Revisiting Automatic Data Curation for Vision Foundation Models in Digital Pathology

Boqi Chen[*12], Cédric Vincent-Cuaz[*3], Lydia A. Schoenpflug[*4], Manuel Madeira[*3], Lisa Fournier[6], Vaishnavi Subramanian[3], Sonali Andani[145], Samuel Ruiperez-Campillo[1], Julia E. Vogt[12], Raphaëlle Luisier[6], Dorina Thanou[3], Viktor H. Koelzer[†45], Pascal Frossard[†3], Gabriele Campanella[†7], and Gunnar Rätsch[†12]

[1] Dept. of Computer Science, ETH Zurich, Zurich, Switzerland
[2] AI Center, ETH Zurich, Zurich, Switzerland
[3] Signal Processing Laboratory (LTS4), EPFL, Lausanne, Switzerland
[4] University of Basel, Basel, Switzerland
[5] University Hospital of Basel, Basel, Switzerland
[6] Idiap Research Institute, Martigny, Switzerland
[7] Icahn School of Medicine at Mount Sinai, New York, United States

Abstract. Vision foundation models (FMs) are accelerating the development of digital pathology algorithms and transforming biomedical research. These models learn, in a self-supervised manner, to represent histological features in highly heterogeneous tiles extracted from whole-slide images (WSIs) of real-world patient samples. The performance of these FMs is significantly influenced by the size, diversity, and balance of the pre-training data. Yet, data selection has been primarily guided by expert knowledge at the WSI level, focusing on factors such as disease classification and tissue types, while largely overlooking the granular details available at the tile level. In this paper, we investigate the potential of unsupervised automatic data curation at the tile-level, taking into account 350 million tiles. Specifically, we apply hierarchical clustering trees to pre-extracted tile embeddings, allowing us to sample balanced datasets uniformly across the embedding space of the pretrained FM. We further show that these datasets are subject to a trade-off between size and balance, potentially compromising the quality of representations learned by FMs. We propose tailored batch sampling strategies to mitigate this effect. We demonstrate the effectiveness of our method through improved performance on a diverse range of clinically relevant downstream tasks.

Keywords: Automatic Data Curation · Pathology Foundation Model

1 Introduction

Large-scale pre-trained self-supervised learning (SSL) models, or foundation models (FMs), have demonstrated a remarkable ability to learn task-agnostic

[*] Equal contribution
[†] Co-corresponding authors: viktor.koelzer@usb.ch; pascal.frossard@epfl.ch; gabriele.campanella@mssm.edu; raetsch@ethz.ch

representations from unlabeled image data, capturing rich domain-specific knowledge [5,12]. Their exposure to vast and diverse data sources enables them to learn highly transferable representations for various downstream tasks [21]. This property is particularly attractive for digital pathology (DP), which involves analyzing high-resolution whole-slide images (WSIs) with significant heterogeneity across different biological scales to assess clinically relevant tasks, such as cancer typing and grading, survival prediction and treatment response assessment.

Data curation plays a crucial role in training FMs, leading to improved performance [17]. Strategies such as data pruning [20,18], active learning [10,22], and nearest neighbor search [17] have proven effective. Recently, clustering-based methods have emerged as competitive alternatives to traditional supervised algorithms, while being fully automated [21]. In contrast, data curation in DP largely relies on expert annotations that provide high-level information at the WSI level (*e.g.*, cancer and tissue types [14]), while potentially overlooking tile-level granularity. To this end, a semi-automatic strategy has been proposed recently, combining WSI labels and weak tile-level supervision based on color statistics [8]. Despite this progress, the need for expert intervention still limits scalability, underscoring the necessity of a fully automated data curation framework tailored to the heterogeneity and annotation challenges of DP.

