The Importance of Downstream Networks in Digital Pathology Foundation Models

Gustav Bredell, Marcel Fischer, Przemyslaw Szostak, Samaneh Abbasi-Sureshjani[0000−0003−4150−6414], and Alvaro Gomariz[0000−0002−6172−5190]

F. Hoffmann-La Roche AG, Basel, Switzerland

Abstract. Digital pathology has significantly advanced disease detection and pathologist efficiency through the analysis of gigapixel whole-slide images (WSI). In this process, WSIs are first divided into patches, for which a *feature extractor* model is applied to obtain feature vectors, which are subsequently processed by an *aggregation model* to predict the respective WSI label. With the rapid evolution of representation learning, numerous new feature extractor models, often termed foundational models, have emerged. Traditional evaluation methods rely on a static downstream aggregation model setup, encompassing a fixed architecture and hyperparameters, a practice we identify as potentially biasing the results. Our study uncovers a sensitivity of feature extractor models towards aggregation model configurations, indicating that performance comparability can be skewed based on the chosen configurations. By accounting for this sensitivity, we find that the performance of many current feature extractor models is notably similar. We support this insight by evaluating seven feature extractor models across three different datasets with 162 different aggregation model configurations. This comprehensive approach provides a more nuanced understanding of the feature extractors' sensitivity to various aggregation model configurations, leading to a fairer and more accurate assessment of new foundation models in digital pathology.

1 Introduction

Digital pathology (DP) has significantly advanced with automated solutions for tasks like breast cancer [7] and metastases detection [15], leveraging gigapixel whole-slide images (WSI) stained with H&E. The challenge of applying standard deep learning models for processing these large images has led to the adoption of the multiple instance learning (MIL) framework. In MIL, as depicted in step 1 in Figure 2, WSIs are divided into patches, also known as tiles. A *feature extractor* model extracts features from each tile to generate embedding vectors. These vectors, collectively referred to as a *bag*, are then processed by an *aggregation model* to predict the WSI label [16, 9]. Popular choices for aggregation models include AttentionMIL [13] and TransMIL [18], which both rely on using attention mechanisms for feature aggregation.

Beyond computational efficiency, feature extractors play a critical role in overcoming the scarcity of labeled data in DP. Using representation learning approaches feature extractors can be trained on large datasets of unlabeled images enabling their use across diverse datasets. Since the pivotal work of Chen et al. [5], which significantly improved visual representations using contrastive learning (SimCLR), a range of novel representation learning approaches has been introduced. SimCLR learns representations by ensuring that the embeddings of images with the same label (positive examples) are close, whereas the embeddings of images with different labels (negative examples) are far apart. Subsquently, Grill et al. [11] showed that self-supervised learning can also be done without negative examples (BYOL). This approach was further improved and combined with transformers leading to DINO [3]. The most recent approaches combine masked autoencoder (MAE) [12] with self-distillation. This is the strategy used by iBOT [23] and is also at the core of DINOv2 [17].

The DP field has adapted these representation learning advancements, notably in CTransPath [21] and REMEDIS [1]. Due to the large number of tiles that can be extracted from a single WSI and the availability of large publicly available datasets, such as TCGA [19], the datasets for CTransPath and REMEDIS contain 16 Mio and 50 Mio tiles, respectively. These large datasets allow the development of superior feature extractors, now commonly known as *foundational models*, that generalize across datasets without the need for re-training. Filiot et al. [8] made a first step in this direction and demonstrated better classification performance compared to CTransPath by extracting feature embeddings using a model trained with iBOT. Chen et al. [4] went a step further and increased the dataset size to 100 Mio tiles while using DINOv2 to train the feature extraction model. Finally, one of the most recent feature extractors, Virchow [20], was also trained using the DINOv2 approach but on a dataset size of 380 Mio tiles.

Foundational models promise to extract informative features from patches across diverse datasets. Ideally, capturing relevant features enhances downstream tasks, such as classification, while poor features hinder it. Feature extractors are often evaluated through their performance in basic classification tasks using models such as linear or K-NN classification [6]. However, in digital pathology, the use of an aggregation model to process embedding vectors and make final predictions can complicate the assessment of the feature extraction quality.

As illustrated with an example in Figure 1, our analysis reveals that, whereas foundational models do indeed have some influence on the classification performance, they are highly sensitive to the aggregation model configuration. Thus, when comparing feature extractors, the sensitivity to the second step of the classification pipeline, namely the aggregation model, is an important variable to control for. Our contribution in this paper is twofold.

