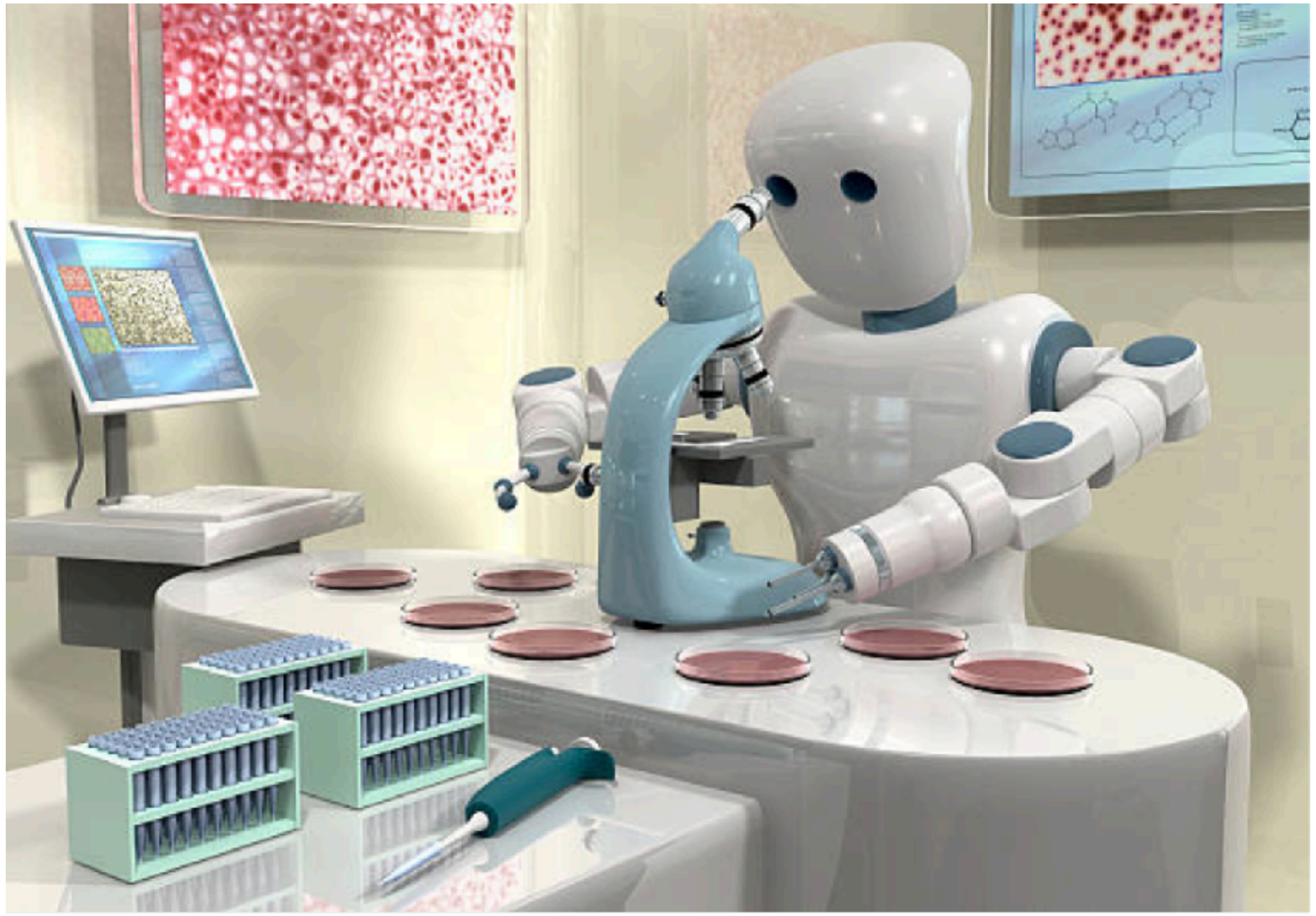
AI-based assessment of cardiac allograft rejections

Jana Lipkova

HARVARD MEDICAL SCHOOL
BWH BRIGHAM AND WOMEN'S HOSPITAL
BROAD INSTITUTE
Mahmood Lab AI for Pathology
OCTOPATH AI for Pathology
UC Irvine Health School of Medicine

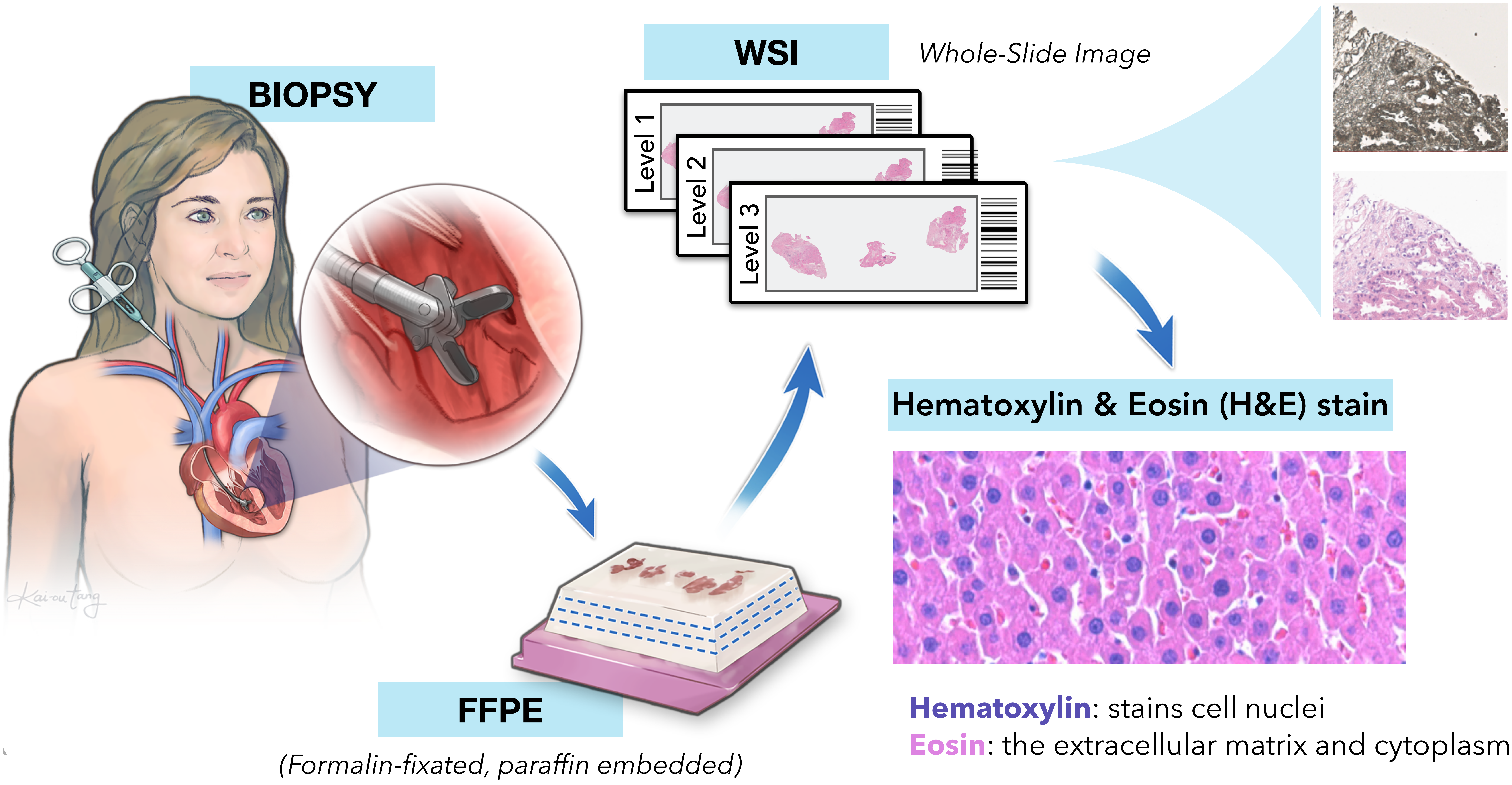
Overview

- **Background:**
  - Histopathology data
  - Cardiac Allograft Rejections
- **AI-based assessment of allograft rejections**



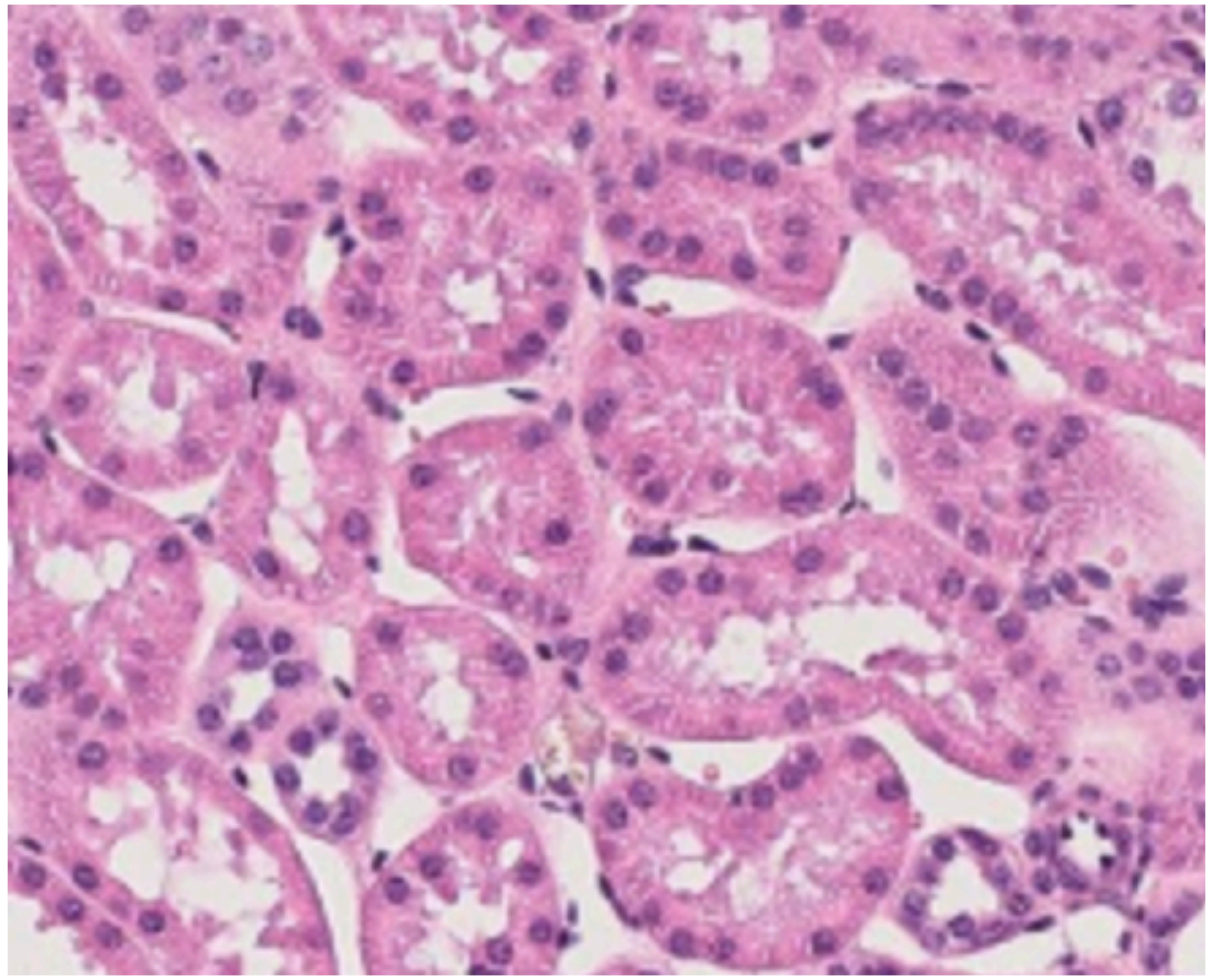
(source GettyImages)

Histology 101



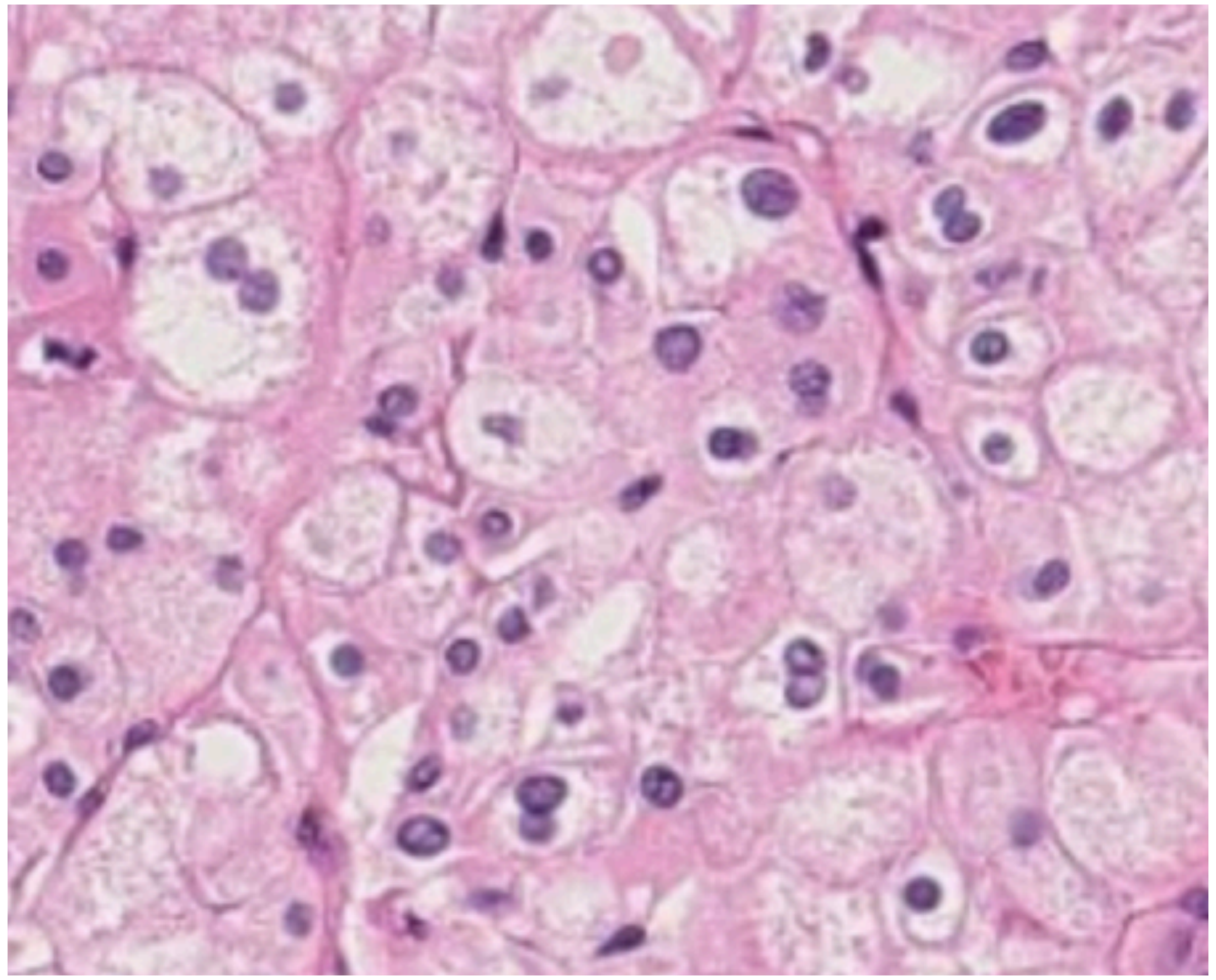
Cancer detection/classification 101

Normal tissue (kidney)



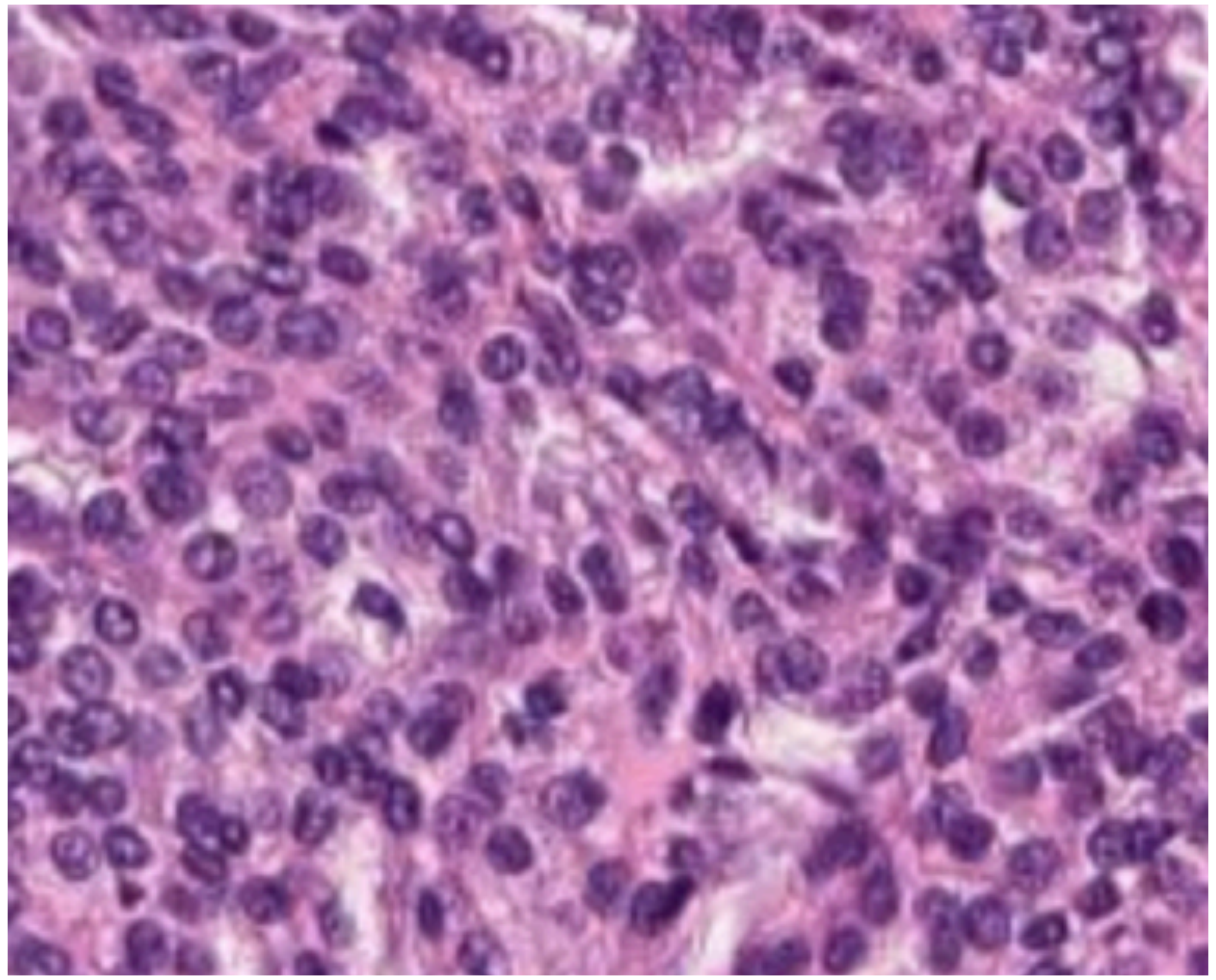
- Symmetric regular structure
- One nuclei per cell
- Cell/nuclei - regular shape

