Learning Structure-Aware Foundational Representation of Rat Testicular Tubules Using Multiple Instance Learning

Vedang Kshirsagar VEDANG.KSHIRSAGAR@AIRAMATRIX.COM
Saketh Juturu SAKETH.JUTURU@AIRAMATRIX.COM
Geetank Raipuria GEETANK.RAIPURIA@AIRAMATRIX.COM
Nitin Singhal NITIN.SINGHAL@AIRAMATRIX.COM
Mumbai, India

Editors: Under Review for MIDL 2026

Abstract

Testicular toxicity is a critical factor in preclinical drug safety assessment, yet automated modelling of testicular abnormalities remains largely unexplored. Unlike liver or kidney tissue, testes tubules vary widely in size and structure, making fixed-resolution patch classification ineffective. We first show that resizing tubules significantly degrades performance especially for larger sized tubules and a Multiple Instance Learning (MIL) model offers substantial improvements. Building on this, we propose TBA-MIL, a transformer-based aggregation model with learnable positional embeddings that encodes the structure of tubules and introduce MIM-MIL, a self-supervised masked instance modelling framework that learns tubule-relevant representations from large-scale unlabeled data. Across four tubule types, TBA-MIL with MIM-MIL outperforms state-of-the-art MIL models and establishing a strong baseline for automated testicular toxicity assessment.
Keywords: Toxicology, Foundational Models, Multiple Instance Learning, Masked Image Modelling, Histopathology, Testes, Testicular Toxicity

1. Introduction

Histopathology based drug induced tissue injury identification is a critical step in a preclinical safety assessment of a potential therapeutic agent. A pathologist identifies, characterizes, and grades any observed microscopic changes on tissue samples from dosed animals. The process provides the tissue morphological data that helps the toxicologic pathologist understand, characterize, and ultimately predict a drug's potential toxic effects. The assessment allows translating potential risks to the human clinical setting and informing appropriate dosing strategies and monitoring requirements.

With advancements in digital pathology and Whole Slide Images (WSI), deep learning models have been shown promising results for identifying a wide variety of tissue injuries (Zingman et al., 2024; Jaume et al., 2024a; Juturu et al., 2025; Zehnder et al., 2022; Linmans et al., 2024; Pocevičiūtė et al., 2025), which assist pathologists in determining drug induced toxicity. However, most of these works have focused primarily on Liver (Hepatic) or Kidney (Renal) toxicity, the major metabolic and excretory organs. While liver and kidney failure are immediately life-threatening, testicular toxicity is critical for often program-ending reasons including unique and irreplaceable organ function, low tolerance for risk in non-life-saving drugs and secondary impact on hormone production. This makes Testicular Toxicity findings a decisive factor in terminating a drug program early, particularly when the drug is intended for a non-life-threatening or chronic condition in a broad patient population.

Models for liver and kidney injury detection have been developed on tissue patches extracted from WSI, as the tissue injuries can be identified without requiring larger tissue context. However, in case of testes tissue, the organ consists of tubules as sub-structures which can reflect varying degrees of toxicity. Figure 1 provides sample of normal and drug affected abnormal tubules. Creating patches from the tissue would not provide sufficient context to identify the injury. Furthermore, tubules in a WSI vary significantly in size and aspect ratio, as seen in figure 1 and 5. This makes modelling tubules for identification of drug induced toxicity non-trivial.

