Biologically-Constrained Multi-Label Classification with Learnable Domain Knowledge

Nabil Mouadden[1,2] NABIL.MOUADDEN@CENTRALESUPELEC.FR
Veronique Verge[3] VERONIQUE.VERGE@GUSTAVEROUSSY.FR
Ahmadreza Arbab[3] AHMADREZA.ARBAB@GUSTAVEROUSSY.FR
Jean-Baptiste Micol[3] JEANBAPTISTE.MICOL@GUSTAVEROUSSY.FR
Elsa Bernard[3] ELSA.BERNARD@GUSTAVEROUSSY.FR
Aline Renneville[3] ALINE.RENNEVILLE@GUSTAVEROUSSY.FR
Stergios Christodoulidis[1,2] STERGIOS.CHRISTODOULIDIS@CENTRALESUPELEC.FR
Maria Vakalopoulou[1,2] MARIA.VAKALOPOULOU@CENTRALESUPELEC.FR

[1] *MICS, CentraleSupelec, Paris-Saclay University, France*
[2] *IHU PRISM, National Center for Precision Medicine in Oncology, Gustave Roussy*
[3] *Gustave Roussy, Villejuif, France*

Editors: Accepted for publication at MIDL 2025

Abstract

Although recent foundation models trained in a self-supervised setting have shown promise in cellular image analysis, they often produce biologically impossible predictions when handling multiple concurrent abnormalities. This is a problem, as the biological information that may be needed for the different clinical-oriented problems is not directly presented in the images. In this study, we present a novel and modular approach to enforce biological constraints in multi-label medical imaging classification. Building on the powerful and rich representations of the DinoBloom hematological foundation model, our method combines learnable constraint matrices with adaptive thresholding, effectively preventing contradictory predictions while maintaining high sensitivity. Extensive experiments on three datasets, two public and one in-house on neutrophil classification, demonstrate significant improvements over different foundation models and the state-of-the-art methods. Through detailed ablation studies and hyperparameter interpretation, we show that our approach successfully captures biological relationships between different abnormalities. Our code is accessible at https://github.com/nabilmouadden/biologically-constrained-classification/.

1. Introduction

Foundation models have revolutionized medical image analysis, with DinoBloom emerging as a powerful model specifically designed for hematological image analysis (Koch et al., 2024). However, while these models excel at general feature extraction from blood cell images, integrating explicit biological constraints and handling multiple concurrent abnormalities remains an open challenge. This limitation is particularly evident in clinical settings where predictions must adhere to known biological impossibilities and relationships. While self-supervised learning (SSL) is a very powerful technique that can capture correlations in large datasets, it often fails to learn critical biological constraints that are usually associated with class-related information. This occurs since, in highly imbalanced medical datasets, rare biological contradictions may be underrepresented or entirely absent in training. Additionally, morphological markers alone may not provide sufficient context to learn mutual exclusivities or complex dependencies between abnormalities.

Neutrophils are a type of white blood cells, and they form the most abundant type of granulocytes of all white blood cells in humans. Neutrophil morphology analysis plays a crucial role in hematological diagnosis, serving as a fundamental tool for identifying various blood disorders and infections (Hoffbrand and Moss, 2016). Traditional manual classification of neutrophil abnormalities is time-consuming and subject to significant inter-observer variability, with reported concordance rates as low as 60% among experts (Briggs et al., 2009). While recently deep learning approaches have shown remarkable promise in medical image analysis (Litjens et al., 2017), existing methods fail to address two critical challenges unique to neutrophil classification: (1) the complex interdependencies between multiple concurrent abnormalities, and (2) the need to enforce explicit biological constraints in predictions (Matek et al., 2019).

In this study, we propose a novel, differentiable framework that enhances foundation models with learnable biological constraints. The main contributions of this study are two folds: *(i)* we introduce a learnable constraint satisfaction module that automatically discovers and enforces biological relationships while maintaining end-to-end differentiability, *(ii)* we propose an adaptive thresholding mechanism that dynamically adjusts to varying degrees of abnormality manifestation. Our extensive experiments and ablations using the DinoBloom foundation model for neutrophil morphology analysis, which included three different datasets with varying numbers of abnormalities, highlight the superiority of our method with respect to the state of the art.

2. Related Work

Foundation Models in Hematology. Hematological image analysis has recently embraced large-scale self-supervision and transformers. Early efforts, such as (Matek et al., 2019) employed supervised convolutional networks on smaller datasets. More recent approaches leverage large data corpora and attention mechanisms: DinoBloom (Koch et al., 2024) emerged as a specialized foundation model for white blood cell (WBC) morphology, while (Wang et al., 2022) proposed unsupervised contrastive transformers for histopathological images. The introduction of domain-specific transformer architectures by (Filiot et al., 2023) further advanced the field through masked image modeling. Although these approaches learn strong representations, they do not explicitly incorporate domain-specific constraints or handle biologically impossible co-occurrences.

Multi-Label Classification in Medical Imaging. Many medical tasks inherently involve multi-label outputs, as conditions often co-exist or overlap. Traditional solutions like binary relevance overlook label correlations, which has prompted research into more holistic methods. Beyond medical imaging, there is substantial work on modeling label co-occurrence in general computer vision. (Pham et al., 2022) proposed graph-based multi-label disease prediction that leverages both data and domain knowledge. Bruton et al. (Bruton et al., 2022) developed label dependency methods for biomedical applications using graph structures. Unlike these approaches which often treat label relationships as fixed, our method learns adaptive constraints that can vary based on the input image. Wang et al. (Wang et al., 2022) proposed unsupervised contrastive transformers for histopathological images, while (Wang et al., 2023a) proposed clustering-guided contrastive learning for cell images (Chen et al., 2019) explored attention mechanisms, yet such methods often assume static or binary inter-label relationships, and do not address clinical uncertainty where co-occurrence can be probabilistic.

Domain-Constrained Learning. Incorporating domain knowledge into deep models has gained traction in medical imaging (Tolkach et al., 2019; Kortum and Armato, 2023), showing that enforcing biologically meaningful constraints can improve both performance and interpretability. However, many techniques rely on rigid rules or postprocessing steps.(Matek et al., 2019) have shown the importance of incorporating morphological constraints in leukemia cell classification. Beyond

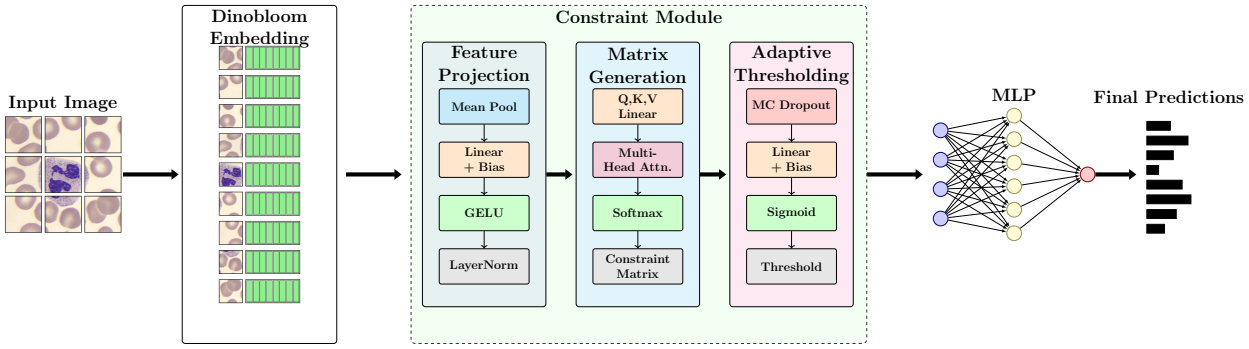

Figure 1: **Biologically-Constrained Multi-Label Classification Architecture.** The proposed model consists of three main components within the constraint module: (A) Feature projection with pooling and normalization, (B) Constraint matrix generation using attention mechanism, (C) Adaptive thresholding with Monte Carlo sampling.

hematology, these studies have demonstrated the value of domain constraints across various medical imaging applications, though they often lack the flexibility to capture probabilistic relationships that arise in real-world clinical settings. In contrast, our approach unifies a powerful transformer-based hematology backbone with learnable constraint matrices and adaptive thresholding, thus capturing both deterministic incompatibilities and nuanced, uncertain relationships that arise in real-world clinical settings.

