

## CytoFM: The first cytology foundation model

Vedrana Ivezić

University of California, Los Angeles, USA

vivezic@g.ucla.edu

Ekaterina Redekop

University of California, Los Angeles, USA

eredekop@g.ucla.edu

Dongwoo Lee

University of California, Los Angeles, USA

rlee8517@g.ucla.edu

Corey Arnold

University of California, Los Angeles, USA

cwarnold@ucla.edu

Ashwath Radhachandran

University of California, Los Angeles, USA

ashwathradha123@g.ucla.edu

Shreeram Athreya

University of California, Los Angeles, USA

shreeram@ucla.edu

Vivek Sant

UT Southwestern Medical Center, Dallas, TX

vivek.sant@utsouthwestern.edu

William Speier

University of California, Los Angeles, USA

Speier@ucla.edu

### Abstract

*Cytology is essential for cancer diagnostics and screening due to its minimally invasive nature. However, the development of robust deep learning models for digital cytology is challenging due to the heterogeneity in staining and preparation methods of samples, differences across organs, and the limited availability of large, diverse, annotated datasets. Developing a task-specific model for every cytology application is impractical and non-cytology-specific foundation models struggle to generalize to tasks in this domain where the emphasis is on cell morphology. To address these challenges, we introduce CytoFM, the first cytology self-supervised foundation model. Using iBOT, a self-supervised Vision Transformer (ViT) training framework incorporating masked image modeling and self-distillation, we pretrain CytoFM on a diverse collection of cytology datasets to learn robust, transferable representations. We evaluate CytoFM on multiple downstream cytology tasks, including breast cancer classification and cell type identification, using an attention-based multiple instance learning framework. Our results demonstrate that CytoFM performs better on two out of three downstream tasks than existing foundation models pretrained on histopathology (UNI) or natural images (iBOT-Imagenet). Visualizations of learned representations demonstrate our model is able to attend to cytologically relevant features. Despite a small pre-training dataset, CytoFM's promising results highlight the ability of task-agnostic pre-training approaches to learn robust and*

*generalizable features from cytology data.*

### 1. Introduction

Cytology, also known as cytopathology, is a specialty of pathology which examines cells from bodily tissues or fluids using a microscope for the purpose of screening and diagnosing cancer [1]. Collecting a cytology biopsy sample compared to a surgical biopsy is safer, faster, and more cost-effective [1]. Cytology tests are used as a screening tool in cancer diagnostics and for conducting follow-ups on suspicious new lesions or fluids from patients with a previous cancer diagnosis [1]. The standard preparation of a cytology sample begins with smearing the sample onto a slide, staining the specimen for cell visualization, and finally, evaluation by a cytopathologist. Unlike histopathology, which focuses on tissue and cell architecture, cytology examines individual cells for their morphology, cohesiveness, and other cellular properties [1, 7]. As a result, a cytopathologist must evaluate each cell under a microscope to identify any diagnostically important alterations which often are present in very few of the examined cells [13]. Given the sparsity of diagnostic cells in the large slides, evaluation of cytology slides is a labor-intensive and expensive task which can delay diagnosis reporting [19]. Recent work has aimed to automate the diagnosis process through deep learning using digitized cytology slides, referred to as whole slide images (WSIs) [6, 12, 15, 24].

Due to their gigapixel size, WSIs are unable to fit at

full resolution into deep learning models. Instead, patch-based methods are commonly used to divide the WSIs into smaller images which are more feasible for model training [6, 23, 24]. Because patch-level labels are often challenging and time-consuming to obtain, multiple instance learning (MIL) is used to aggregate patch level information and is trained on slide-level or patient-level labels as weak supervision [6].

Previous techniques to develop models to learn robust representations of cytology images involve training and evaluating convolutional neural networks on a single dataset. Due to the variations across cytology data, a model trained on a single dataset will struggle to generalize to external datasets [18, 22]. While techniques like stain normalization can be used to address these variations, these methods are not specific to cytology stains and are often not readily available. Rather than relying on normalization techniques, Huang et al. used foundation models which are less susceptible to noise for cytology feature extraction [9]. They proposed to use a frozen feature extractor from foundation models such as BiomedCLIP with a trainable adapter added to the extractor output. The adapter was then trained with parameter-efficient fine-tuning using contrastive learning to increase the specificity of the foundation model features to the task, such as cervical cancer classification. However, because this model was trained to adapt to individual datasets, it must be fully retrained to apply to other datasets and tasks. Using foundation models which are not cytology-specific as the feature extractor and performing task-specific fine-tuning with labeled datasets have also shown promising results [5]. However, Dausort et al. found that histopathology-trained foundation models struggled to generalize to cytology data, likely due to the focus on tissue structure in histology which is not present in cytology [5]. Due to these foundation models not being trained on cytology datasets, they were not able to fully represent and utilize the information present in cytology images.

