On the Use and Evaluation of Computational Pathology Foundation Models for WSI-Based Prediction Tasks

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

The Scandinavian Seminar on Translational Pathology
November 23, 2024

About Me

Postdoctoral researcher in the group of Mattias Rantalainen at Karolinska Institutet (Department of Medical Epidemiology and Biostatistics), since December 2023.

Background:

- 2023: PhD in *Machine Learning*, Uppsala University.
  - Thesis: *Towards Accurate and Reliable Deep Regression Models*.
  - Supervisors: Thomas Schön & Martin Danelljan.
- 2018: MSc in *Electrical Engineering*, Linköping University.
- 2016: BSc in *Applied Physics and Electrical Engineering*, Linköping University.

Machine learning and computer vision for *computational pathology*.

My research focuses on how to build and evaluate *reliable machine learning* models, for applications within *data-driven medicine and healthcare*.

About the Presentation

Will give a short introduction to recent computational pathology foundation models and how they typically are used for WSI-level prediction tasks. Then, I will briefly describe three ongoing projects where we use and evaluate such foundation models:

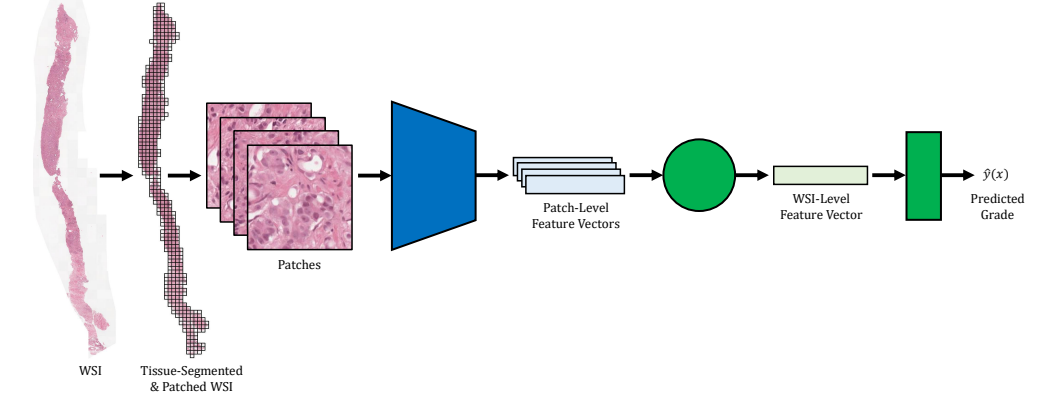
Evaluating Deep Regression Models for WSI-Based Gene-Expression Prediction
Fredrik K. Gustafsson, Mattias Rantalainen
Preprint, 2024-10

Evaluating Computational Pathology Foundation Models for Prostate Cancer Grading under Distribution Shifts
Fredrik K. Gustafsson, Mattias Rantalainen
Preprint, 2024-10

Benchmarking Scanner-Variability Robustness of Computational Pathology Foundation Models
Erik Thiringer, Fredrik K. Gustafsson, Mattias Rantalainen
Ongoing work

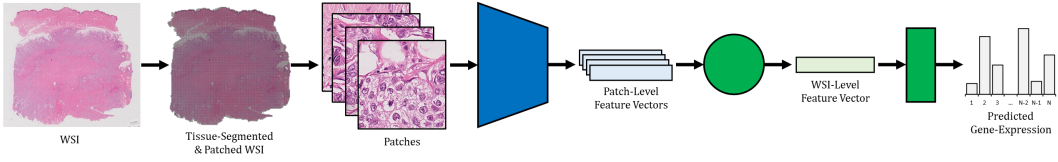
Computational pathology uses machine learning and computer vision to automatically extract useful information from histopathology whole-slide images (WSIs).

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



Foundation Models

Foundation models are large deep learning models (i.e., *very* large machine learning models) trained on *large amounts of unlabeled data* using *self-supervised learning*.

They are intended to be *general-purpose feature extractors*, promising to achieve good performance on a wide range of downstream prediction tasks.

Have recently become a popular research direction within computational pathology:

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

A Visual-Language Foundation Model for Computational Pathology
Nature Medicine, 2024

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

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

Computational Pathology Foundation Models (1/2)

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

- 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. Extracts patch-level feature vectors of dimension 1024.

Prov-GigaPath:

A Whole-Slide Foundation Model for Digital Pathology from Real-World Data (Nature, 2024).

