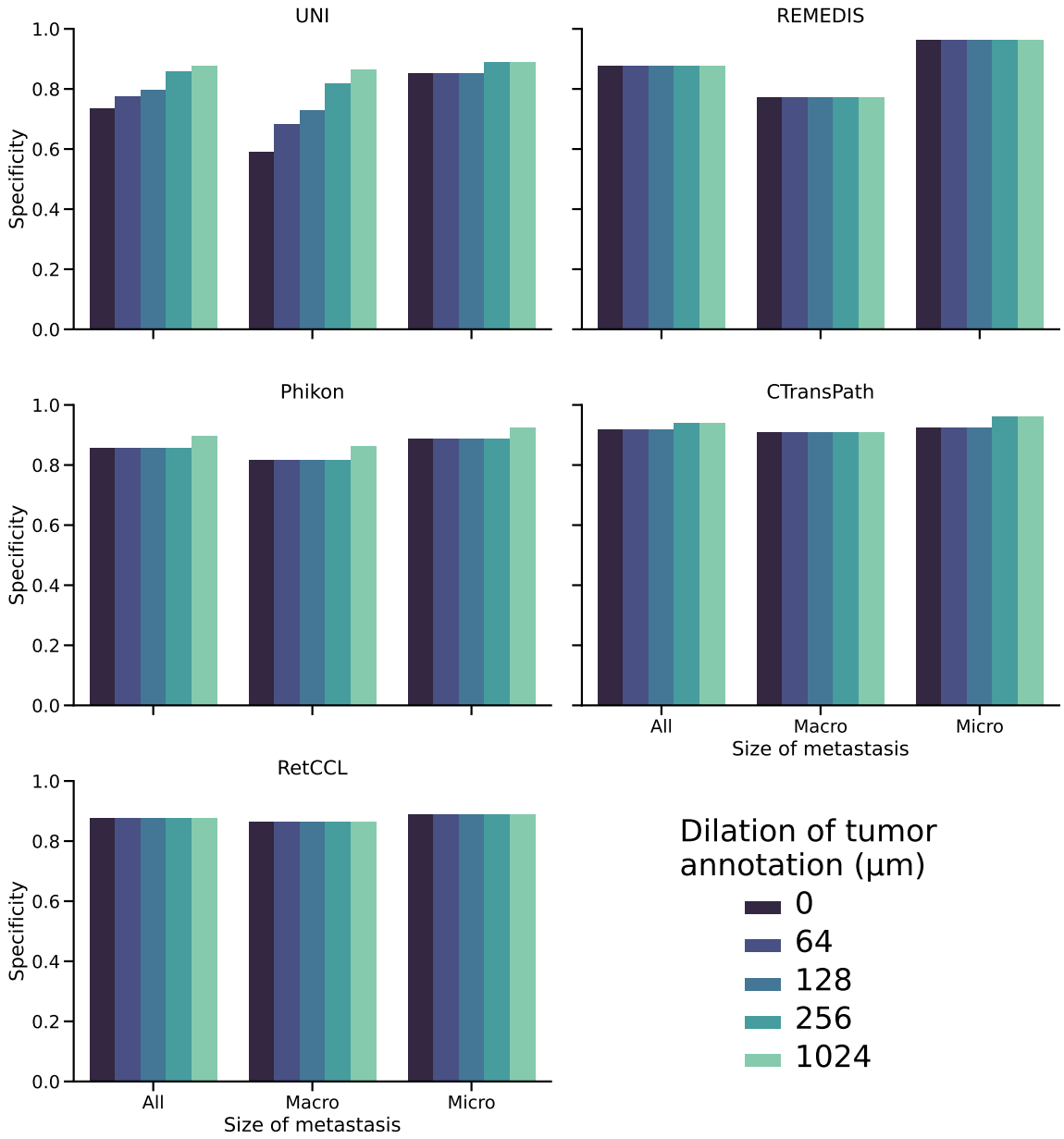
Encoder	Balanced Accuracy	Sensitivity	Specificity	Precision	Weighted F1
UNI	0.982 (0.014)	0.976 (0.017)	0.988 (0.022)	0.980 (0.033)	0.983 (0.015)
REMEDIS	0.922 (0.031)	0.861 (0.062)	0.983 (0.014)	0.968 (0.024)	0.935 (0.025)
Phikon	0.907 (0.083)	0.845 (0.158)	0.970 (0.023)	0.943 (0.049)	0.920 (0.070)
CTransPath	0.858 (0.016)	0.784 (0.045)	0.933 (0.023)	0.879 (0.034)	0.874 (0.013)
RetCCL	0.745 (0.016)	0.567 (0.042)	0.922 (0.049)	0.827 (0.074)	0.777 (0.019)