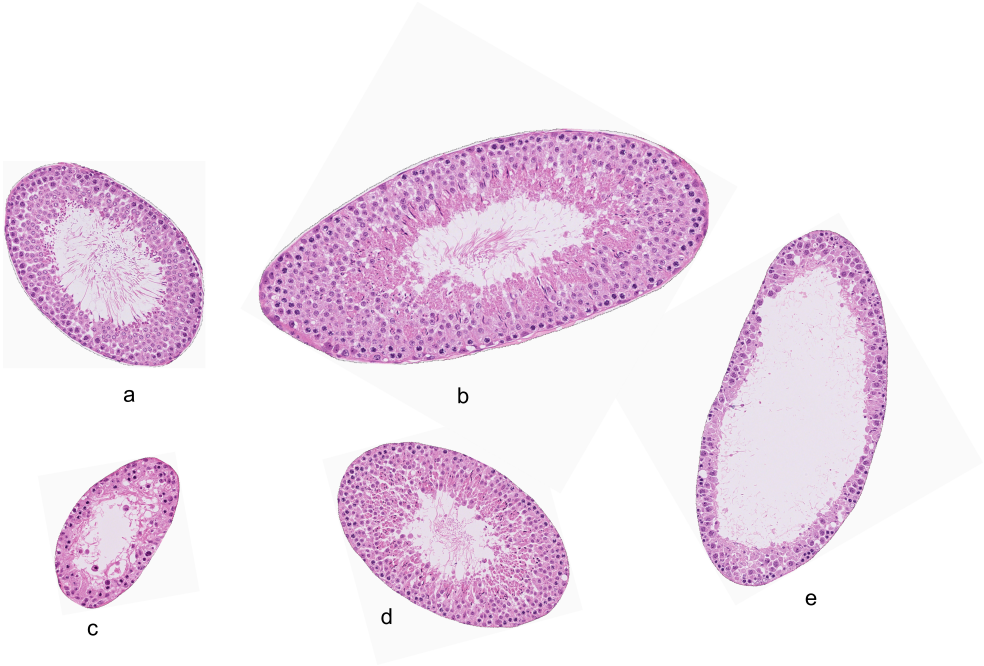

Figure 1: Sample images of testes tubules, resized proportionally to maintain relative size. a-b) Normal tubule (No injuries, intact) c) Tubular Degeneration (Tubular tissue disintegration) d) Germ Cell Degeneration (Cellular level injuries) e) Tubular Dilation (Tubular tissue expansion)

In this work, we first investigate the performance of state-of-the-art convolution and transformer-based feature extractors to classify testes tubules into normal and injury classes. For this, the tubules are resized to a fixed size before passing to the model. We extensively evaluate by ablating for different input image sizes. The analysis shows that resizing tubule images significantly degrades performance on large sized tubules. Next, we model a tubule as a bag-of-words, i.e., Multiple Instance Learning, extracting features of non-overlapping patches from a tubule, followed by feature aggregation. The analysis shows that the MIL model outperforms the resized image-based classifier for all tubule sizes.

Based on this evidence, we propose a MIL model based on the transformer architecture. The MIL model uses positional encoding to retain the relative position of the patch instances, uses a foundational feature extractor trained on testes data, and importantly is itself pre-trained using self-supervised loss to learn tubular representations. We compare the proposed framework with state-of-the-art MIL models and show that it outperforms on tubular injury classification. We also perform an ablation study to show the significance of

pre-training and positional embedding in the MIL model. Figure 4 provides an overview of the proposed representation learning for testicular tubules.

To the best of our knowledge, this work is the first to extensively evaluate testicular injury classification in Wistar rats. Our main contributions are as follows:

1. We demonstrate that modelling each tubule as a bag of features using Multiple Instance Learning (MIL) outperforms fixed-size image classification approaches.
2. We benchmark state-of-the-art MIL models for tubule injury classification
3. We present a new transformer-based aggregation MIL model (TBA-MIL) and a foundational self-supervised pretraining strategy for tubular representation learning - Masked Instance Modelling (MIM-MIL).

2. Related Work

2.1. Drug Induced Injury Detection

Recent works on drug-induced injury detection in digital histopathology emphasize domain-specific representation learning and out-of-distribution (OOD) modelling to address the scarcity and heterogeneity of lesion annotations in toxicology pipelines. (Zingman et al., 2024) and (Dippel et al., 2024) employ supervised patch classification task to train a feature extractor, followed by detecting patches anomalous from normal tissue representation. Other works leverage self-supervised foundational models to extract features for supervised injury classification (Jaume et al., 2024a) and unsupervised anomaly detection (Juturu et al., 2025) using a neighbourhood density (K-Nearest Neighbours) in the latent space of foundation model.

Another direction of work uses generative models to learn normal tissue representation via reconstruction task, and identify anomalous tissue injury as tissue with high reconstruction error. (Zehnder et al., 2022) trained a generative adversarial network with multi-scale patches as input to enhance regional interpretation, whereas (Linmans et al., 2024) train a denoising diffusion probabilistic with a partial diffusion process to learn in-distribution image space, and found it to outperform GAN based models. Lastly, (Pocevičiūtė et al., 2025) and (Juturu et al., 2025) compared reconstruction-based approaches with methods based on latent space of foundation model.

All of the above approaches work on patch level, such that each patches is classified as anomalous or a specific tissue injury type. In case of testicular abnormalities, creating tiles loses context of the tubulular structure making the approach unfit for injury detection.