- Pretrained using self-supervised learning (DINOv2) on a pan-cancer dataset (31 major tissue types) of *1.3 billion patches* from more than *171,000 WSIs*.
  - WSIs are collected from Providence, *"a large US health network comprising 28 cancer centers"*, from more than *30,000 patients*.
- ViT-Giant model, 1.1 billion parameters. Extracts feature vectors of dim 1536.

Computational Pathology Foundation Models (2/2)

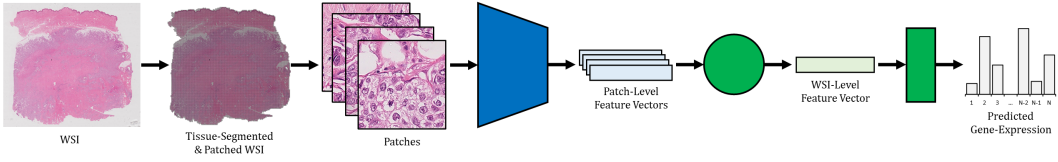
Virchow: *A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection* (Nature Medicine, 2024).

- 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. Extracts feature vectors of dim 2560.

CONCH: Vision-language foundation model.

A Visual-Language Foundation Model for Computational Pathology (Nature Medicine, 2024).

- First pretrained using self-supervised learning on a dataset of *16 million tissue patches* from more than *21,000 WSIs*. Then further pretrained using a vision-language objective on a dataset of more than *1.1 million image-caption pairs* (curated via processing of figures from PubMed articles).
- ViT-Base model, 86 million parameters. Extracts feature vectors of dim 512.



Typical workflow:

- Tissue-segment each WSI and divide it into image patches (e.g. 256×256 pixels).
- Use a *frozen* foundation model to extract feature vectors for all images patches in each WSI (typical range: 1,000 - 20,000 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).

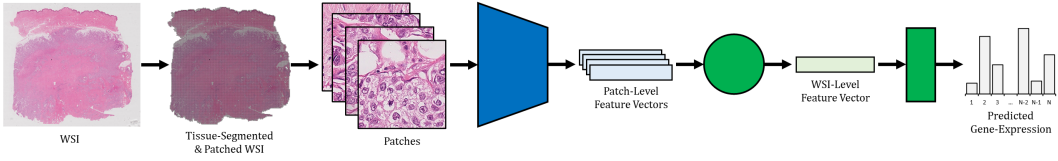
The feature extraction can be quite slow (12 - 24 hours for 1,000 WSIs).

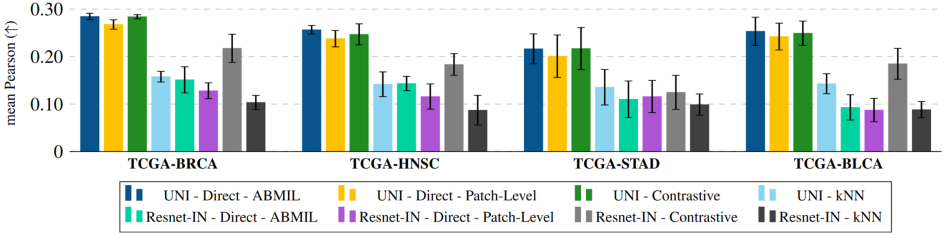
However, once the feature vectors are extracted, the actual model training typically takes less than 30 minutes to run, on a single relatively small GPU.

Evaluating Deep Regression Models for WSI-Based Gene-Expression Prediction

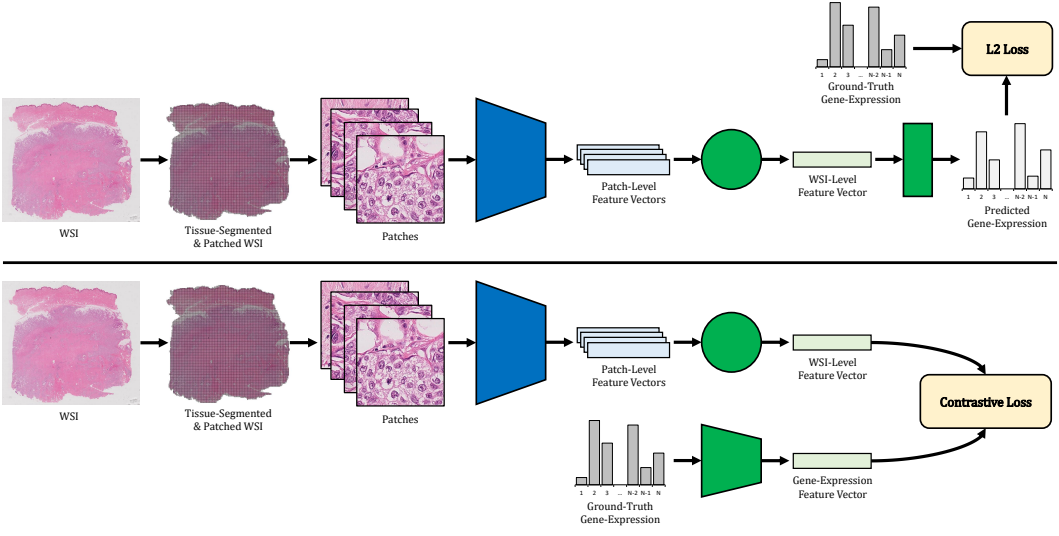
Fredrik K. Gustafsson, Mattias Rantalainen

