BM1b: Data-Efficient Scaling of Bone Marrow Foundation Models through Domain-Specific Self-Supervised Learning and Dense Morphology-Preserving Representation Learning

Nishchal Sapkota[†], Reyhan K. Keser[†], Yanglan Ou, Steven S. Huang, Dong Chen, Eric D. Hsi, Jansen N. Seheult, Wenchao Han

Department of Laboratory Medicine and Pathology
Mayo Clinic, Rochester, MN, USA

Abstract. Bone marrow pathology requires integrating information across multiple spatial scales, from tissue architecture to fine-grained cellular morphology. While recent pathology foundation models have demonstrated promising performance across a wide range of pathology tasks, it remains unclear whether they capture the rich morphological representations required for hematopathology, where diagnosis relies on subtle cellular morphology and complex tissue architecture. In addition, those models are trained primarily with solid organ tissue samples, and the bone marrow specimens are underrepresented or excluded from the training dataset. Motivated by the advances in domain-adapted foundation models, we developed BM1b, a bone marrow biopsy-specific vision foundation model using 11 million multi-resolution image patches from 9,457 H&E bone marrow core biopsy whole slide images (WSIs). BM1b achieves strong performance across clinically relevant bone marrow pathology assessment tasks and outperforms the evaluated state-of-the-art pathology foundation models. During extended training (>200,000 iterations) of a vision transformer (ViT-G), we observe the tradeoff described in DINOv3: dense, morphology-sensitive representations degrade despite stable or improved performance on some of the key global prediction tasks such as disease classification. To mitigate this, we adapt a later-stage training regularization strategy to computational pathology and further extend that to interpolated Gram Anchoring (iGram), which leverages interpolated image representations to preserve local morphology better. Our Gram-regularized variants consistently improved cell detection and instance segmentation performance while preserving performance on region-based tasks. These results demonstrate that disease-specialized foundation models can outperform general-purpose pathology foundation models despite being pretrained on substantially smaller datasets. Moreover, Gram regularization enables larger ViTs to jointly learn fine-grained cellular morphology and global tissue architecture, mitigating the trade-off between dense and global hematopathology representations.

Keywords: Computational pathology · Bone marrow · Foundation models · Self-supervised learning

[†] These authors contributed equally to this work.

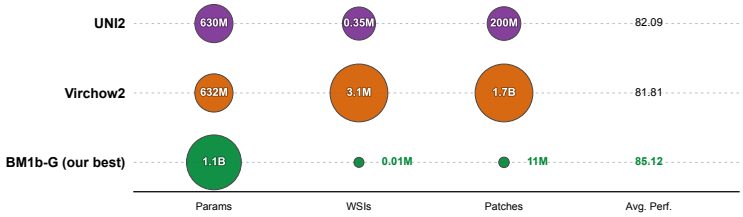

Fig. 1: Comparison of model size (number of parameters), pretraining dataset size (whole-slide images and image patches), and downstream slide-level performance between our best version of the BM1b model (BM1b-G) and state-of-the-art pathology foundation models.

1 Introduction

Bone marrow examination plays a central role in the evaluation of a remarkably diverse range of hematologic diseases, encompassing acute and chronic leukemias of myeloid and lymphoid lineages, plasma cell neoplasms and related disorders, lymphomas, myelodysplastic syndromes, myeloproliferative neoplasms, and numerous non-malignant hematologic disorders. A comprehensive evaluation integrates information from multiple specimen preparations and modalities, including peripheral blood and bone marrow aspirate smears, bone marrow core biopsy sections, and ancillary tests such as flow cytometry, cytogenetic, and molecular testing.

For more than seven decades, morphologic assessment of the bone marrow core biopsy has been a cornerstone of the diagnostic, prognostic, and increasingly, the theragnostic evaluation of hematologic diseases. Unlike aspirate smears, bone marrow core biopsies preserve tissue architecture and marrow microenvironment, enabling assessment of key histomorphologic features such as marrow cellularity, fibrosis, lineage composition, megakaryocyte morphology, patterns of lymphoid infiltration, and focal or diffuse involvement by plasma cell neoplasms and other hematologic malignancies [1, 4, 8]. Accurate bone marrow interpretation requires integration of morphologic information across multiple spatial scales. Low-magnification examination provides a global view of tissue architecture, facilitating assessment of overall cellularity, stromal and bone changes, collagen and reticulin fibrosis, and patterns of involvement in both malignant and benign hematologic diseases. In contrast, high-magnification examination enables detailed evaluation of cellular morphology, including leukemic blasts, plasma cells, dysplastic hematopoietic elements, and megakaryocytes. These complementary observations are integrated by expert pathologists into standardized diagnostic findings and interpretations in the pathology report, guiding disease classification, prognostication, and therapeutic decision-making.

