Explainable AI for computational pathology identifies model limitations and tissue biomarkers

Jakub Kaczmarzyk[1,2,3], Joel Saltz[1], Peter Koo[3]

[1]Department of Biomedical Informatics, Stony Brook University, New York, USA; [2]Medical Scientist Training Program, Stony Brook University, New York, USA; [3]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, New York, USA.

Motivation

- Attention-based multiple instance learning (ABMIL) is the predominant method for modeling specimen-level classification tasks in computational pathology.
- Attention is the de facto standard for interpreting these models, but it does not quantify direct effects on model behavior.

Objective

- Measure the effect of tissue regions on model behavior.
- Test tissue-based hypotheses using trained models.

Methods

- HIPPO generates counterfactuals via the addition or removal of patches for an ABMIL model (**Fig. 1**). We measure the change in model behavior induced by the counterfactual.
- HIPPO search finds patches that either drive the highest or lowest effect for a prediction and can identify the regions that are necessary or sufficient for a prediction.
- We used HIPPO to explain models for metastasis detection, prognosis, and *IDH* mutation classification.

Results

- Quantified the necessity and sufficiency of tumor regions for metastasis detection and identified model limitations (**Fig. 2**).
- Adipose tissue sometimes caused false negatives (**Fig. 3**).
- HIPPO reveals that models learned about prognostic effect of TILs in breast cancer and melanoma (**Fig. 4**).
- HIPPO identified regions that drove false negative *IDH* mutation prediction in glioma (**Fig. 5**).