3. Methods

3.1. Problem Formulation

Given an input image $x \in \mathbb{R}^{H\times W\times 3}$, where H and W represent the height and width of the image respectively, we aim to predict a set of binary labels $\mathbf{y} = [y_1, ..., y_K] \in \{0,1\}^K$, where K is the number of possible classes. The prediction must satisfy two types of biological constraints: *(i) mutual exclusivity constraints:* defined as $\mathcal{C}_{mu} = \{(i,j)|y_i \cdot y_j = 0\}$, where (i,j) represents pairs of classes that cannot co-exist, and *(ii) co-occurrence constraints:* defined as $\mathcal{C}_{co} = \{(i,j,c)|y_i = 1 \implies P(y_j) \geq c\}$, where $c \in [0,1]$ is a threshold probability and models the increased or decreased likelihood of the presence of one class with respect to the rest.

Traditional approaches perform classification in isolation, without considering the rich domain knowledge that biologists leverage in their decision-making process (Pham et al., 2022; Esteva et al., 2019). However, in practice, experts rely on their deep understanding of biological relationships and manifestations - they know that certain abnormalities often co-occur, that some conditions are mutually exclusive, and that the same abnormality can present with varying degrees of severity (Briggs et al., 2009; Hoffbrand and Moss, 2016). Moreover, biologists can discover new relationships between conditions through observations and adjust their confidence based on the strength of different markers (Bruton et al., 2022; Tolkach et al., 2019). Our method aims to emulate this expert reasoning by incorporating learnable biological constraints and adaptive decision thresholds. Our method addresses these limitations by integrating the $\mathcal{C} = \mathcal{C}_{mu} + \mathcal{C}_{co}$ into the training process through learnable constraint and uncertainty-aware adaptive thresholding. An overview of the entire method is presented in Figure 1.

3.2. Learnable Constraint Module

We introduce a learnable constraint module that enhances the model's ability to capture and enforce biological relationships. This module consists of three key components designed to work together:

Projection Layer. The feature projection layer transforms high-dimensional features into a space more suitable for learning biological constraints. We start with the features $\mathbf{h}_L \in \mathbb{R}^{N\times d}$, where N is the number of image patches (196 for 224×224 images with 16×16 patches) and d is the feature dimension. These features are processed through a projection layer to obtain $\mathbf{f} \in \mathbb{R}^d$, which represents a condensed feature vector: $\mathbf{f} = \text{LayerNorm}(\text{GELU}(\mathbf{W}_p\text{Pool}(\mathbf{h}_L) + \mathbf{b}_p))$. where $\text{Pool}(\cdot)$ performs mean pooling over the N patches to create a single d-dimensional feature vector, $\mathbf{W}_p \in \mathbb{R}^{d\times d}$ is a learnable weight matrix that projects the pooled features while preserving dimensionality and $\mathbf{b}_p \in \mathbb{R}^d$ is a learnable bias vector, GELU (Gaussian Error Linear Unit) is a smooth activation function that helps maintain gradient flow and LayerNorm normalizes the features across the feature dimension, stabilizing training by ensuring consistent feature scales. In practice, $\mathbf{f}$ is a d-dimensional feature vector that aggregates the patch-level information from DinoBloom into a single vector representing the entire cell, and it projects the features into a space where biological relationships can be more easily learned through the constraint mechanism.

Constraint Matrix Generation. For the constraint matrix, we employ a transformer-based multi-head attention mechanism to learn the relationships between different abnormalities. This mechanism allows the model to learn both positive (co-occurrence) and negative (mutual exclusivity) relationships dynamically: $\mathbf{Q} = \mathbf{W}_q\mathbf{f}, \quad \mathbf{K} = \mathbf{W}_k\mathbf{f}, \quad \mathbf{V} = \mathbf{W}_v\mathbf{f}$, where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{K\times d}$ are learnable parameter matrices that transform the features into query, key, and value representations and K is the number of possible abnormalities. The constraint matrix $\mathbf{R} \in \mathbb{R}^{K\times K}$ captures pairwise relationships between abnormalities and it is then computed using scaled dot-product, $\mathbf{R} = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d}}\right)\mathbf{V}$. To obtain the final constraint matrix, we implement a linear projection layer that transforms the attention output from $\mathbb{R}^{K\times d}$ to $\mathbb{R}^{K\times K}$ by mapping each class representation to its relationship with all other classes. In practice, after computing the softmax-based attention output for each pair (i, j), we multiply by the learnable value embeddings $\mathbf{V}$ and then apply a re-scaling-and-clamp step so that $\mathbf{R}_{ij} \in [-1, 1]$. Specifically, we first subtract the mean of all resulting entries (centering them), then multiply by a scalar α such that the largest absolute deviation from the mean is mapped to ± 1, and finally clip any remaining out-of-bounds values. Formally, if $\tilde{R}_{ij}$ is the unbounded result of the attention-value product, then $\mathbf{R}_{ij} = \text{clip}\big(\alpha(\tilde{R}_{ij} - \mu), -1, 1\big)$, where μ is the global mean of $\tilde{R}_{ij}$ and α is chosen so that $\max_{i,j} |\tilde{R}_{ij} - \mu|$ maps to 1. This ensures $\mathbf{R}$ remains within $[-1, 1]$ while preserving sign information (negative values for mutual exclusivity, positive for co-occurrence) and preventing large magnitudes from destabilizing training or overshadowing other constraints. We do **not** multiply $\mathbf{R}$ directly into the classification logits. Instead, $\mathbf{R}$ is *regularized* to match the prior $\mathbf{C}$ (see $\mathcal{L}_{\text{con}}$ below), thus shaping the final classification layer and backpropagating constraints. If the model tries to output contradictory labels, large penalty arises unless $\mathbf{R}$ or the logits adjust accordingly. This yields an *end-to-end* effect of encouraging consistency with domain constraints.

Adaptive Thresholding. We introduce an uncertainty-aware adaptive thresholding mechanism that adjusts decision boundaries based on prediction confidence and Monte Carlo dropout sampling. During inference, we apply dropout (with rate 0.5) to the feature vector f and perform M stochastic forward passes through the classification layer: $\hat{\mathbf{y}}^{(m)} = \sigma(\mathbf{W}_c\text{Dropout}(\mathbf{f}) + \mathbf{b}_c)$ where $m = 1, \ldots, M$ with $M = 50$ Monte Carlo samples. The predictive uncertainty for each class is computed as: $U_i = \frac{1}{M}\sum_{m=1}^{M}(\hat{y}_i^{(m)} - \bar{y}_i)^2$, where $\bar{y}_i = \frac{1}{M}\sum_{m=1}^{M}\hat{y}_i^{(m)}$ is the mean prediction for class i. The adaptive threshold for each class is then calculated as: $t_i = \alpha_i \cdot t_{\text{base}} + \beta_i \cdot U_i + \delta_i \cdot p_i$, where each term serves a specific purpose: t_{base} provides a starting point that can be adjusted up or down, U_i incorporates model uncertainty to require higher confidence thresholds when predictions are uncertain, and p_i accounts for class frequency in the training data to adjust for class imbalance. The learnable parameters $\alpha_i, \beta_i, \delta_i$ allow the model to fine-tune the importance of each component

for different abnormalities - crucial since some morphological features require more certainty for positive prediction than others (e.g., subtle chromatin changes versus obvious hypersegmentation).

