Unsupervised Feature Extraction from a Foundation Model Zoo for Cell Similarity Search in Oncological Microscopy Across Devices

Gabriel Kalweit [1 2] Anusha Klett [1] Mehdi Naouar [1 2] Jens Rahnfeld [1 2] Yannick Vogt [1 2]
Diana Laura Infante Ramirez [1 3] Rebecca Berger [3] Jesus Duque Afonso [3] Tanja Nicole Hartmann [3]
Marie Follo [3 4] Michael Luebbert [3 5] Roland Mertelsmann [1 3 6] Evelyn Ullrich [1 7] Joschka Boedecker [1 2 8]
Maria Kalweit [1 2]

Abstract

Acquiring high-quality datasets in medical and biological research is costly and labor-intensive. Traditional supervised learning requires extensive labeled data and faces challenges due to diverse imaging equipment and protocols. We propose Entropy-guided Weighted Combinational FAISS (EWC-FAISS), using foundation models trained on natural images without fine-tuning, as feature extractors in an efficient and adaptive k-nearest neighbor search. Our approach shows superior generalization across diverse conditions, achieving competitive performance compared to fine-tuned DINO-based models and NMTune, whilst reducing computational demands. Experiments validate the effectiveness of EWC-FAISS for efficient and robust cell image analysis.

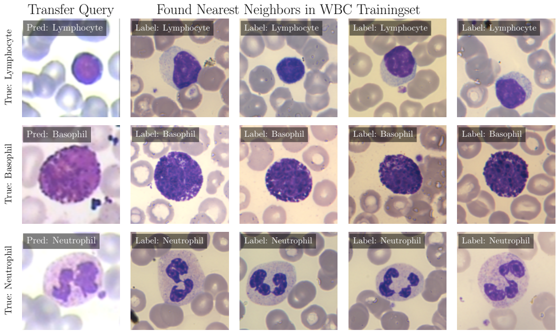


Figure 1: Prediction of EWC-FAISS on query images from the LISC Dataset, the four nearest neighbors in the WBC dataset and the ground truth labels.

1. Introduction

In medical and biological research, data acquisition is often challenging and involves high costs and labor-intensive processes (Johnson & Bourne, 2023). This is especially true in cell image analysis, where traditional approaches rely heavily on supervised learning techniques that require high-quality, large-scale labeled datasets, which are expensive and time-consuming to produce. Additionally, the heterogeneity of imaging equipment (e.g., different microscopes) and protocols (e.g., varying media and lighting conditions) introduces variability, complicating the task and degrading the performance of narrowly trained models. Training these models demands costly GPUs, extensive training time, and frequent retraining for new tasks. Addressing these challenges necessitates methodologies that leverage existing data more efficiently and generalize across diverse imaging conditions without extensive retraining or fine-tuning. Recent advancements in machine learning, particularly in the development of foundation models (Bommasani et al., 2021), present a promising solution. Foundation models, characterized by their vast scale and versatility, are pretrained on a variety of abstract objectives, enabling them to capture a wide array of features applicable across domains. DINO (Caron et al., 2021; Oquab et al., 2024) leverages self-

[1] Collaborative Research Institute Intelligent Oncology (CRIION), Freiburg, Germany [2] Neurorobotics Lab, Department of Computer Science, University of Freiburg, Germany [3] Department of Medicine I, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Germany [4] Lighthouse Core Facility, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Germany [5] Department of Hematology, Oncology and Stem Cell Transplantation, Medical Center University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), Partner Site Freiburg, Freiburg, Germany [6] Mertelsmann Foundation, Freiburg, Germany [7] Goethe University, Department of Pediatrics, Experimental Immunology and Cell Therapy, Frankfurt am Main, Germany; Frankfurt Cancer Institute (FCI), Goethe University, Frankfurt am Main, Germany; University Cancer Center (UCT) Frankfurt; Mildred Scheel Career Center (MSNZ), Hospital of the Goethe University Frankfurt, Germany; German Cancer Consortium (DKTK), partner site Frankfurt/Mainz, Frankfurt am Main, Germany [8] IMBIT//BrainLinks-BrainTools, University of Freiburg, Freiburg, Germany. Correspondence to: Gabriel Kalweit <gabriel.kalweit@intelligent-oncology.org>.

Published at ICML 2024 Workshop on Foundation Models in the Wild. Copyright 2024 by the author(s).