In this work, we fill this gap by studying unsupervised automatic data curation for DP using a large set of 350 million tiles. Inspired by [21], we employ *hierarchical clustering trees* on tile embeddings extracted using existing FMs. These hierarchical clusters enable uniform coverage of the data distribution across the embedding space and facilitate efficient sampling of diverse, curated datasets, while allowing control over their balance and size. This approach mitigates biases stemming from over- or under-represented patterns and permits more efficient selection of representative samples. However, our empirical results indicate that directly applying this curation method in DP does not consistently enhance the discriminative power of learned FM embeddings compared to using uncurated data. We identify the data feeding strategy of the SSL model in [17] as the limiting factor and address it by introducing batch stratification based on hierarchical clusters, which forces models to continuously learn how to represent more diverse samples while mitigating the biases resulting of the imbalance present in our heavy-tailed data distribution. Our full pipeline is depicted in Figure 1. To summarize, our contributions are three-fold: 1) we present, to the best of our knowledge, the first *fully automated* data curation scheme for FM training in DP; 2) we identify batch stratification, alongside hierarchical clustering-based curation, as the key to data distribution uniformization and improved downstream performance; 3) we demonstrate the effectiveness of our method through improved performance on both region of interest (RoI)- and WSI-level benchmarks. Our code can be found on GitHub[8].

[8] https://github.com/swiss-ai/health-pathology/tree/main

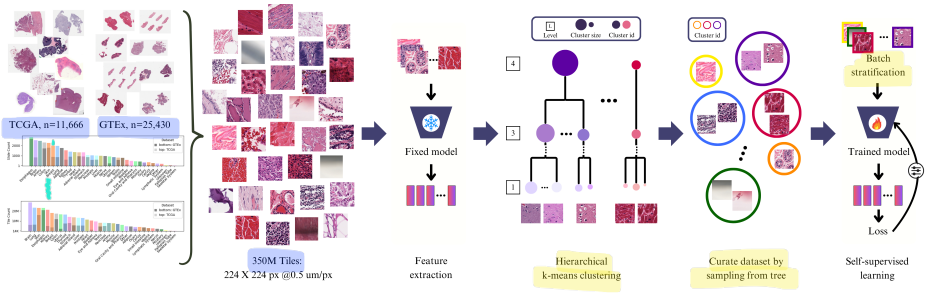

Fig. 1. Overview of our pipeline for automatic data curation and SSL training in DP.

2 Method

In this section, we describe our automatic data curation framework for training FMs in DP. We first detail the pre-training dataset and the curation procedure, and then explain how to effectively utilize the curated data for FM training.

2.1 Hierarchical Clustering for Automatic Data Curation

We build a large pre-training dataset by collecting 11,666 WSIs from The Cancer Genome Atlas (TCGA) cohortspanning 32 cancer types, and 25,430 WSIs from the Genotype-Tissue Expression (GTEx, v7) dataset across 40 healthy tissue sites. Following [4], we preprocess WSIs at $\times 20$ magnification (*i.e.*, 0.5 μm/pixel) by first detecting tissue regions and extracting $\sim$350 million non-overlapping 224×224 pixel tiles. Tissue-specific distributions are shown in Figure 1.

Inspired by [21], we first extract tile embeddings from the pre-training dataset using a well-established FM in DP (*i.e.*, UNI [6]) and construct a hierarchical clustering tree in a bottom-up manner. At the bottom level, K-means is applied to all tile embeddings, generating more clusters and of relatively smaller volumes in dense areas than in sparse ones. Data imbalance is artificially reduced, while the average nearest-neighbor distance is increased, by selecting the cluster centroids as the new data distribution. We recursively apply K-means to these centroids with an exponentially decreasing number of clusters, ensuring more uniform cluster volumes at each hierarchical level. This recursion provably yields *clusters that distribute more uniformly over the data support as one ascends the hierarchy* [17]. Next, a top-down sampling strategy is applied to sample curated subsets of size N. First, sample sizes are allocated to the k top-level clusters $\{c_i\}_{i=1}^{k}$ of respective total sizes $\{s_i\}_{i=1}^{k}$, by finding the size n which minimizes $|N - \sum_i^k min(n, s_i)|$ through a binary search in $\{0, \ldots, N\}$. This process is then repeated for each cluster c_i independently, distributing $min(n, s_i)$ samples across its respective downstream clusters. These two steps define a recursion applied until reaching the bottom-level clusters, where the allocated points are *sampled randomly*. Overall, this method allows us to efficiently sample subsets of great diversity, while controlling the trade-off between balance and size.