- We characterize the feature extractors' sensitivity to various aggregation model configurations, challenging traditional feature extractor evaluation methods in digital pathology.
- We propose a framework for stringent and fair evaluation for state-of-the-art feature extractors.

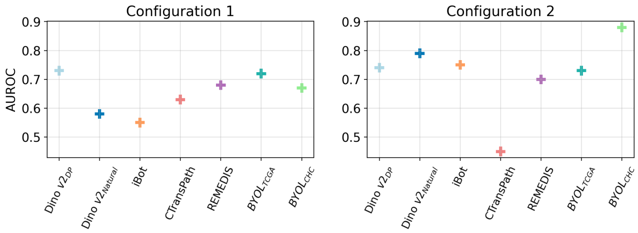

Fig. 1. Typical frameworks for evaluation of feature extraction models use fixed configurations in the aggregation models, leading to substantially different results and hence limited informative value.

2 Experimental Setup and Methods

This section outlines the classification pipeline for whole-slide images (WSIs) and outlines the framework we use for evaluating the sensitivity of feature extractors towards aggregation model configurations.

2.1 Pipeline for classification of WSIs

As depicted in Figure 2, the typical MIL pipeline for DP requires two models to obtain a final classification for a given WSI. First, a feature extraction model leverages recent self-supervised learning advancements and extensive datasets to generalize across tasks and datasets [2, 20, 4]. This model is applied to tiles in a WSI to produce feature embedding vectors. Next, a smaller aggregation model, specific to each dataset, processes the extracted embeddings to aggregate information and classify the WSI. In contrast to the feature extraction model, this aggregation model is re-trained for each dataset.

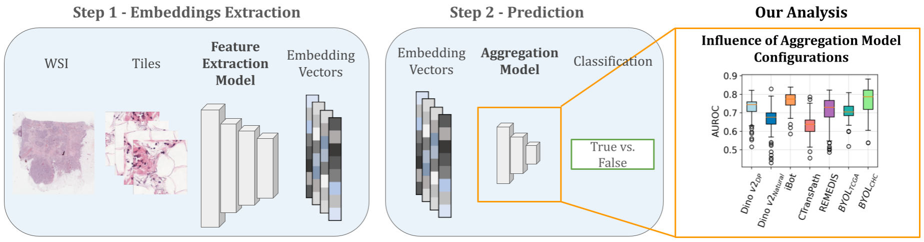

Fig. 2. Illustration of the typical classification pipeline with MIL in digital pathology.

Feature extraction models have been compared under a single aggregation model configuration, i.e. fixed model architecture and hyperparameters [4, 14, 1]. Figure 1 illustrates the significant impact of aggregation model configuration choice on performance, rendering widely adopted evaluation frameworks suboptimal. Indeed, fixed aggregation model configurations can favour some feature extraction models while penalizing others. We outline an experimental setup to thoroughly address two critical questions:

Question 1: Can a single aggregation model configuration optimally support various feature extraction models?

Question 2: How do state-of-the-art feature extraction models perform relative to each other when controlling for different aggregation model configurations?

2.2 Feature Extraction Models

To explore our research questions, we assess seven feature extraction models, with details and characteristics provided in Supplementary Table S1.

We begin by evaluating the ViT-L model from DINOv2, trained on 142 million natural images with 300 million parameters, to assess the applicability of models trained on natural images for DP [17]. Additionally, we explore recently published models specifically designed for DP: CTransPath [21], REMEDIS [1], and iBOT [8] (teacher model), all trained on extensive DP datasets as detailed in Supplementary Table S1. Lastly, we investigate three feature extracting models trained in-house.

We train from scratch a DINOv2 ViT-L model on TCGA and a vast inhouse dataset with diverse tissue types and real-world data. The total training dataset size is 35 million 224×244 tiles. WSIs are usually acquired at different magnifications. Tiles from 20× magnification offer broader content, while 40× magnification tiles provide finer tissue details due to their higher resolution. We train at both 20× and 40× magnifications to capture different features, following successful strategies in the literature [14]. We employ the official DINOv2 repository with the the default parameters for ViT-L/16 training, with a few exceptions. We decrease the batch size to 352 due to computational constraints. Due to the decreased batch size, we increase the number of epochs to 270, warmup epochs to 25 and adjust the learning rate to 1.375×10^{-3} according to the heuristic by Goyal et al. [10].