distillation, allowing the model to *teach* itself by comparing different versions of the same image. Segment Anything (SAM) (Kirillov et al., 2023) focuses on segmentation, learning to identify specific objects within an image based on prompts such as points and bounding boxes. SWIN (Liu et al., 2021; 2022a) builds a layered understanding of the image through hierarchical feature maps and directs its attention to specific regions using a *shifted window* approach. ConvNeXT (Liu et al., 2022b; Woo et al., 2023) rethinks the traditional convolutional neural network architecture. CLIP (Radford et al., 2021) learns to associate image content with natural language descriptions. Finally, ViTMAE (He et al., 2022) employs a masked autoencoder technique, hiding parts of the image and tasking the model to reconstruct them. These diverse objectives and architectures enable these models to extract complementary and orthogonal information from the data, potentially leading to better generalization on unseen data outside the training distribution. In line with the *Platonic Representation Hypothesis* (Huh et al., 2024), we believe this makes them particularly suitable for tasks like cell image analysis, where acquiring large amounts of labeled data can be challenging. Following this rationale, this study explores the utility of various foundation models without fine-tuning for the task of cell image analysis. We develop an automated pipeline, *Entropy-guided Weighted Combinational FAISS* (EWC-FAISS), combining different foundation models as pre-trained feature-extractors to build an approximate Hierarchical Navigable Small World (HNSW) (Malkov & Yashunin, 2020) FAISS index (Douze et al., 2024). To enhance robustness, we propose an entropy-based search for the optimal number of neighbors at runtime, and to alleviate unbalanced settings through distribution-reweighting. By building a FAISS index, model iteration can be executed much faster compared to training a full parameterized classifier while still being able to benefit from the generalization capabilities of sophisticated feature extractors (cf. Figure 1). Our contributions are four-fold. We demonstrate the effectiveness of our approach in multiple scenarios. First, we start by evaluating our approach on the WBC dataset (Bodzas et al., 2023) containing stained blood cell smears, as well as a transfer to the LISC dataset (Rezatofighi et al., 2010). Second, we conduct an analysis of EWC-FAISS on live cell state classification, with a transfer from the Nanolive 3D Cell Explorer to the BioTek Lionheart FX microscopes. Third, we evaluate our approach on live cell type classification, with a transfer from the BioTek Lionheart FX to the Nanolive 3D Cell Explorer. Lastly, we evaluate NMTune (Chen et al., 2024a) in these domains.

2. Related Work

The recent advancements in general foundation models, particularly DINO and SAM, have significantly influenced medical and cellular image processing domains. MedSAM (Ma et al., 2024) has extended the utility of SAM to general medical imaging tasks, while models like UNI (Chen et al., 2024b), WTC-11 DINO (Doron et al., 2023), DINOBloom (Koch et al., 2024) and scDINO (Pfaendler et al., 2023) have adapted DINO-style approaches to histopathology and (multi-channel) cellular image analysis. Israel et al. introduced with CellSAM an adaptation of SAM specifically designed for cell segmentation. Despite these advancements, training foundation models specifically for medical applications often requires substantial computational resources (Ma et al., 2024; Chen et al., 2024b; Kraus et al., 2024), limiting accessibility for multiple iterations during model development. In their work, Doron et al. (2023) showed that ImageNet features *can* generalize in some settings more effectively than fine-tuned models in the cellular domain, especially in (rather) low-data regimes. This study also revealed that DINO features could predict expert-defined cellular phenotypes, enhance the prediction of compound bioactivity, and facilitate unbiased profiling of cellular morphology. Also, self-supervised masked autoencoders have been shown to be capable of capturing cellular biology when trained on massive datasets (Kraus et al., 2024). However, our research indicates that combining features from multiple foundation models, trained on natural images, can outperform single-model approaches, including DINO, in terms of performance and transferability. The scDINO (Pfaendler et al., 2023) model demonstrated that a k-nearest neighbor (k-NN) search using DINO features, fine-tuned and adapted to multi-channel cellular imaging, can be competitive with other methods for cell classification tasks. Recent research has also investigated how to best select foundation models and hyperparameters for cost-efficient fine-tuning for the task at hand (Arango et al., 2024) and how to make the general features learned from foundation models more robust for downstream tasks via covariance and dominant singular value regularization (Chen et al., 2024a). Our proposed approach stands orthogonal to this line of research by leveraging a combination of features from various foundation models as feature extractors, even when trained on non-domain specific data. This methodology aims to achieve better generalization and adaptability in cell image analysis without any fine-tuning typically required.

3. Datasets

We utilized two publicly available datasets and created four new datasets (cf. Figure 3, details in Appendix A):

Stained White Blood Cells: The WBC dataset (Bodzas et al., 2023) includes 14,424 images of stained white blood cells from patients with acute myeloid and lymphoid leukemia, as well as those without leukemic pathology. They are categorized into neutrophil segments, neutrophil bands, eosinophils, basophils, lymphocytes, monocytes,

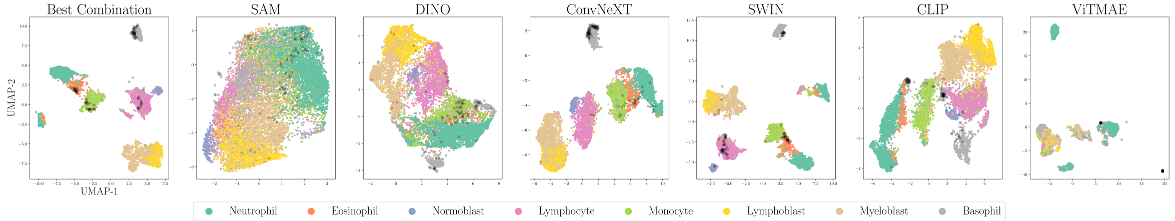


Figure 2: Embeddings for WBC with SAM+ConvNeXT+SWIN+CLIP (Best Combination). Samples from LISC in black.

normoblasts, myeloblasts, and lymphoblasts. The LISC dataset (Rezatofighi et al., 2010) comprises 257 images of white blood cells from healthy individuals, classified into basophils, eosinophils, lymphocytes, monocytes, and neutrophils.