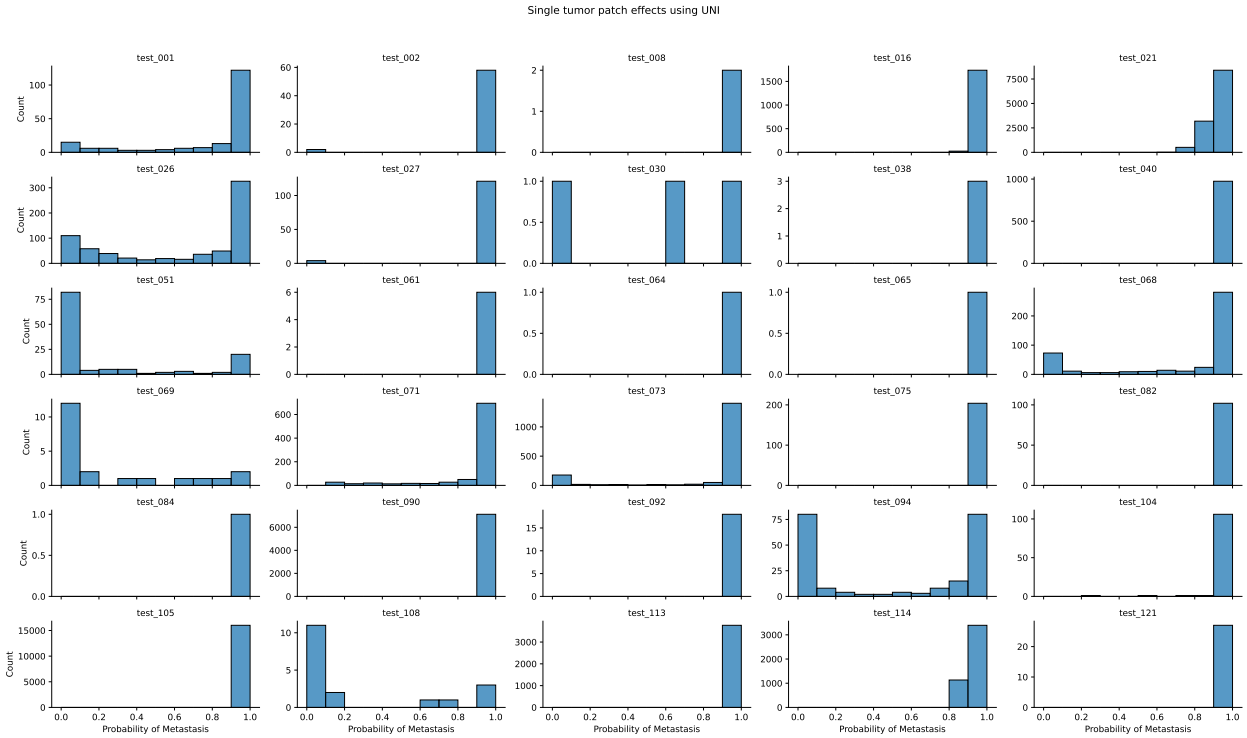
Supplementary Table 1: Performance of metastasis detection models. Values shown are mean (standard deviation) across five random initializations. All models used the same hyperparameters and data splits. The training, validation, and test sets consisted of 243, 27, and 129 specimens, respectively. The test set consisted of 80 negative specimens, 22 samples with macrometastases, and 27 specimens with micrometasetases.

Encoder	Balanced Accuracy	Sensitivity	Specificity	Precision	Weighted F1
UNI	1.000	1.000	1.000	1.000	1.000
REMEDIS	0.949	0.898	1.000	1.000	0.961
Phikon	0.955	0.959	0.950	0.922	0.954
CTransPath	0.885	0.857	0.912	0.857	0.891
RetCCL	0.769	0.612	0.925	0.833	0.799

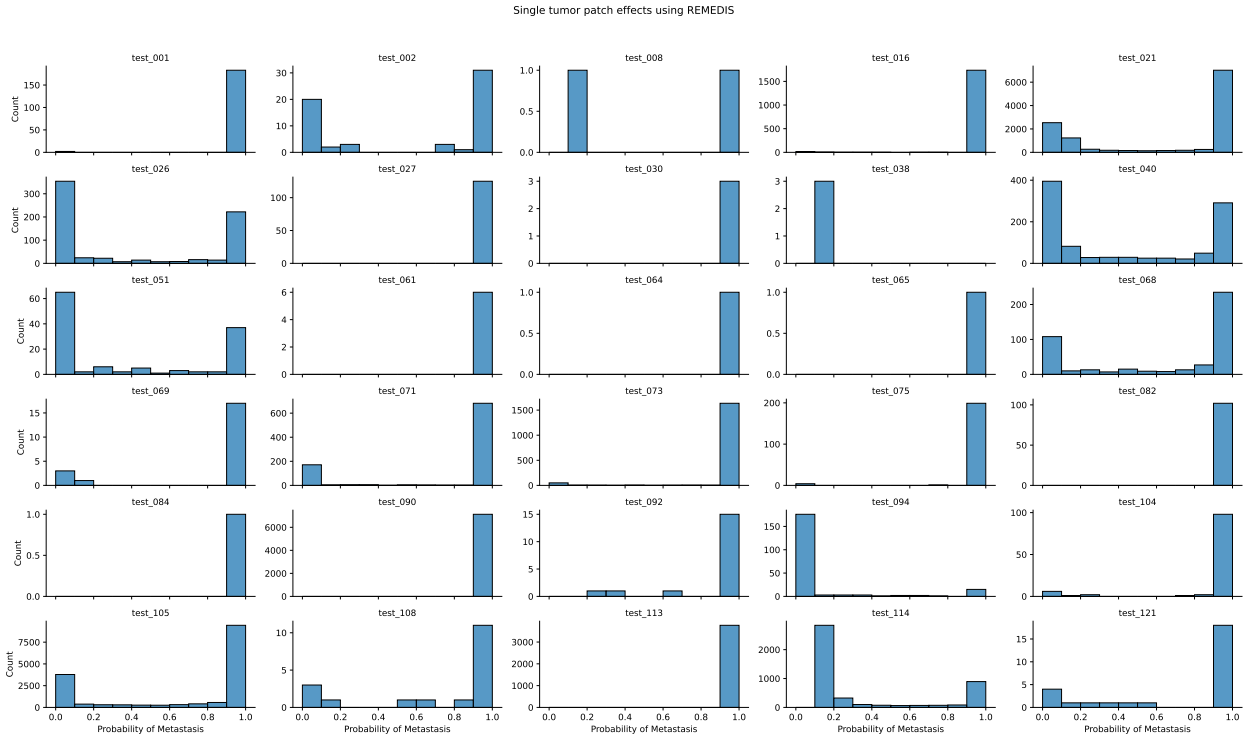
Supplementary Table 2: Performance of the best metastasis detection models for each encoder. This table lists the performance metrics for the single best random initialization for each encoder. Performance was calculated on the CAMELYON16 test set. The models represented here were the ones used for downstream experiments in the present report.