2.2. Multiple Instance Learning

Multiple-instance learning (MIL) has become the dominant computational modelling approach for weakly supervised histopathology as it allows to directly learn slide-level labels from WSI, without needing patch or pixel level labels that are expensive to obtain. A feature encoder is used to produce patch embeddings which are then pooled to produce a bag score for the entire WSI. (Ilse et al., 2018) presented a attention-based MIL that learns a self-attention weighted instance embeddings as representation for the WSI.

The encoder choice strongly affects MIL performance, a histology-tailored pretraining yields superior embeddings that are more sensitive to subtle morphological changes and less sensitive to stain variation(Wölflein et al., 2024). (Shao et al.) and (Wölflein et al., 2024) compared various publicly available foundational models for patch feature extraction and MIL aggregation architectures, on diverse tasks, validating the utility and robustness

of MIL approach for WSI feature representation learning. Both of the works show that ABMIL(Ilse et al., 2018) outperforms aggregation methods.

2.3. Whole Slide Foundation Models

Recently, numerous works have proposed slide-level pre-training to learn WSI representation, allowing transferability to low weakly supervised dataset regimes. These methods can be divided in to two sub-types - unimodal(Chen et al., 2022; Lazard et al., 2023; Xu et al., 2024; Lenz et al., 2025; Shao et al.) and multi-model (Jaume et al., 2024b; Shaikovski et al., 2024; Wang et al., 2024).

Early works (Chen et al., 2022; Lazard et al., 2023) demonstrated that MIL trained with self-supervision can extract useful representations of WSI for down-stream tasks. (Chen et al., 2022) exploit the pyramid structure of WSIs to learn representations that capture both fine-grained morphology and coarse spatial context, by a hierarchical training using self-distillation loss. (Lazard et al., 2023) adapted self-supervised contrastive loss to gigapixel images by creating multiples representations of the same WSI. (Xu et al., 2024) adopted transformer with dilated attention (Ding et al., 2023) as slide encoder to handle massive sequence of images patches from a WSI and generate contextualized embeddings. Other methods (Wang et al., 2024; Shao et al.) employ supervised pre-training to learning aggregation layers that can be used as slide foundational models.

3. Dataset

The dataset used for this work consists of WSIs containing testes tissue from Wistar rats. A total of 648 WSIs were available, of which 148 WSIs were used to create the supervised annotated dataset, whereas the remaining 500 were used for self-supervised training. The following subsections describe the dataset preparation in detail.

3.1. Tubule Data Preparation

This work models testes tubules as the fundamental building blocks of the tissue instead of using equally sized patches. To enable this, we train a segmentation U-Net model for tubule detection, which is used to extract tubule structures from the WSIs at 10× magnification. Each extracted tubule is oriented along the diagonal; that is, we consider each tubule as an ellipse and align its major axis with the diagonal of an imaginary box. This ensures geometric consistency. The region around the tubule is padded with white background using a pixel value of 250, which is similar to the background pixel values on the slide.

3.1.1. Supervised Tubular Classification Data

The supervised dataset is a collection of tubules of varying sizes, sampled and annotated by a panel pathologist from 148 WSI. It consists of a total of 10,880 tubules, assigned to one of four classes — Normal, Germ Cell Degeneration, Tubular Degeneration, and Tubular Dilation — the latter three being common injury types. The Normal class has 5,599 samples, Germ Cell Degeneration has 1,893, Tubular Degeneration has 1,584, and Tubular Dilation has 1,804.

The tubules are split 70:30 for training and testing, stratified by the number of samples in each class. We also ensure that the distribution of tubule height and width is consistent across the train and test sets. The training set is further split 75:25 into training and validation subsets. All the results are reported on test dataset.

3.1.2. Unsupervised data

From the 648 WSI, 500 are used to generate two datasets: **1) Unsupervised tubular dataset**, consisting of about 400,000 tubules including both normal and injury classes. This dataset is used to pre-train the MIL model using self-supervision to learn tubular representations; **2) Unsupervised patch dataset**, consisting of about 12 million tissue patches sampled at 5×, 10×, and 20× magnification, used for training a foundational model for patch-level feature extraction.

3.1.3. Classifier vs Multiple Instance Learning

The initial modelling of testicular tubules is based on a naive approach to resize all tubules to a fixed size, close to the median. The resized tubules image are used for training and evaluating the model. Figure 2 provides results for ConvNextV2(Woo et al., 2023) and ViT(Dosovitskiy, 2020) models, across different model capacities. ConvNext models perform better than ViT, which can be attributed to limited availablility of training data. ConvNext-Tiny performs the best achieving 90.37% balanced accuracy.