Chromophobe renal carcinoma



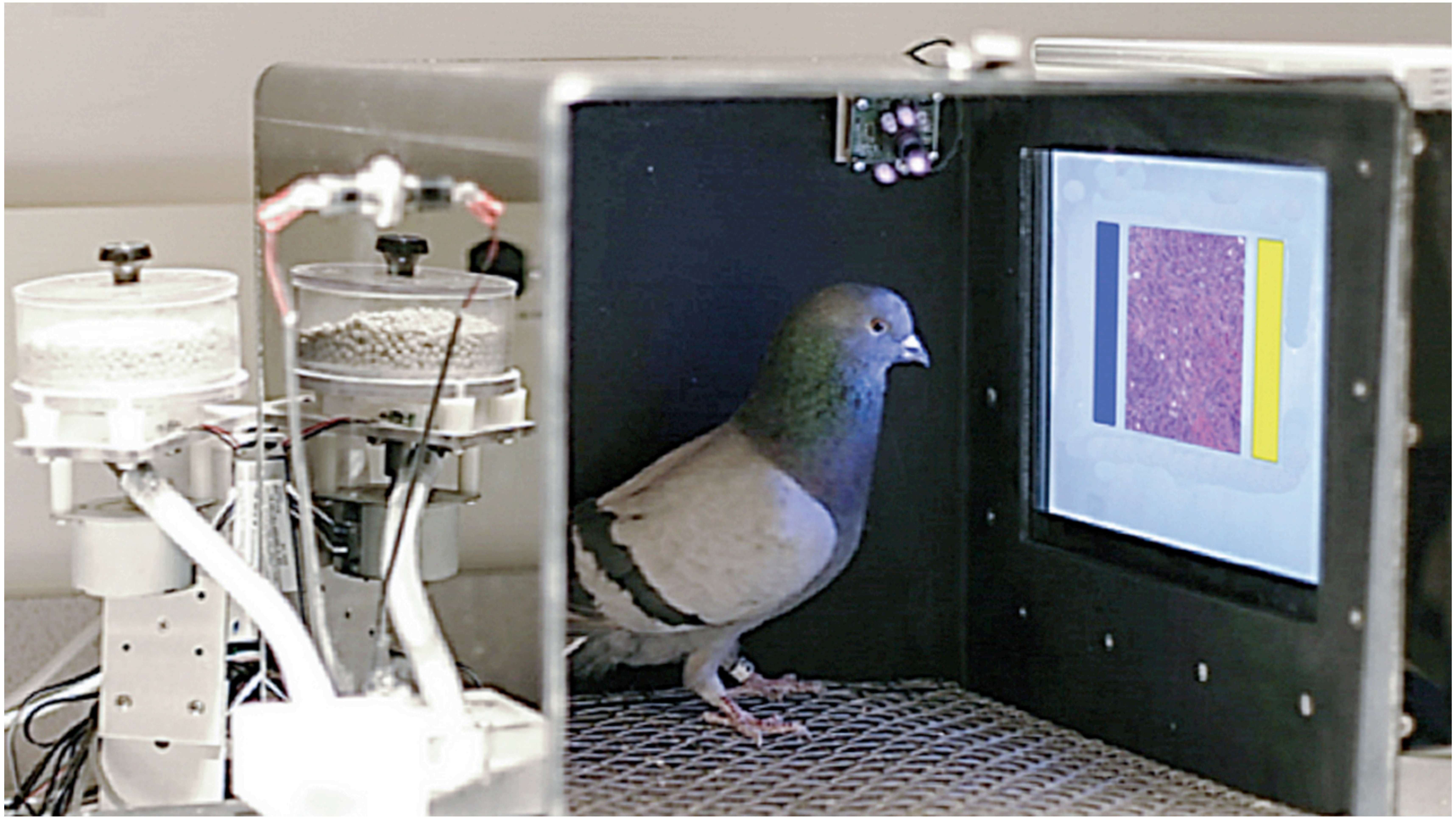
- Enlarged nuclei
- Double nuclei per cell
- Irregular shape

Papillary renal carcinoma



- Papillary cores lined by neoplsatic cells
- Tubulopapillary architecture

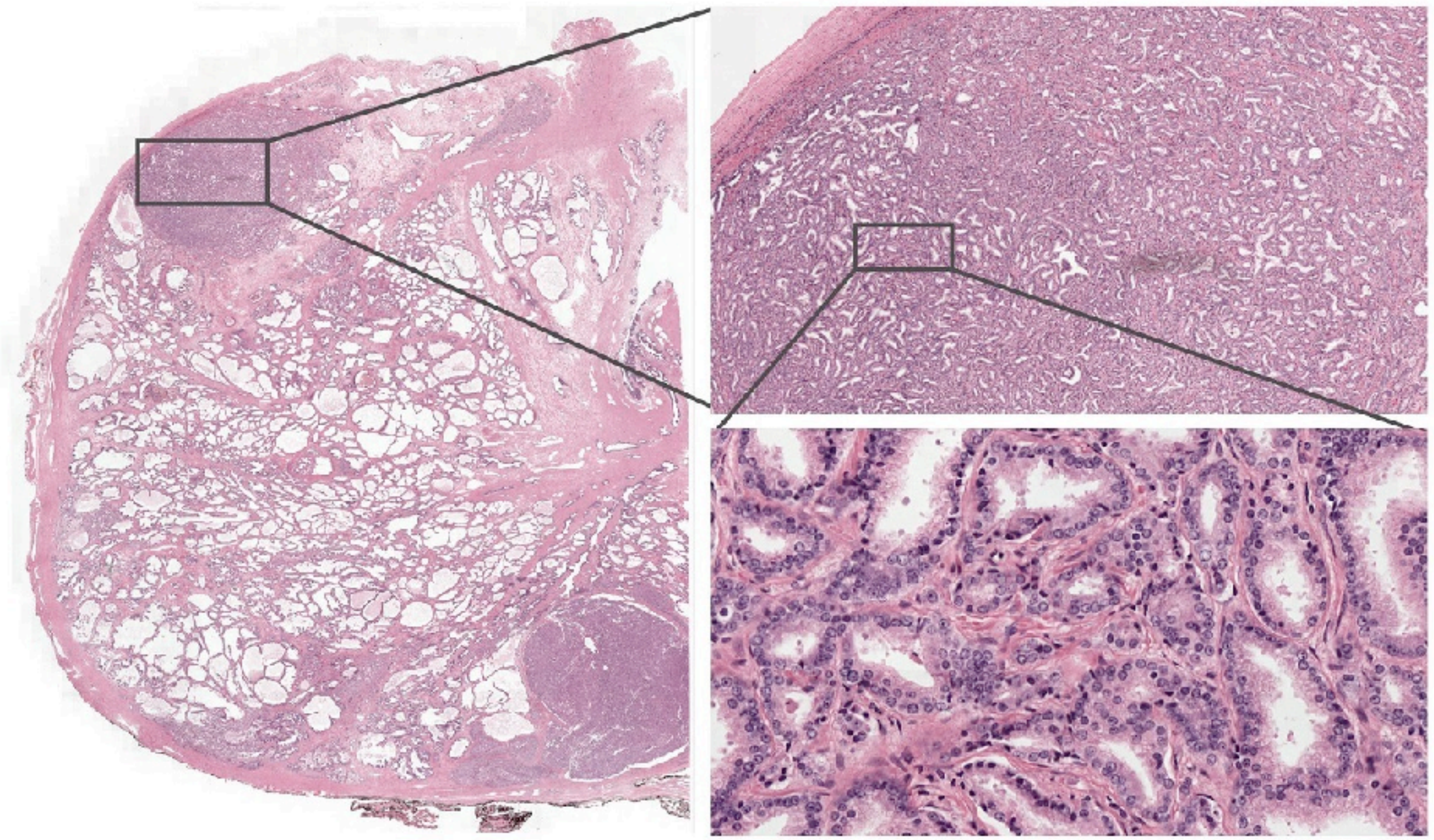
Fun fact: Also pigeons can detect cancer

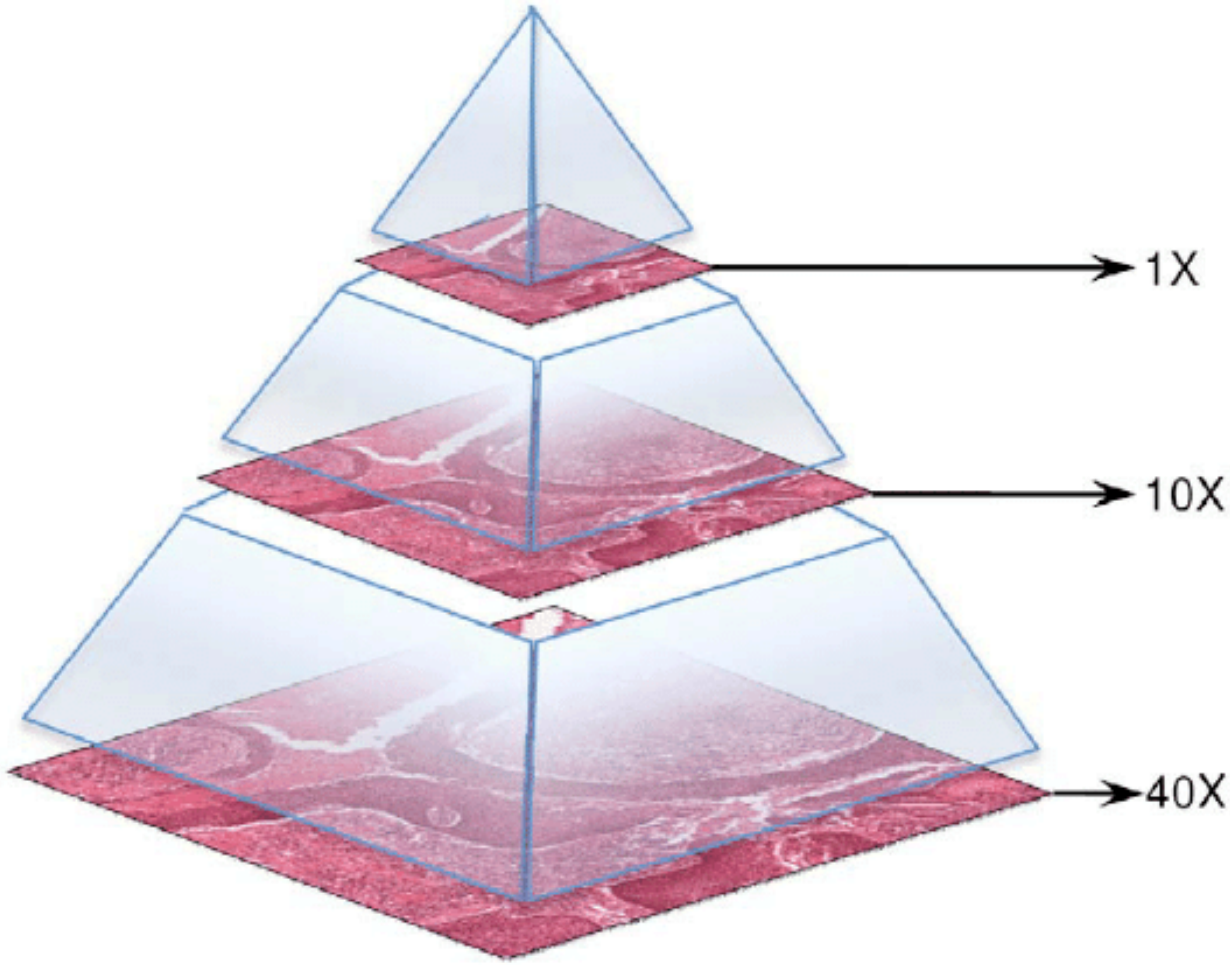


R. Levenson et al: Pigeons (Columba livia) as Trainable Observers of Pathology and Radiology Breast Cancer Images, PloS one, (2015)

Digital Pathology: Whole Slide Images (WSIs)

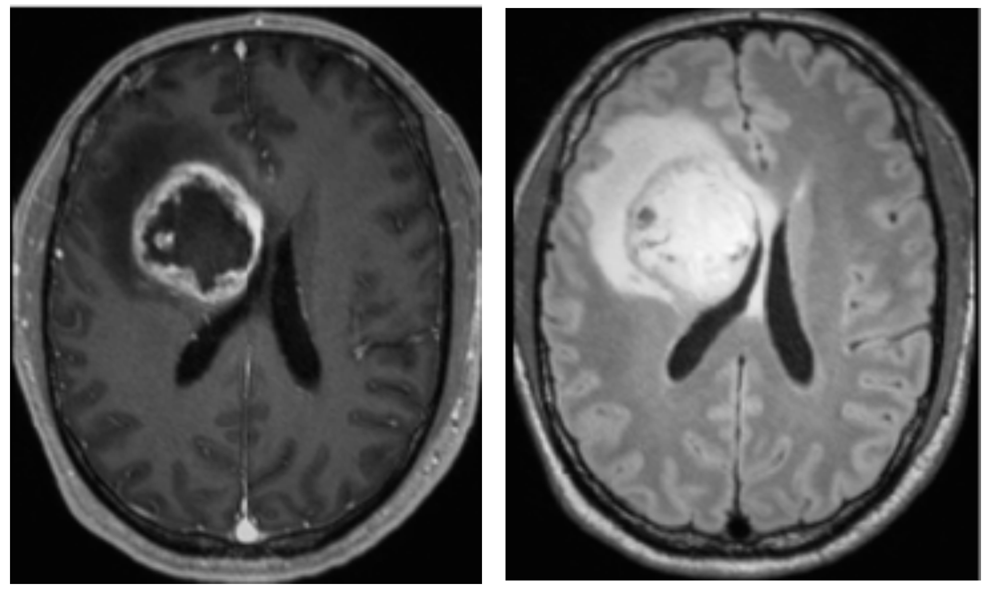
- High resolution scan of an entire tissue section (0.25 - 0.5 microns per pixel)
- Gigapixel image: 100,000 x 100,000 pixels
- **100 WSI have cca same amount of pixels as whole ImageNet**
- Different Stains: H&E, IHC





Medical Data

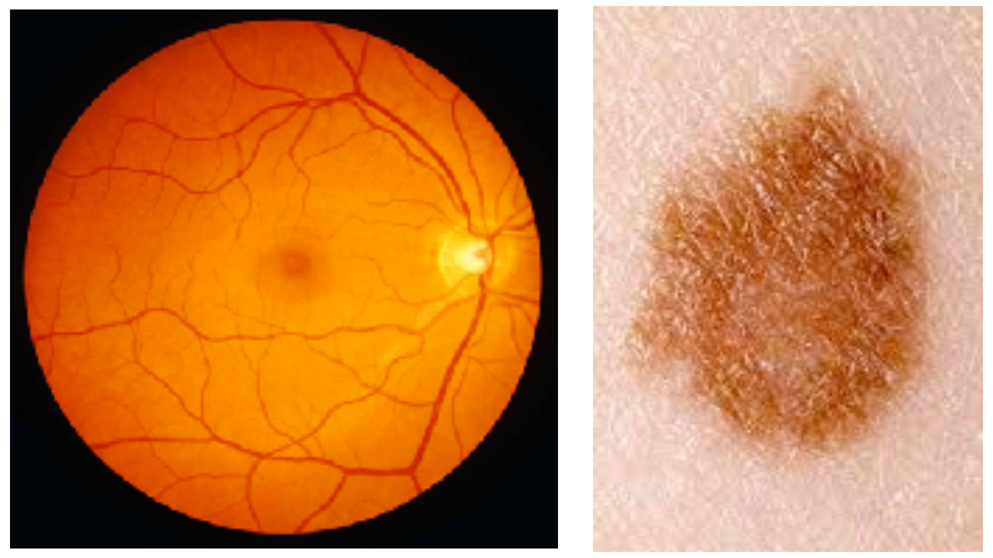
Radiology



(MRI head scan)

- 3D images
- gray-scale images
- resolution: ~1 mm
- size: 256x256x256 voxels

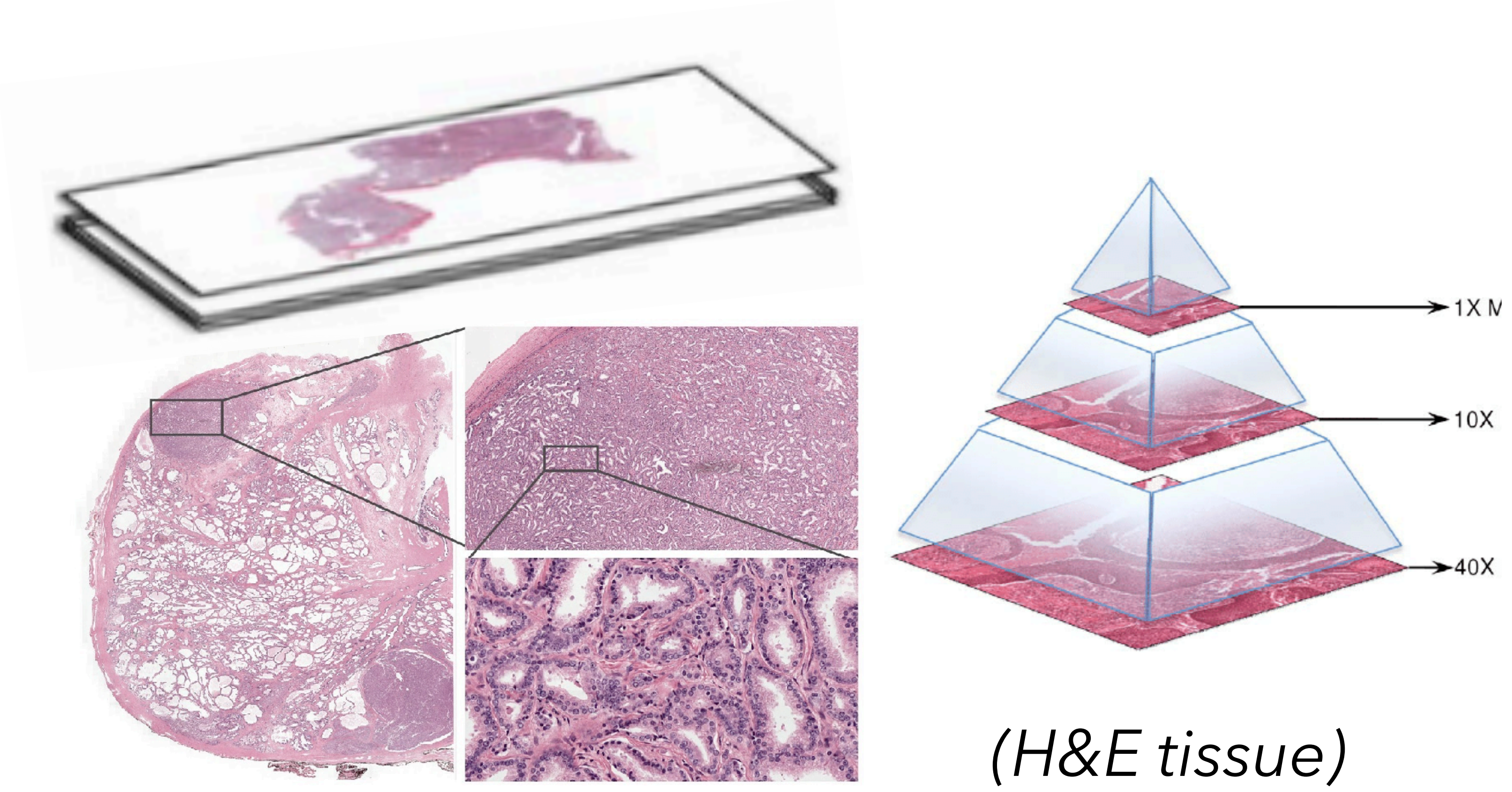
Photography



(Fundus / skin photography)