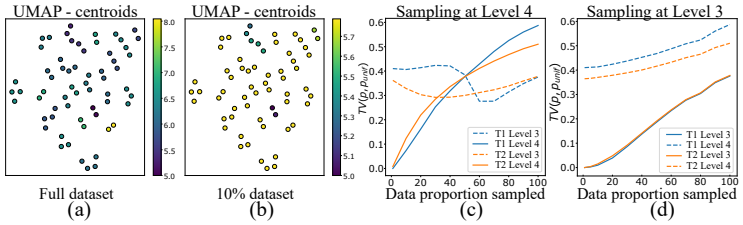

Fig. 2. (a, b) UMAP embeddings of T1 centroids at level 4. Colors represent cluster sizes (logarithmic scale) for the full and the 10% curated dataset, respectively. (c, d) Visualization of hierarchical sampling dynamics for T1 and T2. Solid lines denote the sampling level and dashed lines are the corresponding dynamics at the alternative level.

Since no principled method exists to select an optimal size for a given tree configuration, nor its depth and width that influence this trade-off, we investigate two tree configurations across various subset sizes. First, following [21], we set the depth to 4 and the number of leaves to 1% of our dataset. Then, we select 62 and 2048 as the top-level cluster counts, yielding the following cluster distributions per tree level: **T1** : $\{3.5M, 35k, 350, 62\}$ and **T2** : $\{3.5M, 100k, 10k, 2048\}$. We note that T2 closely mirrors the hyperparameters in [21], with a higher top-level cluster count (equals to the global batch size, see Section 2.2) and branch expansion rates, whereas T1 has fewer top-level clusters, following insights from expert annotations [14], thereby requiring greater compression at each level. In Figure 2, we use these trees to analyze the data distribution and illustrate the effects of hierarchical sampling. Figure 2(a) presents UMAP embeddings of top-level centroids from T1, with hyperparameters chosen to best preserve local relative positions. Centroids are colored based on their effective size without sampling. The results indicate that top-level clusters indeed cover rather uniformly our distribution and that it has heavy tails with 2 modes encompassing approximately two-thirds of the dataset. The results from sampling 10% of the data (Figure 2(b)) show that while global diversity is preserved, the curated subset remains slightly imbalanced. We further analyze the trade-offs between balance and size via the total variation distance (TV) between sampled cluster proportions (p) and uniform weights (p_{unif}) across two clustering levels. Results shown in Figure 2(c) and (d) reveal that the large imbalance in our dataset implies a significant imbalance for all curated subsets with more than 1% data, since the smallest top-level clusters are depleted during sampling. We also observe that sampling from any level results in a nearly maximal imbalance at the other level studied, as densest level 4 clusters contain more clusters at level 3.

2.2 Effective Self-supervised Learning on Curated Data

After the data curation, we proceed to the self-supervised training of FMs using the curated datasets. We follow existing works on FMs in DP [23,24,6,8] and adopt DINOv2 [17] as our SSL framework. We utilize large Vision Transformers

(ViT-L) [9] and train it for 170 thousand iterations with a global batch size of 2048, corresponding to one complete pass over the full dataset (350 million tiles). Typically, these batches are randomly sampled from the pre-training dataset (curated or not), as in most FMs training with automatic data curation [21] or without it [17,23,24,6]. Instead, we propose to structure these batches during training by stratifying them based on the tree clusters from which the tiles were sampled. In particular, each batch is constructed to contain an equal number of tiles from each top-level cluster, leveraging the inherent balancing of curation clusters to guarantee uniform coverage of the data distribution within batches. This approach ensures that differently populated clusters are equally represented during training, prompting models to continually adapt to diverse samples while mitigating biases introduced by dataset imbalance. To further promote intra-cluster diversity in batch construction, we track the samples encountered during training and prioritize those that have been less frequently sampled: for each batch and cluster, we sample exclusively from the least observed tiles within that batch and cluster. This procedure guarantees that every tile in the curated dataset is observed multiple times during training.