We train the best performing model (ConvNextV2-Tiny) with different input image sizes - 512, 768, 1024, and bin the results into categories based on actual tubules size, as seen in figure 3. The tubule of size greater than 1280 would be re-sized to 512 for the ConvNext-512 model, and so on. It can be observed that the performance of all models remain similar or degrade for large sized tubules. This can be due to the models' inability to model larger contextual information. Transformer models, that have better ability to build context over the entire image, are limited by the amount of training data.

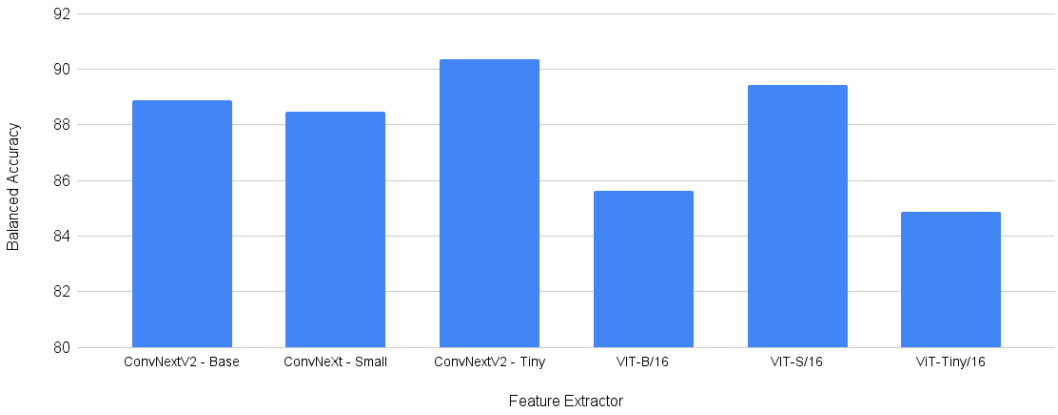

Figure 2: Performance of ConvNext(Woo et al., 2023) and ViT(Dosovitskiy, 2020) model for tubule injury classification

Since the tubule resizing approach gives poor performance on large sized tubules, a modelling approach that doesnt require resizing is needed. We hypothesize that each tubule can be modelled as a bag of patches, following the multiple instance learning (Ilse et al., 2018; Shao et al.). For this we use UNIV2(Chen et al., 2024) as the feature extractor and ABMIL for feature aggregation.

As seen in figure 3, MIL model significantly improves the performance for all tubules sizes. We believe that this is due to a multitude of advantages 1) MIL approach does not

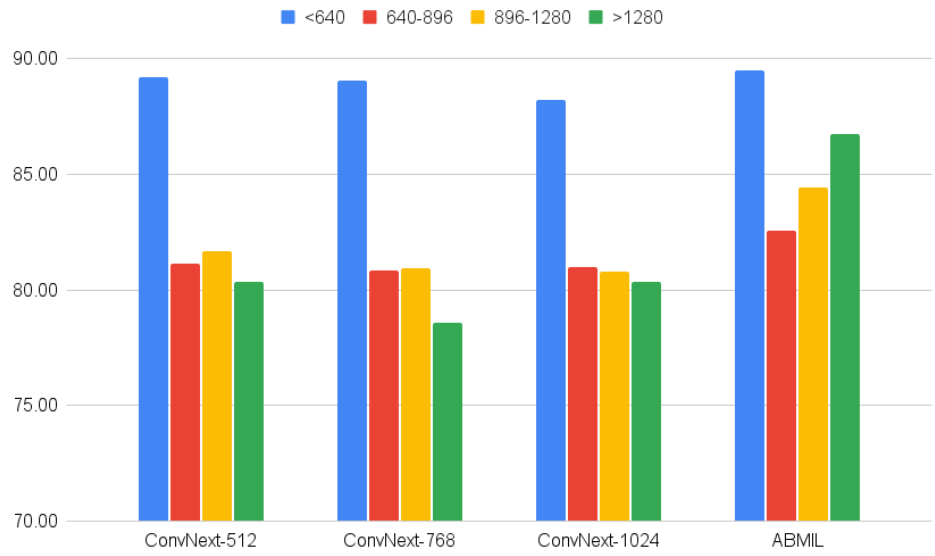

Figure 3: Analysis on performance of different model on tubules binned by their size. All convnext models are ConvNext-Tiny, and the number represents size of the input image used for training and evalution

reduce the spatial dimension of the tubule images, rather creates tiles which are passed through to feature extractor followed by feature aggregation. This would allow better attention to finer morphological features of the tubule 2) the aggregation layers using self-attention that can better extract a global context 3) better features obtained from pre-trained extractor trained on large scale histopathology data.