- 2D images
- RGB
- 10 µm - 1 mm
- size: ~1,700x1,700 pixels

Histology

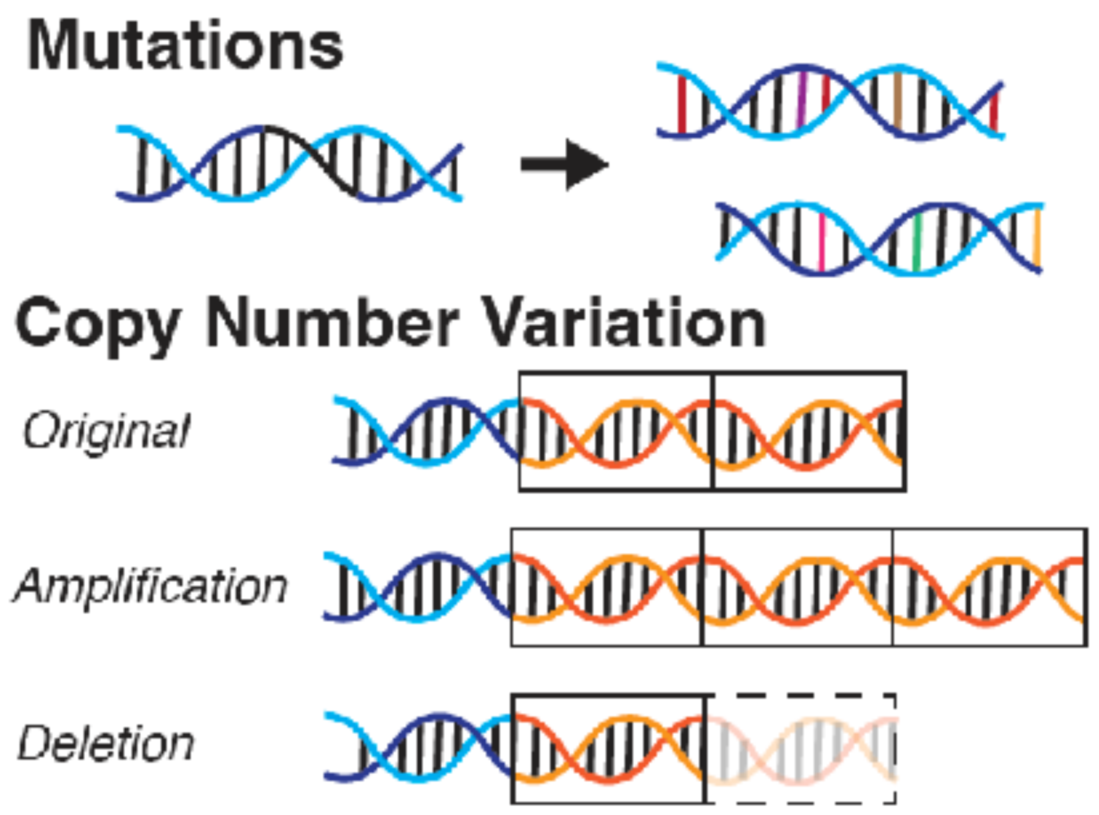


(H&E tissue)

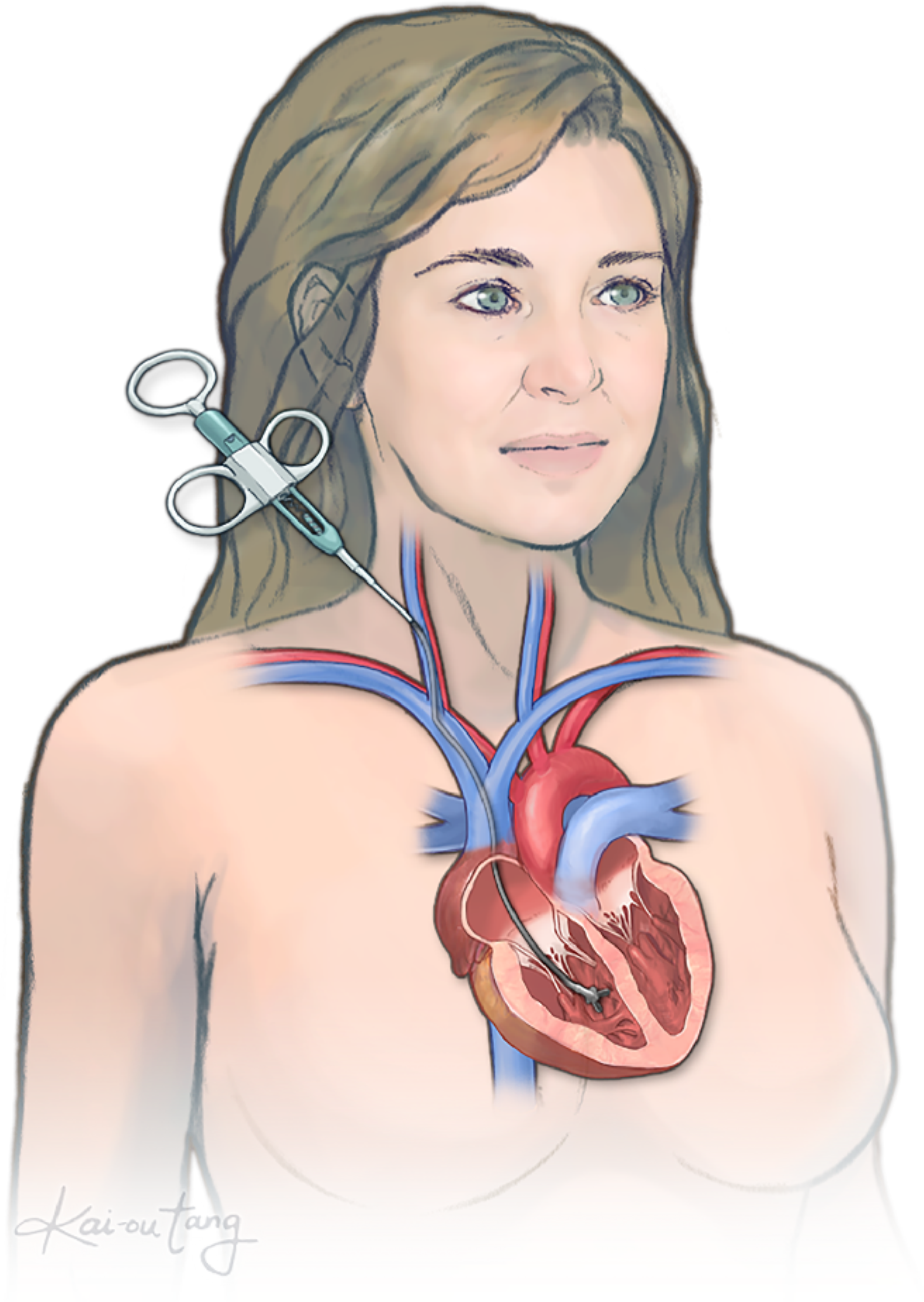
- 2D images
- RGB
- scale: ~0.1µm
- 100,000x100,000 pixels

(varies with magnification, tissue size etc)

Genomics



- 1D array
- float (e.g. '0' wild type, '1' mutation)
- scale: 1µm - 1nm
- ~20,000 protein-coding genes



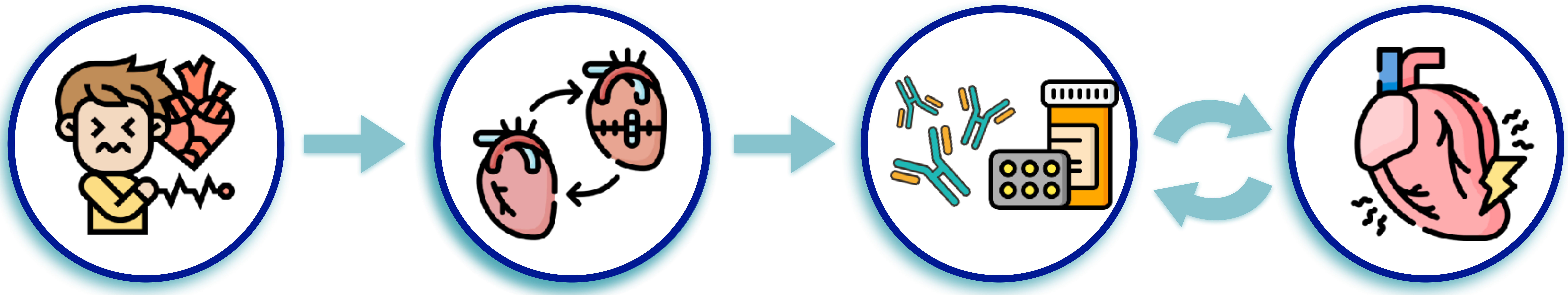
AI-Assessment of Cardiac Allograft Rejections

Jana Lipkova, Tiffany Y Chen, Ming Y Lu, Jingwen Wang, Maha Shady, Mane Williams, Richard Mitchell, Mehmet Turan, Gulfize Coskun, Derya Demir, Deniz Nart, Funda Y Barbet, Katja E Odening, Yara Banz, Faisal Mahmood

Lipkova et al. Nature Medicine (2022)

BACKGROUND

Heart Failure | **Heart Transplant** | **Immune Response** | **Allograft Rejection**



Heart Failure
- Leading cause of hospitalization in USA/EU
- 26 million cases / year

Heart Transplant
- Patients with end-stage failure
- 5000 transplants / year

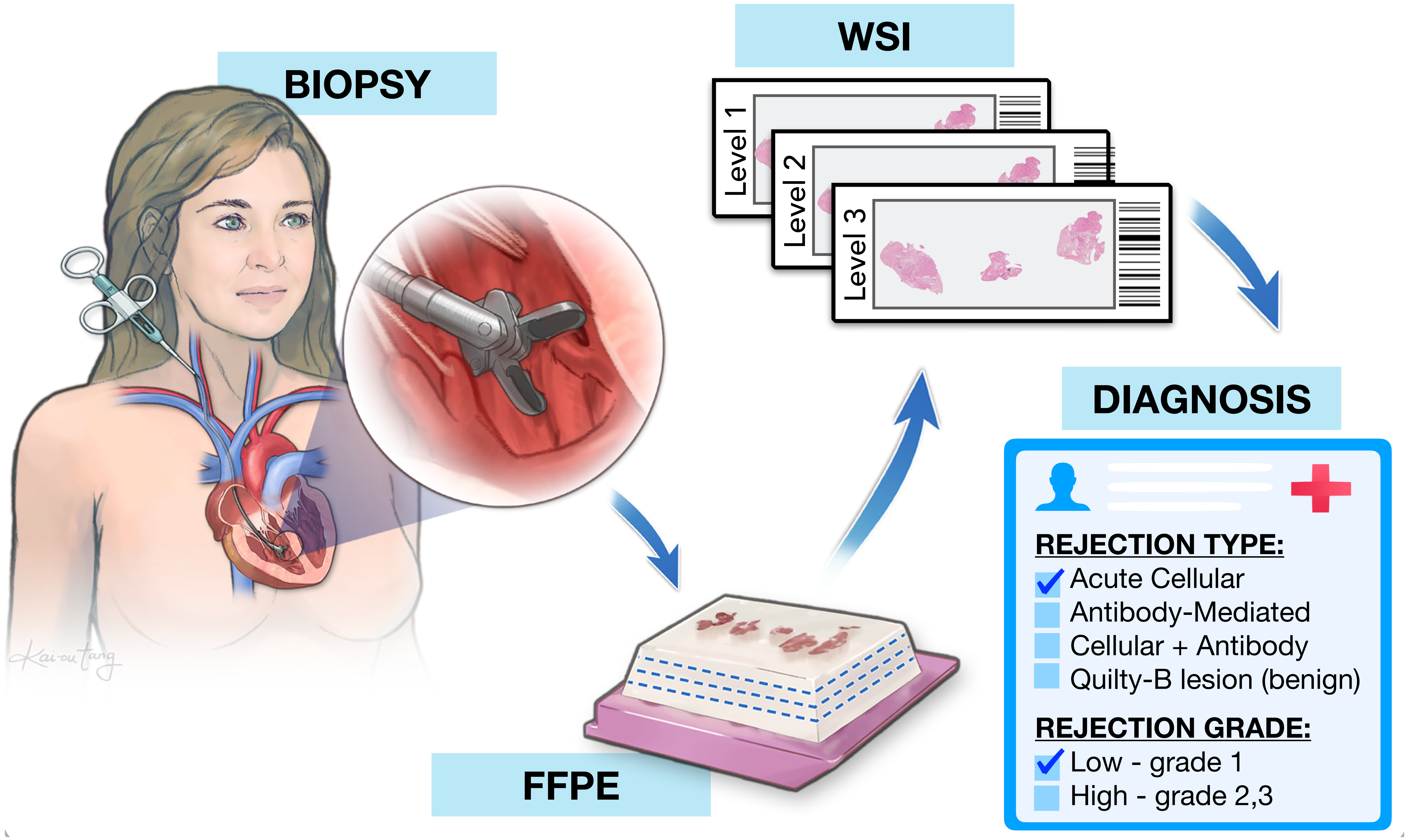
Immune Response
- Immunosuppressives
- Patient-specific set-up

Allograft Rejection
- Main complication & main cause of death
- 40% recipients

MOTIVATION

APPLICATION:

- Early stages of rejections are **asymptomatic** → surveillance **Endomyocardial biopsy** (EMB)
- Gold-standard: manual assessment H&E-stained biopsies:
  - **detection** and **subtyping** of rejections (***acute cellular***, ***antibody-mediate***, ***benign mimickers***) and **grading** (I-III)
- Rejection type & grade **determines the immunosuppressive treatment regime**

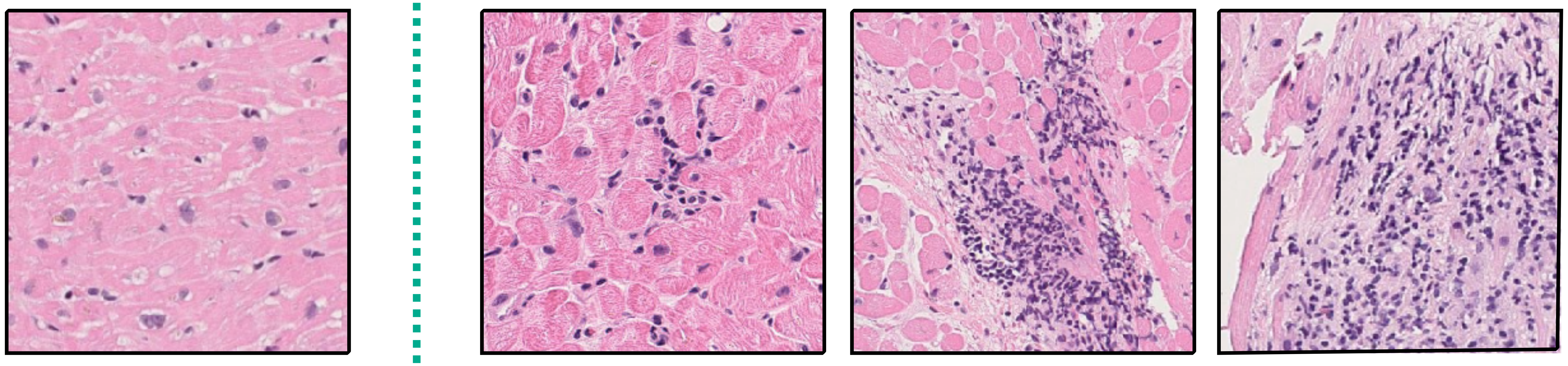


- [1] Concordance among pathologists in the second cardiac allograft rejection gene expression observational study (CARGO II) In: *Transplantation* 94.11 (2012), pp. 1172–1177

Lipkova et al. Nature Medicine (2022)

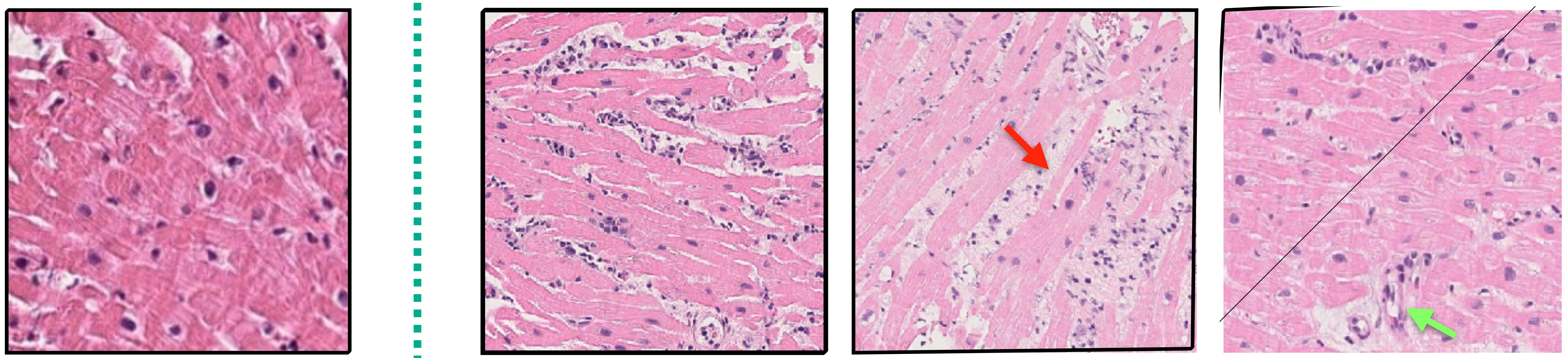
101: Rejection Types

Normal tissue | **Abnormal tissue**



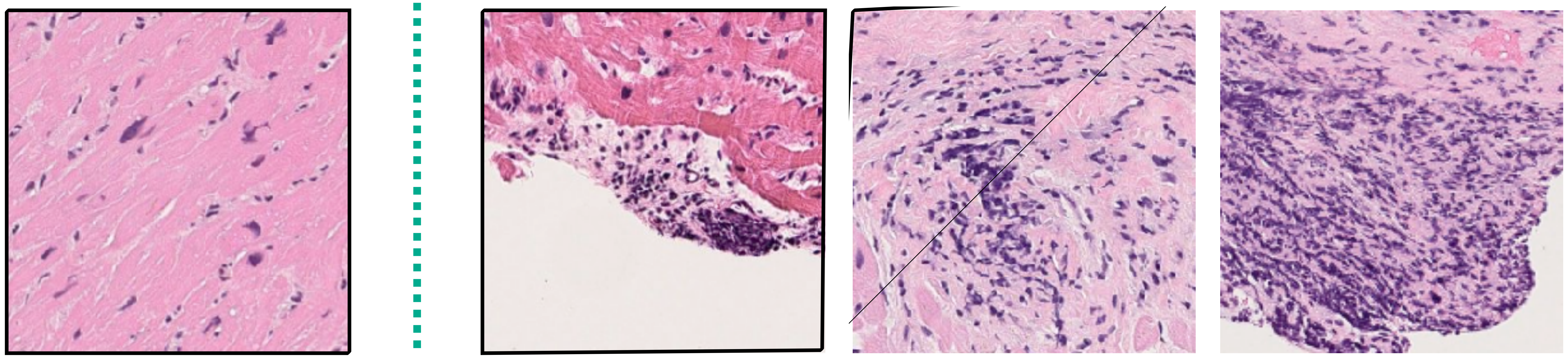
Acute Cellular

- Lymphocyte infiltrates in muscle tissue
- Homogenous structure
- Comprised of T-cells



Antibody Mediated

- Increased extracellular space + **edema**
- **Capillary** endothelial **changes**
- Increase cell damage
- More macrophages and necrosis

