

Extending ENPP3 expression inference from tissue microarrays to whole tissue sections: building a case for automated patient enrichment and IHC scoring

Erik Ames Burlingame¹, Fatemeh Koochaki¹, Tsun-Wen Sheena Yao², Shajo Kunnath-Velayudhan², Chaitanya Parmar¹, Laurie Lenox³, Tommaso Mansi⁴, Joel Greshock⁵, Kristopher Standish¹, and Albert Juan Ramon¹

Johnson & Johnson; ¹Data Science and Digital Health (DSDH), AI/ML, Computer Vision; ²Oncology Translational Research; ³Oncology; ⁴DSDH, AI/ML; ⁵DSDH, Oncology

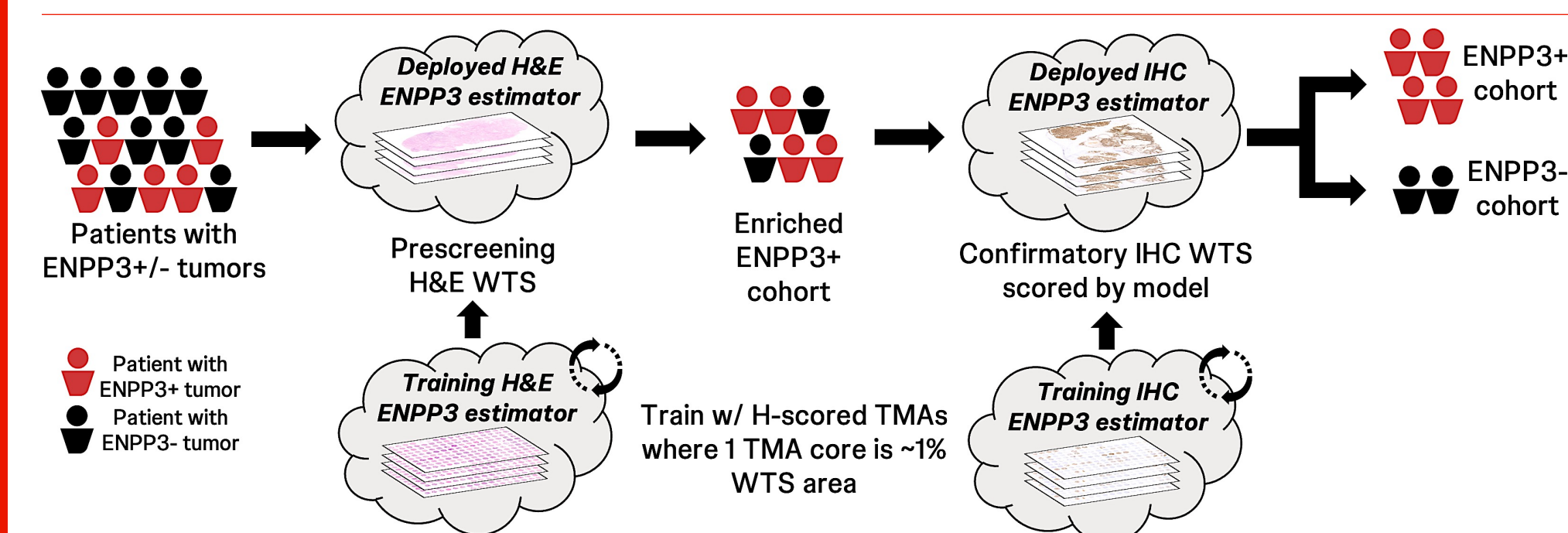


Figure 1. Proposed ENPP3+ patient enrichment strategy to help mitigate potential IHC costs.

Key Takeaway

Building a model to enrich for ENPP3+ patients using standard histopathology images can significantly reduce the number of patients that require IHC diagnostic assessment by pre-excluding those likely to be negative and therefore reduce overall trial-related costs by minimizing IHC-screen failure.