Live Cell State Imaging: The CELL DEATH NANOLIVE dataset contains images of treated JIMT-1 breast cancer cells categorized into living, dead, apoptotic, and necrotic cells. It includes 7,420 with the Roboflow software manually annotated and segmented images captured with a high-resolution Nanolive microscope at 60× magnification. The CELL DEATH LIONHEART dataset contains 59 images recorded with the Lionheart automated microscope at 20×.

Live Cell Type Imaging: The CELL TYPE LIONHEART dataset contains 456,366 images of K562 and Jurkat cancer cells, extracted from homogeneous cell line images recorded with a lower-resolution Lionheart automated microscope at 20×. The CELL TYPE NANOLIVE contains 206,742 images of Jurkat cells captured with the Nanolive at 60×.

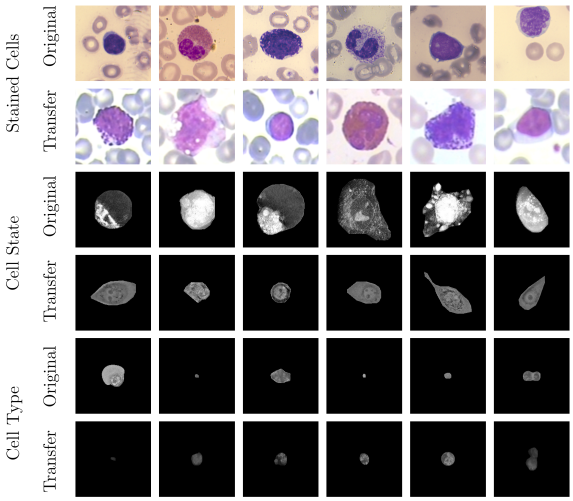


Figure 3: Used datasets: stained white blood cells of the WBC (Original) and LISC dataset (Tansfer), live cell state recorded by the Nanolive (Original) and BioTek Lionheart FX microscopes (Transfer), and two cell lines recorded by BioTek Lionheart FX (Original) and Nanolive (Transfer).

4. Method

Next, we outline how to build, train, and query a latent embeddings database using EWC-FAISS.

Foundation Model Embedding Generation We utilize a set of foundation models as encoders to generate embeddings for our data. Each foundation model M_i in the set $\{M_1, M_2, \ldots, M_n\}$ processes the input data X to produce a corresponding embedding $E_i = M_i(X)$. For a given encoder subset $\mathcal{M} \subseteq \{M_1, M_2, \ldots, M_n\}$, we concatenate the embeddings E_i from each encoder $M_i \in \mathcal{M}$ to form a full feature representation $\mathcal{E} = [||_{m \in \mathcal{M}} E_m]$. This concatenated embedding $\mathcal{E}$ serves as the input for subsequent tasks.

Database and FAISS Index Construction We construct a database $\mathcal{D}$ consisting of embeddings $\{E_i|_{i=1}^{n}\}_c$ and labels for each cell c (cf. Figure 2). This results in $\mathcal{D} = \{E_i|_{i=1}^{n}\}_{c=1}^{N}$, where N is the total number of cells. $\mathcal{D}$ is then used to train a HNSW FAISS index (Malkov & Yashunin, 2020) on concatenated embeddings $\mathcal{E}_c$.

Class Weight and Entropy Calculation To address the issue of class imbalance in our training data, class weights were calculated based on the frequency of each class once after adding the embeddings to the index. The total number of samples was divided by the product of the number of classes and the count of each class. The fixed class weight for class i is given by $w_i = N/(C \cdot N_i)$, where N is the total number of samples, C is the number of classes, and N_i is the number of samples in class i in $\mathcal{D}$. This approach ensures that less frequent classes receive higher weights, thereby reducing the impact of imbalance. We then quantify the normalized entropy of labels from the nearest neighbors to estimate the uncertainty in label distribution independent from the re-weighting by:

$$H_{\text{norm}} = \frac{-\sum_{i=1}^{m} p_i \log(p_i)}{\log(k)}, \tag{1}$$

where p_i is the probability of the i-th class and k is the number of nearest neighbors. By normalizing the entropy, it is scaled between 0 and 1 regardless of the number of drawn neighbors. A lower entropy indicates a higher purity of the

