PLUTO: Pathology-Universal Transformer

Anonymous Authors[1]

Abstract

Pathology images provide a unique challenge for computer-vision-based analysis: a single whole slide image is gigapixel-sized and often contains hundreds of thousands to millions of objects of interest across multiple resolutions. In this work, we propose PathoLogy Universal TransfOrmer (PLUTO): a light-weight pathology foundation model (FM) that is pre-trained on a diverse dataset of 195 million image tiles collected from multiple sites. We design task-specific adaptation heads that utilize PLUTO's output embeddings for tasks ranging from subcellular- to slide-scale, including instance segmentation, tile classification, and slide-level prediction. We find that PLUTO matches or outperforms existing task-specific baselines and pathology-specific FMs, some of which use orders-of-magnitude larger datasets and model sizes. Our findings present a path towards a universal embedding to power pathology image analysis, and motivate further exploration around pathology FMs in terms of data diversity, architectural improvements, sample efficiency, and practical deployability in real-world applications.

1. Introduction

1.1. Foundation Models in Pathology

Pathology as a medical discipline is instrumental in providing diagnostic and prognostic information to clinicians and patients. In a pathology workflow, surgical tissue specimens are collected, stained, and fixed for microscopy. Microscopic analysis of the tissue is used to establish a diagnosis, estimate disease severity, and identify relevant clinical features for treatment (Walk, 2009; Madabhushi & Lee, 2016; Ehteshami Bejnordi et al., 2017). Each whole slide image (WSI) or slide contains up to millions of cells and can be gigapixels in scale, making an exhaustive quantitative manual analysis of WSIs nearly impossible. In addition, information for making pathologic decisions may exist at multiple scales, from several μm to several cm, complicating analysis.

Foundation Models (FMs) are promising for pathology as they can take advantage of large amounts of unlabeled data to build rich representations which can be easily adapted for downstream tasks in a data-efficient manner. The diversity of pre-training data powers these models to generate robust representations, enabling them to generalize better than individual task-specific models trained on smaller datasets. Additionally, these models can be used as a universal backbone across different tasks, reducing the development and maintenance overhead associated with bespoke task-specific models.

Given this prospect, the computational pathology community has made rapid progress in applying self-supervised techniques that have shown promise on natural images such as DINO (Caron et al., 2021), iBOT (Zhou et al., 2021), and DINOv2 (Oquab et al., 2023) to pathology. Most of these efforts have relied on pre-training with a large amount of proprietary data and scaling up the number of backbone parameters used in order to demonstrate high performance on various downstream tasks including tissue classification, disease subtyping classification, and cancer histology segmentation (Kang et al., 2023), (Filiot et al., 2023), (Vorontsov et al., 2023), (Dippel et al., 2024), (Chen et al., 2024).

1.2. Our Approach

We designed and built the *PathoLogy-Universal TransfOrmer*, or PLUTO, a state-of-the-art pathology foundation model that, inspired by the dwarf planet, is based on a novel light-weight ViT backbone that is pre-trained on a diverse dataset from multiple sites and extracts meaningful representations across the levels of the WSI pyramid outlined in Figure 1. The key features of PLUTO are outlined below:

1. **Pre-training Dataset** We compiled a large dataset across a diverse spectrum of histology stains, scanners, and diseases across resolution scales from more than 50 sources. This dataset is augmented with an additional set of samples extracted from over four million manual annotations of 200+ biologically-meaningful objects and region types from board-certified pathologists (Supplementary Section B).

2. **Architecture** We designed the PLUTO backbone to generate multi-scale feature representations from a compact ViT backbone using a self-supervised learning

setup. We extended the FlexiViT scheme (Beyer et al., 2023) for multiple magnifications, and modified the DINOv2 loss with a Masked Autoencoder (MAE) (He et al., 2022) objective and a Fourier-loss-based term to modulate the preservation of low- and high-frequency components (Section 2.1).

3. **Multi-scale Adaptation and Evaluation** We constructed a suite of adaptation heads through efficient fine-tuning to perform diverse, challenging tasks across the levels of the WSI pyramid, and evaluated performance across different biologically-relevant benchmarks (Section 3).

4. **Deployability.** Performing a computational pathology task may require embedding tens to hundreds of thousands of WSI tiles to make a single prediction. To enable this, we focused on developing a model that was efficient (Section 3.2).

