BioX-CPath: Biologically-driven Explainable Diagnostics for Multistain IHC Computational Pathology

Amaya Gallagher-Syed[*†] Henry Senior[*] Omnia Alwazzan[*] Elena Pontarini[*]
Michele Bombardieri[*] Costantino Pitzalis[*] Myles J. Lewis[*] Michael R. Barnes[*]
Luca Rossi[‡] Gregory Slabaugh[*]

Abstract

The development of biologically interpretable and explainable models remains a key challenge in computational pathology, particularly for multistain immunohistochemistry (IHC) analysis. We present BioX-CPath, an explainable graph neural network architecture for whole slide image (WSI) classification that leverages both spatial and semantic features across multiple stains. At its core, BioX-CPath introduces a novel Stain-Aware Attention Pooling (SAAP) module that generates biologically meaningful, stain-aware patient embeddings. Our approach achieves state-of-the-art performance on both Rheumatoid Arthritis and Sjogren's Disease multistain datasets. Beyond performance metrics, BioX-CPath provides interpretable insights through stain attention scores, entropy measures, and stain interaction scores, that permit measuring model alignment with known pathological mechanisms. This biological grounding, combined with strong classification performance, makes BioX-CPath particularly suitable for clinical applications where interpretability is key. Source code and documentation can be found at: https://github.com/AmayaGS/BioX-CPath.

1. Introduction

Whole Slide Image (WSI) scanners capture high resolution, multi-magnification digital images of stained tissue biopsies presented on glass slides. The digitization of these biopsies has spurred the development of computational pathology methods. Analysis of these WSIs currently stands as one of the gold standard diagnostic and subtyping methods for many forms of cancers and autoimmune diseases, such as Rheumatoid Arthritis (RA) and Sjogren's Disease. Different types of staining exist, which highlight different aspects of the tissue samples. Hematoxylin & Eosin (H&E) staining, a traditional and widely used technique, offers a broad view of tissue architecture and cellular morphology, with Hematoxylin staining cell nuclei a deep blue-purple, while Eosin stains cytoplasm and extracellular matrix in shades of pink. In contrast, Immunohistochemistry (IHC) is a more specialized technique that uses antibodies tagged with visual markers to identify specific proteins or cell types within tissue samples, allowing for precise localization and visualization of cell populations present in the tissue [42].

In cancer diagnostics, H&E staining remains the foundation for initial assessment and general diagnosis. However, IHC plays a crucial role in tumor classification, prognosis determination, and treatment selection by pinpointing specific cancer biomarkers, such Human Epidermal Growth Factor Receptor 2 (HER2), Estrogen Receptor (ER) and Progesterone Receptor (PR) [65, 66]. 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 visualizes specific autoantigens targeted by the immune system [44]. In clinical pathology, a tissue sample will be taken and thinly sliced, and different stains applied to these slices, often with a reference H&E slide to verify tissue quality [42]. These multi-stain WSIs stacks are rich in information about cellular types, tissue structures, and spatial patterns which relate to disease presentation and prognosis. Expert pathologists examining these stacks perform a semi-quantitative analysis, efficiently integrating information across both scale, stains, and images.

Most state-of-the-art computational pathology methods so far have focused on H&E and the single-stain domain. Works that has tackled IHC have often done so in the context of cell quantification via cell segmentation [15, 26, 52, 53], as well as prediction and scoring of biomarkers

[*]Queen Mary University of London
[†]Corresponding author: a.r.syed@qmul.ac.uk
[‡]The Hong Kong Polytechnic University
[0]Accepted for publication at CVPR 2025

[26, 27, 48], or registration of multistain stacks [56]. Some works have also explored the potential of H&E to IHC virtual staining techniques [38, 49, 59] or recently of using IHC as views for self-supervised representation learning [26]. Most of these approaches have concentrated on extracting information from IHC slides such that this could be predicted using H&E or on quantification of cell populations in IHC. This is because H&E is an older, more widely available and cost-effective technology. However, the use of IHC and more advanced techniques such as immunofluorescence is only set to grow in the coming years, associated with a decrease in technology cost and more advanced biomarker detection techniques [42]. There is therefore a clear need for methods which explicitly focus on integrating the complex cell landscapes across stains. To the best of our knowledge, few studies have concentrated on the issue of classification of unregistered, unannotated multistain datasets to date, with a single stain graph and mid/late fusion approaches developed in [13] and a multistain attention graph approach proposed in MUSTANG [17]. However, the value of computational pathology extends beyond classification tasks. Given the rich information contained in multi-stain data, these methods must be interpretable and generate actionable biological insights at both the disease and patient level. This interpretability serves two crucial purposes: it advances our understanding of underlying disease mechanisms, and it allows pathologists to verify that model predictions align with established biological knowledge, building trust in the system's outputs.

1.1. Contributions

1. We introduce BioX-CPath, a biologically driven graph-based model tailored to the complex cellular landscapes in multistain datasets. BioX-CPath works across multiple stains using semantic and spatial cues to capture complementary cellular and tissue information.
2. We propose a novel Stain-Aware Attention Pooling (SAAP) module that generates expressive, stain-aware patient embeddings. This module uniquely respects the biological and diagnostic diversity across stains, improving interpretability and diagnostic relevance.
3. We fully leverage the biological interpretability of BioX-CPath via derived metrics: stain attention and entropy scores, stain-stain interaction scores and Graph Neural Networks (GNNs) node heatmaps. These metrics provide detailed insights into stain relevance and inter-stain relationships, uncovering key biological patterns and interactions that contribute to disease pathology.