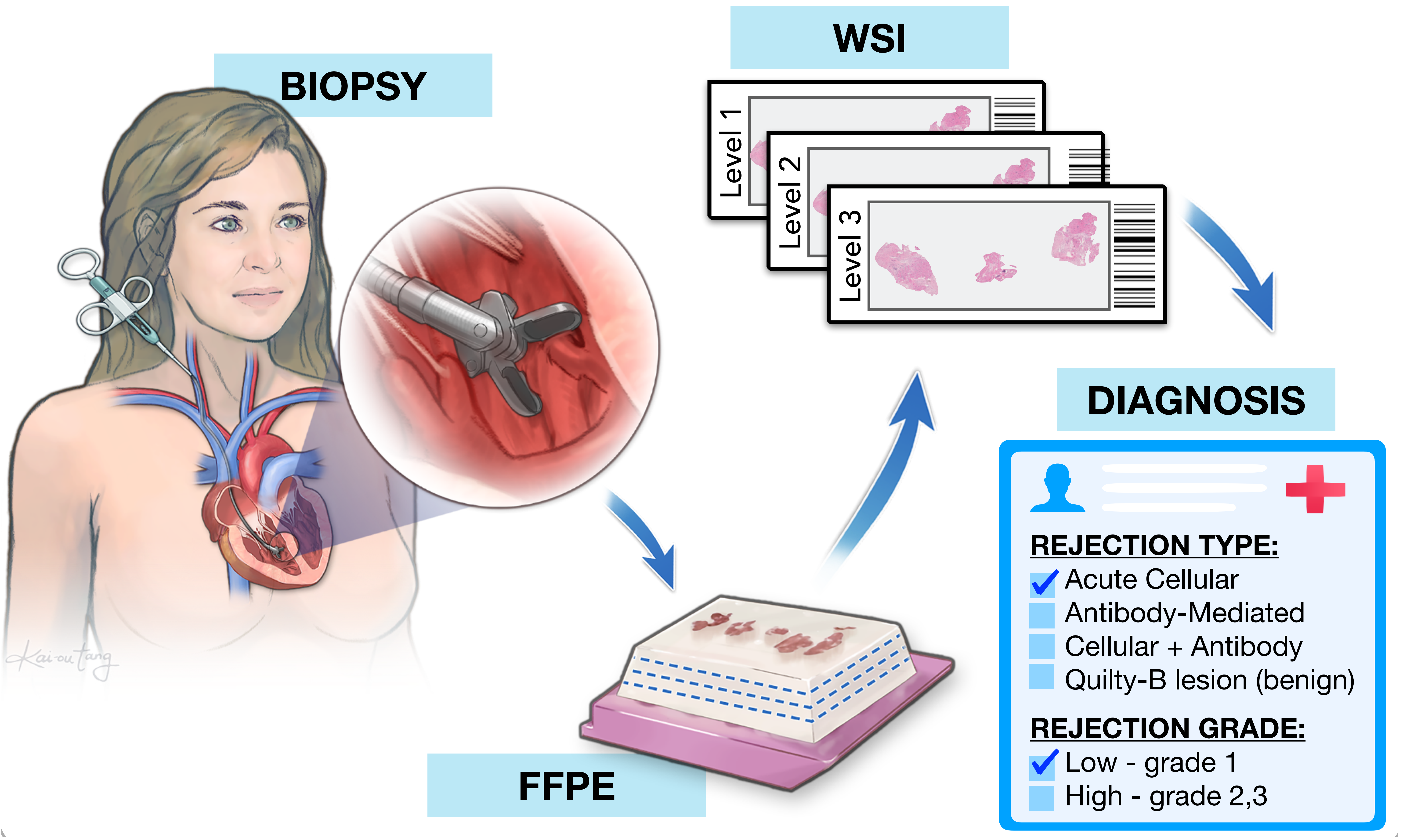


Quilty B Lesions

- Benign lesions
- Mixed B and T-cells, macrophages and plasma cells
- commonly mistaken for cellular rejections

MOTIVATION

- Early stages of rejections are **asymptomatic** → surveillance **Endomyocardial biopsy** (EMB)
- Gold-standard: manual assessment H&E-stained biopsies:
  - **detection** and **subtyping** of rejections (***acute cellular***, ***antibody-mediate***, ***benign mimickers***) and **grading** (I-III)
- Rejection type & grade determines the immunosuppressive treatment regime



CHALLENGES:

- **Substantial inter-rater variability** [1]:
  - <71 % agree if recipient is rejecting the heart
  - <28 % agree on the grade of advance rejections
  - 19 % unable to reach majority agreement
- **Misinterpretation**:
  - under/over treatment with immunosuppressives
  - unnecessary follow-up biopsies
  - worse outcomes

AIM:

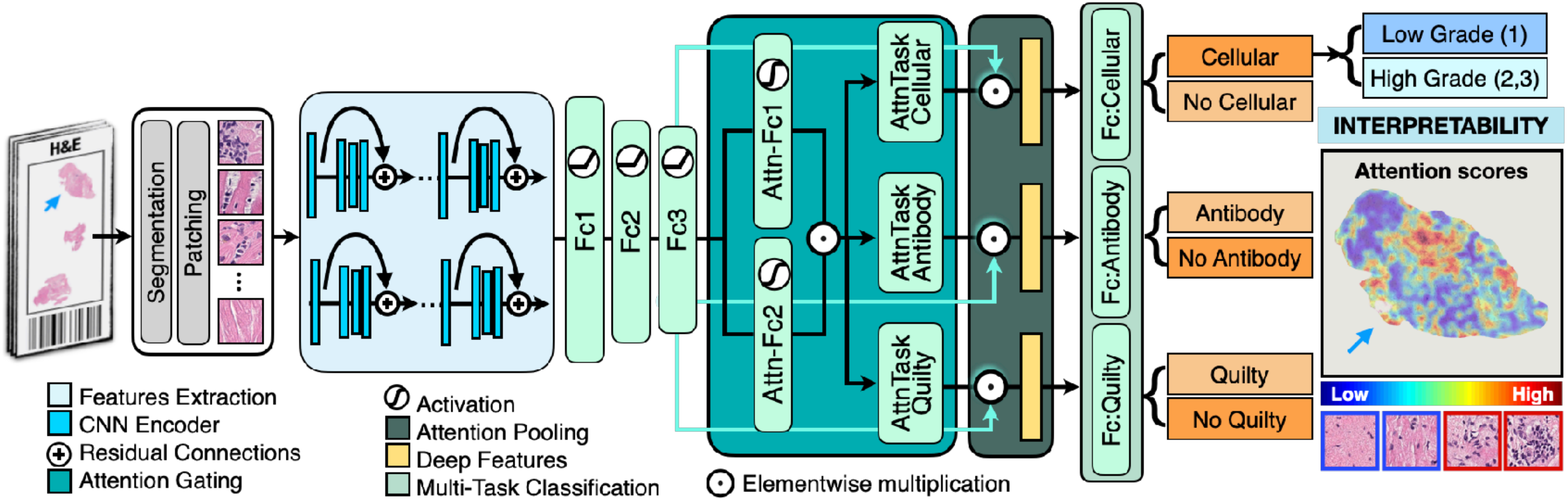
- **Objective and automated EMBs assessment**

[1] Concordance among pathologists in the second cardiac allograft rejection gene expression observational study (CARGO II) In: *Transplantation* 94.11 (2012), pp. 1172–1177

Cardiac Rejection Assessment Neural Estimator

CRANE

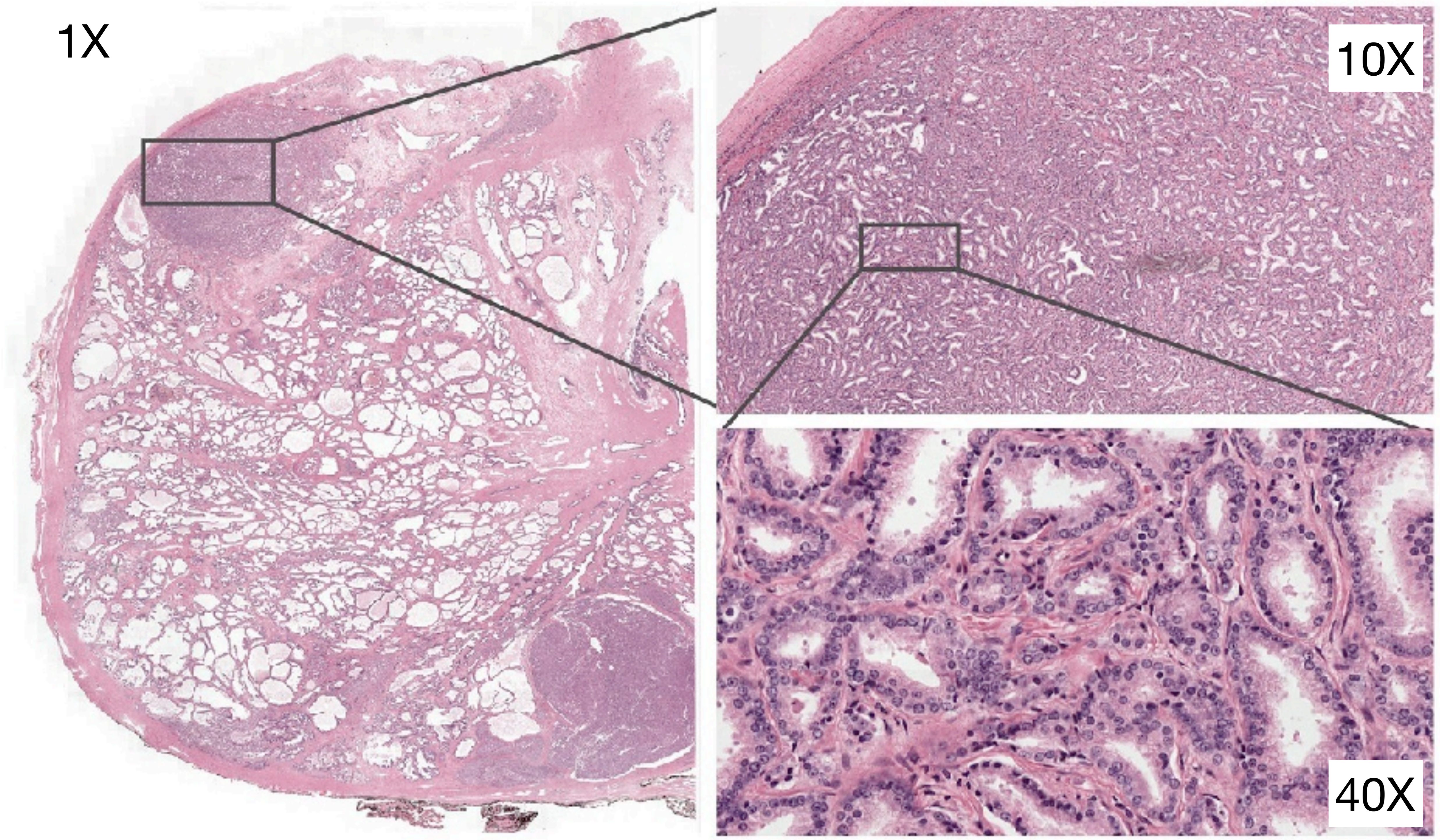
- **Input:** H&E-stained EMBs whole-slide-images (WSIs)
- **Multi-task, multi-label model:** simultaneously identifies **presence** and **type of the rejection** (cellular, antibody, and/or quilty lesions). Separate classifier estimate **rejection grade**
- **Multiple-instance learning:** use **patient diagnosis** as only **label**
  - (avoid pixel-level annotations, supports large-scale deployment)
- **Attention scores,** reflecting relevance of each biopsy region, enable **visual interpretation** of the model's predictions

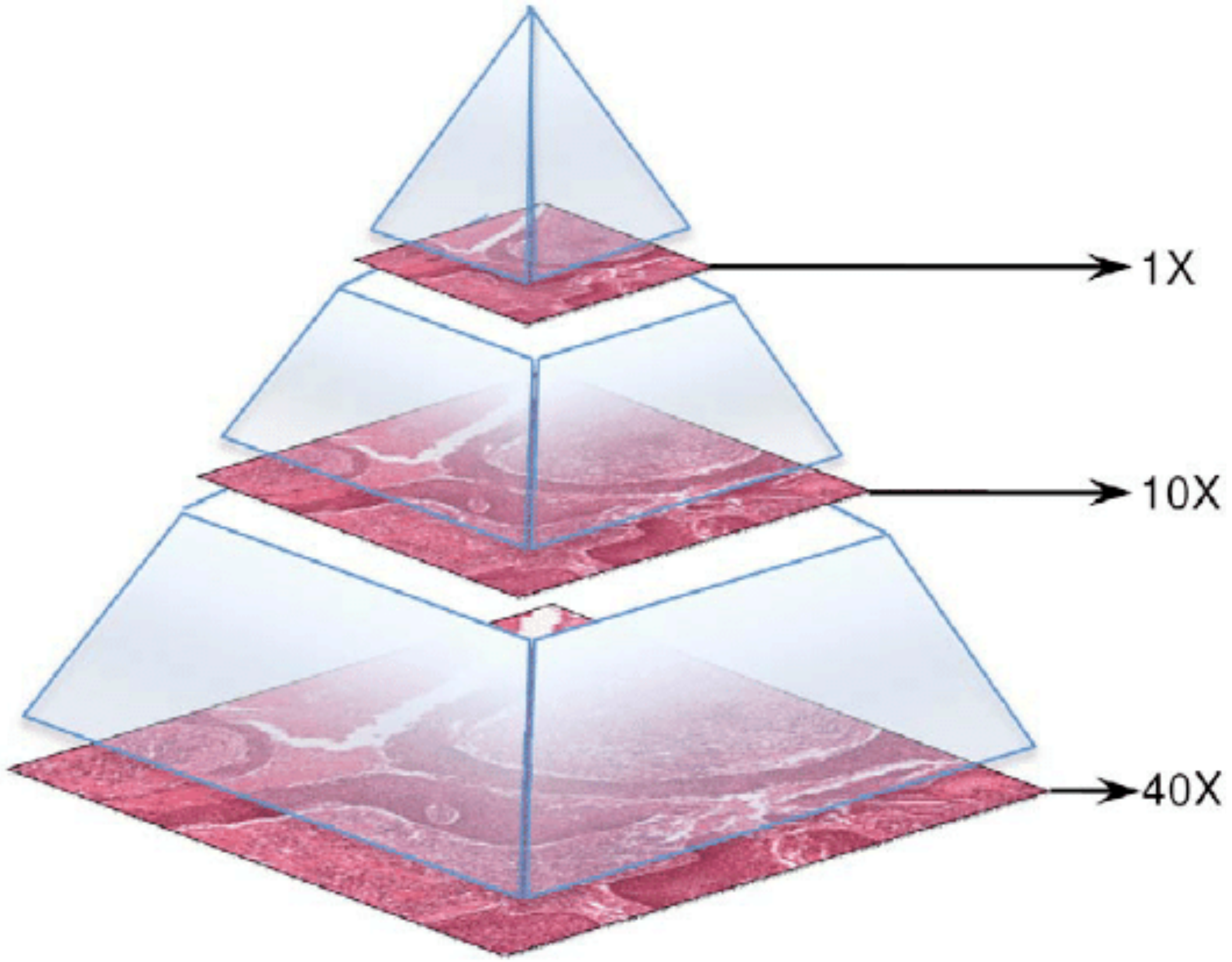


Lipkova et al. Nature Medicine (2022)

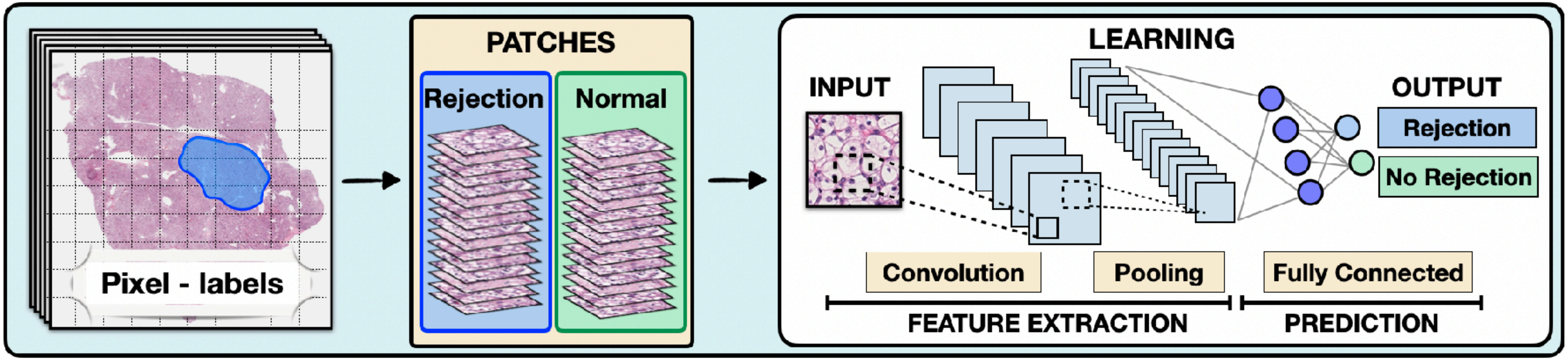
Digital Pathology: Whole Slide Images (WSIs)

- High resolution scan of an entire tissue section (0.25 - 0.5 microns per pixel)
- 1 WSI ~ 1 billion pixels !!!
- **100 WSI has more pixels than <u>whole</u> ImageNet**
- Difficult to train AI directly on WSI





Typical Deep Learning for Pathology

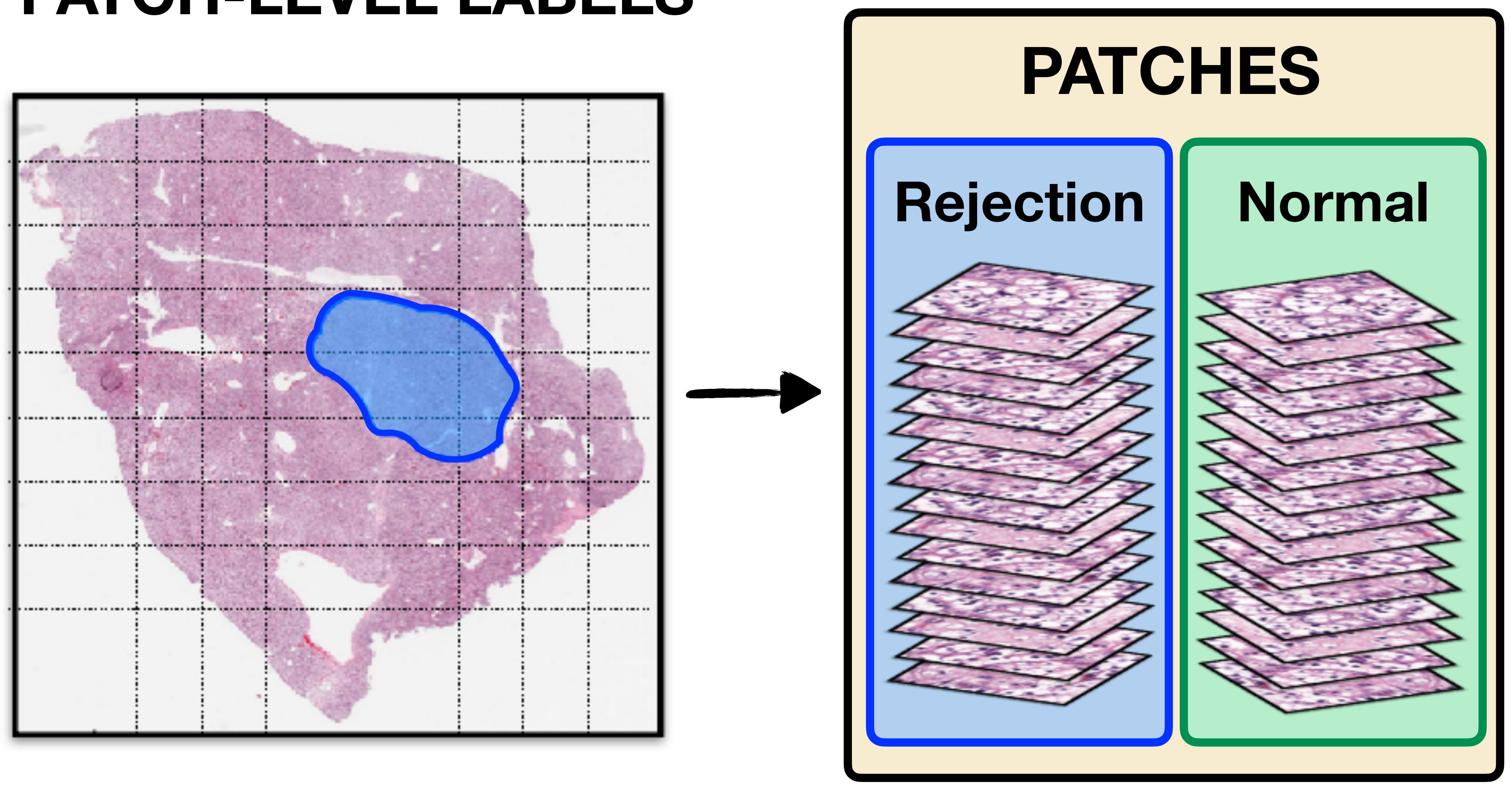


- **Laborious** and **time consuming** to annotate gigapixels large histology images
- Disease borders not always well defined → **inter-rater variability** → **bias**
- **Predictive regions** for some tasks (e.g. treatment response) **might be unknown**
- **Possible data imbalance:** small proportion of image contain the disease (needle-in-haystack problem)
- **Image annotation is not part of standard clinical practice**