2. Methods

2.1. PLUTO Architecture Overview

To design the PLUTO architecture, we start from the DINOv2 architecture which was primarily developed for natural images that are often object-centric. Pathology WSIs have thousands of objects such as nuclei, cells, and glands with different sizes, observed at different image resolutions. To create an encoder which can capture details of objects at different levels of granularity, we add in a MAE objective with multi-scale masking. The MAE setup tries to reconstruct masked regions of the input image from the unmasked regions. We perform masking by varying the patch sizes[1] used for masking while using images across different resolutions of the WSI as shown in Figure 1. In addition to the pixel-level reconstruction loss used in MAE, we add a Fourier reconstruction loss to control the amount of low- and high-frequency information preserved during the pre-training process.

To enable the encoder and decoder to handle varying patch sizes for multi-scale masking, we employ FlexiViT. Since patch size controls the granularity of information captured by the encoder, different downstream tasks may need different patch sizes for optimal performance. FlexiViT allows us to adapt the same backbone to different tasks without needing to train a backbone for every patch size. The patch size also determines the effective sequence length used in ViTs, and FlexiViT allows us to cater to different compute budgets by selecting the most suitable patch size at inference time. Additional training details can be found in Section C.

[1]Note, we use **tile** and **image** interchangeably to refer to the image tiles and **patches** to refer to the patch-tokens obtained by dividing the tile into smaller patches for processing in ViTs.

2.2. PLUTO Adaptation

The backbone training process outlined above learns generic, task-agnostic features. In order to leverage its general capabilities, we add task-specific heads and *adapt* these heads through supervised fine-tuning, while keeping the backbone frozen. This adaptation process is efficient and provides the flexibility to use the same pre-trained backbone for specialized tasks across the biological scales.

We adapt PLUTO to Level 1 slide-level tasks by performing weak supervision on slide-level labels. In particular, MIL (Ilse et al., 2018) is a weakly supervised learning technique where sets of instances are grouped into a "bag" and used to learn bag-level labels. We adapt our PLUTO backbone by using it directly as a featurizer in an AdditiveMIL approach (Javed et al., 2022), which enables interpretable model predictions and class-wise heatmaps.

We adapt PLUTO to tissue-level and cellular/subcellular-level biological scales through fine-tuning either a tile classification or an instance segmentation adaptation head, depending on the context of the dataset and task. We adapt the SSL-pre-trained ViT backbone to instance segmentation tasks via two distinct frameworks: Mask R-CNN (He et al., 2017) and Mask2Former (Cheng et al., 2021). To the best of our knowledge, this is the first work comparing Mask2Former to Mask R-CNN using ViT backbones on histopathology tasks. We also experimented with combining the ViT with a ViT-Adapter (Chen et al., 2022), which has been shown to improve segmentation performance. The output feature maps of the adapter, corresponding to different spatial resolutions of the input image, are used as the input to Mask R-CNN and Mask2Former.

3. Results

3.1. Adaptation Performance

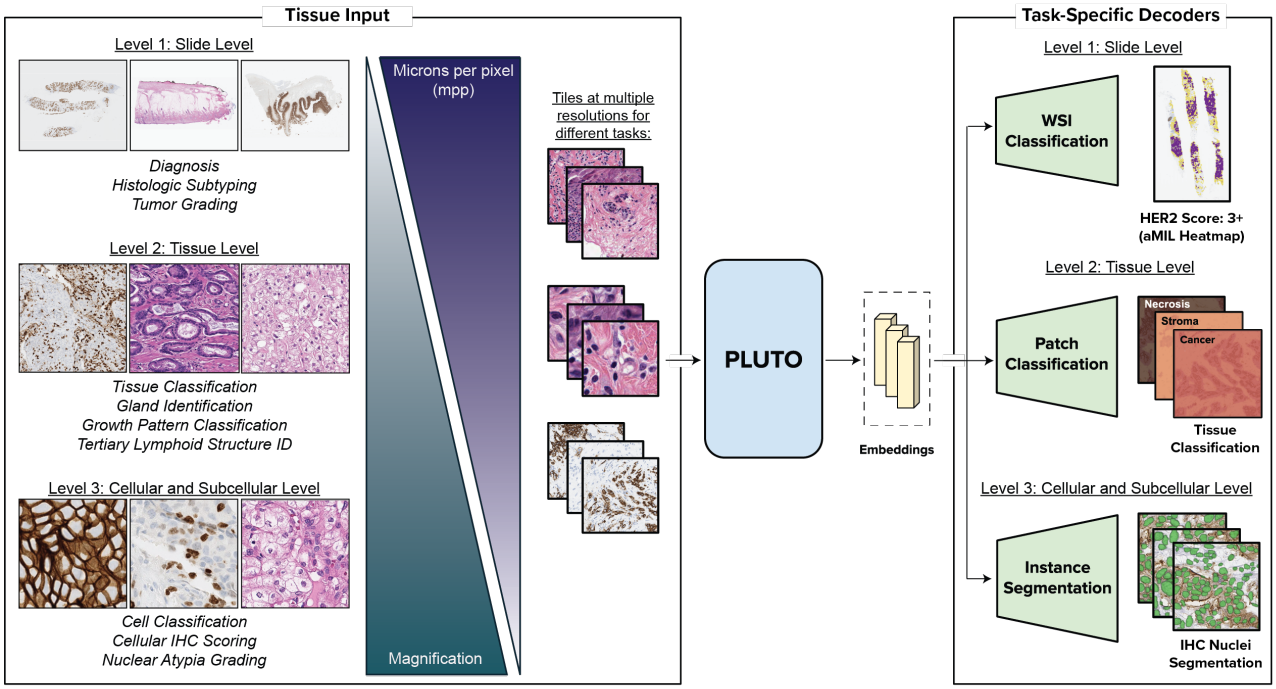
We evaluate the utility of PLUTO on slide-level prediction, tile classification, and instance segmentation tasks[2].

