Performance and Robustness Evaluation of Pathology Foundation Models

Fredrik K. Gustafsson
University of Oxford
Department of Engineering Science
www.fregu856.com

April 29, 2026

About Me

Postdoc in the group of David Clifton at the University of Oxford, since July 2025.

Background:

- Dec 2023 - Jun 2025: Postdoc at Karolinska Institutet in Stockholm, ML/CV for *computational pathology*, in the group of Mattias Rantalainen.
- 2018 - 2023: PhD in *Machine Learning*, Uppsala University.
  - Thesis: *Towards Accurate and Reliable Deep Regression Models*.
    Supervisors: Thomas Schön & Martin Danelljan.
- 2013 - 2018: BSc & MSc in *Electrical Engineering*, Linköping University.
  - 2016 - 2017: Graduate exchange student, Stanford University.

Machine learning for healthcare with a particular focus on biosignals and wearables.

More broadly, I am interested in how to build and evaluate *reliable machine learning* models, for applications within *data-driven medicine and healthcare*.

About the Presentation

I will briefly describe three papers where we evaluate the performance and *robustness* of a range of recent pathology foundation models:

I: Benchmarking Pathology Foundation Models for Breast Cancer Survival Prediction
Fredrik K. Gustafsson, Constance Boissin, Johan Vallon-Christersson, David A. Clifton, Mattias Rantalainen
Preprint, 2026

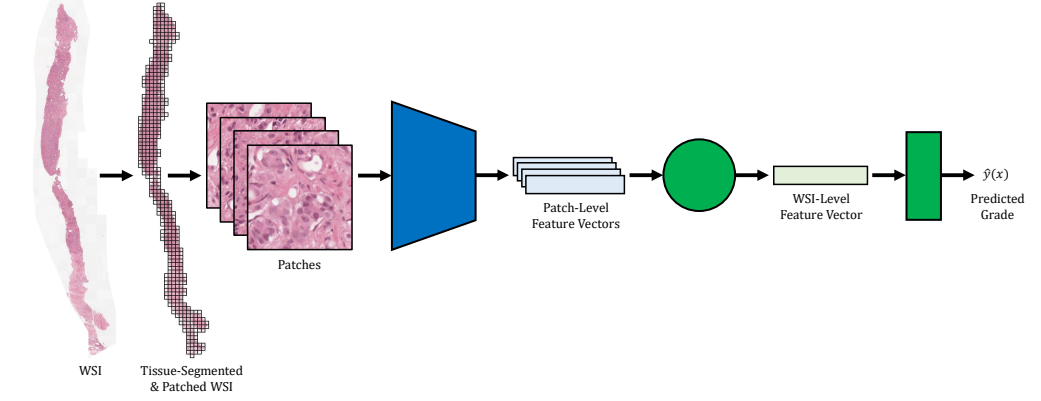
II: Evaluating Computational Pathology Foundation Models for Prostate Cancer Grading under Distribution Shifts
Fredrik K. Gustafsson, Mattias Rantalainen
Preprint, 2026

III: Scanner-Induced Domain Shifts Undermine the Robustness of Pathology Foundation Models
Erik Thiringer, Fredrik K. Gustafsson, Kajsa Ledesma Eriksson, Mattias Rantalainen
Preprint, 2026

Computational Pathology

Computational pathology uses machine learning and computer vision to automatically extract useful information from histopathology whole-slide images (WSIs).

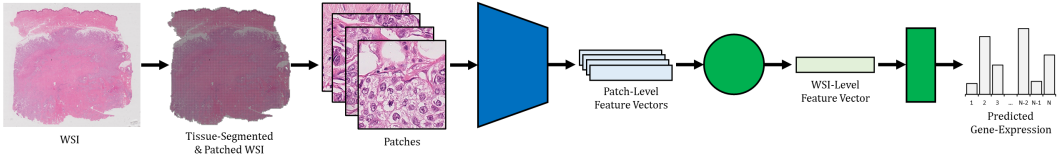
Given datasets of (WSI, label) pairs, models can be trained for applications such as *histological grading*, patient outcome prediction, and prediction of various biomarkers.



Computational Pathology

Computational pathology uses machine learning and computer vision to automatically extract useful information from histopathology whole-slide images (WSIs).

Given datasets of (WSI, label) pairs, models can be trained for applications such as histological grading, patient outcome prediction, and *prediction of various biomarkers*.



Pathology Foundation Models (PFMs)

Foundation models are large models trained on *large amounts of unlabeled data* using *self-supervised learning*. They are intended to be general-purpose feature extractors.

Self-supervised learning enables models to be trained on "raw" unlabeled data. Large collections of unlabeled WSIs – *WSIs without known clinical info, patient outcomes or any other type of annotations* – can thus be directly utilized in model training.