3.3. Training Strategy

We optimize the total loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \lambda_1 \mathcal{L}_{\text{con}} + \lambda_2 \mathcal{L}_{\text{unc}} + \lambda_3 \mathcal{L}_{\text{entropy}}$, where each component serves a distinct purpose in our biological constraint framework:
Binary Cross-Entropy Loss: $\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[y_{ik}\log(\hat{y}_{ik}) + (1-y_{ik})\log\left(1-\hat{y}_{ik}\right)\right]$ ensures accurate classification for each individual abnormality across all N samples and K classes, serving as the primary classification objective.
Constraint Loss: $\mathcal{L}_{\text{con}} = \|\mathbf{R}\mathbf{R}^\top - \mathbf{C}\|_F^2 + \alpha\,\|\mathbf{R}\|_1$ aligns the learned relationships $\mathbf{R}$ with known biological priors $\mathbf{C}$ using the squared Frobenius norm of their difference. The product $\mathbf{R}\mathbf{R}^\top$ captures both direct and transitive relationships, while the ℓ_1 penalty $\|\mathbf{R}\|_1$ encourages sparsity by reducing irrelevant connections. During implementation, we expand the prior matrix $\mathbf{C}$ to match the batch dimension of $\mathbf{R}$ for proper calculation of the Frobenius norm.
Uncertainty Loss: $\mathcal{L}_{\text{unc}} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[\text{KL}(p_{ik}\|\hat{p}_{ik}) + \beta\,\max(0, U_{ik} - \tau)\right]$ performs two critical functions: (1) the KL-divergence term ensures the predicted probability distributions match the ground truth, and (2) the hinge loss term penalizes excessive uncertainty above threshold τ, encouraging confident predictions for clear morphological features.
Entropy Loss: $\mathcal{L}_{\text{entropy}} = -\frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}\left[\mathbf{R}'_{ij}\log\left(\mathbf{R}'_{ij}\right) + (1-\mathbf{R}'_{ij})\log\left(1-\mathbf{R}'_{ij}\right)\right]$, where $\mathbf{R}'$ is $\mathbf{R}$ normalized to $[0,1]$ range. This loss prevents the constraint matrix from becoming overly deterministic by maximizing the binary entropy of each matrix element, allowing flexibility in relationship learning and preventing the model from enforcing excessively rigid constraints.
We set $\lambda_1, \lambda_2, \lambda_3$ to relatively small values (0.1, 0.1, 0.01 respectively) to ensure the constraint terms guide the model without overwhelming the primary classification objective. This balances biological plausibility with classification performance. Similarly, $\alpha = 0.01$ and $\beta = 0.1$ are kept small enough to prevent spurious correlations while maintaining focus on true morphological evidence. During optimization, the constraint and uncertainty terms effectively regularize the model to respect biological relationships while adapting to image-specific evidence.

3.4. Implementation Details

For this study and for efficiency, we utilize DinoBloom-S (Koch et al., 2024) as our feature extraction backbone, which consists of a Vision Transformer (ViT) architecture pretrained on a large corpus of hematological images. The $\mathbf{h}_L$ for the DinoBloom-S model has an embedding dimension of $d = 384$. For more details, please check the original DinoBloom paper. Moreover, the uncertainty threshold τ was set to 0.2 in our experiments, the base threshold t_{base} was set to 0.5, the loss weights as $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.01$ and the hyperparameters $\alpha = 0.01$, $\beta = 0.1$ respectively. During training, different learning rates are employed for each component while keeping the DinoBloom feature extractor frozen: the constraint module uses $\eta_c = 1e-4$, the uncertainty estimation components use $\eta_u = 5e-5$, and the classification head uses $\eta_h = 1e-4$. For the training, we used an AdamW optimizer. Finally, all the information about the constraint matrices (C) per dataset is presented in Appendix A. For computational overhead, our method adds minimal extra parameters (15K for constraint matrix generation, 21 for thresholds) compared to the DinoBloom-S backbone (22M). Regarding inference speed, the biggest factor is the repeated dropout sampling; in practice, 50 MC samples adds approximately 2× inference time overhead compared to a single forward pass. During training, we perform a small number of MC samples (typically 5) at each step or use a differentiable approximation for efficiency. A 50-sample inference is used at test time, increasing the inference time to 0.16 seconds for a single forward pass compared to 0.08 seconds in its absence. Training

Method	Normal		Chromatin		Dohle		Hypergr.		Hyperseg.		Hypogr.		Hyposeg.		Overall	
	wF1	bAcc	wF1	bAcc	wF1	bAcc	wF1	bAcc	wF1	bAcc	wF1	bAcc	wF1	bAcc	wF1	bAcc
DINOv2 ViT-S/14	0.72	0.70	0.43	0.41	0.76	0.74	0.94	0.92	0.79	0.77	0.80	0.78	0.84	0.82	0.68	0.66
+ Ours	0.80	0.78	0.54	0.52	0.84	0.82	0.95	0.93	0.84	0.82	0.86	0.84	0.88	0.86	0.75	0.73
DINOv2 ViT-B/14	0.71	0.69	0.48	0.46	0.77	0.75	0.92	0.90	0.81	0.79	0.82	0.80	0.83	0.81	0.70	0.68
+ Ours	0.82	0.80	0.58	0.56	0.86	0.84	0.95	0.93	0.87	0.85	0.88	0.86	0.89	0.87	0.78	0.76
DINOv2 ViT-L/14	0.72	0.70	0.49	0.47	0.77	0.75	0.89	0.87	0.83	0.81	0.81	0.79	0.84	0.82	0.71	0.69
+ Ours	0.84	0.82	0.60	0.58	0.87	0.85	0.93	0.91	0.89	0.87	0.88	0.86	0.90	0.88	0.80	0.78
DinoBloom-S	0.86	0.84	0.55	0.53	0.86	0.84	0.94	0.92	0.89	0.87	0.90	0.88	0.84	0.82	0.85	0.83
+ Ours	0.93	0.91	0.68	0.66	0.90	0.88	0.98	0.96	0.94	0.92	0.93	0.91	0.90	0.88	0.89	0.87
DinoBloom-B	0.87	0.85	0.61	0.59	0.87	0.85	0.92	0.90	0.91	0.89	0.91	0.89	0.83	0.81	0.86	0.84
+ Ours	0.95	0.93	0.72	0.70	0.91	0.89	0.97	0.95	0.96	0.94	0.94	0.92	0.92	0.90	0.93	0.91
DinoBloom-L	0.88	0.86	0.63	0.61	0.87	0.85	0.91	0.89	0.91	0.89	0.90	0.88	0.84	0.82	0.86	0.84
+ Ours	**0.99**	**0.97**	**0.87**	**0.85**	**0.95**	**0.93**	**1.00**	**1.00**	**0.97**	**0.95**	**0.90**	**0.88**	**0.90**	**0.88**	**0.94**	**0.92**
CTransPath	0.80	0.78	0.52	0.50	0.83	0.81	0.88	0.86	0.80	0.78	0.82	0.80	0.83	0.81	0.74	0.72
+ Ours	0.88	0.86	0.61	0.59	0.89	0.87	0.93	0.91	0.86	0.84	0.87	0.85	0.88	0.86	0.78	0.76
Phikon ViT-B	0.83	0.81	0.54	0.52	0.85	0.83	0.88	0.86	0.82	0.80	0.85	0.83	0.83	0.81	0.76	0.74
+ Ours	0.90	0.88	0.65	0.63	0.91	0.89	0.93	0.91	0.88	0.86	0.90	0.88	0.88	0.86	0.82	0.80

Table 1: Performance comparison on the GR-Neutro dataset showing both weighted F1-score (wF1) and balanced accuracy (bAcc) for each class. The lines with "+ Ours" indicate our method integrated on top of the respective backbone. Best results are in **bold**, second best are underlined.

our full model takes approximately 30 minutes on a single NVIDIA A100 GPU, making it efficient to implement in practice.

