# Generalizable ML for digital pathology

Geneviève Robin

CNRS, Université Paris Cité

June 2026

# Co-authors at Owkin

Antoine Olivier

Arthur Pignet

Alexandre Filiot


Nicolas Dop

Oussama Tchita

Auriane Riou

Simon Grouard

Lucas Fidon

Céline Thiriez

# Two publications presented at MICCAI 2025

- Distilling foundation models for robust and efficient models in digital pathology (Filiot et al. 2025)
- Robust sensitivity control in digital pathology via tile score distribution matching (Pignet et al. 2025)

# Introduction

# Histopathology: Microscopic imaging of tissue samples

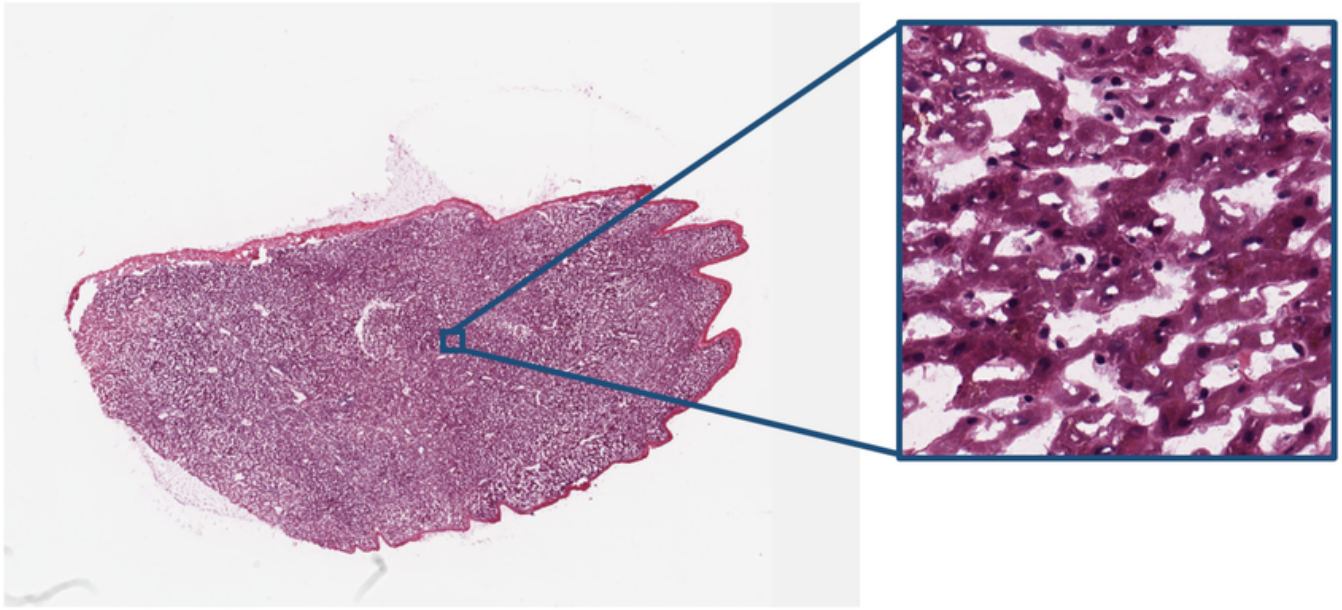


Figure: Example of a H&E whole slide image (WSI) and tile.

- Used routinely in cancer care for tumor detection and subtyping, biomarker quantification, etc.
- Large public data bases of digitized H&E slides/tiles [Filiot et al., 2024]

# Diagnostic tool in colorectal cancer: MSIntuit CRC [Saillard et al., 2023]

- Microsatellite instability (MSI) is a major biomarker in cancer
- Prognosis and eligibility to immunotherapy depend on MSI status in solid tumors [Schlötterer and Harr, 2004, Popat et al., 2005, André et al., 2020]
- MSI status is determined by immunohistochemistry or PCR

# MSIntuit CRC [Saillard et al., 2023]

- MSIntuit CRC is a deep learning model for MSI status prediction from H&E WSI in colorectal cancer (CRC)
- The model detects MSI patients (0.97 sensitivity) and rules out half of MSS patients (0.46 specificity)

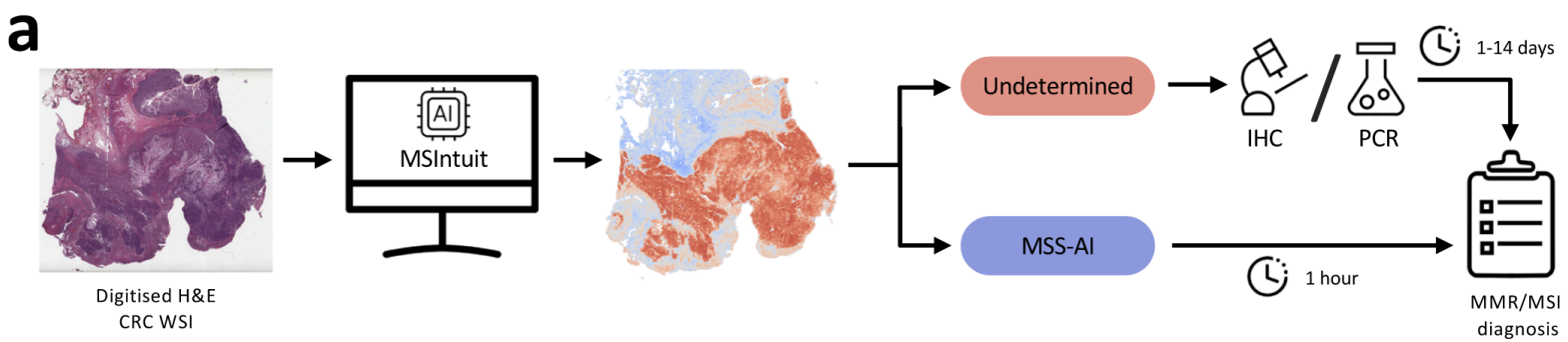



Figure: Clinical workflow of MSI screening with MSIntuit CRC.

# Owkin's digital pathology pipeline

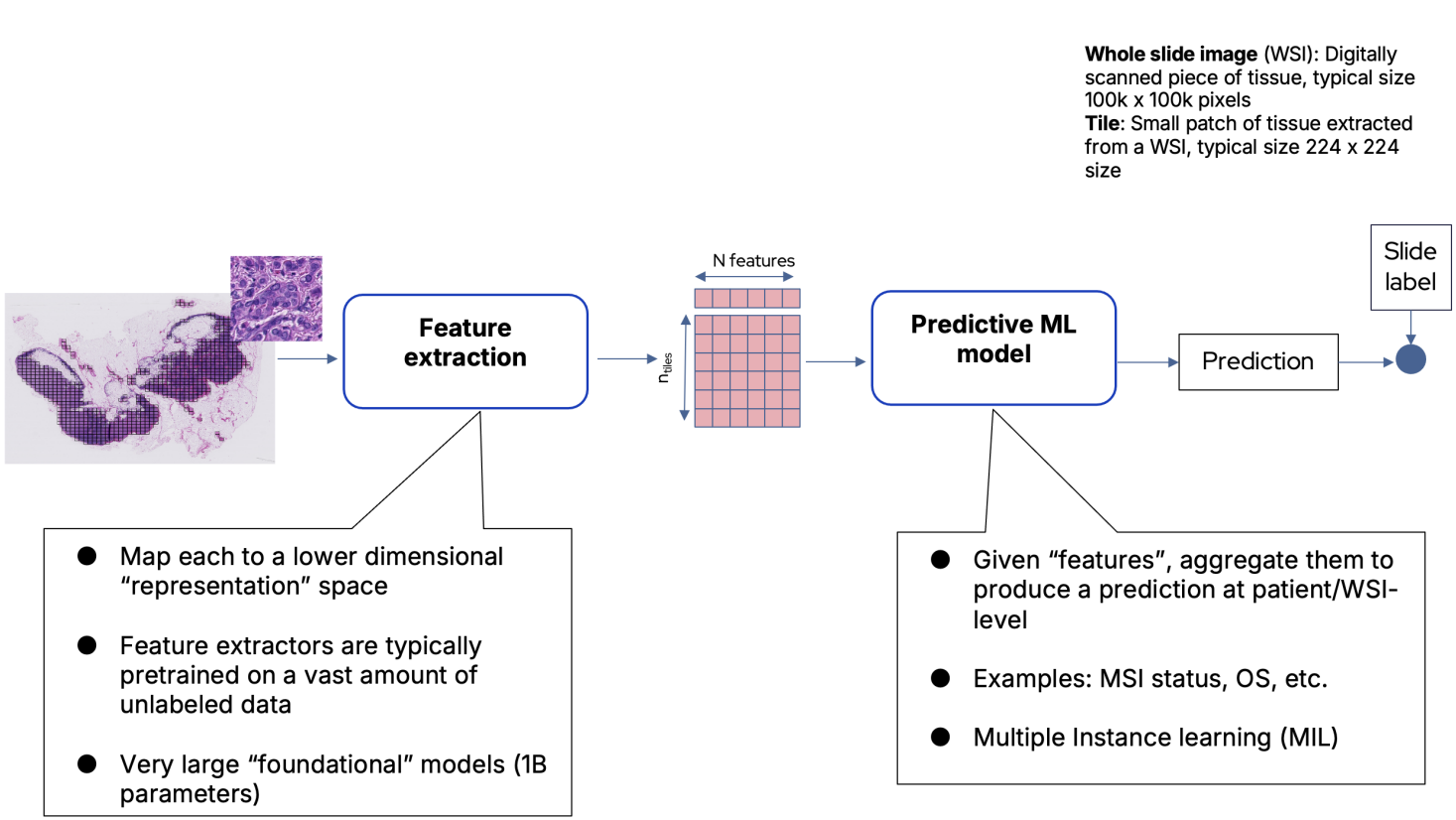


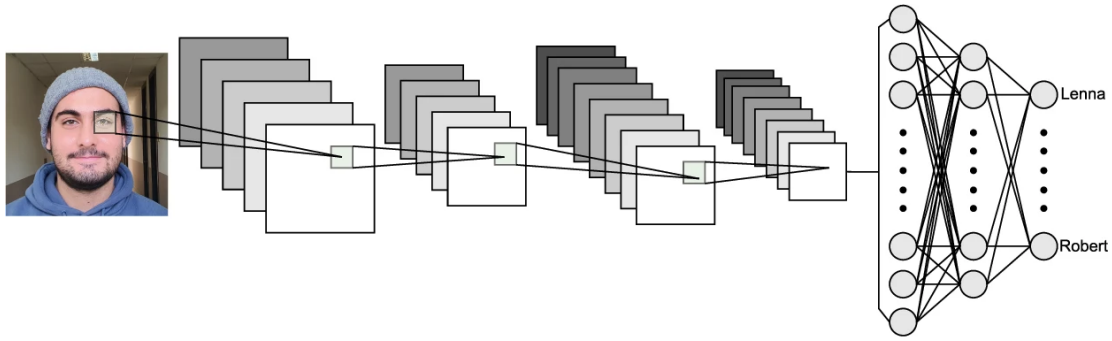

Figure: Overview of Owkin's digital pathology pipeline.

# Distilling foundation models for robust digital pathology

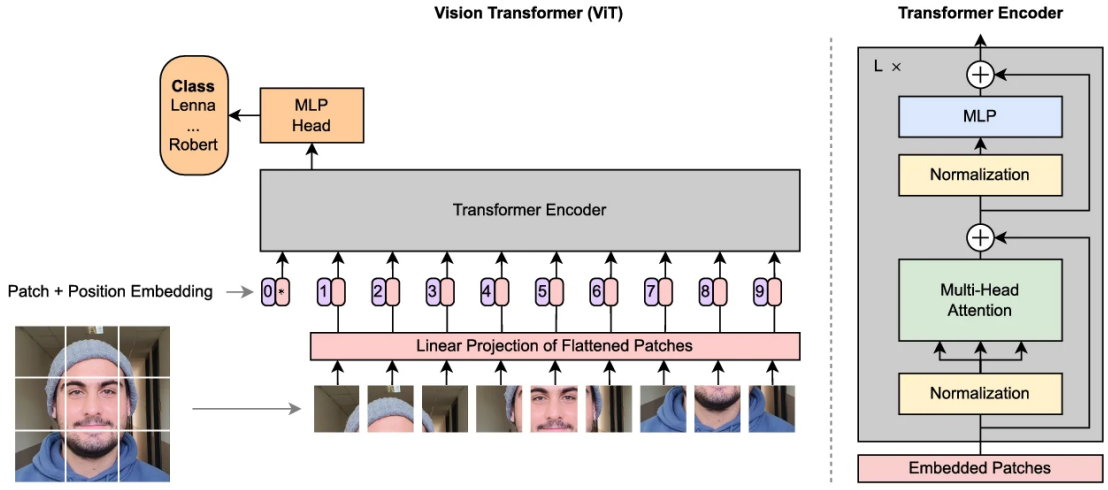
# Foundational models for histopathology

- Large models trained on large data sets which can be applied to a wide range of tasks
- Recent methodological developments
  - Vision Transformers (ViT) architecture [Dosovitskiy et al., 2021]
  - Self-supervised contrastive learning algorithms [Caron et al., 2021, Zhou et al., 2022, Oquab et al., 2023]

# Vision Transformers (ViT) architecture [Dosovitskiy et al., 2021]




**(a)** Common CNN architecture




**(b)** Vision Transformer architecture

Figure: Comparison of CNN and ViT architectures taken from [Rodrigo et al., 2024].