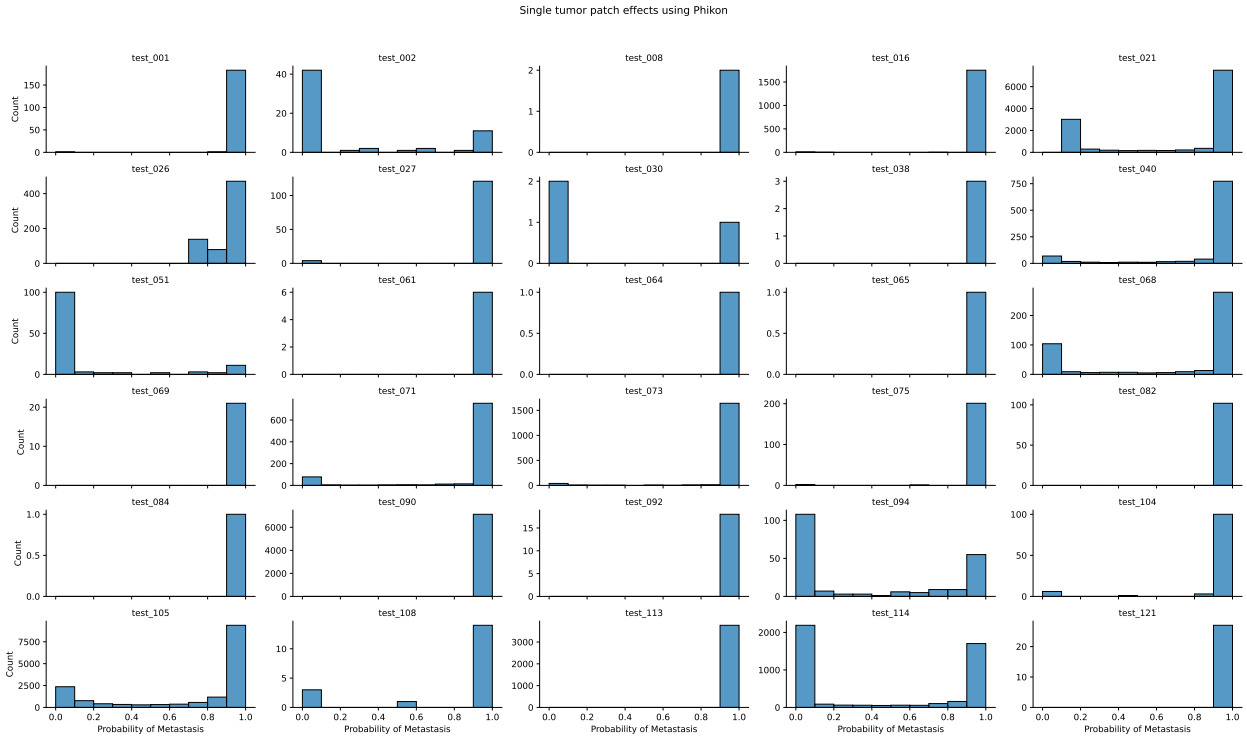
Supplementary Figure 1: Perituromal tissue may affect metastasis detection. The expert tumor annotations in the positive specimens (n=49) of the CAMELYON16 test set were systematically dilated, and then patches that intersected with these dilated regions were removed. This effectively removed tumor tissue and varying amount of peritumoral tissue, rendering the specimens negative for metastasis. Specificity of model outputs (true negative rate) was calculated. Dilation of tumor annotations did not change model outputs in REMEDIS-based or RetCCL-based models, suggesting that peritumoral tissue did not drive model outputs. The largest dilation (i.e., 1024 µm) increased specificity in Phikon-based and CTransPath-based models, suggesting that peritumoral was responsible to some degree for false positive predictions. The UNI-based model demonstrated a graded effect of dilation, suggesting that the tissue surrounding the tumor was driving positive model predictions. This effect was particularly strong in macrometastases (n=22). It appears that the UNI-based model relies on peritumoral tissue to some degree for positive predictions.



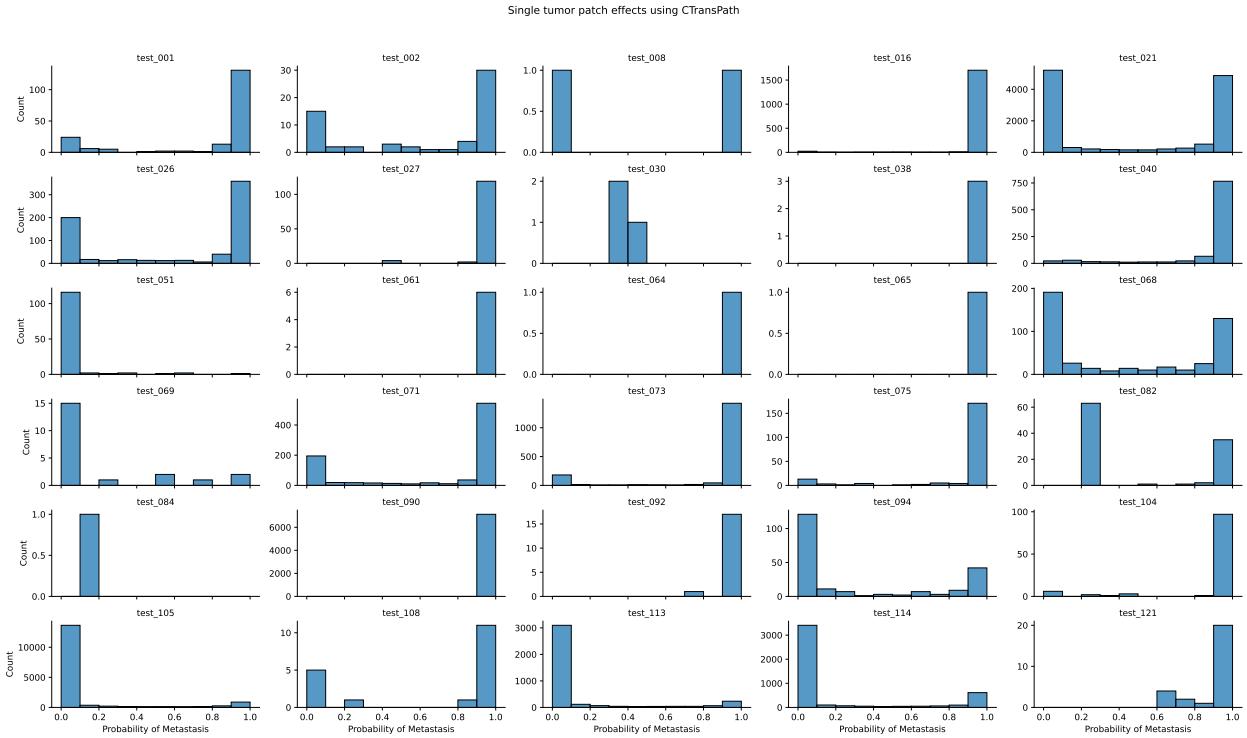
Supplementary Figure 2: Individual tumor patch effects on metastasis detection using the UNI-based model. Histograms show model probabilities of tumor, where the x-axis is model probability and y-axis is the number of examples in the bin. Each histogram represents a different positive specimen (n=49) in the CAMELYON16 test set. First, all patches intersecting the expert tumor annotations were removed. Then, patches full contained within the annotation were added back into the specimen one at a time, and model predictions were recorded.