1.2. Related Work

1.2.1. Multiple Instance Learning

WSIs are gigapixel, heterogeneous image files, which present challenges for computer vision methods given each image can reach over $100k \times 100k$ pixels, generally within a low patient sample setting. A weakly supervised Multiple Instance Learning (MIL) approach is most often employed to address this challenge. The image is divided into a regular grid of smaller patches (e.g., 224×224 pixels), each inheriting slide/patient labels. Patches are then embedded into a feature vector and classified at the slide/patient level using some form of non-trainable (e.g., max or mean) or trainable aggregation on the set of instances. Methods such as ABMIL [25], DS-MIL [36], and CLAM introduced trainable linear attention aggregation layers [39]. TransMIL [51] tackles the issue of long-range dependencies by approximating self-attention operations between patches via the Nyström method [57].

1.2.2. GNNs in Histopathology

Applications in histopathology can be divided into cell, patch, or tissue-level graphs, with both node and graph classification approaches being employed. Patch-graphs can be constructed using features extracted from a WSI or a set of WSIs, then connected via edges [2, 3, 63, 64]. DeepGraphConv [37], PatchGCN [9], GTP [63], CAMIL [16] and HEAT [8] adopt this approach by constructing a graph connecting either the k-nearest neighbors in feature space or region adjacent patches. DeepGraphConv uses spectral graph convolution on a subset of patches, whereas Patch-GCN employs graph convolutional layers with residual connections and a final global attention pooling mechanism layer, which GTP replaces with a Transformer layer. CAMIL [16] combines a spatial neighbor constrained attention module with a transformer layer. HEAT [8] incorporates node and edge attributes in a heterogeneous graph, together with a pseudo-label pooling algorithm based on predicted cell types using KimiaNet, a feature extractor which was pretrained on H&E images from The Cancer Genome Atlas Program (TCGA) [47].

These methods were designed for accurate classification and prognosis in H&E staining and cancer datasets. Because of this, these models are optimized to focus on features and patterns linked to tissue architecture and cell morphology. Notably, because these models were designed for the single-stain cancer domain, they concentrate on spatial awareness which aligns well with the need in cancer to accurately detect tumors and tumor microenvironment based on tissue architecture and cell morphology [8, 9, 34]. Moreover, these approaches provide insight into tumor localization by providing heatmaps overlays showing the attention scores obtained per patch. Other methods, such as TEA-Graph [34] and Slide-Graph [40] also provide insight into interpretable prognostic biomarkers linked to tissue type.

In line with previous work, we adopt a patch-graph approach to efficiently integrate information across multistain WSI stacks. However, although we provide insight into the model decision making process through examination of

layer importance and GNN heatmap, our focus is on providing insight into the alignment of our model with underlying biology and in understanding how the cell populations interact. This approach bridges the gap between performance-based approaches and explainable, insight-driven approaches.

2. Preliminaries

In this section we provide definitions and background on the concepts used throughout this work.

Graph Neural Networks. GNNs are capable of learning representations of graphs by propagating node features through a series of computationally efficient message-passing and aggregation operations [14]. Given a graph over a set of nodes V, during the k-th message-passing iteration, the embedding $\mathbf{h}_u^{(k)}$ corresponding to each node $u \in V$ is updated according to information aggregated from the neighbors of u, i.e.,

$$\begin{aligned}\mathbf{h}_u^{(k+1)} &= \text{UPDATE}^{(k)}\left(\mathbf{h}_u^{(k)}, \mathbf{m}_{\mathcal{N}(u)}^{(k)}\right)\\ \mathbf{m}_{\mathcal{N}(u)}^{(k)} &= \text{AGGREGATE}^{(k)}\left(\left\{\mathbf{h}_v^{(k)}, \forall v \in \mathcal{N}(u)\right\}\right),\end{aligned} \tag{1}$$

where the neighborhood $\mathcal{N}(u)$ is defined as the set of nodes that share an edge with u, UPDATE and AGGREGATE are arbitrary differentiable functions, and $\mathbf{m}_{\mathcal{N}(u)}^{(k)}$ is the "message" that is aggregated from $\mathcal{N}(u)$. At each iteration, the AGGREGATE function takes as input the set of embeddings of the nodes in $\mathcal{N}(u)$ [21]. When each node u of the input graph has an associated d_x-dimensional input feature $\mathbf{x}_u \in \mathbb{R}^{d_x}$, $\mathbf{h}_u^{(0)}$ is set to $\mathbf{x}_u$. As a result, through several message-passing iterations $\mathbf{h}_u^{(k)}$ captures increasingly rich information encapsulating both the topological structure and the features surrounding each graph node u. However, after successive message-passing operation GNNs can suffer from vanishing gradients due to over-smoothing of the signal, leading to increasingly similar node representations [1, 4, 14]. In tasks where long-range interactions between far away nodes are important, this leads to loss of local neighborhood topological information.