For slide-level prediction, we consider the prediction of the cancer subtypes Adenocarcinoma and Squamous cell carcinoma in non-small cell lung carcinoma (NSCLC) H&E-stained WSIs from TCGA, a popular benchmark for slide-level evaluation, and a proprietary out-of-domain (OOD) set. Results are shown in Table 1.

For tile classification, we use two publicly available datasets. The CRC-100K (Kather et al., 2018) dataset consists of images of human colorectal cancer (CRC) and normal tissue classified into one of nine tissue classes. The Camelyon17-WILDS (Bandi et al., 2019; Koh et al., 2021) dataset contains images of breast cancer metastases in lymph node

[2]Additional dataset details can be found in Section D.

A. PLUTO framework for multi-resolution pathology tasks



B. PLUTO training paradigm

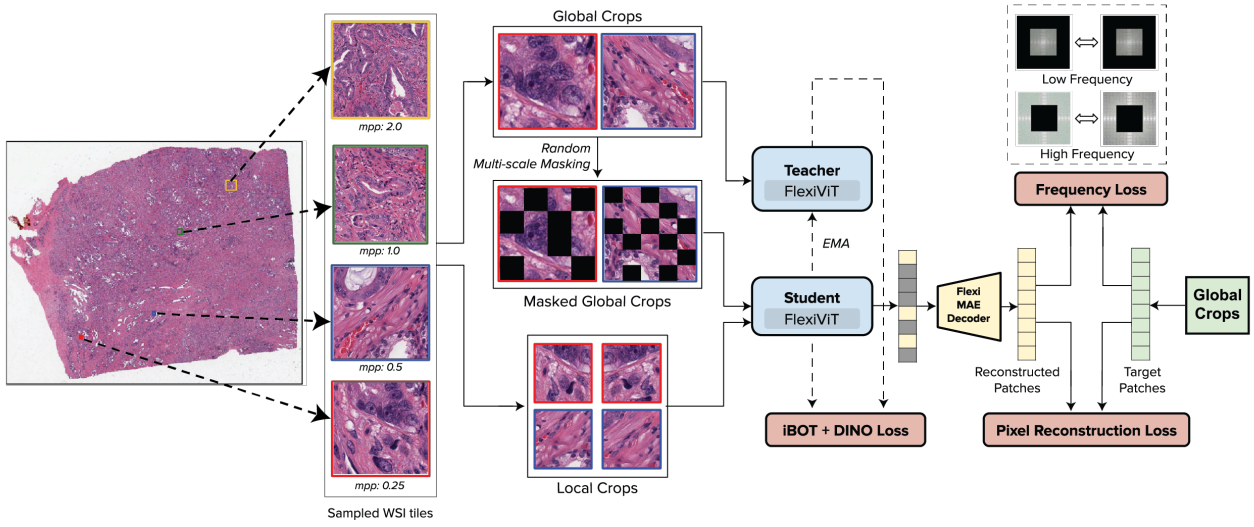


Figure 1. Overview of PLUTO. Panel A) outlines the PLUTO multi-resolution adaptation pipeline. Tiles are extracted from WSIs at multiple resolutions and correspond to scales that capture different biological contexts. We organize pathology tasks according to these biological contexts as slide level, tissue level, and cellular & subcellular level tasks, respectively. PLUTO generates embeddings that are task-agnostic and can be used in a variety of downstream tasks, where adaptation to WSI-level prediction, tile classification, and instance segmentation are shown. Panel B) demonstrates the PLUTO architecture in detail. WSI tiles at multiple resolutions are masked with varying patch sizes and passed to the backbone for self-supervised pre-training. The architecture is optimized for flexibility across multiple scales and patch sizes. In addition to DINO and iBOT losses, MAE and Fourier losses are applied across varying mask sizes to control the amount of low- and high-frequency information that is preserved.

