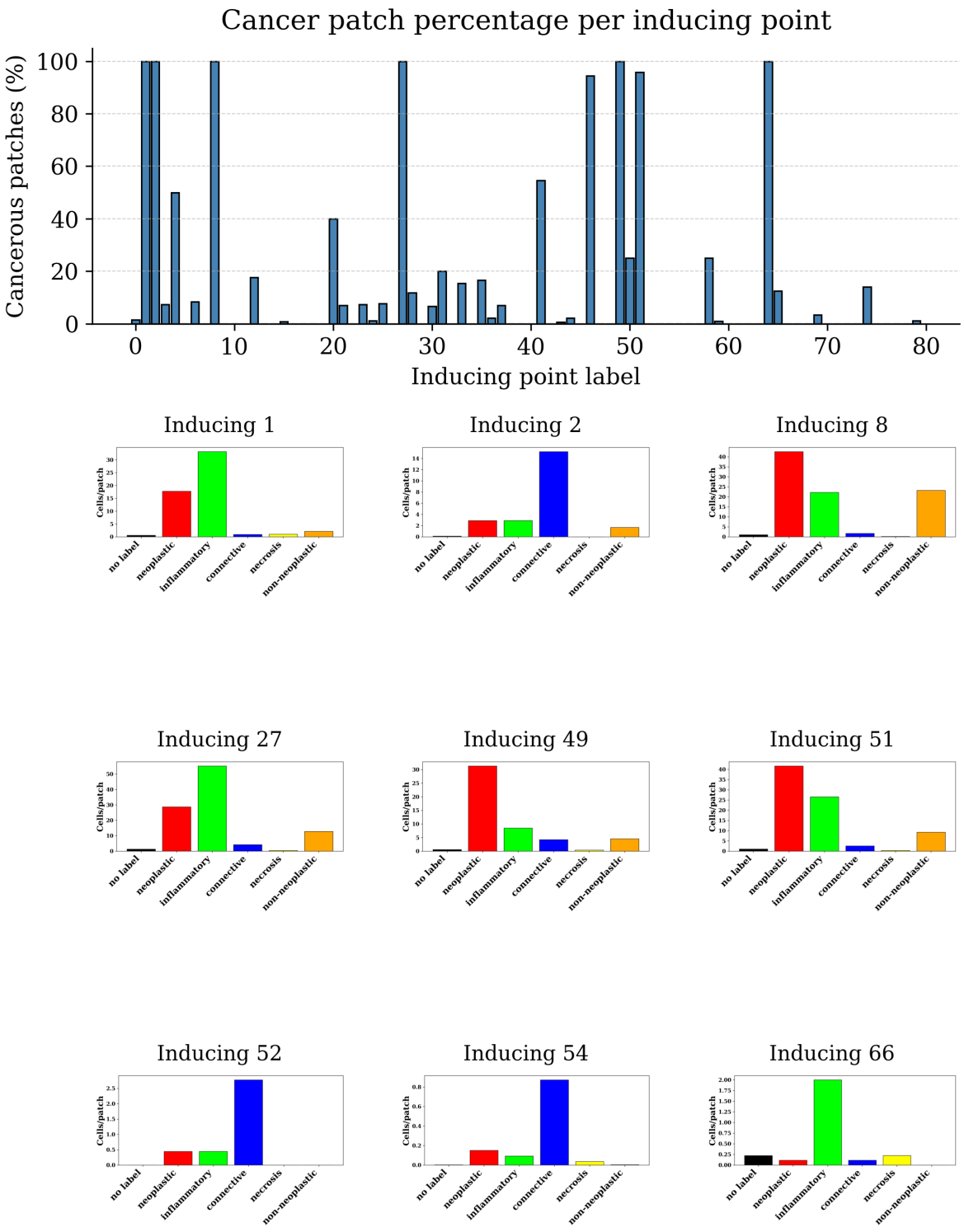

Figure 5. Cancer patch percentage per inducing point (top) and nucleus type distributions (bottom) for selected inducing points in CAMELYON16. Percentages were computed as the fraction of cancerous to total patches assigned to each inducing point in the slide. Inducing points 1, 2, 8, 27, 49, and 51 are highly enriched in cancerous regions and show a predominance of neoplastic nuclei, with inflammatory cells often the next most abundant type. By contrast, inducing points not associated with cancerous patches (52, 54, 66) mainly reflect connective or non-tumor tissue, with sparse isolated tumor cells, consistent with the lack of pixel-level annotations in CAMELYON16.