Unlike previous approaches which rely on explicit annotations, self-supervised learning learns the inherent structure in unlabeled data through pretext tasks, resulting in meaningful representations generalizable to unseen domains. In the field of histopathology where similar WSI related challenges arise, there has been a surge in self-supervised foundation models trained on large and diverse unlabeled datasets [3, 4, 21]. Recent studies have applied self-distillation and masked image modeling to improve the representation learning of Vision Transformers (ViTs) in histology WSIs. These models have proven to be strong feature encoders and robust to differences in staining, preparation style, and other institution-specific differences [4, 21].

Self-supervised foundation models have not been well explored in cytology primarily due to the limited availability of public cytology datasets. Dausort et al. proposed

using four labeled public cytology datasets to fine tune existing foundation models pretrained on medical data. This approach does not fully utilize the publicly available cytology datasets because of the reliance on labels and only uses a single dataset for fine-tuning. With the recent release of new publicly available cytology datasets and the use of a private thyroid cytology dataset, we propose using self-supervised learning to develop a strong and generalizable cytology feature extractor. Self-supervised learning allows the use of both unlabeled and labeled datasets and can learn robust representations from diverse and noisy data. Such an approach can leverage the available cytology datasets to develop a model capable of learning robust, cytology-specific representations.

In this work, we introduce CytoFM (figure 1), the first cytology-specific foundation model. CytoFM is trained on a diverse set of public and private datasets using self-supervised learning. We pretrain a ViT-Base model by leveraging iBOT, a self-supervised framework for pre-training vision transformers (ViT) yielding representations which do not need further fine-tuning [25]. iBOT uses both masked image modeling and self-distillation learning to learn high- and low-level features represented in a single image embedding. iBOT’s rich representation is necessary for cytology where both low-level information, such as mitotic activity, as well as high-level information, like cellularity, are important in diagnostics. We evaluate our model on downstream tasks including breast cancer diagnosis and cell type classification using an attention multiple instance learning (ABMIL) framework. Performance results are compared against both an iBOT model trained on ImageNet (iBOT-Imagenet) [25] and a pathology-specific foundation model, UNI [4]. We summarize our contributions as follows:

- CytoFM is the first foundation model specifically designed for cytology
- Our model learns robust representations of cytology images, capturing important features including cell morphology and mitotic activity
- Despite the limited dataset size, we demonstrate that self-supervised learning on a collection of diverse cytology datasets performs better on two out of three downstream tasks compared to non cytology-specific models in feature extraction and generalization to unseen datasets.

## 2. Methods

### 2.1. Datasets

Foundation models require a large amount of diverse data to learn meaningful and generalizable representations. We combine seven publicly available datasets and one private dataset, creating a multi-institutional dataset comprising 1,393,282 patches which span three types of organs includ-