3 Experiments

In this section, we first examine whether the clusters identified by hierarchical clustering exhibit interpretable biological features. We then present a comprehensive benchmark of data curation and learning strategies, followed by a more detailed investigation the sensitivity of our method to its hyperpameters.

Interpretability of hierarchical clusters. To study the interpretability of hierarchical clusters, we analyze the 1% curated subset using T1 at level 4. We first conduct visual inspections of tile images in the clusters, an example of which is shown in the supplementary video. We then follow [25] to quantitatively report our findings, by extracting basic statistics for handcrafted single-cell features (color intensities, textures, morphologies, spatial arrangements) and including class densities from CellViT [13] and extracellular matrix descriptors. We observe that ∼50% highest-level clusters mainly contain tiles of healthy tissues from GTEx and a few TCGA tiles without neoplastic cells, while the two largest clusters are mainly cancerous TCGA tiles with ∼50% cells neoplastic or inflammatory and the remainder epithelial or connective tissue cells. Next, we cluster the tiles using handcrafted features by applying UMAP, followed by K-means with the same number of clusters as T1's top level. The resulting clusters are then compared to those obtained from hierarchical clustering on raw embeddings using the Adjusted Rand Index (ARI). Our experiments show that both types of clusters match weakly with ARI scores of 7.7% at level 4 and 5.5% at level 3. Moreover, restricting the same process to specific types of handcrafted features reveals that cell colors and spatial arrangements are relatively the most discriminative factors across hierarchical clusters. Overall, these results suggest that while handcrafted features explain the hierarchical clusters to some extent,

they do not fully capture the underlying patterns dictating this organization.

Evaluation framework. In the following, we evaluate different data curation settings at both RoI- and WSI-level. Specifically, 8 RoI-level tasks from independent cohorts are considered, including lung adenocarcinoma classification (LUAD [11]), colorectal tissue and polyp classification (CRC [16], UniToPatho [2], Chaoyang [26]), breast cancer subtyping (BRACS [3]) and tissue classification (BACH [1], BreakHis [19]), and lymph node metastatis classification (PCAM). All RoI images are resized to 224×224 pixels before embedding extraction. We perform linear probing over 1000 nonparametric bootstrap iterations on embeddings extracted from each encoder. At the WSI-level, we assess 9 clinically relevant tasks [7], including breast cancer (BCa) detection, and biomarker prediction for Estrogen Receptor (ER), Progesterone Receptor (PR) and Human Epidermal Growth Factor Receptor (HER2) and breast Homologous Repair Deficiency status (HRD). Further tasks include prediction of lung Epidermal Growth Factor Receptor (EGFR) mutation status and immunotherapy (IO) response in lung cancer patients, and, lastly, detection of inflammatory bowel disease (IBD) versus normal mucosa samples. We represent each WSI as a set of tile embeddings composing its segmented tissue regions and train ABMIL [15] models over 20 Monte Carlo cross-validation runs [15]. We report average balance accuracy (bACC, in %) and area under the receiver operating characteristic curve (AUC, in %) for RoI- and WSI-level tasks, respectively [6,7].

Benchmarked methods. We compare several data curation and batch sampling strategies. The baseline method, F-BR, learns from the *full dataset* with random batch sampling. We also include a *supervised data curation* approach where WSIs are first split into 266 classes, with healthy tissue types for GTEx and combinations of tissue type and primary cancer diagnosis for TCGA, before sampling tiles evenly across classes. Two methods, denoted S-BR and S-BS, learn from these curated datasets using random and stratified batch sampling, respectively. Finally, for the *automatic data curation* introduced in Section 2.1, we select the curated tiles by sampling from T1 and T2 at level 4. For training, we employ both random and stratified batch sampling, denoting these methods as T-BR and T-BS, respectively, with $T \in \{\text{T1}, \text{T2}\}$ indicating the tree used. Unless stated otherwise, we use curated subsets with 10% of the total data.