Conclusions

We have developed a prototype AI estimator of ENPP3 expression based on H&E-stained slides that has the potential of serving as a pre-screening tool to reduce IHC screening cost and expedite turn around time (Figure 4).

H-score regression models show promise as an approach to help automate or augment pathologist H-scoring of IHC-stained slides (Figure 5).

Despite their diminutive scale – 1 TMA core contains ~1% of the tissue of a typical whole slide – this proof-of-concept study suggests TMA cores contain sufficient signal to train models for deployment on WTS.



The QR code is intended to provide scientific information for individual reference, and the information should not be altered or reproduced in any way.

Author contact:
Erik Ames Burlingame
eburlin1@its.jnj.com



Introduction

- Enrollment of patients with target-positive tumors is thought to be a factor of success in CD3-redirected therapy trials, such as in the ongoing Phase I study (NCT06178614) evaluating the safety and preliminary anti-tumor activity of ENPP3xCD3 (JNJ-87890387) in ENPP3-unselected, advanced-stage solid tumors with a high prevalence of ENPP3 expression, including colon adenocarcinoma (COAD) and lung adenocarcinoma (LUAD).
- Immunohistochemistry (IHC)-based target expression analysis and patient selection is often limited by significant testing costs and slow turn around time due to laborious pathologist scoring.
- To help mitigate potential future IHC testing costs, we propose a patient pre-screening approach that uses hematoxylin and eosin (H&E)-stained tissues to estimate tumor ENPP3 levels (Figure 1, left) and a complementary approach that estimates H-scores from ENPP3 IHC slides (Figure 1, right), to help automate the laborious, subjective, and time-consuming scoring done by pathologists.
- We developed proof-of-concept AI models that infer ENPP3 expression in H&E- and ENPP3 IHC-stained tissue microarrays (TMA) containing COAD and LUAD tissues.

Methods & Materials (cont.)

Tissue	# train/val cores	# train/val cases	# test tissues	# test cases
COAD	622	382	29 WTS	29
LUAD	258	258	46 cores	46

Table 1. Data summary. Test set for COAD (LUAD) is WTS (cores).