Table 1. Performance of MIL models with different ViT- and CNN-based featurizers on NSCLC subtyping task. The mean and standard deviation across $1,000$ bootstrapped runs are reported. We note that MIL models that use our frozen PLUTO model as a featurizer tend to outperform models with both frozen and fine-tuned CNN backbones (ShuffleNet) and Imagenet-pre-trained ViT backbones. This is especially evident in OOD performance, highlighting the robustness of PLUTO's embeddings.

Model	Dataset	Patch Size	Tuning	In-domain F1	In-domain AUROC	OOD F1	OOD AUROC
PLUTO	NSCLC	16	Frozen	$\mathbf{90.2(1.9)}$	$\mathbf{94.0(1.6)}$	$\mathbf{86.1(2.8)}$	$\mathbf{91.2(2.5)}$
Meta-DINOv2 ViT-S	NSCLC	14	Frozen	$88.6(2.0)$	$92.0(1.9)$	$72.1(4.1)$	$81.9(3.8)$
ShuffleNet	NSCLC	-	Frozen	$83.6(2.4)$	$90.1(2.0)$	$72.2(4.2)$	$83.5(3.5)$
ShuffleNet	NSCLC	-	Fine-tuned	$88.1(2.2)$	$93.9(1.5)$	$42.5(8.0)$	$90.8(2.1)$

Table 2. PLUTO performance summary on tile classification and segmentation datasets. Using a frozen backbone with various adaptation strategies, PLUTO, despite its smaller size, matches top models on CRC-100K and Camelyon17-WILDS, achieves state-of-the-art on GlaS, and outperforms fully supervised baselines while competing with larger fine-tuned backbones reported on PanNuke.

Model	Adaptation Head	Benchmark Name	Metrics	
			Acc.	Bal. Acc.
PLUTO	**Linear Head**	**CRC-100K**	**96.6**	**95.3**
ResNet50*	N/A	CRC-100K	94.7	N/A
			Acc.	Bal. Acc.
PLUTO	**Linear Head**	**C17-WILDS**	**96.2**	-
DenseNet-121*	N/A	C17-WILDS	70.3	-
			DICE	IoU
PLUTO	**Mask2Former**	**GlaS**	**91.2**	**84.5**
PLUTO	Mask R-CNN	GlaS	88.0	79.6
UNet*	N/A	GlaS	85.5	74.8
			bPQ	mPQ
PLUTO	**HoverNet**	**PanNuke**	**67.1**	**47.7**
PLUTO	Mask R-CNN	PanNuke	58.6	-
(Shui et al., 2023)*	N/A	PanNuke	55.3	36.9

*Fully Supervised Baseline Model

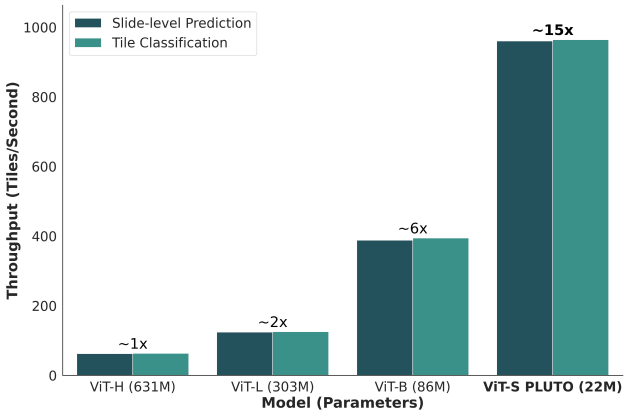


Figure 2. Throughput (tiles/sec) for tile- and slide-level classification tasks using various backbones with a patch size of 16 and tile size of 224×224. Linear probes and AdditiveMIL classifiers are used for tile and slide-level tasks, respectively. Notable pathology FMs use ViT-H (Vorontsov et al., 2023), ViT-L (Chen et al., 2024; Dippel et al., 2024), and ViT-B (Filiot et al., 2023).

sections. The task is to predict whether the central region contains tumor tissue, and the dataset tests for robustness to domain shift.

For instance segmentation, we additionally consider two publicly available datasets. PanNuke (Gamper et al., 2019) consists of exhaustive nuclei labels categorized into five classes. GlaS (Sirinukunwattana et al., 2017) is a gland segmentation dataset with images from colorectal adenocarcinoma. Results for tile classification and instance segmentation in Table 2 demonstrate that PLUTO matches or outperforms task-specific baselines (Nguyen et al., 2024) for these datasets.