3.2. Implementation Details

All the supervised models were executed on **two NVIDIA A100 GPUs**, the use of multiple GPUs was leveraged for efficient distributed training and increased effective batch size. We experiment with learning rates [1e-3, 1e-6] depending upon the training type and models and report the best metrics accordingly. Due to the data imbalance between Normal class and other classes we used Weighted Cross-Entropy loss for better performance. We use an effective batch size 32 with gradient accumulation for all training runs, which allows training even for large image sizes. Mixed Precision was utilized for all runs. We employ the AdamW optimizer with a standard Cosine Learning Rate Scheduler. We use a mix of geometric (rotation, flipping) and color (Brightness Contrast, HSV jitter, grayscaling) augmentations during the training phase. All results are reported as average of three runs.

4. Methodology

Based on the evidence obtained in section 3.1.3, we adopt the Multiple Instance Learning (MIL) paradigm, defining a single tubule ROI as a variable-length bag $X = \{x_1, x_2, ..., x_N\}$, where x_i is a constituent patch. Our objective is to learn a system that processes this dynamic collection of patches to capture the collective structural integrity necessary for

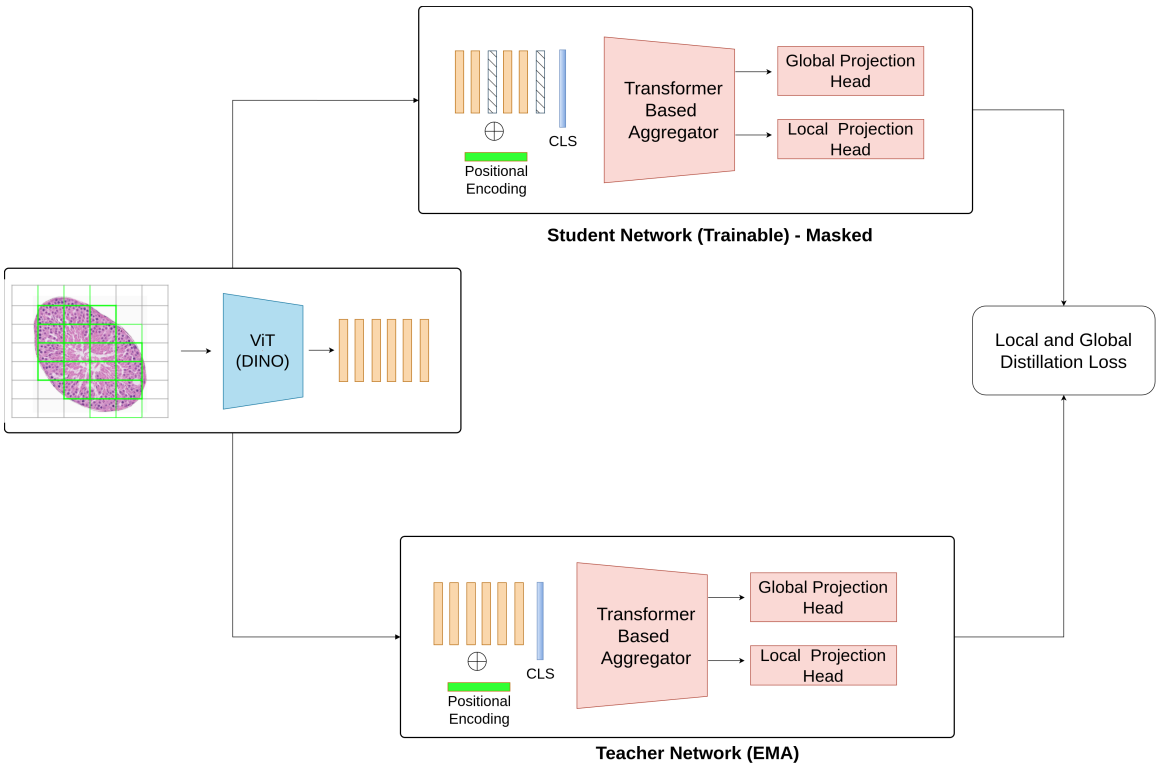

Figure 4: Overview of the proposed framework - Representation Learning for Testicular Tubules. Each tubule is decomposed into a variable-length bag of patch instances and encoded using a foundational ViT feature extractor (Testes-SSL). The resulting features are aggregated using the transformer-based TBA-MIL model with learnable positional embeddings that capture the underlying tubule structure. The aggregation layers are pretrained using Masked Instance Modelling (MIM-MIL), self-supervised learning strategy, that employs student-teacher distillation to learn contextual and morphology-aware representations from large-scale unlabeled tubular data.

accurate abnormality classification. Figure 4 provides an overview of the proposed representation learning for testicular tubules.