Strong vs Weak Supervision

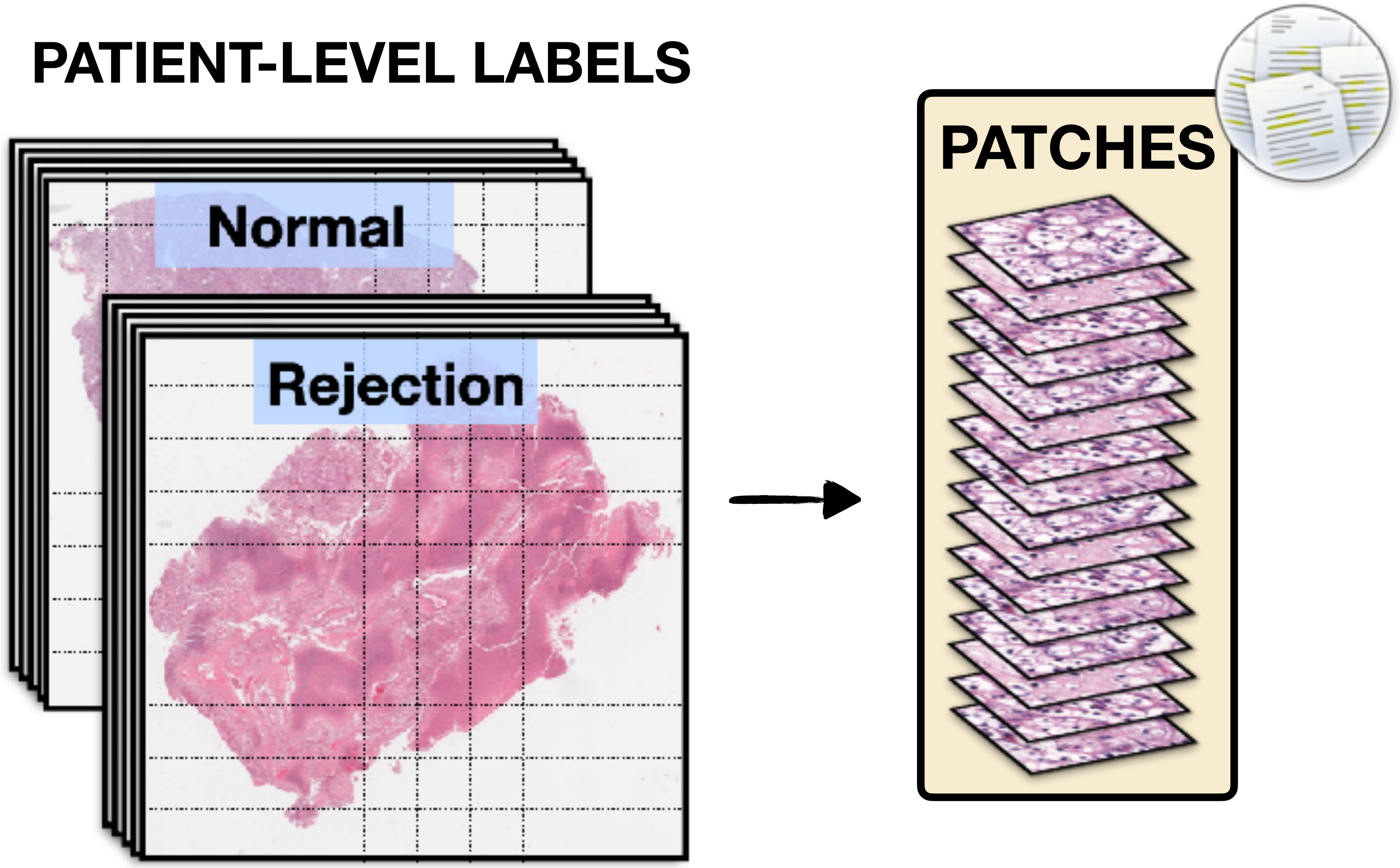
STRONG LABELS

PATCH-LEVEL LABELS



WEAK LABELS

PATIENT-LEVEL LABELS



➡ Model alone must discover which tissue regions and which features are predictive for rejections.

Analogy with Natural Images

STRONG LABELS

- Label for each input

Muffin

Chihuahua

WEAK LABELS

- Label for bag of inputs

Contains Chihuahua

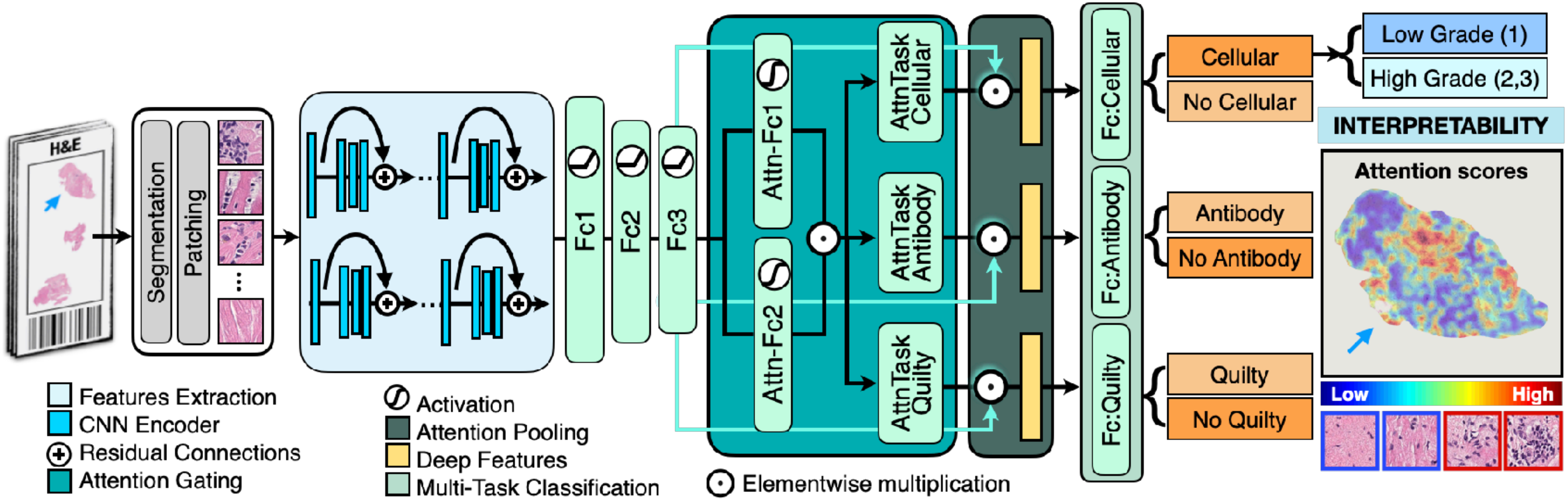
No Chihuahua

➡ The model alone has to discover which image items and features correspond to chihuahua

Cardiac Rejection Assessment Neural Estimator

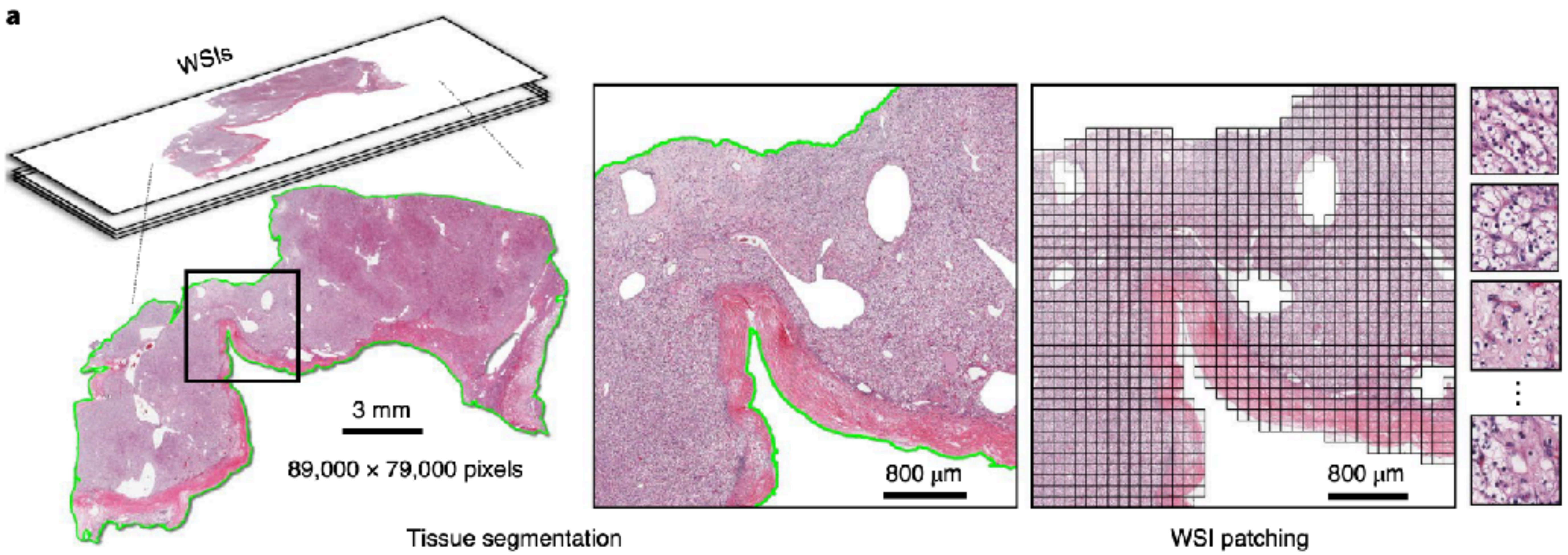
CRANE

- **Input:** H&E-stained EMBs whole-slide-images (WSIs)
- **Multi-task, multi-label model:** simultaneously identifies **presence** and **type of the rejection** (cellular, antibody, and/or quilty lesions). Separate classifier estimate **rejection grade**
- **Multiple-instance learning:** use **patient diagnosis** as only **label**
  - (avoid pixel-level annotations, supports large-scale deployment)
- **Attention scores,** reflecting relevance of each biopsy region, enable **visual interpretation** of the model's predictions

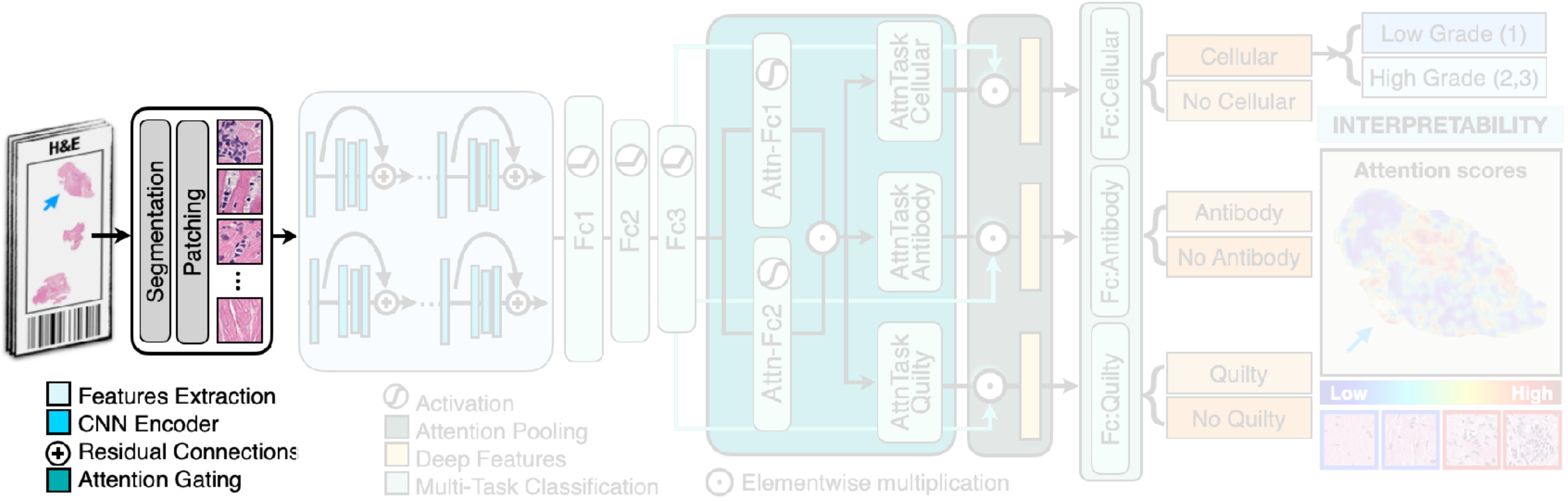


Lipkova et al. Nature Medicine (2022)

PREPROCESSING



≈ 1 Billion Pixels!

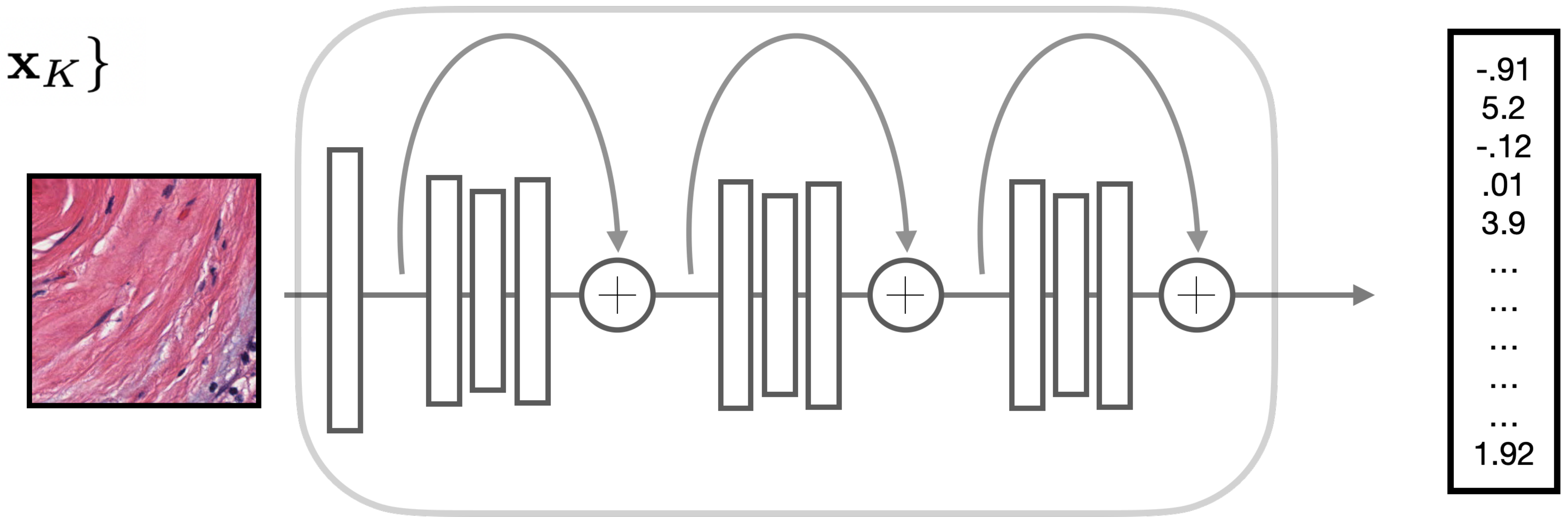


EMBEDDINGS

- Patch-level representation of patch k from {1,...,K}

$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\}$

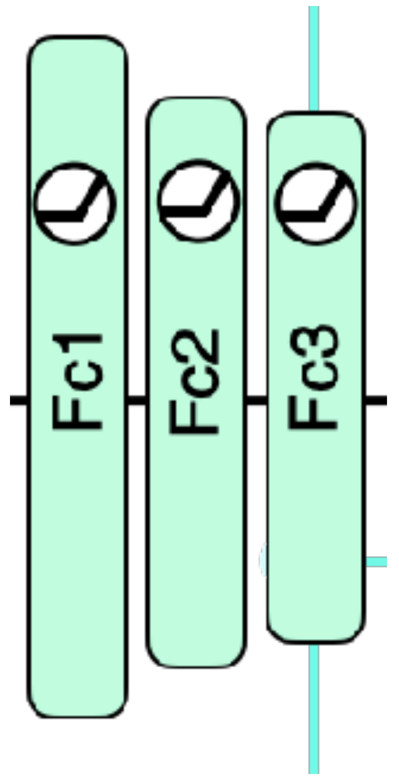
Input x_k :
$256 \times 256 \times 3$



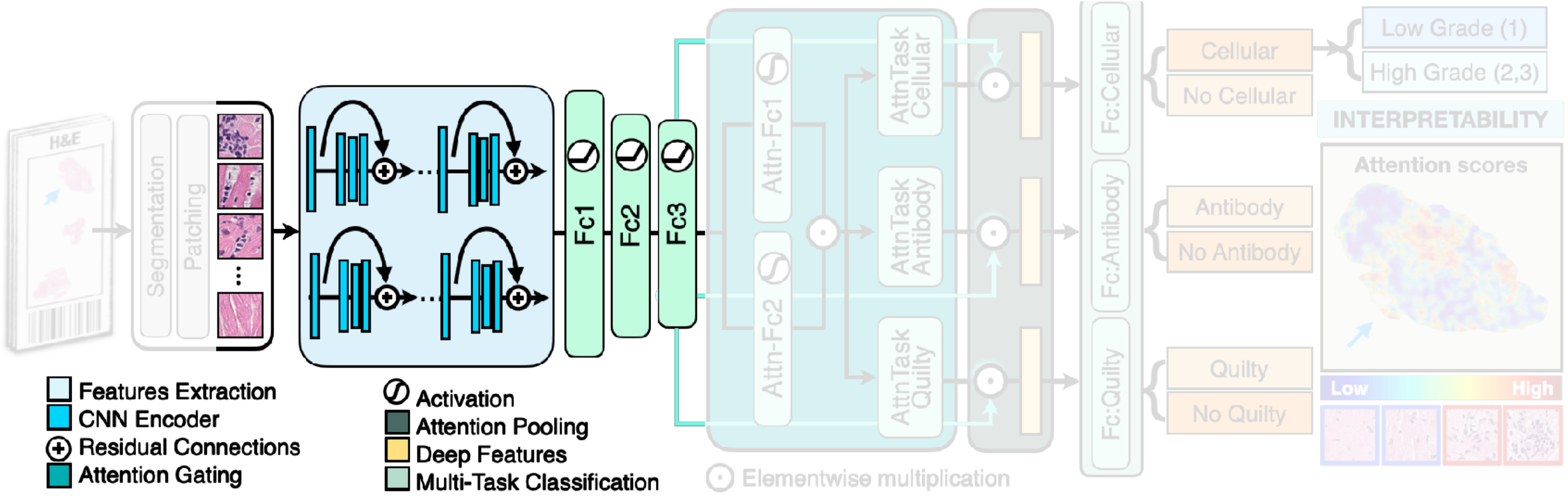
Pretrained Encoder

$f(\cdot;\theta) : \mathbb{R}^{C \times H \times W} \rightarrow \mathbb{R}^{1024}$