Has recently become a popular research direction within computational pathology:

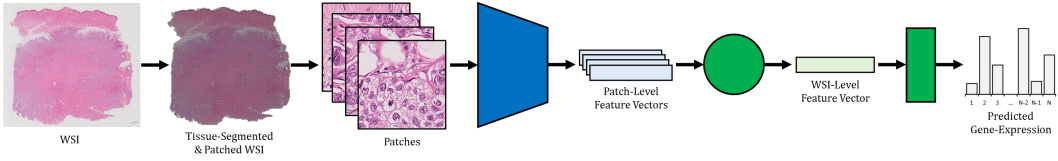
UNI: **Towards a General-Purpose Foundation Model for Computational Pathology**
Nature Medicine, 2024

Prov-GigaPath: **A Whole-Slide Foundation Model for Digital Pathology from Real-World Data**
Nature, 2024

Virchow: **A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection**
Nature Medicine, 2024

⋮

Typical Workflow for Application of PFMs



- Tissue-segment each WSI and divide it into image patches (e.g. 224×224 pixels).
- Use a *frozen* foundation model to extract feature vectors for all images patches in each WSI (*typical range: 5,000 - 25,000 image patches per WSI*).
- Train a small model that, for each WSI, takes the extracted patch-level feature vectors as input and outputs a WSI-level prediction (standard supervised training).

Paper I: Benchmarking PFMs for Survival Prediction

Benchmarking Pathology Foundation Models for Breast Cancer Survival Prediction

Fredrik K. Gustafsson, Constance Boissin, Johan Vallon-Christersson, David A. Clifton, Mattias Rantalainen

Preprint, 2026

Model Name	Architecture	Size	Feature Dimension	Pretraining Data
Resnet-IN [7]	ResNet-50	25M	1024	1.3M natural images
CTransPath [22]	CNN + Swin-T	22M	768	30K WSIs
RetCCL [23]	ResNet-50	25M	2048	32K WSIs
UNI [2]	ViT-L	307M	1024	100K WSIs
UNI2-h [13]	ViT-H	682M	1536	350K WSIs
H-optimus-0 [18]	ViT-G	1.1B	1536	500K WSIs
H-optimus-1 [1]	ViT-G	1.1B	1536	1M WSIs
Prov-GigaPath [25]	ViT-G	1.1B	1536	170K WSIs
Virchow [21]	ViT-H	632M	2560	1.5M WSIs
Virchow2 [26]	ViT-H	632M	2560	3.1M WSIs
H0-mini [5]	ViT-B	86M	768	500K + 6K WSIs
CONCH [11]	ViT-B	86M	512	21K WSIs + 1.1M image-text pairs
CONCHv1.5 [3]	ViT-L	307M	768	N/A

Rank	Model Name	Model Ranks	Mean Model Rank (↓)
1	H-optimus-1	1,1,1,3	1.5
2	H0-mini	2,2,4,5	3.25
3	H-optimus-0	5,5,5,3	4.5
4	CONCHv1.5	8,8,2,1	4.75
5	UNI2-h	4,4,6,6	5
6	Virchow2	3,3,7,8	5.25
7	Virchow	10,10,2,2	6
8	CONCH	6,6,8,7	6.75
9	Prov-GigaPath	6,6,9,9	7.5
10	UNI	9,9,10,10	9.5
11	CTransPath	12,12,11,11	11.5
12	RetCCL	11,11,13,13	12
13	Resnet-IN	13,13,12,12	12.5

I: Benchmarking PFMs for Survival Prediction

We evaluate 13 models: a natural-image baseline, two early pathology-specific models, seven state-of-the-art PFMs, a compact distilled PFM, and two vision-language PFMs.

Downstream survival models are trained on a dataset of *2,315 breast cancer patient*s (SöS-BC-4) and evaluated on two independent external datasets (KS-Solna and SCAN-B-Lund) comprising *3,119 patients* in total.

Models are evaluated under four settings: RFS and PFS, each assessed both for the full cohort ('All Patients') and for the clinically relevant 'ER+ & HER2-' patient subgroup.

The combined evaluation set of 3,119 patients contains 615 RFS events and 233 PFS events, with 2,524 patients (80.9% of the full set), 475 RFS events (77.2%) and 157 PFS events (67.4%) in the 'ER+ & HER2-' patient subgroup.

I: Benchmarking PFMs for Survival Prediction - Main Results

Rank	Model Name	C-index (↑)
1	H-optimus-1	0.678 (0.656 – 0.700)
2	H0-mini	0.676 (0.653 – 0.698)
3	Virchow2	0.675 (0.653 – 0.698)
4	UNI2-h	0.667 (0.644 – 0.689)
5	H-optimus-0	0.664 (0.642 – 0.686)
6	CONCH	0.663 (0.641 – 0.685)
6	Prov-GigaPath	0.663 (0.640 – 0.686)
8	CONCHv1.5	0.660 (0.637 – 0.683)
9	UNI	0.657 (0.635 – 0.680)
10	Virchow	0.656 (0.633 – 0.679)
11	RetCCL	0.645 (0.622 – 0.668)
12	CTransPath	0.632 (0.609 – 0.655)
13	Resnet-IN	0.612 (0.588 – 0.635)

(a) **RFS – All Patients**.