4. Experimental Results

4.1. Datasets

We evaluate our method on three datasets, two public and one in-house. **AML Matek Dataset.** (Matek et al., 2019) consists of 18,365 expert-labeled single-cell images with 15 morphological classes and multiple concurrent abnormalities. Following the original split, we use 15,827 images from 100 AML and 100 non-AML patients for training, with the remaining 2,538 images from 40 patients held out for testing. **BMC Dataset.** (Matek et al., 2021) contains 171,373 cells from bone marrow smears of 945 patients, annotated with 21 distinct cell types. The dataset is divided following the original paper's protocol: 137,098 cells (756 patients) for training and 34,275 cells (189 patients) for testing. This split ensures patient-level separation between train and test sets. The dataset is highly imbalanced, with some rare cell types having as few as 8 samples. **GR-Neutro Dataset.** Our dataset comprises 1,934 high-resolution microscopy images of neutrophils, including both normal cells (878 images) and various abnormalities (582 images). For the construction of the dataset, images extracted from 30 patients were used, and the annotations were performed on high-resolution microscopy images of neutrophils collected by a Sysmex DI-60 system. The dataset was split into training (1,455 images) and test (479 images) sets using stratified sampling to maintain class distribution on the neutrophil level. More details about the dataset and its classes are presented in Appendix B. All datasets were preprocessed following the same protocol: images were resized to 224×224 pixels and normalized using mean and standard deviation computed from the training set. To handle class imbalance, we employ a combination of techniques including oversampling of minority classes using SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002), undersampling of majority classes using random undersampling, and class weights in the loss function proportional to the inverse of class frequencies. For DinoBloom comparison, we used their recommended preprocessing pipeline.

Dataset	**Metric**	DINOv2 ViT-S/14	DINOv2 ViT-B/14	DINOv2 ViT-L/14	Phikon ViT-B	DinoBloom-L
AML Matek	wF1	0.88	0.88	0.89	0.88	0.91
	bAcc	0.82	0.82	0.83	0.83	0.86
BMC	wF1	0.68	0.71	0.71	0.73	0.85
	bAcc	0.45	0.49	0.48	0.54	0.64
Method		(Mustaqim et al., 2023)	(Kassahun et al., 2022)	(Wang et al., 2023b)	(Li and et al., 2023)	**Ours**
AML Matek	wF1	0.94	0.94	0.94	0.94	**0.95**
	bAcc	—	—	—	—	**0.91**
BMC	wF1	—	0.85	0.87	0.86	**0.89**
	bAcc	—	0.73	0.74	0.73	**0.75**

Table 2: Performance on the AML Matek (Matek et al., 2019) and BMC (Matek et al., 2021) datasets. Comparisons with different methods and models are provided.

4.2. Comparison with other models

We evaluate our method on multiple foundation model backbones in addition to comparing against existing baselines. Table 1 shows performance on the GR-Neutro dataset for DINOv2 (Oquab et al., 2023) (ViT-S/14, ViT-B/14, ViT-L/14), DinoBloom (Koch et al., 2024) (S, B, L), and Phikon ViT-B (Filiot et al., 2023), both in their original form and with our approach integrated on top (noted as "+ Ours"). We measure weighted F1-score (wF1) and balanced accuracy (bAcc). In every backbone, the addition of our method significantly boosts performance, indicating that our constraints and adaptive thresholding mechanism provide consistent benefits regardless of the underlying architecture. For instance, adding our approach to DinoBloom-L raises the overall wF1/bAcc from 0.86/0.84 to 0.94/0.92. Even when starting from weaker baselines such as DINOv2 ViT-S/14, we observe substantial improvements (roughly +7 to +10 points in overall wF1). These gains highlight the modular nature of our method: it can be easily attached to any existing foundation model backbone to handle multi-label biological constraints. We obtain the highest absolute scores when combined with DinoBloom-L, reaching near-perfect classification metrics (up to 1.00 wF1 on Hypergranulation). Taken together, these results confirm that our approach consistently outperforms both the standalone baselines and the state-of-the-art methods across all backbone architectures in the GR-Neutro dataset. For completeness, Table 2 presents public-dataset results (AML Matek and BMC). There too, our approach on top of DinoBloom or other backbones yields the best reported figures. Hence, our biologically-constrained multi-label framework not only excels under various data distributions but also adapts seamlessly to any backbone or domain-specific architecture, making it widely applicable in clinical settings. Moreover, Table 2 summarizes the comparison of our approach against several state-of-the-art methods on AML Matek (Matek et al., 2019) and BMC (Matek et al., 2021), including recent Transformer-based techniques (Mustaqim et al., 2023), multitask and self-supervised methods (Kassahun et al., 2022; Wang et al., 2023b; Li and et al., 2023). Our approach on AML Matek achieves a wF1-score of 0.95 and a bAcc of 0.91, outperforming prior methods such as (Wang et al., 2023b), (Kassahun et al., 2022) and (Li and et al., 2023), all of which reported wF1 up to 0.94. Notably, those references do not report balanced accuracy, making our 0.91 bAcc a strong indicator of performance on less frequent cell types. In addition, for the BMC we obtain with our method a wF1 of 0.89 with a 0.75 of bAcc. This outperforms multi-task and self-supervised approaches like Kassahun et al. (Kassahun et al., 2022) (0.85 wF1, 0.73 bAcc) and Wang et al. (Wang et al., 2023b) (0.87 wF1, 0.74 bAcc), confirming that our constraint-based approach manages the extensive class imbalance of the BMC dataset effectively.

The per-class performance analysis on the GR-Neutro dataset presented in Table 7 and Appendix C reveals the robustness of our approach across different abnormality types. Most notably, our method achieves perfect accuracy for hypergranulation detection and near-perfect performance

for normal cell classification. The substantial improvements in challenging cases like chromatin condensation (0.87 vs 0.63 in DinoBloom-L) and Dohle bodies (0.95 vs 0.87) demonstrate the effectiveness of our constraint-based learning approach in handling subtle morphological variations. The consistent performance across all classes, including traditionally challenging ones like hyposegmentation and hypogranulation, highlights the balanced nature of our approach. Some examples from the attention maps obtained by the constraint attention are presented in Appendix D.