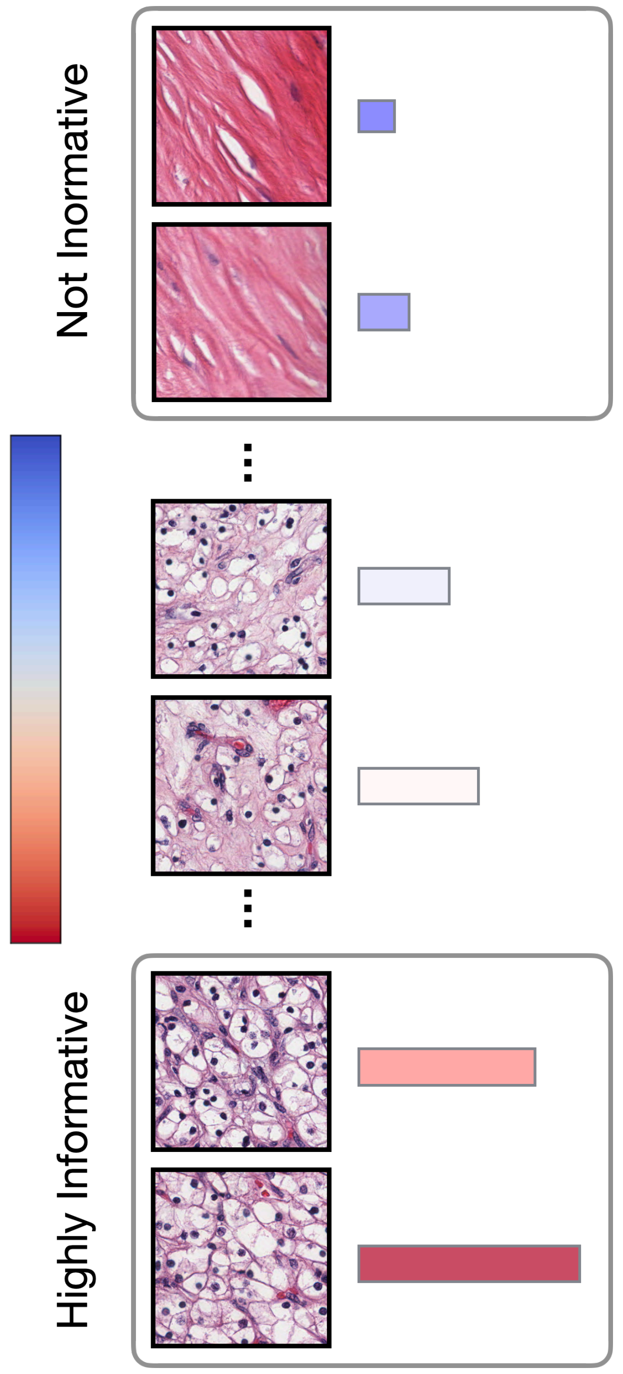
Embedding z_k:
1024



- ResNET50 features $\mathbf{z}_k \in \mathbb{R}^{1024}$
- Three FC layers:
  - FC1: $\mathbf{W}_1 \in \mathbb{R}^{768 \times 1024}$
  - FC2: $\mathbf{W}_2 \in \mathbb{R}^{512 \times 768}$
  - FC3: $\mathbf{W}_3 \in \mathbb{R}^{512 \times 512}$



ATTENTION LEARNING

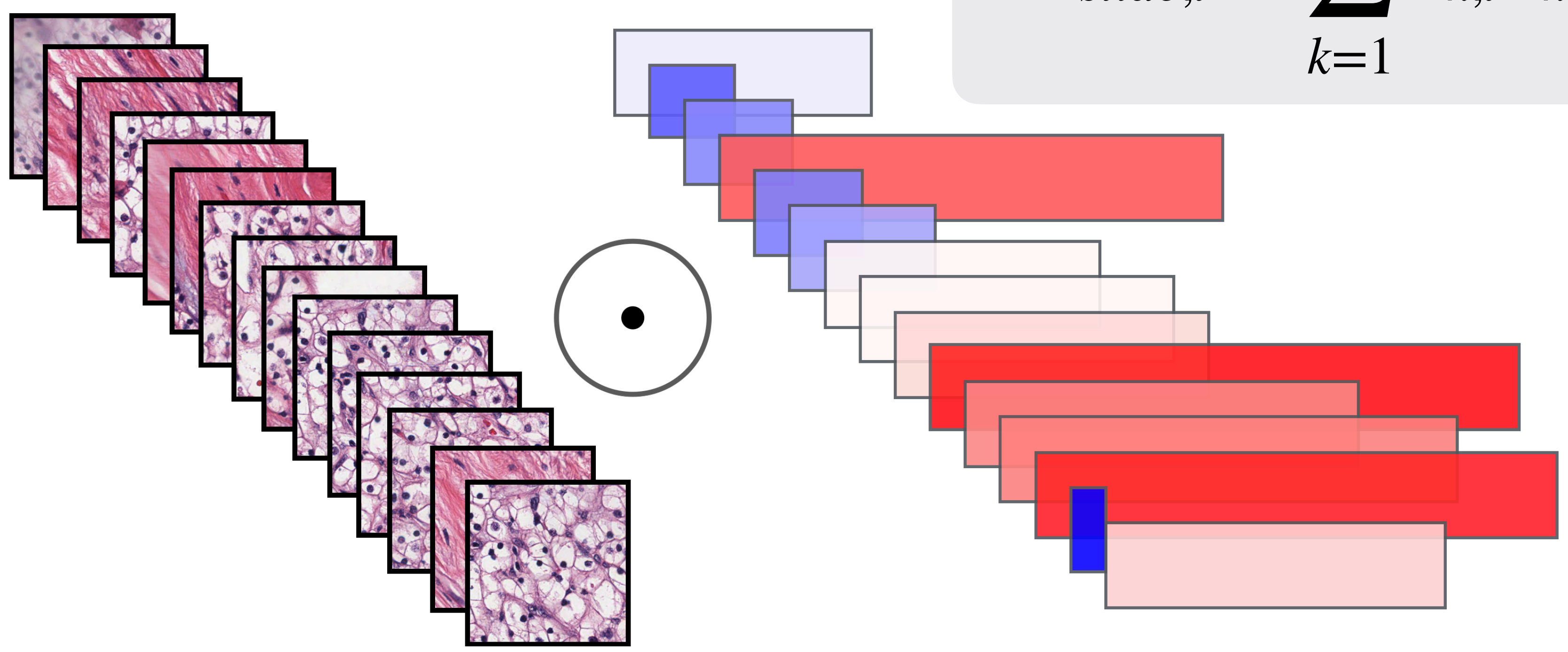


Attention score *(for patch k and task t):*

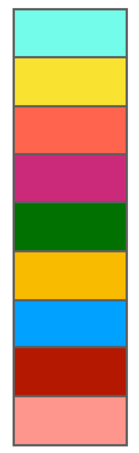
$$a_{k,t} = \frac{\exp\left\{\mathbf{W}_{a,t}\left(\tanh\left(\mathbf{V}_a\mathbf{h}_k\right)\odot \mathrm{sigm}\left(\mathbf{U}_a\mathbf{h}_k\right)\right)\right\}}{\sum_{j=1}^{N}\exp\left\{\mathbf{W}_{a,t}\left(\tanh\left(\mathbf{V}_a\mathbf{h}_j\right)\odot \mathrm{sigm}\left(\mathbf{U}_a\mathbf{h}_j\right)\right)\right\}}$$

Attention-based pooling

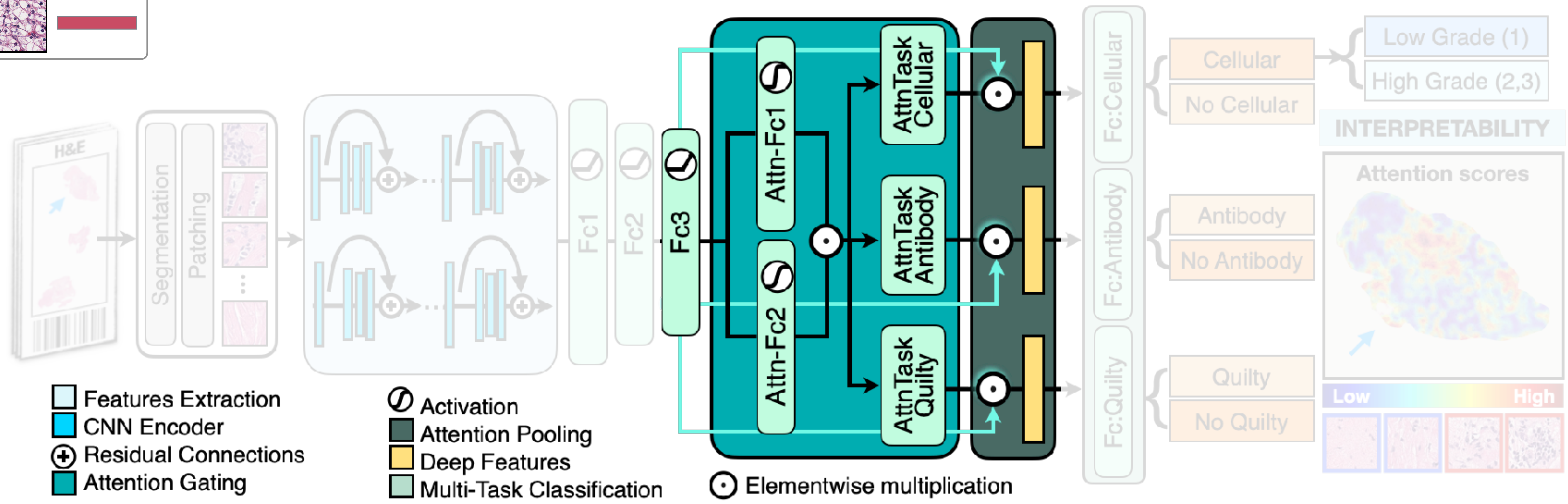
$$\mathbf{h}_{slide,t} = \sum_{k=1}^{K} a_{k,t}\mathbf{h}_k$$



Learned WSI Embedding

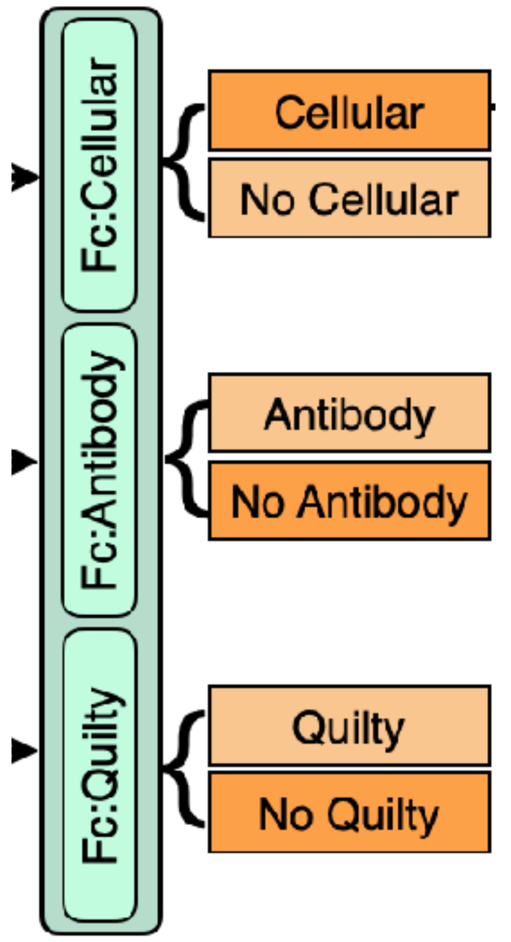


(Ilse et al. ICML 2018)



MULTI-TASK CLASSIFIER

Learned WSI Embedding



- Slide-level representations for task *t*:

$$\mathbf{h}_{slide,t} = \sum_{k=1}^{K} a_{k,t}\mathbf{h}_k$$

- Slide-level predictions for task t:

$$\mathbf{p}_t = \text{Softmax}(\mathbf{W}_{cls,t}\,\mathbf{h}_{slide,t} + \mathbf{b}_{cls,t})$$

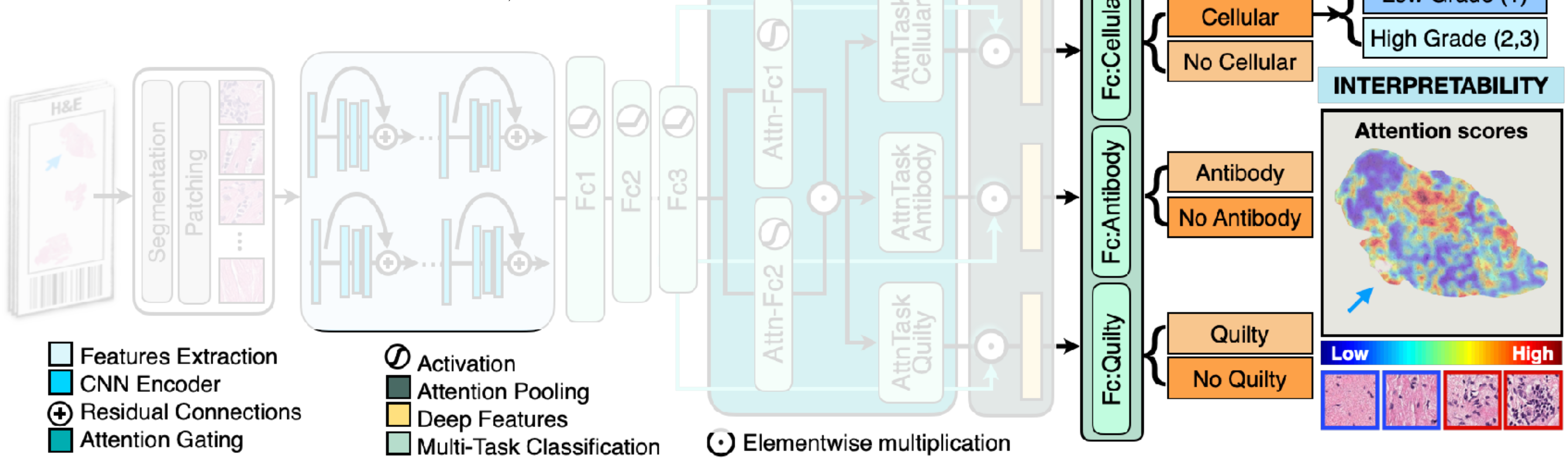
$$\mathbf{W}_{cls,t} \in \mathbb{R}^{2\times512}$$

REJECTION GRADE

- Same MIL model, just single-task

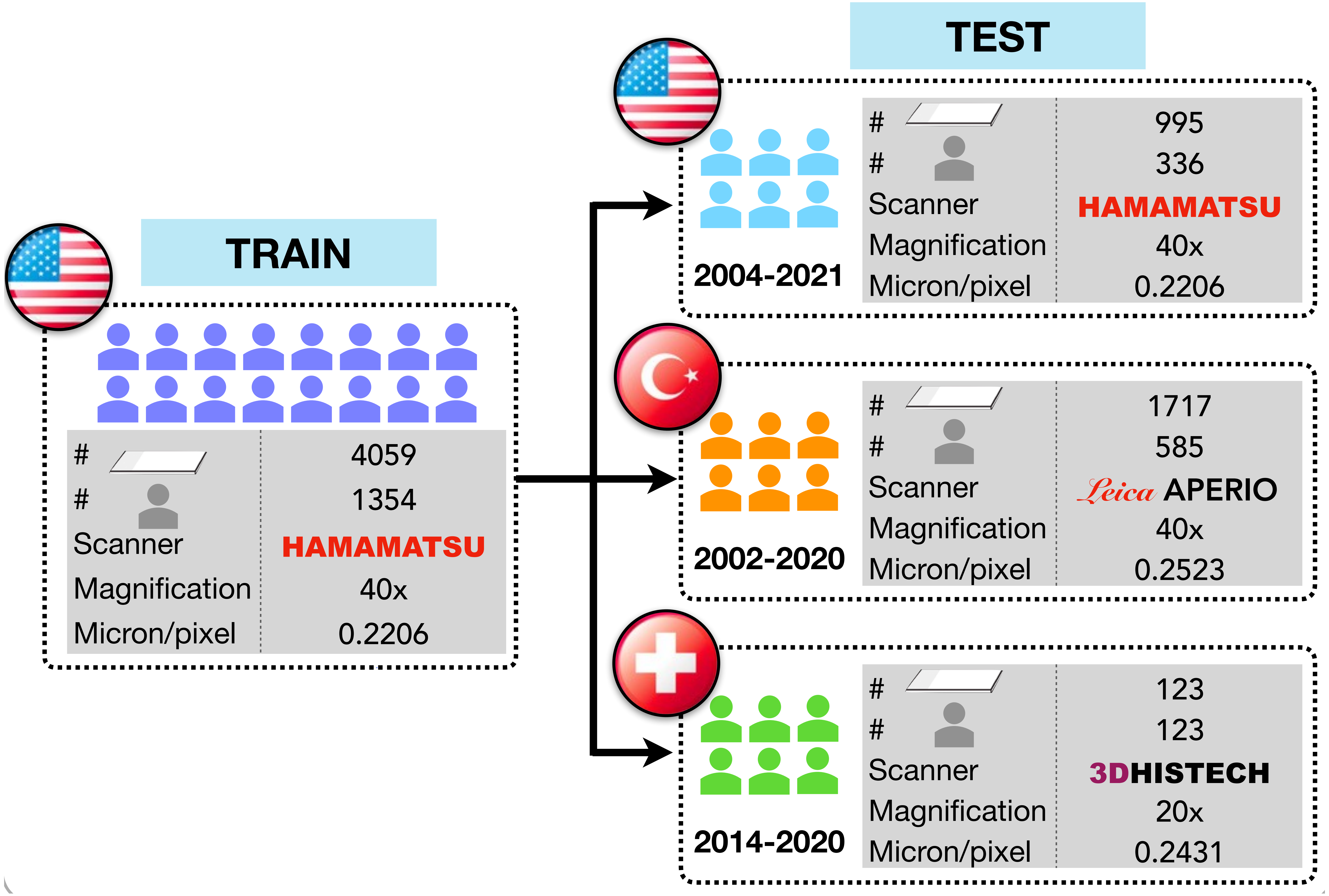
INTERPRETABILITY

- WSI attention heatmaps



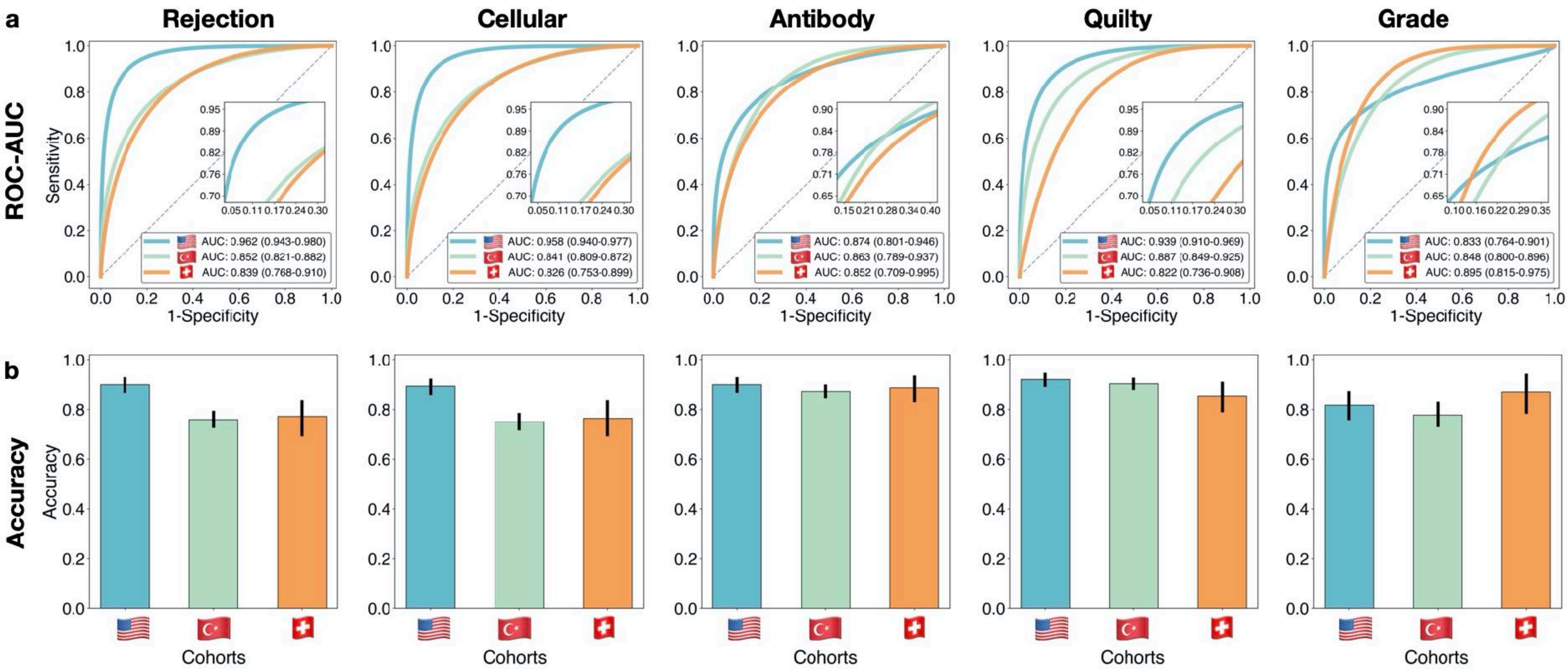
Study Design

- Over 7000 WSI from 3 **independent centers**
- **Large diversity:**
  - *population (geo., pediatric vs adult),*
  - *scanners,*
  - *biopsy protocols,*
  - *staining (manual vs automated),*
  - *noise,*
  - *micron/pixel,*
  - *etc*
- The model is trained on subset of data collected in USA
  - 70/10/20% split (balance diagnosis)
- **Generalization** to external cohorts **without domain-specific adaptations**