4.1. Transformer Based Aggregation: TBA-MIL

Feature Encoder Testes-SSL: A Vision Transformer (ViT-Small), trained using DINO (Caron et al., 2021) self-supervised learning, extracts high-dimensional feature embeddings for each patch x_i.

Structure-Aware Learnable Positional Embeddings To capture the essential radial context of the tubule, we incorporate learnable positional embeddings P. A fixed bank of embeddings $P \in \mathbb{R}^{L_{max} \times D}$ is learned, where $L_{max} = 25$ based on the largest tubule size. This learned encoding is added to the feature embedding z_i of each patch before it enters the Transformer Aggregator:

$$z_i' = z_i + P_i$$

For bags shorter than L_{max}, only the first N embeddings are utilized, imparting necessary spatial awareness.

Patch Aggregator: A Transformer Encoder with 4 layers, 8 attention heads and embedding dimension of size 384 is used to processes the sequence of patch features and the `[CLS]` token.

4.2. Masked Instance Modelling (MIM-MIL)

We propose a self-supervised framework based for pre-training MIL model using knowledge distillation and Masked Instance Modelling (MIM), inspired by previous work on patch level foundational models (Zhou et al., 2021; He et al., 2022). The setup uses two identical networks: a Student (S) and a Teacher (T), both sharing the same architecture as described in section. Additionally, a projection head is added to both teacher and student.

Projection Heads: The output features from the Aggregator are passed to two distinct, linear projection heads that map the high-dimensional features to the prototype space $\mathbb{R}^{D_{\text{proto}}}$ (where D_{proto} is the dimension of the prototypes, e.g., 1024):

- **Global Head (h_g):** Processes the `[CLS]` token (the bag representation) for the global loss $\mathcal{L}_{global}$.
- **Local Head (h_l):** Processes the patch tokens for the local loss $\mathcal{L}_{local}$.

4.2.1. Student-Teacher Distillation

The Teacher network (θ_t) provides stable targets for the Student (θ_s). The Teacher's weights are updated as an Exponential Moving Average (EMA) of the Student's weights, ensuring stability during training:

$$\theta_t \leftarrow \lambda\theta_t + (1 - \lambda)\theta_s$$

where λ follows a cosine schedule, typically starting at 0.996.

4.3. MIM-MIL : Foundational Pre-training using Masked Instance Modelling

Our core self-supervised objective is Masked Instance Modelling, which forces the Student to learn the rules of tissue organization by prediction.

4.3.1. Stochastic Instance Masking

We apply a stochastic binary mask $M \in \{0,1\}^N$ to the bag X. The Student receives a corrupted view $\tilde{X}$, where the patches are masked randomly with mask ratio ranging from 0 to 0.3 and replaced by a learnable `[MASK]` token e_{mask}. The Teacher processes the original, uncorrupted bag X.

4.3.2. Semantic Distillation Loss

We utilize a loss to match the student's prediction to the teacher's semantic assignment (prototypes), with the Teacher acting as an online tokenizer. The total loss ($\mathcal{L}$) combines a Global Loss (CLS token) and a Local Loss (masked patches):

$$\mathcal{L} = \mathcal{L}_{global} + \frac{1}{2}(\mathcal{L}^1_{local} + \mathcal{L}^2_{local})$$

The critical Local Loss ($\mathcal{L}_{local}$) minimizes the Cross-Entropy between the Student's predicted distribution (p_s) and the Teacher's sharpened distribution (p_t), calculated only on the masked tokens:

$$\mathcal{L}_{local} = -\sum_{i \in \text{Masked}} p_t(x_i)^{\tau_t} \cdot \log p_s(\tilde{x}_i)^{\tau_s}$$

where τ_t and τ_s are the temperature parameters. This Cross-Entropy formulation ensures the model learns to identify structural components, optimizing the latent space for downstream classification of abnormalities.