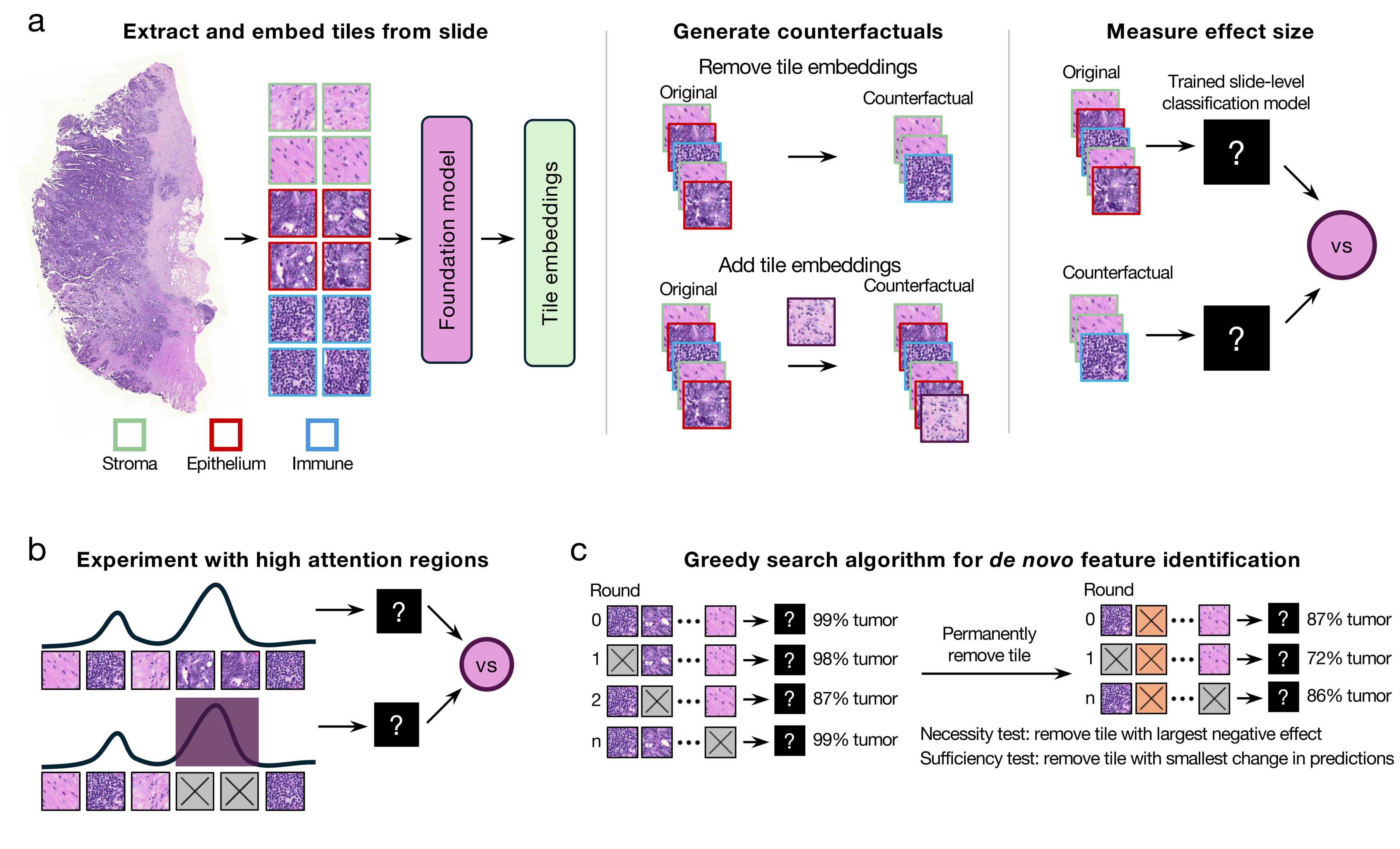

Figure 1. HIPPO explainability framework. **a**, For ABMIL, non-overlapping patches are taken from the whole slide image and embedded with a frozen model. To construct a counterfactual (i.e., "What if?") bag, we add or remove tile embeddings. **b**, The effect of high attention regions can be measured by masking high attention patches and compare model outputs. **c**, We developed search algorithms for *de novo* feature identification.