Graph Attention Network. Graph Attention Networks (GATs) [55] are a type of GNN which incorporate masked self-attention layers [5, 54] into message-passing and use attention weights to define a weighted sum of the neighbors, i.e.,

$$\mathbf{m}_{\mathcal{N}(u)}^{(k)} = \sum_{v \in \mathcal{N}(u)} \beta_{u,v} \mathbf{h}_v^{(k)}, \tag{2}$$

where $\beta_{u,v}$ denotes the attention on neighbor $v \in \mathcal{N}(u)$ when aggregating information at node u.

Graph pooling. Graph pooling methods aim to downsample graphs while preserving essential structural information. There are two different type of approaches: spectral-based and top-k-based methods [60]. Spectral approaches such as DiffPool [60], LaPool [43] or EigenPool [41] transform the graph into a compressed representation through learned soft clustering assignments, producing new abstract node representations. In contrast, top-k methods [62] such as gPool [20], TopKPool [19] or SAGPool [33] directly identify and preserve the most important nodes through various scoring mechanisms. The resulting scores enable direct node selection, maintaining a clear correspondence between the original and pooled graph, which maintains interpretability by producing a subgraph where node identity is conserved. gPool and TopKPool use a learnable vector to calculate projection scores and select the top-ranked nodes, but do not fully take into account graph topology [7, 33]. SAGPool [33] uses the GCN defined in [29] to calculate the self-attention scores $\mathbf{z} \in \mathbb{R}^{N \times 1}$ as follows:

$$\mathbf{z} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}\boldsymbol{\theta}_{att}\right), \tag{3}$$

where $\tilde{\mathbf{A}} \in \mathbb{R}^{N \times N}$ represents the adjacency matrix with self-connections, $\tilde{\mathbf{D}}$ is its degree matrix, $\mathbf{X} \in \mathbb{R}^{N \times F}$ contains node features, and $\boldsymbol{\theta}_{att} \in \mathbb{R}^{F \times 1}$ are the learnable parameters.

By utilizing graph convolutions to obtain self-attention scores, the result of the pooling is based on both graph and topological features, while remaining efficient to calculate in terms of memory and runtime [33]. The node selection method follows [7, 19, 30] by retaining a portion of nodes of the input graph, even when graphs of varying sizes and structures are input. The pooling ratio $k \in (0, 1]$ hyperparameter determines the number of nodes to keep at each pooling layer.

Graph readouts. Graph readout operations are specifically focused on obtaining a fixed-size graph-level representation by aggregating all node features. This is generally done through simple statistical operators such as global mean and global max pooling operations [58]. However, these basic aggregation procedures cause information loss through oversmoothing of the node signals, failing to capture complex topological relationships encoded into graphs. Recent methods have examined how to obtain more expressive graph readouts through the use of clustering [32], attention [9] or variance[50] based techniques. Notably in the histopathology area HEAT [8] proposed to aggregate based on the assignment of tissue type pseudo-labels. However, approaches based on pseudo-cluster can be inconsistent across graphs [8] and fail to align with meaningful and interpretable biology.

Positional encoding. Random walk positional encoding is a technique used to incorporate structural information from a graph into the node embeddings [14]. Specifically, for each node u in the graph, a random walk of fixed length is performed, starting from that node u and considering only the landing probability of transitioning back to the node u itself at each step, i.e., $\mathbf{p}^u_{\text{RWPE}} = \left[RW_{uu}, RW^2_{uu}, \dots RW^l_{uu}\right]^\top \in \mathbb{R}^l$, where $\mathbf{p}^u_{\text{RWPE}}$ represents the random walk positional encoding for node u, RW^l_{uu} is the l-step landing probability of returning to node u after a random walk of length l starting from u, and the positional encoding concatenates these l-step landing probabilities into a vector in $\mathbb{R}^l$. The node random walk positional encoding is then concatenated with its feature vector to obtain a new enriched input feature, i.e., $\mathbf{h}_u = \mathbf{W}_c \left[\mathbf{x}_u \,\|\, \mathbf{p}^u_{\text{RWPE}}\right]$ where $\mathbf{h}_u \in \mathbb{R}^d$ is the final d-dimensional embedding for node u, $\mathbf{x}_u \in \mathbb{R}^{d_x}$ is the initial d_x-dimensional feature vector for node u, $\mathbf{p}^u_{\text{RWPE}} \in \mathbb{R}^l$ is the l-dimensional random walk positional encoding for node u, $\|$ denotes the vector concatenation operation, and $\mathbf{W}_c \in \mathbb{R}^{d\times(d_x+l)}$ is a learnable weight matrix that projects the concatenated node feature and positional encoding to an d-dimensional embedding space. This allows the node embeddings to capture not only the local neighborhood structure around each node, but also higher-order proximity information between nodes that are multiple hops away, potentially improving their ability to capture complex global patterns and dependencies within the graph structure.

3. Methods

Here we introduce our proposed pipeline, which we illustrate graphically in Fig. 1.