3.2. Deployability

To illustrate the efficiency of PLUTO, we compare the throughput efficiency of various ViT backbones (ViT-S, ViTB, ViT-L, ViT-H) for two common pathology tasks: tile classification and slide-level prediction. We note that we have not applied any inference-specific optimizations in this setup. We use the same data-loading pipeline and hardware (A40 GPU) for all the backbones. As seen in Figure 2, for both the tasks, ViT-S is around $2.5\times$ faster than ViT-B, $7.5\times$ faster than ViT-L, and $15\times$ faster than ViT-H.

4. Conclusion

We present in this paper PLUTO: a competitive state-of-the-art pathology Foundation Model based on a light-weight ViT. PLUTO is designed to take advantage of the multi-scale nature of WSIs and provide informative representations across biological scales. We have quantified the performance of PLUTO on a variety of adaptation tasks across biological scales. Our work also demonstrates the importance of incorporation of biological priors in the construction of pre-training datasets and the design of the model architecture for large-scale self-supervised models. We hope that our efforts with PLUTO further motivate building high-performing, deployable FMs; drive FM adoption in pathology; and serve real-world translational research and clinical applications.

References

Bandi, P., Geessink, O., Manson, Q., Dijk, M. V., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F. G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A. B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Cetin, M., Halici, E., Jackson, H., Chen, R., Both, F., Franke, J., Kusters-Vandevelde, H., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J., and Litjens, G. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. *IEEE Transactions on Medical Imaging*, 38:550–560, 2019. doi: 10.1109/TMI.2018.2867350.

Beyer, L., Izmailov, P., Kolesnikov, A., Caron, M., Kornblith, S., Zhai, X., Minderer, M., Tschannen, M., Alabdulmohsin, I., and Pavetic, F. FlexiViT: One Model for All Patch Sizes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14496–14506, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01393. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01393.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging Properties in Self-Supervised Vision Transformers, 2021.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Chen, B., Zhang, A., Shao, D., Song, A. H., Shaban, M., et al. Towards a General-Purpose Foundation Model for Computational Pathology. *Nature Medicine*, 2024.

Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., and Qiao, Y. Vision Transformer Adapter for Dense Predictions. *arXiv preprint arXiv:2205.08534*, 2022.

Cheng, B., Schwing, A. G., and Kirillov, A. Per-Pixel Classification is Not All You Need for Semantic Segmentation. 2021.

Dippel, J., Feulner, B., Winterhoff, T., Schallenberg, S., Gabriel Dernbach, A. K., Tietz, S., Milbich, T., Heinke, S., Eich, M.-L., Ribbat-Idel, J., Krupar, R., Jurmeister, P., Horst, D., Ruff, L., Müller, K.-R., Klauschen, F., and Alber, M. RudolfV: A Foundation Model by Pathologists for Pathologists, 2024.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., , and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 12 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.14585. URL https://doi.org/10.1001/jama.2017.14585.

Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Kain, A. M., Saillard, C., and Schiratti, J.-B. Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling. *medRxiv*, 2023. doi: 10.1101/2023.07.21.23292757. URL https://www.medrxiv.org/content/early/2023/07/26/2023.07.21.23292757.

Gamper, J., Koohbanani, N. A., Benet, K., Khuram, A., and Rajpoot, N. PanNuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European Congress on Digital Pathology*, pp. 11–19. Springer, 2019.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017. doi: 10.1109/ICCV.2017.322.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Ilse, M., Tomczak, J., and Welling, M. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Javed, S. A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., and Prakash, A. Additive MIL: Intrinsically Interpretable Multiple Instance Learning for Pathology. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=5dHQyEcYDgA.

Kang, M., Song, H., Park, S., Yoo, D., and Pereira, S. Benchmarking Self-Supervised Learning on Diverse Pathology Datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3344–3354, June 2023.

Kather, J. N., Halama, N., and Marx, A. 100,000 histological images of human colorectal cancer and healthy tissue, 2018.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *International Conference on Machine Learning*, 2021.

Madabhushi, A. and Lee, G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2016.06.037. URL https://www.sciencedirect.com/science/article/pii/S1361841516301141. 20th anniversary of the Medical Image Analysis journal (MedIA).

Magaki, S., Hojat, S. A., Wei, B., So, A., and Yong, W. H. An introduction to the performance of immunohistochemistry. *Methods Mol. Biol.*, 1897:289–298, 2019.

Molin, J., Bodén, A., Treanor, D., Fjeld, M., and Lundström, C. virchow Stain:Multi-Resolution Feature Enhancement in Pathology Visualization. *preprint arXiv:1610.04141*, 2016.

