PanCyto: A ViT Cytology Foundation Model Trained on a Corpus Spanning Diverse Cell Morphologies

Ethan S. Effendi, Aaditya M. Panchal
Mentors: Joshua J. Levy PhD, Sean McOsker, Louis M. Vaickus MD PhD, Dr. Xiaoying Liu, MD
Emerging Diagnostic and Investigative Technologies, Department of Pathology, Dartmouth Hitchcock Medical Center

Dartmouth Health

INTRODUCTION

- **Cytology** involves a pathologist examining up to millions of cells in a whole-slide image (WSI) sample to screen for diseases.
- Many recent advancements in the field have used **deep learning models** to accelerate this analysis.
- During training, models must first learn **cell morphology** (e.g., N/C ratio, chromatin clumping, etc.) before identifying more complex relationships (e.g., classifying malignancy or cancer types).
- A **cytology foundation model** is a pretrained model that outputs semantically meaningful embeddings (i.e., those that embody morphology in their features) from cell images.
  - Such a model removes the overhead of learning cell morphology in downstream tasks.

OBJECTIVES

- Create a **corpus** of **single-cell images** that can be used to train a cytology foundation model.
- Iteratively perform **self-supervised training** of **ViTs** on subsets of the corpus and evaluate the models.

CORPUS CURATION

Three sets of WSI types were explored over the course of the project.
- Urine (inherited from the AutoParis-X project)
- Cervical

Workflow for Extracting Single-Cell Images from WSIs
1) RGB thresholding to filter away background pixels
2) Connected component analysis to identify all foreground objects
3) Classification model for categorizing objects as different cell types, malignant/benign cells, etc.

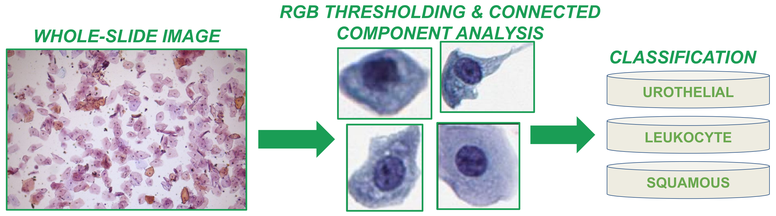


Fig. 1: single-cell image extraction workflow

Metadata
- Urine single-cell images (only isolated cells with 90%+ confidence of classification chosen)
  - Leukocytes: 1,966,869 images
  - Squamous Cells: 2,667 images
  - Urothelial Cells Benign/Atypical: 373,054 images
  - Urothelial Cells w/ Polyomavirus Cytopathic Effect: 260,344 images
- Cervical single-cell images (Pap WSI, objects detected by YOLO, written at level 0 with 15% padding, background masked to white)
  - Squamous single: 1,691
  - Squamous clustered: 142

VISION TRANSFORMER & DINOv3

Vision Transformer (ViT)
- A **ViT model** uses **self-attention** (the technology introduced by Vaswani et al. for text) to capture relationships between the different patches of an image.
  1) Each **patch** of an image is flattened and mapped to a learnable embedding. **RoPE** is used to encode spatial semantics. Additionally, storage and **CLS tokens** are used to embody global information.
  2) **Transformer Encoder Block (Multi-Head Self-Attention)**
  3) Classification MLP Block

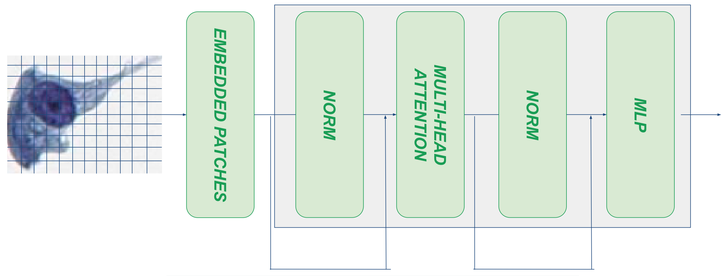


Fig. 2: vision transformers

DINOv3
- A foundation model is trained on examples without labels (i.e., **self-supervised learning**).
  DINOv3 is a state-of-the-art self-supervised learning methodology for training vision transformers.
- DINOv3 works using two ViTs in tandem: a **teacher** and a **student**.
- **Local crops** (smaller regions of input image) and **global crops** (larger regions) are augmented and used as a training signal.
- The teacher's weights are an **exponential moving average (EMA)** of that of the student, which is trained using standard machine learning optimization techniques.
- The goal is for the teacher and student outputs to align for the same image.