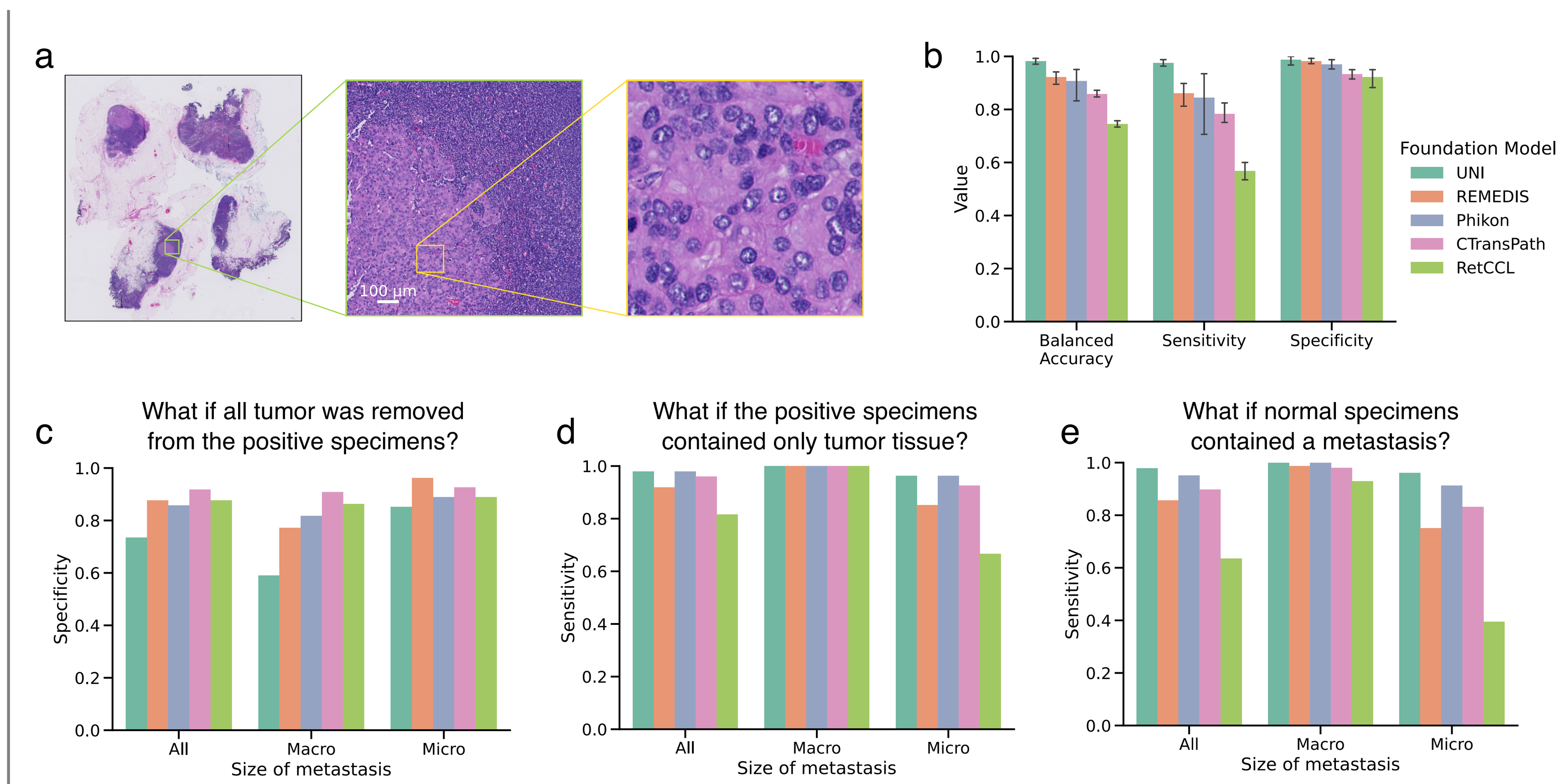

Figure 2. Understanding the role of tumor in metastasis detection. **a**, The CAMELYON16 dataset was used, and **b**, models performed well. **c**, When removing tumor regions, however, some models called many specimens positive. **d**, Removing non-tumor tissue improved sensitivity in four of the five models. **e**, Tumor was sufficient to drive positive detections.