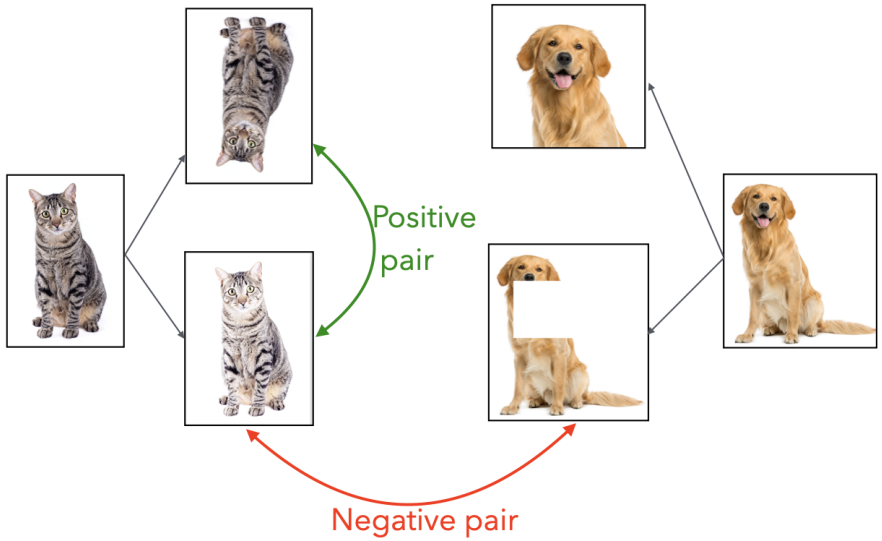
# Self-supervised contrastive learning [Chen et al., 2020]

- Self-supervised learning (SSL) uses the data to produce supervised tasks, instead of external labels.
- Example: Principal Component Analysis (PCA)

$$\text{argmin}_C \|X - XCC^T\|_F^2, \text{ such that } C^T C = I.$$

- **Contrastive** SSL uses data augmentation to produce supervised tasks.

# Self-supervised contrastive learning [Chen et al., 2020]




- Maximize agreement between positives and minimize agreement between negatives.
- Contrastive loss: (normalized temperature-scaled) cross-entropy

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}.$$

# Self-supervised contrastive learning [Chen et al., 2020]

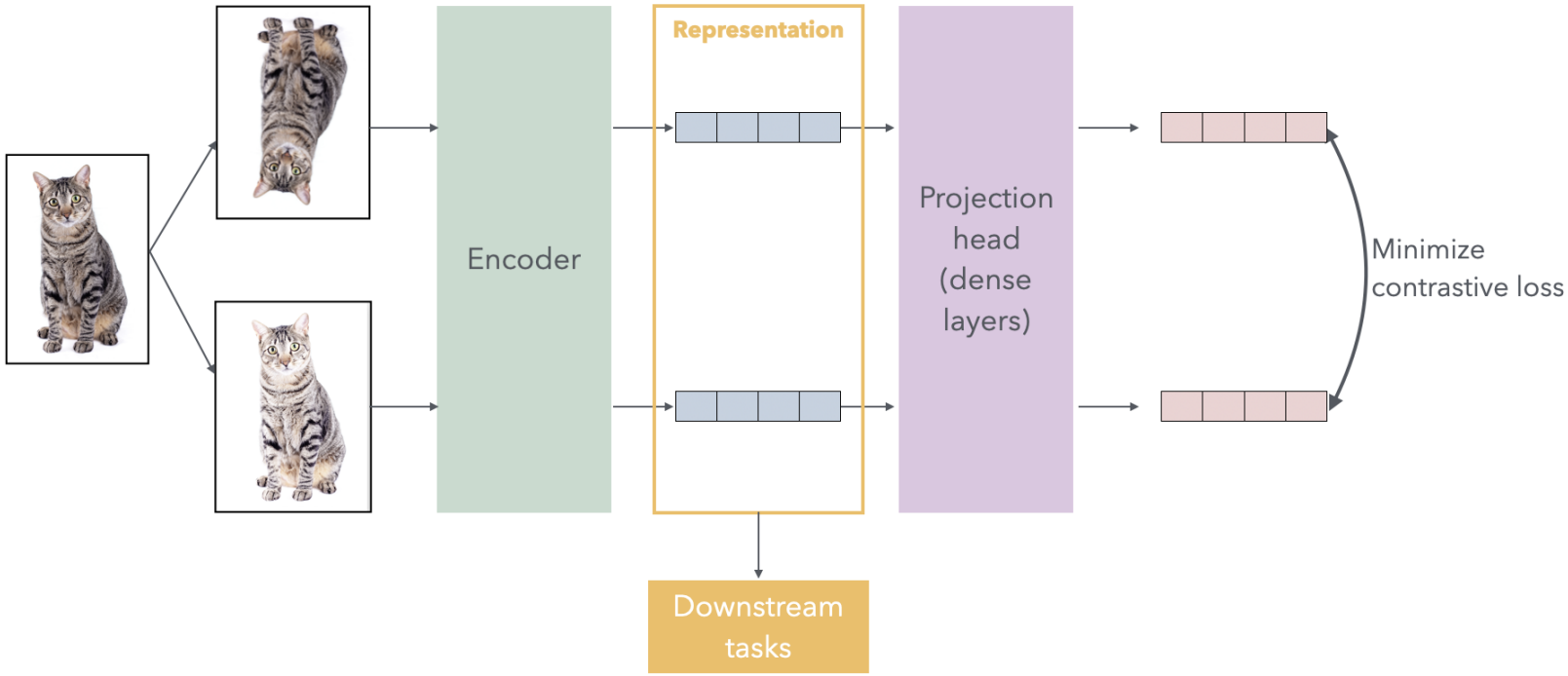



Figure: General architecture of self-supervised contrastive learning.

# Self-supervised contrastive learning [Chen et al., 2020]

- Requires negative samples to avoid collapse (all samples mapped to the same representation).

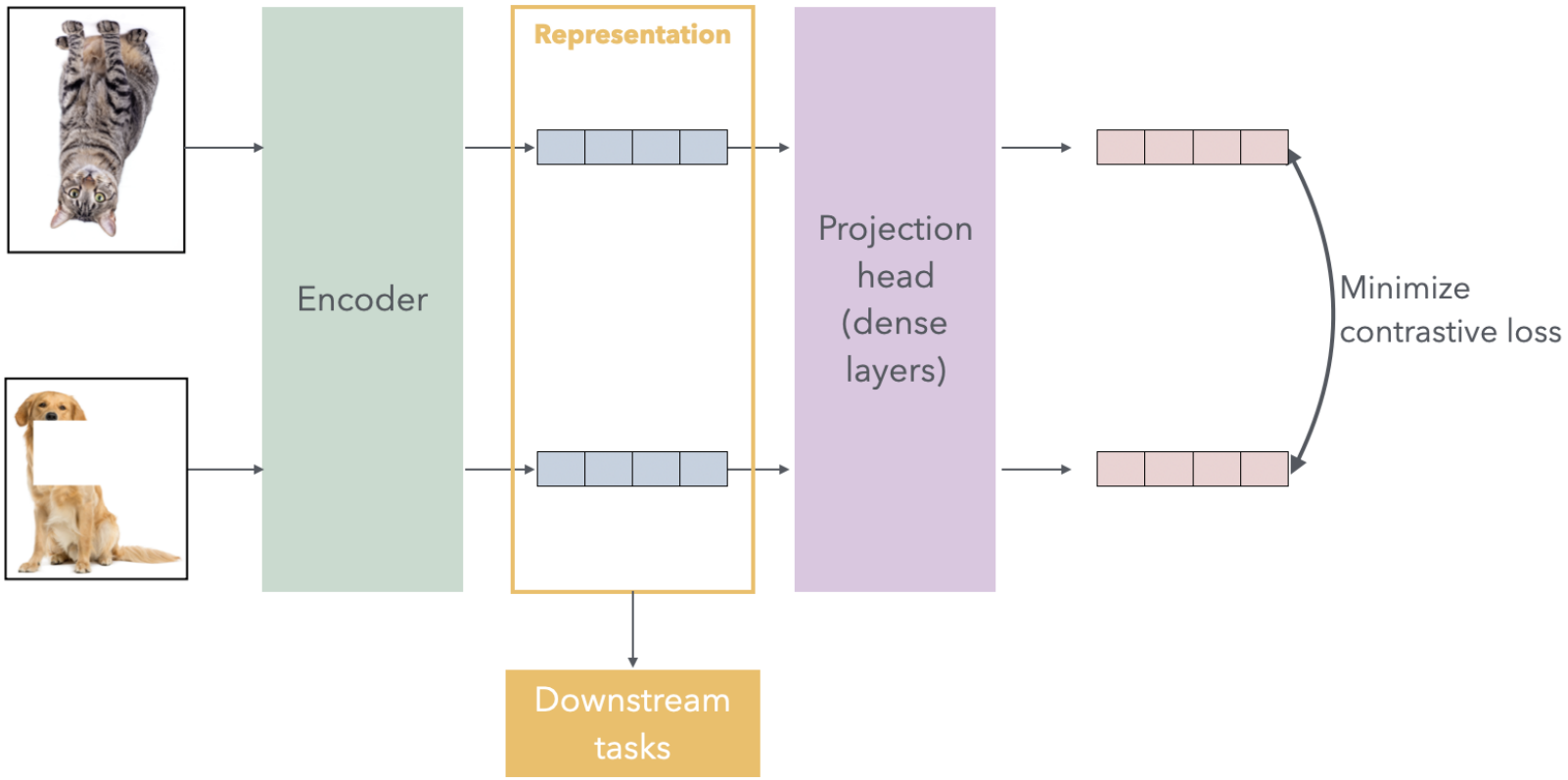



Figure: General architecture of self-supervised contrastive learning.

# Self-distillation with no labels (DINO) [Caron et al., 2021]

- Self-supervised contrastive learning tends to collapse when not enough negative samples are used.
- Computationally expensive, and sensitive to the choice of negatives.
- The use of negative samples can be avoided by using a teacher-student (distillation) architecture.

# Self-distillation with no labels (DINO) [Caron et al., 2021]

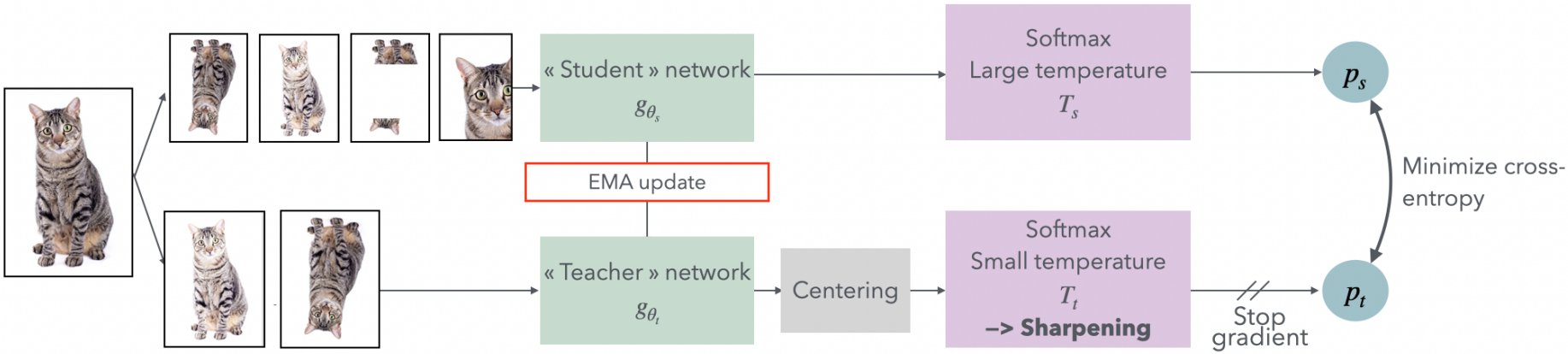



Figure: Self-distillation DINO learning framework.

# Foundational models for histopathology

- Pre-trained feature extractors used to embed tiles.
- State-of-the-art models rely on ViT architecture and DINO learning framework.
- Specific model properties usually derive from the pre-training data set.
- Since 2024: 'race' to improve FM by scaling data sets and model size.
- Most recent models have >1B parameters.