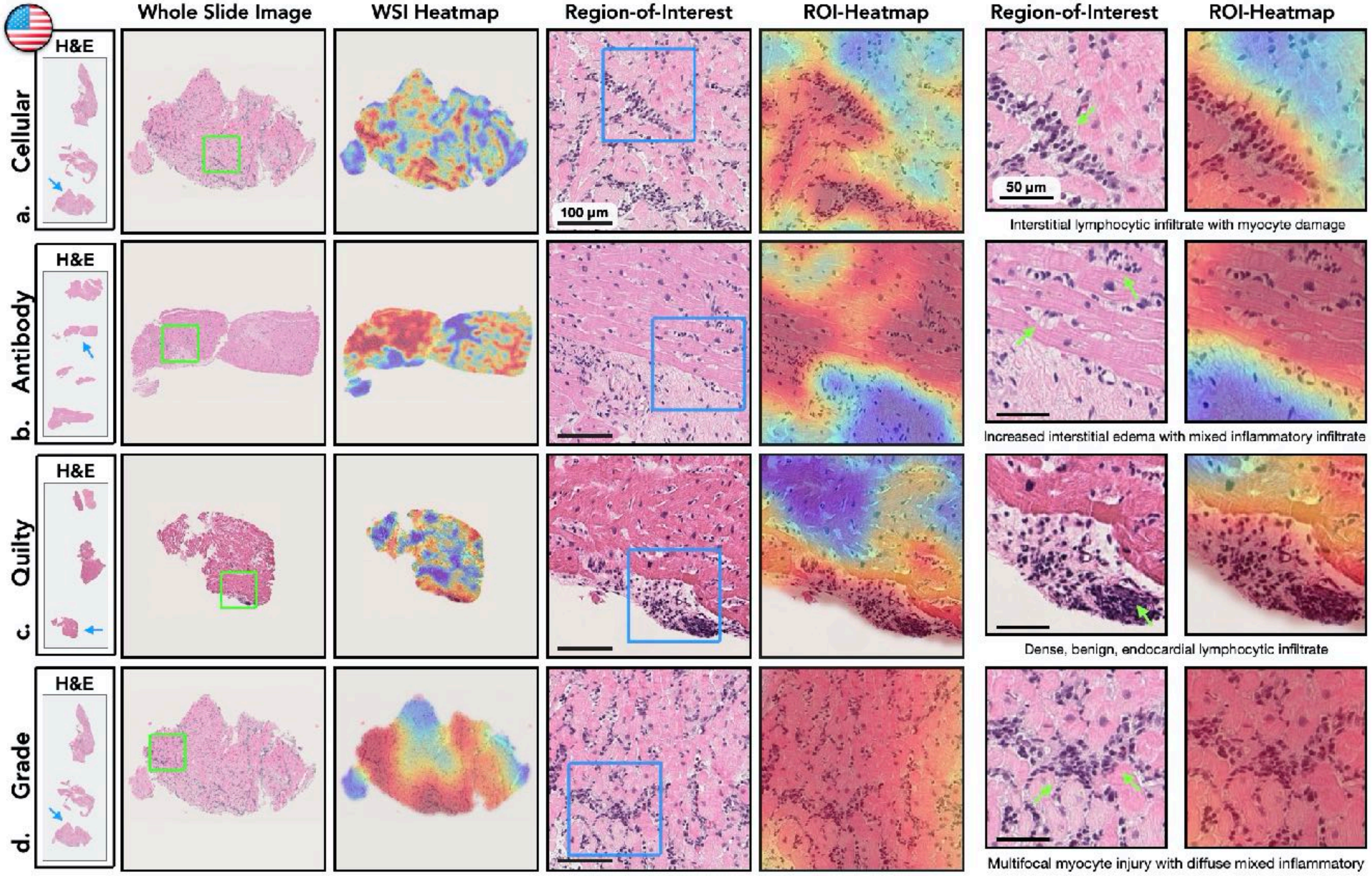
Lipkova et al. Nature Medicine (2022)

Evaluation & Results



Lipkova et al. Nature Medicine (2022)

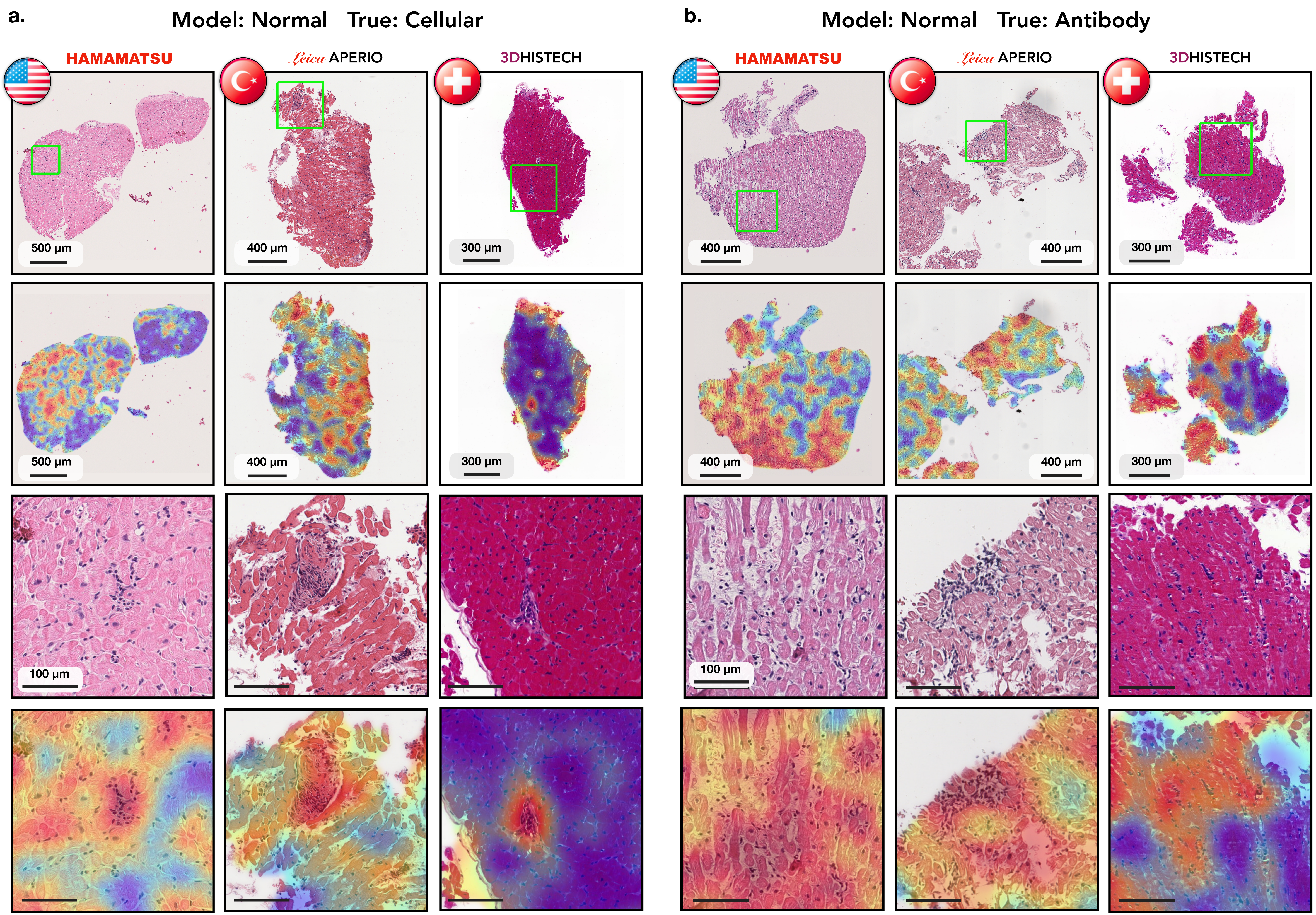
Interpretability



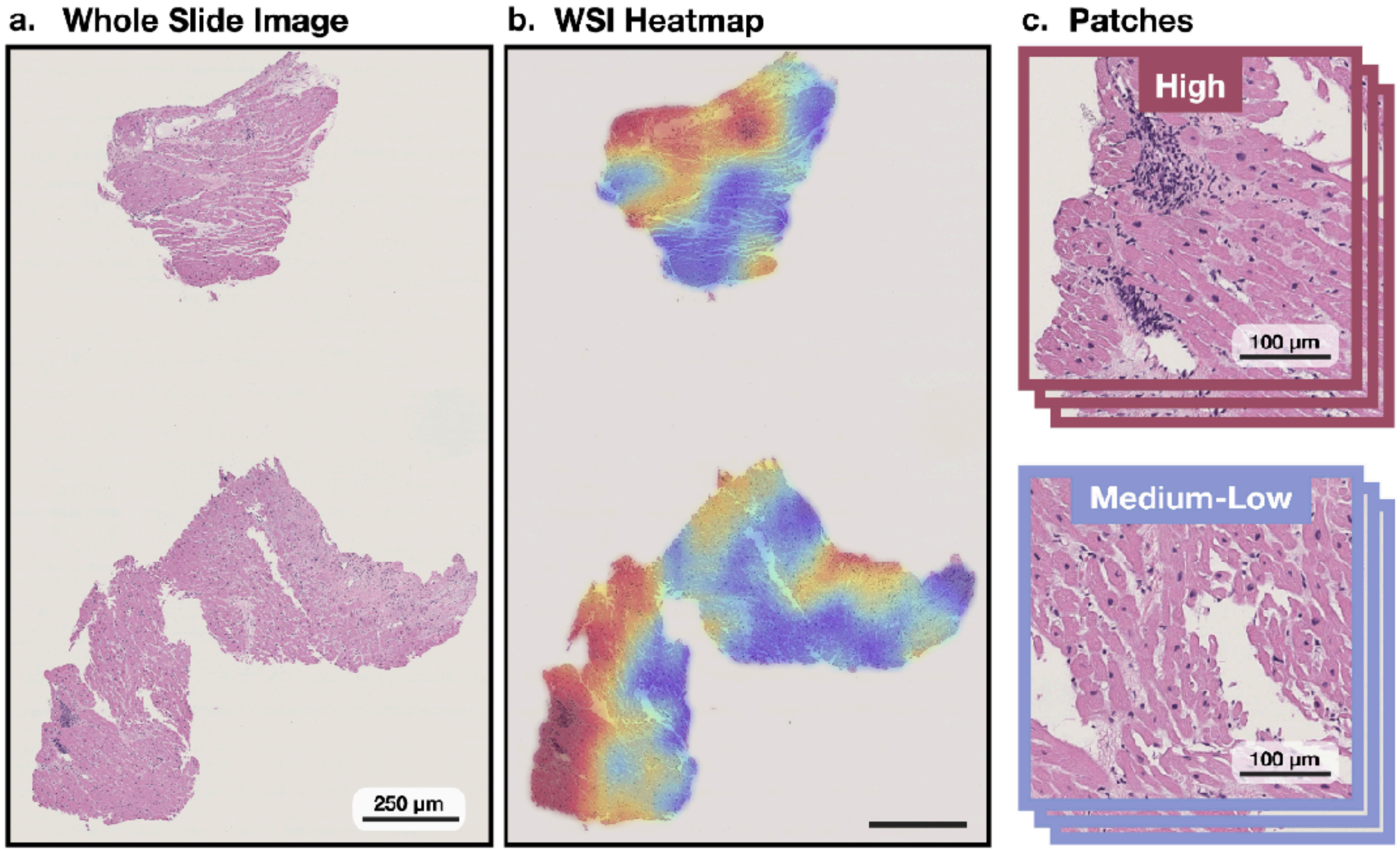
- **High-attention (red)** regions **correspond to rejection morphology used by** pathologist for diagnosis
- **Low-attention (blue)** scores are assigned mostly to **benign tissue**

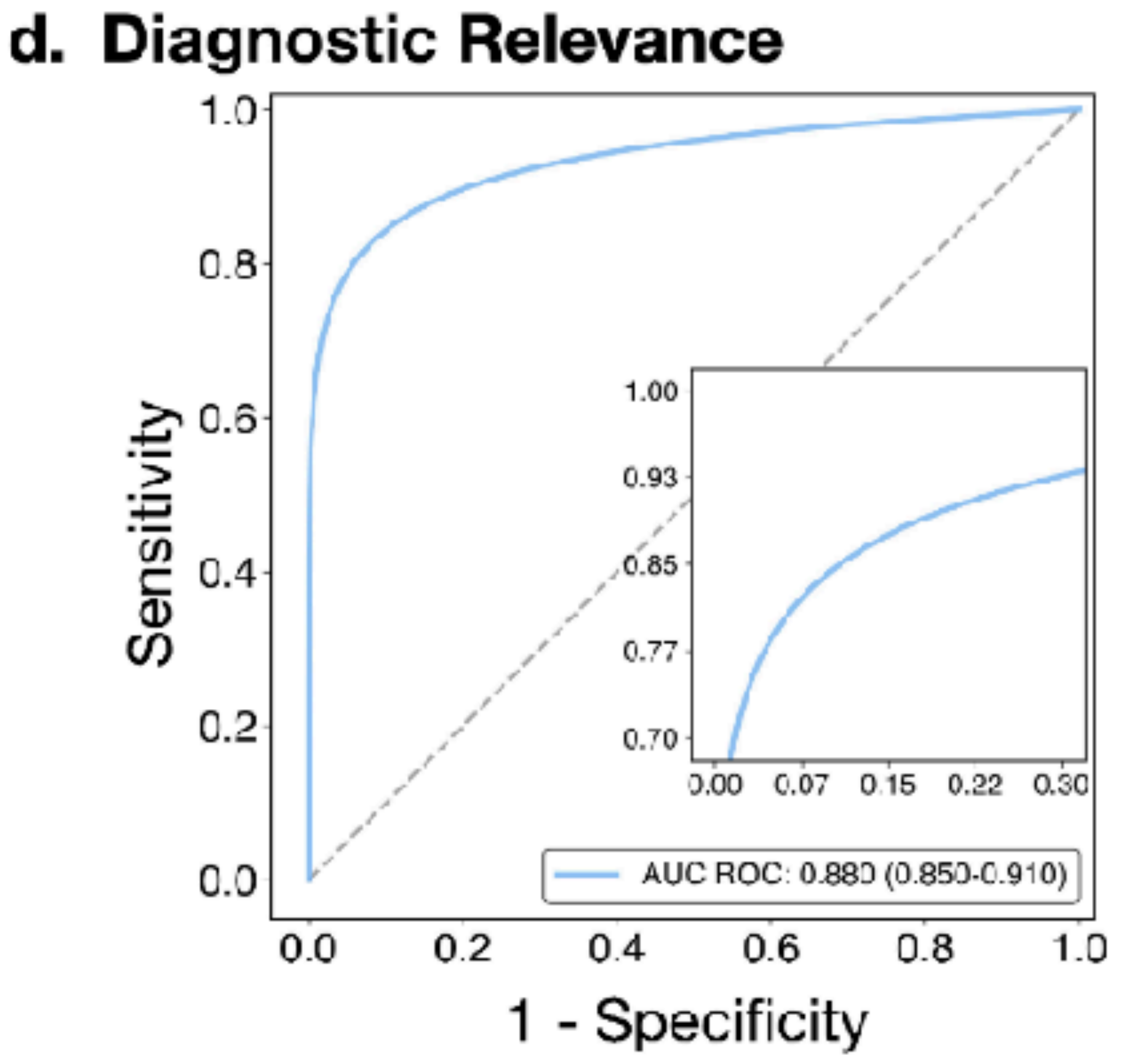
Lipkova et al. Nature Medicine (2022)

Assessment of Failure Cases



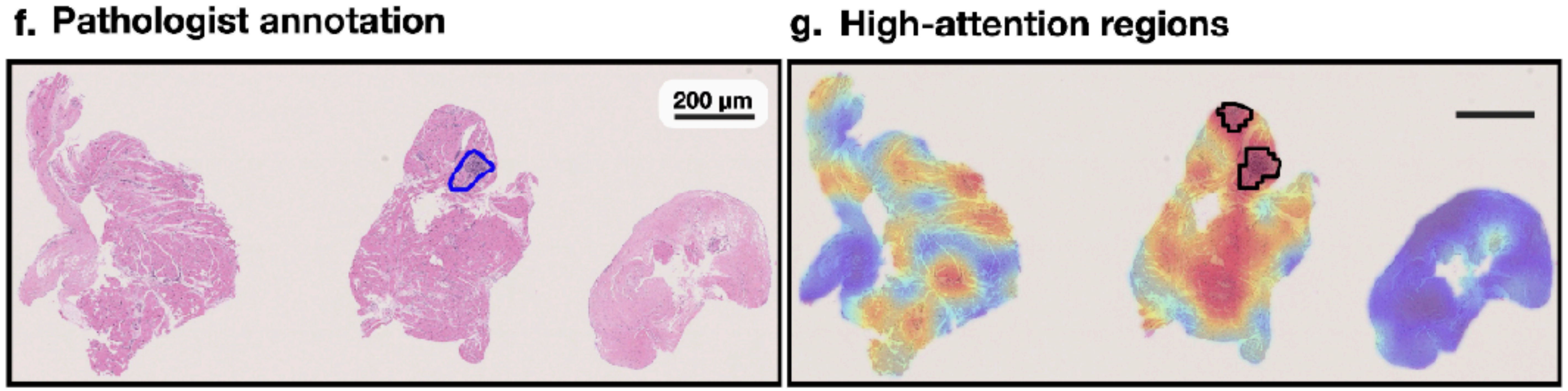
Quantitative Assessment of Interpretability





e. Patch-Level Scores

Tasks:	Accuracy	F1	κ
All	0.873	0.855	0.744
Cellular	0.925	0.914	0.848
Antibody	0.902	0.911	0.802
Quilty	0.809	0.729	0.596



h. Slide-Level Scores

Tasks:	Detection rate
All	0.922
Cellular	0.942
Antibody	0.901
Quilty	0.924

CRANE - Website

crane.mahmoodlab.org

MAIN METADATA

SELECT PATIENT:

