Going Beyond H&E and Oncology: How Do Histopathology Foundation Models Perform for Multi-stain IHC and Immunology?

Amaya Gallagher-Syed **Elena Pontarini** **Myles J. Lewis** **Michael R. Barnes**

Gregory Slabaugh
Queen Mary University of London
a.r.syed, e.pontarini, myles.lewis, m.r.barnes, g.slabaugh@qmul.ac.uk

Abstract

This study evaluates the generalisation capabilities of state-of-the-art histopathology foundation models on out-of-distribution (OOD) multi-stain autoimmune Immunohistochemistry (IHC) datasets. We compare 13 feature extractor models, including ImageNet-pretrained networks, and histopathology foundation models trained on both public and proprietary data, on Rheumatoid Arthritis (RA) subtyping and Sjogren's Disease (SD) diagnostic tasks. Using a simple Attention-Based Multiple Instance Learning classifier, we assess the transferability of learned representations from cancer H&E images to autoimmune IHC images. Contrary to expectations, histopathology-pretrained models did not significantly outperform ImageNet-pretrained models. Furthermore, there was evidence of both autoimmune feature misinterpretation and biased feature importance. Our findings highlight the challenges in transferring knowledge from cancer to autoimmune histopathology and emphasise the need for careful evaluation of AI models across diverse histopathological tasks. The code to run this benchmark is available at https://github.com/AmayaGS/ImmunoHistoBench

1 Introduction

Recent advancements in digital pathology have led to the development of powerful foundation models trained on large-scale H&E cancer datasets [2, 18, 26, 24, 6, 11, 22]. These models have shown remarkable performance in various cancer-related tasks [1, 14, 6, 2, 26, 18]. However, their ability to generalise to other histopathological domains, particularly immunohistochemistry (IHC) staining and autoimmune diseases, remains largely unexplored. This gap is particularly significant given the fundamental differences between cancer and autoimmune pathologies. Indeed, cancer and autoimmune diseases represent two ends of the immune spectrum, with distinct histopathological features reflecting their underlying pathogenesis [19]. In cancer, the immune response is often suppressed or evaded, leading to uncontrolled growth of malignant cells [15]. Conversely, autoimmune diseases are characterised by an overactive immune response against self-tissues [16]. These fundamental differences manifest in contrasting histopathological patterns, aspects of which we exemplify in Figure 1.

1.1 H&E and IHC provide complementary, yet contrasting information

H&E staining, a traditional and widely used technique, offers a broad view of tissue architecture and cellular morphology. Hematoxylin stains cell nuclei a deep blue-purple, while eosin stains cytoplasm and extracellular matrix in shades of pink (see Fig.1 D). In contrast, IHC is a more

Workshop on Advancements In Medical Foundation Models at NeurIPS 2024.

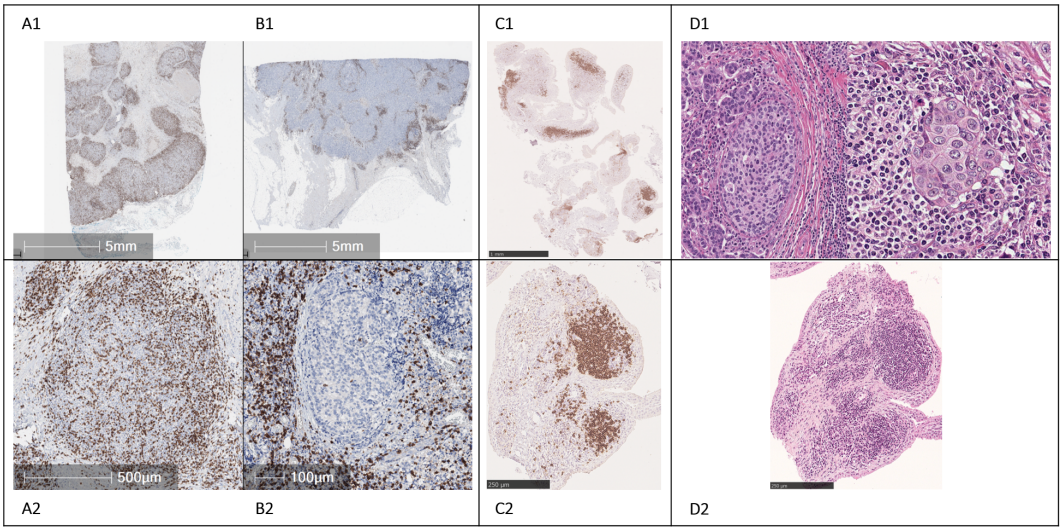

Figure 1: Differences in histological patterns between cancer and autoimmune pathologies. **A-C**: IHC staining, **D**: H&E staining. **A**: Breast cancer tumours (A1), heavily infiltrated with immune cells throughout (**A2**), **B**: Breast cancer tumour (**B1**), characterised by TMEs (**B2**). Tumours with immune infiltrate show better prognosis. **C**: Tissue with strong presence of immune infiltrates in autoimmune disease (**C1**) such as ELS (**C2**), which can correlate to poor prognosis. Note **A2** shows good prognosis, while **C2** shows poor. **D**: differences in H&E staining between cancer (**D1**) and autoimmune (**D2**). In **D1**, we see examples of TMEs surrounded by more darkly stained immune cells unable to infiltrate the tumour. Cancer cells show enlarged cytoplasm to nuclear ratio. **D2** shows darkly stained immune cells.

specialised technique that uses antibodies tagged with visual markers to identify specific proteins or antigens within tissue samples, allowing for precise localisation and visualisation of immune cell population (see Fig.1 A-C). In cancer diagnostics, H&E staining remains the foundation for initial assessment and general diagnosis. However, IHC plays a crucial role in tumour classification, prognosis determination, and treatment selection by pinpointing specific cancer markers and assessing treatment responsiveness indicators. For autoimmune diseases, while H&E staining identifies general patterns of inflammation and tissue damage, IHC becomes essential for a more nuanced understanding of the disease process. It highlights the types of immune cells present in inflammatory infiltrates, detects autoantibody deposits, and visualises specific autoantigens targeted by the immune system.

1.2 Cancer and Autoimmunity: two sides of the immune coin

The immune cell composition and organisation differ significantly between cancer and autoimmune conditions, reflecting their distinct pathogenic mechanisms. In H&E-stained cancer samples, one typically observes striking alterations in cellular and tissue architecture. Cancer cells exhibit uncontrolled proliferation, leading to disorganised growth patterns and invasion into surrounding tissues. These cells often display abnormal morphology, including enlarged nuclei with irregular shapes and altered nuclear-to-cytoplasmic ratios (Fig.1 D1). Furthermore, the immune landscape in cancers forming solid tumours is often characterized by an immunosuppressive tumour microenvironment (TME) that dampens the immune response, allowing cancer cells to evade destruction (Fig.1 A-B). Conversely, autoimmune pathology is characterized by dense inflammatory cell infiltrates, such as Ectopic Lymphoid Structure (ELS), composed of immune cells such as lymphocytes, plasma cells, and macrophages (Fig.1 C1/2) [15, 17, 19, 5]. As shown in Figure 1, these marked differences are reflected in often opposing patterns, with patterns which can indicate a good prognosis in cancer, often indicating a poor one in autoimmune pathologies.

1.3 Why does autoimmunity matter?

These contrasting histopathological patterns reflect the underlying immunological processes at play in cancer and autoimmune diseases, providing valuable insights into their pathogenesis and

potential therapeutic approaches. The intersection of autoimmunity and cancer research offers a rich landscape for knowledge discovery and therapeutic innovation. While these conditions represent opposite ends of the immune dysfunction spectrum, insights from one field often inform the other [15]. For instance, the study of immune checkpoints, originally investigated in autoimmunity, has revolutionised cancer immunotherapy [20]. The presence of ELS structures, typically associated with autoimmunity, has also been linked to improved prognosis in some cancers, highlighting the complex role of immune organisation in disease outcomes [17, 27]. In turn, understanding how cancers evade immune surveillance provides clues about regulating overactive immune responses in autoimmune diseases. Importantly, knowledge gained from treating autoimmune conditions helps manage immune-related adverse events in cancer immunotherapy. This bidirectional flow of insights not only deepens our understanding of immune system dynamics but also paves the way for more nuanced and effective treatments in both cancer and autoimmune diseases. In the decades to come, novel treatment combinations and newly identified druggable targets will only expand the role of immunotherapy in the treatment of cancer and vice versa [23].