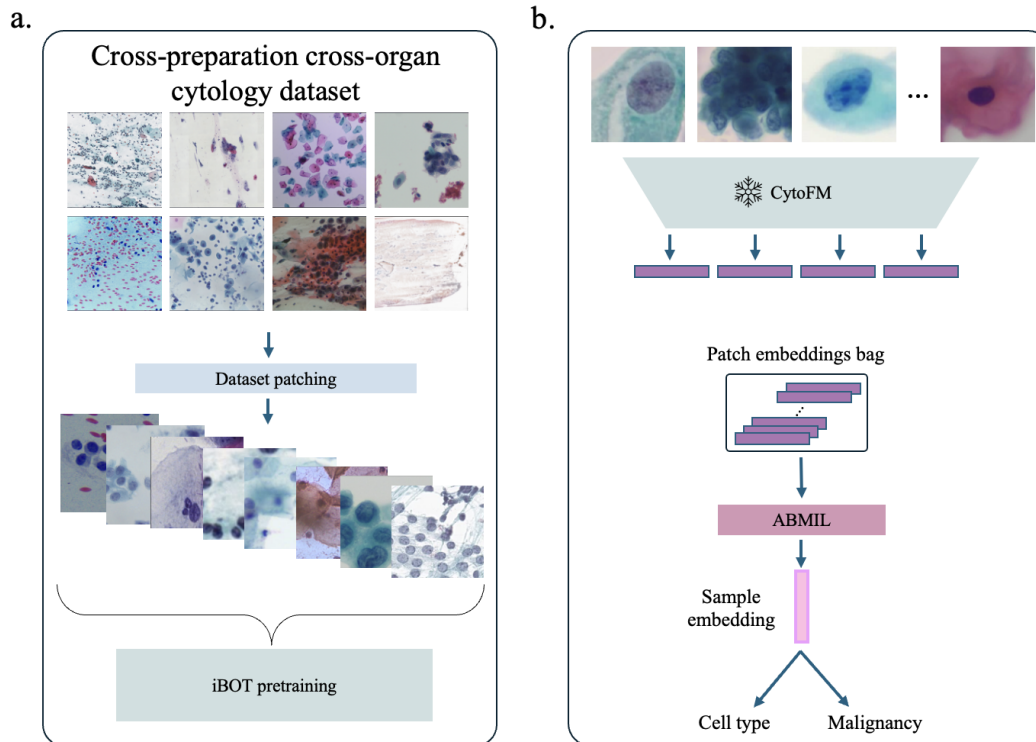


Figure 1. **CytoFM**: the first cytology specific foundation model. Developing this model requires the curation of an unlabeled cytology specific dataset which is used to pre-train a ViT, using the iBOT framework (a). The trained ViT, CytoFM, is then used to extract features for cytology image patches; the patch features for an image are aggregated into a bag and an ABMIL framework creates a single embedding for an image to be used in downstream tasks (b).

ing breast, cervix, and thyroid for pretraining our foundation model (table 1).

The breast cytology dataset we include is the **Saikia et al. fine needle aspiration** (FNAC2019) dataset [20]. Papanicolaou (Pap) stained breast cytology slides were digitized using an ICC50 HD microscope and certified cytopathologists selected and labeled 212 images as benign or malignant.

We include five different publicly available cervical cytology datasets in the pretraining phase.

- The **Mendeley LBC Cervical Cancer** dataset (MLBC) [10] is a liquid-based cytology cervical cancer dataset of 963 Pap smear images at 40x magnification collected from 460 patients using a Leica ICC50 HD microscope. Image level labels for the subcategories of cervical cancer lesions as defined by The Bethesda System for Reporting Cervical Cytology are included.
- The **SiPaKMeD** dataset contains [18] 4049 manually cropped images from cervical Pap smear slides imaged using an Infinity 1 Lumenera CCD camera adapted to an OLYMPUS BX53F optical microscope. Each image is labeled by expert cytopathologists as one of five cell types, which span normal, abnormal, and benign categories.

- The **Brown Multicellular ThinPrep Database** (BMT) [22] consists of 600 clinician selected images taken from 180 cervical ThinPrep® Pap smear slides from 180 patients. Images were obtained using an Olympus BX43 microscope and an Excelis HD color microscopy camera, and were labeled as one of three diagnostic cell types by three board-certified pathologists.
- The **Annotated Pap smear images for Cell Segmentation 2023** (APACS23) [8] is a cervical Pap smear image dataset composed of 3565 images curated from five cytology slides from five patients with cell segmentations provided. The slides were digitized using a 3DHistec Panoramic 1000 digital slide scanner and an Adimec Q-12A-180Fc brightfield camera.
- The **Cervical Cell Edge Detection** (CCEDD) [14] dataset contains 686 images acquired by a Nikon ELIPSE Ci slide scanner, a SmartV350D lens, and a 3-megapixel digital camera, with expert cytologist cell annotations.
- The **Bialystok** cervical cytology dataset [17] contains 162 images from 20 Pap stained slides. Cell type annotations by cytotechnologists were performed for each image, classifying cells into Bethesda defined categories.

We also include a private dataset of **thyroid cytology**

Dataset name	Label	Organ	Pre-training	Evaluation
FNAC2019	binary	Breast	x	x
MLBC	multiclass	Cervix	x	x
SiPaKMeD	multiclass	Cervix	x	
BMT	multiclass	Cervix	x	
APACS23	segmentations	Cervix	x	
CCEDD	segmentations	Cervix	x	
Bialystok	multiclass, segmentations	Cervix	x	
HiCervix	multiclass	Cervix		x
ThyUCLA	binary	Thyroid	x	

Table 1. **Description of datasets** used in the development and evaluation of CytoFM. Datasets are marked if they were used in pre-training the model and if they are also used in evaluation. If a dataset is used in both pre-training and evaluation, it is split into two sets.

**WSIs** from our institution to increase the size and diversity of the dataset. Our dataset contains 496 Pap stained slides from 160 patients digitized with a Leica Aperio GT450 scanner at 40x magnification. The cohort consists of patients with an indeterminate cytology as defined by the Bethesda System for Reporting Thyroid Cytopathology, followed by a suspicious molecular test and surgical resection which provides a final surgical pathology diagnosis at the patient level.

To evaluate the performance and generalizability of CytoFM, we include a large public dataset for downstream evaluation which is completely unseen during the CytoFM training process. The **HiCervix** dataset [2] contains 4,496 Pap stained WSIs which were digitized using either the ZEISS Primostar 3, Olympus microscope BX43, or Sunny-optic RX50 microscope along with a CCD camera. Images of 40,229 cervical cells were extracted from the WSIs and categorized into a three-level hierarchical tree with 29 cell types included.