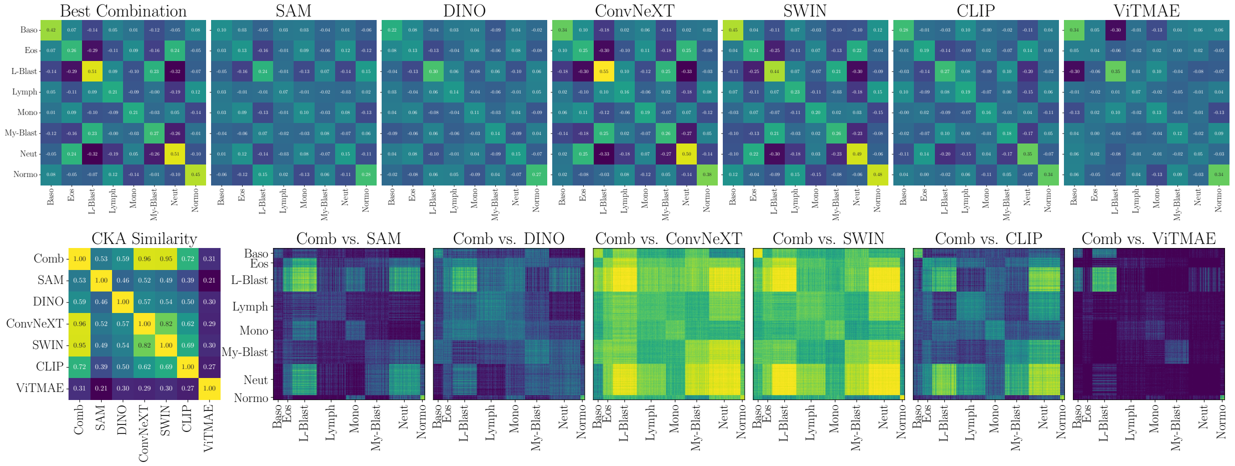


Figure 4: (top) The mean of cosine similarity scores for all cell embeddings per cell type for each foundation model. (bottom) The mean of RBF CKA similarity scores between all foundation models and the RBF CKA similarity scores of the best combination of models (Comb) compared to all single foundation models ordered by cell type (Baso = Basophil, Eso = Eosinophil, L-Blast = Lymphoblast, Lymph = Lymphocyte, Mono = Monocyte, My-Blast = Myeloblast, Neut = Neutrophil, Normo = Normoblast), where blue denotes low similarity and yellow high similarity of two embeddings.

neighborhood in terms of class labels, which is desirable for accurate classification.

Nearest Neighbor Search and Prediction The core of our method involves an adaptive nearest neighbor search to determine the optimal number of neighbors (k) for classification. Starting from a minimum value, k is increased exponentially until the entropy of the nearest neighbors falls below a pre-defined threshold. This adaptive approach balances the need for accuracy and computational efficiency by dynamically adjusting k based on the neighborhood's label distribution. During the search process, the class weights are used to perform a weighted vote among the nearest neighbors to account for minority classes. The predicted class is then determined by the class with the highest weighted vote:

$$y_{\text{pred}} = \arg\max_{l} \sum_{j=1}^{k} w_j \cdot \mathbb{1}\{y_j = l\}, \tag{2}$$

where $\mathbb{1}\{y_j = l\}$ is an indicator function that is 1 if the label of the j-th nearest neighbor y_j is l and w_j is the weight of the j-th nearest neighbor's class label.

5. Experiments

Since most related work uses DINO to represent cellular morphology, we compare EWC-FAISS with the best combination of foundation models (optimized on a validation set) to a finetuned DINOv2-based vision transformer model (DINO FT). We furthermore compare to NMTune, a lightweight addendum to foundation models aiming at making performances more robust on (unseen) downstream tasks. Experimental details can be found in Appendix B.

Embeddings We compare the similarities among cell embeddings within one foundation model exemplary for the WBC dataset in Figure 4 (top) by the mean of the cosine similarity scores of all embeddings per cell type projected via Principal Component Analysis (PCA) to 100 dimensions. The best combination of foundation models achieves the highest intra-class similarity with a mean diagonal similarity of 0.355 and the lowest inter-class similarity with a mean off-diagnoal similarity of -0.043 (details in Appendix C). In contrast, the ViTMAE model shows the noisiest results, indicating less distinct feature separation. Additionally, we study the similarity across foundation model representations, using Radial Basis Function (RBF) Centered Kernel Alignment (CKA) (Kornblith et al., 2019) in Figure 4 (bottom). ConvNeXT and SWIN achieve the highest similarity compared to the best combination of models (being part of the best combination SAM+ConvNeXT+SWIN+CLIP).

Classification An overview of the classification results over five runs is given in Table 1. In terms of inner dataset distribution performance, EWC-FAISS is on par with the baselines (rank second for WBC, rank first for cell state classification from Nanolive and rank third for cell type classification from Lionheart). However, EWC-FAISS can handle domain shift from transfer to different imaging devices and scenarios better than both DINO FT and NMTune. NMTune is on par or better than DINO FT. Furthermore, after embedding generation, building EWC-FAISS is multiple orders of magnitudes faster than training DINO FT ($\sim$ 6 minutes compared to $>$ 10 hours, cf. Figure 5).

Table 1: Classification Results. Models were trained on ORIGINAL and evaluated on test and TRANSFER sets.