Nguyen, T. H., Juyal, D., Li, J., Prakash, A., Nofallah, S., Shah, C., Gullapally, S. C., Yu, L., Griffin, M., Sampat, A., Abel, J., Lee, J., and Taylor-Weiner, A. Contrimix: Scalable stain color augmentation for domain generalization without domain labels in digital pathology, 2024.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. DINOv2: Learning Robust Visual Features without Supervision, 2023.

Sellaro, T. L., Filkins, R., Hoffman, C., Fine, J. L., Ho, J., Parwani, A. V., Pantanowitz, L., and Montalto, M. Relationship between Magnification and Resolution in Digital Pathology Systems. . *J. Pathol. Inform*, 2013.

Shui, Z., Zhang, Y., Yao, K., Zhu, C., Sun, Y., and Yang, L. Unleashing the Power of Prompt-driven Nucleus Instance Segmentation. *arXiv preprint arXiv:2311.15939*, 2023.

Sirinukunwattana, K., Pluim, J. P., Chen, H., Qi, X., Heng, P.-A., Guo, Y. B., Wang, L. Y., Matuszewski, B. J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B. B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D. R., and Rajpoot, N. M. Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis*, 35:489–502, 2017. doi: 0.1016/j.media.2016.08.008.

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., Mathieu, P., van Eck, A., Lee, D., Viret, J., Robert, E., Wang, Y. K., Kunz, J. D., Lee, M. C. H., Bernhard, J., Godrich, R. A., Oakley, G., Millar, E., Hanna, M., Retamero, J., Moye, W. A., Yousfi, R., Kanan, C., Klimstra, D., Rothrock, B., and Fuchs, T. J. Virchow: A Million-Slide Digital Pathology Foundation Model, 2023.

Walk, E. E. The role of pathologists in the era of personalized medicine. *Archives of pathology & laboratory medicine*, 133(4):605–610, 2009.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. iBOT: Image BERT Pre-Training with Online Tokenizer, 2021.

A. Pyramid Structure of Whole Slide Images

WSIs are digitized and stored in a multi-scale pyramidal structure, where the base of the pyramid is the highest-resolution image data as captured by the slide scanner. The resulting scan of a WSI can reach $200,000 \times 200,000$ pixels at a full resolution of 0.25 microns (μm) per pixel (mpp) (Sellaro et al., 2013); however, different "levels" of the pyramid may be accessed for different purposes.

Biological entities observed on WSIs vary dramatically in scale, and therefore pathologists will commonly move between magnifications to assess different aspects of a tissue sample on a pathology slide (Molin et al., 2016). At low magnification, pathologists may scan across slides to identify regions of interest in the tissue, with characteristic lengths of approximately 1 mm–1 cm. At middle magnification (such as 5–10×) pathologists commonly view structures at length scales of 200 μm–1 mm. At this scale, pathologists distinguish between tissue types, glands, tumor growth patterns, histologic subtypes of diseases, or other multicellular entities in the image. At high magnification (20–40×) it is possible to resolve entities 1 μm–50 μm in length, such as individual cell identities, subcellular structural morphology used in determining malignancy, or localization of immunohistochemical (IHC) staining (Magaki et al., 2019).

The hierarchical nature of biological entities necessitates considering the multiple scales at which information must be extracted and used by ML algorithms. For example, passing a 224×224 image tile at 0.25 mpp through an encoder developed for encoding at 1 mpp may completely miss relevant nuclear pleomorphism, whereas passing a 224×224 tile at 1 mpp through an encoder developed for encoding at 0.25 mpp may be unable to adequately distinguish between acinar and lepidic growth patterns. For clarity, we organize pathology tasks according to such biological scales as follows:

- **Level 1: Slide Level** This scale includes tasks that label the entire slide such as predicting driver gene mutations in cancer, histologic subtyping, or tumor grading. However, it is uncommon that slide-level assessments are made at slide-level magnification. Typically, assessments made at this scale are aggregated across evaluation of higher-magnification tiles.
- **Level 2: Tissue Level** This is the scale at which it is possible to identify and characterize tissue regions (e.g. cancer regions and necrotic regions) and many-cellular objects such as glands.
- **Level 3: Cellular and Subcellular Level** This is typically the maximal resolution of a WSI, where cellular and subcellular morphology is evident.

B. Pre-training Data Characteristics

The dataset used for self-supervised pre-training comprises public and proprietary datasets, totaling 195M image tiles sampled at four resolutions from $158,852$ WSIs derived from over 50 source sites (Figure 3). The WSIs span over 16 tissue groups and 28 disease areas, which capture a broad range of benign, malignant, and inflammatory lesions. Additionally, the training set is unique in the representation covering 11 scanners and four stain groups: hematoxylin and eosin (H&E) formalin-fixed paraffin-embedded (FFPE), H&E frozen, IHC (capturing over 100 distinct IHC stains including PD-L1, HER2, and Ki-67) and special stains (including six stains such as trichrome and iron). The base objective magnification of our training set consists of both $20\times$ and $40\times$ slides. Tiles are sampled from regions at up to four different resolutions: $40\times$ (0.25 mpp), $20\times$ (0.5 mpp), $10\times$ (1 mpp), and $5\times$ (2 mpp).