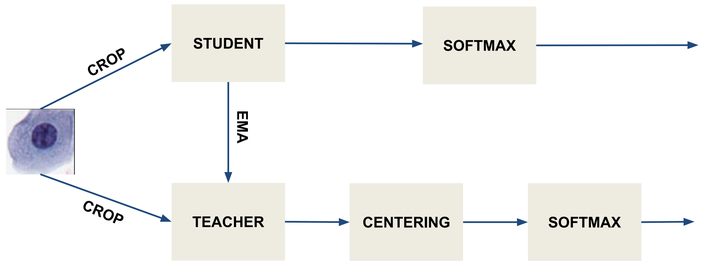
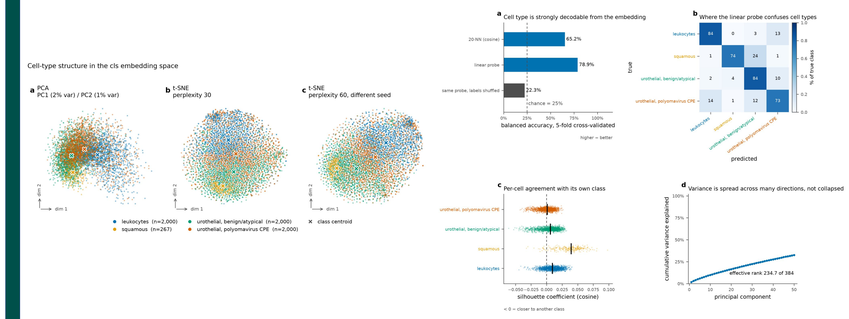


Fig. 3: DINOv3 Self-Supervised ViT Training

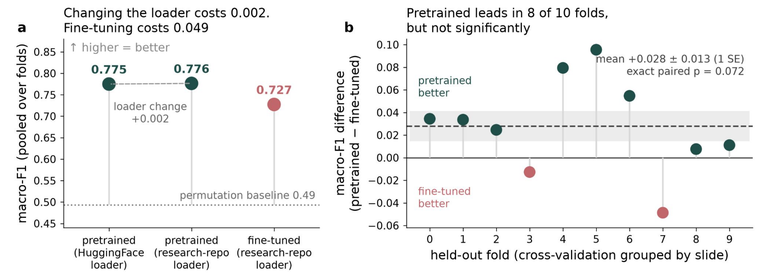
Using the Urine ViT-S as a Case Study
- **ViT-Small (ViT-S):** 12 layers, 22 million parameters
- Trained on 20,000 leukocytes, 20,000 urothelial cells benign/atypical, 20,000 urothelial cells with polyomavirus cytopathic effect, and 2667 squamous cells
- A test set was kept out for later evaluation.

RESULTS

Using the Urine ViT-S as a Case Study



Using the Cervical cells a case study



DISCUSSION & CONCLUSIONS

Gains
- Curated a single-cell image corpus spanning diverse morphologies for cytology foundation model research.
- Trained small ViTs demonstrating promising performance on relatively small subsets of the corpus to provide insight into the development of a production-grade model.
  - 78.9% balanced accuracy across 4 classes of cells from urine WSIs with a linear probe
- Benchmarked classification approaches for Cervical cells through pretrained approaches, along with fine tuning via scratch.

Future Work
- Incorporating other WSI types such as bile duct and fluid cytologies.
- Training a larger, production-grade model.
- Making architectural tweaks to the DINOv3 self-supervised learning setup and exploring which augmentations are actually relevant to cytology.

edit.