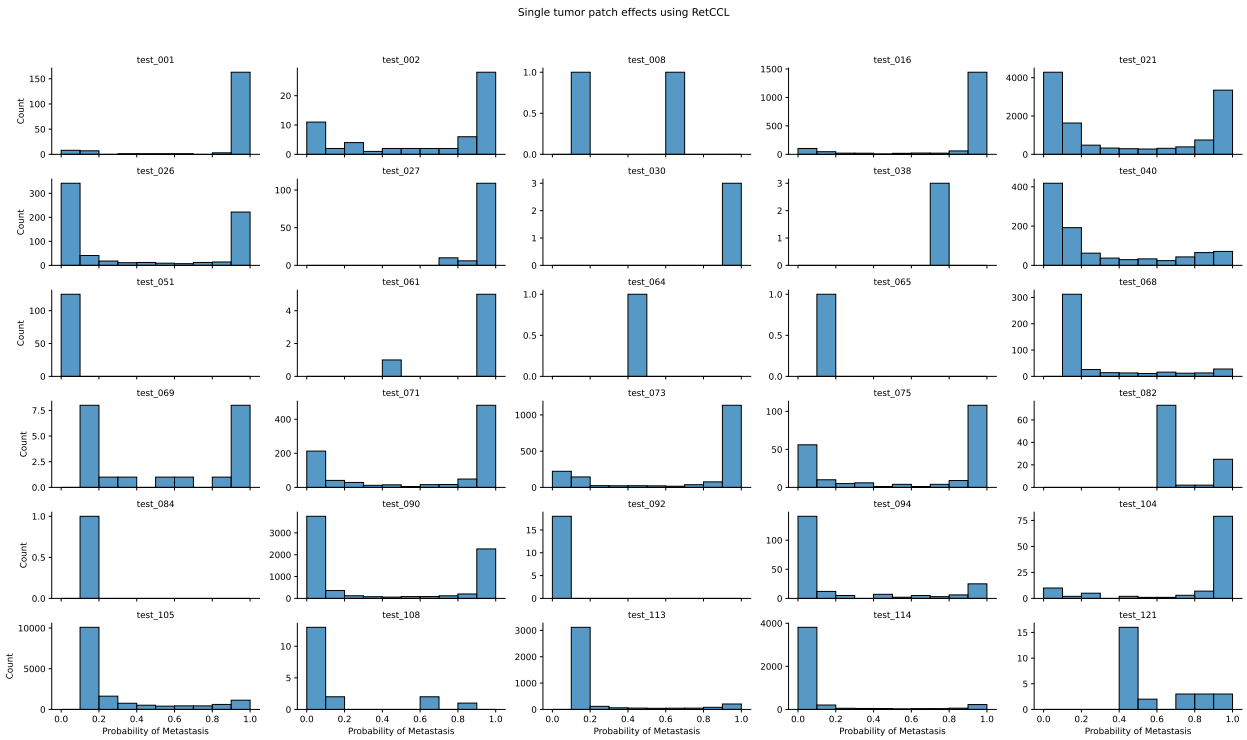
Supplementary Figure 3: Individual tumor patch effects on metastasis detection using the REMEDIS-based model. Histograms show model probabilities of tumor, where the x-axis is model probability and y-axis is the number of examples in the bin. Each histogram represents a different positive specimen (n=49) in the CAMELYON16 test set. First, all patches intersecting the expert tumor annotations were removed. Then, patches full contained within the annotation were added back into the specimen one at a time, and model predictions were recorded.



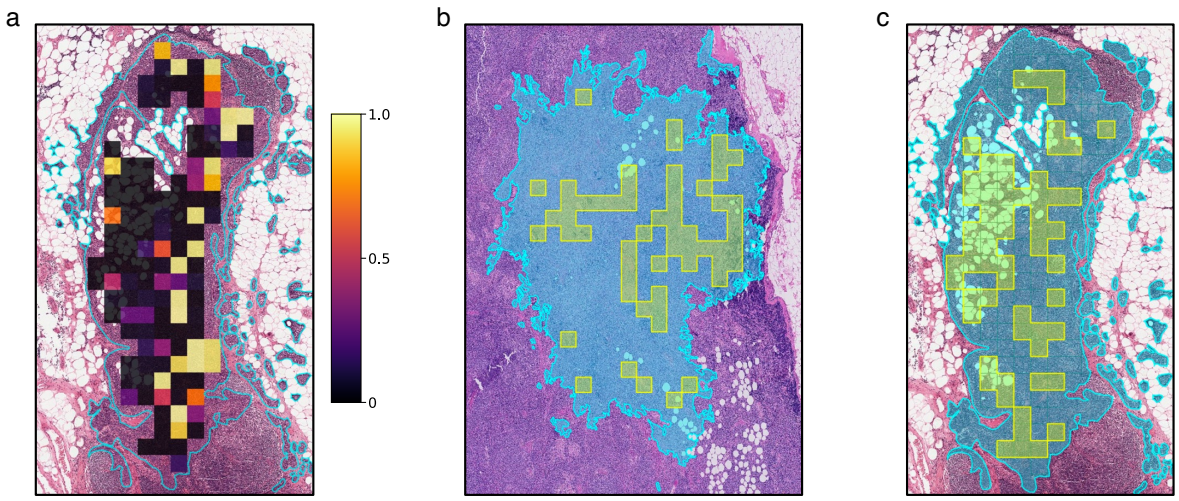
Supplementary Figure 4: Individual tumor patch effects on metastasis detection using the Phikon-based model. Histograms show model probabilities of tumor, where the x-axis is model probability and y-axis is the number of examples in the bin. Each histogram represents a different positive specimen (n=49) in the CAMELYON16 test set. First, all patches intersecting the expert tumor annotations were removed. Then, patches full contained within the annotation were added back into the specimen one at a time, and model predictions were recorded.