4.3. Ablation Study

Model Variant	Accuracy	F1	AUC	Spec.	Sens.
Base Model	0.81	0.79	0.83	0.82	0.80
+ Learnable constraints (R)	0.88	0.87	0.90	0.89	0.88
+ Adaptive Thresholds	0.94	0.93	0.95	0.93	0.94
Constraint Layer Variants					
Fixed Rules (C)	0.85	0.83	0.86	0.84	0.85
Binary Constraints	0.86	0.84	0.87	0.85	0.86
Fully Connected Network	0.85	0.84	0.86	0.84	0.85
Soft Constraints (Ours)	**0.88**	**0.87**	**0.90**	**0.89**	**0.88**
Threshold Mechanism					
Static Threshold	0.89	0.88	0.90	0.88	0.89
Class-Specific	0.90	0.89	0.91	0.90	0.90
Uncertainty-Aware (Ours)	**0.94**	**0.93**	**0.95**	**0.93**	**0.94**

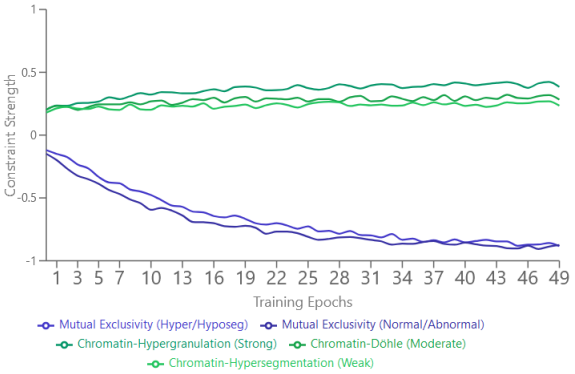

Figure 2: **Left:** Ablation study analysis showing the impact of each model component. **Right:** Evolution of learned constraint relationships during training, showing how the model discovers biologically meaningful patterns - mutual exclusivity constraints converge to strong negative values, while co-occurrence relationships stabilize at different positive strengths based on their biological significance.

To highlight the effectiveness of each component of our method, we conducted an extensive ablation study summarized in Figure 2 (Left). The results are presented in three distinct evaluation settings to demonstrate both the cumulative and individual impacts of our key components. First, we evaluate the progressive addition of components, starting with a base model (0.81 accuracy), then adding learnable constraints (improving to 0.88), and finally incorporating adaptive thresholds (reaching 0.94). This demonstrates the cumulative benefit of our complete architecture.

For constraint mechanisms, we compare four variants in isolation: (i) *fixed rules*, which uses manually predefined constraints (e.g., hardcoding that hyper/hyposegmentation cannot co-occur), achieving 0.85 accuracy, (ii) *binary constraints*, where relationships between abnormalities are limited to strict 0/1 values, reaching 0.86 accuracy, (iii) a *fully connected network* approach where the constraint matrix is generated directly from global features using a two-layer neural network: $\mathbf{R} = \tanh(W_2 \cdot \mathrm{ReLU}(W_1 \cdot \mathbf{f} + b_1) + b_2)$. For the GR-Neutro dataset with $K = 7$ classes, $W_1 \in \mathbb{R}^{28 \times d}$ transforms the feature vector to a 28-dimensional hidden representation, and $W_2 \in \mathbb{R}^{49 \times 28}$ projects this to a vector that is reshaped into the 7×7 constraint matrix. This approach achieved 0.85 accuracy, comparable to fixed rules, and (iv) our *soft constraints* approach, which learns continuous values between 0 and 1 to represent relationship strengths, achieving 0.88 accuracy.

Similarly, for thresholding mechanisms, we evaluate three approaches: (i) *static threshold*, using a fixed threshold (0.5) for all classes, achieving 0.89 accuracy, (ii) *class-specific threshold*, where each class has its own learned threshold, improving to 0.90 accuracy, and (iii) our *uncertainty-aware* approach, which dynamically adjusts thresholds based on prediction confidence, reaching 0.94 accuracy when integrated with the full model. These results demonstrate that both components contribute significantly to model performance, with each achieving their best results when combined in the full architecture.

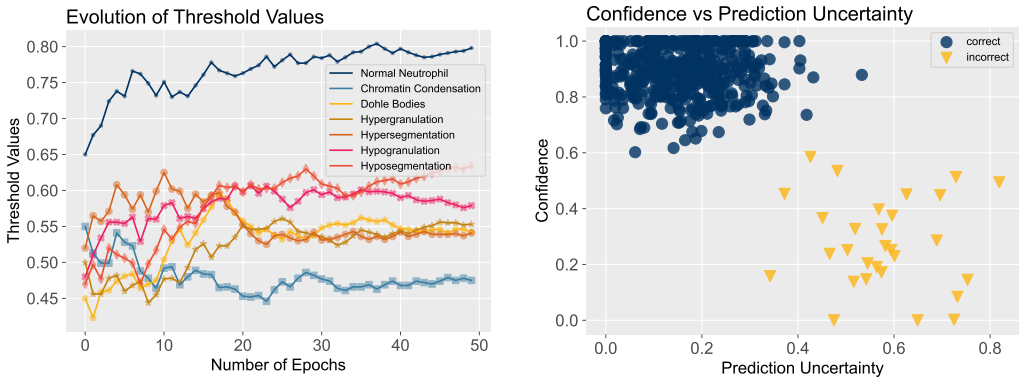

Figure 3: **Left:** Evolution of class-specific adaptive thresholds during training, showing how thresholds adapt to different neutrophil abnormalities based on their characteristics for the GR-Neutro, and **Right:** Relationship between model uncertainty and prediction confidence, showing clear separation between correct (blue) and incorrect (yellow) predictions for the GR-Neutro Dataset.

The evolution of constraints during training (Figure 2, Right) demonstrates how our model discovers and enforces biological relationships in the GR-Neutro Dataset. The mutual exclusivity constraints between hypersegmentation and hyposegmentation converge to strong negative values $\leq$-0.8, indicating the model's clear understanding of incompatible cell states. Moreover, (Figure 2, Right), co-occurrence constraints show more nuanced behavior ($\geq$ 0.45 for chromatin-hypergranulation, $\geq$0.35 for chromatin-Döhle, and $\geq$0.25 for chromatin-hypersegmentation), reflecting the varying strengths of these biological associations. This hierarchy of learned constraints aligns with clinical observations, where mutual exclusivity represents fundamental biological impossibilities while co-occurrences represent more flexible, probabilistic relationships.

Figure(3, Left) shows the evolution of class-specific adaptive thresholds during training. The thresholds start from a common base value (0.5) and gradually diverge based on class-specific characteristics. Normal neutrophils converge to higher threshold values (around 0.7), reflecting the need for higher confidence when classifying normal cells. In contrast, abnormality thresholds settle at different levels (between 0.4-0.6), with chromatin condensation requiring the lowest threshold (0.4) due to its subtle nature, and more obvious features like hypersegmentation maintaining moderate thresholds (0.55). This adaptive behavior enables the model to account for varying degrees of morphological distinctiveness across different abnormalities while maintaining high specificity. Finally, Figure(3, Right) shows the relationship between prediction confidence and uncertainty, where correct predictions (shown in blue) are clearly separated from incorrect ones (shown in yellow). The model demonstrates high confidence ($\geq$0.8) and low uncertainty ($\leq$0.2) for correct predictions, while incorrect predictions show higher uncertainty ($\geq$0.4) and lower confidence values ($\leq$0.6), validating the effectiveness of our uncertainty estimation approach.

5. Conclusion

We presented a novel approach for learning and enforcing biological constraints in multi-label classification of neutrophil abnormalities. Our method not only improves classification accuracy but also provides valuable insights into morphological relationships through learned constraints. The adaptive thresholding mechanism effectively handles varying degrees of abnormality manifestation, while the learnable constraint satisfaction layer prevents biologically impossible predictions. Future work will focus on expanding to larger, multi-center datasets and incorporating temporal dynamics in neutrophil analysis.

Acknowledgment

We thank Qube Research & Technologies for their support. This work was performed using Jean-Zay supercomputer from IDRIS. This work was partially supported by the ANR-23-IAHU-0002, ANR-21-CE45-0007 and ANR-23-CE45-0029.

References

D.A. Arber, A. Orazi, R. Hasserjian, J. Thiele, M.J. Borowitz, M.M. Le Beau, C.D. Bloomfield, M. Cazzola, and J.W. Vardiman. *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*. IARC, revised 4th edition, 2017.