Main results. The performance on the RoI- and WSI-level benchmarks are presented in Tables 1 and 2, respectively. We can observe that our approach, T1-BS, achieves the highest average performance across both benchmarks, followed by S-BS and F-BR. In contrast, T1-BR and S-BR yield the lowest average performance. These results emphasize the effectiveness of jointly applying automatic data curation at the tile level and stratified batch sampling strategies. Nevertheless, rankings per task still vary significantly. For further analysis, we note that our benchmarks predominantly consist of images related to breast, colon, and lung tissues. Thus, we can expect any inherent bias towards these tissue types in benchmarked methods could lead to improved performance. To better disentangle model performance from such biases, we first provide a detailed breakdown of

Table 1. RoI-level evaluation. Best results are in **bold** and second best are underlined.

Setting	Lung	Breast				Colon			Overall
	LUAD	BRACS	BreakHis	BACH	PCAM	CRC	UniToPatho	Chaoyang	
F-BR	**94.2**	66.1	96.9	85.0	**91.1**	86.8	42.4	76.7	79.9
S-BR	93.7	61.9	96.4	88.5	90.9	86.8	37.6	74.8	78.9
S-BS	94.1	62.4	**98.3**	87.4	90.8	90.8	**43.7**	**78.8**	80.8
T1-BR	93.8	63.8	96.6	88.8	90.6	87.4	41.0	76.5	79.8
T1-BS	**94.2**	**69.3**	97.5	**91.2**	90.8	**91.9**	**43.7**	77.1	**82.0**

the dataset compositions used during training. In the full dataset, breast, colon, and lung tissues account for 5.2%, 5.1%, and 6.7% of tiles, with 61% coming from TCGA, which are more likely cancer tissues. In the supervised curation subset, these proportions shift to 6.1%, 3.2%, and 9.3%, with 80% from TCGA, dispersed across 58 of 266 classes. The hierarchical clustering subset contains 2.7%, 3.9%, and 3.9%, with 53% from TCGA, covering all level-4 clusters. From a global perspective, both T1-BS and T1-BR utilize on average the smallest proportions of tiles from these tissues while maintaining competitive performance, underscoring the effectiveness of automatic data curation in enhancing diversity in sampling. Among the most comparable settings, *e.g.* T1-BS and S-BS, we observe strong correlations between performance and both tile proportions and their origins. Both methods use similar tile proportions from GTex for breast and lung tiles, but S-BS uses increasingly more TGCA tiles for each tissue. However, T1-BS outperforms S-BS on RoI-level tasks related to breast tissue, achieving an average balanced accuracy of 87.2% compared to 84.7% for S-BS, while demonstrating slightly better performance on the LUAD lung dataset. Notably, T1-BS leads to the largest improvement on the most challenging task, BRACS, with 7 cancer subtypes. For colon tissues, S-BS uses fewer tiles than T1-BS yet their average performance remains comparable at 71.1% and 70.9%, respectively. T1-BS actually tends to outperform S-BS in benchmarks with greater tissue diversity, achieving superior performance on CRC (9 classes) while performing similarly or slightly worse on UniToPatho (6 classes) and Chaoyang (4 classes), suggesting its potential for training a better generalist FM.

For WSI-level tasks, we recall that AUC considers model calibration and therefore could accentuate the effects of the biases discussed above. Nonetheless, we observe similar trends in performance reinforcing that T1-BS better transfers across tasks. For breast-related biomarker prediction (ER, HER2, PR, HRD) and cancer detection (BCa), T-BS achieves slightly higher average AUC of 86.8% and 97.6%, compared to 86.4% and 97.5% for S-BS. In lung-related benchmarks (EGFR and IO), T1-BS outperforms S-BS by a larger margin of 0.9% in average AUC. Notably, T1-BS shows greatest improvement in the two most challenging tasks: HRD and IO prediction in breast- and lung-related benchmarks.

Sensitivity analysis. We analyze the sensitivity of both T-BS and T-BR to sampling strategies. Figure 3 reports the RoI-level performance of models trained

Table 2. WSI-level evaluation. Best results are in **bold** and second best are underlined.