METHOD	MACRO ACCURACY	MACRO PRECISION
ORIGINAL: WBC (8 CLASSES)		
DINO FT	93.96 ± 0.85	97.77 ± 1.98
NMTUNE	98.43 ± 0.45	98.05 ± 0.88
EWC-FAISS	97.60 ± 0.23	97.94 ± 0.00
TRANSFER: LISC (7 CLASSES)		
DINO FT	16.59 ± 1.37	32.46 ± 2.37
NMTUNE	52.10 ± 10.32	59.44 ± 5.62
EWC-FAISS	78.47 ± 0.34	81.93 ± 0.47
ORIGINAL: CELL DEATH NANOLIVE (4 CLASSES)		
DINO FT	89.49 ± 0.85	88.45 ± 1.23
NMTUNE	88.39 ± 0.61	88.97 ± 1.20
EWC-FAISS	90.07 ± 0.00	91.75 ± 0.00
TRANSFER: CELL DEATH LIONHEART (2 CLASSES)		
DINO FT	64.95 ± 11.44	79.92 ± 21.23
NMTUNE	80.62 ± 7.36	80.14 ± 6.01
EWC-FAISS	86.11 ± 0.98	84.86 ± 0.75
ORIGINAL: CELL TYPE LIONHEART (2 CLASSES)		
DINO FT	91.88 ± 0.14	91.98 ± 0.14
NMTUNE	92.14 ± 0.21	91.91 ± 0.12
FAISS	87.32 ± 0.01	86.50 ± 0.01
TRANSFER: CELL TYPE NANOLIVE (1 CLASS)		
DINO FT	24.54 ± 0.93	–
NMTUNE	62.61 ± 3.97	–
EWC-FAISS	85.08 ± 0.01	–

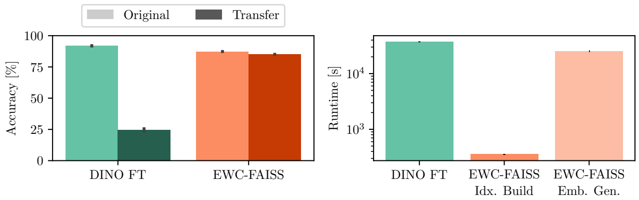


Figure 5: Results on the transfer from a BioTek Lionheart FX to the Nanolive 3D Cell Explorer. (left) EWC-FAISS is robust to distribution shift induced by the second device compared to DINO FT. (right) Once the embeddings are generated (Emb. Gen.), EWC-FAISS is multiple orders of magnitude faster (Idx. Build).

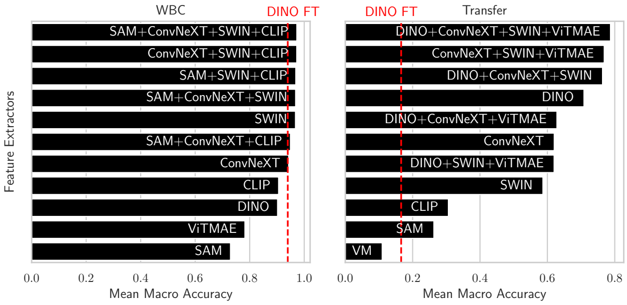


Figure 6: Balanced accuracies of the foundation models and the best combination using EWC-FAISS in black and DINO FT in red on WBC (left) and on the transfer to LISC (right).

Figure 6 shows exemplary for the WBC dataset an evaluation of the best found combinations of foundation models from the power set, as well as an ablation and a comparison to all single-models. As can be seen, combining features clearly leads to better performances.

6. Discussion and Conclusion

Our results show, that adaptive k-NN search on fixed features from combinations of foundation models can yield on par or better performances in the domain of cellular imaging. However, it is no panacea and except for the WBC dataset, we had to adapt the combinations and hyperparameters when transferring to a new device. Nonetheless, we were always able to find a good working combination in the realm of minutes instead of hours of training fully fine-tuned models, even for full iterations over the whole power set. This highlights the versatility and accessibility of the proposed framework, as performing several development cycles is very cost-effective and fast; all experiments have been executed on consumer hardware, specifically an AMD Ryzen 9 7950X3 CPU and an NVIDIA GeForce RTX 4090. NMTune on the best found set of foundation models also is a well-performing alternative to a fully-trained classification model, although the generalization capabilities towards out-of-distribution samples appear to be inferior compared to approximate k-NN search. Given the current pace of development, upcoming or recently introduced foundation models, such as CellSAM, should be incorporated into the proposed framework moving forward.

Impact Statement

Our research introduces Entropy-guided Weighted Combinational FAISS (EWC-FAISS) to tackle the high costs and labor-intensive processes of cell image analysis. By leveraging foundation models without fine-tuning, EWC-FAISS significantly reduces computational and temporal resources, lowering hurdles for researchers from non-technical fields. This approach demonstrates adaptability across diverse imaging settings and devices, making cell image analysis more accessible and efficient. Validated on multiple datasets, EWC-FAISS provides a practical solution for real-world medical research and diagnostics. Its ability to generalize well in various conditions is a step towards robust and reliable performance, offering significant advancements in medical image analysis.