1.4 Challenges for H&E generalisation

The distinct pathogenic mechanisms, immune responses, and tissue alterations in cancer versus autoimmune diseases result in characteristically different histopathological patterns that reflect their underlying biology. Histopathology foundation models trained primarily on cancer-related H&E data might face certain challenges when applied to autoimmune conditions and IHC analysis. We outline some potential pitfalls:

1. **Misinterpretation of unique autoimmune features**: these models may fail to recognise or correctly interpret distinctive autoimmune patterns, such as ELSs or specific inflammatory infiltrates, leading to potential misdiagnosis.
2. **Difficulty with complex IHC staining patterns**: The diverse and intricate staining patterns in IHC, especially in autoimmune diseases, could prove challenging for models trained mainly on H&E images, resulting in inaccurate interpretation of immune cell interactions and signalling molecules.
3. **Bias in feature importance**: These models might inappropriately recognise cancer-related features (e.g., nuclear atypia, TMEs) when analysing autoimmune conditions, potentially skewing diagnosis or disease severity assessment.

1.5 Contribution

This study aims to assess the generalisation capabilities of state-of-the-art histopathology foundation models on out-of-distribution (OOD) multi-stain autoimmune datasets. We compare their performance against traditional ImageNet pretrained feature extraction models using simple Attention-Based Multiple Instance Learning algorithm (ABMIL) [10]. This straightforward, zero-shot approach allows us to directly evaluate the transferability of learned representations from cancer H&E images to autoimmune IHC images without the confounding effects of complex downstream tasks. It serves as an initial step to understanding if cancer-trained models can capture relevant features in autoimmune tissues, despite differences in disease processes and staining techniques.

By bridging the gap between cancer and autoimmune histopathology, this report aims to contribute to the broader goal of developing more comprehensive and versatile AI tools for medical image analysis, ultimately supporting advancements in precision medicine and improved patient care across a wider spectrum of immune-mediated diseases.

2 Methodology

2.1 Pipeline

Figure 2 illustrates our analytical pipeline. The input comprises a set of multi-stain Whole Slide Images (WSIs) per patient, including H&E and various IHC stains listed in Table 1. We apply adaptive thresholding to identify tissue areas and extract 224x224 pixel non-overlapping patches. These patches are then processed through a feature extractor network, generating a matrix of feature

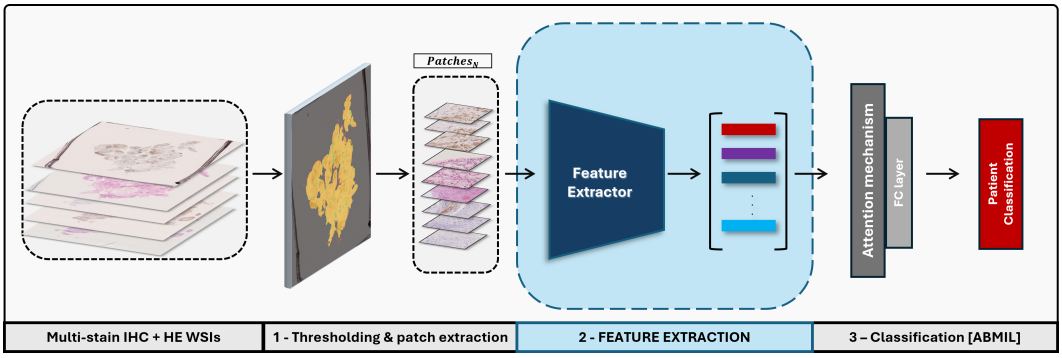

Figure 2: Schematic representation of the multi-stain whole slide image (WSI) analysis pipeline. The process encompasses (1) tissue thresholding and patch extraction, (2) feature extraction yielding patient-level feature matrices, and (3) attention-based multiple instance learning (ABMIL) classification. This study primarily focuses on comparing feature extraction methodologies.

vectors representing each patient's WSI set. For downstream classification, we employ the Attention-Based Multiple Instance Learning (ABMIL) [10] algorithm, chosen for its simplicity and strong benchmark performance. Notably, this study concentrates on comparing different feature extraction techniques within this framework.

2.2 Datasets

To provide a benchmark on autoimmune multi-stain datasets, we use two clinical datasets. One dataset derives from the clinical trial R4RA [9], where patients with difficult to treat Rheumatoid Arthritis (RA) were recruited for treatment with rituximab [13]. The other derives from the routine diagnostic of patients presenting with dry eyes and mouth (sicca) and investigated for Sjogren's Disease (SD). Each dataset is composed of H&E slides, with approximately 3 IHC slides of different immune biomarkers per patient. In Table 1, we give further information on the stains present in each dataset.

Table 1: Metadata and dataset characteristics for Sjogren's Disease and Rheumatoid Arthritis cohorts, including number of patients, WSIs, stains present and average number of stains per patient. We highlight in pink H&E staining and blue IHC.

	Sjogren		**Rheumatoid Arthritis**	
No. Patients	93		153	
No. Slides	347		607	
No. Stains	5		4	
Av. Stains per patient	3.7		3.97	
Magnification	20x		10x	
Total no. patches	237k		275k	
Av. Patches per patient	2 530		1800	
Patches per stain	**Mean**	**Total**	**Mean**	**Total**
HE	650	61055	434	66511
CD3	625	58712	0	0
CD138	377	35416	481	73624
CD20	626	58805	351	53768
CD21	254	23843	0	0
CD68	0	0	535	81915
ML problem type	**Diagnosis**		**Subtyping**	
Labels	Negative	46	Low inflammatory	66
	Positive	47	High inflammatory	87

2.2.1 Rheumatoid Arthritis

This clinical trial dataset comprises 607 Whole Slide Images (WSIs) from 153 Rheumatoid Arthritis patients, categorised into low (N=66) and high (N=87) inflammatory subtypes [9]. As exemplified in

Figure 3, samples were stained with H&E and the IHC markers CD20+ B cells, CD68+ macrophages and CD138+ macrophages. The dataset features a variable number of stains, averaging 3.9 per patient. We perform binary classification on low (N=66) and high (N=87) inflammatory subtypes. We extract non-overlapping patches at a 10x magnification, keeping those with over 40% tissue coverage, totalling approximately 275k patches.

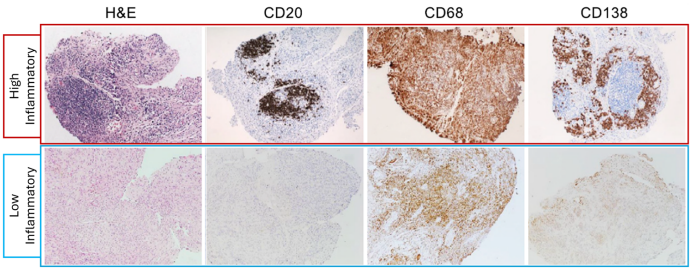

Figure 3: Rheumatoid Arthritis inflammatory pathotypes based on semi-quantitative analysis of synovial tissue biopsies stained with H&E, CD20+ B cells, CD68+ macrophages and IHC+ CD138 plasma cells [9, 7].

2.2.2 Sjogren

Diagnostic dataset consisting of 347 WSIs of labial salivary gland biopsies sampled from 93 patients, with 46 cases of non-specific sicca and 47 cases of Sjogren's Disease (SD). As shown in Figure 4 the samples were stained with H&E and the IHC stains CD20+ B cells, CD3+ T cells, CD21+ follicular dendritic cell network and CD138+ plasma cells. Each patient has a variable set of multi-stain WSIs, averaging 3.7 stains per patient. We perform binary classification on the detection of sicca vs SD+ biopsies. We extract non-overlapping patches at a 20x magnification, keeping those with over 30% tissue coverage, totalling approximately 237k patches.