# Foundational models for histopathology

- Scaling data sets and model size improves performance

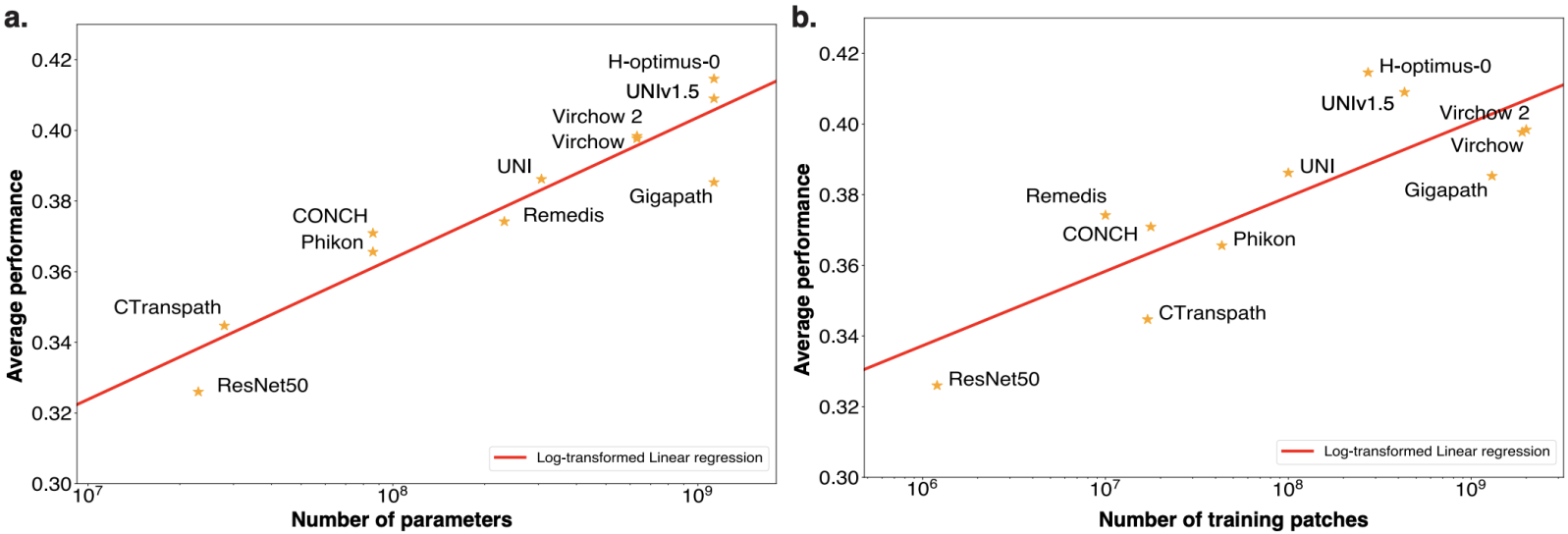
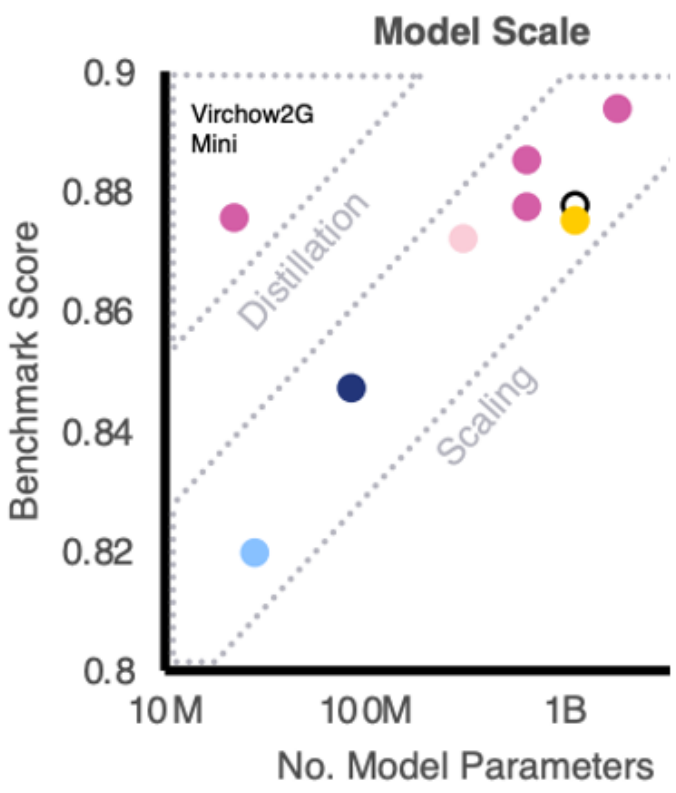


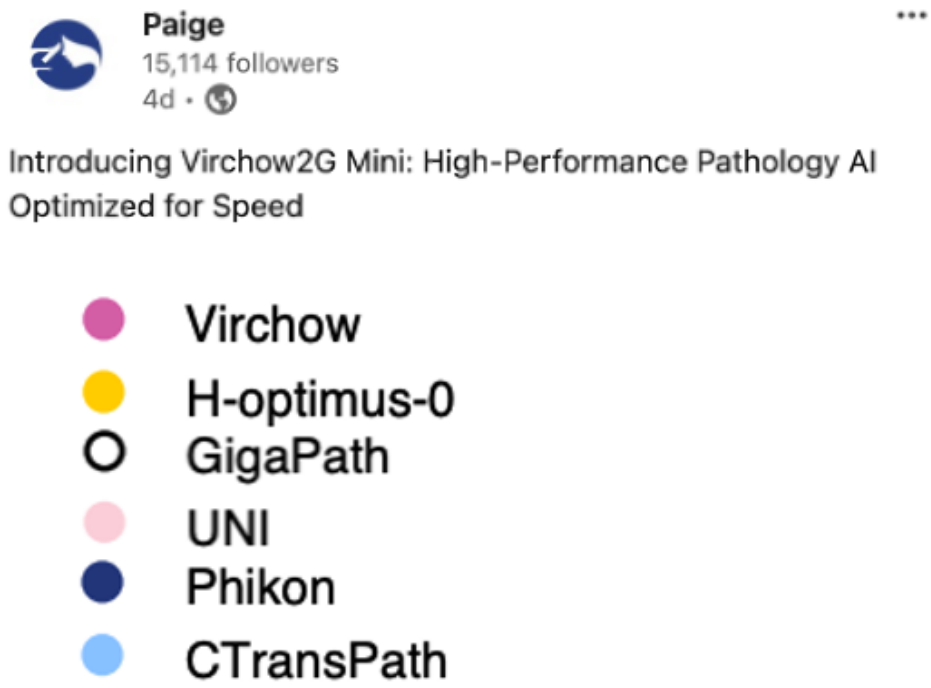

Figure: Scaling law comparing the number of parameters (a) and training patches (b) to average performance [Jaume et al., 2024]

# Methodological challenges

- Objective: include FMs as feature extractors in Owkin's pipeline, and deploy them in labs/hospitals
- Question 1: State-of-the-art FMs are too expansive to be deployed in clinical practice, can we compress them?
- Question 2:Most are pre-trained on large public data sets, can we fine-tune them on smaller, private data?

# Re-distill into smaller models: Example of Virchow2G Mini [Zimmermann et al., 2024]



# Model distillation

- Distillation of neural networks [Hinton et al., 2015] aims at reproducing the outputs of a large model with a smaller one.

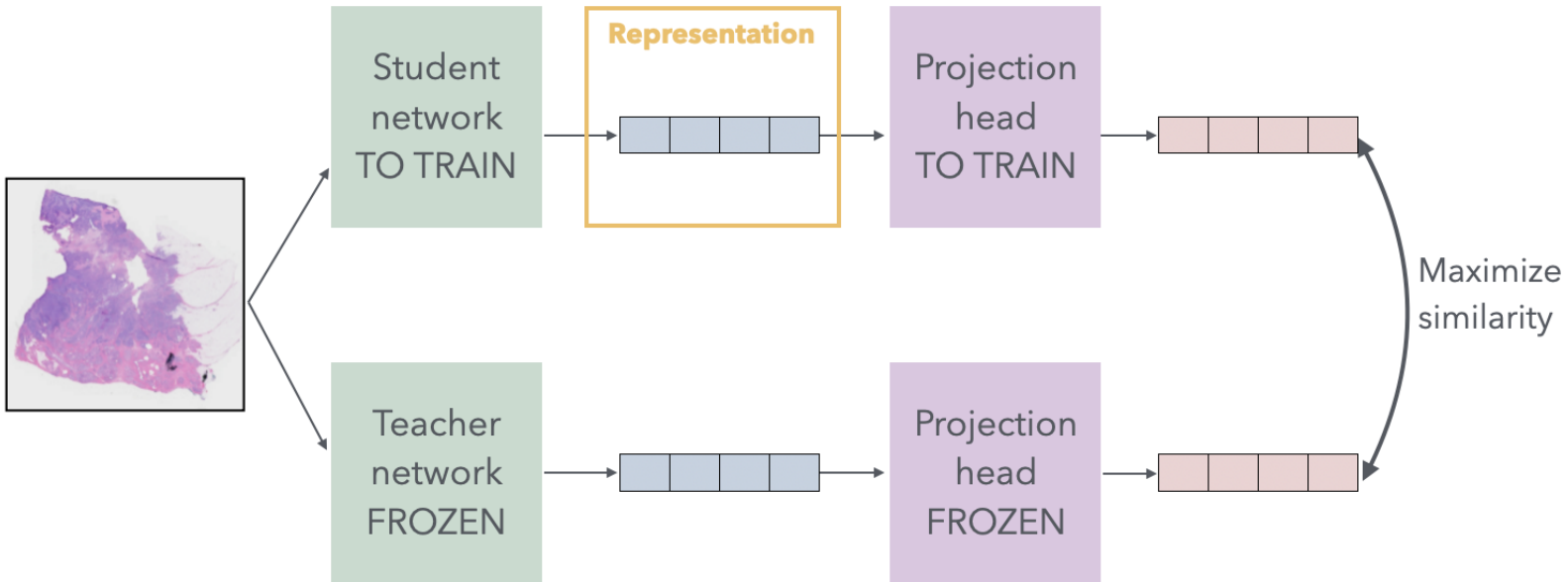



Figure: Classical distillation procedure

# Model distillation

- Distillation of a frozen FM into a smaller ViT model.

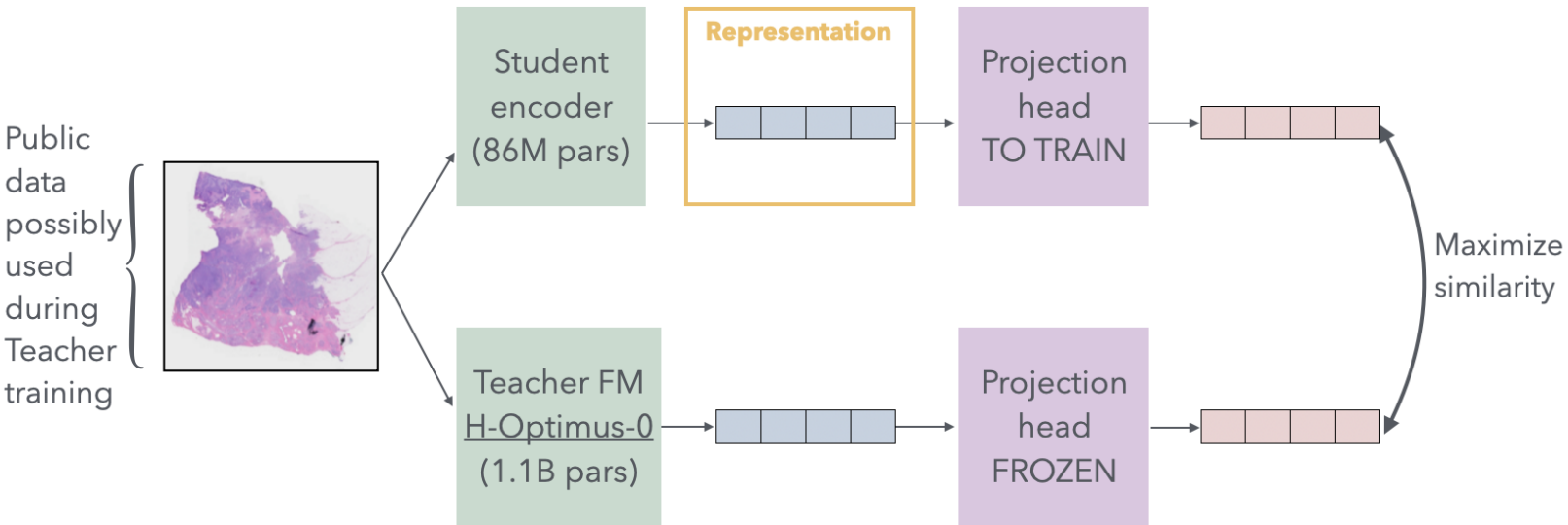



Figure: Distillation of H-Optimus-0[1].