4.4. Implementation Details

The MIM-MIL pre-training was performed on four NVIDIA A100 GPUs, with a learning rate of 1e-4 and batch size of 128, using AdamW optimizer with a standard Cosine Learning Rate Scheduler. Mixed Precision was used to enahance the memory efficiency.

Supervised Fine-tuning: The pre-trained features are used to initialize the final classification model, where the CLS token is passed through a linear layer and classified using standard supervised techniques. All results are reported as average of three runs.

MIL Model	Feature Extractor	Pre-Training	Balanced Accuracy
ABMIL(Ilse et al., 2018)	UNIV2(Chen et al., 2024)	None	91.27
ABMIL(Ilse et al., 2018)	UNIV2	Feather(Shao et al.)	91.74
DFTD(Zhang et al., 2022)	UNIV2	None	84.38
DSMIL(Li et al., 2021)	UNIV2	None	91.73
TRANSMIL(Shao et al., 2021)	UNIV2	None	91.71
TBA-MIL	UNIV2	None	91.28
TBA-MIL	Testes-SSL	None	92.19
TBA-MIL	**Testes-SSL**	**MIM-MIL**	**94.64**

Table 1: The table compares the balanced accuracy for tubular injury classification for various MIL models.

5. Results and Discussion

We investigate the performance of state-of-the-art MIL models for supervised classication of tubular injuries, including ABMIL(Ilse et al., 2018), DFTD(Zhang et al., 2022), DSMIL(Li

et al., 2021), TRANSMIL(Shao et al., 2021) and our proposed TBA-MIL. We also, compare the impact of self-supervised pretraining of ABMIL model using Feather(Shao et al.) and our proposed MIM-MIL. Table 4.4 provides the results. TBA-MIL outperforms all MIL models, using Testes-SSL as the feature extractor and transform based aggregation layers, even without using MIM-MIL pre-training. The MIL pre-training further increases the performance, learning from a large amount of unlabelled tubular data. Due to data imbalance, we use Balanced Accuracy as our comparison metric across all classes.

We also perform an ablation on MIM-MIL pre-training strategy to evaluate the importance of masking and positional embedding. As seen in table 5, both techniques aid in learning better feature representations. Testes tubules have radial structure, as seen in figure 1, which can explain the utility of positional embedding as this allows the MIL model to localize the patches. On the other hand masking helps learning diverse features by compressing redundant visual patterns and thereby enforcing global tissue understanding.

Masking	Pos Embedding	Feature Extractor	Balanced Accuracy
Yes	No	Testes-SSL	90.6
No	Yes	Testes-SSL	91.19
Yes	**Yes**	**Testes-SSL**	**94.64**

Table 2: Ablation for use of Masked Instance Modelling and Positional Embedding in MIM-MIL pre-training

6. Conclusion

This work shows that fixed-size tubule classification is insufficient for modelling the diverse morphology of rat testes tubules. By treating each tubule as a variable-length bag of patches, MIL provides a more effective representation that preserves spatial structure. Our proposed TBA-MIL model further enhances this by incorporating transformer-based aggregation and positional embeddings tailored to tubule geometry. With MIM-MIL pre-training, the model learns structure-aware representations from unlabelled data, yielding substantial performance gains. Together, these contributions set a new benchmark for testicular injury classification and offer a framework for modelling complex, non-uniform tissue structures in digital histopathology.

References

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16144–16155, 2022.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*, 2023.

Jonas Dippel, Niklas Prenißl, Julius Hense, Philipp Liznerski, Tobias Winterhoff, Simon Schallenberg, Marius Kloft, Oliver Buchstab, David Horst, Maximilian Alber, et al. Ai-based anomaly detection for clinical-grade histopathological diagnostics. *Nejm Ai*, 1(11):AIoa2400468, 2024.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

Guillaume Jaume, Simone de Brot, Andrew H Song, Drew FK Williamson, Lukas Oldenburg, Andrew Zhang, Richard J Chen, Javier Asin, Sohvi Blatter, Martina Dettwiler, et al. Deep learning-based modeling for preclinical drug safety assessment. *bioRxiv*, 2024a.

Guillaume Jaume, Anurag Vaidya, Andrew Zhang, Andrew H. Song, Richard J. Chen, Sharifa Sahai, Dandan Mo, Emilio Madrigal, Long Phi Le, and Faisal Mahmood. Multistain pretraining for slide representation learning in pathology. In *European Conference on Computer Vision*, pages 19–37. Springer, 2024b.