Rank	Model Name	C-index (↑)
1	H-optimus-1	0.670 (0.645 – 0.695)
2	H0-mini	0.668 (0.642 – 0.693)
3	Virchow2	0.665 (0.639 – 0.690)
4	UNI2-h	0.663 (0.637 – 0.688)
5	H-optimus-0	0.660 (0.633 – 0.685)
6	CONCH	0.657 (0.631 – 0.681)
6	Prov-GigaPath	0.657 (0.631 – 0.682)
8	CONCHv1.5	0.651 (0.625 – 0.676)
9	UNI	0.648 (0.622 – 0.674)
10	Virchow	0.639 (0.613 – 0.665)
11	RetCCL	0.636 (0.609 – 0.662)
12	CTransPath	0.627 (0.601 – 0.652)
13	Resnet-IN	0.600 (0.572 – 0.627)

(b) **RFS – Patient Subgroup (ER+ & HER2-)**.

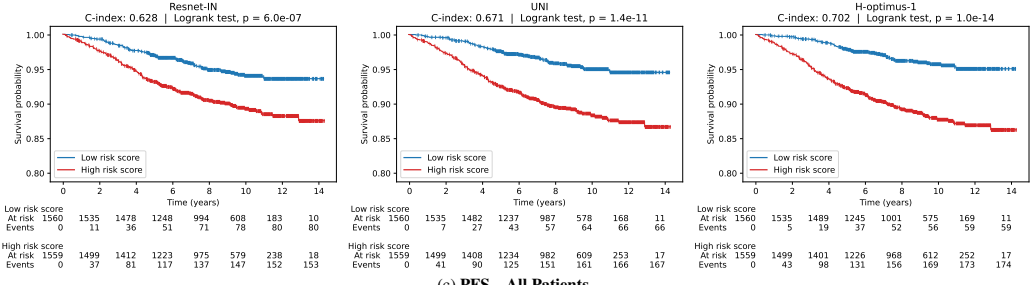
Rank	Model Name	C-index (↑)
1	H-optimus-1	0.702 (0.666 – 0.737)
2	Virchow	0.695 (0.660 – 0.729)
2	CONCHv1.5	0.695 (0.660 – 0.729)
4	H0-mini	0.692 (0.655 – 0.728)
5	H-optimus-0	0.686 (0.651 – 0.722)
6	UNI2-h	0.683 (0.647 – 0.719)
7	Virchow2	0.681 (0.645 – 0.718)
8	CONCH	0.675 (0.639 – 0.710)
9	Prov-GigaPath	0.673 (0.636 – 0.710)
10	UNI	0.671 (0.634 – 0.708)
11	CTransPath	0.647 (0.609 – 0.686)
12	Resnet-IN	0.628 (0.590 – 0.666)
13	RetCCL	0.627 (0.589 – 0.666)

(c) **PFS – All Patients**.

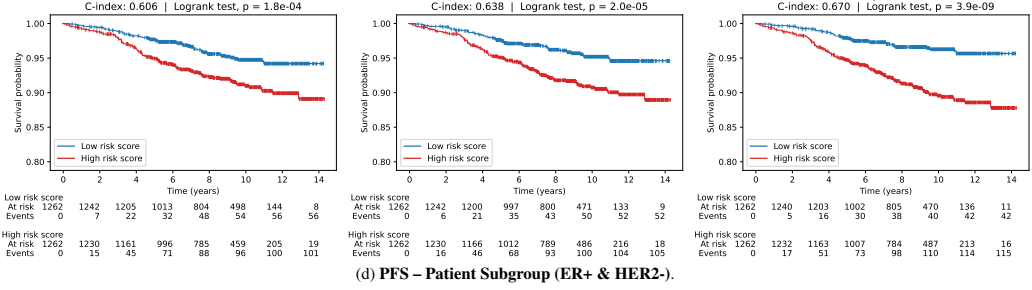
Rank	Model Name	C-index (↑)
1	CONCHv1.5	0.680 (0.636 – 0.721)
2	Virchow	0.678 (0.634 – 0.720)
3	H-optimus-1	0.670 (0.626 – 0.715)
3	H-optimus-0	0.670 (0.624 – 0.714)
5	H0-mini	0.666 (0.619 – 0.711)
6	UNI2-h	0.657 (0.611 – 0.702)
7	CONCH	0.656 (0.612 – 0.700)
8	Virchow2	0.655 (0.608 – 0.700)
9	Prov-GigaPath	0.641 (0.594 – 0.687)
10	UNI	0.638 (0.592 – 0.683)
11	CTransPath	0.622 (0.573 – 0.670)
12	Resnet-IN	0.606 (0.557 – 0.655)
13	RetCCL	0.604 (0.555 – 0.653)

(d) **PFS – Patient Subgroup (ER+ & HER2-)**.

I: Benchmarking PFMs for Survival Prediction - Risk Stratification

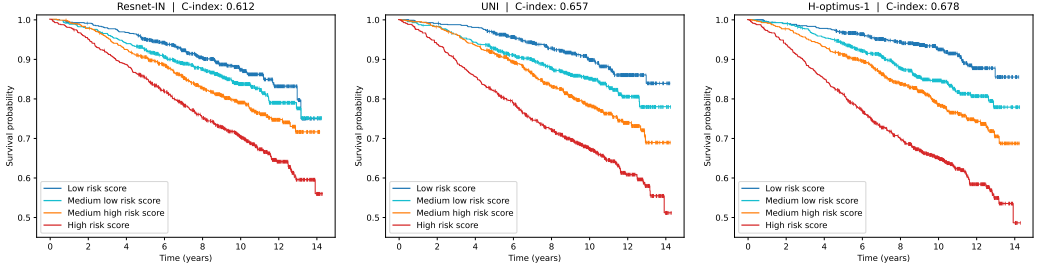


(c) **PFS – All Patients**.