[1] Developed by Bioptimus, partner company

# Model distillation as fine-tuning

- Distillation of a frozen FM using small proprietary data sets

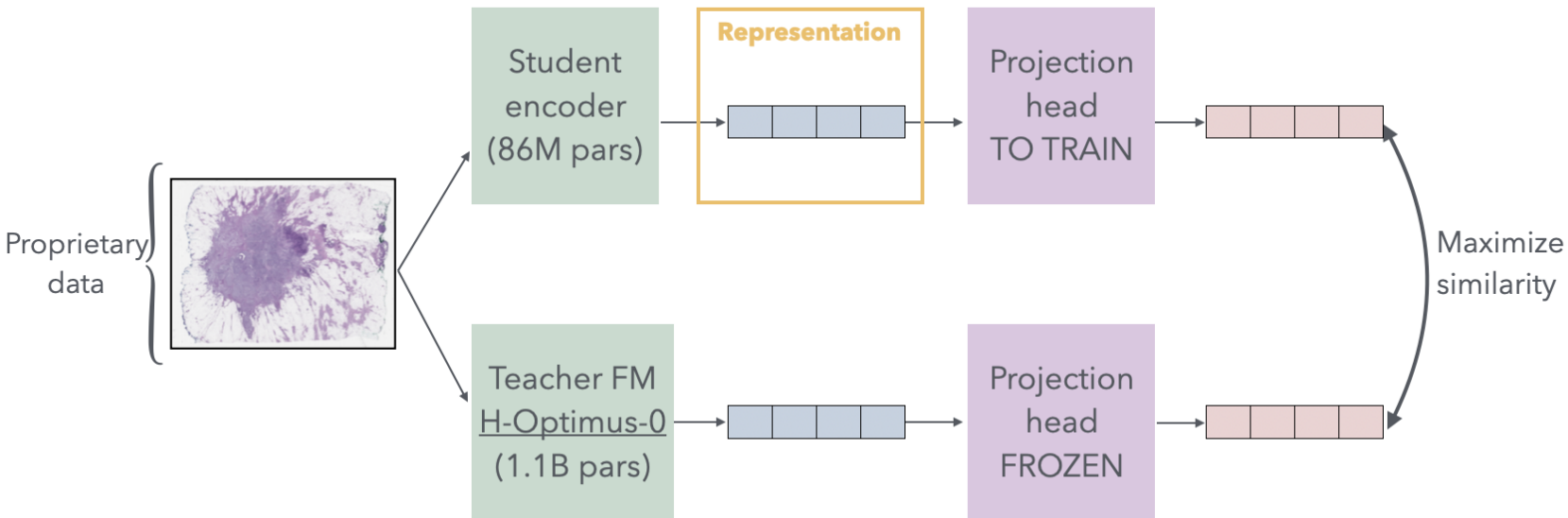



Figure: Distillation as fine-tuning on Owkin's proprietary data.

# Image-level vs patch-level information

- "Similarity" combines image-level loss (DINO cross-entropy) with a patch-level loss (iBoT)

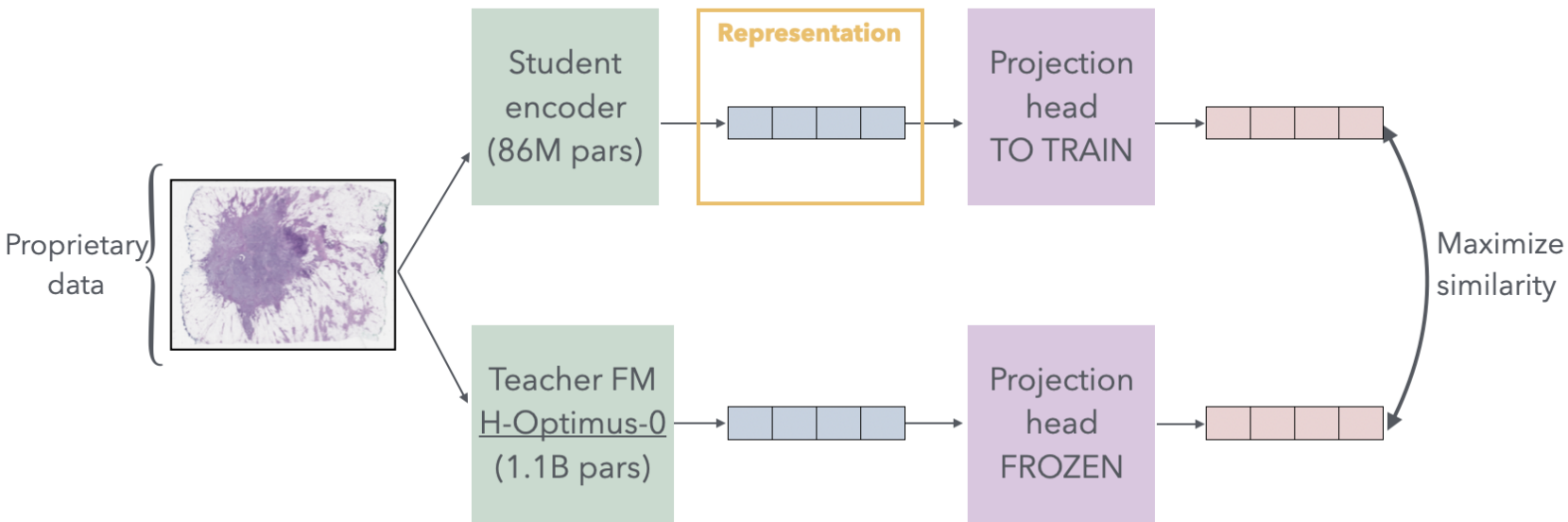



Figure: Distillation as fine-tuning on Owkin's proprietary data.

# Summary of the procedure

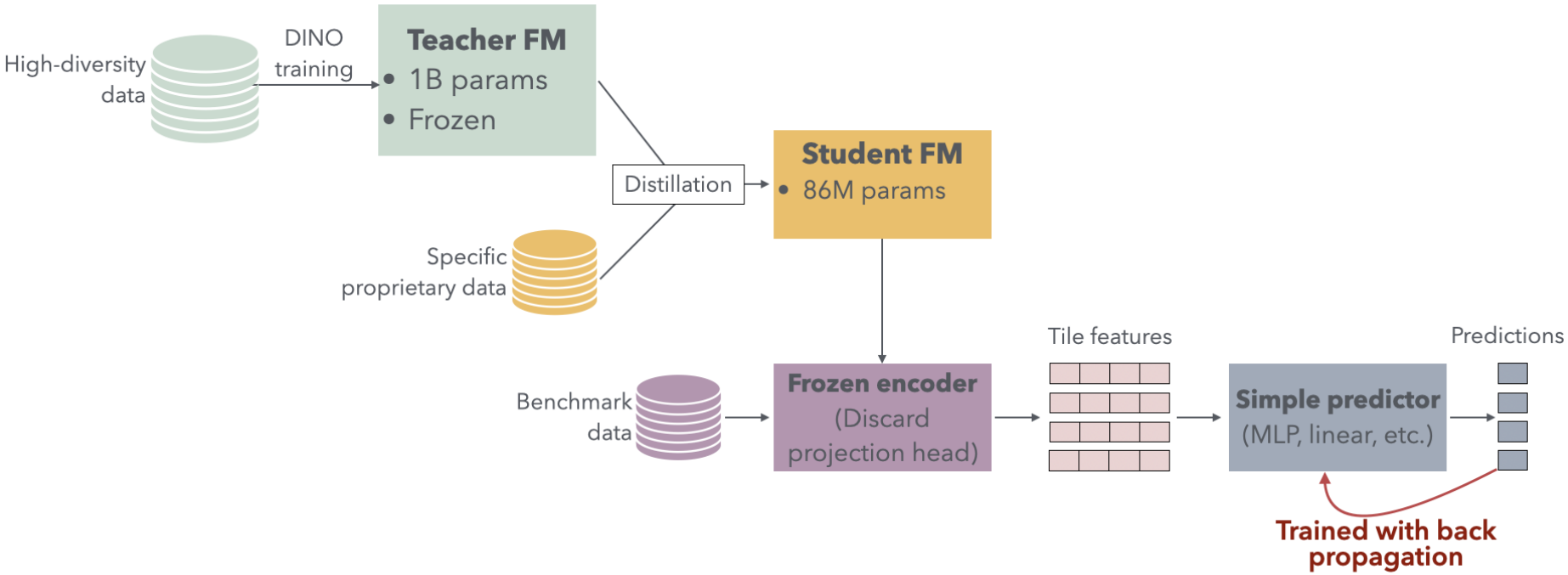



Figure: Complete model distillation procedure.

# Results on EVA benchmark [kaiko.ai et al., 2024]

| | | Patch-level classification | | | | Slide-level classification | | Segmentation | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Size** | **Bach** | **Crc** | **Mhist** | **Pcam** | **Cam16** | **Panda** | **Consep** | **Monusac** | **All** |
| Virchow2 | 632M | 0.880 | **0.966** | **0.858** | 0.936 | **0.864** | 0.642 | 0.630 | 0.663 | **0.794** |
| UNI2-h | 682M | **0.914** | 0.965 | 0.820 | 0.949 | 0.855 | 0.672 | 0.632 | 0.642 | 0.791 |
| H0 | 1,100M | 0.758 | 0.958 | 0.839 | 0.942 | 0.820 | 0.645 | 0.637 | **0.679** | 0.789 |
| Gigapath | 1,100M | 0.761 | 0.952 | 0.829 | 0.945 | 0.814 | 0.664 | 0.621 | 0.672 | 0.785 |
| GPFM | 307M | 0.830* | 0.952 | 0.811 | 0.945 | 0.851* | 0.647* | 0.639 | 0.640 | 0.784 |
| UNI | 307M | 0.797 | 0.947 | 0.844 | 0.936 | 0.834 | 0.656 | 0.628 | 0.638 | 0.783 |
| **H0-mini** | 86M | 0.774 | 0.961 | 0.790 | 0.942 | 0.842 | 0.667 | 0.629 | 0.643 | 0.782 |
| Hibou $L_{16}$ | 307M | 0.816 | 0.931 | 0.826 | **0.951** | 0.832 | 0.633 | 0.642 | 0.658 | 0.782 |
| Kaiko $B_8$ | 86M | 0.858 | 0.957 | 0.823 | 0.918 | 0.818 | 0.638 | **0.645** | 0.675 | 0.782 |
| Phikon | 86M | 0.722 | 0.936 | 0.799 | 0.922 | 0.797 | 0.640 | 0.629 | 0.644 | 0.767 |
| PhikonV2 | 307M | 0.727 | 0.939 | 0.775 | 0.893 | 0.808 | 0.635 | 0.630 | 0.639 | 0.760 |

Figure: Performance on EVA benchmark for tile-level and slide-level classification tasks.

# Results on HEST-1K benchmark [Jaume et al., 2024]

| Model | Idc | Prad | Paad | Skcm | Coad | Read | Ccrcc | Luad | L. Idc | Mean |
|---|---|---|---|---|---|---|---|---|---|---|
| UNI2-h | 0.6054 | 0.3753 | **0.5231** | **0.6829** | **0.3319** | 0.2265 | 0.2662 | 0.5743 | **0.2743** | **0.4292** |
| H0 | **0.6106** | 0.3621 | 0.5106 | 0.6614 | 0.3089 | **0.2401** | 0.2669 | 0.5754 | 0.2664 | 0.4224 |
| **H0-mini** | 0.5909 | 0.3633 | 0.5068 | 0.6125 | 0.2700 | 0.2047 | 0.2643 | 0.5633 | 0.2640 | 0.4044 |
| Virchow2 | 0.5971 | 0.3528 | 0.4778 | 0.6404 | 0.2580 | 0.2073 | 0.2604 | 0.5685 | 0.2568 | 0.4019 |
| Hibou $L_{16}$ | 0.5945 | 0.3231 | 0.4758 | 0.6059 | 0.3128 | 0.1823 | **0.2777** | 0.5720 | 0.2490 | 0.3992 |
| Gigapath | 0.5707 | **0.3841** | 0.4920 | 0.5823 | 0.3076 | 0.186 | 0.2277 | 0.5579 | 0.2499 | 0.3952 |
| GPFM | 0.5796 | 0.3733 | 0.4686 | 0.5839 | 0.2801 | 0.1769 | 0.2510 | 0.5522 | 0.2391 | 0.3940 |
| Kaiko $B_8$ | 0.5710 | 0.3827 | 0.4727 | 0.5904 | 0.3105 | 0.1726 | 0.2664 | **0.5883** | 0.2362 | 0.3912 |
| UNI | 0.5851 | 0.3274 | 0.4882 | 0.6235 | 0.2583 | 0.1757 | 0.2463 | 0.5558 | 0.2576 | 0.3907 |
| PhikonV2 | 0.5677 | 0.3793 | 0.4771 | 0.5845 | 0.2561 | 0.1865 | 0.2607 | 0.5502 | 0.2476 | 0.3897 |
| Phikon | 0.5481 | 0.3452 | 0.4639 | 0.5555 | 0.2668 | 0.1667 | 0.2496 | 0.5679 | 0.2387 | 0.3780 |

Figure: Performance on HEST-1k benchmark for tile-level gene expression prediction.