Acknowledgements

This project was funded by the Mertelsmann Foundation.

References

Arango, S. P., Ferreira, F., Kadra, A., Hutter, F., and Grabocka, J. Quick-tune: Quickly learning which pretrained model to finetune and how. In *The Twelfth International Conference on Learning Representations*, 2024.

Bodzas, A., Kodytek, P., and Zidek, J. A high-resolution large-scale dataset of pathological and normal white blood cells. *Sci. Data*, 10(1):466, July 2023.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L. E., Goel, K., Goodman, N. D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T. F., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M. S., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S. P., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J. F., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y. H., Ruiz, C., Ryan, J., R'e, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K. P., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M. A., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models. *ArXiv*, 2021. URL https://crfm.stanford.edu/assets/report.pdf.

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021.

Chen, H., Wang, J., Shah, A., Tao, R., Wei, H., Xie, X., Sugiyama, M., and Raj, B. Understanding and mitigating the label noise in pre-training on downstream tasks. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=TjhUtloBZU.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., Williams, M., Oldenburg, L., Weishaupt, L. L., Wang, J. J., Vaidya, A., Le, L. P., Gerber, G., Sahai, S., Williams, W., and Mahmood, F. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024b. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL http://dx.doi.org/10.1038/s41591-024-02857-3.

Doron, M., Moutakanni, T., Chen, Z. S., Moshkov, N., Caron, M., Touvron, H., Bojanowski, P., Pernice, W. M., and Caicedo, J. C. Unbiased single-cell morphology with self-supervised vision transformers. June 2023.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. The faiss library. 2024.

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

Huh, M., Cheung, B., Wang, T., and Isola, P. The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

Israel, U., Marks, M., Dilip, R., Li, Q., Schwartz, M., Pradhan, E., Pao, E., Li, S., Pearson-Goulart, A., Perona, P., et al. A foundation model for cell segmentation. *bioRxiv*.

Johnson, T. R. and Bourne, P. E. The biological data sustainability paradox, 2023.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2023.

Koch, V., Wagner, S. J., Kazeminia, S., Sancar, E., Hehr, M., Schnabel, J. A., Peng, T., and Marr, C. Dinobloom: A foundation model for generalizable cell embeddings in hematology. *CoRR*, abs/2404.05022, 2024. doi: 10.48550/ARXIV.2404.05022. URL https://doi.org/10.48550/arXiv.2404.05022.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Kraus, O., Kenyon-Dean, K., Saberian, S., Fallah, M., McLean, P., Leung, J., Sharma, V., Khan, A., Balakrishnan, J., Celik, S., Beaini, D., Sypetkowski, M., Cheng, C. V., Morse, K., Makes, M., Mabey, B., and Earnshaw, B. Masked autoencoders for microscopy are scalable learners of cellular biology. *CoRR*, abs/2404.10242, 2024.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 9992–10002. IEEE, 2021.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. Swin transformer v2: Scaling up capacity and resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022a.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022b.

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. Segment anything in medical images. *Nat. Commun.*, 15(1): 654, January 2024.

Malkov, Y. A. and Yashunin, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(4):824–836, 2020.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Pfaendler, R., Hanimann, J., Lee, S., and Snijder, B. Self-supervised vision transformers accurately decode cellular state heterogeneity. *bioRxiv*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Rezatofighi, H., Khaksari, K., and Soltanian-Zadeh, H. Automatic recognition of five types of white blood cells in peripheral blood. volume 6112, pp. 161–172, 06 2010. ISBN 978-3-642-13774-7. doi: 10.1007/978-3-642-13775-4_17.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S. ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023.

A. Dataset Details

We split the original datasets to 90% train, 9% validation, and 1% test sets and used the transfer datasets only for evaluation.

WBC The WBC dataset includes 14,424 cell images from microscopic blood smear images from 36 leukemic and 45 non-leukemic peripheral blood smears, collected from 78 anonymized patients. This cohort includes 18 patients with acute myeloid leukemia, 15 with acute lymphoid leukemia, and 45 with no leukemic pathology. Blood smears were stained using May-Grünwald and Giemsa-Romanowski solutions, and blast cell lineage was determined by flow cytometry. Images were captured using an Olympus BX51 microscope with a Basler acA5472-17uc camera, achieving a resolution of approximately 42 pixels per $1\mu m$ under a magnification of 100×. The dataset contains nine different annotated blood cell types: neutrophil segments, neutrophil bands, eosinophils, basophils, lymphocytes, monocytes, normoblasts, myeloblasts, and lymphoblasts. Due to the low number of neutrophil bands, we have merged them with the neutrophil segments.

