Explainable AI for computational pathology identifies model limitations and tissue biomarkers

Jakub R. Kaczmarzyk
Department of Biomedical Informatics
Stony Brook University
Stony Brook, NY 11794
jakub.kaczmarzyk@stonybrookmedicine.edu

Joel H. Saltz
Department of Biomedical Informatics
Stony Brook University
Stony Brook, NY 11794
joel.saltz@stonybrookmedicine.edu

Peter K. Koo
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724
koo@cshl.edu

Abstract

Deep learning models have shown promise in histopathology image analysis, but their opaque decision-making process poses challenges in high-risk medical scenarios. Here we introduce HIPPO, an explainable AI method that interrogates attention-based multiple instance learning (ABMIL) models in computational pathology by generating counterfactual examples through tissue patch modifications in whole slide images. Applying HIPPO to ABMIL models trained to detect breast cancer metastasis reveals that they may overlook small tumors and can be misled by non-tumor tissue, while attention maps—widely used for interpretation—often highlight regions that do not directly influence predictions. We also used HIPPO with prognostic models to identify prognostic tissue regions and to experiment with interventions that affect prognosis. These findings demonstrate HIPPO's capacity for comprehensive model evaluation, bias detection, and quantitative hypothesis testing. HIPPO greatly expands the capabilities of explainable AI tools to assess the trustworthy and reliable development, deployment, and regulation of weakly-supervised models in computational pathology.

1 Introduction

Digital pathology has emerged as a transformative force in medicine, ushering in an era where computational methods can augment and enhance the diagnostic and prognostic capabilities of pathologists. By digitizing whole slide images (WSIs) of tissue specimens, this field has opened up new avenues for applying advanced machine learning techniques to analyze complex histological patterns and features. The potential impact of computational pathology is far-reaching, promising to improve diagnostic accuracy, standardize interpretation, and uncover novel biomarkers that may inform personalized treatment strategies [1–13].

Recently, attention-based multiple instance learning (ABMIL) [14] has emerged as a powerful approach to analyze WSIs for various pathological tasks, demonstrating performance that often rivals or surpasses that of expert pathologists [15]. ABMIL models treat each WSI as a collection of smaller image patches (instances) and use attention mechanisms to identify and focus on the most relevant regions for the task at hand. Importantly, multiple instance learning allows ABMIL models to learn from specimen-level labels, not requiring exhaustive pixel-level annotations, which are

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

time-consuming and costly to obtain[15]. This feature makes ABMIL models particularly well-suited for tasks such as cancer detection [16, 17], diagnosis [18–21], identification of primary cancer origin [22], grading [17, 23, 24], genomic aberration detection [25–30], molecular phenotyping [31–33], treatment response prediction [34–36], and prognostication [35, 37–39].

However, the widespread adoption of ABMIL models in clinical settings is hindered by challenges in model interpretability and trustworthiness [9, 10, 40, 41]. A key limitation lies in the heavy reliance of interpretations based on ABMIL's attention, which is often used as a proxy for understanding model behavior. While attention highlights regions of interest within a WSI, they do not necessarily reflect the direct influence of these regions on model predictions [42, 43]. This disconnect between attention and model output can lead to misinterpretations of model behavior, potentially eroding trust in the model's decisions and limiting its clinical utility [44–47]. In addition, post hoc model explanations via attribution methods, such as LIME [48] and SHAP [49], make restrictive additive or linear assumptions of individual pixels, which have been argued to not reflect a model's decision making process [50].

To address these challenges, we introduce HIPPO (Histopathology Interventions of Patches for Predictive Outcomes), an explainable AI method designed to enhance trust in ABMIL models and provide deeper insights into their decision-making processes. HIPPO goes beyond traditional attention-based interpretations by quantitatively assessing the impact of specific tissue regions on model predictions. By simulating targeted interventions through the occlusion or inclusion of individual or groups of patches, HIPPO enables a more nuanced understanding of how different histological features influence ABMIL model outputs.

We demonstrate the utility of HIPPO by applying it to two clinically important tasks in computation pathology: metastasis detection and prognostication. For metastasis detection, we evaluated five foundation models in pathology using the CAMELYON16 dataset [16]. Our analysis uncovers model-specific limitations and biases that would have remained hidden using attention mechanism alone. We reveal that some models rely heavily on extratumoral tissue for metastasis detection, while others are surprisingly insensitive to small tumor regions. With prognostic models, we used HIPPO to identify the regions that drive prognostic predictions, and we perform experiments to measure the effect of tumor-infiltrating lymphocytes (TILs) on predicted prognosis (Appendix). These findings highlight the importance of rigorous model evaluation beyond standard performance metrics and underscore the potential of HIPPO in identifying when and why models might fail.

As computational pathology continues to advance, the need for robust, interpretable, and trustworthy AI models becomes increasingly critical. HIPPO represents a significant step forward in this direction, offering a powerful tool for uncovering the strengths, limitations, and potential biases of ABMIL models in pathology. By providing a more comprehensive understanding of model behavior, HIPPO not only enhances the interpretability of existing models but also paves the way for developing more reliable and clinically relevant AI tools in pathology. As we demonstrate in metastasis detection, HIPPO has the potential to accelerate the translation of computational pathology into clinical practice, ultimately improving patient care and outcomes.

2 Results

2.1 HIPPO: Histopathology Interventions of Patches for Predictive Outcomes

HIPPO is a specimen-level perturbation toolkit that explains weakly-supervised models in computational pathology (Fig. 1a). The fundamental goal of HIPPO is to explore counterfactual (i.e., "what if") scenarios that are infeasible to realize in actual tissue samples. For instance, it would be impractical to directly manipulate the tumor microenvironment of a tissue specimen to understand its effect on a prognostic model. Instead, we can digitally modify a WSI that simulates this intervention. HIPPO enables virtual interventions through the occlusion or inclusion of single or multiple patches, utilizing the resulting ABMIL model predictions as counterfactual outcomes. HIPPO provides quantitative insights into how specific tissue alterations impact pathological assessments through the lens of the AI model. These assessments can include but are not limited to, patient prognosis, treatment response prediction, metastasis detection, inference of spatial transcriptomics, gene mutation detection, and microsatellite instability identification. Applying HIPPO to ABMIL models enables researchers, regulators, and clinicians to elucidate model behavior and assess the reliability of model outputs in high-risk clinical contexts.

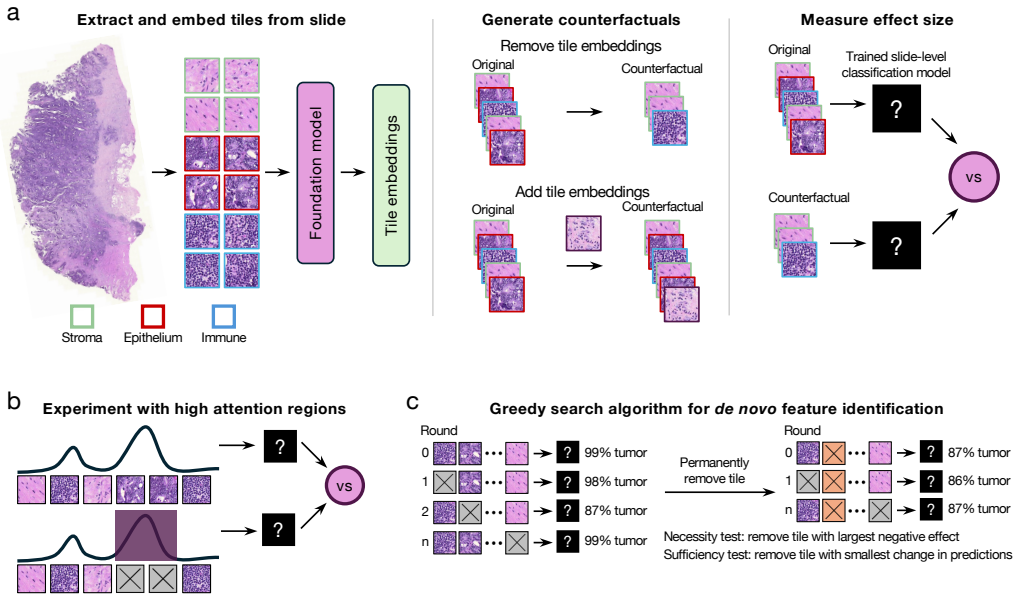

Figure 1: **HIPPO explainable AI toolkit.** HIPPO enables quantitative assessment of how specific tissue regions impact model predictions, enhancing interpretation and validation of AI models. **a**, Schematic of attention-based multiple instance learning. Whole slide images are divided into patches and embedded using a pretrained foundation model. ABMIL learns specimen-level labels from these bags of patches, assigning attention weights to each patch. Leveraging ABMIL's invariance to patch order and count, we can create counterfactual specimens by adding or removing tissue regions within patches. Model outputs are then compared between original and counterfactual specimens to measure effects. **b**, HIPPO quantifies the effect of high-attention regions by removing them and measuring the resulting change in model outputs. **c**, HIPPO implements greedy search algorithms to identify necessary or sufficient tissue regions *de novo*.