The inherently multi-scale nature of bone marrow evaluation, along with the extreme morphologic diversity within and between hematologic disorders, presents a compelling opportunity for artificial intelligence in clinical practice.

Computational pathology models capable of jointly learning global tissue architecture and fine-grained cellular morphology may more closely emulate expert diagnostic bone marrow workup, while providing reproducible quantitative assessment of relevant morphologic features [4].

Recent advances in self-supervised learning have enabled the development of foundation models that learn generalizable visual representations from large-scale unlabeled natural image datasets [2, 10, 13]. These representations allow a single pretrained encoder to support a wide range of downstream applications, often outperforming narrow task-specific models trained from scratch. In computational pathology, these advances have been extended to large-scale self-supervised pretraining on pathology whole slide images (WSIs), producing general-purpose pathology foundation models that transfer effectively across diverse diagnostic and prognostic pathology tasks [14,15]. Models such as UNI/UNI2 [3] and Virchow/Virchow2 [14,18] have demonstrated the effectiveness of scaling pretraining with large ViTs [6] on hundreds of thousands to millions of WSIs and millions to billions of image patches.

Despite significant advances towards a general-purpose histopathology foundation model, key gaps remain. Existing pathology foundation models are primarily pretrained on broad surgical pathology and solid organ datasets, in which bone marrow specimens are underrepresented or absent. This is a critical limitation since the diagnostic signal in a trephine bone marrow core resides at the single-cell scale (blast enumeration, dysplasia, megakaryocyte morphology) and in tissue features underrepresented in general pathology training datasets, including bony trabeculae, the hematopoietic-to-adipocyte ratio that defines cellularity, and decalcification-induced alterations in chromatin and staining. Moreover, the demonstrated performance gains observed with pathology-specific foundation models over natural-image foundation models suggest that additional specialization within distinct and highly heterogeneous pathology domains may yield even more effective, clinically relevant, and biologically grounded visual representations.

To address these limitations, we propose domain-specialized self-supervised pretraining on digital WSIs of bone marrow biopsy specimens to develop a foundation model optimized for bone marrow pathology. We train BM1b, a 1.1-billion-parameter ViT-G bone marrow foundation model, with 11 million multi-resolution image patches (256 x 256 pixels) sampled at 5×, 10×, 20×, and 40× magnifications from a contemporaneous cohort of 9,457 bone marrow specimens using one digitized hematoxylin and eosin (H&E)-stained section per case. Although BM1b is a 1.7× larger backbone compared with publicly available pathology foundation models UNI2 or Virchow2 in parameter count (1.10B vs. 0.63B parameters), it was trained using approximately 35× fewer WSIs than UNI2 and 310× fewer than Virchow2 (Fig. 1). Despite the substantially smaller pretraining corpus, BM1b achieves the strongest performance across downstream bone marrow tasks, including region-level and cell-level prediction tasks. These findings demonstrate that carefully curated, disease-specific pathology data can

enable data-efficient training of large vision encoders optimized for specialized pathology subdomains.

Foundation models have scaled both training data and Vision Transformer architectures to improve downstream performance. However, recent work in computer vision has shown that prolonged self-supervised pretraining of large ViTs may improve semantic representations while degrading dense visual features [13]. During continued training of BM1b, we observe a similar phenomenon: dense morphology-sensitive representations progressively decline despite improving performance on some high-level semantic prediction tasks such as disease classification. These findings suggest a tradeoff between semantic abstraction and morphology preservation that is particularly relevant to hematopathology, where many diagnostically important tasks rely on subtle cellular morphology in addition to global tissue architecture. Addressing this tradeoff is important because pathology foundation models are typically released as a single pretrained checkpoint that is expected to support a wide range of downstream applications, from slide-level diagnosis to dense prediction tasks such as cell detection and segmentation.