Preprint, 2024-10

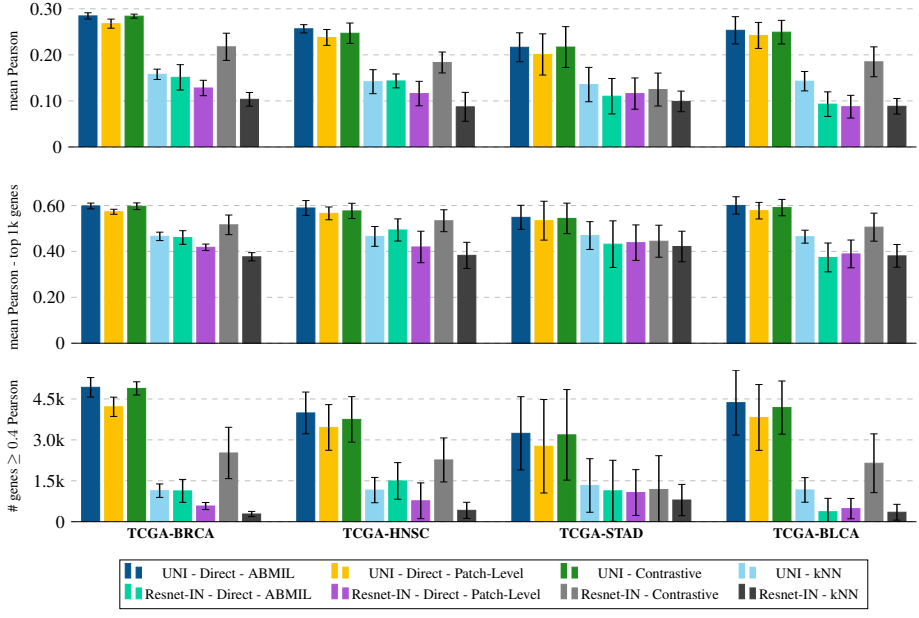




Project I: Gene-Expression Prediction - Regression Models



Project I: Gene-Expression Prediction - Results



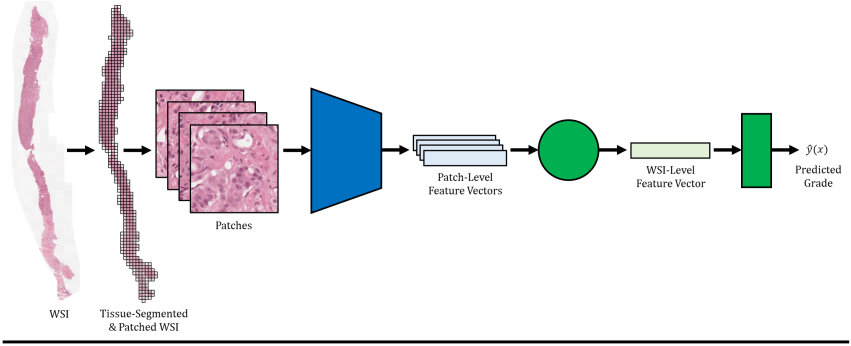
(1/4) Training regression models on top of UNI features gives accurate WSI-based models for gene-expression prediction (TCGA-BRCA: 4,927 genes with Pearson correlation ≥ 0.4, mean Pearson of 0.56 for PAM50 genes).

(2/4) Despite conceptual differences, *Direct* - *ABMIL* and *Contrastive* achieve very similar performance and should both be considered go-to regression models.