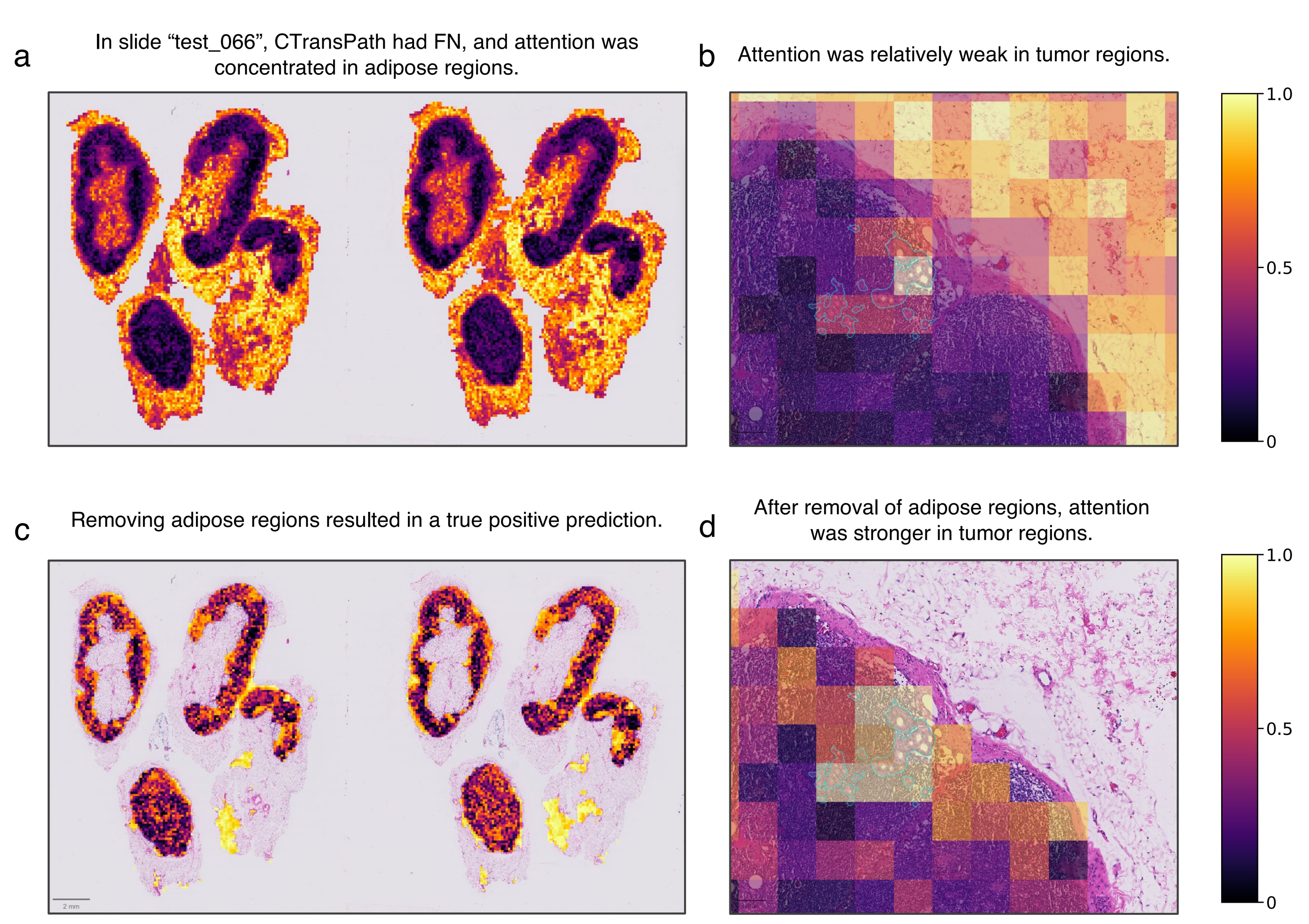

Figure 3. Adipose tissue may cause ABMIL models to miss metastases. **a**, **b**, Attention was high in adipose regions in a false-negative slide. **c**, **d**, After removing adipose regions, the true positive prediction was rescued, and attention was high on tumor regions.