To address this, we adapt Gram Anchoring (GA) [13] to computational pathology as a later-stage training regularization strategy. Unlike DINOv3, which applies Gram Anchoring to uniformly sampled image resolutions with their specific high-resolution counterparts, our pathology pretraining data naturally contain multi-resolution image patches, requiring a modified formulation tailored to whole-slide image pretraining. Building on this adaptation, we further introduce interpolated Gram Anchoring (iGram), an extension that we believe is particularly well suited to pathology, which computes Gram constraints from interpolated resampled image feature representations (512 x 512 pixels) to better preserve fine-grained local morphology (see Section 2.3). While GA consistently improves slide-level prediction performance by mitigating the loss of dense morphologic information during continued self-supervised training, iGram further enhances dense prediction tasks by strengthening local feature representations without sacrificing global semantic understanding. Our main contributions are:

1. We demonstrate an effective strategy for scaling pathology foundation models through domain-specific pretraining. We introduce BM1b, a 1.1B-parameter hematopathology foundation model that achieves competitive, and often superior, performance using over 35-310× fewer pretraining whole-slide images (WSIs) than state-of-the-art generalist pathology foundation models.
2. We identify a previously underexplored challenge in scaling large pathology Vision Transformers: prolonged self-supervised pretraining improves semantic abstraction for some of the clinically important tasks, while progressively degrading morphology-sensitive dense representations.
3. To address this challenge, we adapt Gram Anchoring to computational pathology and introduce interpolated Gram Anchoring (iGram), an extension that we believe is particularly well suited to pathology, improving the recovery of morphologic representations degraded during prolonged pretraining without compromising slide-level semantic performance.

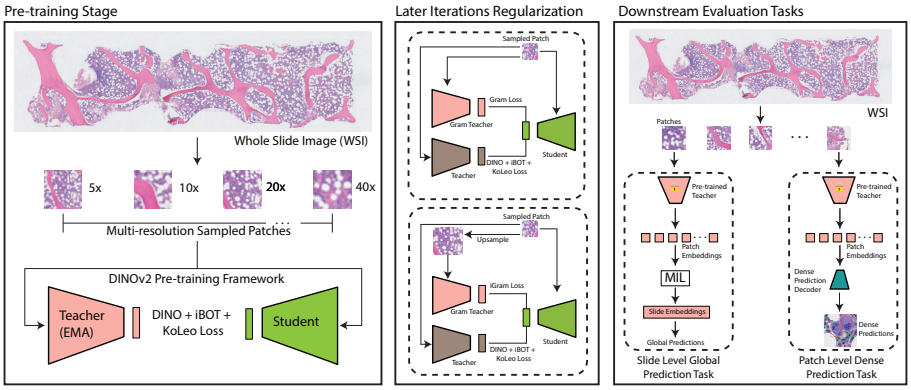

Fig. 2: Main Diagram: Overview of the BM1b training and evaluation pipeline. **Pre-training Stage**: A ViT-G encoder is pretrained using DINOv2 on multi-resolution bone marrow WSI patches. **Later Iterations Regularization**: The model is further optimized using Gram Anchoring (GA) to recover degraded dense representations. We further introduce interpolated Gram Anchoring (iGram), which uses interpolated image representations to better preserve fine-grained cellular morphology. **Downstream Evaluation Tasks**: The resulting models are evaluated on slide-level diagnostic tasks using multiple instance learning (MIL) and dense prediction tasks using trainable decoders.

4. On two bone marrow diagnostic tasks that are largely absent from solid-tissue pathology, disease classification and megakaryocyte morphology classification, our best BM1b model surpasses Virchow2 by 5.08% and UNI2 by 3.88%. Overall, across all case-level tasks, on average, our BM1b improves performance by approximately 3.31% over Virchow2 and 3.03% over UNI2.

2 Method

In this section, we discuss our pre-training dataset, model architecture, pre-training framework, and a regularization strategy for continued self-supervised pretraining.

2.1 Pre-training Dataset

We use Whole Slide Images (WSIs) of hematoxylin and eosin (H&E)-stained slides scanned on two types of scanners (Leica GT 450 and Pramana SpectralHT). To capture the complementary visual cues at different fields of view and magnification of the tissue, we construct a large-scale multi-resolution pathology dataset by sampling image patches at 5×, 10×, 20×, and 40× magnifications. A total of 9,457 bone marrow whole-slide images (one WSI per case) spanning a three-year period (2022 - 2025) were selected. Tissue regions are first identified using a pre-trained U-Net [12] segmentation model. Image patches are sampled

using a content-aware strategy that preferentially selects patches with heterogeneous visual patterns while down-sampling homogeneous regions. The resulting pretraining corpus contains approximately 11 million 256×256 image patches spanning a broad spectrum of bone marrow morphology.