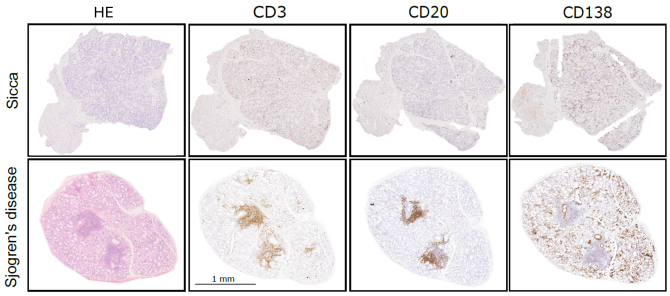

Figure 4: Example of sicca vs Sjogren's Disease presentation in H&E and IHC stains. On top, a patient diagnosed with sicca, on bottom a patient diagnosed with Sjogren's disease. Here we show samples stained with IHC stains CD3+ T cells, CD20+ B cells and CD138+ plasma cells.

2.3 Feature Extraction Models

We evaluate a comprehensive and diverse set of 13 feature extractor models to assess their efficacy on Autoimmune IHC datasets. As detailed in Table 2.3, these models span three distinct categories: 5 general computer vision models pretrained on ImageNet [21, 8, 25, 4], 5 histopathology-specific foundation models trained on publicly available data SSL_ResNet18 [3], SSL_ResNet50 [11], CTransPath [24], Lunit [11] and Phikon-v2 [6], as well as three of the latest, largest models trained on proprietary data UNI, H-Optimus-0 and GigaPath [2, 18, 26]. This selection enables us to systematically compare the performance of generic visual features against domain-specific representations learned from histopathological data. Our analysis encompasses both Convolutional Neural Network (CNN) architectures and Self-Attention (SA) based models, including state-of-the-art Vision Transformers. The models vary significantly in scale, ranging from 11 million to 1.1 billion parameters, and in their pretraining data, from generic image datasets to specialised collections of up to 1.3 billion histopathology image patches.

Model	Operation type	Backbone Architecture	Parameters	Training type	Training Data Origin	Tissue types	Magnifications	WSIs (patches)	Stains
VGG16 [21]	CNN	-	138M	Supervised	ImageNet-1k	-	-	-	-
ResNet18 [8]	CNN	-	11M	Supervised	ImageNet-1k	-	-	-	-
ResNet50 [8]	CNN	-	23M	Supervised	ImageNet-1k	-	-	-	-
ConvNeXt [25]	CNN	-	88M	Supervised	ImageNet-1k	-	-	-	-
ViT [4]	SA	-	86M	Supervised	ImageNet-21k/1k	-	-	-	-
SSL_ResNet18 [3]	CNN	ResNet18	11.1M	SimCLR	TCGA, CPTAC, Multiple Public	Cancer Normal	Multiple	25 000 (400K)	H&E IHC
SSL_ResNet50 [11]	CNN	ResNet50	23.5M	Barlow Twins	TCGA/Internal	Cancer	20/40x	36 666 (32M)	H&E
CTransPath [24]	CNN-SA	SwinT	27.5M	MoCO-v3	TCGA/PAIP	Cancer Normal	20x	32 220 (15M)	H&E
Lunit [11]	SA	ViT-S	21.6M	DINO	TCGA/Internal	Cancer	20/40x	36 666 (32M)	H&E
Phikon-v2 [6]	SA	ViT-L	307M	DINOv2	TCGA, CPTAC, GTeX, Multiple Public	Cancer Normal	20x	58 359 (456M)	H&E
H-Optimus-0 [18]	SA	ViT-G	1.1B	DINOv2	Internal	Cancer	20x	500 000	H&E
UNI [2]	SA	ViT-L	307M	DINOv2	Internal/GTeX	Cancer Normal	20x	100 426 (100M)	H&E
GigaPath [26]	SA	ViT-G	1.1B	DINOv2	Internal	Cancer	20x	171 189 (1.3B)	H&E

Table 2: Comparative analysis of feature extractor models for digital pathology, contrasting ImageNet-pretrained models with histopathology-specific foundation models. Models are categorised into three groups: (orange) general computer vision models pretrained solely on ImageNet; (blue) histopathology foundation models pretained on publicly available datasets; and (green) large-scale histopathology foundation models trained on proprietary datasets. This classification facilitates comparison between generic visual features and domain-specific representations. The table details backbone architecture, CNN vs Self-Attention (SA), parameter number, training paradigms and the origin, tissue and staining type of training data.

2.4 Implementation Details

Feature Extraction. In total we extract 26 patient level feature representations for the RA and SD datasets, obtained from each of the 13 feature extraction networks with frozen weights. We maintain the embedding dimension of each feature extraction network. See Appendix 5 for further details.

Experimental Setup and Evaluation Metrics. We separate a random label stratified 20% hold out test set and perform 5-fold random label stratified cross validation on the remaining data (train:val:test / 60:20:20). Models were trained for a maximum 200 epochs, with a patience set to 15 such that early stopping was called if no change was observed in either the loss, accuracy or AUC score for 15 epochs. Weights were kept for the model obtaining the best accuracy score on each validation set while ensuring there was no under-fitting or over-fitting of the models. Each of the 5 trained model was applied to the hold-out test. We report the mean and standard error (SE) of the results obtained on the hold-out test set for Accuracy, Macro F1-score, Precision, Recall, AUC and Average Precision.

ABMIL. For classification we use the original one-layer ABMIL implementation [10], with input embeddings projected to a hidden layer of dimensions 128 and sigmoid activation function.

Training schedule. All models were trained using cross-entropy loss, with the AdamW optimizer set to $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$, with a learning rate $1e^{-5}$ and weight decay $L_2 = 0.01$. No learning scheduler was used. We conducted an initial search on learning rate ($1e^{-2}$, $1e^{-3}$, $1e^{-4}$, $1e^{-5}$, settling for $1e^{-5}$ as we observed the smoothest loss on average across models and datasets. Training was conducted on an NVidia A100 GPU (40Gb) [12].

3 Results

3.1 Rheumatoid Arthritis

In Figure 5 and Table 3, we show the results of our comparative analysis for RA, with scores organised from most to least performant within each type of model (here labelled as ImageNet, TCGA and Internal for convenience). For each metric, the top three performers are highlighted with a red box.

The top three performers are SSL_ResNet18 [3], UNI [2] and VGG16 [21] across Accuracy, AUC and Average Precision, however there are no significant differences between the three. Overall, there are no substantial performance gain between ImageNet and histopathology-pretrained networks in general, contrary to findings from cancer-based benchmarking studies [1, 14, 6, 2].

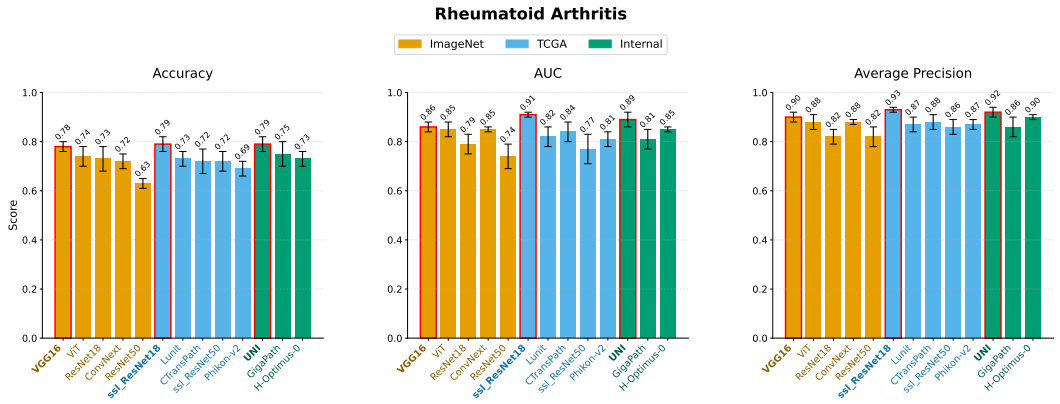

Figure 5: Performance comparison of feature extractor models on Rheumatoid Arthritis Subtyping. ImageNet pretrained models are shown in orange, Histopathology pretrained models in blue for models trained on publicly available datasets and green for proprietary data .Within each category and metric models are organised from most to least performant, with the top three performers highlighted with a red box.