## 2.2. Preprocessing

To ensure consistency in size and magnification, images from each dataset are split into non-overlapping 256x256 pixel patches at 40x magnification. Patches at the edges of images are zero padded or dropped if more than half of the patch requires padding. In the case where a dataset does not provide the magnification level, we calculate it based on the average pixel size of a nucleus compared to the average micrometer size of a nucleus. Using the calculated magnification we rescale the image to 40x magnification. While histopathology foundation models generally patch WSIs at 20x magnification [4], we choose a higher magnification because cytology focuses on cellular features and small clusters of cells rather than broader cell and tissue structure.

Because cytology slides contain sparse cells, removing slide preparation debris and background areas is essential for creating an informative dataset that allows the model to learn meaningful representations. In the public datasets

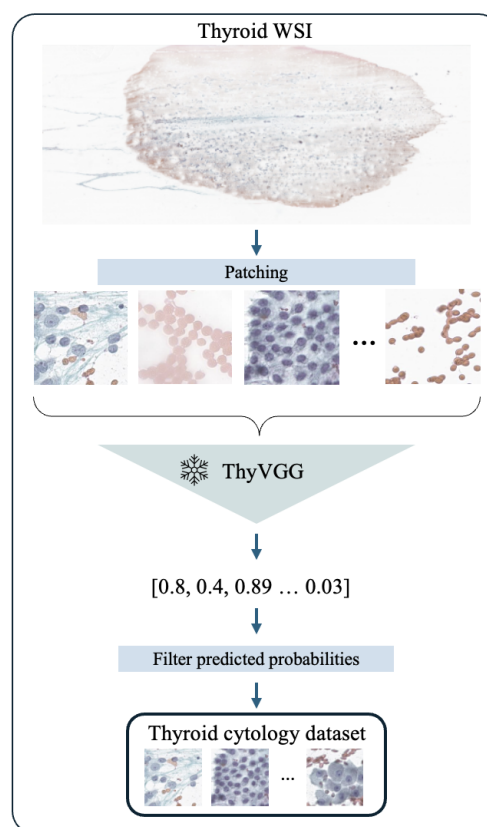


Figure 2. **Filtering of patches from a WSI** for our private thyroid dataset. A WSI is patched and ThyVGG is used to predict the probability that the patch contains relevant information. The top 1500 patches with the highest probability from a slide are used in the thyroid cytology dataset.

we used, this was achieved by previous preprocessing of WSIs into smaller images of cytologically relevant cells by cytologists. However, in our private dataset, this was not possible. To achieve comparable preprocessing, we employ our ThyVGG model (figure 2) to predict the proba-



bility of a patch containing diagnostically relevant information. ThyVGG is a VGG-16 model initialized with ImageNet weights and fine-tuned on a set of 5000 patches annotated for the presence of diagnostic cells. For a given slide, the frozen ThyVGG predicts for each patch the probability that it contains relevant cells. These probabilities are then used to add the top 1500 most relevant patches to our thyroid dataset.

While cytology images are characterized by variations in preparation styles, staining intensities, scanner types, and other institutional differences, we do not apply any normalization techniques related to these differences. The pretraining method of our CytoFM inherently learns to ignore such subtle differences and noise, instead focusing on the cytology relevant information [25].

### 2.3. ViT pre-training

For pretraining CytoFM, we use iBOT [25], a state-of-the-art self supervised learning method for training ViTs using a teacher-student framework with self-distillation and masked image modeling (MIM) using an online tokenizer. iBOT uses self-distillation to learn a semantically-meaningful online tokenizer through minimizing the cross-entropy loss between the predictive categorical distributions from the teacher and student network. Given an image from the dataset,  $x \sim \mathbb{D}$  and two randomly augmented views of the image  $u$  and  $v$ , the cross-entropy loss is computed across views ( $u$  compared to  $v$ ) of the predicted classification tokens from the student,  $u_s^{[CLS]}$  and  $v_s^{[CLS]}$ , and the teacher,  $u_t^{[CLS]}$  and  $v_t^{[CLS]}$ . Self-distillation encourages the model to focus on the important information from an image while ignoring noise, such as variations in stain intensity. High-level context and features are learned by iBOT through MIM using self-distillation, where the model aims to predict masked areas of an image using the remaining contextual information. Specifically, the student network is given an augmented image with blockwise masking while the teacher network is given the augmented image. The student produces projections of the patch tokens for the masked image and iBOT aims to match these embeddings to those produced by the teacher from the unmasked augmented image. Through inferring missing patches, spatial relationships must be learned resulting in the model recognizing both low level details like local structures and cellular morphology as well as high-level semantics like cytoplasm clusters. Following iBOT pretraining, a standalone ViT can be initialized using weights from the teacher ViT, which is a more stable and generalized version of the student, and then frozen. The resulting ViT will be able to produce image embeddings that encode both large-scale and fine-grained patterns.