2.2 Model and Self-supervised Training

Recent advances in self-supervised learning have shown that foundation model performance can improve through the coordinated scaling of model capacity, training strategy, and pretraining data [2, 10, 13]. In particular, DINOv3 [13] demonstrated that large-scale self-supervised pretraining can produce increasingly capable visual foundation models, motivating the use of large Vision Transformer (ViT) architectures together with carefully designed training strategies. However, effectively scaling large Vision Transformers remains challenging, as larger models require increasingly effective optimization and high-quality pretraining data to fully realize their capacity. Recent pathology foundation models further illustrate this point: despite using significantly fewer training slides, UNI2 achieves competitive performance to Virchow2 [18] on many downstream tasks, suggesting that scaling performance depends not only on dataset size but also on effective pretraining strategies and domain specificity.

Motivated by these observations, we developed BM1b, a large-scale ViT trained effectively using bone marrow biopsies, with a content-aware sampling approach. Our BM1b model adopts the ViT-G/14 architecture introduced in [16] and used in the self-supervised learning setting by DINOv2 [10] framework, consisting of approximately 1.1B parameters, 40 transformer blocks, a hidden dimension of 1536, and SwiGLU feed-forward networks. We additionally incorporate eight register tokens [5], which have been shown to improve representation quality and training stability in large self-supervised ViTs.

BM1b is pretrained using the DINOv2 framework [10], which combines the DINO self-distillation objective for learning invariant global representations, the iBOT objective [17] for masked patch-level representation learning, and the KoLeo regularization objective [10] for promoting a uniform utilization of the embedding space. The pretraining loss is defined as

$$\mathcal{L}_{\text{pre}} = w_D \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{iBOT}} + w_{DK} \mathcal{L}_{\text{DKoleo}}, \tag{1}$$

where $\mathcal{L}_{\text{DINO}}$, $\mathcal{L}_{\text{iBOT}}$, and $\mathcal{L}_{\text{DKoleo}}$ denote the DINO, iBOT, and KoLeo objectives, respectively, and w_D and w_{DK} control the relative contributions of the DINO and KoLeo terms.

2.3 Training Regularization with Gram and iGram Anchoring

Recent work has shown that prolonged self-supervised training of very large ViTs can improve semantic representations while degrading dense visual representations [13]. We observe a similar phenomenon in BM1b (Fig. 3), where dense

prediction performance progressively declines despite stable or improving performance on some slide-level prediction tasks, such as disease classification. This trade-off is particularly problematic in bone marrow biopsy evaluation, where downstream applications require both global contextual understanding and fine-grained cellular morphology.

To preserve dense morphology-sensitive representations, we adopt Gram Anchoring (GA) [13], which regularizes token-level spatial relationships using a frozen anchor model. Given student token features $X_S \in \mathbb{R}^{N\times D}$ and anchor features $X_G \in \mathbb{R}^{N\times D}$, where N is the number of tokens and D is the embedding dimension, the Gram Anchoring loss is defined as

$$\mathcal{L}_{\text{Gram}} = \left\| X_S X_S^\top - X_G X_G^\top \right\|_F^2 . \tag{2}$$

The corresponding training objective is

$$\mathcal{L}_{\text{Reg}}^{\text{Gram}} = \mathcal{L}_{\text{pre}} + w_{\text{Gram}} \mathcal{L}_{\text{Gram}}, \tag{3}$$

where $\mathcal{L}_{\text{pre}}$ denotes the default self-supervised pretraining objective and w_{Gram} controls the strength of Gram regularization. By matching the pairwise relationships encoded by the student and anchor Gram matrices, GA encourages the preservation of local spatial structure during continued training.

While DINOv3 employs additional high-resolution views during training, our pretraining corpus already contains native multi-resolution pathology images spanning 5× to 40× magnifications. We therefore adapt Gram-based regularization to computational pathology, where native multi-resolution images are already available.

Building on this formulation, we propose *interpolated Gram Anchoring* (iGram), an extension of GA designed to further preserve fine-grained morphological representations. Rather than extracting anchor features directly from the original image patch, iGram first upsamples the input image and extracts anchor features from the interpolated view. Interpolation changes the tokenization of the same underlying image, providing a denser representation of its morphology without requiring additional high-resolution image acquisitions or training data. By exposing complementary local token relationships and aligning them through Gram regularization, iGram provides additional supervisory signal for preserving morphology-sensitive representations during prolonged pretraining. This is particularly relevant to bone marrow pathology, where diagnostically important cellular structures may occupy only a small number of pixels at the original input resolution. Given interpolated anchor features $X_{iG} \in \mathbb{R}^{N\times D}$, the iGram loss is defined as

$$\mathcal{L}_{\text{iGram}} = \left\| X_S X_S^\top - X_{iG} X_{iG}^\top \right\|_F^2 . \tag{4}$$

The corresponding iGram training objective is

$$\mathcal{L}_{\text{Reg}}^{\text{iGram}} = \mathcal{L}_{\text{pre}} + w_{\text{iGram}} \mathcal{L}_{\text{iGram}}, \tag{5}$$

where w_{iGram} controls the strength of iGram regularization.