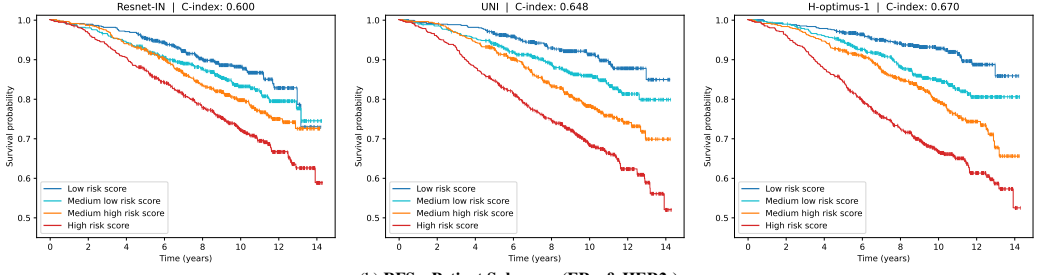


(d) **PFS – Patient Subgroup (ER+ & HER2-)**.

I: Benchmarking PFMs for Survival Prediction - Risk Stratification

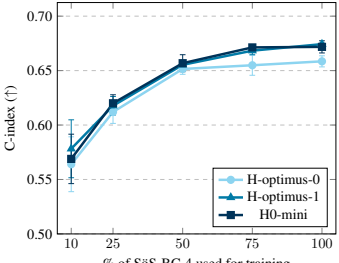


(a) **RFS – All Patients**.

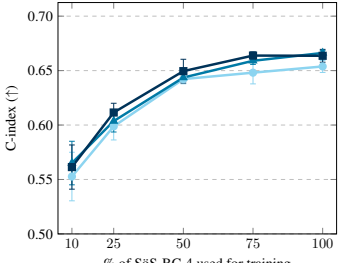


(b) **RFS – Patient Subgroup (ER+ & HER2-)**.

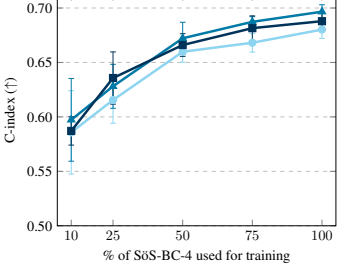
I: Benchmarking PFMs for Survival Prediction - Detailed Comparison



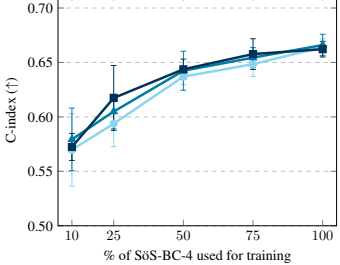
(a) **RFS – All Patients**.



(b) **RFS – Patient Subgroup (ER+ & HER2-)**.

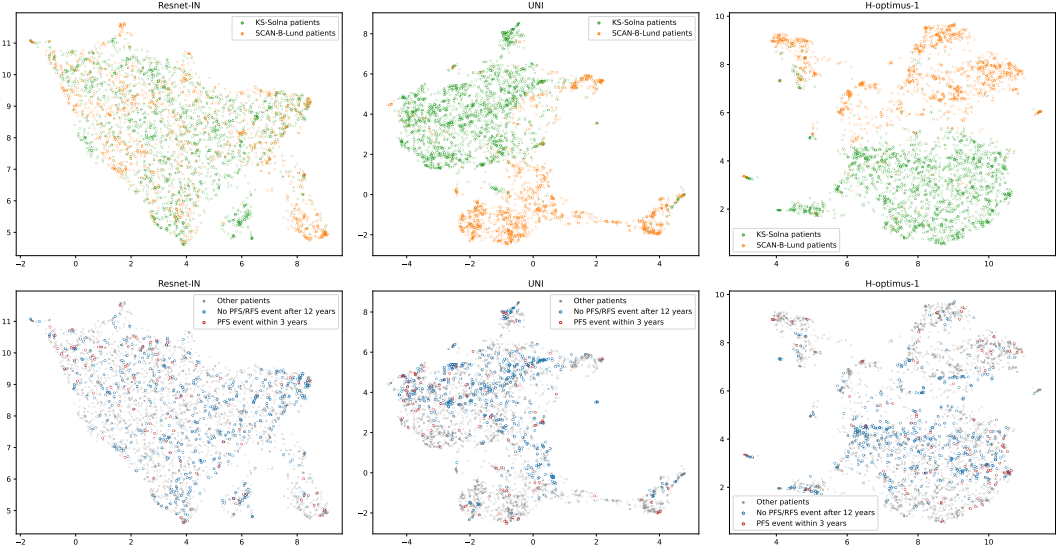


(c) **PFS – All Patients**.



(d) **PFS – Patient Subgroup (ER+ & HER2-)**.