3.1. Preprocessing

Feature extraction. We start by preprocessing each stack of patient multistain WSIs by thresholding tissue areas from background and extracting patches. For each extracted patch, the (x, y)-coordinates are saved. Each patch is then processed by a feature extractor to obtain an embedded feature vector. Here we use the UNI feature encoder [11] as it has shown reasonable performance on IHC benchmarking tasks [18]. This produces a feature matrix $\mathbf{X}_p \in \mathbb{R}^{N\times d}$ which represents the stack of WSIs for a given patient p, with d the embedding dimension of the feature encoder. See SM. E for further details.

Graph initialization. Given our feature matrix $\mathbf{X}_p$, we first construct a k-Nearest Neighbor (k-NN) graph in feature space. This feature space graph G_{FS} contains relationships between semantically similar patches, regardless of their spatial relationship, and has an adjacency matrix denoted $\mathbf{A}_{FS}[i, j]$ where $\mathbf{A}_{FS}[i, j] = 1$ if patch j is among k nearest neighbors of i in feature space. We then construct a region adjacency graph G_{RA} using the extracted (x, y) coordinates, with adjacency matrix $\mathbf{A}_{RA}[i, j]$ where $\mathbf{A}_{RA}[i, j] = 1$ if patch j is among k region adjacent nearest neighbors of i both on the (x, y) plane (same WSIs) and z-axis of the WSIs stack. We illustrate these two types of connectivity in Fig. 1B. We then combine $\max(\mathbf{A}_{FS}, \mathbf{A}_{RA})$ to obtain our full $\mathbf{A}_{FRA} \in \{0, 1\}^{N\times N}$, which we use to initialize our input graph $G_{FRA} = (V, E)$. For each node, we store as a categorical node attribute their stain type S, while for each edge we store the edge type. The combination of feature and spatial proximity was chosen to connect stains across the stack and permit information flow during message passing operations.

Positional encoding. For each node in G_{FRA}, a fixed length random walk is performed [13], starting from a given node and considering only the landing probability of transitioning back to the initial node at each step. The random walk positional encoding vector is appended to the initial feature vector of its associated node and re-appended through each layer of our backbone. We employ this approach to alleviate issues with long-range cross WSIs stack connectivity by providing global topological information to the graph.

3.2. Patient Level Encoding

Hierarchical graph blocks. To obtain patient-level encoding, we use as our backbone a hierarchal graph approach as presented in [7, 60], with the aim of attenuating oversmoothing issues. Our patient level encoder backbone consists of alternating GAT layers [55] and our proposed Stain-Aware Attention Pooling (SAAP) module, which refines the node features whilst selecting the most relevant ones - using an importance score - to be forwarded to the next layer [33]. Finally, we apply multi-head self-attention (MHSA) to the concatenated stain-aware patients encoding returned by the SAAP module at each layer, with the resultant features passed to a fully connected classification head. Our backbone architecture choice is motivated by the desire to obtain the most expressive representation of patient encoding [8, 17, 22, 46]

Stain-Aware Attention Pooling module. The SAAP algorithm, illustrated in Fig. 1, begins with calculating node attention scores $\mathbf{a} \in \mathbb{R}^N$. Here we use the SAGPool algorithm as defined in 2. Briefly, the attention scores are computed as $\mathbf{a} = \text{GNN}(\mathbf{X}, \mathbf{A}_{\mathbf{FRA}})$. These node attention scores represent the importance of each node in the graph based on both their features and graph topology. Both the node attention scores $\mathbf{a}$ and the feature matrix $\mathbf{X}$ are sorted and a subset $\mathbf{X}'$ is selected based on the top k nodes wrt the