Patient1

Select Attention map:

Cellular Antibody

Quilty Grade

Modes

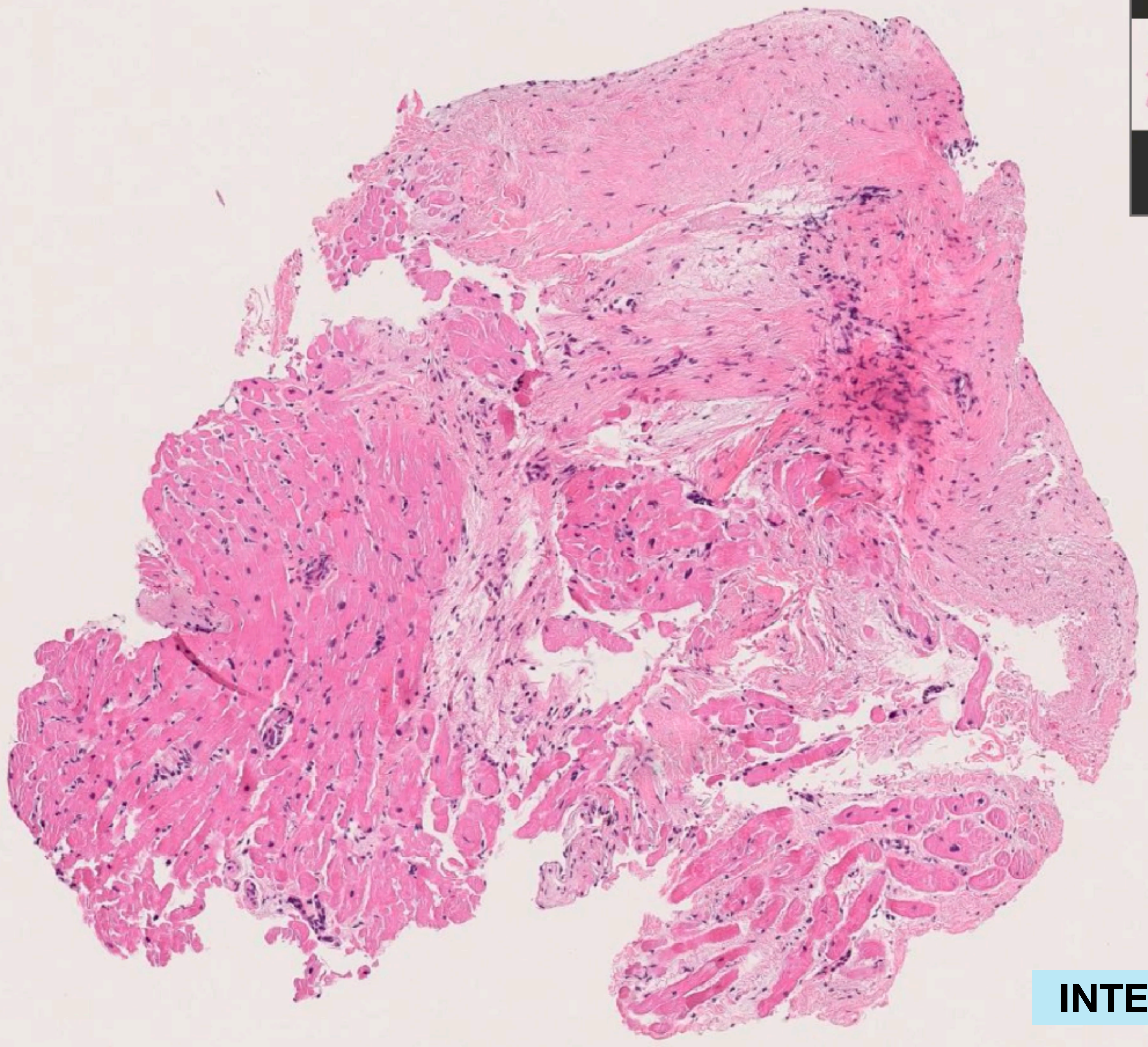
Curtain Side By Side

Overlay

Diagnosis	Confidence
Cellular:	0.9840
Antibody	0.0040
Quilty	0.0141
Grade	0.9186

CRANE

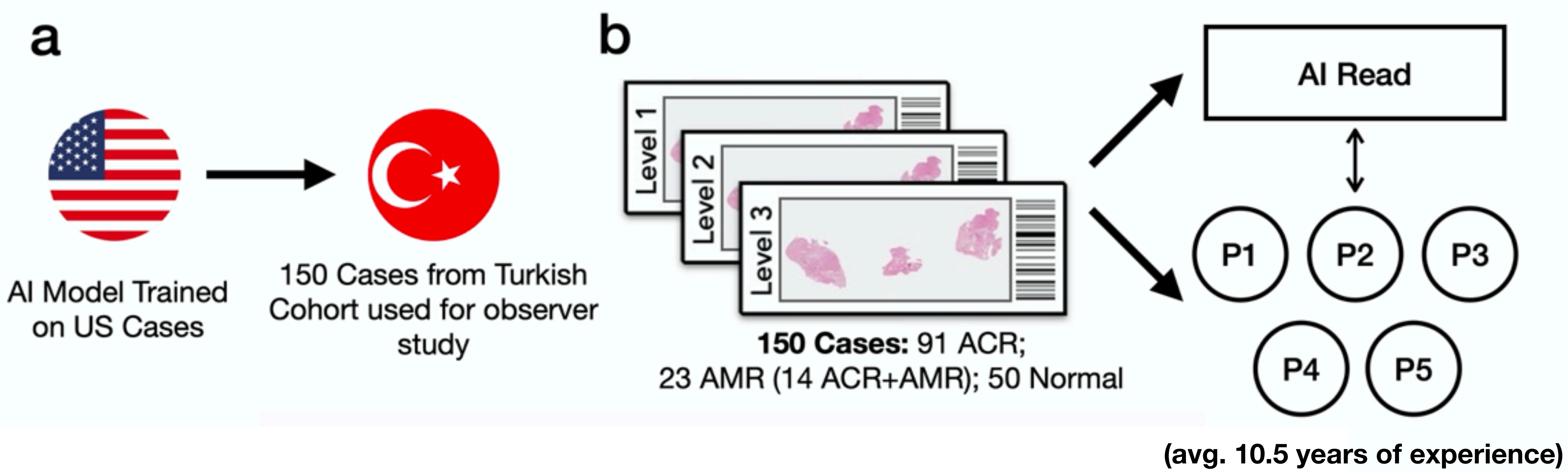
www.mahmoodlab.org
fmahmood@broadinstitute.org

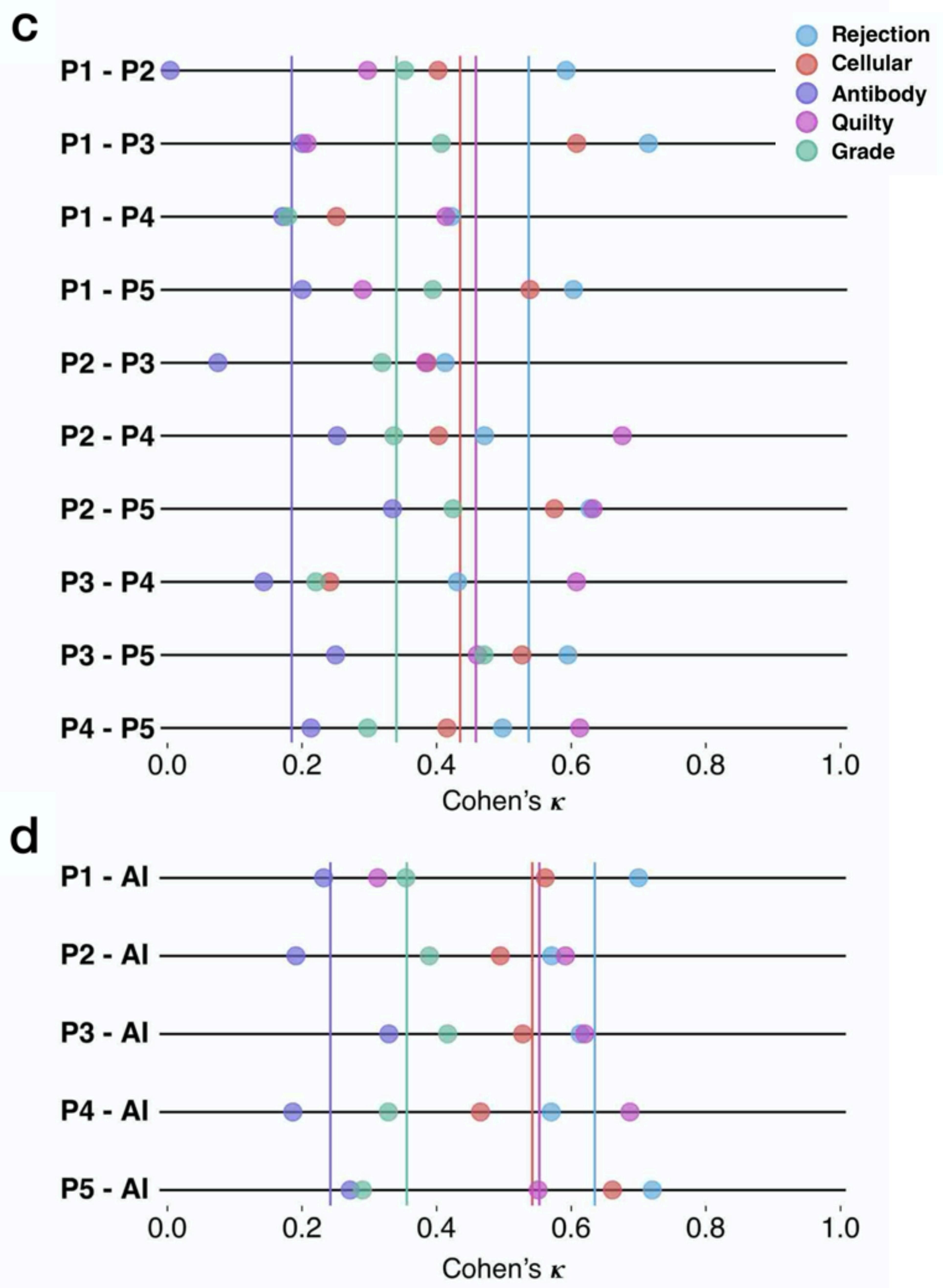


INTERACTIVE DEMO

crane.mahmoodlab.org

Comparison with human readers





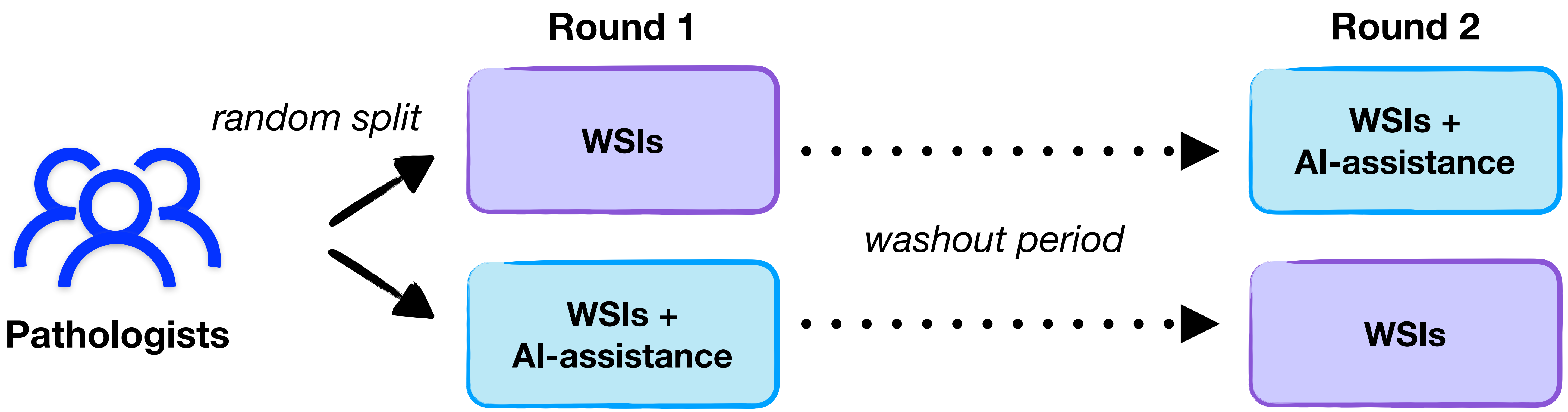
- **Cohens' κ** (-1 to 1): inter-observer agreement:
- Agreement between expert is comparable to previous studies

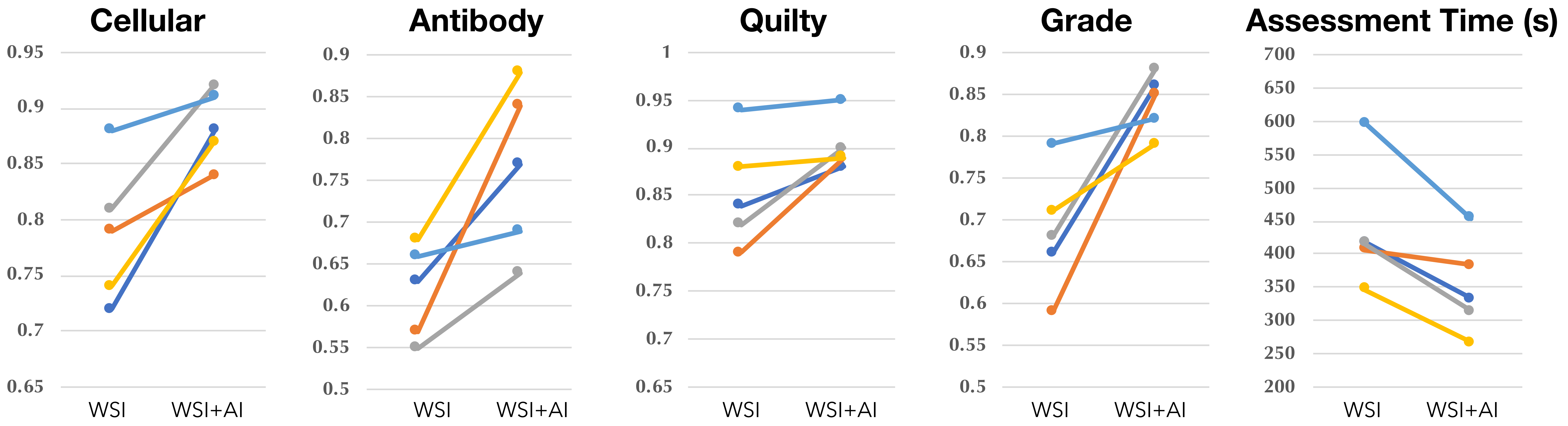
- For all tasks **AI-predictions are <u>not inferior</u> to human experts:**
  - avg. agreement on **rejection** between **pathologists** κ = 0.537 (moderate agreement)
  - avg. agreement between **pathologists and model** κ = 0.639 (substantial agreement)

Clinical Potential

- **Ground-truth labels**: consensus of readers from the first study
- **AI-assistance**: attention heatmaps as semi-transparent layer at the top of H&E slide

- **For all readers:**
  - **Increase accuracy**
  - (i.e. reduce inter-rater variability)
  - **Decrease assessment time**





Lipkova et al. Nature Medicine (2022)

Study Design Flow Chart

The Mahmood Lab

HARVARD MEDICAL SCHOOL

BROAD INSTITUTE

BWH BRIGHAM AND WOMEN'S HOSPITAL

Dana-Farber Cancer Institute

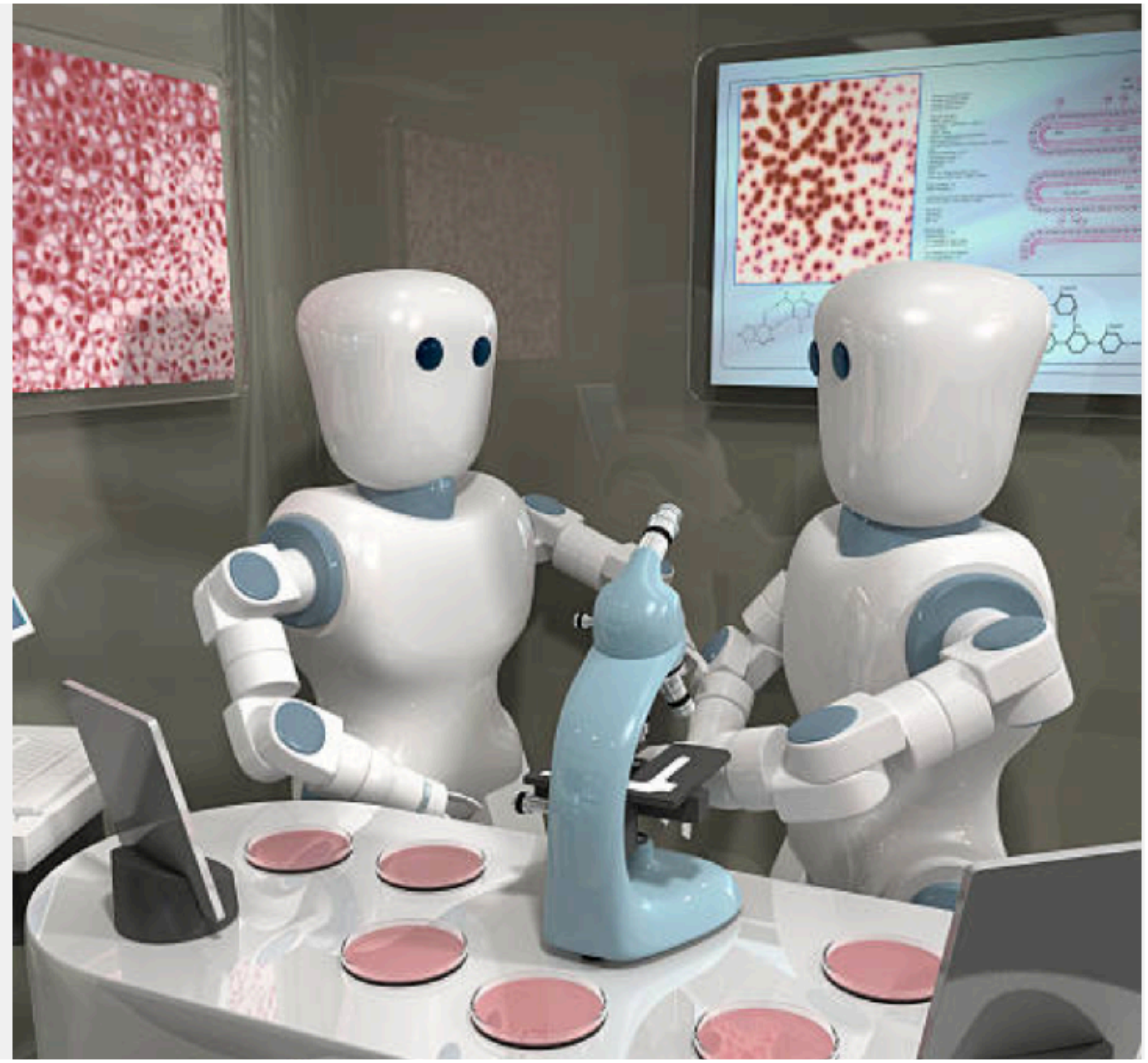
Mahmood Lab
AI for Pathology

OCTOPATH

AI for Pathology

Jana's Lab

SCAN ME



UC Irvine Health

School of Medicine