B. J. Bain. *Blood Cells: A Practical Guide*. Wiley-Blackwell, 6th edition, 2015.

Carol Briggs, Ian Longair, Michael Slavik, Kathryn Thwaite, Roger Mills, Vasanth Thavaraja, Anna Foster, Daniel Romanin, and Samuel J. Machin. Quality counts: new parameters in blood cell counting. *International Journal of Laboratory Hematology*, 31(3):277–297, 2009. URL https://doi.org/10.1111/j.1751-553X.2009.01160.x.

Andrew Bruton, Nataša Sladoje, and Joakim Lindblad. Multi-label learning in biomedical image analysis using graph-based label dependencies. *Pattern Recognition*, 127:108628, 2022. URL https://doi.org/10.1016/j.patcog.2022.108628.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. URL https://doi.org/10.1613/jair.953.

Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 2019. URL https://doi.org/10.1016/j.media.2019.101539.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark A. DePristo, Katherine Chou, Claire Cui, Greg S. Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019. URL https://doi.org/10.1038/s41591-018-0316-z.

Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling, 2023. URL https://doi.org/10.1101/2023.07.21.23292757. medRxiv preprint 2023.07.21.23292757.

A. Victor Hoffbrand and Paul A. H. Moss. *Hoffbrand's Essential Haematology*. Wiley-Blackwell, Chichester, UK, 7th edition, 2016. ISBN 978-1118408674.

Y. Kassahun, S. Soomro, N. Khan, G. Fink, and J. Ostermann. Deep multi-task learning for bone marrow cell classification. *IEEE Access*, 10, 2022.

Valentin Koch, Sophia J. Wagner, Salome Kazeminia, Ece Sancar, Matthias Hehr, Julia A. Schnabel, Tingying Peng, and Carsten Marr. Dinobloom: A foundation model for generalizable cell embeddings in hematology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, Proceedings, Part XII*, volume 15012 of *Lecture Notes in Computer Science*, pages 520–530. Springer, 2024. URL https://doi.org/10.1007/978-3-031-72390-2_49.

Richard R. Kortum and Samuel G. Armato. Enforcing morphological constraints in deep learning-based segmentation of histopathology images. *IEEE Access*, 11:46490–46501, 2023. URL https://doi.org/10.1109/ACCESS.2023.1007922.

Y. Li and et al. A comparative study of vision transformers and convolutional models for bone marrow cell morphology, 2023. Preprint, submitted to BMC Bioinformatics.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra A. Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. URL https://doi.org/10.1016/j.media.2017.07.005.

Christian Matek, Simone Schwarz, Karsten Spiekermann, and Carsten Marr. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nature Machine Intelligence*, 1(11):538–544, 2019. URL https://doi.org/10.1038/s42256-019-0101-9.

Christian Matek, Sebastian Krappe, Christian Münzenmayer, Torsten Haferlach, and Carsten Marr. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood*, 138(20):1917–1927, 2021. URL https://doi.org/10.1182/blood.2020010568.

Tanzilal Mustaqim, Chastine Fatichah, and Nanik Suciati. Identification of acute lymphoblastic leukemia subtypes on a microscopic image of multicellular blood using object detection model with swin transformer. In *Proceedings of the 2023 7th International Conference on Medical and Health Informatics*, pages 280–286, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193. arXiv preprint arXiv:2304.07193.

Thuan Pham, Xiaohui Tao, Ji Zhang, Jianming Yong, Yuefeng Li, and Haoran Xie. Graph-based multi-label disease prediction model learning from medical data and domain knowledge. *Knowledge-based systems*, 235:107662, 2022.

R.E. Raskin, M. Bishop, and K.E. Lamb. Cytoplasmic and nuclear changes in toxic neutrophils. *Veterinary Clinical Pathology*, 49(1):9–19, 2020.

C. L. Swaggerty. Granulocytic dysplasia: Hypersegmented and hyposegmented neutrophils. *Clinics in Laboratory Medicine*, 39(1):69–80, 2019.

Yevgen Tolkach, Theresa Dohmgörgen, and Glen Kristiansen. Protein expression-based subtyping of clinically relevant b-cell lymphoma categories by deep learning. *Scientific Reports*, 9:7784, 2019. URL https://doi.org/10.1038/s41598-019-44178-3.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022. URL https://doi.org/10.1016/j.media.2022.102559.

Xiyue Wang, Yuexi Du, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Retccl: Clustering-guided contrastive learning for whole-slide image retrieval. *Medical Image Analysis*, 83:102645, 2023a. URL https://doi.org/10.1016/j.media.2022.102645.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Masked autoencoders for self-supervised bone marrow cell classification. *Computers in Biology and Medicine*, 152:106233, 2023b. URL https://doi.org/10.1016/j.compbiomed.2023.106233.

Appendix A. Constraint Prior Matrices for Each Dataset

In this appendix, we provide details on the prior constraint matrices $\mathbf{C}$ used for each dataset (AML Matek, BMC, and GR-Neutro). Recall that $C_{ij} = -1$ indicates mutual exclusivity between classes i and j, $C_{ij} = 0$ indicates no direct relationship, and $C_{ij} = c_{ij} > 0$ indicates a soft co-occurrence with strength $c_{ij} \in [0, 1]$. **Why these specific values?** Negative entries (-1) denote biologically impossible co-occurrences such as hypersegmentation vs. hyposegmentation or blasts vs. fully segmented neutrophils; references (Hoffbrand and Moss, 2016; Bain, 2015; Briggs et al., 2009; Arber et al., 2017; Swaggerty, 2019; Matek et al., 2019; Raskin et al., 2020) affirm these strict incompatibilities. Zero entries indicate no documented interaction (e.g., RBC artifacts vs. granulocytic abnormalities). Meanwhile, moderate positive values (0.2–0.4) reflect partial or probabilistic overlaps; for example, Döhles bodies can co-occur with other "toxic changes" (Raskin et al., 2020), or adjacent stages of myeloid maturation can appear together (Hoffbrand and Moss, 2016; Bain, 2015). These numeric strengths were chosen to encode the relative likelihood of either co-occurrence or mutual exclusivity, without implying absolute determinism.

A.1 GR-Neutro Dataset (7×7)

Table 3: Constraint matrix $\mathbf{C}$ for the GR-Neutro dataset.

	Normal	Chromatin	Döhle	Hypergran.	Hyperseg.	Hypogran.	Hyposeg.
Normal	0	-1	-1	-1	-1	-1	-1
Chromatin	-1	0	0.3	0.4	0.3	0.2	0.3
Döhle	-1	0.3	0	0.3	0.2	0.2	0.2
Hypergran.	-1	0.4	0.3	0	0.3	-1	0.2
Hyperseg.	-1	0.3	0.2	0.3	0	0.2	-1
Hypogran.	-1	0.2	0.2	-1	0.2	0	0.2
Hyposeg.	-1	0.3	0.2	0.2	-1	0.2	0

A.2 AML Matek Dataset (15×15)

A.3 BMC Dataset (21×21)

Key relationships include: (i) sequential maturation stages have positive co-occurrence (0.3), (ii) distinct lineages are mutually exclusive (-1), (iii) RBC/platelet artifacts or "rare" cells show weak correlations (0.15–0.2) to certain lineages, and (iv) early precursors can weakly co-occur (0.2–0.3) with adjacent developmental stages.

Table 4: Constraint matrix **C** for the AML Matek dataset.