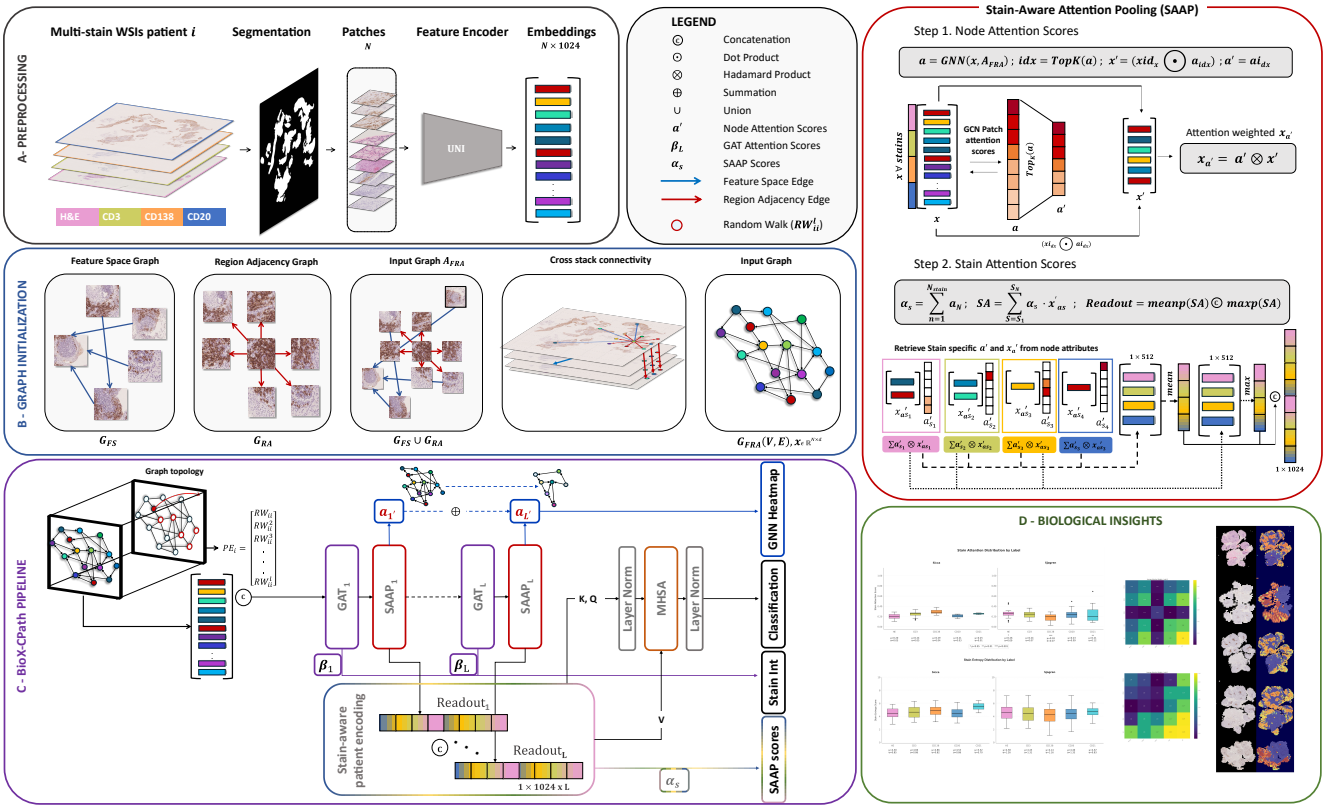

Figure 1. **Architecture**: Our approach begins by preprocessing the WSIs into patch features using UNI [11] (Section A). The resultant features are combined into two graphs, G_{FS} and G_{RA}, representing the feature space similarity and region adjacency respectively. Given that the node sets of the two graphs are shared, we join the edge sets together, yielding graph G_{FRA} (Section B). G_{FRA} is then passed through hierarchical GNN blocks (Section C) consisting of a Graph Attention Network (GAT) [55] and our proposed Stain-Aware Attention Pooling (SAAP) (detailed in the top right), which updates the node features while selecting the most relevant ones using an importance score. We obtain stain-aware patient encoding, which we pass through a final MHSA layer, before classification. Derived from both SAAP and GAT layers we propose metrics which provide biological insights into the model's predictions (Section D).

attention scores, forming the subgraph G'_{FRA}. This preserves the most relevant nodes while reducing the computational complexity. The attention scores of the k nodes are then used to scale the node features $\mathbf{X}'$ through element-wise multiplication. This injects relevance ranking in the feature matrix $\mathbf{X}'$, such that more relevant nodes now have higher weight. For each stain s, a stain-level weight α_s is then calculated as the sum of the normalized attention scores $\mathbf{a}'$ of the N_s nodes belonging to that stain, i.e.,

$$\alpha_s = \sum_{n=1}^{N_s} \mathbf{a}'_n, \tag{4}$$

The algorithm then pools weighted features by stain. Stain attention (SA) scores are then calculated as

$$\text{SA scores} = \sum_{s \in S} \alpha_s \cdot \mathbf{X}'_s \tag{5}$$

where S is the set of stains, α_s is the attention weight for stain s and $\mathbf{X}'_s$ represents the features specific to stain s. Finally, we obtain stain-aware readouts Readout $=$ $[\text{meanp(SA)} \| \text{maxp(SA)}]$ where meanp and maxp represent mean pooling and max pooling operations respectively, and $\|$ represents the vector concatenation operation. SAAP explicitly handles multiple stain modalities by computing stain-specific weights (α_s), allowing the model to learn the relative importance of different stains for downstream tasks. With this we aim to maximize expressiveness, while aligning it with relevant biological information.

Biological insight. Based on our SAAP module and the proposed backbone architecture, we introduce a number of derived metrics which allow us to verify if the model aligns with known biology and can help provide clinical insights. These metrics are:

- **SAAP scores**, defined above. This score informs us on which stains were most diagnostically relevant for the downstream task.
- **Stain entropy scores**, $H_s = -\sum_{n=1}^{N_s}(\mathbf{a}'_n \cdot \log(\mathbf{a}'_n))$

where H_s is the entropy for stain s and $\mathbf{a}'_n$ are the normalized attention scores of the N_s nodes belonging to stain s. This measures how uniformly distributed the attention scores are within each stain type, with lower entropy values indicating more concentrated, focused attention patterns aligning with organized, localized cellular structures, while higher entropy represents uniformly distributed attention corresponding to diffuse, disorganized cellular structures present throughout the tissue.

- **Stain-stain interaction scores**, $\mathbf{I}$ are defined as $\mathbf{I}_{i,j} = \mathbf{I}_{j,i} = \frac{1}{|P_{i,j}|} \sum_{p \in P_{i,j}} \beta_p$ where i, j are indices in the set of unique stains S, $P_{i,j}$ is the set of all pairs between stains s_i and s_j and β_p represents the GAT attention weights for pair p, extracted from the model's attention mechanism. $|P_{i,j}|$ is the number of pairs between stains s_i and s_j. This score quantifies the importance of edge connections between nodes of different stain types.