Saketh Juturu, Geetank Raipuria, Raghav Amaravadi, Aman Srivastava, Malini Roy, and Nitin Singhal. Unsupervised cellular anomaly detection in toxicological histopathology. In *Medical Imaging with Deep Learning*, 2025.

Tristan Lazard, Marvin Lerousseau, Etienne Decencière, and Thomas Walter. Giga-ssl: Self-supervised learning for gigapixel images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, 2023.

Tim Lenz, Peter Neidlinger, Marta Ligero, Georg Wölflein, Marko van Treeck, and Jakob N Kather. Unsupervised foundation model-agnostic slide-level representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30807–30817, 2025.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

Jasper Linmans, Gabriel Raya, Jeroen van der Laak, and Geert Litjens. Diffusion models for out-of-distribution detection in digital pathology. *Medical Image Analysis*, 93:103088, 2024.

Milda Pocevičiūtė, Yifan Ding, Ruben Bromée, and Gabriel Eilertsen. Out-of-distribution detection in digital pathology: Do foundation models bring the end to reconstruction-based approaches? *Computers in Biology and Medicine*, 184:109327, 2025.

George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D Kunz, Juan A Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024.

Daniel Shao, Richard J Chen, Andrew H Song, Joel Runevic, Ming Y Lu, Tong Ding, and Faisal Mahmood. Do multiple instance learning models transfer? In *Forty-second International Conference on Machine Learning*.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024.

Georg Wölflein, Dyke Ferber, Asier Rabasco Meneghetti, Omar SM El Nahhas, Daniel Truhn, Zunamys I Carrero, David J Harrison, Ognjen Arandjelović, and Jakob Nikolas Kather. A good feature extractor is all you need for weakly supervised pathology slide classification. In *European Conference on Computer Vision*, pages 68–87. Springer, 2024.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024.

Philip Zehnder, Jeffrey Feng, Reina N Fuji, Ruth Sullivan, and Fangyao Hu. Multiscale generative model using regularized skip-connections and perceptual loss for anomaly detection in toxicologic histopathology. *Journal of Pathology Informatics*, 13:100102, 2022.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18802–18812, 2022.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Igor Zingman, Birgit Stierstorfer, Charlotte Lempp, and Fabian Heinemann. Learning image representations for anomaly detection: application to discovery of histological alterations in drug development. *Medical Image Analysis*, 92:103067, 2024.

Appendix A. Sizes Distribution of Testes Tubules

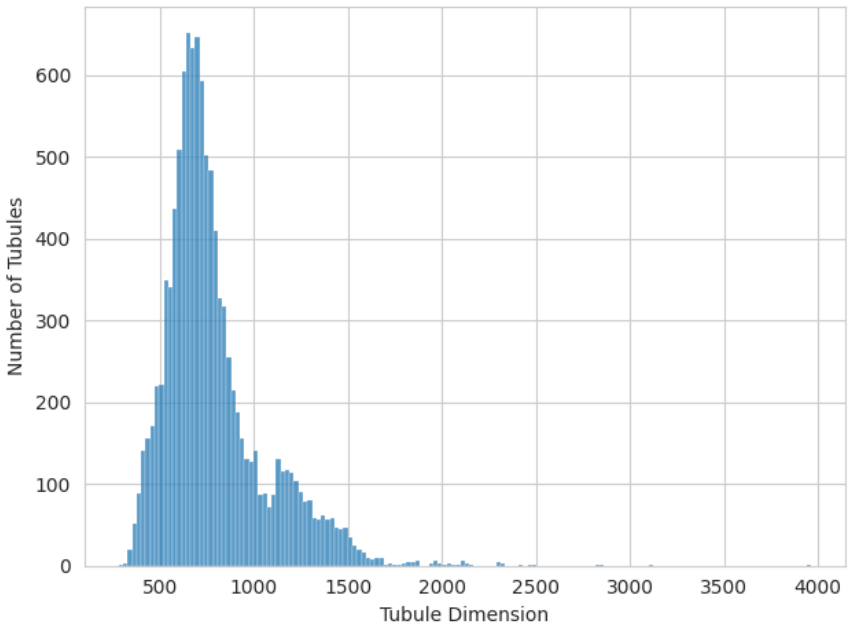

Figure 5: Histogram for tubule sizes from train and test dataset. Median tubule size is 718 pixels, at 10x magnification.