Table 3: Detailed performance metrics for various feature extraction models on Rheumatoid Arthritis classification tasks. Results include Accuracy, Macro F1-score, Precision, Recall, AUC, and Average Precision for each model, with standard errors in parentheses. Models are grouped into ImageNet-pretrained (orange), publicly available histopathology-pretrained (blue), and large-scale proprietary histopathology-pretrained (green) categories. Top performing models highlighted in a gradient of blue.

	Rheumatoid Arthritis					
	Accuracy	**Macro F1-score**	**Precision**	**Recall**	**AUC**	**Average Precision**
VGG16	0.78 (0.02)	0.77 (0.02)	0.79 (0.02)	0.76 (0.01)	0.86 (0.02)	0.90 (0.02)
ResNet18	0.73 (0.05)	0.68 (0.07)	0.73 (0.05)	0.69 (0.06)	0.79 (0.04)	0.82 (0.03)
ResNet50	0.63 (0.02)	0.54 (0.05)	0.58 (0.07)	0.58 (0.03)	0.74 (0.05)	0.82 (0.04)
ConvNext	0.72 (0.03)	0.65 (0.06)	0.80 (0.02)	0.67 (0.04)	0.85 (0.01)	0.88 (0.01)
ViT	0.74 (0.04)	0.69 (0.06)	0.76 (0.05)	0.70 (0.05)	0.85 (0.03)	0.88 (0.03)
ssl_ResNet18	0.79 (0.03)	0.76 (0.04)	0.86 (0.01)	0.76 (0.04)	0.91 (0.01)	0.93 (0.01)
ssl_ResNet50	0.72 (0.04)	0.66 (0.07)	0.67 (0.09)	0.69 (0.05)	0.77 (0.06)	0.86 (0.03)
CTransPath	0.72 (0.05)	0.67 (0.07)	0.77 (0.05)	0.69 (0.05)	0.84 (0.04)	0.88 (0.03)
Lunit	0.73 (0.03)	0.71 (0.04)	0.73 (0.04)	0.71 (0.04)	0.82 (0.04)	0.87 (0.03)
Phikon-v2	0.69 (0.03)	0.65 (0.05)	0.69 (0.03)	0.67 (0.04)	0.81 (0.03)	0.87 (0.02)
H-Optimus-0	0.73 (0.03)	0.69 (0.04)	0.77 (0.04)	0.69 (0.03)	0.85 (0.01)	0.90 (0.01)
UNI	0.79 (0.03)	0.78 (0.03)	0.80 (0.02)	0.79 (0.03)	0.89 (0.03)	0.92 (0.02)
GigaPath	0.75 (0.05)	0.71 (0.06)	0.77 (0.05)	0.73 (0.06)	0.81 (0.04)	0.86 (0.04)

3.2 Sjogren's Disease

For Sjogren's Disease, Gigapath [26], ConvNext [25] and CTransPath [24] are the top three overall performers. However, as with RA we observe a similar trend of marginal differences between models, with considerable overlap in error bars and no significant differences between top performers for ImageNet vs histopathology pretrained models indicating the learned feature representation extracted from histopathology foundation models are not conferring real downstream benefits for diagnosis or subtyping of autoimmune diseases.

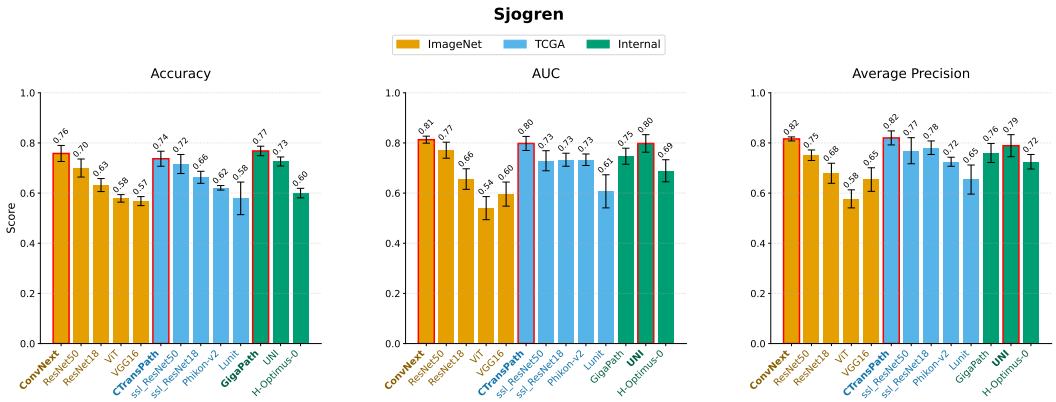

Figure 6: Performance comparison of feature extractor models on Rheumatoid Arthritis Subtyping and Sjogren's Disease Diagnostic tasks. ImageNet pretrained models are shown in orange, Histopathology pretrained models in blue for models trained on publicly available datasets and green for proprietary data.Within each category and metric models are organised from most to least performant, with the top three performers highlighted with a red box.

Table 4: Detailed performance metrics for various feature extraction models on Sjogren's Disease classification tasks. Results include Accuracy, Macro F1-score, Precision, Recall, AUC, and Average Precision for each model, with standard errors in parentheses. Models are grouped into ImageNet-pretrained (orange), publicly available histopathology-pretrained (blue), and large-scale proprietary histopathology-pretrained (green) categories. Top performing models highlighted in a gradient of blue.

	Sjogren					
	Accuracy	**Macro F1-score**	**Precision**	**Recall**	**AUC**	**Average Precision**
VGG16	0.57 (0.02)	0.57 (0.02)	0.57 (0.02)	0.57 (0.02)	0.60 (0.05)	0.65 (0.05)
ResNet18	0.63 (0.03)	0.62 (0.03)	0.63 (0.03)	0.63 (0.03)	0.66 (0.04)	0.68 (0.04)
ResNet50	0.70 (0.03)	0.68 (0.04)	0.71 (0.03)	0.69 (0.04)	0.77 (0.03)	0.82 (0.02)
ConvNext	0.76 (0.03)	0.75 (0.04)	0.78 (0.03)	0.75 (0.03)	0.81 (0.01)	0.82 (0.01)
ViT	0.58 (0.01)	0.55 (0.03)	0.63 (0.04)	0.58 (0.02)	0.54 (0.05)	0.58 (0.04)
ssl_ResNet18	0.66 (0.02)	0.66 (0.02)	0.67 (0.02)	0.66 (0.02)	0.73 (0.03)	0.78 (0.03)
ssl_ResNet50	0.72 (0.04)	0.71 (0.04)	0.72 (0.04)	0.72 (0.04)	0.73 (0.04)	0.77 (0.05)
CTransPath	0.74 (0.03)	0.73 (0.03)	0.76 (0.03)	0.74 (0.03)	0.80 (0.03)	0.82 (0.03)
Lunit	0.58 (0.06)	0.57 (0.07)	0.57 (0.07)	0.58 (0.07)	0.61 (0.07)	0.65 (0.06)
Phikon-v2	0.62 (0.01)	0.62 (0.01)	0.64 (0.02)	0.62 (0.01)	0.73 (0.02)	0.72 (0.02)
H-Optimus-0	0.60 (0.02)	0.59 (0.02)	0.62 (0.02)	0.61 (0.02)	0.69 (0.04)	0.73 (0.03)
UNI	0.73 (0.02)	0.72 (0.02)	0.74 (0.02)	0.73 (0.02)	0.80 (0.03)	0.79 (0.04)
GigaPath	0.77 (0.02)	0.77 (0.02)	0.77 (0.02)	0.77 (0.02)	0.75 (0.03)	0.76 (0.04)