GNN Heatmap. Extending on the use of attention scores, we design a simple GNN heatmap visualization method. The attention scores calculated for each node at the first SAAP layer are extracted and successively updated after each the pooling procedure. The final attention scores are min-max normalized and mapped back to their spatial location to obtain an attention heatmap of node importance. The resulting heatmap overlay provides a visual interpretation of the GNN model attention, highlighting the regions of the image that are considered most important for the downstream task.

4. Experiments

4.1. Datasets

We test our pipeline on two autoimmune multi-stain datasets, one for Rheumatoid Arthritis and the other for Sjogren's Disease. Each dataset is composed of H&E slides, with approximately 3 IHC slides of different immune biomarkers per patient. In SM. A, we give further information on the stains present in each dataset.

4.1.1. Rheumatoid Arthritis

This dataset consists of 607 Whole Slide Images (WSIs) from 153 RA patients, categorized into low (N=66) and high (N=87) inflammatory subtypes [24]. Samples were stained with H&E and the IHC markers CD20+ B cells, CD68+ macrophages, and CD138+ macrophages (see Fig. SM. A for details). 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, totaling approximately 275k patches.

4.1.2. Sjogren

This dataset consists of 347 WSIs of labial salivary gland biopsies sampled from 93 patients, with 46 cases of non-specific Sicca and 47 cases of Sjpgren. 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 (see SM. B for details). Each patient has a variable set of multi-stain WSIs, averaging 3.7 stains per patient. We perform detection of inflammatory patterns. We extract non-overlapping patches at a 20x magnification, keeping those with over 30% tissue coverage, totaling approximately 237k patches.

4.2. Implementation 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 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.

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 $1e^{-3}$ and weight decay $L_2 = 0.01$. No learning scheduler was used. We show the hyperparameters used in Table SM.3. Training was conducted on an NVidia A100 GPU (40Gb). See SM. C, SM. D for hyperparameters used and peak VRAM and memory use.

Benchmarking and ablation studies. We compare our method against seven SOTA methods, ABMIL [25], CLAM-SB [39], DeepGraphConv [37] PatchGCN [9], TransMIL [51], GTP [63] and MUSTANG [17]. We perform ablation on the different components of our pipeline: the SAAP module, the RW positional encoding, and the Multi Head Self-Attention layer.

5. Results

In Table 1 we present the results obtained by BioX-CPath on both datasets. On the RA dataset, our model achieved 0.90 ($\pm$0.019) accuracy, representing a 4 percent point improvement over the next best performing model, MUSTANG (0.86 $\pm$0.021). BioX-CPath did not outperfrom MUSTANG in AUC (0.96 $\pm$0.007) and average precision (0.98 $\pm$0.004), however did well compared to other methods. On the Sjogren dataset BioX-CPath achieved 0.84

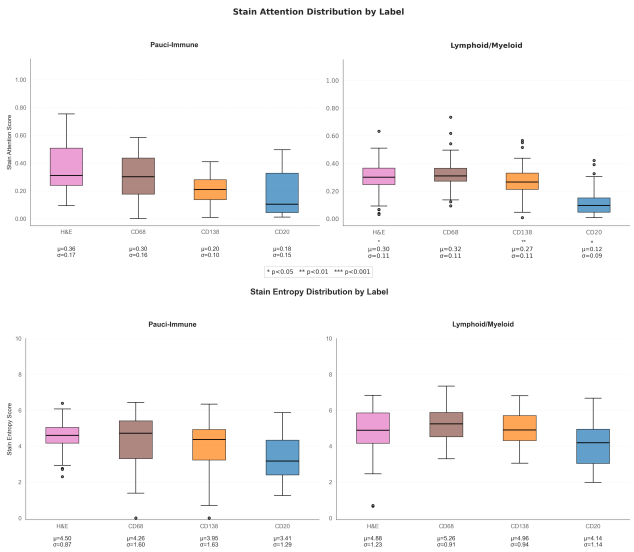

Figure 2. **RA Dataset Explainability**: The top row shows box plots of the SAAP scores distribution for different stain types (H&E, CD68, CD138, and CD20) for each classification label in the RA dataset (Pauci-Immune and Lymphoid/Myeloid). The bottom row shows the entropy score distributions for each of the stain types according to the classification label.

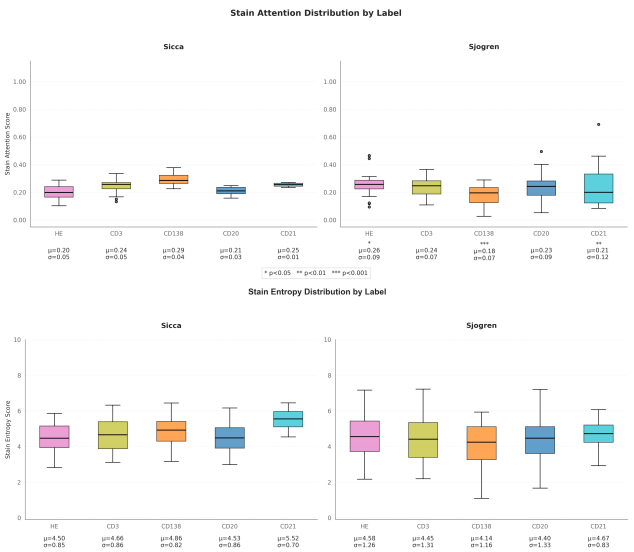

Figure 3. **Sjogren Dataset Explainability**: The top row shows box plots of the SAAP scores for different stain types (HE, CD3, CD138, CD20, and CD21) for each classification label in the Sjogren dataset (Sicca and Sjogren). The bottom row shows the entropy score distributions for each of the stain types according to the classification label.