(3/4) Training a single model to regress all $N = 20\,530$ genes is a computationally efficient and very strong baseline, this should be the starting point given new datasets.

(4/4) Training one model for each individual gene incurs an extremely high computational cost yet achieves comparatively low regression accuracy.

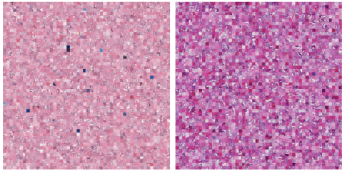
Evaluating Computational Pathology Foundation Models for Prostate Cancer Grading under Distribution Shifts. *Fredrik K. Gustafsson, Mattias Rantalainen*. Preprint, 2024-10



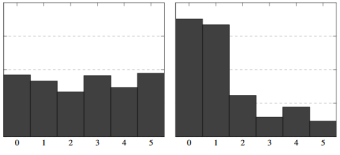
The PANDA dataset was collected from two different sites: *Radboud* University Medical Center in the Netherlands, and *Karolinska* Institutet in Sweden.

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.

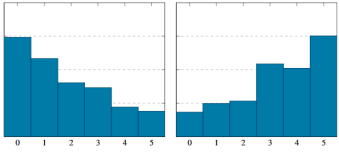
By creating further subsets of the PANDA dataset, we are also able to evaluate robustness in terms of shifts in the label distribution over the ISUP grades 0 - 5.



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



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

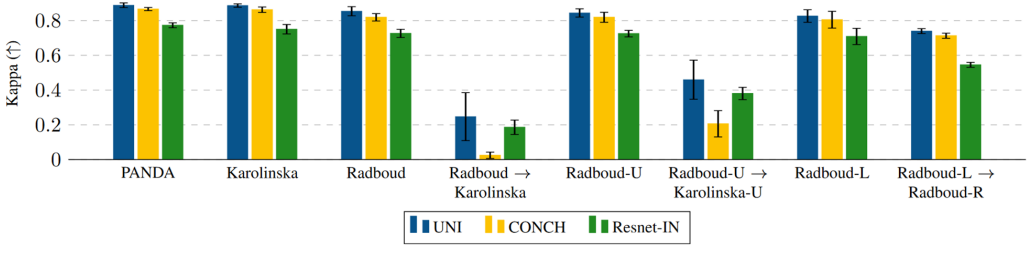


(c) Grade label shift, *Radboud-L* → *Radboud-R*.

Project II: Prostate Cancer Grading - Results

When models are trained and evaluated on the full PANDA dataset, both UNI and CONCH perform well (0.89 kappa for UNI), and similar results are achieved also when models are both trained and evaluated exclusively on data from just one of the sites.

However, when models are trained on Radboud data and evaluated on Karolinska data, the performance drops drastically (0.25 kappa for UNI).



(1/3) While the computational pathology foundation models UNI and CONCH achieve very strong performance *relative* to a baseline model pretrained on natural images, the *absolute* performance can still be far from satisfactory in certain settings.

(2/3) The fact that UNI and CONCH have been trained on very large and varied datasets does *not* guarantee that downstream prediction models always will be robust to commonly encountered distribution shifts.

(3/3) Even within the paradigm of powerful pathology-specific foundation models, the quality of the data utilized to fit downstream prediction models is a crucial aspect.

- If this data has limited variability (in terms of the number of data collection sites or utilized scanners), downstream models can still become sensitive to distribution shifts.

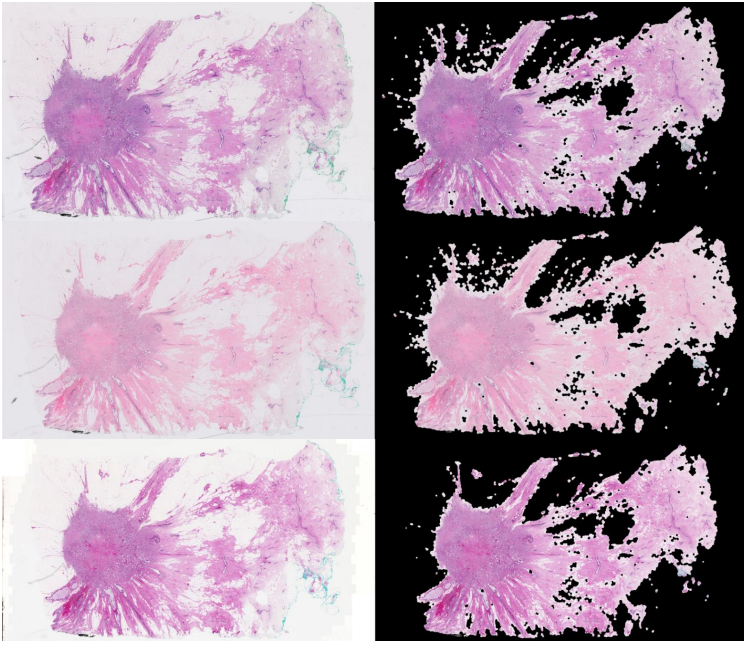
Benchmarking Scanner-Variability Robustness of Computational Pathology Foundation Models
Erik Thiringer, Fredrik K. Gustafsson, Mattias Rantalainen
Ongoing work