DINOv2 [16] expands on iBOT and has shown superior performance on ImageNet and in histopathology [4] with

stronger linear probe performance. iBOT was developed for pretraining ViT-Base and ViT-Large models using smaller datasets than DINOv2 which was developed for ViT-Large and ViT-Giant pretraining [4]. Given the smaller size of the cytology dataset, we choose to use ViT-Base with iBOT pre-training and initialize the model with ImageNet-1K iBOT pretrained weights.

## 3. Experiments and results

### 3.1. Implementation details

iBOT initialized with weights from ImageNet-1K pretraining, is used to pretrain a ViT-Base model with  $\sim 1.4$  million cytology patches from seven datasets spanning three organs and seven institutions. CytoFM is the resulting trained ViT which is frozen and used to extract features for downstream classification tasks.

CytoFM performance is compared against iBOT-ImageNet, an iBOT-Base model initialized with weights from pre-training only on ImageNet-1K. We also compare performance against the histopathology specific foundation model, UNI. No additional pre-training or fine-tuning was performed for either of these models and weights were taken from the respective data releases.

We perform downstream classification on three datasets. Two of the datasets, FNAC2019 and MLBC, are split into two non-overlapping sets, one set used during pre-training and the other completely unseen during pre-training and held out for downstream analysis. Malignancy classification is done on a held out set of patches from 100 images in the FNAC2019 dataset where there is an even split of benign and malignant examples. Four class cell type classification is performed on a held out set of patches from 140 samples from the MLBC dataset where there is an even split of cell types. The held out portions of both datasets were split into train, validation, and test sets with a 60:20:20 ratio stratified on the class label. The HiCervix dataset is completely held out from pretraining of the feature extractor model. Cell type classification with four cell types is performed. The train, validation, and test set splits used are those included in the public release of the dataset. Image level labels provided in the public release of the datasets are used in the classification tasks.

Each image in the testing datasets is patched with the same methods as the datasets used for pretraining. Stain normalization is not applied to any of the downstream testing datasets. Using the frozen CytoFM, we extract features from each patch. ABMIL is then used to aggregate the extracted features from patches of each image into a bag for each image using learned attention weights. A linear layer is then used to produce a final predicted probability for that image for the specified classification task. The ABMIL model is fully trainable and metrics are computed on

	FNAC2019		MLBC		HiCervix	
	Acc.	AUC	Acc.	AUC	Acc.	AUC
iBOT-Imagenet	<b>0.946 <math>\pm</math> 0.05</b>	<b>0.991 <math>\pm</math> 0.01</b>	0.879 $\pm$ 0.06	0.983 $\pm$ 0.01	0.803	0.956
UNI	0.927 $\pm$ 0.06	0.983 $\pm$ 0.02	0.895 $\pm$ 0.06	0.986 $\pm$ 0.01	0.800	0.952
CytoFM	0.908 $\pm$ 0.06	0.979 $\pm$ 0.02	<b>0.930 <math>\pm</math> 0.05</b>	<b>0.993 <math>\pm</math> 0.01</b>	<b>0.844</b>	<b>0.968</b>

Table 2. **Downstream performance** results of iBOT-Imagenet, UNI, and CytoFM on two cell type and one malignancy classification tasks. The best performance is in bold. HiCervix is a completely unseen dataset in the training of CytoFM while FNAC2019 and MLBC have a portion of their dataset seen and a portion completely unseen during CytoFM training.

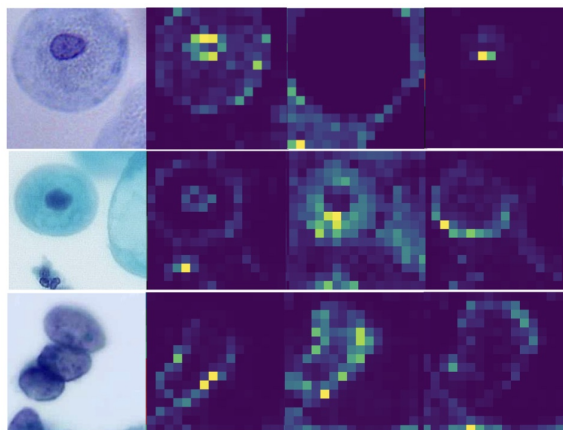


Figure 3. **CytoFM attentions.** Selected attention maps from the last layer of the model. The model attends to relevant cytological features such as the nuclei, mitotic activity, nuclei boundaries, and morphology.

the test set using the model with the best performance on the validation set. Hyperparameters for the ABMIL model are kept the same for each downstream task and across all feature extractor model comparisons.

### 3.2. Results

For the binary classification task we report the accuracy at a threshold of 0.5 and the area under the receiver operating characteristic (AUROC) curve. In the case of multiclass classification, the top predicted probability is taken as the class prediction and the micro-averaged AUROC is reported. Due to the small size of the two partially held out datasets (FNAC2019 and MLBC), we train and evaluate the downstream model on 100 random train/test/validation splits and report the average over the metrics. For the large HiCervix dataset which is completely unseen during pretraining, we report only the metrics from the train/validation/test set specified by the dataset creators.