3.3 Attention heatmaps highlight misinterpretation and bias in feature importance

In Figure 7 we show an example of a SD+ set of multi-stain WSIs (H&E, CD3 and CD21), with attention heatmaps obtained from the ABMIL models trained with features extracted from ConvNext (left) and GigaPath (right), the two top performer models for SD. Observing these heatmaps we see clear evidence of the pitfalls mentioned in the introduction: in all three WSIs we see inflammatory aggregates present throughout the tissue being explicitly less attended to by the GigaPath foundation model. This is particularly clear in the H&E stained slide, were for GigaPath we see inflammatory aggregates appear as darker areas of low attention scores in the middle of highly attented areas. In contrast, ConvNext has a more evenly spread out attention map covering areas of immune cell infiltration. In CD3 and CD21, we again see this pattern of more spread out attention covering areas with immune cells in ConvNexT, whereas GigaPath is picking up strongly on areas which appear to bear high level similarities to TMEs, but which are not structures present in autoimmune disease. This is particularly striking in the top 5 patches with highest and lowest attentions scores plotted below each WSIs: GigaPath consistently assigns low attention scores to areas with immune cell populations and high attention scores to tissue with morphologies which bear some resemblance

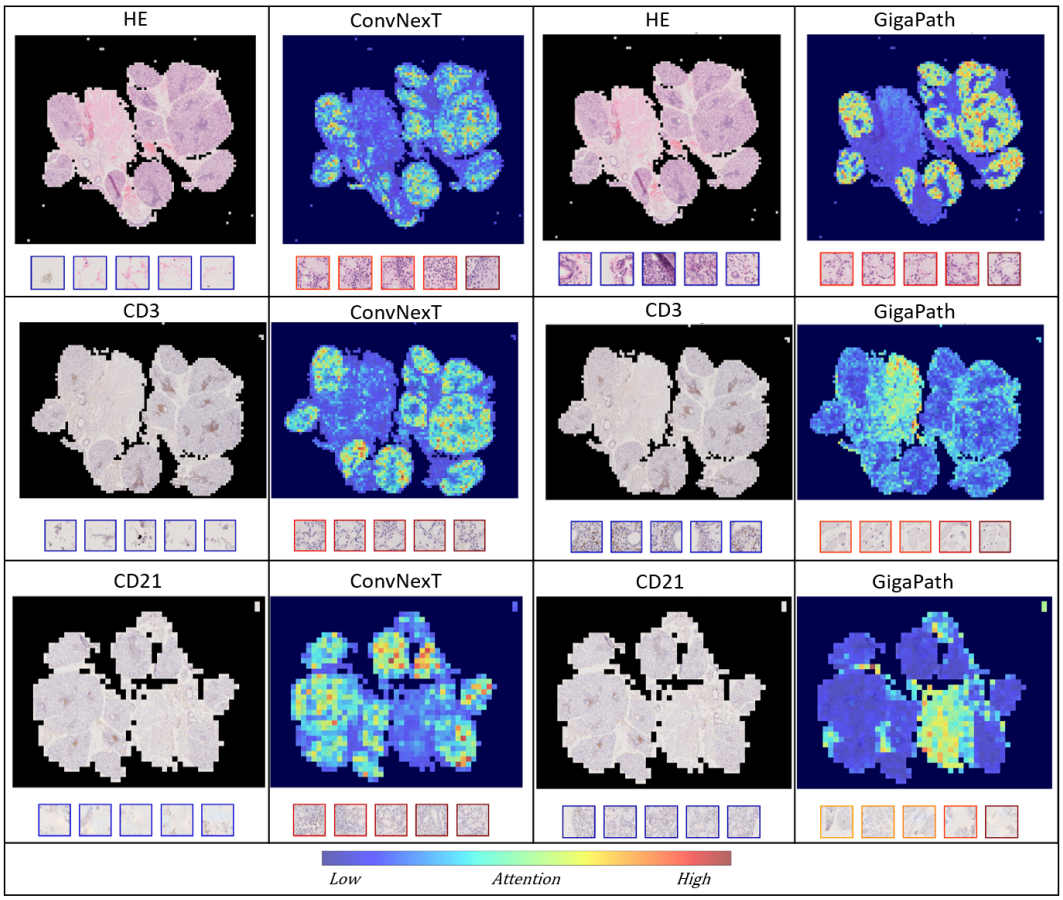

Figure 7: Attention heatmaps obtained from the ABMIL models trained with features extracted from ConvNext (left) and GigaPath (right) for a multi-stain (H&E, CD3 and CD21) set of SD+ WSIs. These results highlight bias in feature importance, with immune cell populations being less attended to by the foundation model, as well as misinterpretation of features, with the foundation model concentrating instead on areas which bear a high level resemblance to TMEs and abnormal cancer cells. See Appendix 8 9 10 for higher resolution images

to abnormal cancer cells, while ConvNexT more accurately assigns low attention scores to areas with little tissue coverage or with WSIs artefacts present and higher attention scores to areas with immune cell populations. We further explore attention heatmaps for UNI and CTransPath, as well as show higher resolution images in appendix 8 9 10. Notably, both UNI and CTransPath show similar attention patterns, with attention scores which are less spread out and which do focus on immune cell populations. This indicates GigaPath is focusing on more cancer specific patterns, in line with its very large and highly specialised training corpus. For researchers working on IHC and non-cancer datasets wishing to use a histopathology foundation model, UNI and CTransPath might therefore currently represent a good compromise choice.

3.4 Conclusion

Our findings highlight the complexity of transferring knowledge from cancer histopathology to the autoimmune context and underscores the need for careful evaluation of model performance across diverse histopathological tasks, as this will be key for use in clinical practice. We believe future work should focus on incorporating a broader scope of immune-mediated diseases into histopathological foundation models, and handling complex IHC staining patterns to develop more versatile and accurate AI tools for a broader spectrum of diseases.

By understanding how the immune system behaves in these different contexts, we can gain deeper insights into immune evasion, surveillance, and modulation. This knowledge not only bridges the gap between cancer biology and immunology but also opens new avenues for disease prevention, diagnosis, and treatment in both fields. Future AI models in digital pathology should be designed to capture and interpret these complex immunological features across different disease states, enhancing our ability to leverage insights from both autoimmunity and cancer research.

Acknowledgments and Disclosure of Funding

We wish to thank Dr. Dovile Zilenaite for her insightful comments, knowledge and help in making Figure 1. A.G.S. receives funding from the Wellcome Trust [218584/Z/19/Z].

References

[1] Jack Breen et al. "Histopathology Foundation Models Enable Accurate Ovarian Cancer Subtype Classification". In: *arXiv preprint arXiv:2405.09990* (2024).

[2] Richard J. Chen et al. "Towards a general-purpose foundation model for computational pathology". en. In: *Nature Medicine* 30.3 (Mar. 2024), pp. 850–862. ISSN: 1546-170X. DOI: 10.1038/s41591-024-02857-3. URL: https://www.nature.com/articles/s41591-024-02857-3 (visited on 09/26/2024).

[3] Ozan Ciga, Tony Xu, and Anne L. Martel. *Self supervised contrastive learning for digital histopathology*. Tech. rep. arXiv:2011.13971 [cs, eess] type: article. arXiv, Sept. 2021. DOI: 10.48550/arXiv.2011.13971. URL: http://arxiv.org/abs/2011.13971 (visited on 09/26/2024).

[4] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. Tech. rep. arXiv:2010.11929 [cs] type: article. arXiv, June 2021. DOI: 10.48550/arXiv.2010.11929. URL: http://arxiv.org/abs/2010.11929 (visited on 09/26/2024).

[5] Zeev Elkoshi. "Cancer and Autoimmune Diseases: A Tale of Two Immunological Opposites?" In: *Frontiers in Immunology* 13 (Jan. 2022), p. 821598. ISSN: 1664-3224. DOI: 10.3389/fimmu.2022.821598. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8822211/ (visited on 09/22/2024).