I: Benchmarking PFMs for Survival Prediction - Feature Space Analysis



I: Benchmarking PFMs for Survival Prediction - Main Takeaways

(1/4) H-optimus-1 achieves the strongest overall performance, but absolute differences between top-performing PFMs are small and confidence intervals substantially overlap.

(2/4) Across model families, consistent generational improvements are observed, with second-generation PFMs (H-optimus-1, CONCHv1.5, UNI2-h, Virchow2) outperforming their first-generation counterparts. However, these gains remain modest despite substantial increases in pretraining data scale, suggesting diminishing returns from scaling alone.

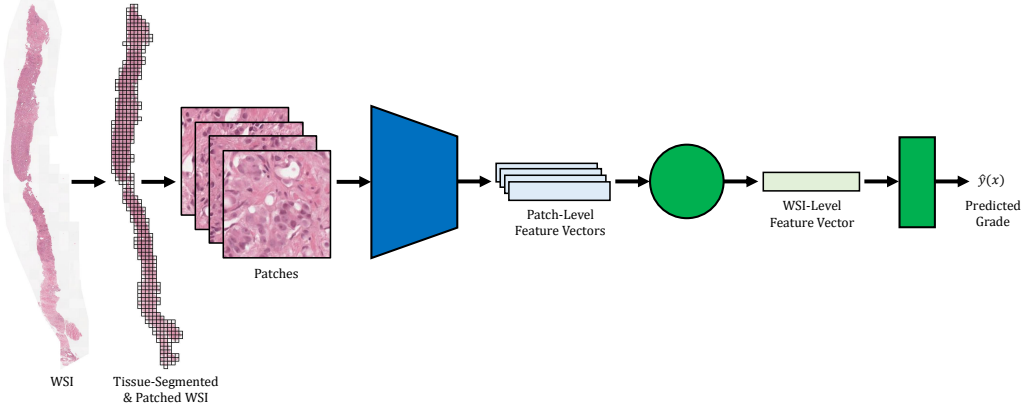
(3/4) Model size is not a reliable predictor of performance, as smaller and more efficient models can match or exceed much larger architectures, emphasizing the importance of training strategy and pretraining data quality over model scaling.

(4/4) The distilled H0-mini achieves the second-best overall ranking and slightly outperforms its teacher model H-optimus-0, while using less than 8% of the parameters and enabling much faster feature extraction. Knowledge distillation can thus yield highly efficient PFMs without sacrificing performance, making it a particularly promising approach for practical deployment.

Evaluating Computational Pathology Foundation Models for Prostate Cancer Grading under Distribution Shifts

Fredrik K. Gustafsson, *Mattias Rantalainen*

Preprint, 2026

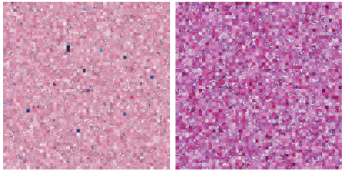


II: Evaluating PFMs under Distribution Shifts

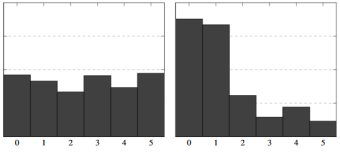
The PANDA dataset was collected from two different sites: *Radboud* University Medical Center in the Netherlands, and *Karolinska* Institutet in Sweden.

Radboud and Karolinska differ in terms of both the pathology lab procedures and utilized scanners, creating a clear distribution shift for the WSI image data.

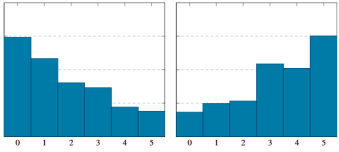
By creating further subsets of the PANDA dataset, we are also able to evaluate robustness in terms of shifts in the label distribution over the ISUP grades 0 - 5.



(a) WSI image data shift, *Radboud* → *Karolinska*.

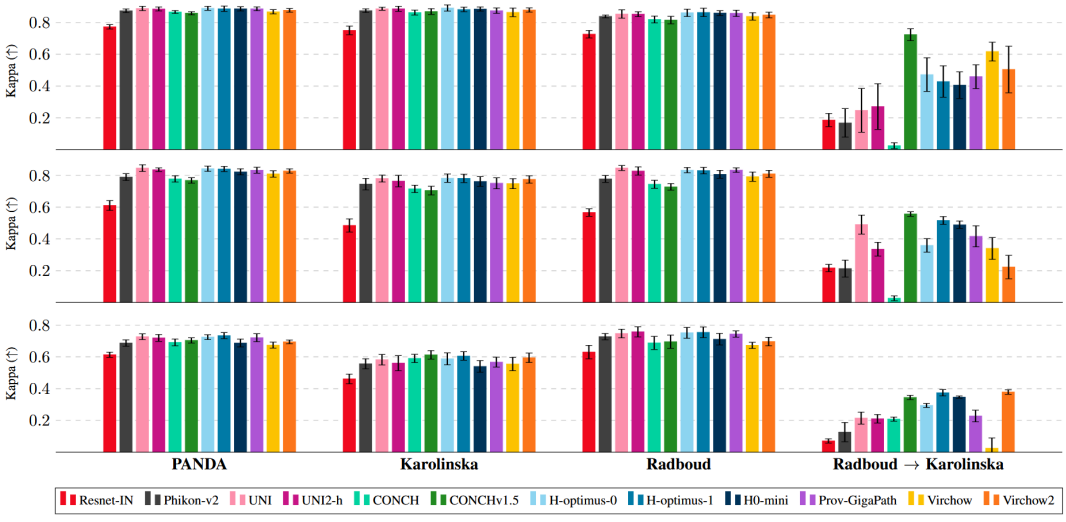


(b) Grade label shift, *Radboud* → *Karolinska*.