Traditional approaches to digital interventions in medical imaging often require precise segmentation of objects for occlusion or inclusion [51, 52], as well as sophisticated inpainting techniques to maintain image integrity [53–56]. Alternatively, generative AI can generate counterfactual images [57, 58], but the quality of the generated images has not been thoroughly evaluated for histopathology. These manual or AI-assisted methods can introduce covariate shifts when imperfectly executed [59], potentially leading to unreliable model predictions. The key insight for HIPPO is based on how data flows through ABMIL models. A WSI is treated as a bag of permutation-invariant patches, where the number and order of patches are allowed to vary [14]. Thus, an intervention can be achieved through two primary perturbation mechanisms: (1) removing specific patches, effectively excising tissue from the input specimen, or (2) including specific patches, simulating the addition of new tissue into the specimen. HIPPO leverages unique properties of multiple instance learning models to facilitate the generation of counterfactual images bypassing the complexities of direct image manipulation by creating hypothetical scenarios such as the introduction or removal of tumor patches or regions of tumor-infiltrating lymphocytes (TILs) from a patient's specimen. Understanding when ABMIL models alter their predictions due to interventions provides quantitative insights into their decision making process, revealing important features and potential biases learned.

There are several ways to choose the regions to occlude or include, and the choice of region depends on whether spatial annotations are available. If annotations are available (e.g., pixel-level tissue type), then patches may be selected based on their annotation. In the present report, we used the CAMELYON16 dataset, which includes expert annotations. We also used TCGA data and chose patches based on HoVer-Net nucleus detections [60]. We name the process of selecting patches based on annotations *HIPPO-knowledge*, as this is an intervention based on prior knowledge (Fig. 1a). However, we acknowledge the difficulty in acquiring fine annotations. Given this, we developed search algorithms to identify patches of interest in a data-driven fashion. The two search algorithms are *HIPPO-search-high-effect* and *HIPPO-search-low-effect* (Fig. 1c). These algorithms identify the

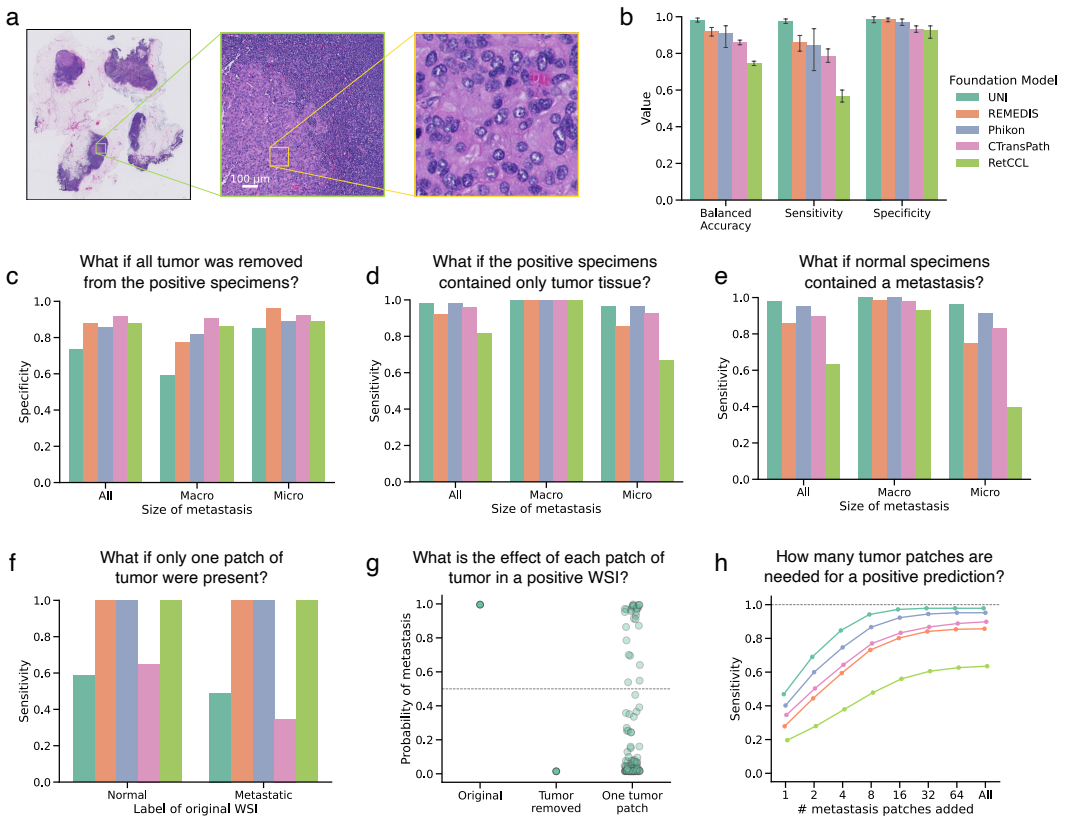

Figure 2: **Understanding the role of tumor in detecting metastases. a**, Example WSI from the CAMELYON16 dataset containing a macrometastasis (specimen `test_001`), with a 128 × 128 μm patch highlighted. **b**, Bar plot of balanced accuracy, sensitivity, and specificity on the CAMELYON16 test set (n=129, 80 negative, 49 positive) across five random initializations and five encoders, with mean values and 95% confidence intervals. The best-performing model for each encoder was used in subsequent experiments. **c-d**, Bar plots showing specificity when tumor-containing patches are removed (c) and sensitivity when only tumor tissue remains (d) in positive specimens (n=49, 22 macrometastases, 29 micrometastases), quantifying necessity and sufficiency of tumor regions for metastasis detection. **e**, Bar plot of sensitivity after adding metastases to negative specimens (3920 counterfactuals: 80 negative × 49 positive), further quantifying tumor sufficiency. **f**, Bar plot showing sensitivity of counterfactuals with a single 128 × 128 μm tumor patch in normal (n=80) and metastatic (n=49) specimens. **g**, Strip plot of model probabilities for tumor patches in specimen `test_051` using the UNI-based ABMIL model, comparing original, tumor-removed, and single-tumor-patch (n=125) conditions. **h**, Line plot relating tumor size to model sensitivity, with each point representing 3920 counterfactuals (80 negative × 49 positive) as tumor patches are added to negative specimens.

patches that, when removed, have a strong negative effect on model predictions (i.e., are necessary) or have little effect on predictions (i.e., are not necessary), respectively. These search strategies complement attention heatmaps, in that they identify regions considered important by the model, but they do so through measuring model effects. Another way to choose regions for occlusion or inclusion is with attention maps. To quantify the effect of high attention regions, one may occlude high attention regions and measure the change in model outputs. One may also include these high attention regions into other tissue specimens and measure the effects on model outputs. We name this patch selection strategy *HIPPO-attention* (Fig. 1b).

2.2 Do MIL models think tumor is necessary for breast cancer metastasis detection?

Metastasis detection is a well-studied task, with well-defined features (i.e., tumor cells) that drive the label of whether or not a specimen contains metastasis. In a clinical setting, it is critical that

metastases are identified; a false negative is unacceptable. Recent studies have shown that ABMIL models have strong performance in metastasis detection[61]. However, previous studies have also found that computer vision models can make the correct predictions for the wrong reasons, such as short-cut features or spurious correlations [62, 63]. Thus, the degree to which AI models rely on the tumor regions remains to be seen, even for a relatively straightforward task like tumor detection. Understanding this is critical to elucidate the strengths and limitations of ABMIL models for metastasis detection, including potential biases.

To evaluate this, we trained several ABMIL models for breast metastasis detection using the CAMELYON16 dataset [16] (Fig. 2a). Several pathology foundation models have recently emerged, demonstrating near-human levels in metastasis detection. Here we consider five pathology foundation models (UNI [61], REMEDIS [64], Phikon [65], CTransPath [66], and RetCCL [67]). We trained five ABMIL models for each foundation model to distinguish whether or not a specimen contained metastasis. Similar to previously reported results [61], UNI achieved a mean balanced accuracy of 0.982, REMEDIS 0.922, Phikon 0.907, CTransPath 0.858, and RetCCL 0.745. (Fig. 2b). For HIPPO explainability experiments, we used the best-performing model (out of 5 random initializations) on the test set for each foundation model. The best UNI model achieved balanced accuracy of 1.00, REMEDIS 0.949, Phikon 0.955, CTransPath 0.885, and RetCCL 0.769.