Table 1: Downstream evaluation tasks and sample sizes for global case-level prediction and local dense feature-related prediction tasks.

Task	Details	# Samples
	Global Case-Level Prediction Tasks	# Cases
Disease Classification (Diagnosis)	7-category diagnosis: lymphoma, acute myeloid leukemia, plasma cell neoplasm, myeloproliferative neoplasm, myelodysplastic syndrome, MDS/MPN overlap, or negative.	220
Megakar. Morphology	Classifies megakaryocytes as normal or dysplastic.	562
Fibrosis Grading	Grades marrow fibrosis as mild, moderate, or severe.	268
Cellularity Assessment	Classifies marrow cellularity as hypo-, normo-, or hypercellular.	587
M:E Ratio Estimation	Classifies myeloid-to-erythroid ratio as decreased, normal, or increased.	493
	Local Dense Prediction Tasks	# Patches
Multiclass Cell Detection (MCD)	Detects individual cells, Megakaryocytes, and Megakaryocyte nuclei (MK Nuclei) from WSI patches at higher magnification	1447
Binary Cell Instance Segmentation (BCIS)	Segments individual cell instances to evaluate the preservation of cellular boundaries and morphology-rich dense features.	1447

3 Experiments and Results

In this section, we first describe the training setup, baselines, and evaluation protocol. We then analyze the effects of continued scaling on representation quality, followed by evaluations on global case-level prediction tasks and a comprehensive ablation study of Gram-based regularization for local dense prediction tasks.

3.1 Training Details, Baselines and Evaluation Protocol

The BM1b pretraining stage was performed on 32 NVIDIA H200 GPUs using a total batch size of 2,688 for 450,000 iterations. Gram Anchoring and iGram were conducted for an additional 50,000 iterations, with most downstream tasks reaching their best performance within the first 15,000 regularization iterations. Unless otherwise specified, all training hyperparameters follow the official DINOv2 implementation [10].

We benchmark BM1b against UNI2 and Virchow2, two widely used open-weight pathology foundation models pretrained on substantially larger pathology datasets (Fig. 1). Although the authors of Virchow2 [18] also trained a larger ViT (i.e., ViT-G, called Virchow2G), its pretrained model weights are not publicly available for our benchmark experiments. For all models, patch embeddings are extracted from their respective frozen pretrained encoders and evaluated using the same downstream training framework, data splits, and evaluation protocol.

The main slide-level benchmark uses ten-fold cross-validation, while the ablation experiments use three-fold cross-validation. Cross-validation splits were performed at the case level. To prevent data leakage, all cases used for downstream evaluation were excluded from the pretraining corpus. Scanner type and acquisition period were not explicitly used for stratification and therefore followed their natural distributions across folds.

For slide-level prediction tasks, image patch embeddings are aggregated using a variant of the multiple instance learning (MIL) framework [9]. The same