Classification performance on the held out portions of the FNAC2019 and MLBC datasets are shown in table 2. CytoFM outperforms UNI and iBOT-Imagenet on MLBC, but not on FNAC2019. CytoFM performed the best in cell type classification of the MLBC dataset with an ac-

curacy of  $0.930 \pm 0.05$  (vs. UNI:  $p < 0.001$ , vs iBOT-Imagenet:  $p < 0.001$ ) and an AUROC of  $0.993 \pm 0.01$  (vs. UNI:  $p < 0.001$ , vs iBOT-Imagenet:  $p < 0.001$ ). UNI and iBOT-Imagenet both had lower classification performance with accuracies of  $0.895 \pm 0.06$  and  $0.879 \pm 0.06$  and AUROC values of  $0.986 \pm 0.01$  and  $0.983 \pm 0.01$ , respectively. For malignancy classification, all three models achieve accuracies over 90 percent. iBOT-Imagenet had the highest accuracy,  $0.946 \pm 0.05$  (vs. UNI:  $p = 0.026$ , vs CytoFM:  $p < 0.001$ ), and AUROC,  $0.991 \pm 0.01$  (vs. UNI:  $p < 0.004$ , vs CytoFM:  $p < 0.001$ ). UNI and CytoFM both had lower performance with accuracies of  $0.927 \pm 0.06$  and  $0.908 \pm 0.06$  and AUROC values of  $0.983 \pm 0.02$  and  $0.979 \pm 0.02$ , respectively.

CytoFM had the highest performance on the classification task on the HiCervix dataset, which was completely unseen during model training. On this unseen dataset, CytoFM achieved an accuracy of 0.844 and an AUROC of 0.968. iBOT-Imagenet and UNI perform worse, with accuracies of 0.803 and 0.800 and AUROCs of 0.956 and 0.952, respectively.

## 4. Discussion

Our model outperforms the non cytology foundation models on two of the three downstream tasks, and achieves high performance on the third task. The promising performance of CytoFM demonstrates that the model is learning relevant cytological details and the generalizability of our foundation model.

To assess the ability of CytoFM to learn relevant features, we visualize the embeddings from each model for the two held out datasets using uniform manifold approximation projection (UMAP) (figure 4). Across the three models, we observe that CytoFM produces more defined clusters while the other models exhibit more dispersed feature distributions. The higher dispersion of features suggests that the non cytology-specific models may not have fully captured the cytology-specific features, particularly in the task of cell type classification. We also visualize attention maps (figure 3) from the final layer of CytoFM to evaluate which areas of cytology images the model attends to. The maps show that the model attends to characteristics such as cell