In this dataset, expert pathologists finely annotated metastatic regions. This allows us to use HIPPO-knowledge to determine whether metastatic regions are necessary for detecting breast cancer metastasis. Specifically, for patients who were positive for metastasis, we removed the patches that intersected with the tumor annotations, effectively creating a version of the specimen that does not contain metastasis. We compared model predictions before and after the intervention. Specificity was calculated as the ratio of true negatives to all negative samples. In this set of counterfactuals, all specimens were negative, so the specificity represented the proportion of correct negative predictions by the models. Notably, the UNI-based model exhibited the lowest specificity (0.73) in these counterfactual examples despite achieving the highest balanced accuracy on the original test set (1.00). This discrepancy was particularly pronounced in counterfactual specimens that originally contained macrometastases (specificity 0.59), suggesting that the UNI-based ABMIL model uses tissue outside of the tumor region to drive positive metastasis predictions. The REMEDIS-based model exhibited a similar trend, with a specificity of 0.77 in counterfactuals derived from macrometastases. In contrast, the other models showed less dependence on extratumoral tissue (sensitivity of Phikon-based, 0.86; CTransPath-based, 0.92; RetCCL-based, 0.88), indicating that their predictions are primarily driven by tumor epithelial cells rather than other tissue components (Fig. 2c). In summary, HIPPO enabled the quantitative exploration of peritumoral tissue on metastasis detection.

2.3 Is tumor sufficient for breast cancer metastasis detection?

While necessity assesses the importance of a feature or feature set, it does not inform whether the feature set is sufficient for model predictions. Metastasis detection models must be able to detect tumor regions no matter how small. Using HIPPO-knowledge, we tested the sufficiency of metastatic regions using two methods: removing all non-tumor patches and measuring model outputs and adding tumor regions to normal specimens and measuring model outputs.

First, we constructed counterfactual specimens (n=49) by removing all non-tumor tissue (i.e., removing patches that did not intersect with expert tumor annotations) and measuring model outputs. With only the tumor present, the true label for these images was "positive", and the foundation models had the following sensitivity (true positive rate): UNI-based 0.98, REMEDIS-based 0.92, Phikon-based 0.98, CTransPath-based 0.96, RetCCL-based 0.82 (Fig. 2d). There is evidence to suggest that extratumoral tissue caused false negative predictions. Four of the five foundation models improved sensitivity when using only tumor tissue in micrometastases compared to the original positive samples, suggesting that extratumoral tissue drove false negative predictions. The sensitivity of CTransPath increased by 25%, Phikon by 4%, REMEDIS by 5%, and RetCCL by 100%. For UNI, however, using original WSIs resulted in a sensitivity of 1.0 on micrometastasis. However, when using only the tumor tissue, one false negative prediction suggested that the UNI-based model may use tissue outside of the metastatic region in its predictions. Critically, this demonstrated that the tumor was insufficient for a positive prediction in this specimen with the UNI-based model and that extratumoral tissue was solely driving the positive prediction. RetCCL had a true positive rate in macrometastases of 0.95 (21 predicted positive of 22 positive specimens). When using only

tumor tissue, all macrometastases were detected successfully, demonstrating that tissue outside the metastatic region caused a false negative prediction.

We also evaluated whether tumor was sufficient for metastasis detection by embedding tumor regions in normal specimens. We embedded all patches intersecting with tumor annotations into normal specimens, resulting in 3,920 positive counterfactual examples (80 normal slides $\times$ 49 positive slides). Model outputs for these examples were recorded. The UNI-based model had a sensitivity of 0.98, REMEDIS-based 0.86, Phikon-based 0.95, CTransPath-based 0.90, and RetCCL-based 0.63. Positive counterfactuals made with micrometastases were less likely to be detected by most models (UNI-based achieved sensitivity of 0.96, REMEDIS-based 0.75, Phikon-based 0.91, CTransPath-based 0.93, and RetCCL-based 0.40), suggesting that smaller tumors in the context of normal tissue are insufficient for positive metastasis detection (Fig. 2e).

The average treatment effect for each metastatic slide was calculated by averaging the model's probability of metastasis across all negative samples. This informs which positive slides can drive positive predictions across individuals. 100% of macrometastases (n=22) led to true positives in UNI-based, REMEDIS-based, Phikon-based, and CTransPath-based models. In the RetCCL-based model, 90% (n=20) of macrometastases had an average true positive effect. Micrometastases (n=27) were less likely to induce positive predictions on average, with 96% (n=26) positive in UNI, 93% (n=25) in Phikon, 81% (n=22) in CTransPath, 74% (n=20) in REMEDIS, and 37% (n=10) in RetCCL.

2.4 Foundation models may miss small breast cancer metastases

To evaluate the sensitivity of ABMIL models to detect metastasis based on the size of the metastasis in a specimen, we analyzed the metastasis-positive specimens from the CAMELYON16 test set. Our methodology involved initially removing all tile embeddings that intersected with expert tumor annotations, effectively rendering the slide negative for metastases. A 128×128 μm region of tumor (shown in the right-hand side of Fig. 2a) was added to 80 normal specimens and 49 metastasis-removed positive specimens. When the single-patch tumor region was embedded in normal specimens, the REMEDIS-, Phikon-, and RetCCL-based ABMIL models detected 100% of counterfactuals as positive, highlighting their robustness to this small region of tumor. The UNI-based model, on the other hand, failed to detect 41% (n=33) of positive counterfactuals (n=80), and the CTransPath-based models failed to detect 35% (n=28) of positive counterfactuals. A similar trend was observed when the tumor region was embedded into the context of metastatic specimens (i.e., the positive specimen with metastasis removed). The REMEDIS-, Phikon-, and RetCCL-based models detected 100% of positive counterfactuals (n=49), whereas the UNI-based model missed 51% (n=25) and CTransPath-based missed 65% (n=32) of positive counterfactuals specimens (Fig. 2f). This result is surprising because the UNI-based model had perfect sensitivity in the original test set (Fig. 2b) as well as the highest sensitivity when larger tumors were embedded into normal tissue (Fig. 2e). This highlights that high classification performance on the held-out test set is insufficient to assess generalization to more nuanced downstream applications.

We also sought to quantify the sensitivity of models to each tumor patch in positive specimens, which can shed light on whether tumor patches carry different levels of informativeness for machine learning classifiers. To accomplish this, all tumor patches intersecting with expert tumor annotations were removed. Then, we reintroduced tiles fully within the expert tumor annotation, one at a time, to the tumor-removed specimen and evaluated the model outputs. These model outputs were compared to those when all tumor was removed. While some tumor patches could drive a positive prediction on their own, many could not (representative example shown in Fig. 2g).

To further quantify the effect of tumor size in metastasis detection, we added tumor patches into normal slides in a graded fashion and measured the sensitivity. All models exhibited a graded effect of tumor size, and UNI exhibited the highest sensitivity (Fig. 2h). Models tended to plateau in sensitivity at $0.262\,144\,\mathrm{mm}^2$ of tumor (16 patches) added. The RetCCL-based model showed the lowest sensitivity and the least sensitivity to smaller tumors.