[6] Alexandre Filiot et al. *Phikon-v2, A large and public feature extractor for biomarker prediction*. Tech. rep. arXiv:2409.09173 [cs, eess] type: article. arXiv, Sept. 2024. DOI: 10.48550/arXiv.2409.09173. URL: http://arxiv.org/abs/2409.09173 (visited on 09/26/2024).

[7] Amaya Gallagher-Syed et al. "Multi-Stain Self-Attention Graph Multiple Instance Learning Pipeline for Histopathology Whole Slide Images". In: *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. URL: https://papers.bmvc2023.org/0789.pdf.

[8] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Tech. rep. arXiv:1512.03385 [cs] type: article. arXiv, Dec. 2015. DOI: 10.48550/arXiv.1512.03385. URL: http://arxiv.org/abs/1512.03385 (visited on 09/26/2024).

[9] Frances Humby et al. "Rituximab versus tocilizumab in anti-TNF inadequate responder patients with rheumatoid arthritis (R4RA): 16-week outcomes of a stratified, biopsy-driven, multicentre, open-label, phase 4 randomised controlled trial". English. In: *The Lancet* 397.10271 (Jan. 2021), pp. 305–317. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(20)32341-2. URL: https://www.thelancet.com/journaltitles/lancet/article/PIIS0140-6736(20)32341-2/fulltext (visited on 04/05/2023).

[10] M. Ilse, J. Tomczak, and M. Welling. "Attention-based Deep Multiple Instance Learning". In: *Proc. 35th ICML*. Vol. 80. 2018, pp. 2127–2136. URL: https://proceedings.mlr.press/v80/ilse18a.html.

[11] Mingu Kang et al. "Benchmarking Self-Supervised Learning on Diverse Pathology Datasets". en. In: 2023, pp. 3344–3354. URL: https://openaccess.thecvf.com/content/CVPR2023/html/Kang_Benchmarking_Self-Supervised_Learning_on_Diverse_Pathology_Datasets_CVPR_2023_paper.html (visited on 09/26/2024).

[12] Thomas King, Simon Butcher, and Lukasz Zalewski. "Apocrita - High Performance Computing Cluster for Queen Mary University of London". In: (Mar. 2017). URL: https://zenodo.org/record/438045 (visited on 05/11/2023).

[13] Myles J. Lewis et al. "Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment Response Phenotypes". en. In: *Cell Reports* 28.9 (Aug. 2019), 2455–2470.e5. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2019.07.091. URL: https://www.sciencedirect.com/science/article/pii/S2211124719310071 (visited on 04/05/2023).

[14] Mayur Mallya et al. "Benchmarking histopathology foundation models for ovarian cancer bevacizumab treatment response prediction from whole slide images". In: *arXiv preprint arXiv:2407.20596* (2024).

[15] Davide Mangani, Dandan Yang, and Ana C. Anderson. "Learning from the nexus of autoimmunity and cancer". English. In: *Immunity* 56.2 (Feb. 2023), pp. 256–271. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2023.01.022. URL: https://www.cell.com/immunity/abstract/S1074-7613(23)00034-1 (visited on 09/22/2024).

[16] David S. Pisetsky. "Pathogenesis of autoimmune disease". en. In: *Nature Reviews Nephrology* 19.8 (Aug. 2023), pp. 509–524. ISSN: 1759-507X. DOI: 10.1038/s41581-023-00720-1. URL: https://www.nature.com/articles/s41581-023-00720-1 (visited on 09/22/2024).

[17] Costantino Pitzalis et al. "Ectopic lymphoid-like structures in infection, cancer and autoimmunity". en. In: *Nature Reviews Immunology* 14.7 (July 2014), pp. 447–462. ISSN: 1474-1741. DOI: 10.1038/nri3700. URL: https://www.nature.com/articles/nri3700 (visited on 09/22/2024).

[18] Charlie Saillard et al. *H-optimus-0*. 2024. URL: https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0.

[19] Justyna Sakowska et al. "Autoimmunity and Cancer—Two Sides of the Same Coin". In: *Frontiers in Immunology* 13 (May 2022), p. 793234. ISSN: 1664-3224. DOI: 10.3389/fimmu.2022.793234. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9140757/ (visited on 09/22/2024).

[20] Padmanee Sharma and James P. Allison. "The future of immune checkpoint therapy". In: *Science* (2015).

[21] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Tech. rep. arXiv:1409.1556 [cs] type: article. arXiv, Apr. 2015. DOI: 10.48550/arXiv.1409.1556. URL: http://arxiv.org/abs/1409.1556 (visited on 09/26/2024).

[22] Eugene Vorontsov et al. "A foundation model for clinical-grade computational pathology and rare cancers detection". In: *Nature Medicine* (2024).

[23] Alex D. Waldman, Jill M. Fritz, and Michael J. Lenardo. "A guide to cancer immunotherapy: from T cell basic science to clinical practice". en. In: *Nature Reviews Immunology* 20.11 (Nov. 2020), pp. 651–668. ISSN: 1474-1741. DOI: 10.1038/s41577-020-0306-5. URL: https://www.nature.com/articles/s41577-020-0306-5 (visited on 09/22/2024).

[24] Xiyue Wang et al. "Transformer-based unsupervised contrastive learning for histopathological image classification". In: *Medical Image Analysis* 81 (Oct. 2022), p. 102559. ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102559. URL: https://www.sciencedirect.com/science/article/pii/S1361841522002043 (visited on 09/26/2024).

[25] Sanghyun Woo et al. *ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders*. Tech. rep. arXiv:2301.00808 [cs] type: article. arXiv, Jan. 2023. DOI: 10.48550/arXiv.2301.00808. URL: http://arxiv.org/abs/2301.00808 (visited on 09/26/2024).

[26] Hanwen Xu et al. "A whole-slide foundation model for digital pathology from real-world data". en. In: *Nature* 630.8015 (June 2024), pp. 181–188. ISSN: 1476-4687. DOI: 10.1038/s41586-024-07441-w. URL: https://www.nature.com/articles/s41586-024-07441-w (visited on 09/26/2024).

[27] Dovile Zilenaite-Petrulaitiene et al. "Intratumoral heterogeneity of Ki67 proliferation index outperforms conventional immunohistochemistry prognostic factors in estrogen receptor-positive HER2-negative breast cancer". In: *Virchows Archiv* (2024), pp. 1–12.

A Supplemental material

A.1 Model feature size & extraction time

Table 5: We show vector feature size and extraction time, as well as peak RAM and VRAM usage for each model on both datasets.

	Feature vector size	Sjogren			RA		
		Extraction time (m)	Peak RAM (GB)	Peak VRAM (GB)	Extraction time (m)	Peak RAM (GB)	Peak VRAM (GB)
VGG16 [21]	4096	39	40	1.7	49	45	1.7
ResNet18 [8]	512	38	40	1.2	43	33	1.2
ResNet50 [8]	2048	59	20	1.2	64	45	1.2
ConvNeXt [25]	1024	99	22	2.7	118	37	2.6
ViT [4]	768	79	45	0.4	87	45	0.4
SSL_ResNet18 [3]	512	39	40	1.2	43	34	1.2
SSL_ResNet50 [11]	2048	57	38	1.2	64	30	1.2
CTransPath [24]	768	88	34	0.3	99	40	0.3
Lunit [11]	384	60	46	0.1	99	40	0.3
Phikon-v2 [6]	1024	156	37	1.2	190	50	1.2
H-Optimus-0 [18]	1536	300	64	4.6	336	50	4.6
UNI [2]	1024	123	43	1.2	150	50	1.2
GigaPath [26]	1536	300	64	4.6	325	52	4.6

A.2 Attention heatmaps

We show higher resolution images for ConvNexT and GigaPath in 8 and 9, as well as further heatmaps for CTransPath and UNI models in 10 and 11.