Setting	Biomarker							Detection		Overall
	Site 1					Site 2		Site 1		
	ER	HER2	PR	HRD	EGFR	EGFR	IO	BCa	IBD	
F-BR	96.5	80.5	91.3	74.4	**73.3**	**76.2**	57.3	**97.6**	96.5	82.6
S-BR	96.1	79.6	91.1	72.8	71.6	75.3	56.7	97.2	96.1	81.8
S-BS	**96.6**	**81.0**	**91.7**	76.2	72.0	75.3	57.9	97.5	**97.0**	82.8
T1-BR	95.7	80.1	89.9	76.2	70.1	75.8	52.9	97.1	95.9	81.5
T1-BS	96.2	80.8	91.2	**79.1**	73.0	75.9	**59.0**	**97.6**	96.7	**83.3**

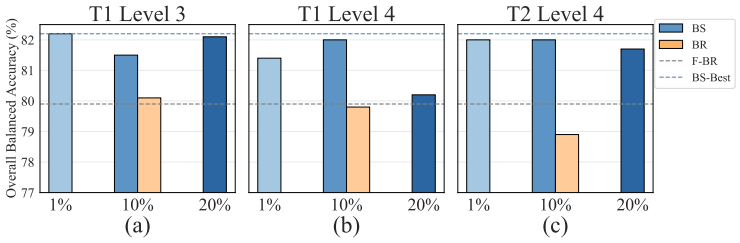

Fig. 3. RoI-level performances across clustering trees and batch sampling strategies.

on subsets curated using different trees, sampling levels and data proportions. First, with 10% subsets from different (sub-)trees, our stratified batch sampling consistently outperforms the random one. Focusing on T-BS, top performances remain comparable across trees, with 82.2% average bACC when sampling from T1 at level 3 and 82.0% otherwise. However, for a given tree, performance appears sensitive to the subset sizes, particularly with fewer clusters (T1) for batch stratification. We observe that this sensitivity stems from the alignment of tissue proportions between the pre-training curated subsets and those in the downstream evaluation datasets. Hence, improving FMs as generalist models motivates sampling strategies that maximize both diversity and balance, with narrower trees offering easier control as their tissue proportions converge more slowly to the full dataset. Optimizing this trade-off can be achieved with a minimal subset size by allocating samples accordingly to the volume of each bottom-level cluster (clusters covering smaller volumes get fewer samples), as attained by uniform sampling from T1 at level 4. Additionally, since the relative positions of embeddings dictate the cluster coverage of the data support, more informative bottom-level cluster sampling (beyond random) and batch construction (beyond top-level cluster-based) are promising directions for future work.

4 Conclusion

We investigate automatic data curation for FM training in DP, and propose a tailored batch sampling strategy to better convey to models the data diversity captured by hierarchical clusters and mitigate biases from data imbalance. Our results demonstrate the effectiveness of our method and highlight the relevance of data heterogeneity in DP. The study of our method is based on the assumption that both data curation and batch sampling leverage UNI's informative embeddings. However, the confidence attributed to their relative positions is low, given that sampling within bottom-level clusters is random and batch stratification does not take into account the full tree hierarchy. We plan to further investigate these effects using different pre-trained FMs, additional tree configurations and more informative sampling strategies for both curation and learning. Finally, we intend to study in more detail the transformations that our approach implies in the embedding and its applicability to continual SSL.

Acknowledgments. We gratefully acknowledge funding from the ETH AI Center, the Idiap Research Institute, and the Swiss Federal Institutes of Technology strategic focus on personalized health and related technologies, as well as the EPFL and ETH core funding. We also acknowledge funding from the Novartis Foundation for Medical-Biological Research (grant no. 22B104) and from the Swiss AI Initiative through a grant by the Swiss National Supercomputing Centre (CSCS) under project ID a02 on Alps.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to this article.