# PLISM data set and benchmark [Ochi et al., 2024, Filiot et al., 2025]

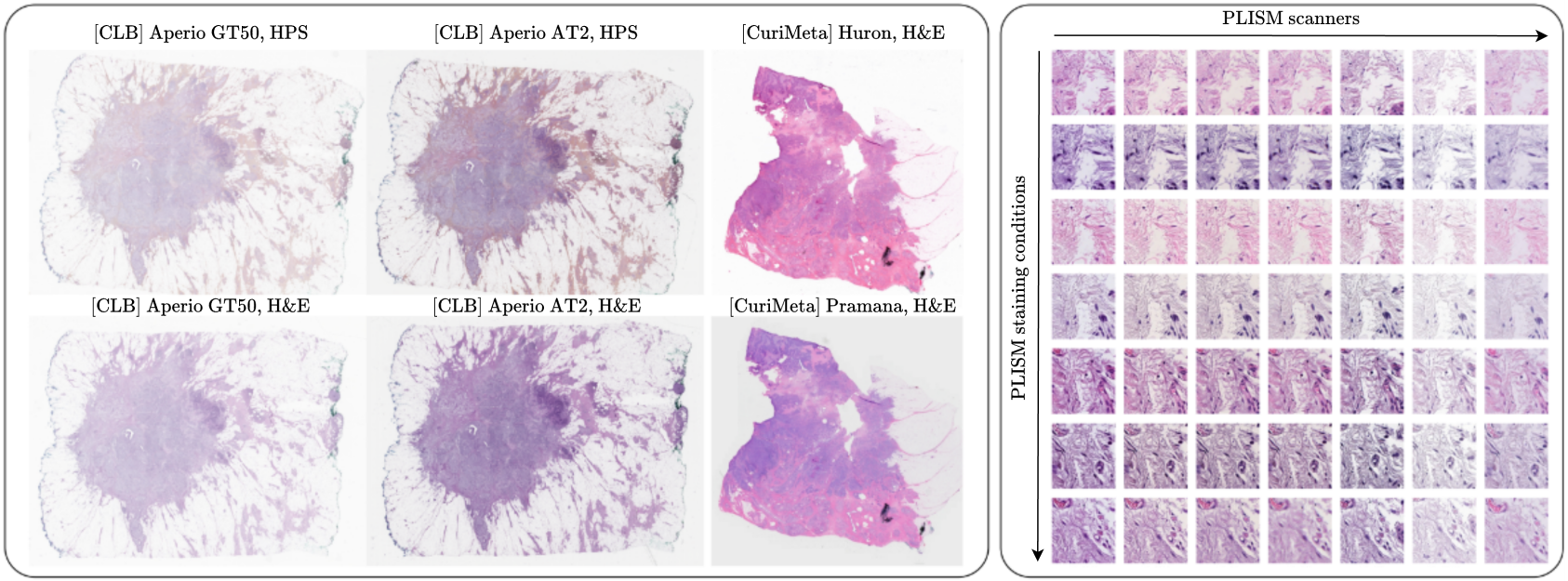



Figure: Visualizations of BreastBm (left) and PLISM (right) datasets. For PLISM, we only display 7 of the 13 different stainings on the y-axis.

# PLISM data set and benchmark [Ochi et al., 2024, Filiot et al., 2025]

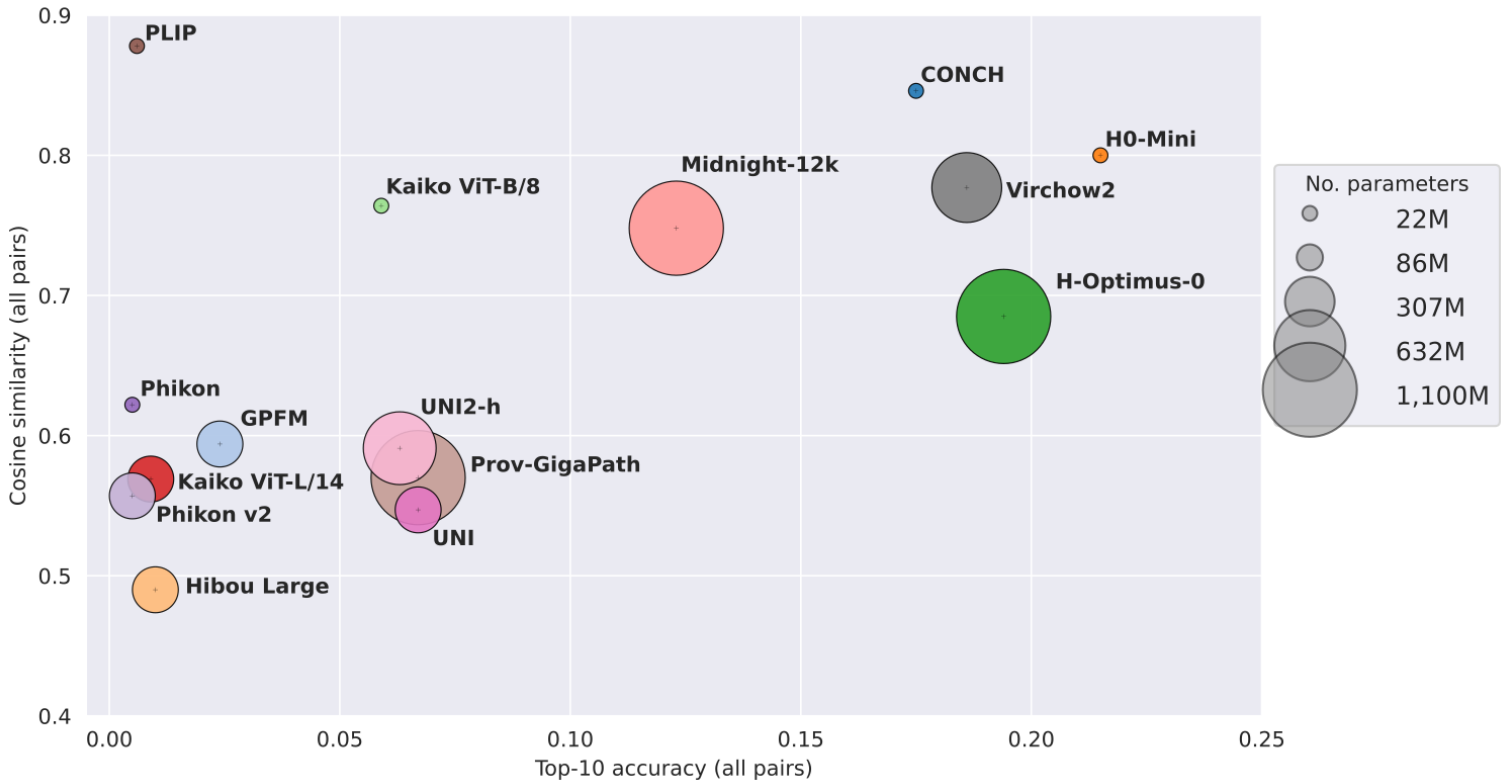



Figure: Robustness of digital pathology representation models to staining and scanner variations

# PLISM data set and benchmark [Ochi et al., 2024, Filiot et al., 2025]

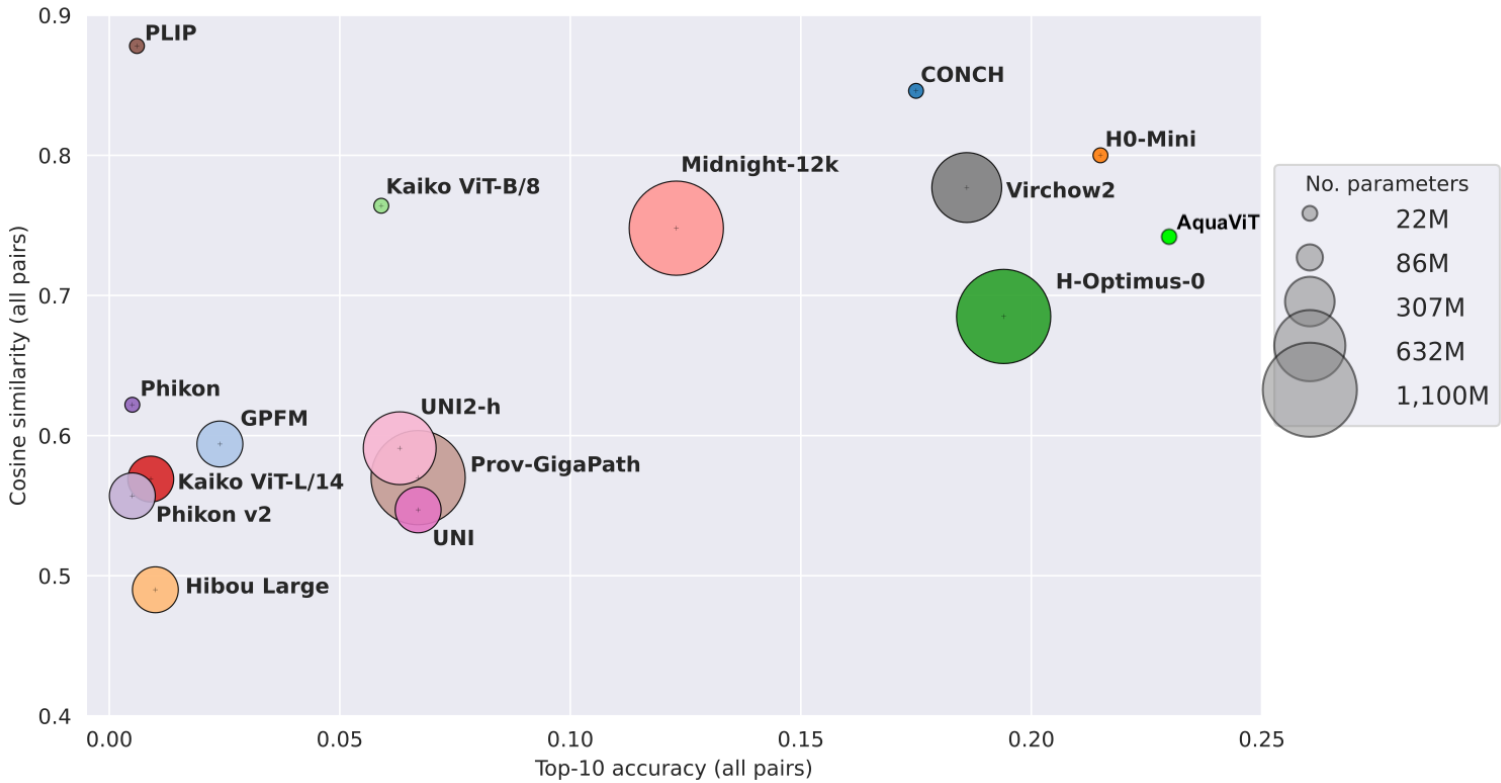



Figure: Robustness of digital pathology representation models to staining and scanner variations

# Conclusion on model distillation

- Foundational models for digital pathology have reached remarkable performance on a wide range of tasks
- They are too large for clinical deployment, and face robustness issues
- Model distillation enables simultaneous compression and fine-tuning of foundational models