($\pm$0.018) accuracy, showing a significant improvement over both CLAM-SB and MUSTANG (0.80 $\pm$0.018). The model also demonstrated stronger AUC (0.88 $\pm$0.023) and average precision (0.86 $\pm$0.032) compared to all baseline methods.

Ablation results shown in Tables 2 and 3 highlight the contribution of each component in our model. On the Sjogren dataset, the baseline model achieved 0.756 ($\pm$0.059) accuracy, while adding the RW positional encoding improved the performance to 0.80 ($\pm$0.038), indicating the importance of adding long-range topological information to the graph. The addition of SAAP provided another substantial boost, bringing the accuracy to 0.84 ($\pm$0.018). Similarly for the RA dataset, while the positional encoding improved the accuracy from 0.79 ($\pm$0.018) to 0.86 ($\pm$0.018), the full model with SAAP achieved the best performance at 0.90 ($\pm$0.019). We note the addition of the MHSA brought a slight decrease in performance. However, given the gains in model interpretability we do not view this as a significant disadvantage (we discuss this further in SM.8).

5.1. Biological Interpretability

5.1.1. RA

The Pauci-Immune pathotype exhibited lower attention scores for CD138 ($\mu = 0.20$, $\sigma = 0.10$, $p < 0.01$) and CD20 ($\mu = 0.18$, $\sigma = 0.15$, $p < 0.05$) markers, reflecting the characteristic scarcity of lymphocytic and plasma cell infiltrates in this disease subset. The lower entropy values observed in these samples (CD20: $\mu = 3.41$, $\sigma = 1.29$; CD138: $\mu = 3.95$, $\sigma = 1.63$) quantitatively capture the more ordered tissue architecture and sparse inflammatory foci associated with this RA pathotype. Conversely, Lymphoid/Myeloid samples showed more balanced attention distribution across CD68 ($\mu = 0.32$, $\sigma = 0.11$) and CD138 ($\mu = 0.27$, $\sigma = 0.11$) with consistently higher entropy values (CD68: $\mu = 5.26$, $\sigma = 0.91$; CD138: $\mu = 4.96$, $\sigma = 0.94$), reflecting the established role of plasma cells and macrophages in driving severe disease through autoantibody production and pro-inflammatory cytokine secretion [12, 61]. These computational findings provide quantitative support for the histological classification of RA subtypes, where Lymphoid/Myeloid pathotypes demonstrate abundant but disorganized immune cell infiltrates, while Pauci-Immune samples show more limited inflammatory patterns and more ordered tissue architecture [23, 35]. The relatively high attention scores for H&E staining in Pauci-Immune ($\mu = 0.36$, $\sigma = 0.17$, $p < 0.05$) align with the understanding that when specific immune cell infiltrates are less prominent, general tissue architecture becomes more informative for pathotype classification, reflecting the heterogeneous nature of RA synovitis and its immunological basis.

5.1.2. Sjogren

Sjogren's samples show a balanced attention across immune markers, with CD3 ($\mu = 0.24$, $\sigma = 0.07$), CD20

Table 1. Performance comparison of BioX-CPath against SOTA methods on the RA and Sjogren datasets. We report accuracy, AUC, and average precision (AP) with standard error shown in parentheses. The best results for each metric are shown in bold, with the second best underlined.

	RA			Sjogren		
	Accuracy (↑)	**AUC (↑)**	**AP (↑)**	**Accuracy (↑)**	**AUC (↑)**	**AP (↑)**
ABMIL [25]	0.79 (0.028)	0.89 (0.027)	0.92 (0.019)	0.73 (0.018)	0.80 (0.035)	0.79 (0.044)
CLAM-SB [39]	0.81 (0.026)	0.92 (0.011)	0.95 (0.008)	<u>0.80 (0.018)</u>	<u>0.85 (0.017)</u>	<u>0.85 (0.026)</u>
TransMIL [51]	0.80 (0.025)	0.87 (0.024)	0.91 (0.021)	0.75 (0.018)	0.73 (0.011)	0.74 (0.017)
DeepGraphConv [37]	0.81 (0.025)	0.88 (0.009)	0.92 (0.007)	0.77 (0.038)	0.83 (0.031)	0.83 (0.039)
Patch-GCN [10]	0.83 (0.015)	0.91 (0.019)	0.94 (0.014)	0.77 (0.019)	0.85 (0.015)	0.83 (0.030)
GTP [63]	0.79 (0.020)	0.87 (0.012)	0.92 (0.007)	0.62 (0.048)	0.73 (0.031)	0.72 (0.024)
MUSTANG [17]	<u>0.86 (0.021)</u>	<u>0.96 (0.010)</u>	<u>0.97 (0.006)</u>	<u>0.80 (0.018)</u>	<u>0.85 (0.019)</u>	0.84 (0.026)
BioX-CPath [ours]	**0.90 (0.019)**	**0.96 (0.007)**	**0.98 (0.004)**	**0.84 (0.018)**	**0.88 (0.023)**	**0.86 (0.032)**

Table 2. Ablation on model components shown on the Sjogren dataset.