2.5 HIPPO identifies shortcut learning when attention struggles

Identifying spurious correlations in deep learning models for medical imaging is crucial to ensure reliable and clinically relevant results. To test HIPPO's ability to identify spurious correlations, we conducted an experiment where we deliberately introduced an artificial bias into the CAMELYON16

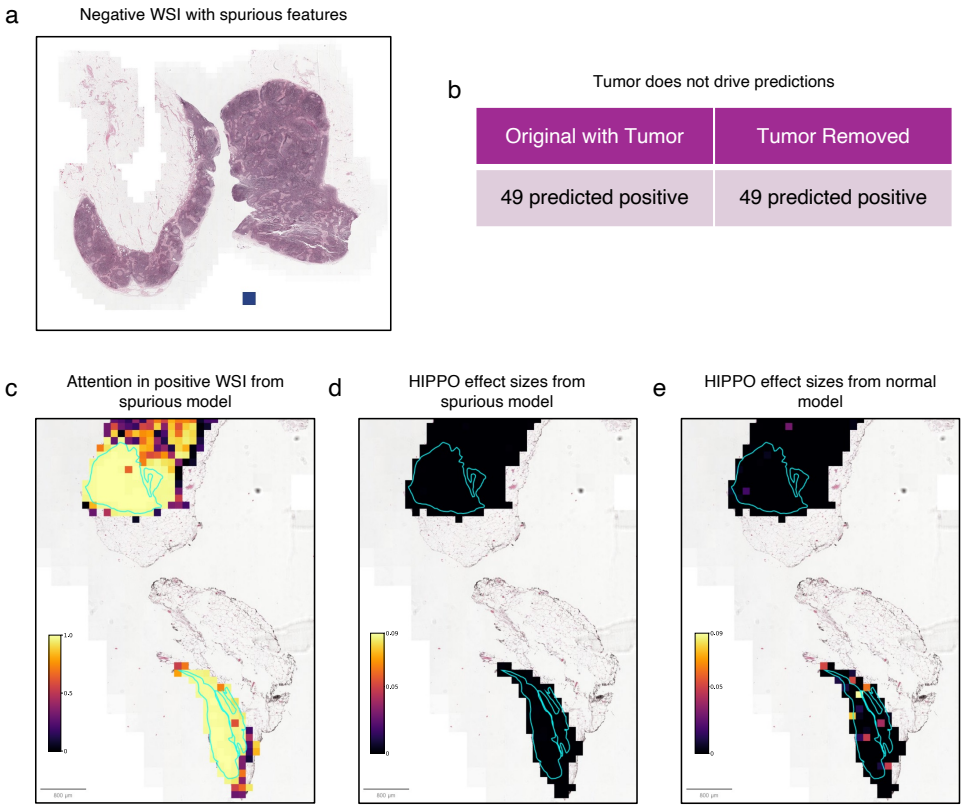

Figure 3: **HIPPO identifies shortcut learning when attention struggles. a**, Thumbnail of a negative specimen (`normal_009`) with a 768×768 µm blue square added. A blue square was added to all negatives specimens (n=239) in the CAMELYON16 dataset to promote shortcut learning. The UNI foundation model 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.

dataset (Figs. 3a and 3b). Specifically, 768×768 µm blue squares were added to all negative images. This mimics the plausible scenario in which a pathologist marks certain slides with a blue marker. However, in doing so, it introduces a strong spurious correlation with labels. We hypothesized that the models would learn that slides were negative if a blue region was present and that slides lacking this blue region are positive (as blue regions are easier to identify compared to more variable tumor regions).

An ABMIL model was trained on the modified training data using UNI embeddings. The model achieved a balanced accuracy of 1.0 on the test set, suggesting the spurious correlations created a trivial prediction task. By performing standard model interpretation using attention, we found that metastatic regions were considered highly important (Fig. 3c). However, removing these regions using HIPPO did not alter the model predictions, demonstrating that tumor regions were not important for model predictions despite a strong attention assignment. This highlights an important weakness of attention: the disconnect between attended regions and model predictions.

Knowing that the metastatic regions did not affect model outputs, we used the search algorithm *HIPPO-search-high-effect* to identify the regions that maximally drove positive tumor predictions in

both models using one positive specimen, `test_002`. Given that the model trained with spurious correlations uses the lack of a blue square as a cue for positive specimens, we expected that no individual patches would drive the positive metastasis output and that tumor regions would not have a high effect on the prediction. Indeed, effect sizes were small and evenly distributed across the WSI (minimum 2.1×10^{-5}, maximum 0.02, mean 9.4×10^{-5}, and median 5.5×10^{-5}), indicating that no single region contributed strongly to the model prediction (Fig. 3d). By contrast, applying this search algorithm to the model trained on the original CAMELYON16 dataset, we found that patch effect sizes were higher (minimum 3.7×10^{-8}, maximum 0.09, mean 1.3×10^{-4}, and median 4.9×10^{-8}), and high effect patches were within expert tumor annotations (Fig. 3e). By tying interpretation analysis directly to predictions, HIPPO-based interpretations may provide more reliable explanations of model predictions.

Shortcut learning is an important bias that must be identified and addressed in deep learning on medical images. In this case, model performance and attention were insufficient to diagnose the shortcut learning. Observational analysis based on attention maps could easily mislead an observer to believe that tumor regions drive model predictions. Quantifying effect sizes of tumor regions using HIPPO addressed these limitations and diagnosed the shortcut learning.

2.6 HIPPO quantifies the effect of prognostic biomarkers in breast and skin cancer

Moving beyond metastasis detection, we used HIPPO to evaluate prognostic models in breast cancer and cutaneous melanoma. The methods and results for these experiment are described in the Appendix. In brief, we used the data-driven search strategy *HIPPO-search-high-effect* to identify the patches that most strongly drove prognosis predictions. We found that tumor-infiltrating lymphocytes (TILs), a known prognostic biomarker, were present in many of the patches identified by HIPPO. In addition, these patches identified by HIPPO contained a greater proportion of TILs than the patches identified by attention (Appendix Fig. 4a,b). We also evaluated the effect of TILs on high-risk patients by extracting TIL-positive patches from low-risk specimens and embedded them into high-risk specimens. We measured a significant decrease in predicted risk upon the addition of TIL-positive patches (Appendix Fig. 4c). Removing TILs from low-risk specimens also increased predicted risk (Appendix Fig. 4d). Last, we used HIPPO as a framework for virtual experiments and evaluated the dosage effect of TILs in high-risk patients (Appendix Fig. 5a). In general, predicted risk of death decreased as more TILs were added (Appendix Fig. 5b). HIPPO enabled us to create counterfactual examples to study the effect of TILs on predicted survival.

3 Methods

Deep neural network development. We employed ABMIL to learn specimen-level labels from whole slide images for metastasis detection. Five patch encoders were evaluated: UNI [61], REMEDIS [64], CTransPath [66], Phikon [65], and RetCCL [67]. These embedded non-overlapping 128×128 μm patches. ABMIL model hyperparameters, adapted from Chen et al. [61], included two hidden layers (512 and 384 units) with gated attention; dropout rate of 0.25; binary classification output layer; cross-entropy loss and Adam optimizer (learning rate: 1×10^{-4}); cosine learning rate scheduler; batch size of 1 without gradient accumulation; 20 epochs maximum, with best model selected by highest validation ROC AUC. We trained five models with different random seeds for each encoder, selecting the initialization with the highest balanced accuracy on the CAMELYON16 test set for further experiments. Attention heatmaps were visualized using QuPath [68]. Models were implemented in PyTorch and trained on NVIDIA RTX 2080 Ti GPUs.

Breast cancer metastasis dataset. We used the CAMELYON16 dataset [16] to study breast cancer metastasis. This dataset consists of 399 images and has fine-grained tumor annotations made by expert pathologists. The training set was split into 90% training and 10% validation, stratified by the label of the specimen (i.e., normal or tumor). Training set consisted of 143 negative and 100 positive WSIs (52 macrometastases and 48 micrometastases). The validation set consisted of 16 negative and 11 positive WSIS (6 macrometastases and 5 micrometastases). We used the pre-defined test set, which consisted of 80 negative and 49 positive WSIs (22 macrometastases and 27 micrometastases). In the entire dataset, there were 160 metastasis-positive specimens. There was an average tumor area of $12.26\,\mathrm{mm}^2$ (std. dev. $34.04\,\mathrm{mm}^2$; minimum $0.008\,\mathrm{mm}^2$; and maximum $276.09\,\mathrm{mm}^2$). All

399 slides had pixel spacings between 0.226 and 0.243 $\frac{\mu m}{px}$ (MPP). The WSIs had $10,250 \pm 6,672$ patches (mean $\pm$ standard deviation), where each patch was 128×128 μm.

Whole slide image processing. We used the CLAM toolkit to extract 128×128 μm patches from whole slide images. Patches were encoded using five foundation models: UNI [61], REMEDIS [64], Phikon [65], CTransPath [66], and RetCCL [67]. This was performed using NVIDIA RTX 2080 Ti GPUs and took several days to complete.

Diagnosing shortcut learning. We evaluated HIPPO's ability to uncover shortcut learning and compared it to attention analysis. We modified the CAMELYON16 dataset by adding a blue square (768×768 μm, color code `#284283`) to normal specimens, simulating pathologist markings. This was done by replicating a UNI model [61] embedding of a 128×128 μm blue square 36 times. We hypothesized that the ABMIL model would learn to distinguish normal from metastatic specimens based on the blue region's presence. To assess tumor regions' impact on positive specimens, we removed patches intersecting tumor annotations and recorded model outputs. Attention maps were visualized. We used the *HIPPO-search-high-effect* strategy to identify regions with the highest effect sizes. This process was repeated using a UNI-based ABMIL model trained on the original, unaltered CAMELYON16 dataset with identical hyperparameters and random seed.