Two ResNet-50 models are trained using BYOL: The first, $\text{BYOL}_{\text{TCGA}}$, utilizes 2 million tiles randomly sampled from the TCGA dataset at 20× magnification, featuring a smaller dataset and model size for comparison. The second, BYOL_{CHC} (according to the first letter of each of the three evaluation datasets), is also trained on 2 million tiles but randomly sampled from the training set of the evaluation datasets. Thus even though the training dataset is small compared to the other published models, there is no domain gap between the dataset on which it is trained and evaluated on. This strategy ensures direct relevance to the evaluated datasets, potentially offsetting the smaller scale of the model with its domain specificity.

2.3 Aggregation Model Configurations

To investigate the performance fluctuation of the MIL pipeline when the feature extraction model is fixed and the aggregation model configuration change, we use different network hyperparameters and two well adopted aggregation model architectures: AttentionMIL [13], which uses an attention mechanism to aggregate tile information and assumes no interdependency between the tiles. TransMIL [18], which learns inter-tile dependencies by using the self-attention mechanism of transformers, in particular the Nyströmformer [22]. We change four hyperparameters with three distinct values each as shown in Table 1. These are decided heuristically with preliminary experiments assessing the influence and effective range of each hyperparameter. The resulting 162 different configurations (81 for each of the 2 architectures) are outlined below.

Table 1. Set of hyperparameter values for each aggregation model. Layers refer to fully connected layers in AttentionMIL and to attention blocks in TransMIL.

Hyperpar.	**AttentionMIL**	**TransMIL**
Learn. rate	1e-4, 1e-3, 1e-2	1e-5, 1e-4, 1e-3
Bag size	128, 1024, 8192	128, 1024, 2048
Layers	(512), (512, 384, 384), (512, 256, 128, 64, 32)	1, 2, 3
Dropout	0.00, 0.25, 0.50	0.00, 0.25, 0.50

AttentionMIL: When creating a batch for the aggregation model during training, there are two relevant parameters. One is the *bag size*, which determines the amount of tiles that is sampled from a particular WSI. The second is the *bags per batch*, determining how many bags from different WSIs are collected to form a batch. Here, we vary only the bag size parameter since it showed a larger influence. The final batch size=*bag size***bags per batch*. The *Layers* parameter corresponds to the number of nodes in the fully connected (FC) layers in the aggregation model. The list of numbers in Table 1 indicate the number of nodes for each layer. Lastly, the dropout parameter refers to the dropout which is applied at every layer of the aggregation model.

TransMIL: The selected hyperparameters for TransMIL are different due to the model architecture being a transformer, which does not employ FC layers. *Layers* refers to the number of Nyströmformer attention blocks. We also reduce the maximal bag size to 2048 due to computational limitations.

Both models share fixed training parameters: a weight decay of 10^{-5}, four *bags per batch*, AdamW optimizer, weighted cross entropy loss, and a cosine annealing scheduler. Aggregation models are trained for 50 epochs to ensure convergence within our configuration range.

2.4 Evaluation Datasets

Our study evaluates binary classification performance of feature extraction models across three distinct DP datasets. Thereby providing a more generalizable answer to our research questions. These datasets, comprising H&E-stained histopathology slides WSIs, allow us to assess each feature extractor under 162 different aggregation model configurations. This comprehensive approach, covering 7 feature extractors, 162 aggregation model configurations, and 3 datasets, culminates in a total comparison of $7 \times 162 \times 3 = 3402$ experimental configurations.

Performance metrics include the area under the receiver operating characteristic curve (AUROC) and average precision (AP), both ranging from 0 to 1. A higher AUROC indicates superior distinction between the positive and negative classes, while a higher AP reflects more accurate predictions of positive instances across all recall levels, effectively balancing precision and recall. Performance metrics are derived from the test set, using the aggregation model's epoch with best validation score during training.

COO: Binary classification of cell of origin (COO). Each image contains the COO prediction label of activated B-cell like (ABC) or germinal center B-cell like (GCB) tumors in diffuse large B-cell lymphoma (DLBCL). 709 WSIs from two internal datasets were used. This data closely mirrors real-world data, since it is crucial to assess classification approaches in DP using such data and tasks. The WSIs (40× magnification) have been scanned by Ventana DP200 scanners. The artifact-free tissue tiles of this dataset were combined and randomly split into 70% training set, 15% validation set and 15% test set.