	Myelo.	Promyelo.	Myelo.	Meta.	Band	Segm.	Eos.	Baso.	Mono.	Lymph.	Plasma	Erythro.	RBC/Plt	Rare	Art.
Myeloblast	0	0.3	-1	-1	-1	-1	-1	-1	0.2	-1	-1	0	-1	0.2	0
Promyelocyte	0.3	0	0.3	-1	-1	-1	0.2	0.2	-1	-1	-1	0	-1	0.2	0
Myelocyte	-1	0.3	0	0.3	-1	-1	0.2	0.2	-1	-1	-1	0	-1	0.2	0
Metamyelocyte	-1	-1	0.3	0	0.3	-1	-1	-1	-1	-1	-1	0	-1	0.2	0
Band Neut.	-1	-1	-1	0.3	0	0.3	-1	-1	-1	-1	-1	-1	-1	0.2	0
Segm. Neut.	-1	-1	-1	-1	0.3	0	-1	-1	-1	-1	-1	-1	-1	0.2	0
Eosinophil	-1	0.2	0.2	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	0.2	0
Basophil	-1	0.2	0.2	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	0.2	0
Monocyte	0.2	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	0.2	0
Lymphocyte	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0.2	-1	-1	0.2	0
Plasma Cell	-1	-1	-1	-1	-1	-1	-1	-1	-1	0.2	0	-1	-1	0.2	0
Erythroblast	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	0	0.3	0.2	0
RBC/Platelet	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0.3	0	0.2	0
Rare/Atypical	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0	0.2
Artifact	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0

Table 5: Constraint matrix **C** for the BMC dataset.

	Myelo.	Prom.	Myelo.	Meta.	Band	Seg.	Eos.	Baso.	Mono.	Lymph.	Plas.	Eryth.	Mega.	Pro-E	Baso-E	Poly-E	Ortho-E	RBC	Art.	Smudge	Other
Myeloblast	0	0.3	-1	-1	-1	-1	-1	-1	0.2	-1	-1	0	0	0	0	-1	-1	-1	0	0.15	0
Promyelocyte	0.3	0	0.3	-1	-1	-1	0.2	0.2	-1	-1	-1	0	0	0	0	-1	-1	-1	0	0	0
Myelocyte	-1	0.3	0	0.3	-1	-1	0.2	0.2	-1	-1	-1	0	0	0	0	-1	-1	-1	0	0	0
Metamyelocyte	-1	-1	0.3	0	0.3	-1	-1	-1	-1	-1	-1	0	0	0	0	-1	-1	-1	0	0	0
Band Neut.	-1	-1	-1	0.3	0	0.3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0
Segm. Neut.	-1	-1	-1	-1	0.3	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0
Eosinophil	-1	0.2	0.2	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0
Basophil	-1	0.2	0.2	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0
Monocyte	0.2	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0
Lymphocyte	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0.2	-1	-1	-1	-1	-1	-1	-1	0	0.15	0
Plasma Cell	-1	-1	-1	-1	-1	-1	-1	-1	-1	0.2	0	-1	-1	-1	-1	-1	-1	-1	0	0	0
Erythroblast	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	0	-1	0.3	0.3	0.3	0.3	0.3	0	0	0
Megakaryocyte	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	0.2	0	0	0
Pro-Erythro.	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	0.3	-1	0	0.3	-1	-1	-1	0	0	0
Baso-Erythro.	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	0.3	-1	0.3	0	0.3	-1	-1	0	0	0
Poly-Erythro.	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0.3	-1	-1	0.3	0	0.3	-1	0	0	0
Ortho-Erythro.	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0.3	-1	-1	-1	0.3	0	0.3	0	0	0
RBC	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0.3	0.2	-1	-1	-1	0.3	0	0	0	0
Artifact	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.2	
Smudge	0.15	0	0	0	0	0	0	0	0	0.15	0	0	0	0	0	0	0	0	0.2	0	0.2
Other	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0.2	0

Appendix B. Details about the GR-Neutro dataset

The **GR-Neutro** dataset is composed of 7 different classes including normal neutrophils, nuclear chromatin condensation, Döhle bodies (basophilic cytoplasmic inclusions), hypergranulation (increased cytoplasmic granulation), hypersegmentation (increased nuclear lobes), hypogranulation (decreased cytoplasmic granulation), and hyposegmentation (decreased nuclear lobes). Table 6 includes the number of each class for the training and testing splits.

Split	**Normal**	**Chromatin**	**Dohle**	**Hypergr.**	**Hyperseg.**	**Hypogr.**	**Hyposeg.**
# Train	658	277	56	45	46	279	253
# Test	220	93	19	15	16	93	84

Table 6: Distribution of samples across training and test splits in the GR-Neutro dataset.

Appendix C. Cross validation results for the GR-Neutro dataset

We performed comprehensive 5-fold cross-validation to the **GR-Neutro** dataset to ensure robust evaluation of our approach. Table 7 presents detailed results across all folds. Results show consistent performance improvements across all folds, with low variance in key metrics.

Appendix D. Grad-CAM Visualizations

Figures 4 illustrates example Grad-CAM attention maps comparing our full model (constraint module + adaptive thresholds) with the baseline DinoBloom-S classifier across different neutrophil abnormalities. We observe that our model (middle column) consistently focuses on morphologically relevant regions, while the baseline model (right column) often attends to irrelevant background areas or fails to concentrate on diagnostically important features.

Particularly noteworthy is our model's ability to maintain focus on the nuclear region for nuclear-related abnormalities (chromatin condensation, hypersegmentation, hyposegmentation) and on the cytoplasmic region for granular abnormalities (hypergranulation, hypogranulation) and Döhle bodies. This alignment with clinical diagnostic practice demonstrates how incorporating biological

Table 7: 5-Fold Cross-Validation Results

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.93	0.94	0.92	0.95	0.93
Macro F1	0.92	0.93	0.91	0.94	0.92
AUC-ROC	0.94	0.95	0.93	0.96	0.94
Per-Class F1-Scores					
Normal	0.98	0.99	0.98	0.99	0.98
Chromatin	0.86	0.88	0.85	0.89	0.87
Döhle	0.94	0.96	0.93	0.97	0.95
Hypergran.	0.99	1.00	0.98	1.00	0.99
Hyperseg.	0.96	0.98	0.95	0.98	0.97
Hypogran.	0.89	0.91	0.88	0.92	0.90
Hyposeg.	0.89	0.91	0.88	0.92	0.90

constraints helps the model develop more interpretable attention patterns that mirror expert reasoning.

Appendix E. Sensitivity Analysis

An extensive sensitivity analysis is presented in this section performed in the GR-Neutro dataset.

Appendix E.1 Dropout Rate Selection Sensitivty

We performed a sensitivity analysis for different dropout rates $\in \{0.2, 0.3, 0.4, 0.5, 0.6\}$ on a held-out validation subset. Table 8 summarizes the results. A lower rate (0.2–0.3) tended to underestimate uncertainty and yielded slightly lower macro-F1, while higher rates (0.6) degraded feature quality and classification metrics. We found that 0.5 provided the best balance between performance and well-calibrated uncertainty.

Table 8: Sensitivity analysis of different dropout rates on the validation subset. 'Avg. Unc.' stands for the mean predicted uncertainty across classes.

Dropout Rate	Accuracy (%)	Macro-F1 (%)	Avg. Unc.	AUROC (%)
0.2	90.5	89.2	0.12	92.1
0.3	91.0	89.8	0.15	92.4
0.4	92.2	90.7	0.19	93.6
0.5	**93.5**	**91.5**	**0.20**	**94.2**
0.6	91.2	88.9	0.23	91.7

Appendix E.2 Monte Carlo Sampling Sensitivity

We analyzed how many Monte Carlo (MC) samples were required for stable uncertainty estimates without excessive computational overhead. As shown in Table 9, moving from 10 to 50 MC samples improved the macro-F1 from 90.5% to 92.0%. Increasing further to 100 changed the mean uncertainty by less than 0.5% but roughly doubled the inference cost compared to 50 samples.