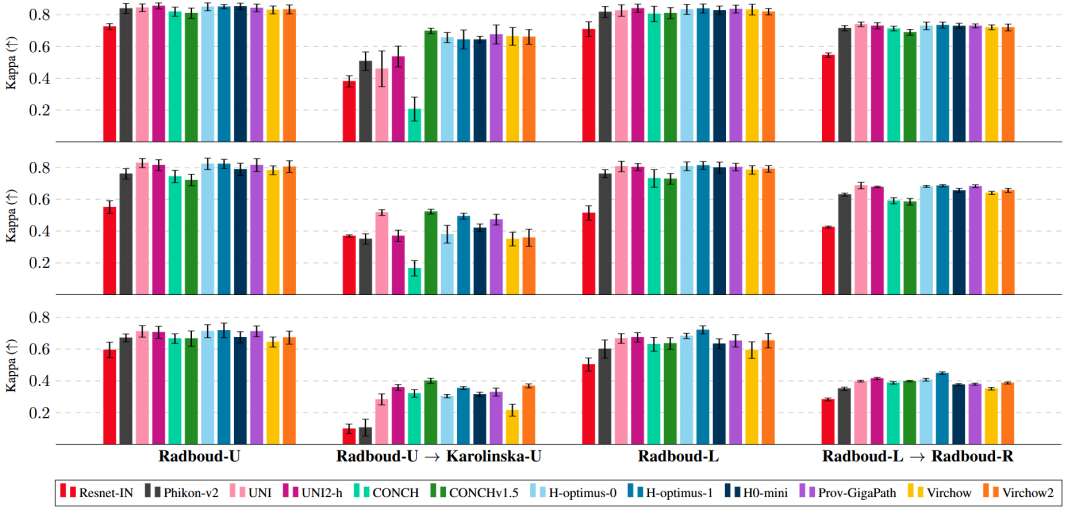


(c) Grade label shift, *Radboud-L* → *Radboud-R*.

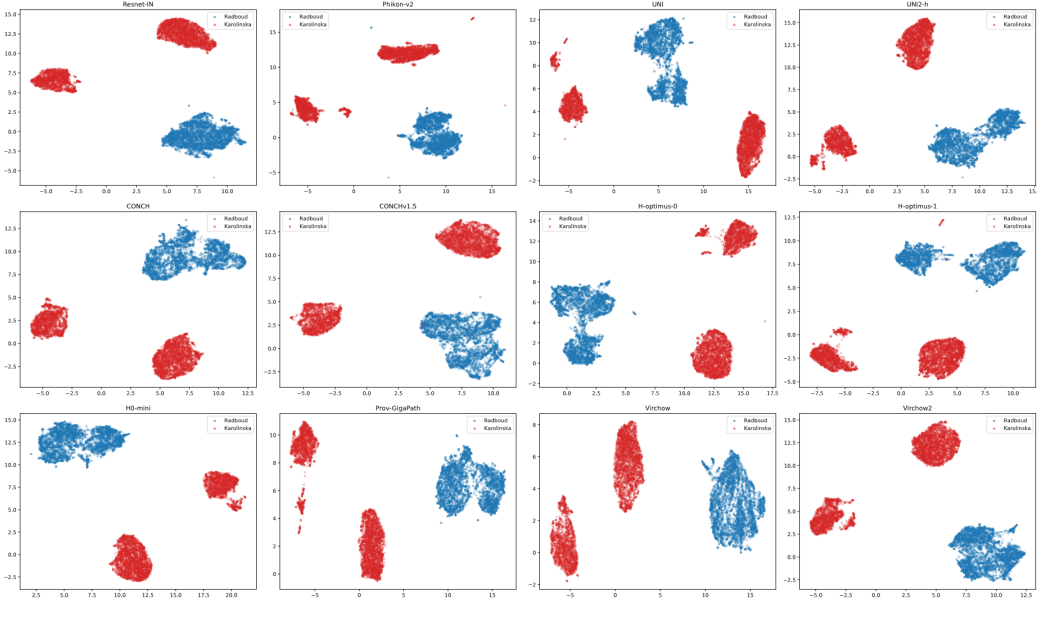
II: Evaluating PFMs under Distribution Shifts - Results