References

1. Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S.S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., et al.: Bach: Grand challenge on breast cancer histology images. Medical image analysis **56**, 122–139 (2019) 6
2. Barbano, C.A., Perlo, D., Tartaglione, E., Fiandrotti, A., Bertero, L., Cassoni, P., Grangetto, M.: Unitopatho, a labeled histopathological dataset for colorectal polyps classification and adenoma dysplasia grading. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 76–80. IEEE (2021) 6
3. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022**, baac093 (2022) 6
4. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. **25**(8), 1301–1309 (Aug 2019) 3
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 2

6. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024) 3, 4, 5, 6
7. Chen, S., Campanella, G., Elmas, A., Stock, A., Zeng, J., Polydorides, A.D., Schoenfeld, A.J., Huang, K.l., Houldsworth, J., Vanderbilt, C., et al.: Benchmarking embedding aggregation methods in computational pathology: A clinical data perspective. arXiv preprint arXiv:2407.07841 (2024) 6
8. Dippel, J., Feulner, B., Winterhoff, T., Milbich, T., Tietz, S., Schallenberg, S., Dernbach, G., Kunft, A., Heinke, S., Eich, M.L., et al.: Rudolfv: a foundation model by pathologists for pathologists. arXiv preprint arXiv:2401.04079 (2024) 2, 4
9. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 5
10. Geifman, Y., El-Yaniv, R.: Deep active learning over the long tail. arXiv preprint arXiv:1711.00941 (2017) 2
11. Han, C., Pan, X., Yan, L., Lin, H., Li, B., Yao, S., Lv, S., Shi, Z., Mai, J., Lin, J., et al.: Wsss4luad: Grand challenge on weakly-supervised tissue semantic segmentation for lung adenocarcinoma. arXiv preprint arXiv:2204.06455 (2022) 6
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 2
13. Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., et al.: Cellvit: Vision transformers for precise cell segmentation and classification. Medical Image Analysis **94**, 103143 (2024) 5
14. Hosseini, M.S., Chan, L., Tse, G., Tang, M., Deng, J., Norouzi, S., Rowsell, C., Plataniotis, K.N., Damaskinos, S.: Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11747–11756 (2019) 2, 4
15. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018) 6
16. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine **16**(1), e1002730 (2019) 6
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 2, 3, 4, 5
18. Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., Morcos, A.: Beyond neural scaling laws: beating power law scaling via data pruning. Advances in Neural Information Processing Systems **35**, 19523–19536 (2022) 2
19. Spanhol, F.A., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. Ieee transactions on biomedical engineering **63**(7), 1455–1462 (2015) 6
20. Toneva, M., Sordoni, A., Combes, R.T.d., Trischler, A., Bengio, Y., Gordon, G.J.: An empirical study of example forgetting during deep neural network learning. arXiv preprint arXiv:1812.05159 (2018) 2

21. Vo, H.V., Khalidov, V., Darcet, T., Moutakanni, T., Smetanin, N., Szafraniec, M., Touvron, H., Couprie, C., Oquab, M., Joulin, A., et al.: Automatic data curation for self-supervised learning: A clustering-based approach. arXiv preprint arXiv:2405.15613 (2024) 2, 3, 4, 5
22. Vo, H.V., Siméoni, O., Gidaris, S., Bursuc, A., Pérez, P., Ponce, J.: Active learning strategies for weakly-supervised object detection. In: European Conference on Computer Vision. pp. 211–230. Springer (2022) 2
23. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., et al.: Virchow: A million-slide digital pathology foundation model. arXiv preprint arXiv:2309.07778 (2023) 4, 5
24. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature pp. 1–8 (2024) 4, 5
25. Zhao, S., Chen, D.P., Fu, T., Yang, J.C., Ma, D., Zhu, X.Z., Wang, X.X., Jiao, Y.P., Jin, X., Xiao, Y., et al.: Single-cell morphological and topological atlas reveals the ecosystem diversity of human breast cancer. Nature Communications **14**(1), 6796 (2023) 5
26. Zhu, C., Chen, W., Peng, T., Wang, Y., Jin, M.: Hard sample aware noise robust learning for histopathology image classification. IEEE transactions on medical imaging **41**(4), 881–894 (2021) 6