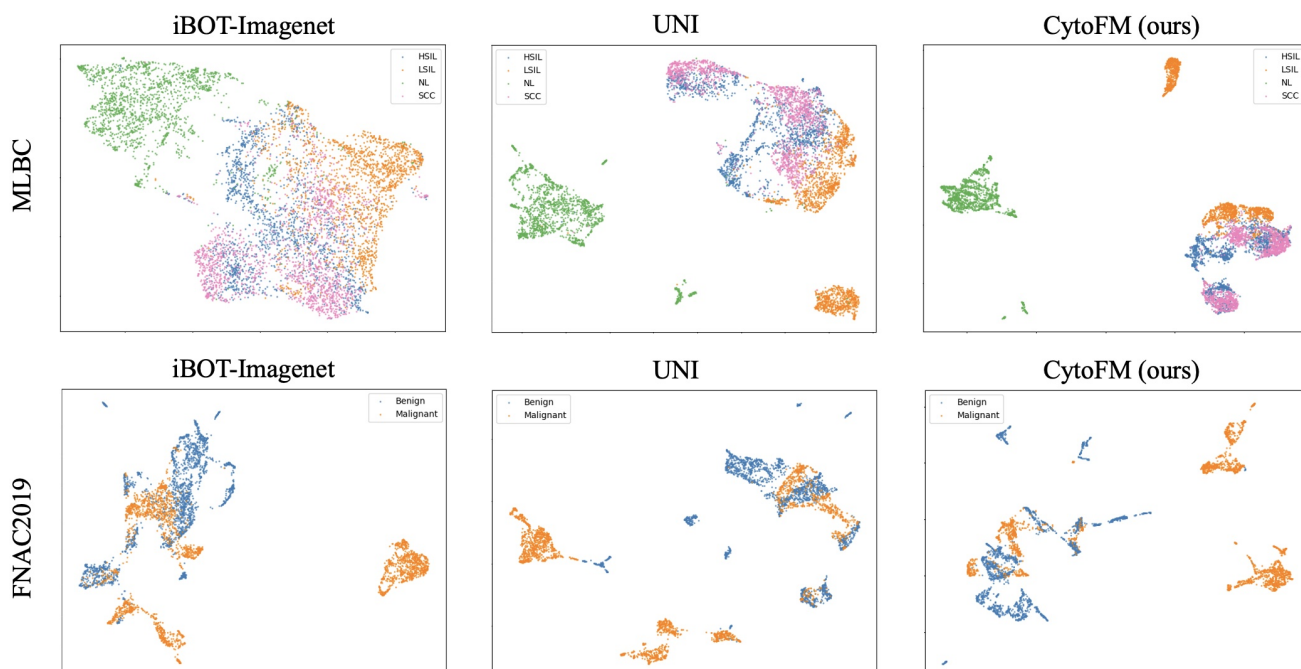


Figure 4. **UMAP of extracted features** for the MLBC (top row) and FNAC2019 (bottom row) datasets. Less dispersion in the features is seen in our model, CytoFM.

morphology and areas of mitotic activity which are important in cytology classification. Therefore, cytology-specific foundation models are able to harness the complex information present in cytology images. In future work, we will explore the impact of attending to these cytology-specific features in downstream applications.

CytoFM outperforms iBOT-Imagenet and UNI on two out of the three downstream tasks. The lower performance on the FNAC2019 dataset compared to the other models may be a result of the disproportional amount of breast cytology in the pretraining of CytoFM. Due to limited dataset availability, only 2,544 images in the pretraining set represented breast cytology compared to close to a million cervix cytology images. Given the high performance on the cervix datasets, MLBC and HiCervix, the model likely became more cervix cytology focused. In future work, we will explore the use of resampling methods in pretraining that could improve the generalizability of the model to other domains [11] and ablations to evaluate dataset importance.

The performance of CytoFM on the HiCervix dataset is slightly lower compared to the current HiCervix state of the art which uses a HeirSwin and has an accuracy of 0.921. The HeirSwin was both trained and evaluated on the HiCervix dataset which limits the generalization ability of the model to external datasets. CytoFM, on the other hand, never sees the dataset when learning feature representations. In order to present a fair comparison with a focus

on the features, we use a simple classification model and do not perform any calibration of the results or tune hyperparameters for the downstream tasks. These modifications could provide an increase in performance resulting in metrics more similar to the reported state of the art. Nevertheless, CytoFM still performs competitively and demonstrates the ability of our model to generate robust representations on unseen datasets from different institutions without the need for additional fine-tuning. The rich representations that CytoFM learns could be leveraged in multimodal applications, enabling the integration of cytology with other data modalities, such as omics and molecular testing to enhance diagnostic accuracy and clinical insights.

## 5. Conclusions

CytoFM is the first cytology foundation model. The model is trained in a self-supervised manner using a multi-institutional and multi-organ dataset containing  $\sim 1.4$  million patches. In two out of three downstream tasks, our model outperforms foundation models trained on natural images and histology. We highlight that our model learns robust and meaningful cytology-specific features which generalize to unseen datasets. Given the competitive performance of CytoFM, we will explore additional training techniques and more extensive ablations and evaluations in future work. CytoFM's performance demonstrates the promise that cytology foundation models can act as robust and generalizable feature extrac-

tors, crucial for the application of deep learning to cytology.