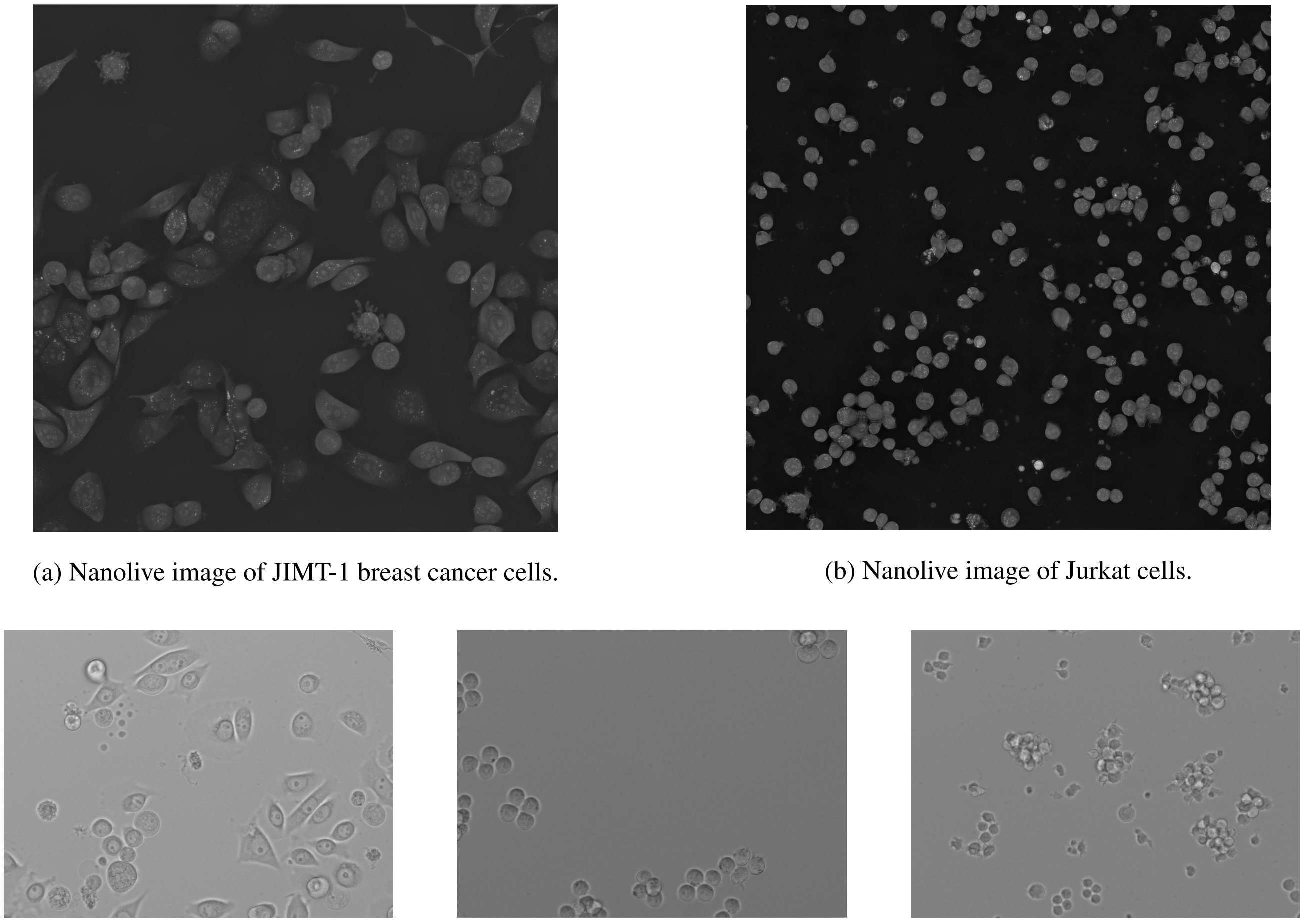
LISC The LISC dataset contains hematological images from peripheral blood of 8 healthy individuals, resulting in 257 white blood cell images from 100 microscope slides. These slides were stained using the Gismo-Right technique, imaged with a Microscope-Axioskope 40 at 100× magnification, and recorded by a Sony SSCDC50AP digital camera in BMP format. Each image was collected from the Hematology-Oncology and BMT Research Center at Imam Khomeini Hospital, Tehran. A hematologist classified the images into five normal leukocyte categories (basophil, eosinophil, lymphocyte, monocyte, and neutrophil).

CELL DEATH NANOLIVE This dataset comprises 7,420 images of JIMT-1 cells, captured at 40× magnification using a Nanolive 3D microscope. The dataset is categorized into Dead (728), Living (4,613), Apoptotic (707), and Necrotic (1,372) cells. An additional test set includes 1,122 images, with Dead (255), Living (500), Apoptotic (99), and Necrotic (268) cells. The images are 2D projections from the 3D microscope. JIMT-1 cells were cultivated using Dulbecco's modified eagle medium (DMEM) FluoroBrite (Gibco), supplemented with 10% fetal bovine serum (FBS; Gibco), 1x L-glutamine (Gibco), and 1% Pen Strep (10,000 Units/ml penicillin, 10,000 µg/ml streptomycin; Gibco). For each image, we use contrast limited adaptive histogram equalization (CLAHE) to normalize its contrast. Examples can be found in Figure 7.