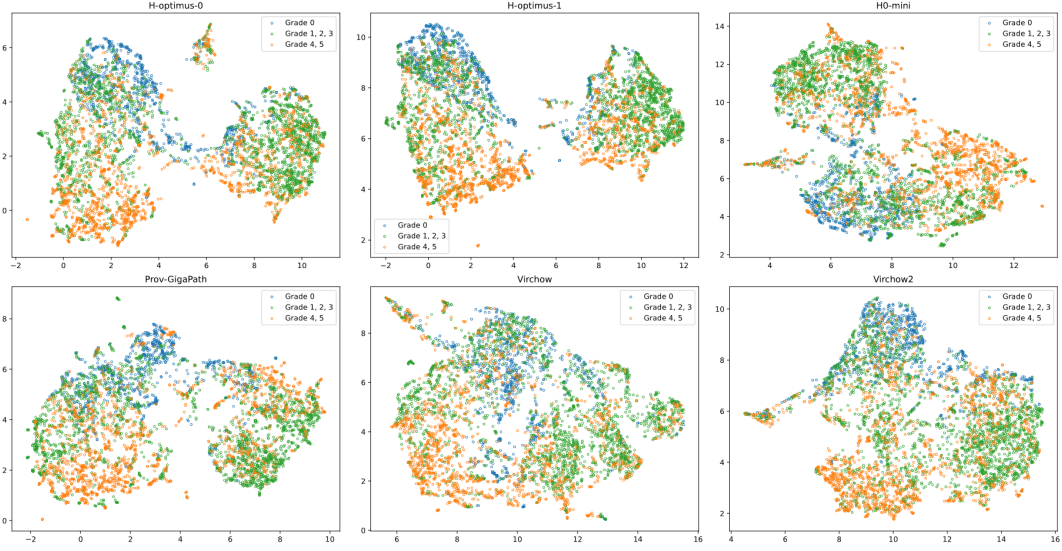
II: Evaluating PFMs under Distribution Shifts - Results



II: Evaluating PFMs under Distribution Shifts - Feature Space Analysis



II: Evaluating PFMs under Distribution Shifts - Feature Space Analysis



II: Evaluating PFMs under Distribution Shifts - Main Takeaways

(1/5) PFMs substantially outperform a natural-image baseline for prostate cancer grading, but absolute performance under cross-site shifts can still deteriorate severely.

(2/5) Large-scale pretraining of PFMs on diverse datasets does not guarantee downstream cross-site generalization, and increasing model size or pretraining data alone does not reliably improve robustness.

(3/5) PFMs are significantly more sensitive to image-domain shifts than to label distribution shifts; acquisition-related variability seems to be the dominant challenge.

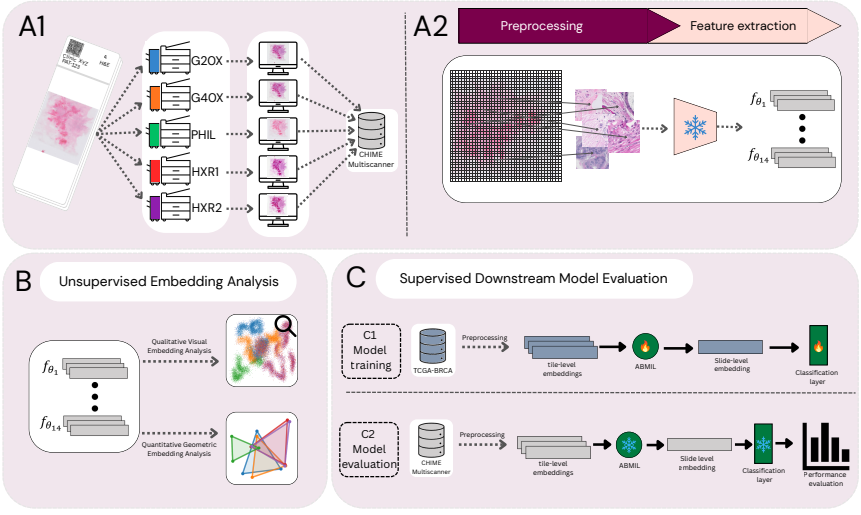
(4/5) Feature space analysis reveals persistent domain separation, suggesting that current PFMs encode site- and scanner-specific variation more strongly than pathology-relevant signals.

(5/5) Strong PFMs and high-quality downstream data complement rather than replace each other: even in the foundation-model era, the quality and diversity of the data used to train downstream prediction models remains critical for reliable deployment.

Scanner-Induced Domain Shifts Undermine the Robustness of Pathology Foundation Models

Erik Thiringer, Fredrik K. Gustafsson, Kajsa Ledesma Eriksson, Mattias Rantalainen

Preprint, 2026

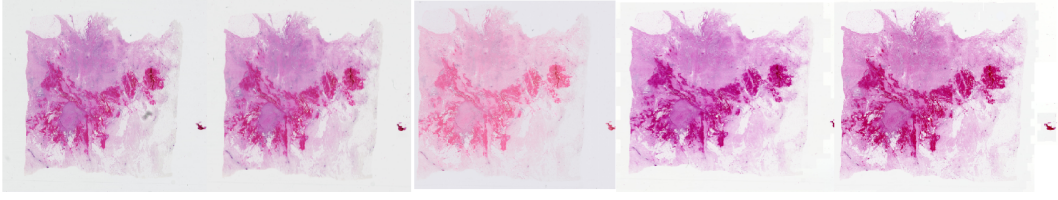


III: PFM Scanner-Variability Robustness

We systematically evaluate the robustness of 13 PFMs to scanner-induced variability.