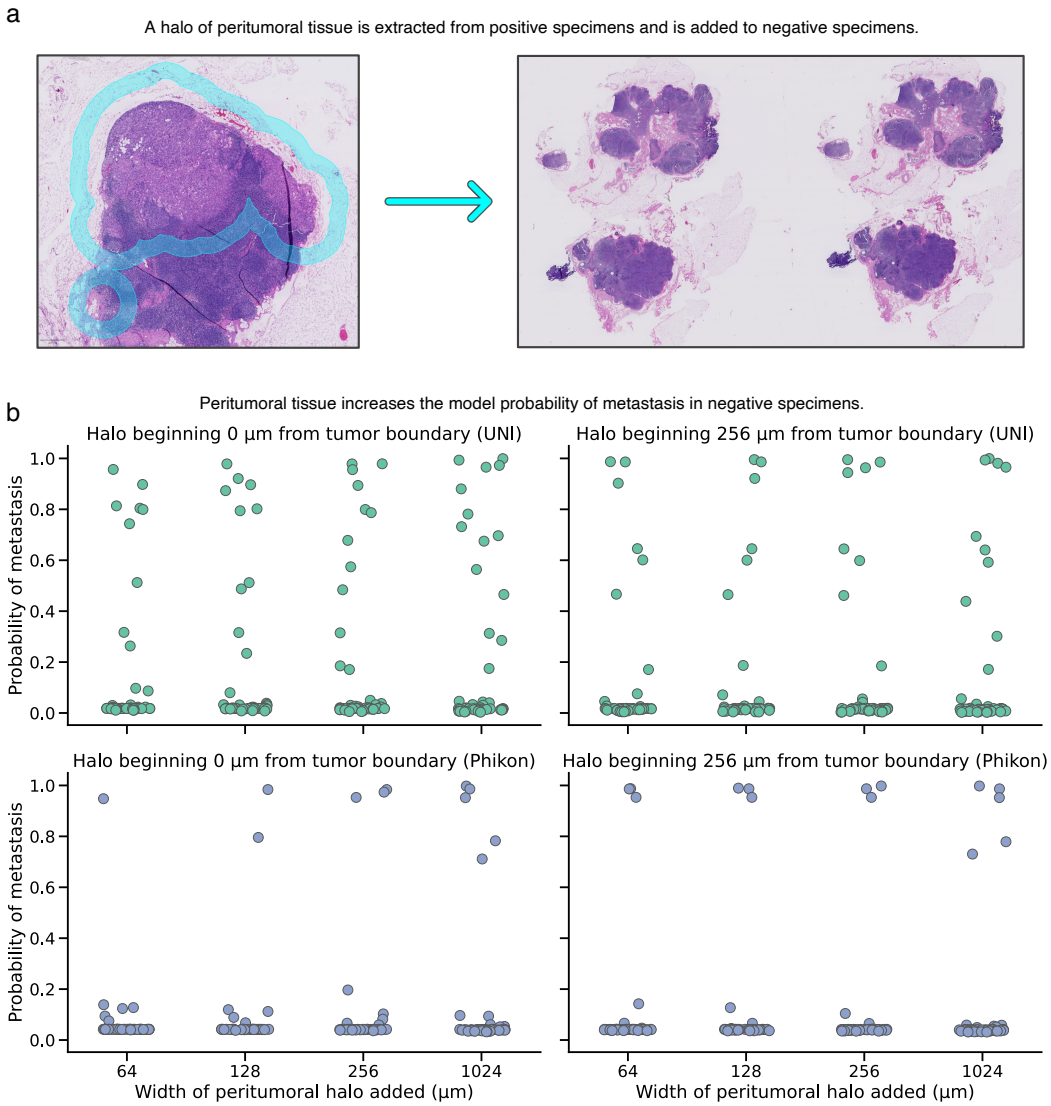
Supplementary Figure 5: Individual tumor patch effects on metastasis detection using the CTransPath-based model. Histograms show model probabilities of tumor, where the x-axis is model probability and y-axis is the number of examples in the bin. Each histogram represents a different positive specimen (n=49) in the CAMELYON16 test set. First, all patches intersecting the expert tumor annotations were removed. Then, patches full contained within the annotation were added back into the specimen one at a time, and model predictions were recorded.