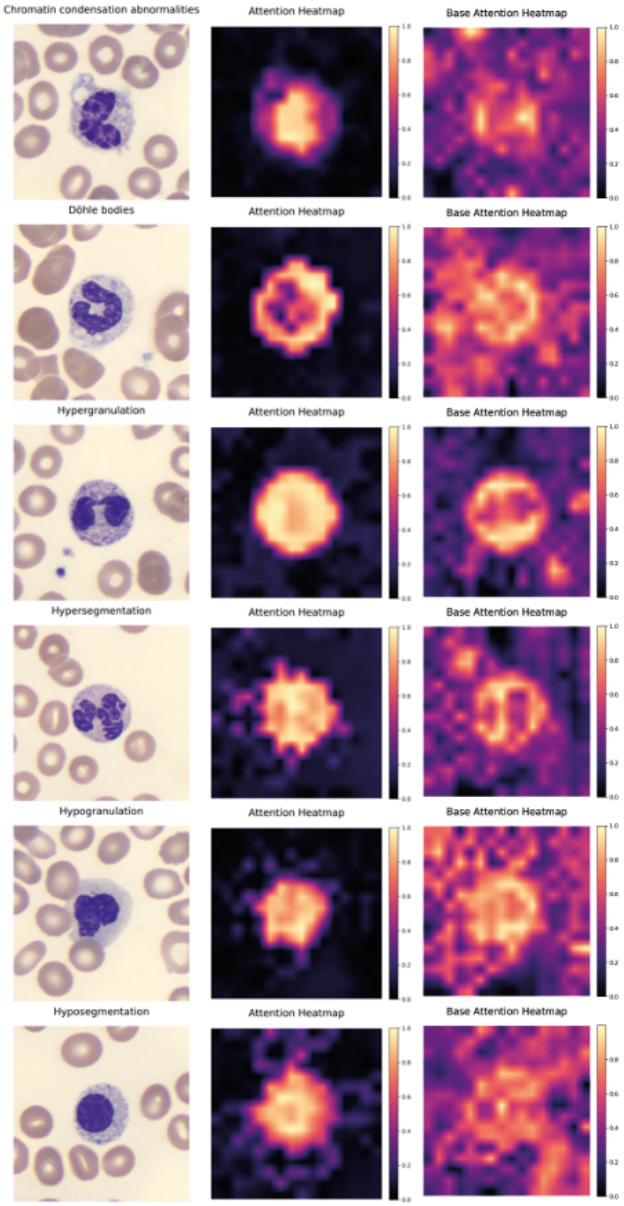

Figure 4: **Grad-CAM attention maps** on sample neutrophil images from different abnormality classes. (Left) Original neutrophil images; (Middle) Our constrained model based on DinoBloom-S, correctly focuses on informative regions. (Right) Baseline DinoBloom-S showing less consistent attention patterns with significant focus on background or irrelevant regions.

From Table 9, we conclude that 50 MC samples provides near-optimal performance with an acceptable inference-time overhead (2.0× compared to a single forward pass). Increasing to 100 samples yields diminishing returns (only +0.2% in macro-F1 and -0.1% in mean uncertainty) yet doubles the computational load relative to 50 samples.

Table 9: Sensitivity analysis of different numbers of Monte Carlo samples. 'Inference Time×' is the relative inference cost compared to a single forward pass (1.0×).

MC Samples	Accuracy (%)	Macro-F1 (%)	Mean Unc. (%)	Inference Time×
1 (no MC)	89.0	88.2	–	1.0×
10	91.5	90.5	3.1	1.3×
20	92.0	91.0	2.8	1.6×
50	93.0	92.0	2.4	2.0×
100	93.3	92.2	2.3	4.0×

E.3 Sensitivity to Random Seeds and Hyperparameters

To assess model robustness, we repeated experiments with different random seeds and hyperparameter settings. Table 10 presents these results.

Table 10: Sensitivity analysis for the final model on GR-Neutro. (A) Accuracy under different random seeds (5 runs). (B) Impact of varying λ-hyperparameters.

(A) Random Seeds (5 runs)		
Seed	**Accuracy (%)**	**Macro-F1 (%)**
Seed 1	93.0	92.4
Seed 2	94.0	93.2
Seed 3	92.8	92.0
Seed 4	93.2	92.5
Seed 5	93.7	92.8
Mean	93.3	92.6
Std.	±0.6	±0.5

(B) Varying Hyperparameters			
λ_1	λ_2	λ_3	Macro-F1 (%)
0.1	0.1	0.01	93.0
0.12	0.12	0.012	92.3
0.08	0.08	0.008	92.0
0	0.1	0.01	85.0
0.1	0	0.01	88.7
0.1	0.1	0	91.5

The random seed experiments (Table 10A) show minimal variance in performance with a standard deviation of only ±0.6% in accuracy, demonstrating the model's stability across different initializations. For hyperparameter sensitivity (Table 10B), we tested various combinations, finding that small changes to all parameters simultaneously result in only minor performance decreases (92.0-92.3% vs. 93.0% baseline). More importantly, completely removing individual constraint components reveals their impact: setting $\lambda_1 = 0$ (removing the constraint loss) causes the most significant drop to 85.0%, with contradictory predictions appearing; setting $\lambda_2 = 0$ (removing uncertainty awareness) reduces performance to 88.7%; and setting $\lambda_3 = 0$ (removing entropy regularization) has the smallest impact, with performance at 91.5%.

E.4 Uncertainty Threshold Selection

The uncertainty threshold τ controls how much predictive uncertainty is acceptable before the model is penalized in the loss function. Finding the optimal τ is critical for balancing sensitivity to subtle abnormalities with specificity in classification. Table 11 shows our systematic evaluation of different threshold values.

The threshold $\tau = 0.2$ achieved optimal performance with 94.0% accuracy and 93.0% F1-score. Lower thresholds ($\tau = 0.1$) were too restrictive, causing the model to miss subtle abnormalities

Table 11: Impact of uncertainty threshold τ on model performance (GR-Neutro dataset)

τ Value	Accuracy (%)	F1-Score (%)	False Positives (%)
0.1	91.5	90.8	3.2
0.2	**94.0**	**93.0**	**5.1**
0.3	92.7	91.5	7.8
0.4	90.2	89.3	11.3
0.5	88.5	87.1	14.7

despite having only 3.2% false positives. Higher thresholds ($\tau \geq 0.3$) progressively degraded performance by allowing excessive uncertainty in predictions, leading to increased false positive rates (7.8% at $\tau = 0.3$, rising to 14.7% at $\tau = 0.5$).

Appendix F. Computational Complexity

With respect to the computational complexity, our method adds minimal extra parameters (approximately $15K$ for constraint matrix generation, 21 for thresholds) compared to the DinoBloom-S backbone ($22M$). For inference speed, MC dropout sampling adds a $2\times$ overhead. Alongside the time cost, we also examined the approximate FLOPs needed for inference. Using 50 Monte Carlo samples effectively doubles the FLOPs compared to a single forward pass of the chosen backbone, consistent with the increased number of forward evaluations. In our typical implementation, our method runs at about 2×10^{11} FLOPs per image with 50 samples (vs. 1×10^{11} FLOPs for a single pass), though exact figures may vary depending on hardware optimizations and specific backbone architectures. As before, reducing the Monte Carlo samples proportionally lowers the FLOPs, allowing a trade-off between computational overhead and uncertainty quantification.