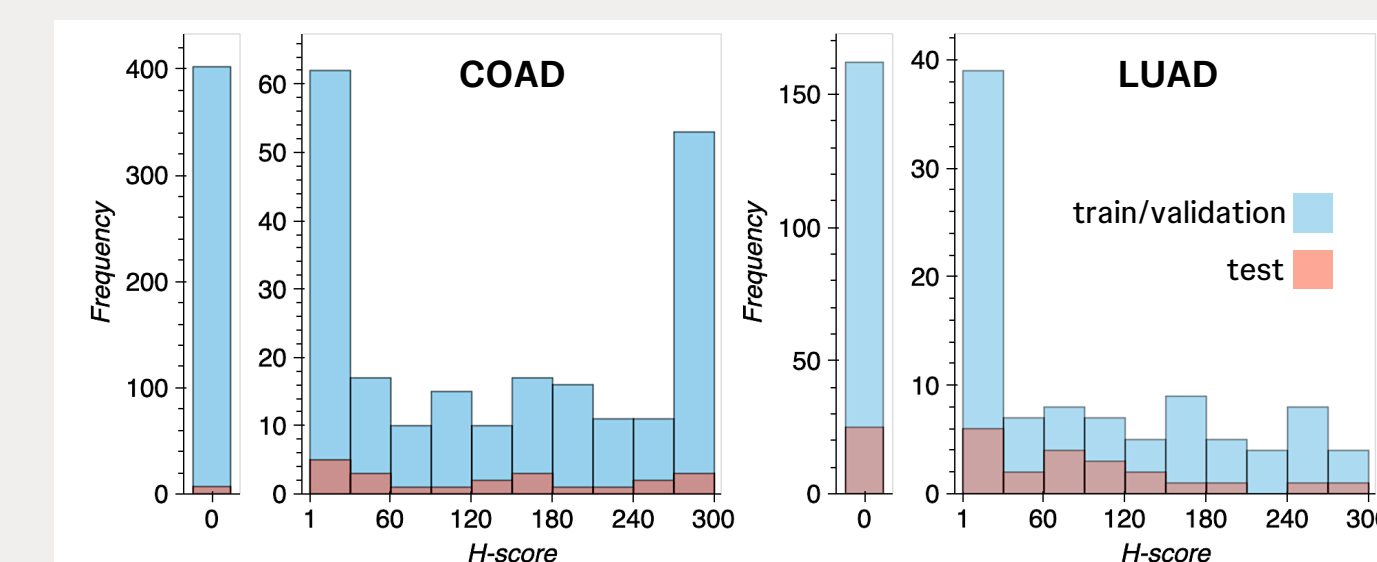


Figure 2. Histograms showing distribution of H-scores for train/validation sets with test sets overlaid. H-score=0 bin separated to visualize scale. All H-scores are from one pathologist scorer.

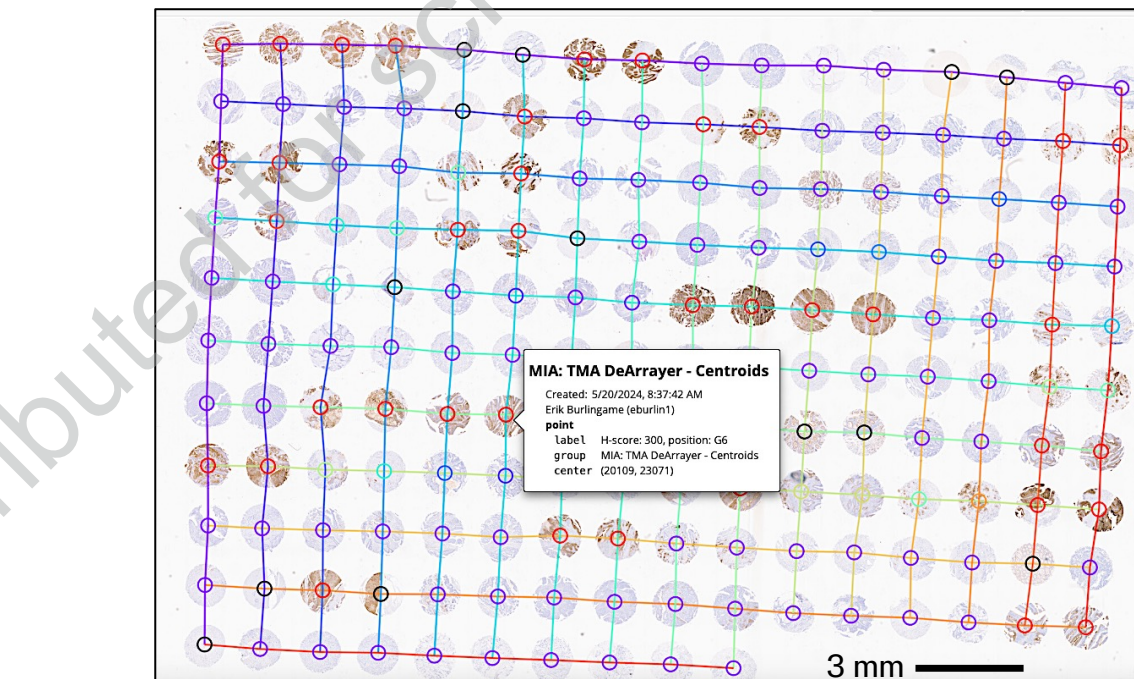
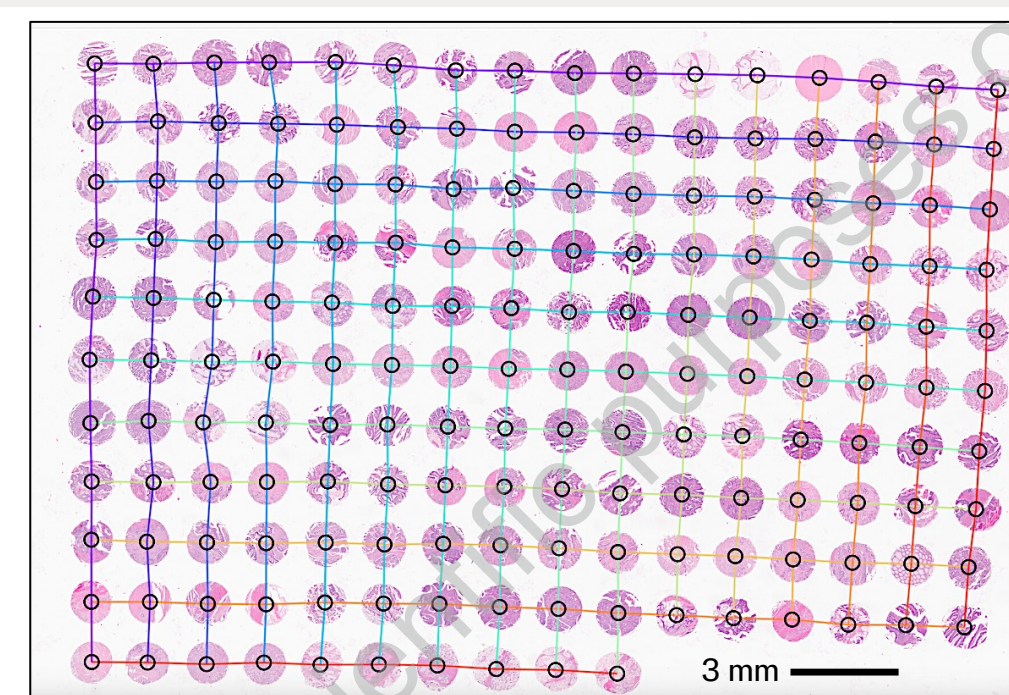