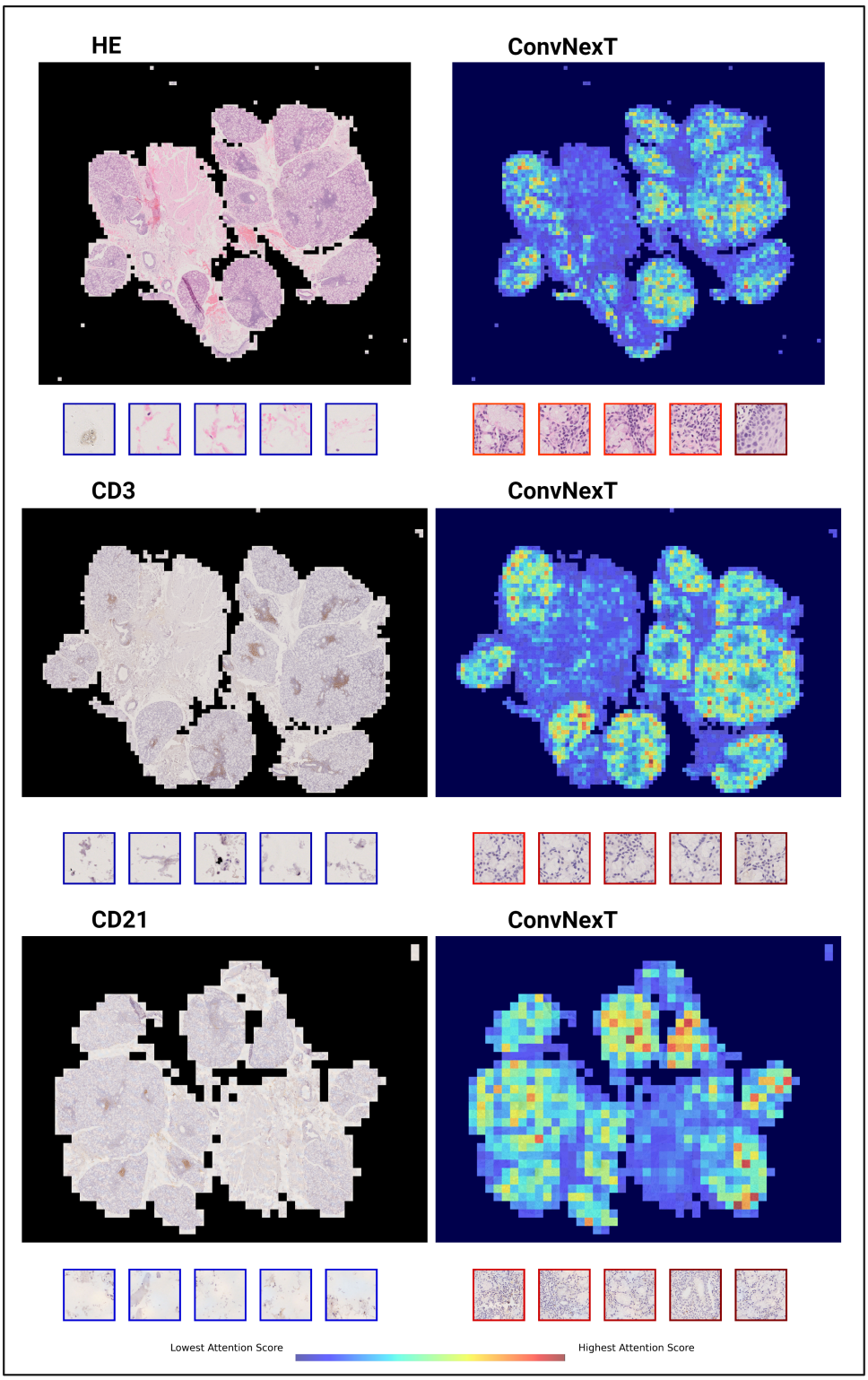

Figure 8: Attention heatmaps obtained from the ABMIL models trained with features extracted from ConvNext for a multi-stain (H&E, CD3 and CD21) set of SD+ WSIs. Overall we see a more spread out attention pattern, covering areas of immune cell populations.

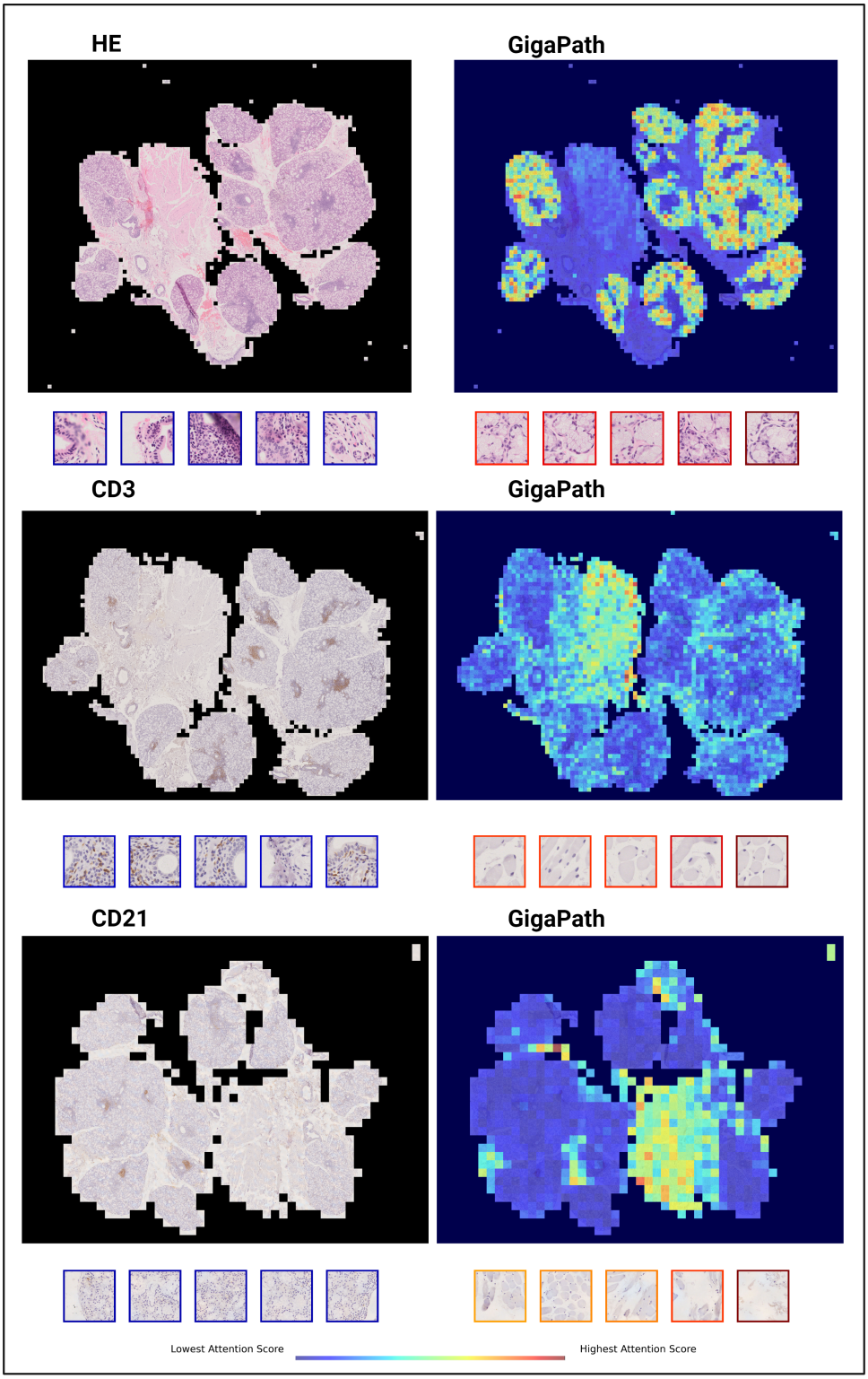

Figure 9: Attention heatmaps obtained from the ABMIL models trained with features extracted from GigaPath for a multi-stain (H&E, CD3 and CD21) set of SD+ WSIs. We notice attention score tend to cluster more tightly around areas with no inflammatory aggregates.