Camelyon16: Binary classification of cancer metastases vs. healthy in H&E images of lymph node tissue. The Camelyon16 dataset [15] consists of 400 WSIs of sentinel lymph nodes. The dataset is publicly available. For our evaluation, all artifact-free tissue tiles were used as well as the official train-test split. 20% of the training data was used as the validation set.

Herohe: Binary classification of breast cancer human epidermal growth factor receptor 2 (HER2) using the publicly available Herohe [7] dataset. Each H&E stained WSI is either labeled as HER2 positive or HER2 negative. We use the artifact-free tiles from tumor regions detected with an in-house tumor segmentation model. The 508 WSIs are split according to the official train-test split. 20% of the training data is used as the validation set.

3 Results

This section addresses our initial inquiries, first assessing if a universally optimal aggregation model configuration exists for multiple feature extraction (foundation) models, and then comparing different feature extractors considering performance variability across aggregation model setups.

3.1 Aggregation Model Configuration Influence

Our analysis begins by evaluating the sensitivity of feature extractors, or foundation models, to various aggregation model configurations.

The heatmap in Figure 3 displays the classification performance across all aggregation model configurations. Trends are consistent across both AUROC

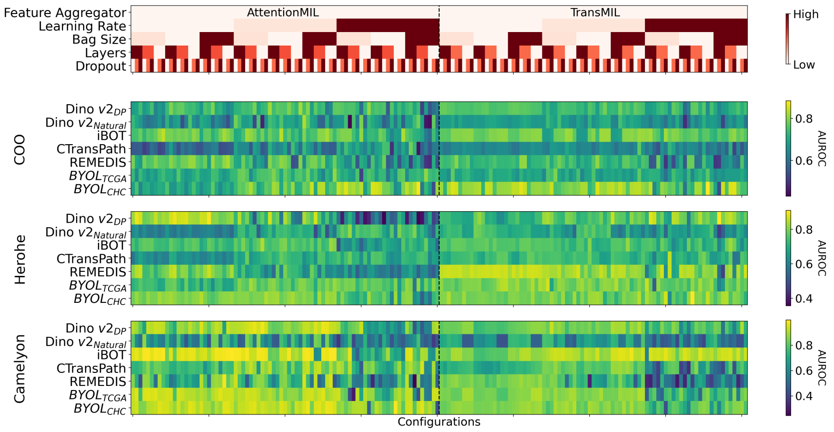

Fig. 3. The heatmap shows the performance of every aggregation model configuration set for each feature extraction model. The red colored legend shows how the configurations are ordered on the heatmap.

and AP scores. The heatmap legend aids in identifying patterns, such as configurations with the lowest learning rate positioned on the left of each feature aggregator, which tend to yield lower performance in the COO dataset when using CTransPath but not for other features extractors. Analysis of these heatmaps reveals:

Lack of a universal configuration: No single aggregation model configuration consistently outperforms across all feature extractors, as indicated by the absence of a uniformly bright column across models.

Dataset-specific configurations: Optimal configurations for a given feature extractor vary by dataset. While certain parameters like learning rate for AttentionMIL and the number of attention blocks for TransMIL show some dataset-specific importance, no definitive pattern emerges across datasets or models, suggesting the need for investigation of model-specific configurations.

These results highlight the need for evaluating a diverse range of configurations in the aggregation model. This approach would ascertain that any observed superiority of one feature extraction model over another is not simply attributed to the specific aggregation model setup selected.

3.2 Feature Extractor Comparison

Figure 4 diverges from the standard practice of showing a single outcome for a fixed aggregation model setup by presenting feature extractor model performance across all 162 configurations for various datasets. Through box plots, we observe substantial performance overlap among feature extraction models despite the variance across configurations. Key insights include:

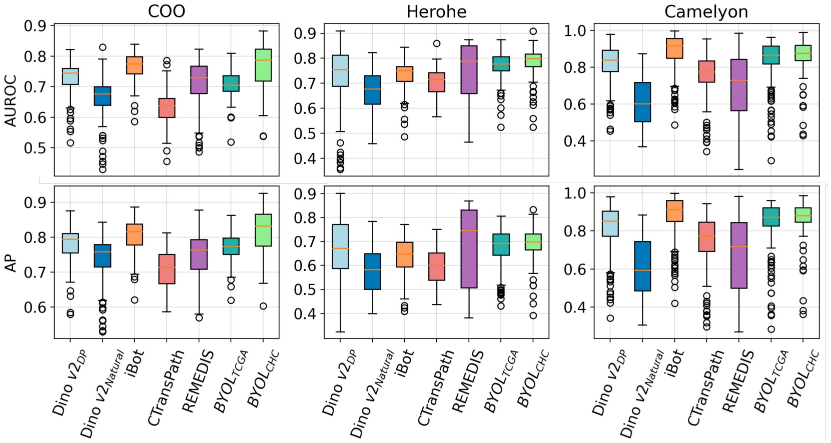

Fig. 4. Comparison of 7 feature extraction models across 162 different aggregation model configurations, which include 2 architectures with 81 parameters each.