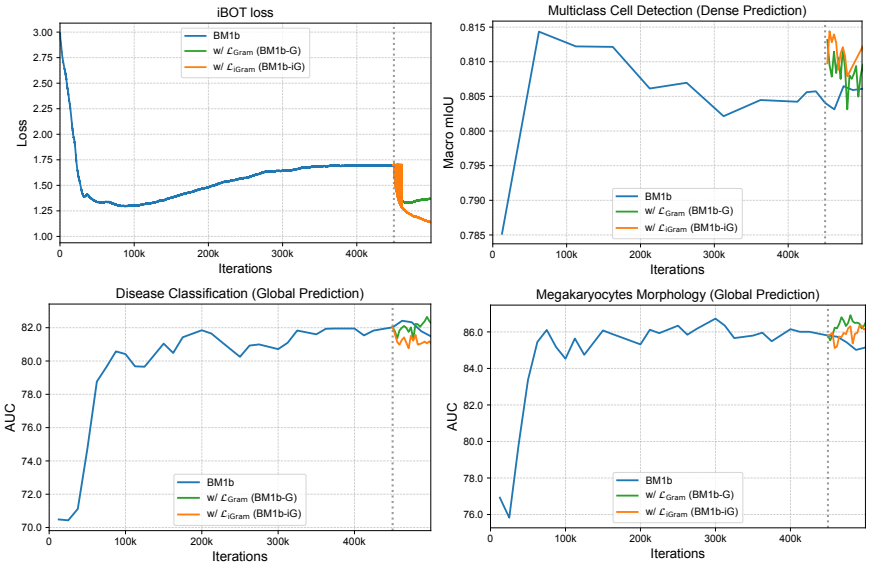

Fig. 3: Effects of continued self-supervised pretraining and Gram-based regularization on training dynamics, slide-level prediction tasks, and dense prediction. Panels show the iBOT training loss, performance on two representative slide-level tasks (disease classification and megakaryocyte morphology classification), and a representative dense prediction task (cell detection).

MIL architecture and training configuration are used across all models to ensure a consistent comparison, and performance is reported as the average AUC across cross-validation folds. The benchmark comprises clinically relevant case-level tasks of bone marrow biopsy assessment, including disease classification, fibrosis grading, cellularity assessment, megakaryocyte morphology classification, and M:E ratio estimation. Details for each task and sample size are provided in Table 1.

To evaluate dense visual representations, we attach task-specific prediction heads to the frozen foundation model encoders. A Faster R-CNN [11] head is used for multiclass cell detection, while a Mask R-CNN [7] head is used for binary cell instance segmentation. Because these tasks require precise localization and delineation of individual cellular structures, they provide a direct assessment of morphology preservation. For multiclass cell detection, we report F1 and mean intersection over union (mIoU), while for binary cell instance segmentation, we report mIoU. For our models, evaluation checkpoints are selected based on the best validation performance for each downstream task.

3.2 Effects on the Training Losses

Fig. 3 shows the evolution of training losses during continued optimization. Consistent with observations from DINOv3, prolonged training leads to degradation

Table 2: Main Results: Bone marrow case-level performance in AUC, reported as mean ± standard deviation across 10 cross-validation folds.

Model	Diagnosis	Megakar.	Fibrosis	Cellularity	M:E Ratio	Average
UNI2	77.44 ± 3.63	85.62 ± 2.34	82.80 ± 4.13	91.61 ± 1.04	72.99 ± 4.40	82.09 ± 1.50
Virchow2	75.93 ± 4.18	84.73 ± 2.48	83.84 ± 3.46	**91.85** ± 1.32	72.73 ± 4.23	81.81 ± 1.49
BM1b	82.97 ± 4.59	86.40 ± 2.62	85.21 ± 3.64	91.60 ± 0.99	76.80 ± 3.73	84.59 ± 1.34
w/ $\mathcal{L}_{Gram}$	**83.28** ± 4.47	**87.54** ± 2.75	**85.89** ± 3.98	**91.64** ± 1.12	**77.27** ± 3.73	**85.12** ± 1.35
w/ $\mathcal{L}_{iGram}$	82.44 ± 4.61	86.82 ± 2.66	85.72 ± 3.87	**91.64** ± 1.30	76.80 ± 3.84	84.69 ± 1.38

of the iBOT objective despite continued improvement of semantic representations. GA stabilizes training and successfully recovers the iBOT loss, which is a proxy for the model's ability to capture dense features. The Gram loss decreases steadily throughout training, indicating successful alignment between student and anchor representations. Similar trends are observed for iGram, suggesting that Gram-based regularization effectively counteracts dense feature degradation during prolonged self-supervised training.

3.3 Global Features and Performance

Table 2 summarizes the performance of the baselines, our BM1b model, and the proposed Gram-based regularization strategies on the five slide-level prediction tasks for bone marrow biopsy specimens. Despite being pretrained on only 10K WSIs, BM1b establishes a strong baseline with an average AUC of 84.59, outperforming both UNI2 (82.09) and Virchow2 (81.81), which were pretrained on substantially larger and more diverse pathology datasets. This result demonstrates that carefully curated disease-specific pretraining can be an effective approach for scaling large ViT backbones. Disease classification exhibits the largest improvement over prior foundation models, whereas M:E ratio shows more modest gains, likely because it is conventionally assessed from aspirate smears rather than core biopsy images. Nevertheless, our models still outperform both UNI2 and Virchow2 on this biopsy-based prediction task.