Following findings from DINOv2 (Oquab et al., 2023) highlighting the significant value of incorporating curated data into self-supervised pre-training, this dataset is augmented with an additional set of samples extracted from over four million manual annotations from board-certified pathologists. These hand-drawn pathologist annotations correspond to hundreds of different types of biological entities at various scales (e.g., lymphocyte, blood vessel, Gleason pattern 3 prostate cancer, tumor bed). During pre-training, the labels are discarded, but the inclusion of pathologist-curated regions covering a wide range of biological patterns provides an implicit data diversity in the pre-training process. This source of biological diversity, combined with the broad range of stains, organs, diseases, and source sites, makes this one of the most diverse large-scale digital pathology datasets to date.

C. Backbone Training Details

We observe slightly better performance with the teacher over the student, and thus use the teacher for all downstream tasks. We use ViT-S for the student and teacher encoders, and a shallower model is used for the MAE decoder. For training, we use AdamW with a base learning rate of 0.002 and a learning rate warmup for the first 5 epochs. We use a distributed training setup to scale the training across 64 NVIDIA A40 GPUs.

D. Adaptation Datasets

D.1. NSCLC subtyping

For NSCLC subtyping, we use slides from the publicly available TCGA Adenocarcinoma (LUAD) and Squamous Cell Carcinoma (LUSC) groups. We use 500 slides for model development and 247 (128 LUAD / 119 LUSC) slides for test