# Robust sensitivity control in digital pathology

# Owkin's digital pathology pipeline

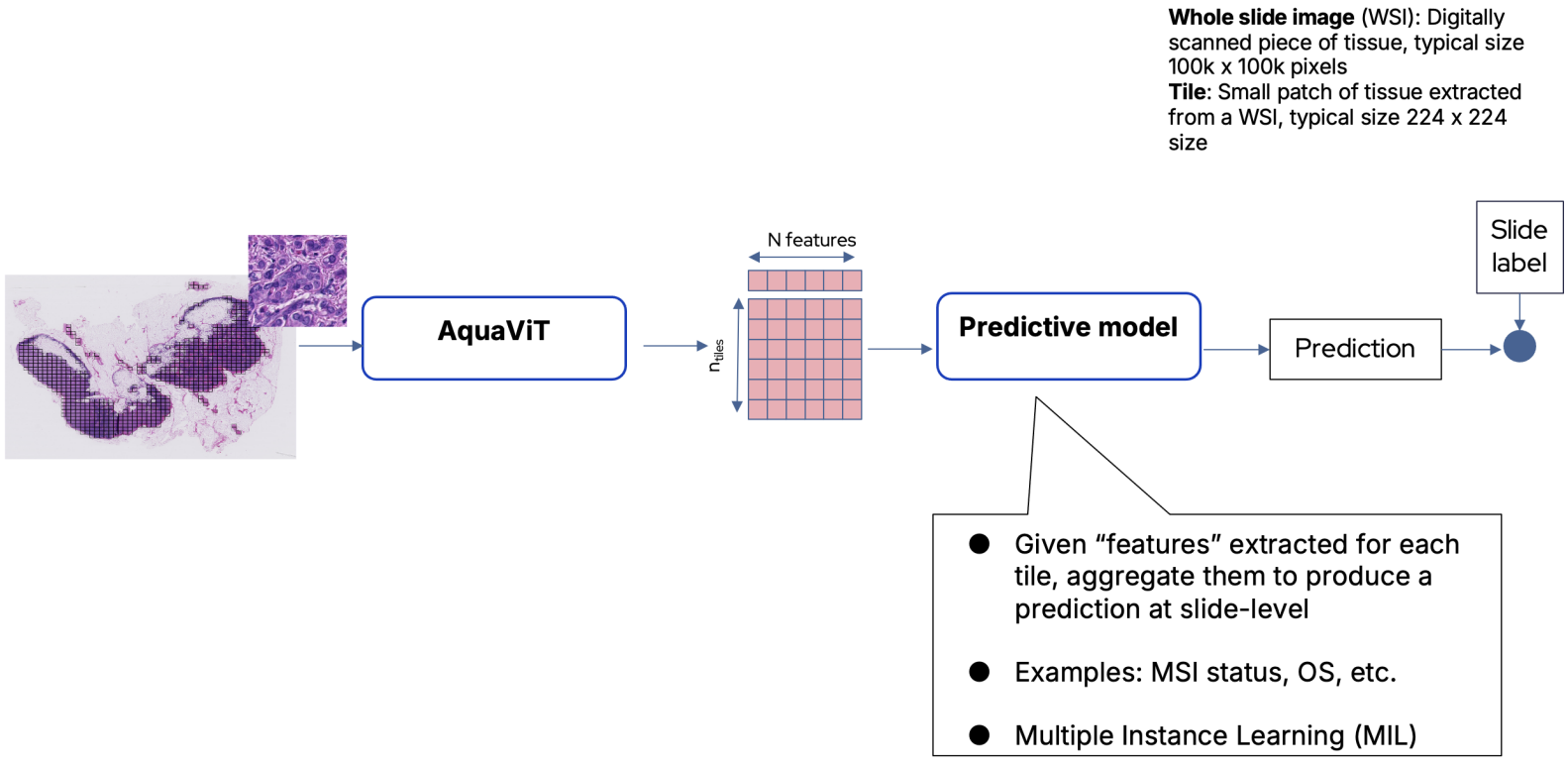
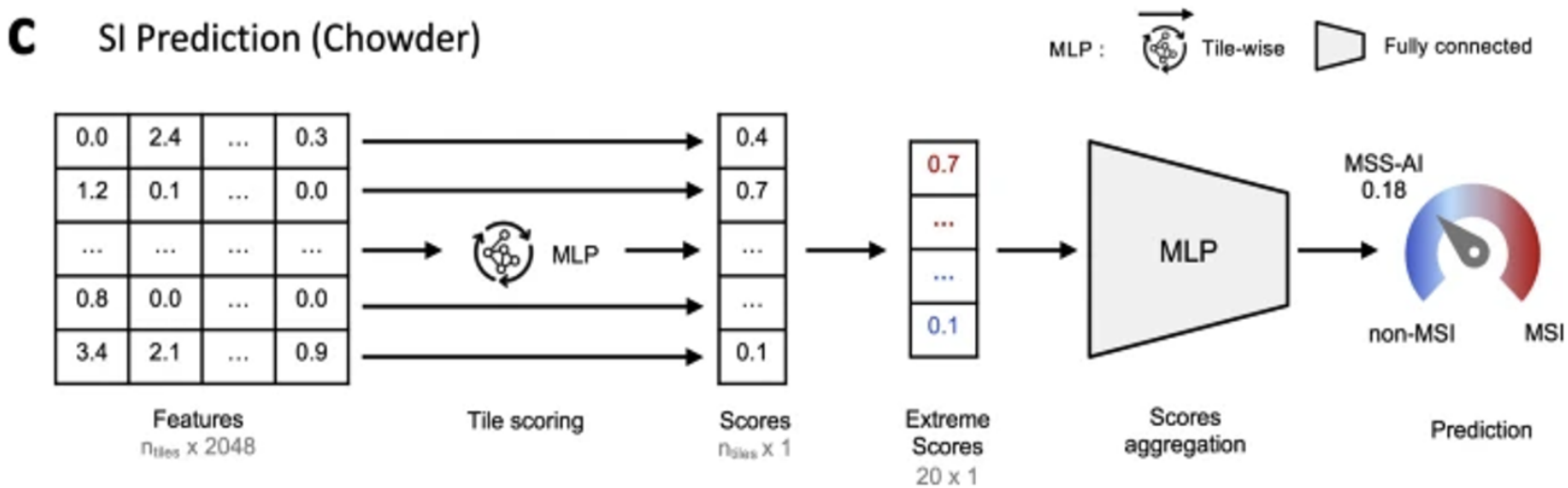



Figure: Overview of Owkin's digital pathology pipeline.

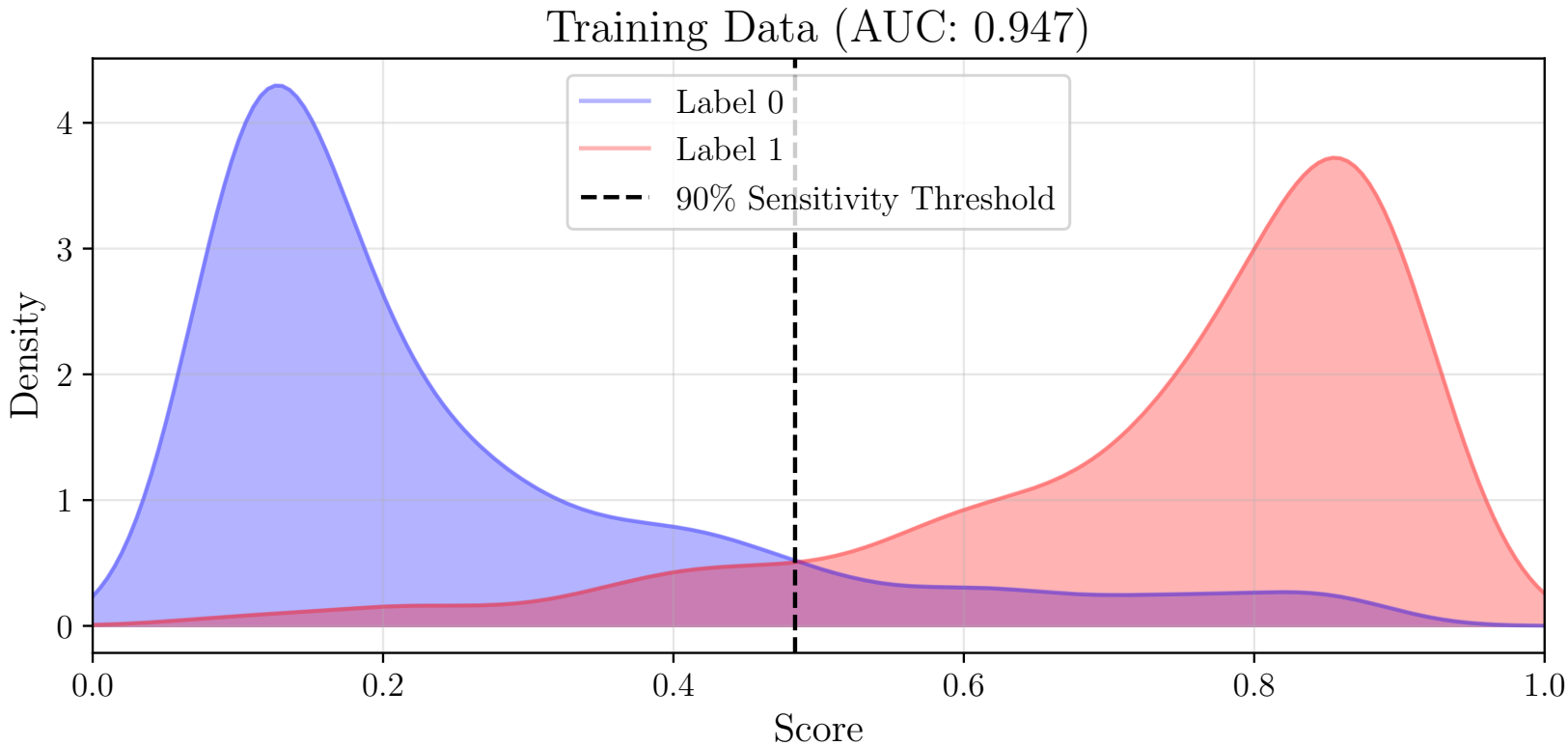
# Chowder Multiple Instance Learning (MIL) model [Courtiol et al., 2018]

- Chowder predicts a score for each tile, selects the top 10 highest and lowest scores and aggregates them into a slide level score



# From Chowder scores to MSI status prediction

- Chowder WSI scores are binarized using a threshold $\tau$ so as to achieve a certain sensitivity level on the training data



# Thresholds transfer poorly to external cohorts even when AUC is robust

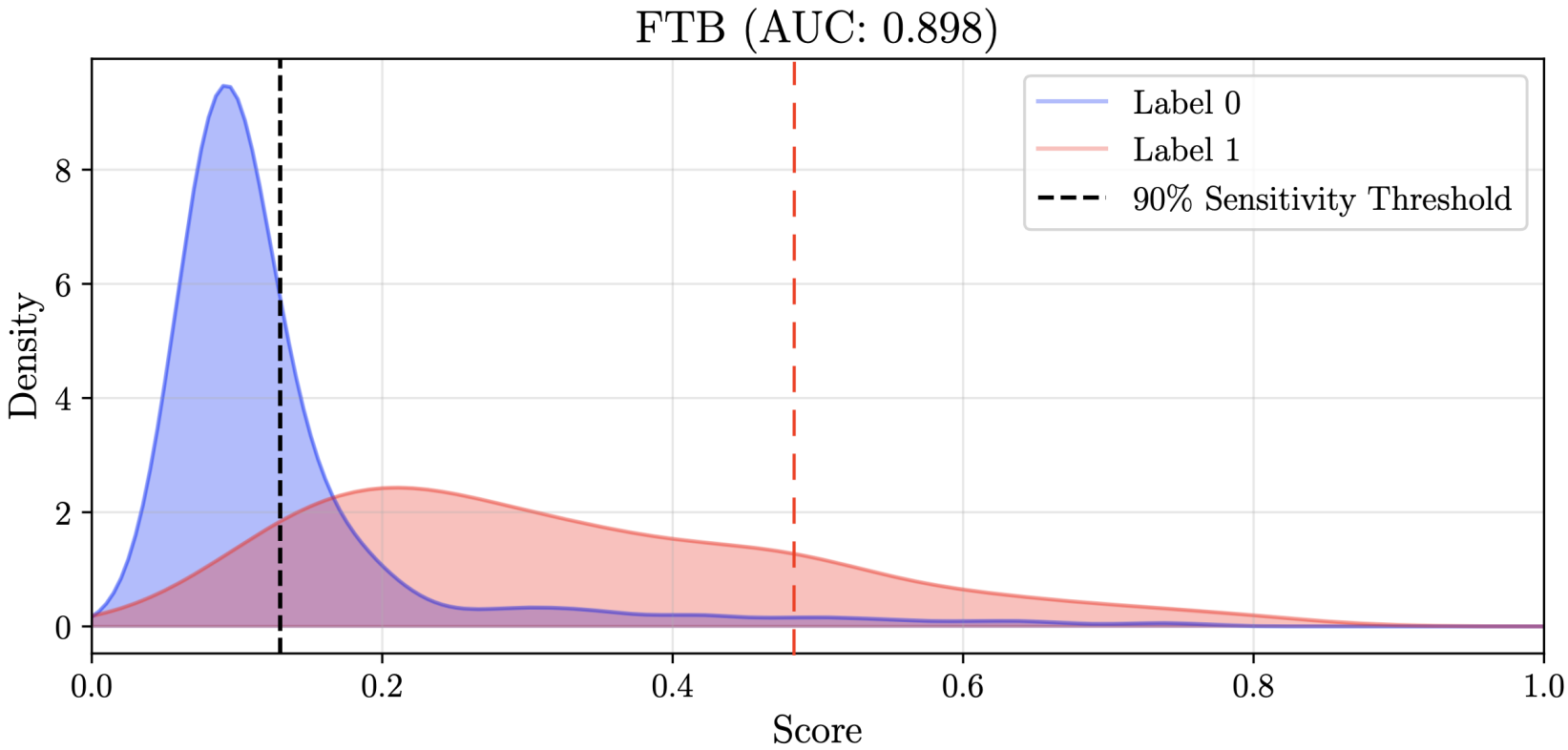



Figure: Threshold gap between training and external cohorts.