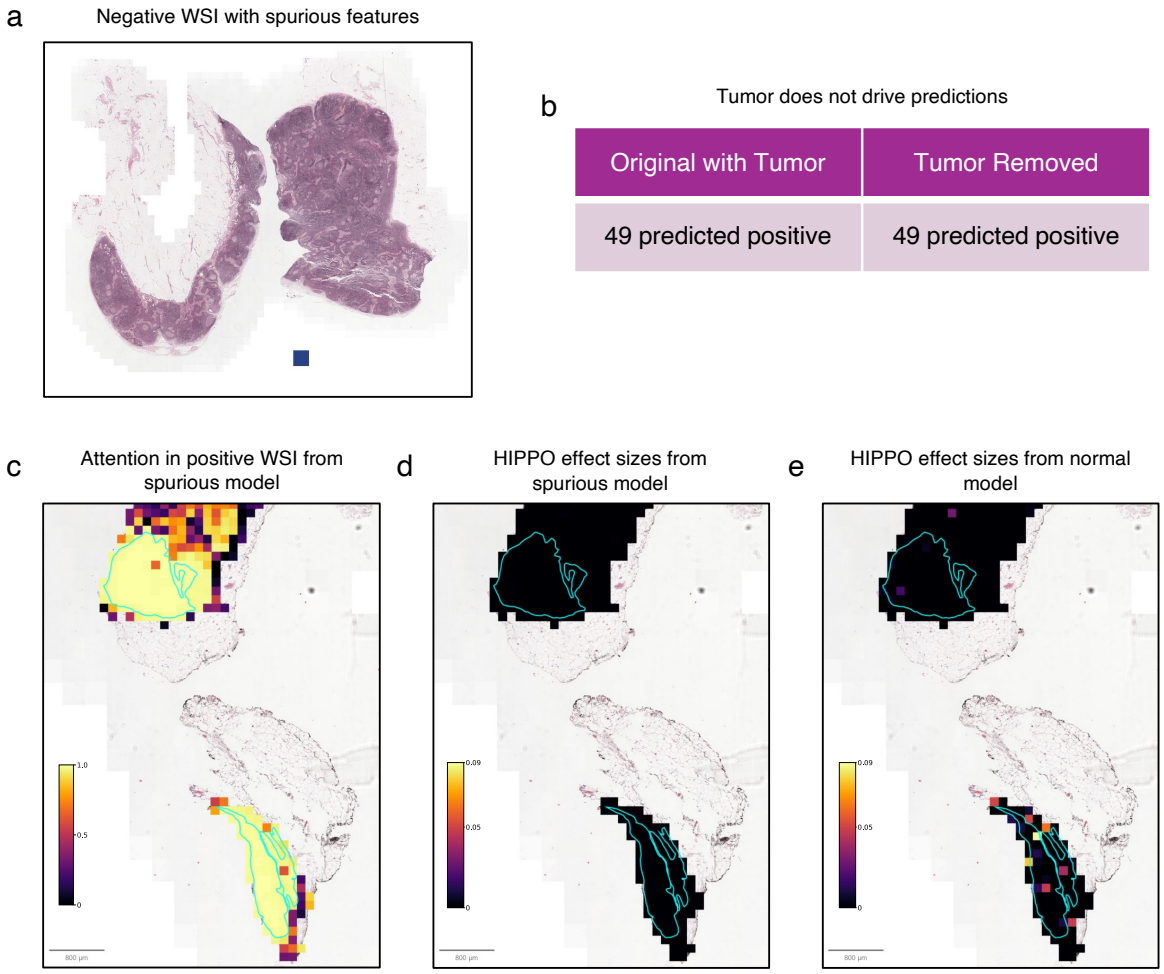
Supplementary Figure 6: Individual tumor patch effects on metastasis detection using the RetCCL-based model. Histograms show model probabilities of tumor, where the x-axis is model probability and y-axis is the number of examples in the bin. Each histogram represents a different positive specimen (n=49) in the CAMELYON16 test set. First, all patches intersecting the expert tumor annotations were removed. Then, patches full contained within the annotation were added back into the specimen one at a time, and model predictions were recorded.



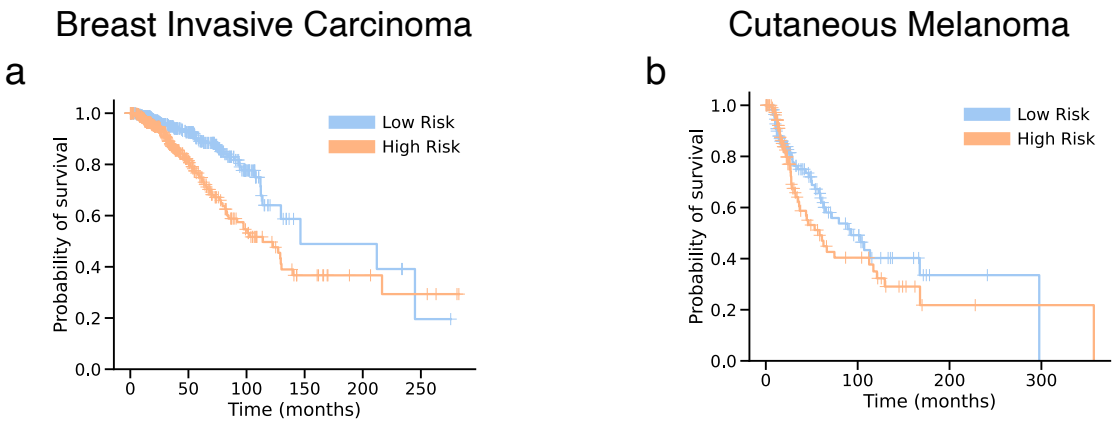
Supplementary Figure 7: Tumor patches have variable effects on metastasis detection, and some tumor regions go undetected entirely. To evaluate the effect of each individual tumor patch on metastasis detection, first all patches intersecting with expert tumor annotations was removed in the positive specimens (n=49) of the CAMELYON16 test set. Then, patches that were fully contained in the tumor annotation were introduced into the specimen one at a time, and the model probability of metastasis was recorded. (a) shows a representative example of model probabilities of metastasis for each tumor patch, using the UNI-based model in specimen "test_051". The expert tumor annotation is outlined in cyan. A subset of patches was sufficient to drive a positive tumor prediction (model probability > 0.5), but mahy tumor patches were insufficient to drive a positive prediction on their own. Some of these insufficient patches contained tumor epithelial cells along with adipose cells, but many did not. We also used a version of the search algorithm *HIPPO-search-low-effect* to identify the largest set of tumor patches that can be added to a negative counterfactual while still maintaining a negative prediction. First, all tumor patches intersecting the tumor boundary were removed. Then, we iteratively added tumor patches back into the specimen, and kept the tumor patch that drove the lowest probability of metastasis. This continued until the model probability was greater than 0.5. These "unseen" tumor regions are highlighted in yellow in (b) and (c), and the tumor region is highlighted in cyan. In (b) (specimen "test_094"), we identified a $0.95\,\mathrm{mm}^2$ area of tumor that was undetected by the UNI-based model, and in (c), we identified a $1.0\,\mathrm{mm}^2$ area of tumor that was undetected by the UNI-based model.