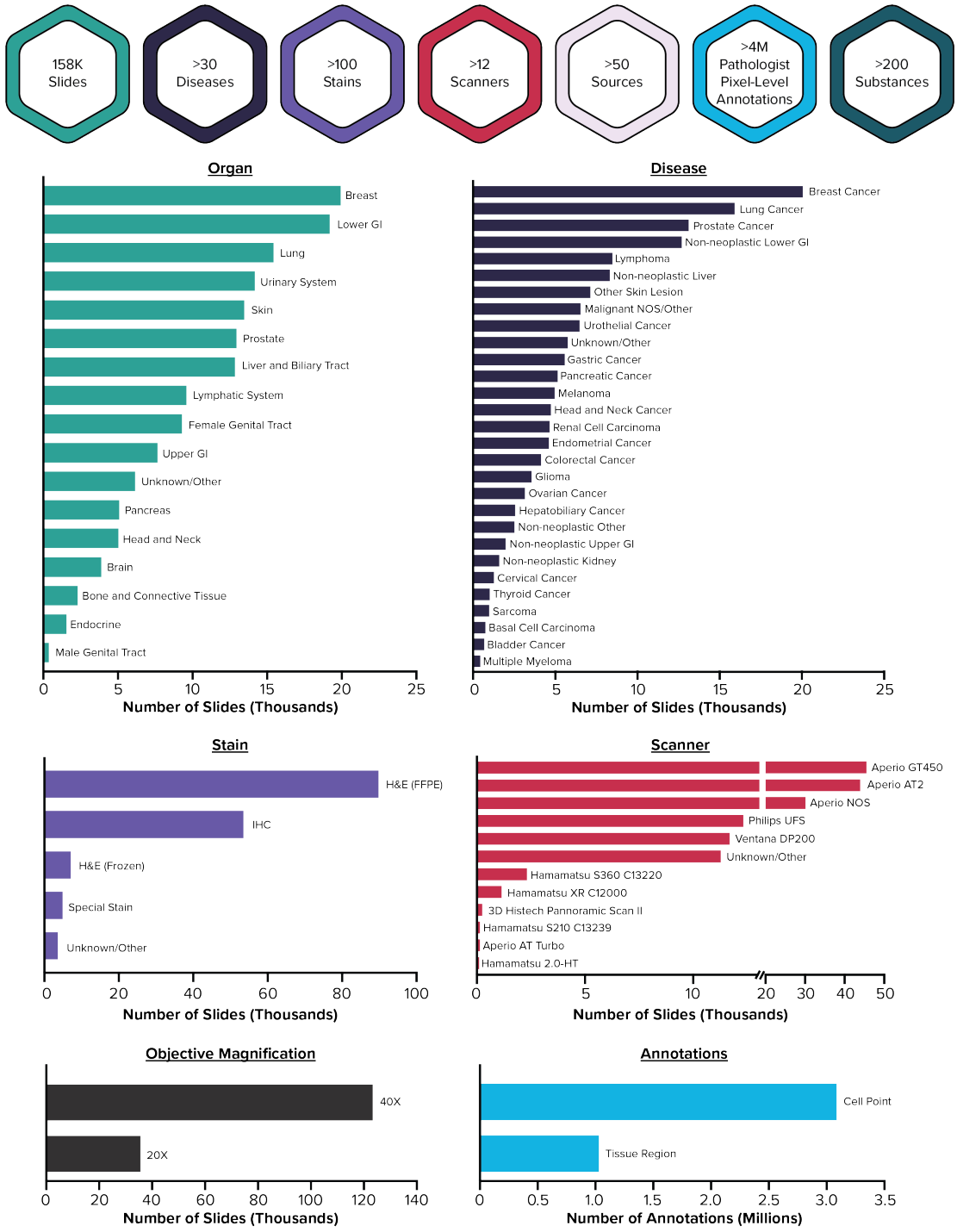


Figure 3. Dataset characterization for the pre-training dataset. The distribution of the dataset by organ, disease, stain, scanner, and objective magnification is shown, as well as the distribution of cell point and tissue region annotations which augment the pre-training dataset (NOS: Not Otherwise Specified). Aggregate data characteristics are summarized above these distributions which also indicate the number of biologically-meaningful objects and region types, which we term *substances* (e.g. lymphocyte, blood vessel, Gleason pattern 3 prostate cancer, tumor bed). The large number of source sites (50+) guarantees large diversity during PLUTO self-supervised pre-training.

set evaluation. We evaluate out-of-distribution (OOD) performance using a proprietary dataset of 205 WSIs (162 Adenocarcinoma WSIs, 45 Squamous Cell Carcinoma WSIs) collected from a different source site with varying image acquisition and processing steps, resulting in visual differences from the TCGA WSIs. Since slide-level prediction tasks are often limited by the number of slides available for development, we limit our development set to 500 slides for both of these tasks and evaluate model performance on in-distribution (ID) and OOD test sets.

D.2. CRC-100k

The CRC-100K (Kather et al., 2018) dataset consists of $107,180$ images (224×224 at 0.5 mpp) of human colorectal cancer (CRC) and normal tissue extracted from 136 H&E histopathology WSIs from the NCT Biobank and the UMM pathology archive, classified into one of nine tissue classes. The training set consists of $100,000$ images (referred as NCT-CRC-HE-100K) and the evaluation set consists of $7,180$ images (referred as CRC-VAL-HE-7K). Performance was measured using accuracy (Acc.) and balanced accuracy (Bal. Acc.).

D.3. Camelyon17-WILDS

The Camelyon17-WILDS (Bandi et al., 2019; Koh et al., 2021) dataset contains $455,954$ images (96×96 pixels at 1 mpp, downsampled from 0.25 mpp slides) from 50 WSIs of breast cancer metastases in lymph node sections from five different hospitals. The task is a binary classification of whether the central 32×32 region contains tumor tissue. The training set consists of $302,436$ tiles from 30 WSIs from three hospitals, the ID validation set of $33,560$ from the same 30 WSIs, the OOD validation set of $34,904$ from 10 WSIs from the fourth hospital, and the OOD test set of $85,054$ from 10 WSIs from the fifth hospital. Each split has a $50/50$ class balance. Performance was evaluated using accuracy in the OOD test set, measuring robustness to shifts across hospitals.

D.4. GlaS

We evaluated the performance on the GlaS (Sirinukunwattana et al., 2017) dataset, which consists of 85 images for training and 80 images for testing for a total of 165 images derived from 16 H&E-stained sections of stage T3 or T42 colorectal adenocarcinoma. These slides were scanned using a Zeiss MIRAX MIDI Slide Scanner with a resolution of 0.465 mpp and varying image sizes (most commonly 775×522). Performance was measured using dice coefficient (Dice) and Intersection over Union (IoU) in the test set.

D.5. PanNuke

The PanNuke dataset (Gamper et al., 2019) consists of 481 visual fields across 19 different tissue types from WSIs from TCGA and a local hospital, with a total of $189,744$ exhaustive nuclei labels categorized into five classes. The visual fields were randomly sampled from more than $20,000$ WSIs that were scanned at either $40\times$ or $20\times$ and re-sized to $40\times$. Following the original publishers of this dataset, we report binary panoptic quality (bPQ) and multi-class panoptic quality (mPQ). For the ablation study comparing different adaptation heads, experiments were conducted in the binary configuration where nuclei were not classified, and therefore only bPQ is reported. The experimental setup used an inference patch size of 16 for the HoverNet architecture due to its design. All conducted experiments were thus performed using this specified patch size.