CELL DEATH LIONHEART This dataset contains 59 annotated test images, categorized into dead (23) and living cells (36). The breast carcinoma cell line JIMT-1 (ACC 589, DSMZ) was used as adherent cells. JIMT-1 cells were cultivated using Dulbecco's modified eagle medium (DMEM) (Gibco), supplemented with 10% fetal bovine serum (FBS; Gibco), 1x L-glutamine (Gibco), and 1% Pen Strep (10,000 Units/ml penicillin, 10,000 µg/ml streptomycin; Gibco). Cells were incubated at 37°C in a humidified atmosphere containing 5% CO2 and passaged twice a week. For treatment, JIMT-1 cells were seeded in an 8-well chip (Ibidi) and either left untreated or treated with 25 µM of Etoposide (Sigma Aldrich) for 72 hours. Propidium Iodide (0.25 µg/ml, Sigma Aldrich) was used as a fluorescence marker to stain dead cells. Brightfield and fluorescence images were acquired every 2 hours using a Biotek Lionheart Fx automated microscope. Examples can be found in Figure 7.

CELL TYPE LIONHEART The CELL TYPE LIONHEART dataset includes 456,366 images of homogeneous K562 (264,904) and Jurkat cell images (191,462), captured using a Lionheart automated microscope at 20× magnification, and 206,742 images of Jurkat cells captured with a 3D Nanolive microscope. Each image was segmented using SAM, and the type of each crop was assigned accordingly. Contrast normalization was applied to each image using CLAHE. Jurkat and K562 cells were cultivated in RPMI 1640 medium (Gibco), supplemented with 10% FBS, 1x L-glutamine, and 1% Pen Strep. All cells were incubated at 37°C in a humidified atmosphere containing 5% CO2 and passaged twice a week. The cell images were segmented and masked using SAM, where we cropped 224 × 224 pixel crops around the centers of the found cell masks. Examples can be found in Figure 7.

Magnification Adaption We adapt the resolution between Lionheart and Nanolive microscopes. The Lionheart microscope, operating at 20× with a field of view of 291x394 μm^2, generates images of 904x1224 pixels, corresponding to approximately 0.322 µm/pixel. In contrast, the Nanolive microscope, set at 60× with a field of view of 85x85 μm^2, produces images of 448x448 pixels, resulting in a resolution of about 0.19 µm/pixel. To match the resolution of the images from the Nanolive microscope to those from the Lionheart microscope, a scaling factor of 0.59 is applied, calculated by dividing the Nanolive pixel size by the Lionheart pixel size.



(a) Nanolive image of JIMT-1 breast cancer cells.

(b) Nanolive image of Jurkat cells.

(c) Lionheart image of JIMT-1 cells.

(d) Lionheart image of K562 cells.

(e) Lionheart image of Jurkat cells.

Figure 7: Stitched images from our Nanolive and Lionheart microscopes showing various cell types.

B. Hyperparameters

DINO FT Since most related work are using DINO to represent cellular morphology, as baseline, we use a pre-trained DINOv2 vision transformer model, which was finetuned for 50 epochs using the Adamw optimizer with cosine learning rate decay with warm-up (from lr=10^{-5} to lr=0). We employ horizontal flipping, normalization and color jitter as data augmentations.

FM Zoo + NMTune As proposed in (Chen et al., 2024a), we set $\lambda = 0.01$ for feature consistency, covariance, and dominant singular value regularization and use a two-layer MLP with 800 units for all our experiments. To save costs, we perform PCA on the embeddings to 200 components. We use Adam with lr=10^{-3} for 10 epochs.

EWC-FAISS The best-performing index for the WBC dataset used a combination of SAM, ConvNeXT, SWIN, and CLIP with k between 3 and 1000, an entropy threshold of 0.3, and L2-distance. For the transfer to LISC, we use a combination of DINO, ConvNeXT, SWIN, and ViTMAE. For the cell state classification from the Nanolive microscope, we use a combination of SAM, ConvNeXT, SWIN, CLIP and ViTMAE with k between 3 and 100, an entropy threshold of 0.6 and Canberra distance. For the transfer to the Lionheart microscope, we set k between 10 and 1000 and an entropy threshold of 0.1. For the cell type classification from the Lionheart microscope, as well as for the transfer to the Nanolive microscope, we use a combination of SAM, DINO, ConvNeXT, and CLIP, with k between 20 and 1000, an entropy threshold of 0.2 and L2-distance.

C. Feature Analysis

Table 2 shows the mean intra-class similarities and Table 3 the mean inter-class similarities (in terms of cosine similarity) of all foundation models and their best combination for the embeddings of the WBC dataset.

Model	Intra-class Similarity
Best Combination	0.355
SWIN	0.349
ConvNeXT	0.33
CLIP	0.244
DINO	0.183
ViTMAE	0.179
SAM	0.144

Table 2: Mean intra-class similarities.

Model	Inter-class Similarity
Best Combination	-0.043
SWIN	-0.041
ConvNeXT	-0.04
CLIP	-0.027
ViTMAE	-0.021
SAM	-0.019
DINO	-0.014

Table 3: Mean inter-class similarities.