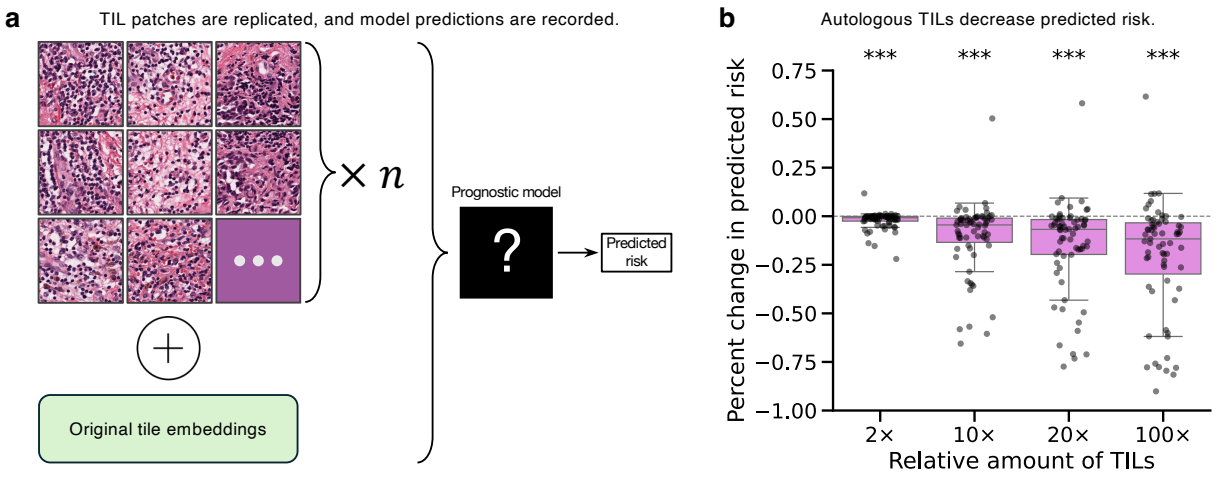
Supplementary Figure 8: Non-tumor tissue is sufficient for positive detections in specimens without tumor. (a) To evaluate the sufficiency of non-tumor tissue from positive specimens to drive positive detections in negative specimens, halos of peritumoral tissue were selected. These halos did not intersect with the expert tumor annotations, and as such were considered to be entirely non-tumor. The patches intersecting with the halo but not intersecting with tumor annotations were added to normal specimens, resulting in 3,920 counterfactual examples (80 negative $\times$ 49 positive specimens). (b) The model's probability of metastasis was averaged across each negative specimen to evaluate the global effect of the peritumoral halo on model outputs. Four widths of halos were evaluated (i.e., 64, 128, 256, and 1024 µm), beginning at either the outer edge of the expert tumor annotation (left column) or 256 µm outside of the annotation (right column). Two foundation models were evaluated: UNI (top row) and Phikon (bottom row). Multiple halos of non-tumor tissue were sufficient to drive false positive metastasis detection. In the UNI-based model, for example, a 64 µm halo beginning at the tumor annotation border from 7 positive specimens was sufficient to drive false positives. In the Phikon-based model, a 1024 µm halo was sufficient for false positive predictions from 5 positive specimens.



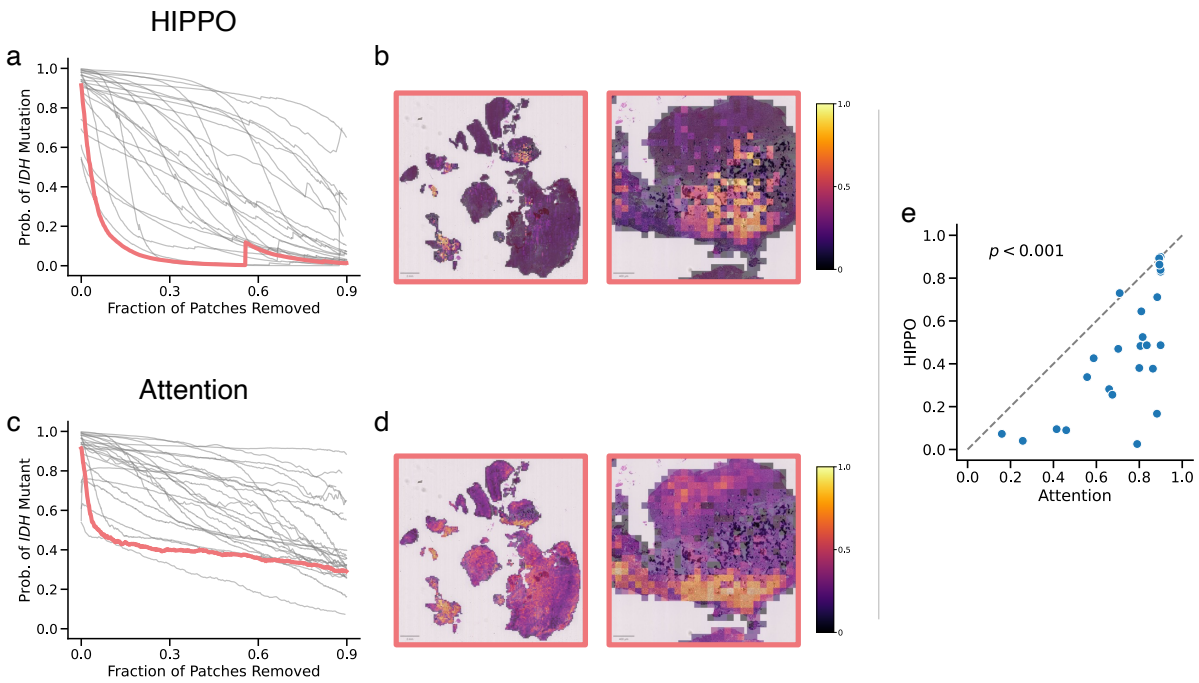
Supplementary Figure 9: **HIPPO identifies shortcut learning when attention struggles.** **a**, Thumbnail of a negative specimen (`normal_009`) with a 768×768 µm blue square added. A blue square was added to all negatives specimens (n=239) in the CAMELYON16 dataset to promote shortcut learning. The UNI foundation model was used to embed the tissue and the blue squares. Positive samples were unaltered. **b**, All positive specimens were predicted as positive, and removal of tumor regions did not change model predictions. This suggested that the ABMIL models learned that if a blue patch is absent, the specimen is positive for metastasis. **c**, Attention heatmap for specimen `test_002`, with expert tumor annotation in cyan. Despite tumor having no effect on model predictions, there was strong attention on tumor regions. **d**, Heatmap of patch effect sizes in specimen `test_002` using the ABMIL model trained on deliberate spurious specimens. Using *HIPPO-search-high-effect*, we searched for the patches with highest effect on model outputs. **e**, Heatmap of patch effect sizes in specimen `test_002` using the original ABMIL model, trained without deliberate spurious specimens.