Training regularization further improves global performance. Gram Anchoring ($\mathcal{L}_{Gram}$) achieves the highest average AUC of 85.12, outperforming UNI2 and Virchow2 by 3.03 and 3.31 AUC points, respectively. Although iGram ($\mathcal{L}_{iGram}$) is designed to improve dense morphology preservation, it maintains a comparable average AUC of 84.69. Both Gram-based variants outperform the base BM1b model as well as UNI2 and Virchow2. Together, these results demonstrate that Gram-based regularization improves dense representations without sacrificing slide-level predictive performance.

3.4 Ablation Study: Dense Local and Global Hematopathology Features and Gram Regularization

As seen in Fig. 3, dense prediction performance initially improves during pretraining before gradually deteriorating with continued optimization, despite sta-

Table 3: Ablation study of BM1b with Gram Anchoring and iGram Anchoring on global and dense prediction tasks. Dense prediction performance is reported using F1 and mIoU for multiclass cell detection (MCD) and mIoU for binary cell instance segmentation (BCIS), with the overall dense score computed as the average of MCD macro mIoU and BCIS mIoU. Global prediction performance is reported as the average AUC across five hematopathology tasks.

	MCD F1			MCD mIoU			BCIS	Overall Avg.	
Model	**Macro**	**Cell**	**MK Nuclei**	**Macro**	**Cell**	**MK Nuclei**	**mIoU**	**Dense**	**Global**
BM1b	81.03	69.06	83.75	80.76	76.22	81.81	75.40	78.08	84.59
w/ $\mathcal{L}_{Gram}$	**81.61**	69.73	**85.44**	81.23	76.66	82.29	75.85	78.54	**85.12**
w/ $\mathcal{L}_{iGram}$	81.56	**70.18**	84.29	**81.44**	**76.74**	**82.51**	**76.35**	**78.90**	84.69

ble or improving global prediction performance on some tasks like disease classification. This behavior suggests that prolonged training shifts representations toward greater semantic abstraction at the expense of morphology-rich dense features.

Table 3 shows the ablation study results for multiclass cell detection and binary cell instance segmentation with and without GA and iGram. Both GA and iGram effectively recover dense representations while preserving strong slide-level prediction task performance. GA achieves the highest average AUC (85.12) across all five slide-level bone marrow biopsy tasks, whereas iGram achieves the strongest dense prediction performance, with an average mIoU of 78.90 across MCD and BCIS. Specifically, iGram achieves the highest MCD macro mIoU (81.44), Cell mIoU (76.74), MK Nuclei mIoU (82.51), and BCIS mIoU (76.35), indicating superior preservation of fine-grained morphology.

The differences in performance of our models become apparent in morphologically challenging cases (Fig. 4). Qualitative results further support these findings. In the first row, BM1b-G fragments the megakaryocyte nucleus into two detections, whereas BM1b-iG correctly preserves it as a single object (see the region indicated by the yellow arrows in the ground-truth image and the corresponding region in the model predictions). In the second row, our base model BM1b incorrectly splits a cell with a multi-lobed nucleus into two separate detections, while both Gram-based variants correctly recognize it as a single cell. Overall, iGram consistently recovers additional small-cell detections in densely populated regions while preserving the morphology of larger cellular structures, with fewer false positives.

4 Conclusion

We present BM1b, a 1.1B-parameter foundation model pretrained on bone marrow biopsy whole-slide images and demonstrate that domain-specific pretraining provides a highly data-efficient strategy for scaling pathology foundation models. Despite being pretrained on only 11 million multi-resolution image patches derived from approximately 10,000 whole-slide images, BM1b achieves competitive

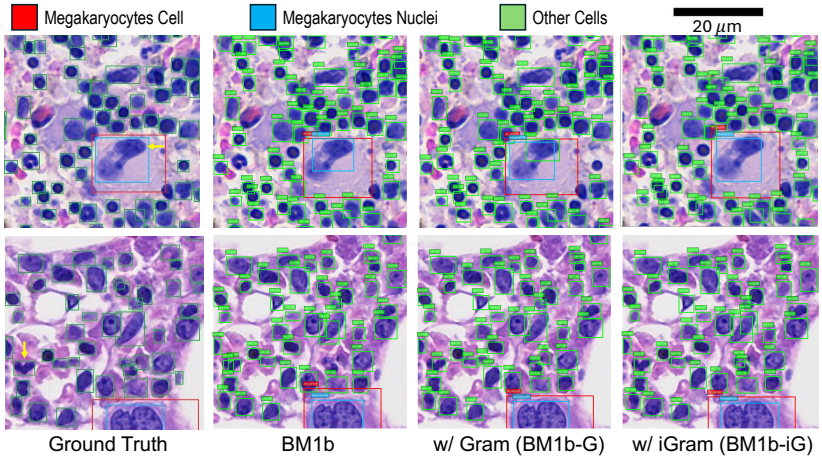

Fig. 4: Object detection results with and without Gram-based regularization for BM1b.

and often superior performance compared with state-of-the-art generalist pathology foundation models trained on substantially larger and more diverse pathology datasets. These findings suggest that domain specificity, multi-resolution pretraining, and morphology-aware regularization can be as important as dataset scale for developing effective pathology foundation models.