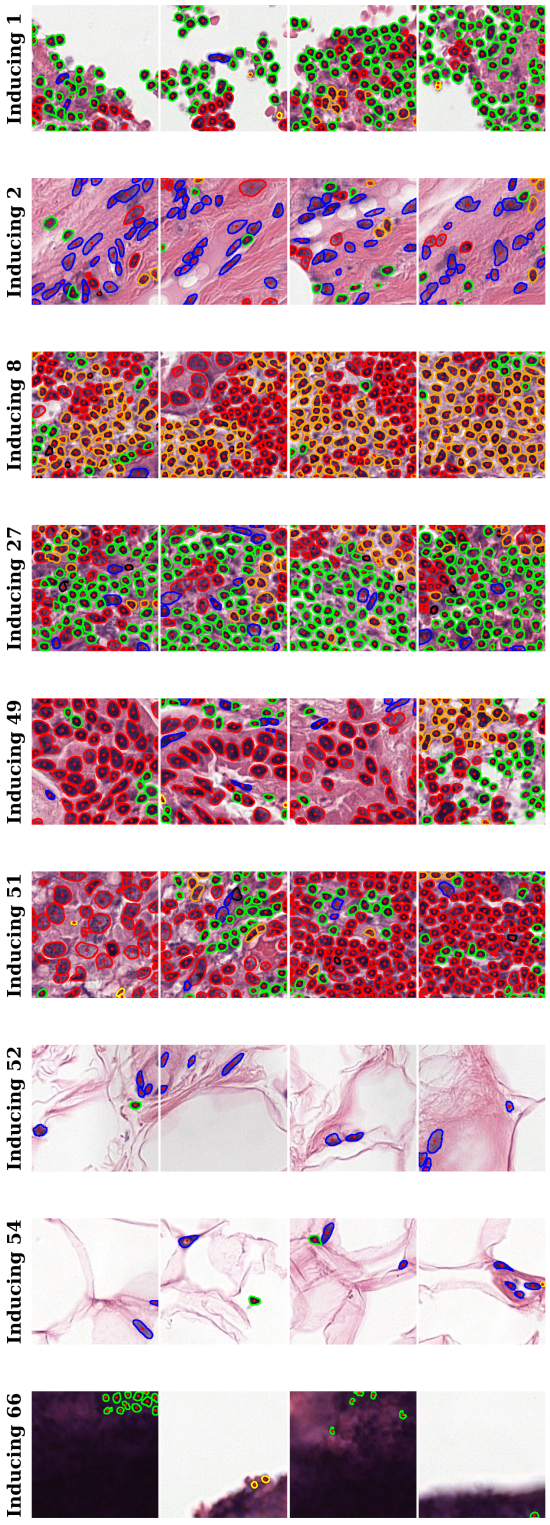

Figure 6. Representative image patches most similar to inducing points in CAMELYON16 slide 75. Each row corresponds to one inducing point (labels shown on the left), with four example patches illustrating the tissue context and associated nuclear types as segmented by HoverNet. Inducing points 1, 2, 8, 27, 49, and 51 predominantly align with tumor-rich regions, while others (e.g., 52, 54, 66) capture connective or non-tumor tissue, occasionally containing isolated tumor cells.

	ACC (p)	AUC (p)	ACE (p)
ABMIL	$\mathbf{1.81e{-3}}$	$\mathbf{3.75e{-2}}$	$\mathbf{1.53e{-2}}$
CLAM	$\mathbf{3.11e{-2}}$	$\mathbf{4.64e{-2}}$	$\mathbf{7.92e{-3}}$
TransMIL	$\mathbf{7.29e{-4}}$	$\mathbf{1.30e{-2}}$	$\mathbf{8.27e{-3}}$
DGRMIL	$\mathbf{3.07e{-2}}$	$\mathbf{8.13e{-3}}$	$\mathbf{6.68e{-3}}$
BayesMIL	$\mathbf{3.69e{-3}}$	$\mathbf{8.01e{-3}}$	$\mathbf{1.03e{-2}}$
MixMIL	$\mathbf{8.88e{-4}}$	$\mathbf{2.94e{-2}}$	$\mathbf{1.49e{-4}}$
AGP	$\mathbf{4.81e{-4}}$	$\mathbf{3.85e{-3}}$	1.80e−1

Table 8. p-values from one-sided paired t-tests comparing SGPMIL with baseline models across folds on BRACS (slide level). Bold values indicate $p < .05$.

	ACC (p)	AUC (p)	ACE (p)
ABMIL	$\mathbf{1.27e{-4}}$	$\mathbf{3.84e{-3}}$	$\mathbf{6.41e{-4}}$
CLAM	$\mathbf{8.92e{-4}}$	$\mathbf{2.29e{-3}}$	$\mathbf{1.60e{-4}}$
TransMIL	$\mathbf{4.53e{-5}}$	$\mathbf{1.14e{-3}}$	$\mathbf{4.08e{-2}}$
DGRMIL	$\mathbf{8.89e{-3}}$	$\mathbf{4.82e{-2}}$	$\mathbf{6.19e{-4}}$
BayesMIL	$\mathbf{2.31e{-4}}$	$\mathbf{1.12e{-4}}$	$\mathbf{1.51e{-5}}$
MixMIL	$\mathbf{3.03e{-5}}$	$\mathbf{3.49e{-3}}$	$\mathbf{3.96e{-7}}$
AGP	$\mathbf{4.15e{-3}}$	$\mathbf{6.15e{-3}}$	9.75e−1

Table 9. p-values from one-sided paired t-tests comparing SGPMIL with baseline models across folds on CAMELYON16 (instance level). Bold values indicate $p < .05$.