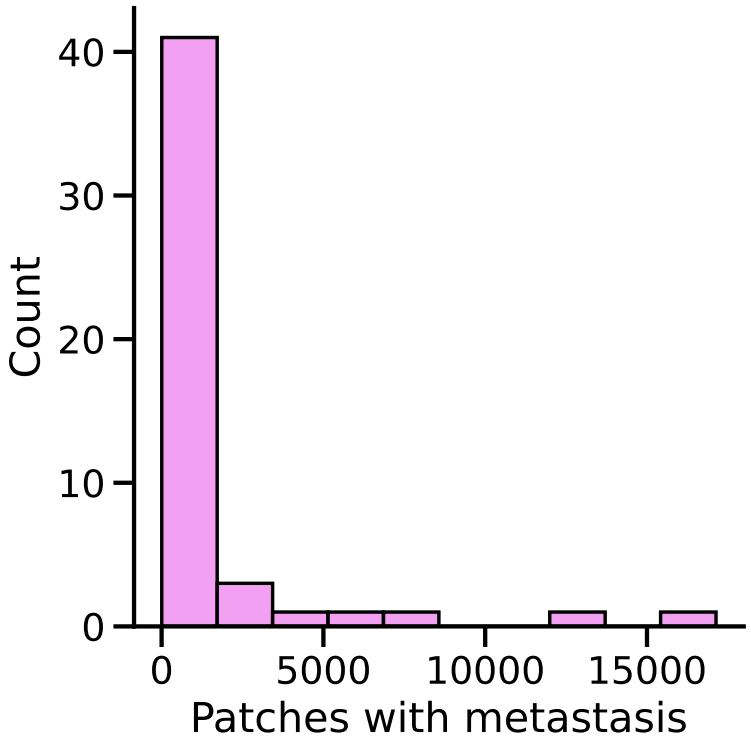
Supplementary Figure 10: **a, b**, Kaplan Meier plots for breast cancer (BRCA) (**a**) and cutaneous melanoma (SKCM) (**b**) in The Cancer Genome Atlas. Prognostic attention-based multiple instance learning models were trained to learn overall survival from whole slide images (WSIs), and risk scores were used to stratify patients. If a patient had multiple WSIs, the predicted prognoses were averaged across WSIs to arrive at a single predicted risk score per patient. Risk scores were then median split into low risk and high risk. BRCA overall survival had concordance index of 0.667 ($p < 0.005$, log-rank test), and for SCKCM, concordance index was 0.557 ($p > 0.05$, log-rank test). Please note that for experiments in the main text, low and high risk were defined as the first and fourth quartiles of risk scores, respectively.



Supplementary Figure 11: **Autologous TILs improve predicted prognosis.** In high-risk slides of cutaneous melanoma (TCGA-SKCM, n=67), TIL-positive patches were identified using a heuristic from [**?**]. High risk was defined as slides with the top 25% of predicted risk scores. **a**, The embeddings of TIL-positive regions were replicated and concatenated with the original embeddings (the ellipsis denotes that the displayed TIL patches are a representative sample of a larger set). Model predictions are then recorded for this counterfactual with additional autologous TILs. **b**, Box plot showing the difference in model predictions, relative to the original specimens. Differences are shown on the y-axis and were calculated as the predicted risks with autologous TILs minus the original predicted risk (negative values indicate that autologous TILs decreased predicted risk). The x-axis shows the amount of TILs relative to the original specimens. The sample size in each box is 67. Box plots show the first and third quartiles, the median (central line) and the range of data with outliers removed (whiskers), and significance is shown (***: $p < 0.001$).



Supplementary Figure 12: **HIPPO outperforms attention in identifying regions that drive positive predictions.** The strategy `HIPPO-search-high-effect` was used to identify the patches that were most responsible for positive predictions. Ten patches were removed at each iteration of the HIPPO search to reduce running time. For attention, ten patches were removed at a time, in order of descending attention, for comparison with HIPPO. **(a, c)**, line plots showing the probability of *IDH* mutation on the vertical axis and the ratio of patches removed on the horizontal axis, where patches are removed by **(a)** HIPPO search or **(c)** attention. **(b, d)**, heatmap of patches found by HIPPO **(b)** and heatmap of attention weights **(d)**, both normalized to range [0, 1]. **(e)** scatter plot showing the ratio of patches removed to decrease the predicted *IDH* mutation to 0.4. HIPPO more effectively identified the patches driving positive predictions, requiring fewer patch removals to reduce the probability to 0.4 compared to attention ($p < 0.001$, independent t-test).



Supplementary Figure 13: Distribution of the number of metastasis-containing patches in the CAMELYON16 test set. The test consists of 49 specimens with metastasis (plotted here) and 80 specimens negative for tumor. The positive specimens contained an average of 1320 patches that intersected the tumor annotation, where each patch was 128×128 µm. This distribution was heavily right-skewed. The median number of tumor patches was 102, and the 25th and 75th percentiles were 18 and 616, respectively.