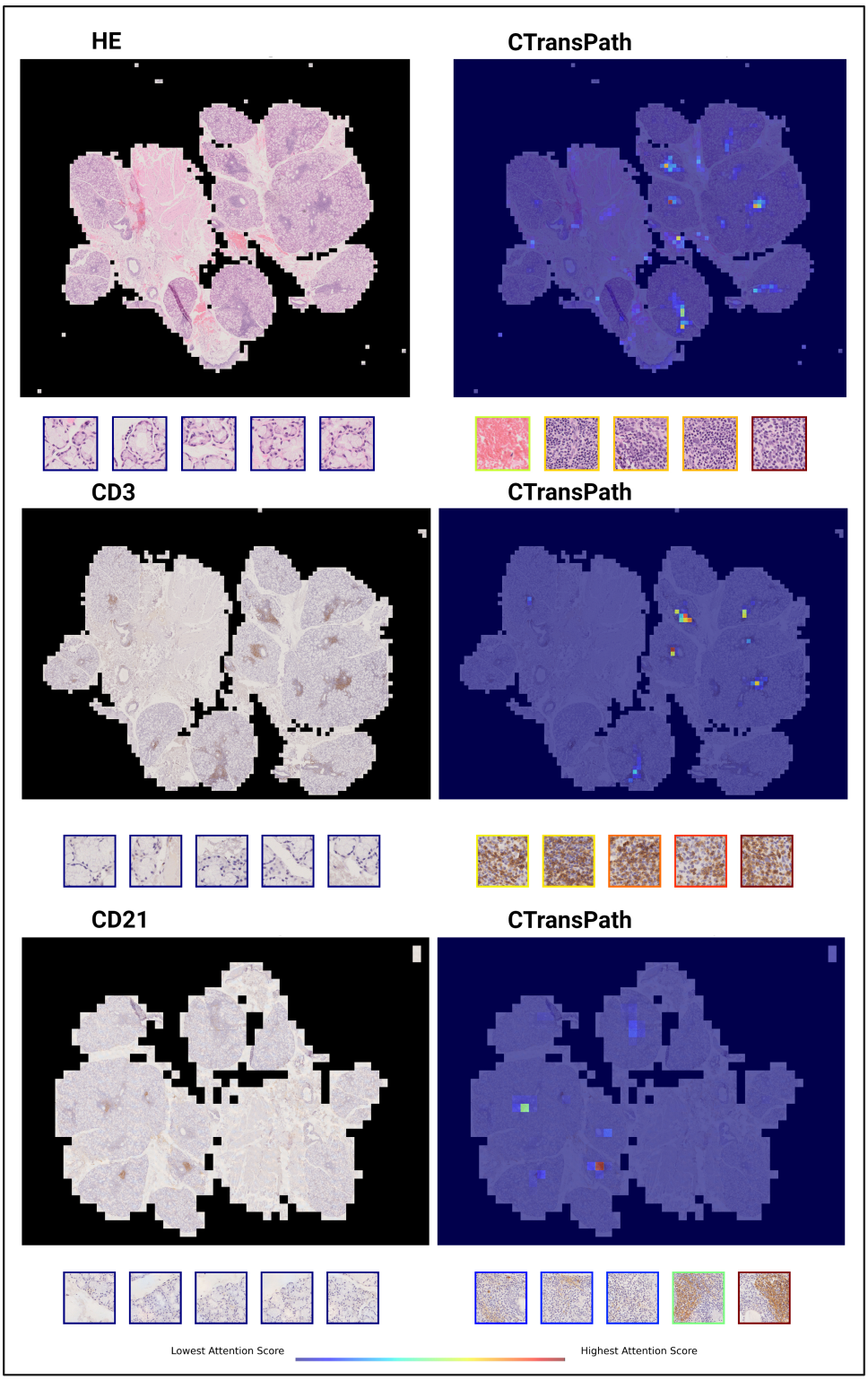

Figure 10: Attention heatmaps obtained from the ABMIL models trained with features extracted from CTransPath for a multi-stain (H&E, CD3 and CD21) set of SD+ WSIs. The attention score pattern is very focused, with emphasis on a few well delimitated immune cell aggregates.

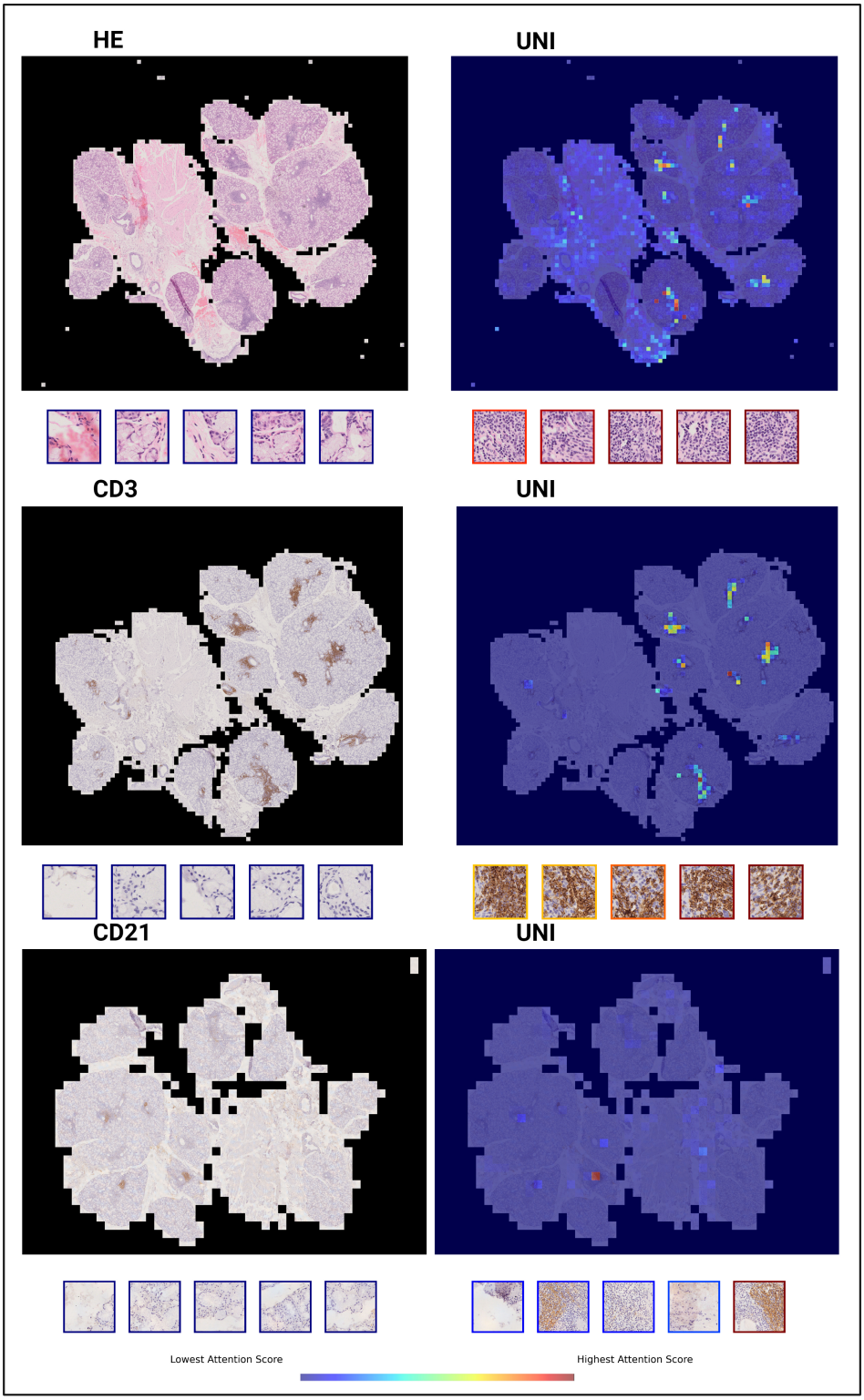

Figure 11: Attention heatmaps obtained from the ABMIL models trained with features extracted from UNI for a multi-stain (H&E, CD3 and CD21) set of SD+ WSIs. Similarly to CTransPath, the attention score pattern is very focused, with emphasis on a few well delimitated immune cell aggregates.