	ACC (p)	AUC (p)	ACE (p)
ABMIL	$\mathbf{1.03e{-5}}$	$\mathbf{2.26e{-5}}$	8.75e−2
CLAM	$\mathbf{1.50e{-5}}$	$\mathbf{2.17e{-4}}$	7.37e−2
TransMIL	$\mathbf{3.75e{-3}}$	$\mathbf{6.65e{-3}}$	$\mathbf{4.17e{-3}}$
DGRMIL	$\mathbf{1.79e{-4}}$	$\mathbf{1.20e{-4}}$	$\mathbf{8.47e{-3}}$
BayesMIL	$\mathbf{9.11e{-5}}$	$\mathbf{3.22e{-4}}$	$\mathbf{5.40e{-3}}$
MixMIL	$\mathbf{1.01e{-5}}$	$\mathbf{2.81e{-6}}$	9.11e−1
AGP	$\mathbf{3.17e{-2}}$	$\mathbf{3.17e{-3}}$	$\mathbf{8.10e{-7}}$

Table 10. p-values from one-sided paired t-tests comparing SGPMIL with baseline models across folds on BRACS (instance level). Bold values indicate $p < 0.05$.

	WSI			INSTANCE		
U	ACC	AUC	ACE	ACC	AUC	ACE
16	.971	.977	.045	<u>.893</u>	.963	.271
32	<u>.979</u>	<u>.985</u>	<u>.026</u>	.880	<u>.970</u>	**.046**
64	.964	.981	.041	.832	.910	.210
80	**.980**	**.986**	**.021**	**.914**	**.973**	<u>.051</u>

Table 11. Performance of SGPMIL on CAMELYON16 for varying inducing points.

Model	Training (s)	Inference (s)	Params (M)
ABMIL	5.5*	0.8*	0.66
CLAM	7.0	0.9	0.92
TransMIL	13.4	1.2	2.67
DGRMIL	16.7	1.5	4.34
BayesMIL	9.5	1.1	1.32
MixMIL	11	1.5	1.57
AGP	31.0	6.0	1.21
SGPMIL	9.0**	1.0**	1.21

Table 12. Training and inference times (in seconds) and model sizes (number of trainable parameters in millions, M). Training times are averaged over 30 epochs, while inference times correspond to processing the full test set of 129 slides. For MixMIL, we report the full variational posterior variant (non-mean field approximation). * marks the overall fastest model (ABMIL) and ** highlights the fastest probabilistic model among AGP, BayesMIL, MixMIL and SGPMIL.

	WSI		
Kernel	ACC	AUC	ACE
RBF	$.980_{.007}$	$.986_{.005}$	$.021_{.005}$
Linear	$.973_{.010}$	$.985_{.005}$	$.048_{.008}$
Matern	$.985_{.006}$	$.988_{.003}$	$.024_{.006}$

Table 13. Performance of SGPMIL on CAMELYON16 for various kernels, across folds.

	INSTANCE		
Kernel	ACC	AUC	ACE
RBF	$.914_{.006}$	$.973_{.004}$	$.051_{.008}$
Linear	$.851_{.017}$	$.950_{.020}$	$.113_{.016}$
Matern	$.850_{.011}$	$.960_{.012}$	$.062_{.014}$

Table 14. Performance of SGPMIL on CAMELYON16 for various kernels, across folds.

	Bag-level		Instance-level			
	AUC	ACE	F1	FROC	AUC	ACE
ABMIL	0.986	0.030	0.958	0.816	0.987	0.033
CLAM	0.989	0.021	0.952	0.813	**0.999**	0.039
TransMIL	0.990	0.021	**0.980**	0.826	0.993	0.025
DGRMIL	**0.996**	<u>0.016</u>	<u>0.979</u>	<u>0.827</u>	0.996	**0.013**
BayesMIL	0.990	0.019	0.968	0.824	0.994	<u>0.019</u>
AGP	0.994	0.046	0.957	0.820	0.981	0.083
SGPMIL	<u>0.995</u>	**0.009**	0.978	**0.830**	<u>0.998</u>	0.050

Table 15. Bag-level and instance-level performance for the MNIST-bags dataset. Bags are comprised of 9 instances each. All baselines are trained end-to-end with a CNN as initial feature extractor.