Figure 3: Alignment of paired adjacent H&E and ENPP3 IHC TMA images in the TMA DeArrayer workflow. The workflow consists of 3 steps: 1) automatic core centroid detection (see points on core centroids—color for IHC centroids indicates pathologist H-score); 2) automatic assignment of each core to the alphanumeric grid used during pathologist scoring (see grid lines); 3) automatic extraction of individual core image patches and their associated H-scores for downstream feature extraction and modeling.

Methods & Materials

TMA pre-processing

- Training models typically requires a large set of whole tissue sections (WTS), which can be challenging to score, and may not be available in early clinical studies. Alternatively, TMAs can incorporate hundreds of cases on a single slide and are readily available through commercial vendors, reducing the number of slides needed for model development. As TMA cores are typically orders of magnitude smaller than WTS, we test the hypothesis that models trained on TMA cores can be deployed on WTS. The data are described in Table 1 and Figure 2.
- We developed a human-in-the-loop workflow that dearrays TMAs and aligns individual tissue cores with their associated metadata (Figure 3).

Foundational model pre-training

- A vision transformer-based model was pre-trained using the DINOv2 framework¹ on 55,000 H&E-stained WTS spanning tissue types. This foundation model (FM) serves as an image feature extractor preconditioned for histopathology tasks.

ENPP3 estimation modeling

- The FM is used to extract features from tiled image patches, whether from TMA core (~10-100 patches) or WTS (~1,000-10,000 patches).
- Separate weakly supervised attention-based aggregation models² are trained to either estimate whether a core's ENPP3 H-score is greater than 0 (H-score > 0 classification) or estimate the H-score directly (H-score regression) using either H&E or IHC patch features as input.
- We employ a 5-fold cross validation approach and use the median of the model ensemble output when testing.

Results

- H-score > 0 classification is tractable with H&E COAD (Figure 4) and IHC models, but H-score regression is only tractable with IHC models (Table 2).
- H-score regression models trained on IHC-stained tissues approximate pathologist-determined H-scores (Figure 5).
- H&E-trained models attend to ENPP3+ tumor regions during inference (Figure 6).
- Despite the FM being trained exclusively on H&E-stained tissues, the features it derives from IHC-stained tissues are sufficient for H-score model development until a dedicated IHC FM model is developed.

