Building and Evaluating Computational Pathology Foundation Models for Breast Cancer

Fredrik K. Gustafsson
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics, Rantalainen Group
www.fregu856.com

Annual SeRC Meeting
May 15, 2025

About the Presentation

I'm a postdoc in the group of Mattias Rantalainen at Karolinska Institutet (Department of Medical Epidemiology and Biostatistics), since December 2023.

Will give a snapshot of some ongoing work in the group, focusing on how we train and evaluate *computational pathology foundation models* on in-house breast cancer data.

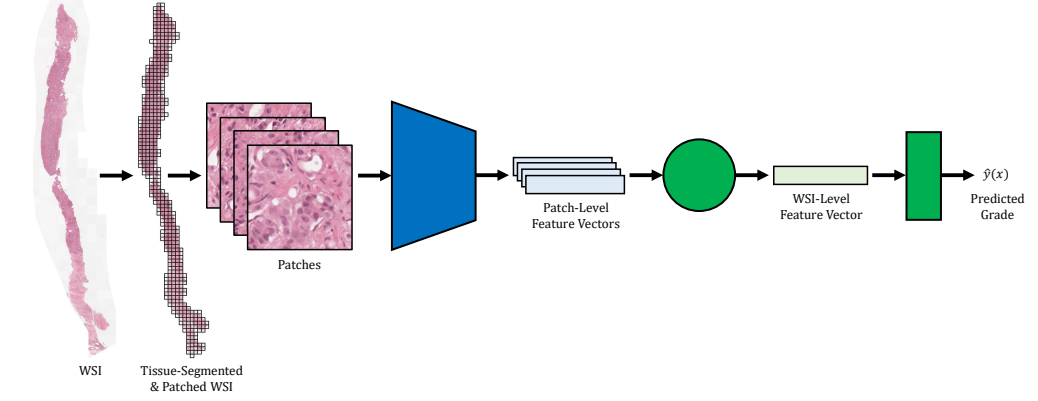
Rantalainen Group:

Mattias Rantalainen.

Constance Boissin, Kajsa Ledesma Eriksson, Bojing Liu, Francisco J. Peña, Abhinav Sharma, Erik Thiringer, Duong Tran, Yujie Xiang, Anne-May Österholm.

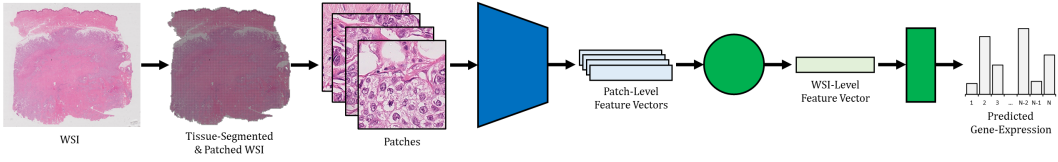
Computational pathology uses machine learning and computer vision to automatically extract useful information from histopathology whole-slide images (WSIs).

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 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*.



Foundation Models

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

⋮

Details for two recent computational pathology (CPATH) foundation models:

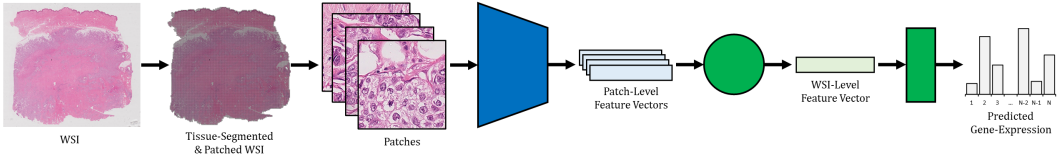
UNI:

- Pretrained using self-supervised learning (DINOv2) on a *pan-cancer dataset* (20 major tissue types) of *100 million tissue patches* from more than *100,000 WSIs*.
- Most WSIs are collected from the Massachusetts General Hospital and Brigham and Women's Hospital in Boston, USA.
- Vision transformer ViT-Large model, 303 million parameters.

Virchow:

- Pretrained using self-supervised learning (DINOv2) on a *pan-cancer dataset* (17 major tissue types) of *2 billion patches* from more than *1.4 million WSIs*.
- WSIs are collected from the Memorial Sloan Kettering Cancer Center (New York, USA), from more than *119,000 patients*.
- ViT-Huge model, 632 million parameters.