4 Discussion

In this study, we introduce HIPPO, an explainable AI method designed to enhance the interpretability and trustworthiness of ABMIL models in computational pathology. Our results demonstrate HIPPO's ability to uncover hidden biases, quantify the impact of specific tissue regions on model predictions, and bridge the gap between computational outputs and clinically relevant insights. These findings may have significant implications for the development, regulation, and clinical application of AI in pathology.

One of the key strengths of HIPPO lies in its capacity to reveal model-specific limitations that are not apparent from performance metrics or attention mechanisms alone. In our evaluation of metastasis detection models, we uncovered surprising variations in how different foundation models process histological information. For instance, some models showed a strong reliance on extratumoral tissue, while others demonstrated unexpected insensitivity to small tumor regions. These findings underscore the importance of rigorous model evaluation beyond standard performance metrics and highlight potential pitfalls in clinical deployment.

While our study demonstrates the potential of HIPPO, several limitations must be acknowledged. First, the counterfactual scenarios generated by HIPPO, while informative, may not always reflect biologically plausible tissue alterations. In particular, adding patches from one specimen into another specimen may not always be a realistic intervention. Future work should focus on refining these interventions to more closely mimic realistic tissue changes. Second, our analysis was limited to a specific set of foundation models and datasets. Broader evaluation across diverse pathology tasks and model architectures is needed to fully characterize the generalizability of our findings. In addition the interpretations offered by HIPPO are inherently bound by the underlying model's capabilities and potential shortcomings in representing complex biological systems.

Looking ahead, several avenues for future research emerge from this work. The integration of HIPPO with multi-modal data, including genomic and clinical information, could provide even richer insights into model behavior and biological relevance. Additionally, exploring the use of HIPPO in guiding model refinement, such as targeted fine-tuning based on identified weaknesses, represents a promising direction for improving model robustness and clinical applicability.

In conclusion, HIPPO represents a major advance in the ability to interpret AI models in computational pathology. By providing a quantitative framework for assessing the impact of specific tissue regions on model predictions, HIPPO offers a powerful tool for uncovering model limitations, verifying biological relevance, and biomarker discovery for various clinical applications. As the field of computational pathology continues to evolve, quantitative methods like HIPPO will be crucial in ensuring that AI tools are deployed responsibly and effectively in healthcare settings.

References

[1] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

[2] Amelie Echle, Niklas Timon Rindtorff, Titus Josef Brinker, Tom Luedde, Alexander Thomas Pearson, and Jakob Nikolas Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer*, 124(4):686–696, 2021.

[3] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[4] Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023.

[5] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. Digital pathology and artificial intelligence. *The lancet oncology*, 20(5):e253–e261, 2019.

[6] Emad A Rakha, Michael Toss, Sho Shiino, Paul Gamble, Ronnachai Jaroensri, Craig H Mermel, and Po-Hsuan Cameron Chen. Current and future applications of artificial intelligence in pathology: a clinical perspective. *Journal of clinical pathology*, 74(7):409–414, 2021.

[7] Miao Cui and David Y Zhang. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4):412–422, 2021.

[8] Sandra Morales, Kjersti Engan, and Valery Naranjo. Artificial intelligence in computational pathology–challenges and future directions. *Digital Signal Processing*, 119:103196, 2021.

[9] Hamid Reza Tizhoosh and Liron Pantanowitz. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of pathology informatics*, 9(1):38, 2018.

[10] Khoa A Tran, Olga Kondrashova, Andrew Bradley, Elizabeth D Williams, John V Pearson, and Nicola Waddell. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine*, 13:1–17, 2021.

[11] Didem Cifci, Gregory P Veldhuizen, Sebastian Foersch, and Jakob Nikolas Kather. Ai in computational pathology of cancer: improving diagnostic workflows and clinical outcomes? *Annual Review of Cancer Biology*, 7(1):57–71, 2023.

[12] Artem Shmatko, Narmin Ghaffari Laleh, Moritz Gerstung, and Jakob Nikolas Kather. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nature cancer*, 3(9):1026–1038, 2022.

[13] Raquel Perez-Lopez, Narmin Ghaffari Laleh, Faisal Mahmood, and Jakob Nikolas Kather. A guide to artificial intelligence for cancer researchers. *Nature Reviews Cancer*, pages 1–15, 2024.

[14] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[15] Michael Gadermayr and Maximilian Tschuchnig. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, page 102337, 2024.

[16] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 12 2017.

[17] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester Van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022.

[18] Shuhao Qiu, Yao Guo, Chuang Zhu, Wenli Zhou, and Huang Chen. Attention based multi-instance thyroid cytopathological diagnosis with multi-scale feature fusion. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3536–3541. IEEE, 2021.

[19] Joshua Butke, Tatjana Frick, Florian Roghmann, Samir F El-Mashtoly, Klaus Gerwert, and Axel Mosig. End-to-end multiple instance learning for whole-slide cytopathology of urothelial carcinoma. In *MICCAI Workshop on Computational Pathology*, pages 57–68. PMLR, 2021.

[20] Rocío Del Amor, Pablo Meseguer, Tommaso Lorenzo Parigi, Vincenzo Villanacci, Adrián Colomer, Laëtitia Launet, Alina Bazarova, Gian Eugenio Tontini, Raf Bisschops, Gert De Hertogh, et al. Constrained multiple instance learning for ulcerative colitis prediction using histological images. *Computer methods and programs in biomedicine*, 224:107012, 2022.

[21] Rocío Del Amor, Laëtitia Launet, Adrián Colomer, Anaïs Moscardó, Andrés Mosquera-Zamudio, Carlos Monteagudo, and Valery Naranjo. An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images. *Artificial intelligence in medicine*, 121:102197, 2021.

[22] Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. Ai-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110, 2021.

[23] Ziyu Su, Thomas E. Tavolara, Gabriel Carreno-Galeano, Sang Jin Lee, Metin N. Gurcan, and M.K.K. Niazi. Attention2majority: Weak multiple instance learning for regenerative kidney grading on whole slide images. *Medical Image Analysis*, 79:102462, 2022.

[24] Jiayun Li, Wenyuan Li, Anthony Sisk, Huihui Ye, W. Dean Wallace, William Speier, and Corey W. Arnold. A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in Biology and Medicine*, 131:104253, 2021.

[25] Oliver Lester Saldanha, Chiara ML Loeffler, Jan Moritz Niehues, Marko van Treeck, Tobias P Seraphin, Katherine Jane Hewitt, Didem Cifci, Gregory Patrick Veldhuizen, Siddhi Ramesh, Alexander T Pearson, et al. Self-supervised attention-based deep learning for pan-cancer mutation prediction from histopathology. *NPJ Precision Oncology*, 7(1):35, 2023.

[26] Gabriel Dernbach, Daniel Kazdal, Lukas Ruff, Maximilian Alber, Eva Romanovsky, Simon Schallenberg, Petros Christopoulos, Cleo-Aron Weis, Thomas Muley, Marc A Schneider, et al. Dissecting ai-based mutation prediction in lung adenocarcinoma: a comprehensive real-world study. *European Journal of Cancer*, page 114292, 2024.

[27] Qingyuan Zheng, Xinyu Wang, Rui Yang, Junjie Fan, Jingping Yuan, Xiuheng Liu, Lei Wang, Zhuoni Xiao, and Zhiyuan Chen. Predicting tumor mutation burden and vhl mutation from renal cancer pathology slides with self-supervised deep learning. *Cancer Medicine*, 13(16):e70112, 2024.

[28] Yoni Schirris, Efstratios Gavves, Iris Nederlof, Hugo Mark Horlings, and Jonas Teuwen. Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer. *Medical Image Analysis*, 79:102464, 2022.

[29] Jakob Nikolas Kather, Alexander T Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature medicine*, 25(7):1054–1056, 2019.

[30] Jakob Nikolas Kather, Lara R Heij, Heike I Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M Niehues, Kai AJ Sommer, Peter Bankhead, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature cancer*, 1(8):789–799, 2020.

[31] Renan Valieris, Luan Martins, Alexandre Defelicibus, Adriana Passos Bueno, Cynthia Aparecida Bueno de Toledo Osorio, Dirce Carraro, Emmanuel Dias-Neto, Rafael A Rosales, Jose Marcio Barros de Figueiredo, and Israel Tojal da Silva. Weakly-supervised deep learning models enable her2-low prediction from h &e stained slides. *Breast Cancer Research*, 26(1):124, 2024.