Input stain	Tissue	Magnification	Model task	Median test performance
H&E	COAD	20X	cls	AUC = 0.79 AP = 0.94 (0.76)
			reg	ICC = 0.25 [-0.12, 0.56]
	LUAD	20X	cls	AUC = 0.44 AP = 0.42 (0.46)
			reg	ICC = 0.10 [-0.20, 0.37]
		40X	cls	AUC = 0.61 AP = 0.58 (0.46)
			reg	ICC = -0.02 [-0.31, 0.27]
ENPP3 IHC	COAD	20X	cls	AUC = 0.90 AP = 0.97 (0.76)
			reg	ICC = 0.76 [0.55, 0.88]
	LUAD	20X	cls	AUC = 0.97 AP = 0.97 (0.46)
			reg	ICC = 0.81 [0.68, 0.89]
		40X	cls	AUC = 0.90 AP = 0.92 (0.46)
			reg	ICC = 0.88 [0.80, 0.93]

Table 2. Model performance. cls: H-score>0 classification; reg: H-score regression; AUC: area under the receiver operating curve; AP: average precision; ICC: intraclass correlation coefficient. For AP entries, the value in parentheses represents the chance level. For ICC entries, the values in the square brackets represent the 95% confidence interval. Bold entries correspond to results presented in Figure 4 (top entry) and Figure 5 (bottom entry).