# MSIntuit CRC requires calibration data from the deployment site

- Clinical requirements: at least 95% sensitivity, high specificity, evaluated on blind external cohorts
- MSIntuit CRC computes the threshold $\tau$ on every new center by using 30 MSI positive calibration samples
- This procedure applicable in CRC where $\sim$15% of patients are MSI positive
- Many solid tumors (endometrial, gastric, etc.) have low MSI prevalence, and obtaining 30 positive samples is unrealistic

# Chowder MIL model

- Let $(S^t, Y^t)$ denote a labeled WSI from the reference cohort
- Let $(S^c, Y^c)$ denote a labeled WSI from the calibration cohort
- For a WSI $S$, we denote $S = (T_1, \ldots, T_N)$ its partition into $N$ tiles
- The Chowder prediction function can be written as

$$f(S) = h\left(r\left(g\left(T_1\right), \ldots, g\left(T_N\right)\right)\right),$$

where $g$ is the tile-level score, $r$ selects the top 10 highest and lowest scores, and $h$ aggregates them through an MLP.

# General idea

- Final model predictions are given by

$$\mathbf{1}\{f(S) \geq \tau\},$$

where $\tau$ is adjusted to satisfy a fixed sensitivity level $\sigma$ on the training data

$$\mathbb{P}_{S,Y}(f(S) > \tau | Y = 1) = \sigma.$$

- Tile Score Distribution Matching (TSM) matches the distribution of tile-level scores $g(T_i)$ between training and calibration, to improve the transferability of the threshold $\tau$.

# Tile Score Distributions

- Assume that tiles are i.i.d. conditionally to the slide label. For $1 \leq i, j \leq N$, $k \in \{t, c\}$ and $l \in \{0, 1\}$

$$\mathbb{P}\left(T_i^k | Y^k = l\right) = \mathbb{P}\left(T_j^k | Y^k = l\right)$$

- By continuity of $g$, tile scores $X_i^k = g(T_i^k)$ are also conditionally i.i.d. Denote $\omega^k = \mathbb{P}(Y^k = 1)$, the density of tile scores in cohort $k$ is given by

$$\rho_X^k = \omega^k \rho_{X|Y=1}^k + (1 - \omega^k)\rho_{X|Y=0}^k$$

# Tile Score Distribution Matching (TSM)

- Tile Score Distributions $\rho^t$ and $\rho^c$ are matched using optimal transport (OT) up to an adjustment w.r.t. prevalence

- The optimal map is computed using the Monge formulation

$$M^* = \underset{M \in \mathcal{M}_{a \to b}}{\arg\min} \int_{\mathbb{R}} |M(x) - x|^2 da(x),$$

where $a = \rho^c_X$ and

$$b = \omega^c \rho^t_{X|Y=1} + (1 - \omega^c)\rho^t_{X|Y=0}.$$

- Since $X$ is one-dimensional, the closed-form solution is given by quantile matching

# Sensitivity control with TSM

**Theorem ([Pignet et al., 2025])**

*Let $\tau \in \mathbb{R}$ and denote $sens_{train}(\tau)$ and $sens_{val}(\tau)$ the sensitivities associated to threshold $\tau$ on the training and validation cohorts. Assume $\omega^c = 1$, i.e., only positive samples are used for calibration. Then,*

$$sens_{train}(\tau) = sens_{val}(\tau).$$

# Sensitivity control with TSM

## Theorem ([Pignet et al., 2025])

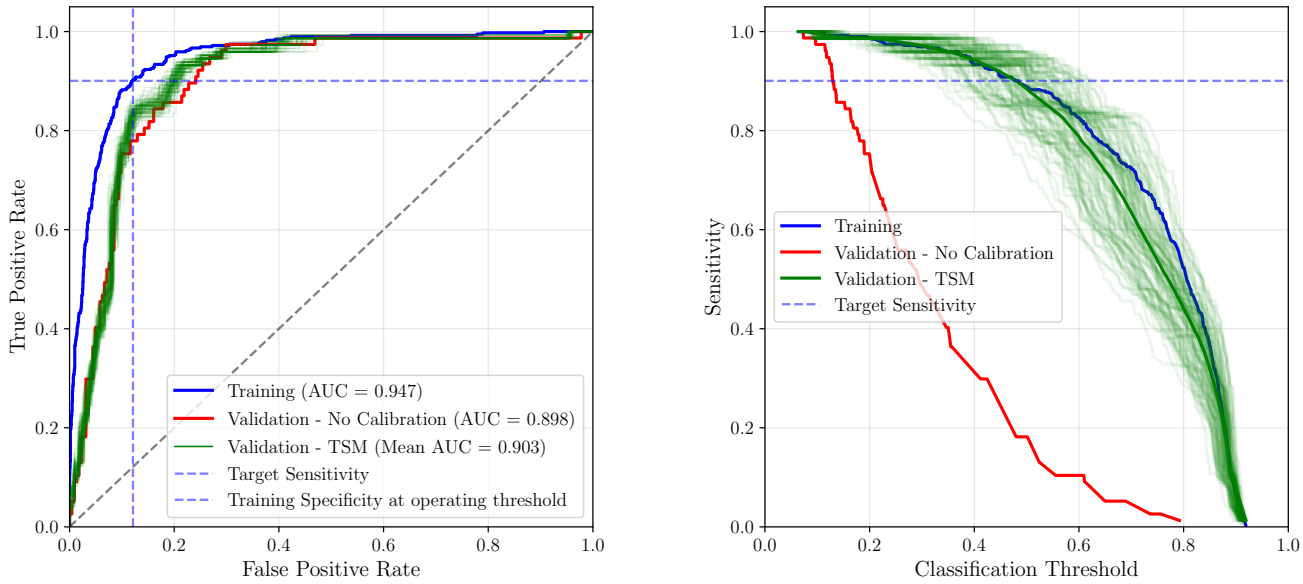
*Let $\tau \in \mathbb{R}$ and denote $sens_{train}(\tau)$ and $sens_{val}(\tau)$ the sensitivities associated to threshold $\tau$ on the training and validation cohorts. Assume $\omega^c = 1$, i.e., only positive samples are used for calibration. Then,*

$$sens_{train}(\tau) = sens_{val}(\tau).$$

- In practice, $\rho_X^c$ and $\rho_X^t$ are estimated using the wieghted sum of Dirac masses centered on each data point

$$\widehat{\rho_X^c} = \frac{1}{n_c} \sum_{i=1}^{n_c} \delta_{x_{(i)}^c}, \text{ and } \widehat{\rho_X^t} = \frac{1}{n_t} \sum_{i=1}^{n_t} \delta_{x_{(i)}^t}$$

# Application to MSI status prediction




(a) ROC curves.

(b) Sensitivity vs thresholds.

Figure: Metrics computed on training and FTB validation cohorts for the MSI classification task, using 30 positive samples for calibration.

# TSM controls sensitivity in low prevalence regimes

- Sensitivity control is improved by leveraging negative samples

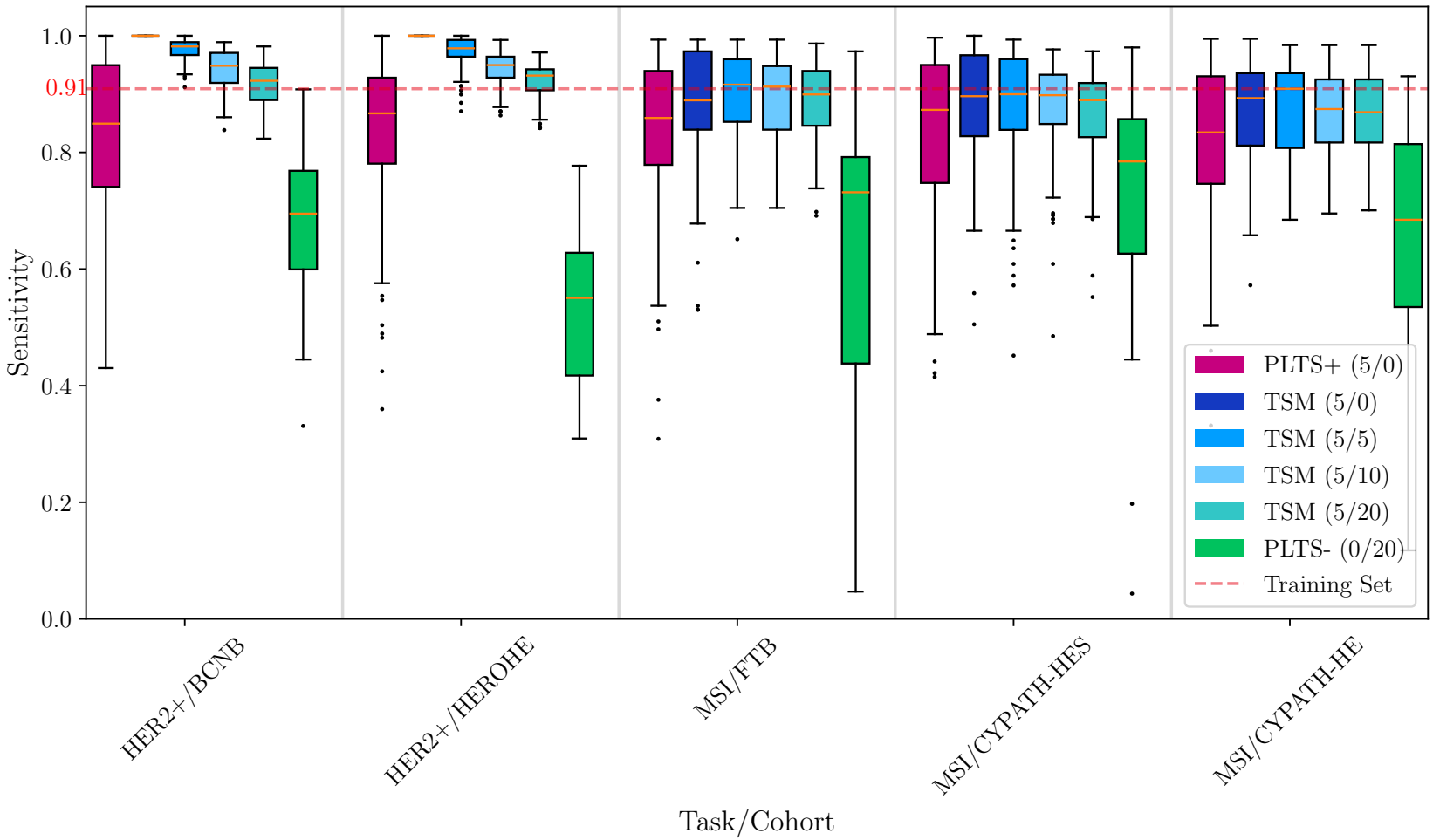



Figure: Sensitivity achieved by PLTS+ and TSM using 5 positive samples for calibration.

# Conclusion on TSM

- TSM controls sensitivity in low prevalence digital pathology classification tasks
- **Limitation:** The method is specific to the Chowder MIL model, which is based on tile-level scores aggregation

# General conclusion

- Model generalizability is key to deploy digital pathology tools in the clinical setting
- Generalization should be enhanced both in terms of model performance and statistical properties
- Model distillation provides compressed, robust foundational models
- Tile Score Distribution Matching (TSM) controls sensitivity in low prevalence regimes
- Next milestone: Diagnostic tools for low MSI prevalence solid tumors are in development based on AquaViT and TSM

# Extension to Spatial transcriptomics?

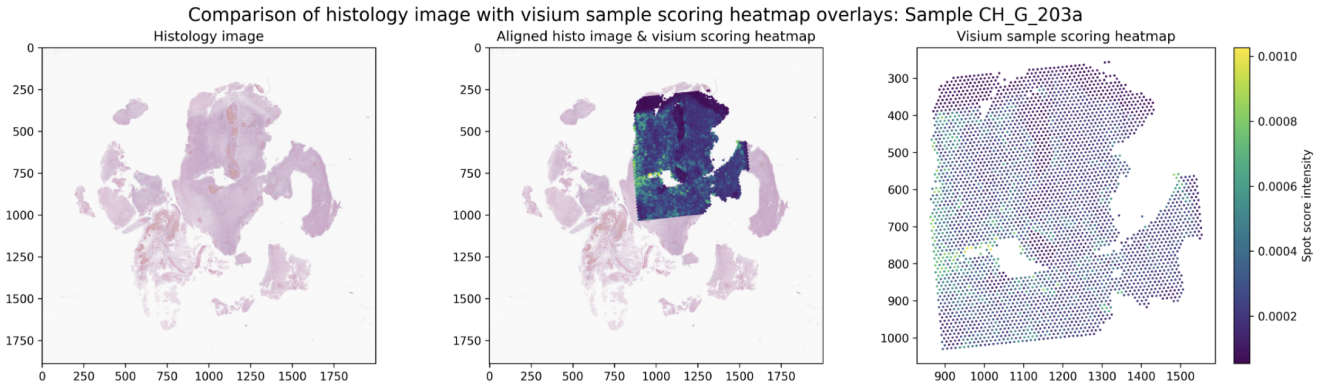
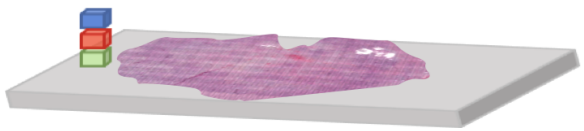
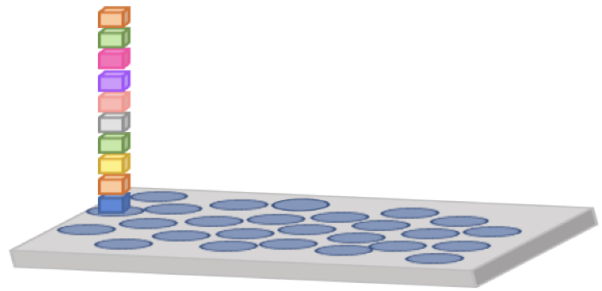



Figure: Example of a 10x Genomics Visium slide.