Beyond introducing BM1b, our study identifies a previously underexplored challenge in scaling large pathology Vision Transformers. Consistent with recent observations in DINOv3, we observe that, for the clinically important bone marrow task of disease classification, continued self-supervised pretraining progressively improves slide-level semantic representations, despite simultaneously degrading morphology-sensitive dense representations, thereby hampering dense prediction tasks such as megakaryocyte detection and segmentation. To address this challenge, we adapt Gram Anchoring to computational pathology and introduce interpolated Gram Anchoring (iGram), an extension that we believe is particularly well suited to pathology, using interpolated image feature representations to recover fine-grained cellular morphology information lost during continued pretraining. Importantly, iGram achieves this by extracting additional supervisory signal from the existing pretraining data, requiring neither additional annotations nor higher-resolution image acquisitions. Together, our results show that, on the evaluated tasks, morphology-sensitive representations can be preserved without sacrificing slide-level predictive performance, enabling a single foundation model checkpoint to support both slide-level diagnostic tasks and dense morphology-driven applications.

Our work suggests that the next generation of pathology foundation models should focus not only on scaling model size and pretraining data, but also on improving data efficiency and preserving clinically relevant visual representations. We anticipate that these principles will facilitate the development of increas-

ingly specialized foundation models for individual pathology domains, support multimodal foundation models that integrate morphology with ancillary clinical and molecular data, and motivate standardized benchmarks that evaluate both semantic understanding and fine-grained morphologic representations. Additionally, while iGram improves upon Gram Anchoring through interpolation-based anchor generation on dense prediction tasks, its performance may be further enhanced with more powerful interpolation algorithms. Future work should also investigate a two-stage regularization strategy that first applies Gram Anchoring and subsequently iGram to combine the strengths of both.

Limitations: Although our dataset includes bone marrow specimens collected from multiple sources affiliated with Mayo Clinic and WSIs acquired using two scanner types, external validation using data from independent institutions, patient populations, scanners, and staining protocols is needed to establish broader generalizability. Moreover, the clinical nature of the data restricts the public release of the pretraining dataset, presenting challenges for full reproducibility. The substantial computational cost of training a 1.1B-parameter ViT-G limits extensive hyperparameter and data augmentation studies. Consequently, the behavior of BM1b and its Gram-based variants across different training configurations has yet to be systematically investigated. Finally, our evaluation focuses on a selected set of slide-level and dense prediction tasks, and broader hematopathology benchmarks are needed to assess generalization across additional clinically relevant applications.

References

1. American Society of Hematology: Demystifying the bone marrow biopsy: A hematopathology primer. `https://www.hematology.org/education/trainees/fellows/hematopoiesis/2021/demystifying-the-bone-marrow-biopsy-a-hematopathology-primer` (2021), accessed: 2026-06-10
2. 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)
3. 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)
4. College of American Pathologists: Bone marrow pathology teaching presentation. `https://documents.cap.org/documents/bone-marrow-teaching-presentation.pdf` (nd), accessed: 2026-06-10
5. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: International conference on learning representations. vol. 2024, pp. 2632–2652 (2024)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)

8. Jain, S., DeZern, A.E.: Laboratory evaluation of bone marrow. In: StatPearls. StatPearls Publishing, Treasure Island, FL (2024), https://www.ncbi.nlm.nih.gov/books/NBK603716/
9. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
10. 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. Transactions on Machine Learning Research (2024)
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
13. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
14. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., et al.: A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine **30**(10), 2924–2935 (2024)
15. 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 **630**(8015), 181–188 (2024)
16. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12104–12113 (2022)
17. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. International Conference on Learning Representations (2021)
18. Zimmermann, E., Vorontsov, E., Viret, J., Casson, A., Zelechowski, M., Shaikovski, G., Tenenholtz, N., Hall, J., Klimstra, D., Yousfi, R., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology. arXiv preprint arXiv:2408.00738 (2024)