## References

- [1] Mousa A. Al-Abbadi. Basics of cytology. *Avicenna Journal of Medicine*, 1(1):18–28, 2011. [1](#)
- [2] De Cai, Jie Chen, Junhan Zhao, Yuan Xue, Sen Yang, Wei Yuan, Min Feng, Haiyan Weng, Shuguang Liu, Yulong Peng, Junyou Zhu, Kanran Wang, Christopher Jackson, Hongping Tang, Junzhou Huang, and Xiyue Wang. HiCervix: An Extensive Hierarchical Dataset and Benchmark for Cervical Cytology Classification. *IEEE transactions on medical imaging*, 43(12):4344–4355, 2024. [4](#)
- [3] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning, 2022. arXiv:2206.02647 [cs]. [2](#)
- [4] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. [2](#), [4](#), [5](#)
- [5] Manon Dausort, Tiffanie Godelaine, Maxime Zanella, Karim El Khoury, Isabelle Salmon, and Benoît Macq. Exploring Foundation Models Fine-Tuning for Cytology Classification, 2024. arXiv:2411.14975 [eess]. [2](#)
- [6] David Dov, Shahar Z. Kovalsky, Serge Assaad, Jonathan Cohen, Danielle Elliott Range, Avani A. Pendse, Ricardo Henao, and Lawrence Carin. Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images. *Medical Image Analysis*, 67:101814, 2021. [1](#), [2](#)
- [7] Enrico Giarnieri and Simone Scardapane. Towards Artificial Intelligence Applications in Next Generation Cytopathology. *Biomedicine*, 11(8), 2023. [1](#)
- [8] Balazs Harangi, Gergo Bogacsóvics, Janos Toth, Ilona Kovacs, Erzsebet Dani, and Andras Hajdu. Pixel-wise segmentation of cells in digitized Pap smear images. *Scientific Data*, 11(1):733, 2024. [3](#)
- [9] Jialong Huang, Gaojie Li, Shichao Kan, Jianfeng Liu, and Yixiong Liang. An efficient framework based on large foundation model for cervical cytopathology whole slide image screening, 2024. arXiv:2407.11486 [cs]. [2](#)
- [10] Elima Hussain, Lipi B. Mahanta, Himakshi Borah, and Chandana Ray Das. Liquid based-cytology Pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. *Data in Brief*, 30:105589, 2020. [3](#)
- [11] Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, and Yi Guo. USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical Image Analysis*, 96:103202, 2024. [7](#)
- [12] Michael S. Landau and Liron Pantanowitz. Artificial intelligence in cytopathology: a review of the literature and overview of commercial landscape. *Journal of the American Society of Cytopathology*, 8(4):230–241, 2019. [1](#)
- [13] L. J. Layfield, E. E. Mooney, B. Glasgow, S. Hirschowitz, and A. Coogan. What constitutes an adequate smear in fine-needle aspiration cytology of the breast? *Cancer*, 81(1):16–21, 1997. [1](#)
- [14] Jiawei Liu, Huijie Fan, Qiang Wang, Wentao Li, Yandong Tang, Danbo Wang, Mingyi Zhou, and Li Chen. Local Label Point Correction for Edge Detection of Overlapping Cervical Cells. *Frontiers in Neuroinformatics*, 16, 2022. Publisher: Frontiers. [3](#)
- [15] Stefano Marletta, Darren Treanor, Albino Eccher, and Liron Pantanowitz. Whole-slide imaging in cytopathology: state of the art and future directions. *Diagnostic Histopathology*, 27(11):425–430, 2021. [1](#)
- [16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2024. arXiv:2304.07193 [cs]. [5](#)
- [17] Antonina Pater, Krzysztof Siemion, Karol Deptuch, Łukasz Roszkowiak, Jakub Żak, Katarzyna Jakubowska, Stanisław Sulkowski, Marek Baltaziak, Mariusz Koda, and Anna Korzyńska. Conventional Cervical Cytology Image Dataset with Cell Outline Annotations. In *2023 International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 1–6, 2023. ISSN: 1849-2266. [3](#)
- [18] Marina E. Plissiti, P. Dimitrakopoulos, G. Sfikas, Christophoros Nikou, O. Krikoni, and A. Charchanti. Sipakmed: A New Dataset for Feature and Image Based Classification of Normal and Pathological Cervical Cells in Pap Smear Images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3144–3148, 2018. ISSN: 2381-8549. [2](#), [3](#)
- [19] Andrew A. Renshaw, Dawn Underwood, Ghada Aramoni, Beverly Cash, Maureen Croyle, Dave Deeds, Sandra Dolar, Stephen Gmitro, Nancy Ray, Debbie Sabo, Julie A. Shorie, Bridgette Springer, Dana Weber Moffsinger, and Tarik M. Elsheikh. Time consumed by microscopic and nonmicroscopic tasks in image-assisted gynecologic screening: Implications for workload assessment. *Cancer Cytopathology*, 124(7):501–507, 2016. [1](#)
- [20] Amartya Ranjan Saikia, Kangkana Bora, Lipi B. Mahanta, and Anup Kumar Das. Comparative assessment of CNN architectures for classification of breast FNAC images. *Tissue and Cell*, 57:8–14, 2019. [3](#)
- [21] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi,



- Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Hannah Wen, Juan A. Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, Siqi Liu, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30(10):2924–2935, 2024. [2](#)
- [22] E. Celeste Welch, Chenhao Lu, C. James Sung, Cunxian Zhang, Anubhav Tripathi, and Joyce Ou. BMT: A Cross-Validated ThinPrep Pap Cervical Cytology Dataset for Machine Learning Model Training and Validation. *Scientific Data*, 11(1):1444, 2024. [2](#), [3](#)
- [23] Ling Zhang, Le Lu, Isabella Nogues, Ronald M. Summers, Shaoxiong Liu, and Jianhua Yao. DeepPap: Deep Convolutional Networks for Cervical Cell Classification. *IEEE Journal of Biomedical and Health Informatics*, 21(6):1633–1643, 2017. arXiv:1801.08616 [cs]. [2](#)
- [24] Beidi Zhao, Wenlong Deng, Zi Han (Henry) Li, Chen Zhou, Zuhua Gao, Gang Wang, and Xiaoxiao Li. LESS: Label-efficient multi-scale learning for cytological whole slide image screening. *Medical Image Analysis*, 94:103109, 2024. [1](#), [2](#)
- [25] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer, 2022. arXiv:2111.07832 [cs]. [2](#), [5](#)