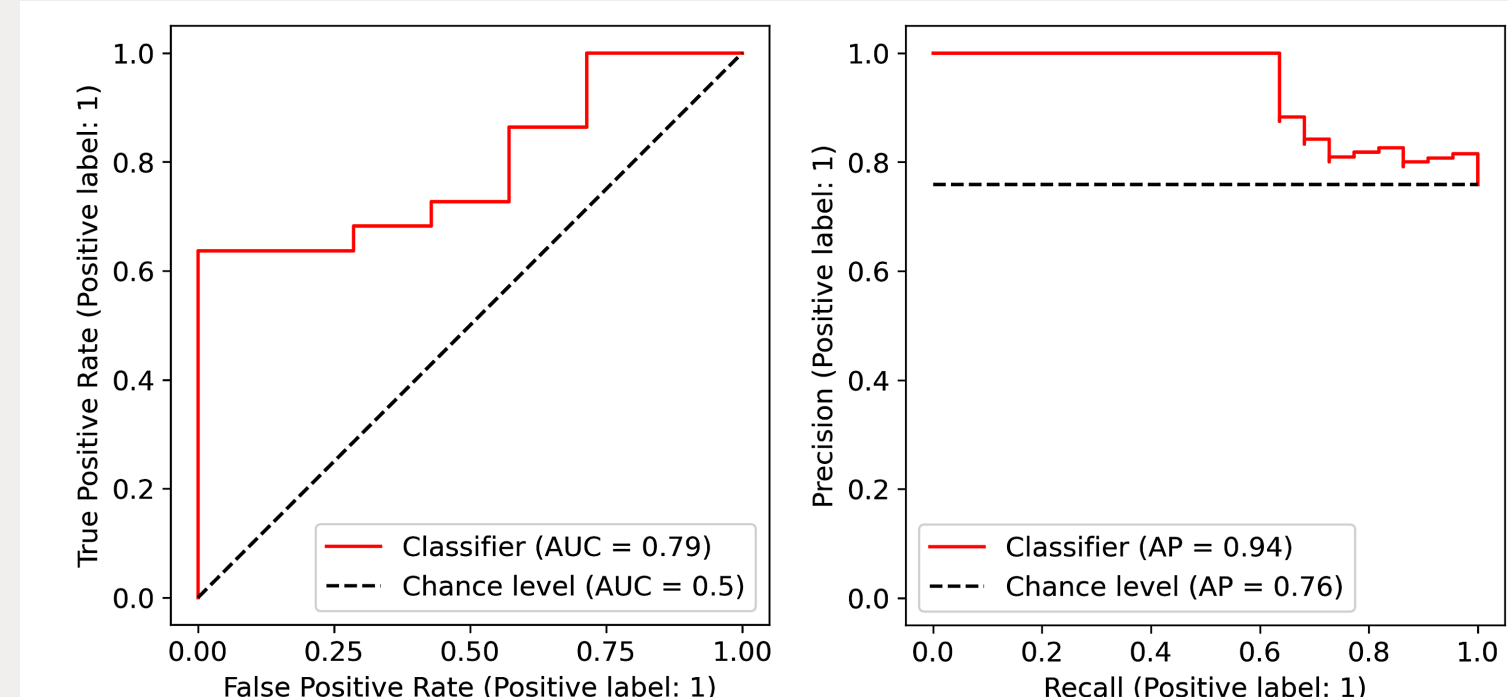


Figure 4. Receiver operator curve (left) and precision-recall curve (right) for H&E COAD 20X H-score>0 classification model trained on TMAs then tested on WTS.

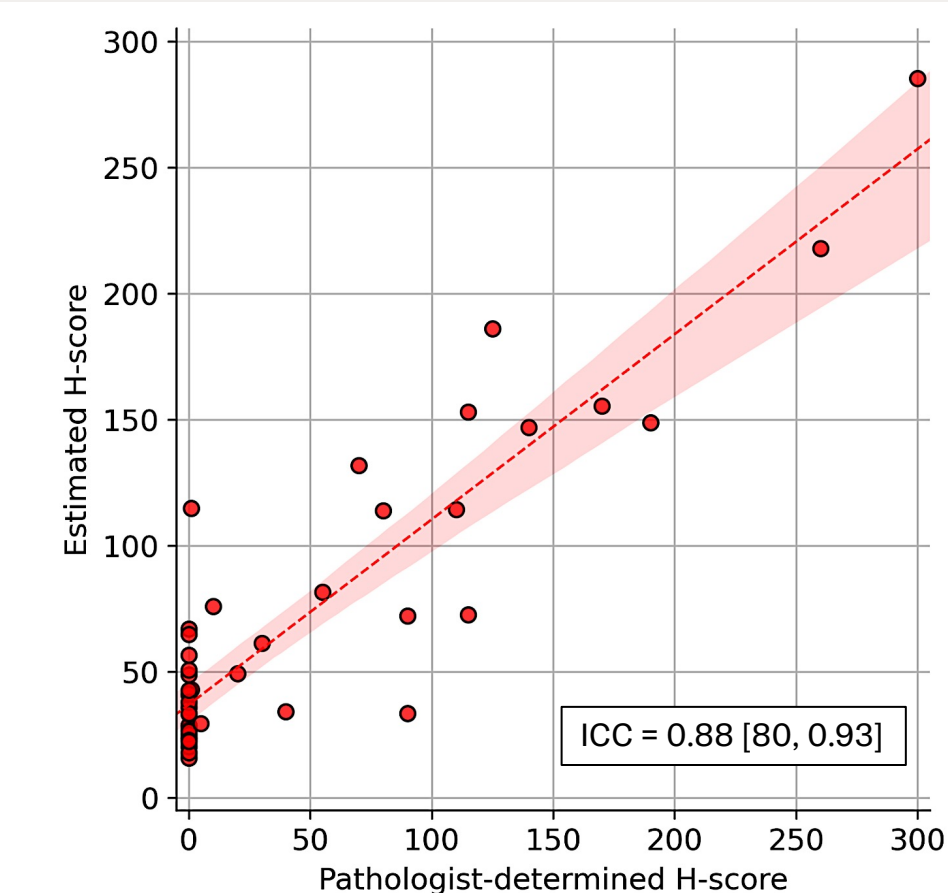


Figure 5. Pathologist-determined H-scores vs H-score regression model estimates from IHC LUAD 40X model. Dotted line is linear regression line and shaded area is 95% confidence interval from 1000 bootstrap resamples.