[32] Kevin M. Boehm, Omar S. M. El Nahhas, Antonio Marra, Pier Selenica, Hannah Y. Wen, Britta Weigelt, Evan D. Paul, Pavol Cekan, Ramona Erber, Chiara M. L. Loeffler, Elena Guerini-Rocco, Nicola Fusco, Chiara Frascarelli, Eltjona Mane, Elisabetta Munzone, Silvia Dellapasqua, Paola Zagami, Giuseppe Curigliano, Pedram Razavi, Jorge S. Reis-Filho, Fresia Pareja, Sarat Chandarlapaty, Sohrab P. Shah, and Jakob Nikolas Kather. Multimodal histopathologic models stratify hormone receptor-positive early breast cancer. *bioRxiv*, 2024.

[33] Omar SM El Nahhas, Chiara ML Loeffler, Zunamys I Carrero, Marko van Treeck, Fiona R Kolbinger, Katherine J Hewitt, Hannah S Muti, Mara Graziani, Qinghe Zeng, Julien Calderaro,

et al. Regression-based deep-learning predicts molecular biomarkers from pathology slides. *nature communications*, 15(1):1253, 2024.

[34] Jana Lipkova, Tiffany Y Chen, Ming Y Lu, Richard J Chen, Maha Shady, Mane Williams, Jingwen Wang, Zahra Noor, Richard N Mitchell, Mehmet Turan, et al. Deep learning-enabled assessment of cardiac allograft rejection from endomyocardial biopsies. *Nature medicine*, 28(3):575–582, 2022.

[35] Yuming Jiang, Zhicheng Zhang, Wei Wang, Weicai Huang, Chuanli Chen, Sujuan Xi, M Usman Ahmad, Yulan Ren, Shengtian Sang, Jingjing Xie, et al. Biology-guided deep learning predicts prognosis and cancer immunotherapy response. *Nature Communications*, 14(1):5135, 2023.

[36] Mayur Mallya, Ali Khajegili Mirabadi, Hossein Farahani, and Ali Bashashati. Benchmarking histopathology foundation models for ovarian cancer bevacizumab treatment response prediction from whole slide images. *arXiv preprint arXiv:2407.20596*, 2024.

[37] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4025, 2021.

[38] Richard J Chen, Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022.

[39] Jonas Ammeling, Lars-Henning Schmidt, Jonathan Ganz, Tanja Niedermair, Christoph Brochhausen-Delius, Christian Schulz, Katharina Breininger, and Marc Aubreville. Attention-based multiple instance learning for survival prediction on lung cancer tissue microarrays. In *BVM Workshop*, pages 220–225. Springer, 2023.

[40] Markus Plass, Michaela Kargl, Tim-Rasmus Kiehl, Peter Regitnig, Christian Geißler, Theodore Evans, Norman Zerbe, Rita Carvalho, Andreas Holzinger, and Heimo Müller. Explainability and causability in digital pathology. *The Journal of Pathology: Clinical Research*, 9(4):251–260, 2023.

[41] Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

[42] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: Intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022.

[43] Tianhang Nan, Yong Ding, Hao Quan, Deliang Li, Mingchen Zou, and Xiaoyu Cui. Establishing truly causal relationship between whole slide image predictions and diagnostic evidence subregions in deep learning. *arXiv preprint arXiv:2407.17157*, 2024.

[44] Xin Liu, Weijia Zhang, and Min-Ling Zhang. Attention is not what you need: Revisiting multi-instance learning for whole slide image classification. *arXiv preprint arXiv:2408.09449*, 2024.

[45] Edward Raff and James Holt. Reproducibility in multiple instance learning: a case for algorithmic unit tests. *Advances in Neural Information Processing Systems*, 36, 2024.

[46] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839, 2023.

[47] Milda Pocevičiūtė, Gabriel Eilertsen, and Claes Lundström. Survey of xai in digital pathology. *Artificial intelligence and machine learning for digital pathology: state-of-the-art and future challenges*, pages 56–88, 2020.

[48] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[49] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[50] Zachary C Lipton. The mythos of model interpretability. *arXiv 1606.03490*, 2016.

[51] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[52] Chris Lin, Ian Covert, and Su-In Lee. On the robustness of removal-based feature attributions. *Advances in Neural Information Processing Systems*, 36, 2024.

[53] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[54] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

[55] Mara Graziani, Thomas Lompech, Henning Müller, and Vincent Andrearczyk. Evaluation and comparison of cnn visual explanations for histopathology. In *Proceedings of the AAAI Conference on Artificial Intelligence Workshops (XAI-AAAI-21), Virtual Event*, pages 8–9, 2021.

[56] Pardis Afshar, Sajjad Hashembeiki, Pouya Khani, Emad Fatemizadeh, and Mohammad Hossein Rohban. Ibo: Inpainting-based occlusion to enhance explainable artificial intelligence evaluation in histopathology. *arXiv preprint arXiv:2408.16395*, 2024.

[57] Alex J DeGrave, Zhuo Ran Cai, Joseph D Janizek, Roxana Daneshjou, and Su-In Lee. Auditing the inference processes of medical-image classifiers by leveraging generative ai and the expertise of physicians. *Nature Biomedical Engineering*, pages 1–13, 2023.

[58] Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, pages 1–8, 2024.

[59] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[60] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019.

[61] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.

[62] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[63] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

[64] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S Sara Mahdavi, Ellery Wulczyn, et al. Robust and efficient medical imaging with self-supervision. *arXiv preprint arXiv:2205.09723*, 2022.

[65] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023.

[66] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

[67] Xiyue Wang, Yuexi Du, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Retccl: Clustering-guided contrastive learning for whole-slide image retrieval. *Medical Image Analysis*, 83:102645, 2023.

[68] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman,

et al. Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):1–7, 2017.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[70] Amod A. Sarnaik, Omid Hamid, Nikhil I. Khushalani, Karl D. Lewis, Theresa Medina, Harriet M. Kluger, Sajeve S. Thomas, Evidio Domingo-Musibay, Anna C. Pavlick, Eric D. Whitman, Salvador Martin-Algarra, Pippa Corrie, Brendan D. Curti, Judit Oláh, Jose Lutzky, Jeffrey S. Weber, James M. G. Larkin, Wen Shi, Toshimi Takamura, Madan Jagasia, Harry Qin, Xiao Wu, Cecile Chartier, Friedrich Graf Finckenstein, Maria Fardis, John M. Kirkwood, and Jason A. Chesney. Lifileucel, a tumor-infiltrating lymphocyte therapy, in metastatic melanoma. *Journal of Clinical Oncology*, 39(24):2656–2666, 2021. PMID: 33979178.

[71] Natalie Healey. Tumor infiltrating lymphocyte approval heralds new era for precision cancer immunotherapy. *Nature Medicine*, 30(7):1795—1796, July 2024.

5 Appendix

6 Results

6.1 Refining the search for prognostic tissue biomarkers

Having demonstrated HIPPO's effectiveness in metastasis detection, where the regions of interest are well-defined and were previously annotated by expert pathologists, we extended our investigation to the more complex domain of cancer prognosis. Unlike the clear delineation of tumor regions in metastasis detection, prognostic factors in WSIs are multifaceted and less clearly defined. We applied HIPPO to prognostic models that generate risk scores from WSIs, aiming to identify the tissue regions driving these predictions. Our experiments with HIPPO yielded two key insights. First, HIPPO's search algorithms demonstrated superior ability in identifying tissue patches that consistently and significantly influence risk predictions compared to conventional attention-based methods. While attention mechanisms yielded mixed effects — potentially identifying regions that counterintuitively drive lower risk in otherwise high-risk specimens — HIPPO provided a more consistent, reliable, and quantitative assessment of the regions that drive risk. Second, HIPPO's unique features enable *in silico* experiments to measure the effects of targeted tissue interventions on prognostic outcomes through the lens of the ABMIL model. HIPPO's potential to accelerate the discovery and validation of prognostic tissue biomarkers is an exciting development in cancer research, potentially bridging the gap between computational predictions and clinical actionability.