We have an in-house dataset of primary H&E WSIs from *more than 300 breast cancer patients* who underwent surgery at Södersjukhuset in Stockholm during 2015.

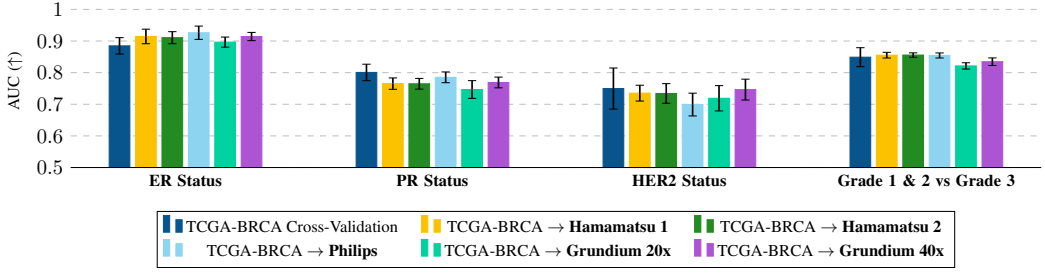
For each patient, *we have scanned the physical WSI using five different scanners*, from three different manufacturers (Hamamatsu 1 & 2, Philips, Grundium 20x & 40x).

Utilizing our multi-scanner dataset, together with the publicly available TCGA-BRCA as an external training dataset, we aim to answer the following research questions:
1. *How much does the performance of recent computational pathology foundation models vary across different scanners?* 2. *Are there clear differences in scanner-variability robustness among different foundation models?*

Project III: Scanner-Variability Robustness - Multi-Scanner WSI



Project III: Scanner-Variability Robustness - Early UNI Results



Some observations:

- Overall, the model generalizes quite well across all scanners, for all tasks (binary classification of ER/PR/HER2 status and grade 1 & 2 vs 3, from the H&E WSI).
- The performance on Hamamatsu 1 & 2 is basically identical. Given that these are two different scanners but of the same scanner model, this is encouraging.
- The performance on the Grundium 40x scanner is consistently a bit better than on Grundium 20x, which seems reasonable.

Summary & Main Takeaways

(1/4) A number of foundation models have been published, just within the last year. These are large deep learning models (*303 million - 1.1 billion parameters*) which have been trained on large amounts of unlabeled WSIs (*100,000 - 1.4 million WSIs*).

(2/4) These foundation models are typically used as *frozen* patch-level feature extractors, on top of which small supervised models are trained. The resulting models have demonstrated strong performance on a range of WSI-level prediction tasks.

(3/4) There are many published foundation models but not yet a lot of comprehensive benchmarking studies. UNI seems to consistently be among the top-performing models, but more detailed analysis is still required.

(4/4) The raw number of WSIs used in pretraining does *not* seem to be the most important factor for model performance: UNI (100,000 WSIs) often outperforms Virchow (1.4 million WSIs). Data *quality* matters, but not yet clear exactly how.

Contact & Acknowledgements

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

Rantalainen Group:

Mattias Rantalainen.

Abhinav Sharma, Ariane Buckenmeyer, Bojing Liu, Constance Boissin, Duong Tran, Erik Thiringer, Francisco J. Peña, Kajsa Ledesma Eriksson, Yujie Xiang.

Karolinska Institutet
CHIME the Cancer Histopathology IMage Epidemiology project
MedTechLabs
ABCAP
Vetenskapsrådet
SWElife
CANCERFONDEN
#ERAPerMed
Cancer Research KI
SeRC Swedish e-Science Research Centre
SWAIPP SWEDISH AI PRECISION PATHOLOGY
AstraZeneca
Stratipath
VINNOVA