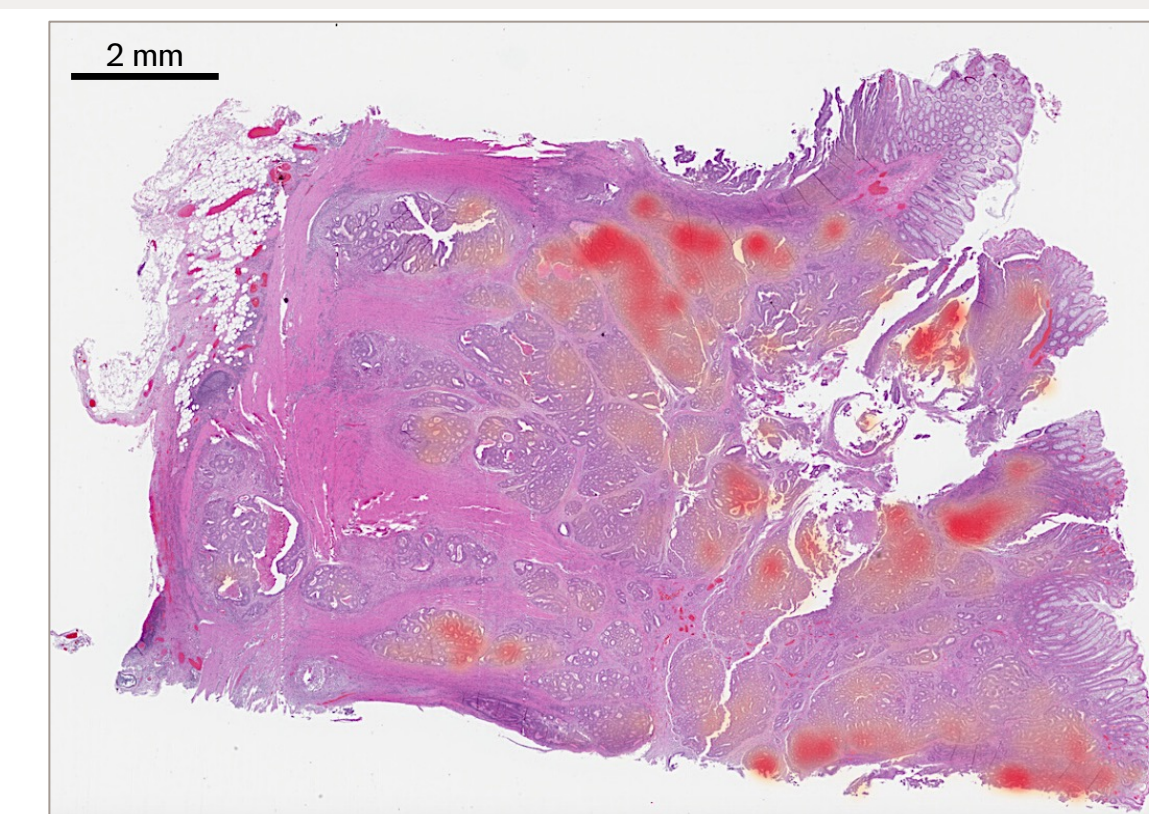


Figure 6. Attention heatmap indicates regions of test COAD H&E WTS that were influential to the model's $p(\text{H-score} > 0)$ estimate. The model attends to regions in the H&E slide (top) that correspond to ENPP3+ tumor regions in the adjacent ENPP3 IHC-stained slide (bottom).

References

- [1] Oquab, Maxime, et al. "Dinov2: Learning robust visual features without supervision." *arXiv preprint arXiv:2304.07193* (2023).
- [2] Ilse, Maximilian, Jakob Tomczak, and Max Welling. "Attention-based deep multiple instance learning." *International conference on machine learning*. PMLR, 2018.