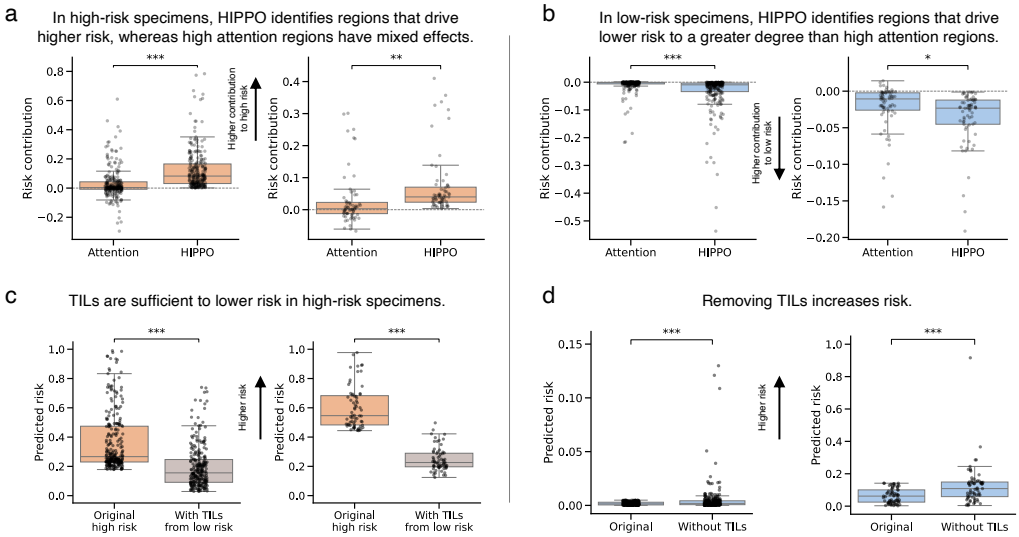

Figure 4: **HIPPO outperforms attention in identifying prognostic tissue regions.** We studied prognostic ABMIL models in invasive breast carcinoma (BRCA) and cutaneous melanoma (SKCM) from The Cancer Genome Atlas. **a, b**, Box plots of the prognostic effects of patches selected using attention and HIPPO in high-risk (**a**) and low-risk (**b**) specimens. The y-axis depicts the risk contribution, which is calculated as the original predicted risk minus the predicted risk when using a specimen with high-attention or high-HIPPO patches removed. Positive values indicate contribution to higher risk (**a**), and negative values indicate contribution to lower risk (**b**). The x-axis is the method of patch selection (either the top 1% of attended patches or the top 1% of patches found using *HIPPO-search-high-effect*). **c**, Box plots showing the predicted risk scores before and after adding tumor-infiltrating lymphocytes (TILs) to high-risk BRCA (left, n=256) and SKCM (right, n=67) specimens. Orange boxes show the original risk scores, and gray boxes show risk scores after adding TILs from low-risk specimens and averaging across low-risk specimens. Lower risk scores indicate improved prognosis. **d**, Box plots showing the predicted risk scores before and after removing TILs from low-risk BRCA (left, n=256) and SKCM (right, n=67) specimens. 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 (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Sample sizes in high-risk (**a, c**) and low-risk (**b, d**) are n=256 for BRCA (left) and n=67 for SKCM (right).

We trained prognostic ABMIL models using the PORPOISE framework [38], a computational tool designed for predicting survival outcomes from histopathology images, to predict overall survival from WSIs in breast cancer (TCGA-BRCA) and cutaneous melanoma (TCGA-SKCM) (Supplementary Fig. 9). The same training and validation splits were used as in the original publication. Non-overlapping 128×128 µm patches from WSIs were embedded using the UNI model [61] (in the original PORPOISE publication, a truncated ResNet50 [69] was used). Low and high risk were defined as the first and fourth quartiles of risk scores. High attention regions were defined as the top 1% of attended patches, and HIPPO search algorithms were also used to identify the top 1% of patches by effect size.

High attention regions drove counterintuitive effects in many specimens, while *HIPPO-search-low-effect* and *HIPPO-search-high-effect* identified more robust and consistent drivers of risk. High attention regions in high-risk cutaneous melanoma specimens (n=67) drove lower risk in 45% (n=30) of specimens. *HIPPO-search-high-effect*, on the other hand, identified regions that all drove higher risk and that more greatly contributed to high-risk predictions ($t = 3.03$, $p < 0.01$, independent t-test). High attention in high-risk breast cancer specimens (n=256) drove lower risk in 40% (n=102) specimens. Again, *HIPPO-search-high-effect* consistently identified regions that drove higher risk in the high-risk specimens ($t = 8.83$, $p < 0.0001$, independent t-test) (Fig. 4a). High attention regions in low-risk SKCM specimens (n=67) drove higher risk in 10% (n=7). *HIPPO-search-low-effect* identified regions that all drove lower risk and more strongly contributed to lower risk predictions ($t = -2.30$, $p < 0.05$, independent t-test). High attention regions in low-risk BRCA specimens (n=256) drove higher risk predictions in 8% (n=20) specimens. *HIPPO-search-low-effect* identified patches that consistently drove lower risk predictions ($t = -5.43$, $p < 0.0001$, independent t-test) (Fig. 4b). This counterintuitive effect underscores that attention scores may not directly relate to model predictions. Thus, interpretations that solely rely on these features may be misguided. HIPPO search algorithms reliably identified the regions that drove risk predictions and may have value as a tool for prognostic biomarker search.

TILs are a well-known prognostic biomarker. We evaluated the necessity and sufficiency of TILs for low-risk predictions in BRCA and SKCM. To test sufficiency, we extracted TIL-positive patches from low-risk specimens and placed them in high-risk specimens. For each high-risk slide, we embedded the TILs from each low-risk slide, and we averaged the model predictions across the low-risk slides to compute the average treatment effect of TILs for each high-risk slide. In high-risk BRCA specimens (n=253, three specimens failed cell detection), the addition of TILs from low-risk specimens decreased the risk by 46% ($t = 17.95, p < 0.0001$, paired t-test) from 0.37 (std. dev. 0.20) to 0.20 (std. dev. 0.15). In SKCM (n=67), the addition of TILs significantly decreased risk by 59% ($t = -22.53, p < 0.0001$, paired t-test) from 0.60 (std. dev. 0.14) to 0.25 (std. dev. 0.08) (Fig. 4c). To evaluate the necessity of TILs, we removed TIL-positive patches from low-risk specimens and measured the change in predictions. If TILs were necessary, then risk predictions would increase upon removal of TILs. In BRCA (n=254, two specimens failed cell detection), the removal of TILs significantly increased risk by 179% ($t = 3.83, p < 0.001$, paired t-test) from 0.002 (std. dev. 0.001) to 0.005 (std. dev. 0.014). In SKCM (n=67), the removal of TILs increased risk by 98% ($t = 4.27, p < 0.0001$, paired t-test) from 0.064 (std. dev. 0.045) to 0.126 (std. dev. 0.123) (Fig. 4d). The removal of TILs did increase risk predictions, but the risk predictions did not reach the level of high-risk slides, suggesting that other features in the WSIs were also driving the low-risk predictions. HIPPO facilitated a quantitative evaluation of the role of TILs on prognosis, providing insights beyond those achievable through the attention mechanism of ABMIL.

6.2 Generating hypotheses of which patients may benefit from autologous TIL therapy

Lifileucel is a promising immunotherapy for melanoma that involves isolating TILs from a patient's tumor, replicating the TILs, and infusing them back into the patient[1]. In a phase II clinical trial, over 30% of patients responded to the therapy [70]. Identifying the patients that might respond to this therapy has the potential to improve patient outcomes and decrease costs (a single treatment may cost over $\$500\,000$ [71]). Therefore, we sought to explore whether we could emulate this with ABMIL and HIPPO. We conducted *in silico* experiments to measure the effect of autologous TILs on prognosis. We used the prognostic model for cutaneous melanoma described above, and we

[1] `https://www.fda.gov/news-events/press-announcements/fda-approves-first-cellular-therapy-treat-patients-unresectable-or-metastatic-melanoma`

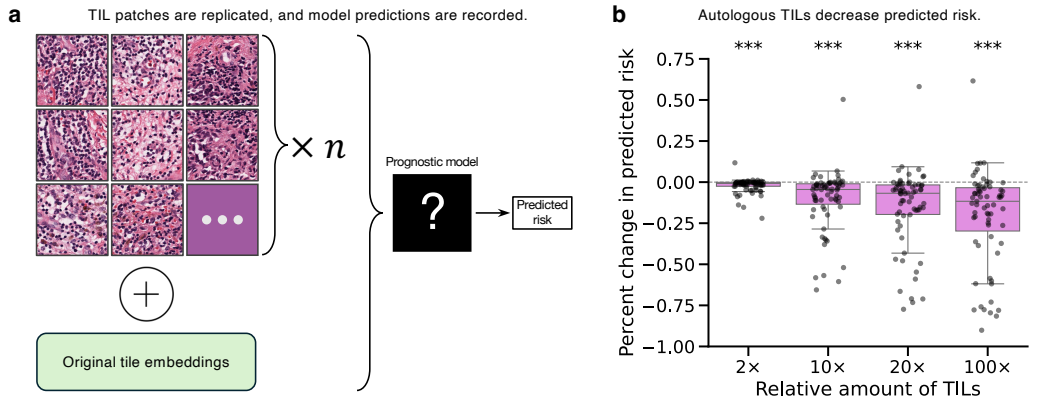

Figure 5: **Autologous TILs improve predicted prognosis.** In high-risk slides of cutaneous melanoma (TCGA-SKCM, n=67), TIL-positive patches were identified using a heuristic from [38]. High risk was defined as slides with the top 25% of predicted risk scores. **a**, The embeddings of TIL-positive regions 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 (***: $p < 0.001$).

studied the high-risk specimens in TCGA-SKCM (n=67 WSIs, n=54 patients). Counterfactuals were designed to model the injection of autologous TILs. In each specimen, TIL-positive patches were replicated $2\times$, $10\times$, $20\times$, and $100\times$ (Fig. 5a). TIL-positive patches were defined using the same heuristic as above (see Methods). The change in model predictions between original specimens and autologous counterfactuals was recorded to measure the effect of additional TILs on prognosis. Cohen's *d* was also calculated to quantify effect sizes. Importantly, we do not claim to demonstrate the efficacy of autologous TIL therapy through HIPPO and TCGA-SKCM. Rather, we aim to show a proof-of-principle that HIPPO may be used for hypothesis generation.