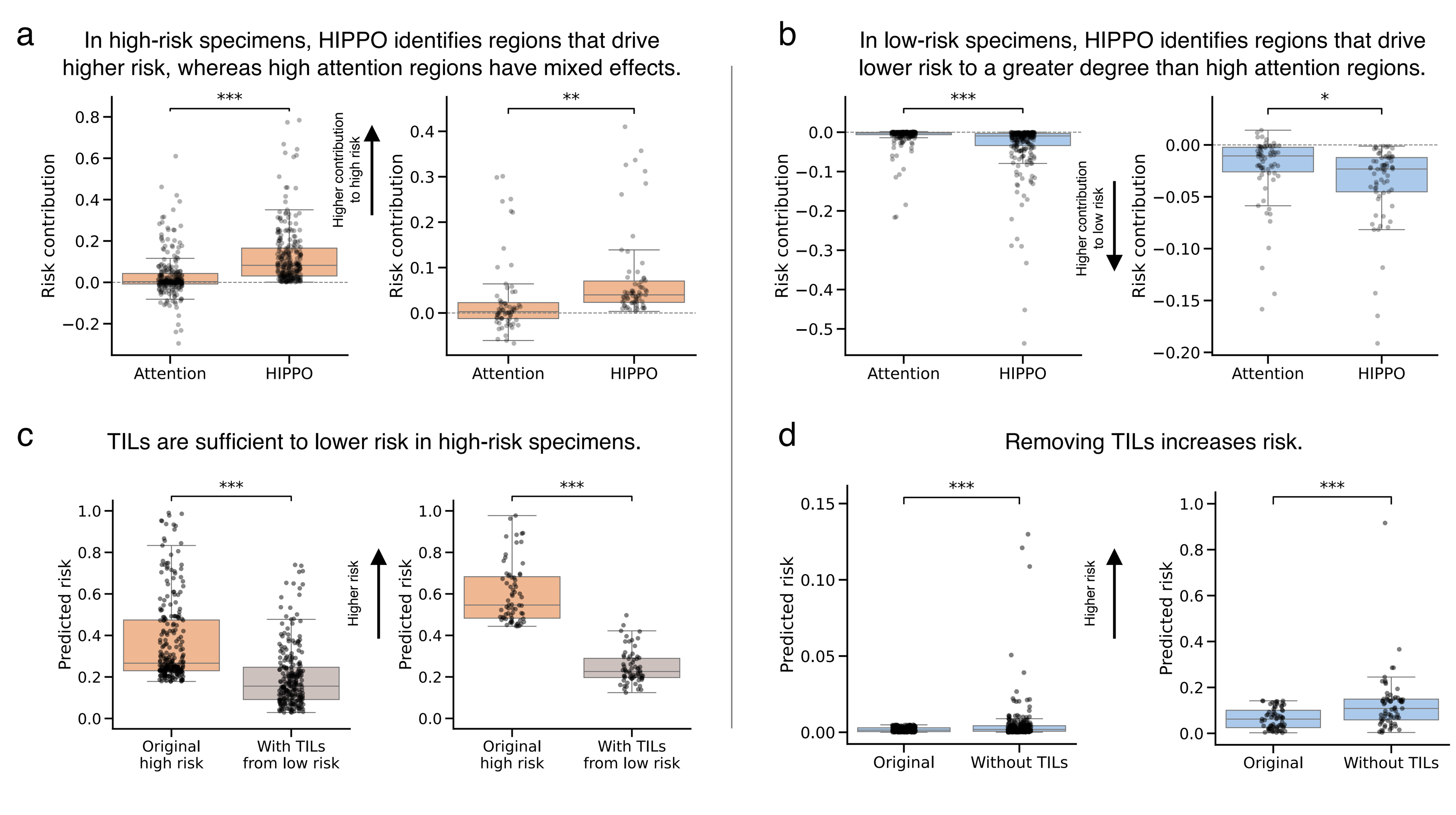

Figure 4. HIPPO outperforms attention in identifying prognostic tissue regions. We trained prognostic models using PORPOISE on TCGA-BRCA and TCGA-SKCM for overall survival. Then we used HIPPO to measure the effects of tissue regions on predicted prognosis. **a**, In low-risk specimens, HIPPO greedy search identified regions that drove more consistent and more negative risk scores than attention (BRCA left, SKCM right). **b**, In high-risk specimens, top 1% attention regions sometimes drove lower risk scores. HIPPO greedy search, on the other hand, identified consistent drives. **c**, TILs were sufficient to lower risk, and **d**, TILs were necessary for low risk to a degree.