Using CPATH Foundation Models



Typical workflow:

- 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).

Observation:
In various downstream breast cancer-related tasks, CPATH foundation models trained on pan-cancer histopathology image data (e.g. UNI & Virchow) significantly outperform regular foundation models trained on natural images.

Hypothesis:
In various downstream breast cancer-related tasks, *tissue-specific* foundation models trained on *breast-specific* histopathology image data will outperform pan-cancer CPATH foundation models (e.g. UNI & Virchow).

Approach:
Train ViT model using DINOv2 on an *in-house dataset of more than 60,000 WSIs from Swedish breast cancer patients*, compare with UNI and other pan-cancer models.

Training CPATH Foundation Models for Breast Cancer - Details

Dataset details:

- More than *60,000 WSIs*, *100 TB* of data in total.
- More than *1.1 billion* 224×224 image patches after preprocessing.
- Data from six different sites, 80% of WSIs are from Kalmar, Örebro or Jönköping.
- Only H&E-stained WSIs (future work: Utilize the IHC-stained WSIs as well).
- All WSIs have been digitized/scanned in-house.

First approach: Randomly sample 1,000 tissue patches per WSI to create the final training dataset, resulting in more than *60 million* image patches (800 GB of data).

Initial experiments are ongoing, training ViT-Large and ViT-Base models.

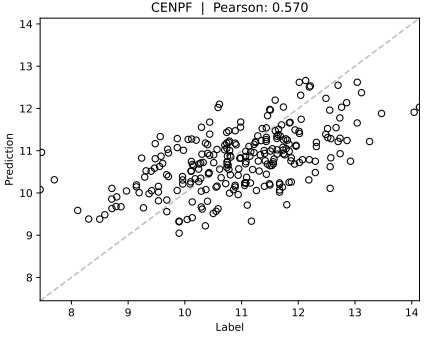
Evaluating CPATH Foundation Models for Breast Cancer (1/3)

Two examples of tasks we are benchmarking public CPATH foundation models on:

Benchmarking task 1: **Gene-expression prediction**.

Using the public TCGA-BRCA dataset, containing WSIs and corresponding gene-expression labels of *20,000 genes* for more than *1,000 patients*. Train model to predict all 20,000 genes, evaluate on subset of *50 breast cancer-related genes (PAM50)*.

Rank	Model name	PAM50 mean Pearson (↑)
1	H-optimus-1	0.595±0.016
2	H0-mini	0.591±0.014
3	H-optimus-0	0.587±0.012
4	UNI2-h	0.583±0.016
5	Virchow2	0.582±0.019
6	CONCHv1.5	0.576±0.021
7	CONCH	0.574±0.019
8	Prov-GigaPath	0.571±0.009
9	Virchow	0.563±0.020
10	UNI	0.562±0.026
11	CTransPath	0.517±0.029
12	RetCCL	0.449±0.034
13	Resnet-IN	0.379±0.034

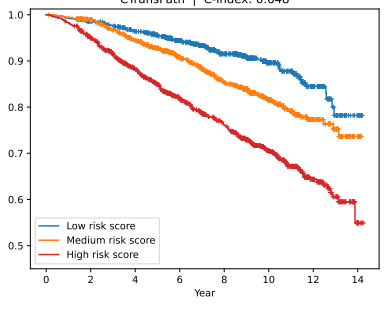


Benchmarking task 2: **Survival prediction**.

Using in-house datasets with WSIs and corresponding patient outcomes (overall survival + recurrence events) for Swedish breast cancer patients from 3 different sites.

- Training: *2,300 patients*, mean follow-up time of *7.5 years*, 350 events.
- Evaluation: *3,100 patients*, mean follow-up time of *7.6 years*, 510 events.

Rank	Model name	C-index (↑)
1	H0-mini	0.689±0.012
2	H-optimus-1	0.687±0.012
3	Virchow2	0.682±0.012
4	H-optimus-0	0.677±0.012
4	UNI2-h	0.677±0.012
6	CONCH	0.675±0.012
7	Prov-GigaPath	0.674±0.012
8	CONCHv1.5	0.672±0.012
9	Virchow	0.671±0.013
10	UNI	0.666±0.012
11	RetCCL	0.659±0.012
12	CTransPath	0.648±0.013
13	Resnet-IN	0.633±0.013

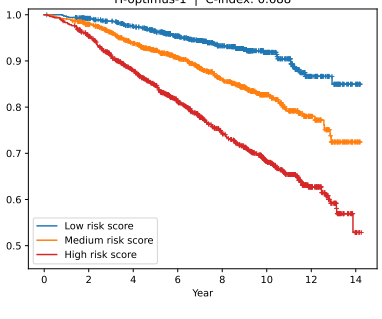


Benchmarking task 2: **Survival prediction**.