Autologous TILs significantly lowered predicted risk in a dose-dependent manner. Risk decreased by $-2.18\,\%$ ($d = -0.50$) at $2\times$ dose ($t = -4.06$, $p < 0.001$, paired t-test), $-10.8\,\%$ ($d = -0.56$) at $10\times$ dose ($t = -4.59$, $p < 0.0001$, paired t-test), $-15.3\,\%$ ($d = -0.62$) at $20\times$ dose ($t = -5.06$, $p < 0.0001$, paired t-test), and $-20.8\,\%$ ($d = -0.67$) at $100\times$ dose ($t = -5.49$, $p < 0.0001$, paired t-test) (Fig. 5b). Increasing the number of TILs by $100\times$ decreased predicted risk scores by over half in $18\,\%$ of high-risk specimens. Together, we demonstrated a proof-of-principle in which we use HIPPO to identify patients who may benefit from autologous TIL therapy through improved predicted prognosis following the replication of their TILs.

7 Methods

7.1 Prognostic neural network model development

For prognostic models, we used the ABMIL models defined in [38]. The model was composed of a linear layer with 512 units, dropout with a rate of 0.25, and a second linear layer of 256 units. Gated attention was used. The model had four outputs, representing hazards at four points in time. Risk scores were calculated as in ref. [38] and were in range $[0, 1]$, where 0 indicates lowest probability of survival. Models were all implemented in PyTorch, and training was performed on NVIDIA RTX 2080 Ti GPUs.

7.2 Prognostic datasets

Prognostic models were trained and evaluated using the invasive breast carcinoma (BRCA) and cutaneous melanoma (SKCM) studies from The Cancer Genome Atlas. In TCGA BRCA, 1,022 WSIs from 956 patients were used (130 death events), and in TCGA SKCM, 268 slides from 230 patients were used (89 death events). Overall survival time and censoring was used and retrieved from the code repository[2] of ref. [38]. The training and validation splits for cross validation were accessed from the same code repository. The WSIs in TCGA BRCA had $11,260 \pm 6,544$ patches (mean $\pm$ standard deviation). The WSIs in TCGA SKCM had $14,153 \pm 7,471$ patches.

7.3 HIPPO experiment details

Testing the necessity of tumor regions. To assess tumor regions' influence on ABMIL models for metastasis detection, we removed all tumor patches from 49 tumor-positive specimens. Embeddings of patches intersecting expert tumor annotations were removed, and specimens were reclassified as "negative". Model outputs were recorded, and true negative rate (specificity) was calculated for all tested patch embeddings.

Testing the sufficiency of tumor regions. We evaluated tumor region sufficiency in two ways: (1) Using only tumor tissue from positive specimens (n=49); (2) Embedding metastatic patches from positive specimens (n=49) into negative specimens (n=80). For method 1, we removed all non-tumor patches. For method 2, we created 3920 counterfactual examples by adding tumor patches from each positive slide to each negative slide. Sensitivity was measured as the proportion of positive model predictions in both cases.

Testing the effect of tumor size. We explored tumor size effects using in three different ways: (1) a single 128×128 μm tumor region was added to normal specimens (n=80) and positive specimens with tumor removed (n=48); (2) individual tumor patch effect, Removed all tumor patches from positive slides (n=49), then added back one at a time. (3) Incremental tumor size, randomly sampled and added back increasing numbers of tumor patches (i.e., 1, 2, 4, 8, 16, 32, 64) to positive slides with tumor removed. Sensitivity was evaluated as the proportion of positive predictions for each scenario.

Identifying prognostic regions and comparing with attention. We sought to compare the effectiveness of attention and HIPPO for identifying tissue regions related to predicted prognosis. TCGA BRCA and SKCM data were used in these experiments. For attention, regions assigned the top 1% of attention scores were selected. For HIPPO, the search strategy *HIPPO-search-high-effect* was used to identify the regions most contributing to high risk in high-risk specimens, and the search strategy *HIPPO-search-low-effect* was used to identify the regions most contributing to low risk in low-risk specimens. Low and high risk were defined as the first and fourth quartiles of predicted risk scores, respectively. The first 1% of patches identified by the HIPPO search algorithms were selected for evaluation. To quantify the effect of the selected regions on predicted prognosis, we calculated the difference between the predicted prognosis on the original specimens and the predicted prognosis on the specimens with the selected regions removed.

$$\text{Risk contribution of ROI} = \text{Risk using original WSI} - \text{Risk when ROI is removed} \quad (1)$$

Positive values indicated that the regions contributed to higher risk, and negative values indicated that the regions contributed to lower risk. Independent t-tests were used to assess significance of differences between attention and HIPPO.

Effect of TILs on prognostic models. In prognostic models, we measured the effects of tumor-infiltrating lymphocytes (TILs) on model behavior. The number of TILs was quantified using the same approach as Ref. [38]. Briefly, HoVer-Net [60] was used to outline and label the nuclei in TCGA BRCA and SKCM WSIs. The model labels nuclei as one of six categories: tumor epithelium, lymphocyte, stroma, necrosis, normal epithelium, and unknown. Each 128×128 μm was called TIL-positive if it contained more than 20 cells, more than 10 immune cells, and more than 5 tumor cells. In TCGA BRCA, HoVer-Net failed for 12 WSIs, some of which were missing pixel spacing information.

[2] `https://github.com/mahmoodlab/PORPOISE`

We measured the effect of TIL patches on predicted prognosis in TCGA BRCA AND SKCM by either removing TILs from low-risk specimens or adding TILs to high-risk specimens, where low-risk was defined as samples in the first quartile of predicted risk and high-risk were samples in the fourth quartile of predicted risk. The predicted prognoses were compared before and after the intervention. To evaluate the sufficiency of TILs for predicting low risk, we added TIL patches from low-risk specimens to high-risk specimens. Risk predictions of the model were recorded, and differences were tested using paired t-tests. To assess the necessity of TIL regions, we removed TIL-positive patches from low risk specimens and measured risk predictions. Differences were tested using paired t-tests.

Evaluating autologous TILs. Autologous TIL therapy is a promising immunotherapy. We explored how HIPPO could be used for hypothesis generation in the context of autologous TILs in high-risk SKCM specimens (n=67). We sought to assess the degree to which prognostic ABMIL models are effected by the number of TILs in a specimen. We do not claim to assess the efficacy of autologous TILs through HIPPO. The embeddings of TIL-positive regions were replicated $2\times$, $10\times$, $20\times$, or $100\times$, and the change in predicted risk was measured:

$$\text{Change in Risk} = \text{Risk with autologous TILs} - \text{Risk with original WSI} \tag{2}$$

Negative values indicated that the addition of TILs decreased risk. The change in risk from baseline was assessed using paired t-tests.

8 NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the abstract and introduction are supported by experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We describe multiple limitations of this work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We do not report theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the methods and datasets in detail, so that readers may reproduce the experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We include the links from where datasets can be downloaded. However, we were unable to upload the code as supplemental information to OpenReview. We are happy to share the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We list the hyperparameters and details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tests of statistical significance and error bars are present where appropriate.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We list the compute resources used to perform the experiments described in our report.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform to the NeurIPS Code of Ethics. All datasets are public and anonymized.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, we discuss the broad potential impacts of this work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We believe that this paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All original datasets and code are attributed.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We distribute code and documnetation on how to use the code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This research did not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This research does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.