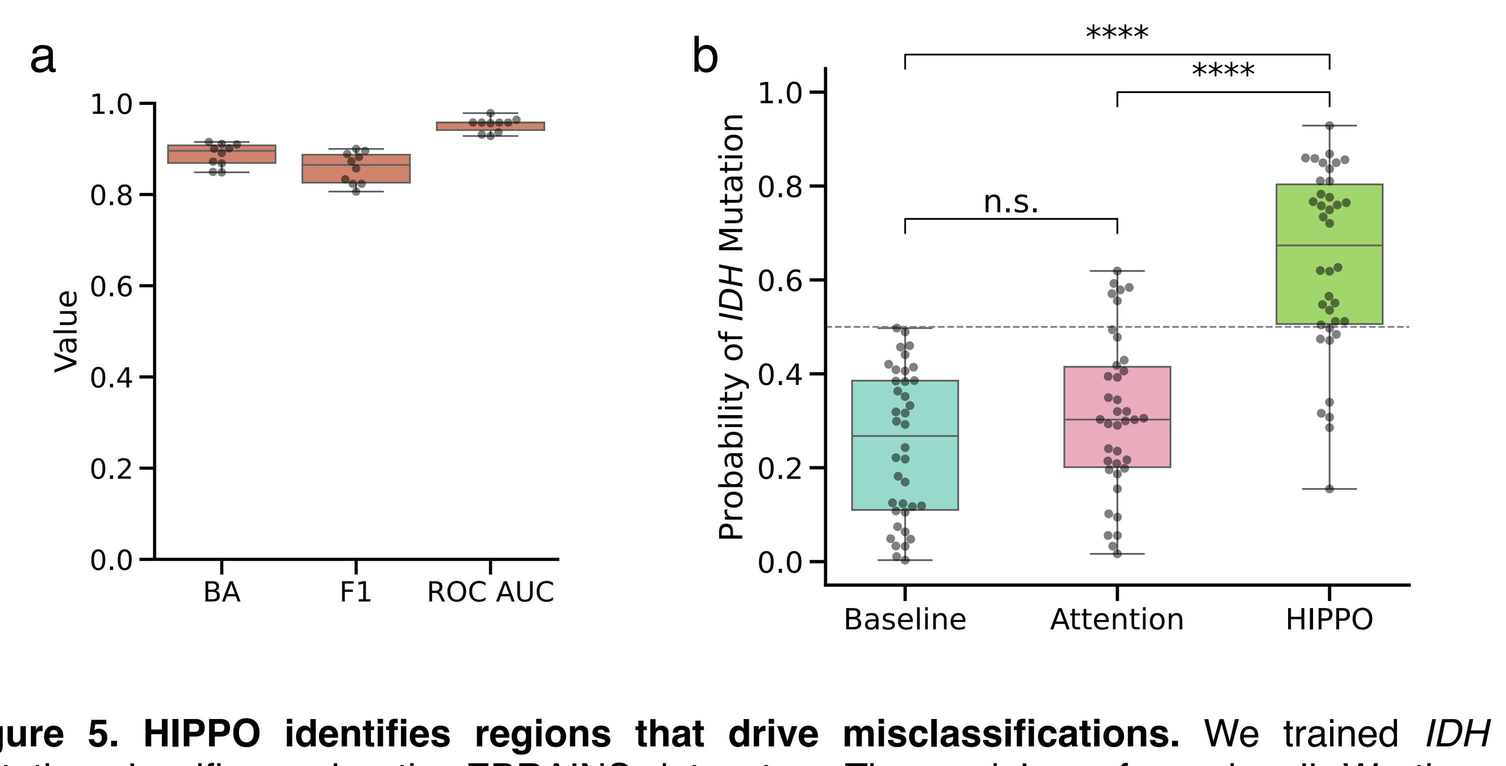

Figure 5. HIPPO identifies regions that drive misclassifications. We trained *IDH* mutation classifiers using the EBRAINS dataset. **a**, The models performed well. We then applied HIPPO greedy search to the false negative classified specimens to identify the regions that drove the misclassifications. **b**, Removing high-attention regions did not significantly change model outputs. Removing patches identified by HIPPO, however, significantly increased the model probability of *IDH* mutation.

HIPPO
Histopathology Interventions of Patches for Predictive Outcomes

Conclusions

- We introduce HIPPO, an explainable AI method designed to address limitations of attention.
- HIPPO enables rigorous model evaluation, bias detection, and quantitative hypothesis testing.
- 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.

Acknowledgements

We are grateful for the support of the Department of Biomedical Informatics at Stony Brook University, the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory, and the Medical Scientist Training Program at Stony Brook University.

References

Chen, R. J. *et al.* Pan-cancer integrative histology-genomic analysis via Multimodal Deep Learning. *Cancer Cell* **40**, (2022).

Ehteshami Bejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 318, 2199 (2017).

Ilse, M., Tomczak, J. M. & Welling, M. Attention-based Deep Multiple Instance Learning. Proceedings of the 35th International Conference on Machine Learning 80, 2127–2136 (2018).

Roetzer-Pejrimovsky, T., Moser, A.-C., Atli, B., Vogel, C. C., Mercea, P. A., Prihoda, R., Gelpi, E., Haberler, C., Höftberger, R., Hainfellner, J. A., Baumann, B., Langs, G., & Woehrer, A. (2022). The Digital Brain Tumour Atlas, an open histopathology resource. *Scientific Data*, 9(1).