Training on DP datasets is necessary: The DINOv2 model trained on natural images performs poorly compared to all other models, consistently for all the datasets.
Comparable Performance Across Model Sizes: The relatively small model BYOL$_{\mathrm{TCGA}}$ matches the performance of larger ones, suggesting that larger models are not necessarily better for DP. This echoes Filiot et al.'s [8] findings that a ViT-B model can outperform a ViT-L model.
Feature extractors generalize well: The BYOL$_{\mathrm{CHC}}$ model, trained on WSIs from evaluation datasets, shows good performance across all datasets as expected. Interestingly, its performance is not much higher than that of other models such as DINOv2$_{\mathrm{DP}}$, iBOT and BYOL$_{\mathrm{TCGA}}$. This observation confirms that the feature extraction models have the capability to generalize well.

4 Conclusion

In this study, we challenge the prevailing methodology for comparing foundation models in digital pathology literature, demonstrating that it may yield misleading results. We show that due to the high sensitivity of feature extraction models to downstream aggregation model configurations, relying solely on a single aggregation model configuration can disproportionately favor certain feature extractor models while disadvantaging others. Hence, we propose evaluating foundation models across different configurations for fairer comparisons. Our comprehensive analysis, taking into account performance variations across multiple configurations of the aggregation model, reveals a considerable overlap in performance between different foundation or feature extractor models. Significantly, we find no universal aggregation model configuration that is uniformly effective for all feature extractors. Our work is limited though by only looking at classification tasks. In addition, the DINOv2 model we trained on digital pathology images might be subpar to other models due to computational and dataset limitations. Nevertheless, we believe this work will contribute to a more nuanced evaluation of foundation models that will help gain insight and further accelerate this rapidly evolving field.

References

1. Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrović, J., Strachan, P., et al.: Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. Nature Biomedical Engineering pp. 1–24 (2023)
2. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
3. 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)
4. 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.: A general-purpose self-supervised model for computational pathology. arXiv preprint arXiv:2308.15474 (2023)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
6. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems **33**, 22243–22255 (2020)
7. Conde-Sousa, E., Vale, J., Feng, M., Xu, K., Wang, Y., Della Mea, V., La Barbera, D., Montahaei, E., Baghshah, M., Turzynski, A., Gildenblat, J., Klaiman, E., Hong, Y., Aresta, G., Araújo, T., Aguiar, P., Eloy, C., Polónia, A.: Herohe challenge: Predicting her2 status in breast cancer from hematoxylin&eosin whole-slide imaging. Journal of Imaging **8**(8) (2022). https://doi.org/10.3390/jimaging8080213, https://www.mdpi.com/2313-433X/8/8/213
8. Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Mac Kain, A., Saillard, C., Schiratti, J.B.: Scaling self-supervised learning for histopathology with masked image modeling. medRxiv pp. 2023–07 (2023)
9. Gadermayr, M., Tschuchnig, M.: Multiple instance learning for digital pathology: A review on the state-of-the-art, limitations & future potential. arXiv preprint arXiv:2206.04425 (2022)
10. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
11. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
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)

13. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
14. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3344–3354 (2023)
15. Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., Manson, Q.F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk, M., van der Laak, J.: 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience **7**(6), giy065 (05 2018). https://doi.org/10.1093/gigascience/giy065, https://doi.org/10.1093/gigascience/giy065
16. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. Advances in neural information processing systems **10** (1997)
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)
18. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
19. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: Review the cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia **2015**(1), 68–77 (2015)
20. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Mathieu, P., van Eck, A., Lee, D., Viret, J., et al.: Virchow: A million-slide digital pathology foundation model. arXiv preprint arXiv:2309.07778 (2023)
21. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)
22. Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V.: Nyströmformer: A nyström-based algorithm for approximating self-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14138–14148 (2021)
23. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)