Using a controlled multiscanner dataset comprising 384 breast cancer WSIs from as many patients, *scanned on five different whole-slide scanners*, we isolate scanner-induced effects from all other sources of variation.

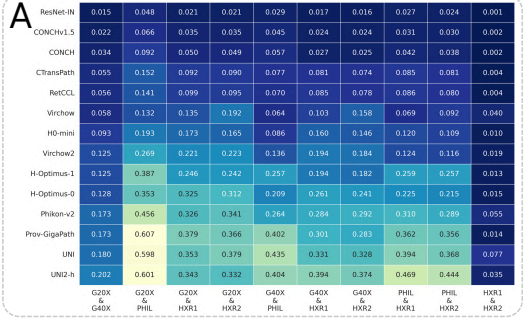
Robustness is assessed through complementary unsupervised analyses of the feature embedding space and a set of clinicopathological supervised prediction tasks.

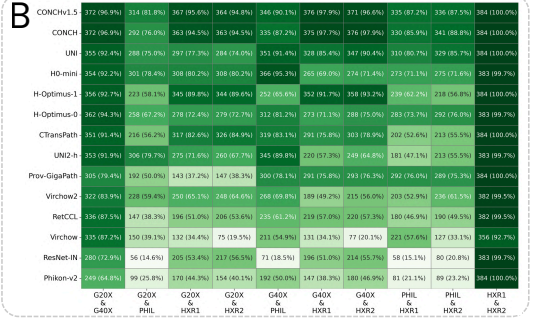


III: PFM Scanner-Variability Robustness - Patch-level Feature Space

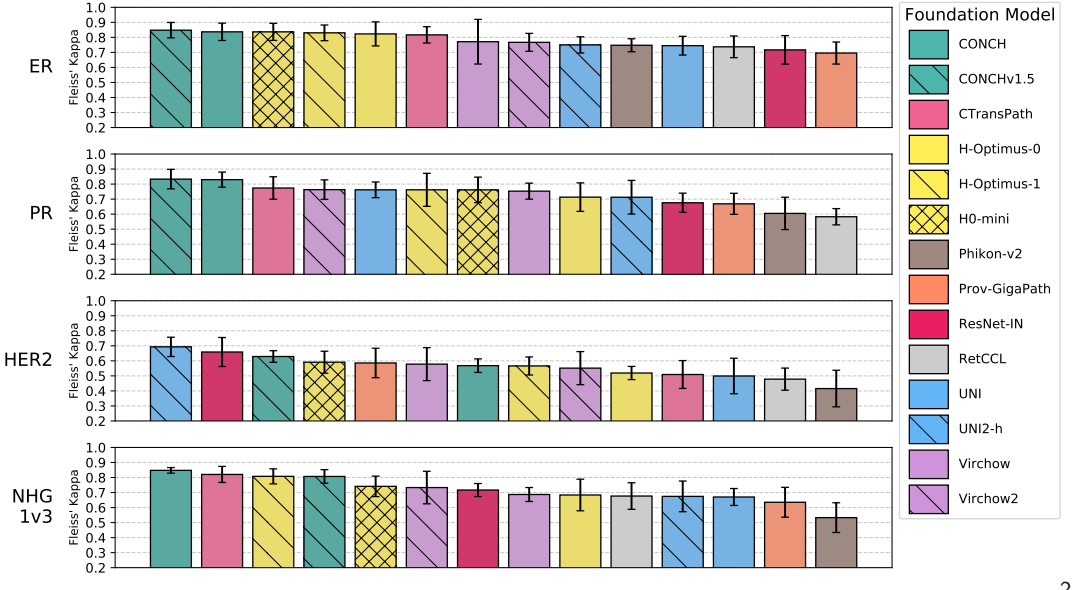
III: PFM Scanner-Variability Robustness - Slide-level Feature Space

III: PFM Scanner-Variability Robustness - Quantitative Feature Analysis





III: PFM Scanner-Variability Robustness - Prediction Consistency



III: PFM Scanner-Variability Robustness - Main Takeaways

(1/5) Current state-of-the-art PFMs are not invariant to scanner-induced domain shifts. Most models encode pronounced scanner-specific variability in their feature embedding space.

(2/5) Although predictive performance largely remains similar when measured by AUC, this masks an important failure mode: scanner variability systematically alters the embedding space and impacts calibration of downstream model predictions, resulting in scanner-dependent bias.

(3/5) Robustness is not a simple function of training data scale, model size, or model recency.

(4/5) The models trained on the most diverse data, vision-language models, appear to have an advantage with respect to robustness. More systematic analysis is however needed, comparing vision-only and multimodal pretraining under controlled settings.

(5/5) Targeted robustness-oriented training strategies appear promising, as the distilled model H0-mini consistently outperforms its larger teacher model H-optimus-0 across multiple aspects of embedding stability and downstream prediction consistency.

Questions?

Fredrik K. Gustafsson
University of Oxford
fredrik.gustafsson@eng.ox.ac.uk
www.fregu856.com