**H&E slide**: high-resolution image (~1Gpixel) with 3 channels (RGB)



**SpT sample**: low-resolution image (~5000 spots) with large nb of channels (~1e3 genes)

# Foundational models for SpT are disappointing

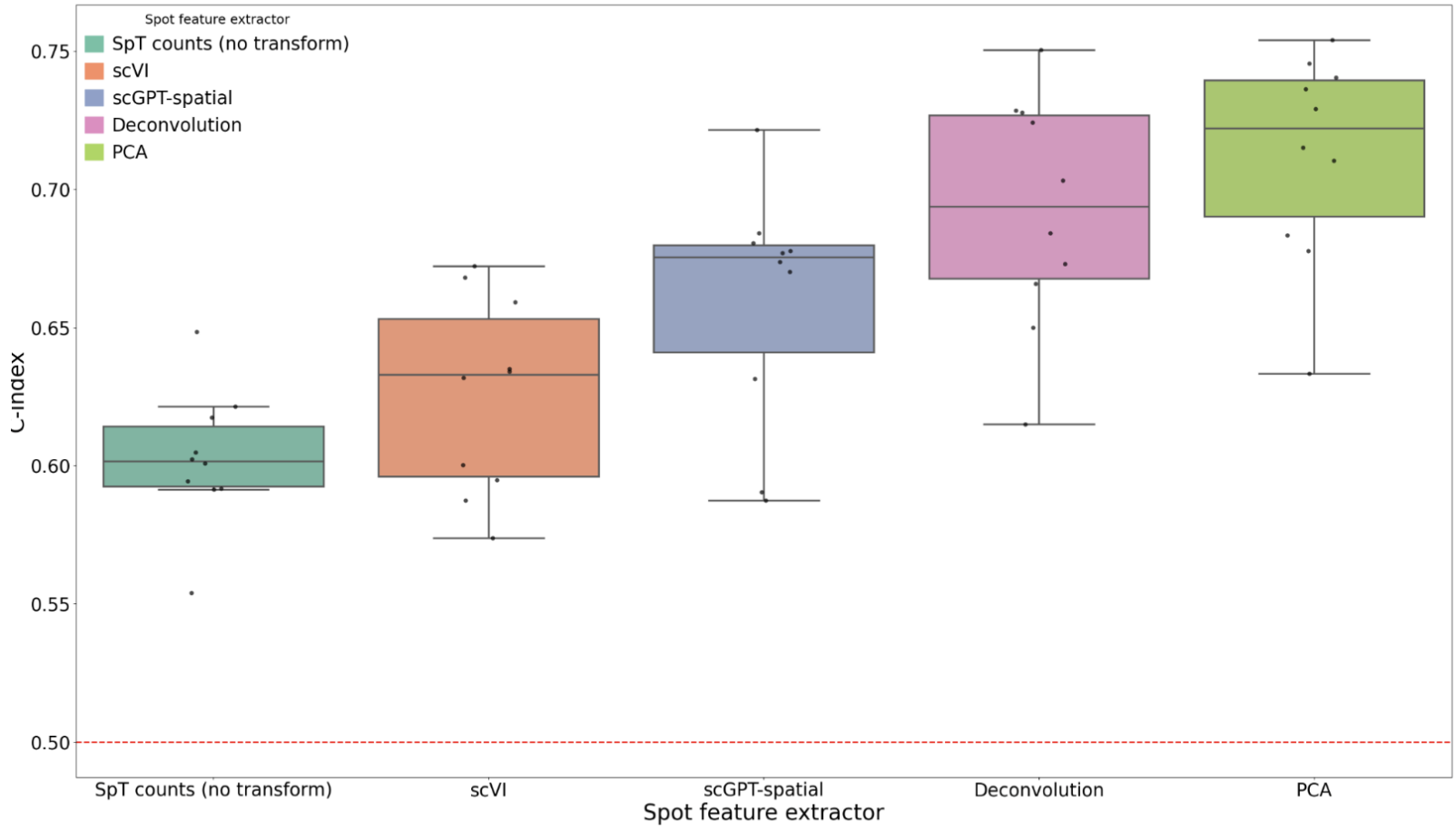



Figure: SpT representation with PCA clearly outperforms foundational models.

# Distillation can be used to perform multimodal prediction (HE and SpT)

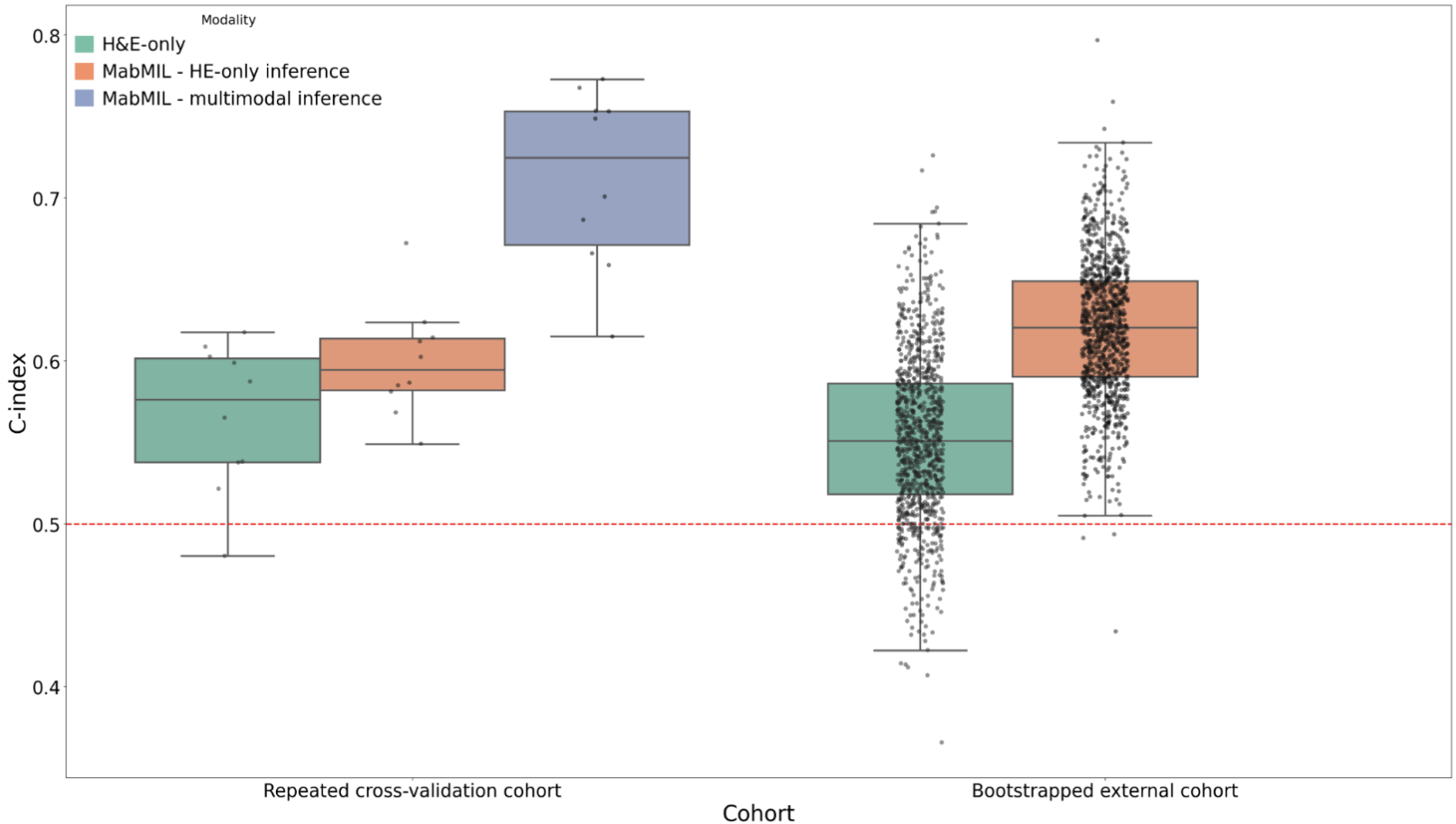



Figure: Distilling SpT models (teacher) into HE model (student) improves performance.

# Thank you!

Figure: Ballade au marais Poitevin (F.Roch-CRTNA).

# References I


André, T., Shiu, K.-K., Kim, T. W., Jensen, B. V., Jensen, L. H., Punt, C., Smith, D., Garcia-Carbonero, R., Benavides, M., Gibbs, P., De La Fouchardiere, C., Rivera, F., Elez, E., Bendell, J., Le, D. T., Yoshino, T., Van Cutsem, E., Yang, P., Farooqui, M. Z., Marinello, P., and Diaz, L. A. (2020).
Pembrolizumab in Microsatellite-Instability–High Advanced Colorectal Cancer.
*New England Journal of Medicine*, 383(23):2207–2218.

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021).
Emerging Properties in Self-Supervised Vision Transformers.
In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, Montreal, QC, Canada. IEEE.

# References II



Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020).
A simple framework for contrastive learning of visual representations.
In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.


Courtiol, P., Tramel, E. W., Sanselme, M., and Wainrib, G. (2018).
Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach.
Version Number: 2.


Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021).
An image is worth 16x16 words: Transformers for image recognition at scale.
In *International Conference on Learning Representations*.

# References III



Filiot, A., Dop, N., Tchita, O., Riou, A., Dubois, R., Peeters, T., Valter, D., Scalbert, M., Saillard, C., Robin, G., and Olivier, A. (2025).
Distilling foundation models for robust and efficient models in digital pathology.
Version Number: 2.


Filiot, A., Jacob, P., Kain, A. M., and Saillard, C. (2024).
Phikon-v2, A large and public feature extractor for biomarker prediction.
arXiv:2409.09173 [eess].


Hinton, G., Vinyals, O., and Dean, J. (2015).
Distilling the Knowledge in a Neural Network.
Version Number: 1.

# References IV



Jaume, G., Doucet, P., Song, A. H., Lu, M. Y., Almagro-Perez, C., Wagner, S. J., Vaidya, A. J., Chen, R. J., Williamson, D. F. K., Kim, A., and Mahmood, F. (2024).
Hest-1k: A dataset for spatial transcriptomics and histology image analysis.
In *Advances in Neural Information Processing Systems*.


kaiko.ai, Gatopoulos, I., Känzig, N., Moser, R., and Otálora, S. (2024).
eva: Evaluation framework for pathology foundation models.
In *Medical Imaging with Deep Learning*.


Ochi, M., Komura, D., Onoyama, T., and Ishikawa, S. (2024).
Pathology Images of Scanners and Mobilephones (PLISM) Dataset.

# References V



Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2023).
DINOv2: Learning Robust Visual Features without Supervision.
Version Number: 2.


Pignet, A., Klein, J., Robin, G., and Olivier, A. (2025).
Robust sensitivity control in digital pathology via tile score distribution matching.


Popat, S., Hubner, R., and Houlston, R. (2005).
Systematic Review of Microsatellite Instability and Colorectal Cancer Prognosis.
*Journal of Clinical Oncology*, 23(3):609–618.

# References VI



Rodrigo, M., Cuevas, C., and García, N. (2024).
Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks.
*Scientific Reports*, 14(1):21392.


Saillard, C., Dubois, R., Tchita, O., Loiseau, N., Garcia, T., Adriansen, A., Carpentier, S., Reyre, J., Enea, D., Von Loga, K., Kamoun, A., Rossat, S., Wiscart, C., Sefta, M., Auffret, M., Guillou, L., Fouillet, A., Kather, J. N., and Svrcek, M. (2023).
Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides.
*Nature Communications*, 14(1):6695.


Schlötterer, C. and Harr, B. (2004).
Microsatellite Instability.
In *Encyclopedia of Life Sciences*. Wiley, 1 edition.

# References VII


Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. (2022).
ibot: Image bert pre-training with online tokenizer.
*International Conference on Learning Representations (ICLR)*.

Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., Fuchs, T., Fusi, N., Liu, S., and Severson, K. (2024).
Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology.
Version Number: 3.