Using in-house datasets with WSIs and corresponding patient outcomes (overall survival + recurrence events) for Swedish breast cancer patients from 3 different sites.

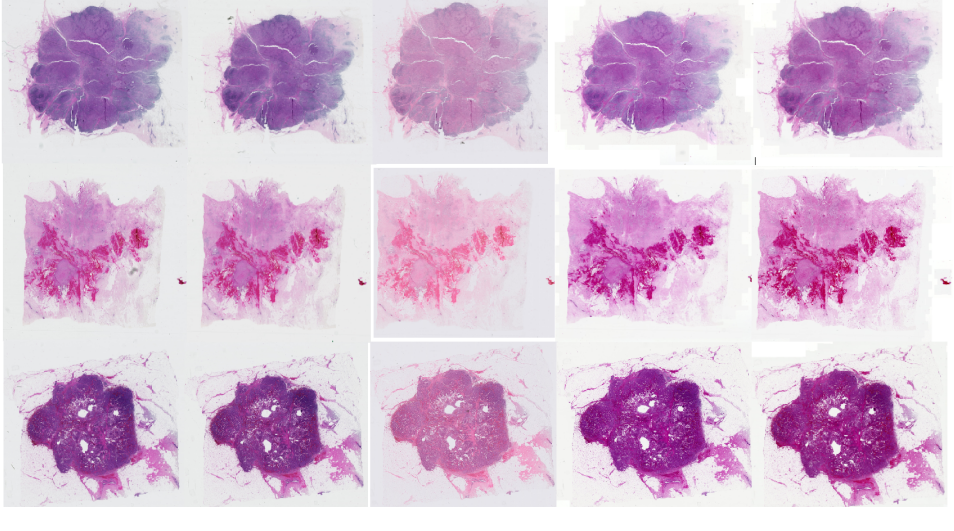
- Training: *2,300 patients*, mean follow-up time of *7.5 years*, 350 events.
- Evaluation: *3,100 patients*, mean follow-up time of *7.6 years*, 510 events.

Rank	Model name	C-index (↑)
1	H0-mini	0.689±0.012
2	H-optimus-1	0.687±0.012
3	Virchow2	0.682±0.012
4	H-optimus-0	0.677±0.012
4	UNI2-h	0.677±0.012
6	CONCH	0.675±0.012
7	Prov-GigaPath	0.674±0.012
8	CONCHv1.5	0.672±0.012
9	Virchow	0.671±0.013
10	UNI	0.666±0.012
11	RetCCL	0.659±0.012
12	CTransPath	0.648±0.013
13	Resnet-IN	0.633±0.013



Evaluating CPATH Foundation Models for Breast Cancer (3/3)

We are also evaluating **scanner-variability robustness** of CPATH foundation models, using an in-house dataset of WSIs digitized/scanned with five different scanners.



Main Takeaways

(1/4) We are training foundation models specifically for breast cancer using an *in-house dataset of more than 60,000 WSIs from Swedish breast cancer patients*.

(2/4) Our hypothesis is that such *tissue-specific* foundation models will outperform current pan-cancer models in important breast cancer-related CPATH applications.

(3/4) Being able to train in-house foundation models also enables the group to explore strategies for improved model robustness (different scanners, labs & hospitals), train models for IHC-stained WSIs, develop new self-supervised learning methods tailored for pathology data, study how performance scales with the model and dataset size, etc.

(4/4) Our in-house foundation models will serve as the backbone for various *breast cancer precision diagnostics* solutions developed by the group moving forward.

Contact & Acknowledgements

Fredrik K. Gustafsson
fredrik.gustafsson@ki.se
www.fregu856.com

Rantalainen Group:

Mattias Rantalainen.

Constance Boissin, Kajsa Ledesma Eriksson, Bojing Liu, Francisco J. Peña, Abhinav Sharma, Erik Thiringer, Duong Tran, Yujie Xiang, Anne-May Österholm.