	Accuracy (↑)	**AUC (↑)**	**AP (↑)**
Baseline	0.756 (0.059)	0.849 (0.024)	0.84 (0.036)
+ MHSA	0.736 (0.049)	0.849 (0.021)	0.86 (0.036)
+ RW	0.80 (0.038)	0.84 (0.035)	0.81 (0.034)
+ SAAP	**0.84 (0.018)**	**0.88 (0.023)**	**0.86 (0.032)**

Table 3. Ablation on model components shown on the RA dataset.

	Accuracy (↑)	**AUC (↑)**	**AP (↑)**
Baseline	0.79 (0.018)	0.87 (0.011)	0.92 (0.010)
+ MHSA	0.78 (0.025)	0.88(0.024)	0.92 (0.018)
+ RW	0.86 (0.018)	0.95 (0.010)	0.98 (0.007)
+ SAAP	**0.90 (0.019)**	**0.96 (0.007)**	**0.98 (0.004)**

($\mu = 0.23$, $\sigma = 0.09$), and CD21 ($\mu = 0.21$, $\sigma = 0.12$) receiving balanced attention, along with HE staining ($\mu = 0.26$, $\sigma = 0.09$). This reflects characteristic organized lymphocytic infiltrates with a mix of B-cells, plasma cells, and T-cells [6], as well as the importance of changes in tissue architecture. Most notably CD138 shows significantly lower attention in the Sjogren's group ($\mu = 0.18$, $\sigma = 0.07$) compared to Sicca ($\mu = 0.29$, $\sigma = 0.04$, $p < 0.001$), with lower entropy scores ($\mu = 4.14$, $\sigma = 1.16$) suggesting that specific plasma cell organization patterns, rather than overall abundance, are distinctive for Sjogren's pathology, which is consistent with the formation of ectopic lymphoid structures typical in Sjogren's [45]. Additionally, CD21 shows significant attention differences between groups (Sjogren's $\mu = 0.21$, $\sigma = 0.12$; Sicca $\mu = 0.25$, $\sigma = 0.01$, $p < 0.01$), with notable outliers in the Sjogren's group suggesting well-formed follicular dendritic networks in some cases. These patterns align with current understanding where Sicca represents non-inflammatory dryness with more homogeneously distributed immune cells (higher entropy), while Sjogren's demonstrates organized autoimmune infiltrates with more concentrated immune cell groupings (lower entropy). The model appears to have learned biologically relevant features that align with known pathological mechanisms.

In Supplementary materials, we conduct further analysis of model interpretability, looking at stain interaction scores (SM. F), GNN heatmaps (SM. G) and Layer Importance (SM. H).

6. Conclusion

BioX-CPath is an explainable GNN-based architecture for multistain IHC analysis, that bridges computational pathology and biological interpretability. By integrating multistain histopathological data into a unified framework, our approach not only achieves state-of-the-art accuracy but also provides mechanistic insights that align with established pathological mechanisms. This work establishes a foundation for developing and extending explainable computational pathology to other complex autoimmune and inflammatory diseases where multistain tissue analysis is essential for accurate diagnosis and subtyping.

Acknowledgments and Disclosure of Funding

We wish to thank Dr. Dovile Zilenaite for her insightful comments and knowledge, in particular discussing stain-stain interaction and entropy scores. A.G.S. receives funding from the Wellcome Trust [218584/Z/19/Z]. This paper utilized Queen Mary's Andrena HPC facility [28]. This work also acknowledges the support of the National Institute for Health and Care Research Barts Biomedical Research Centre (NIHR203330), a delivery partnership of Barts Health NHS Trust, Queen Mary University of London, St George's University Hospitals NHS Foundation Trust and St George's University of London.

BioX-CPath: Biologically-driven Explainable Diagnostics for Multistain IHC Computational Pathology

Supplementary Material

The content of the supplementary material are as follows: in A we describe IHC and the different immune markers used in this study and in B, we give further details on the datasets. We go over hyperparameters, memory usage and further technical clarification in sections C, D and E. Finally, we conduct further analysis of model interpretability, by looking at stain interaction scores in F, GNN heatmaps in G and Layer Importance H.

A. Immunohistochemistry Staining

IHC serves as a critical molecular mapping tool in clinical diagnostics and research, enabling precise identification and localization of disease-specific markers. The technique's power lies in its ability to reveal the molecular and cellular landscape of pathological processes, providing crucial information for diagnosis, prognosis, and treatment decisions.

In autoimmune disease diagnosis and monitoring, IHC enables detailed immune cell profiling through the characterization of inflammatory infiltrates and quantification of specific immune cell populations. This information reveals patterns of autoantibody deposits, complement activation, and tissue-specific autoantigen expression. The technique proves particularly valuable in assessing disease activity through the evaluation of inflammatory marker expression and monitoring tissue damage and repair processes.

IHC's integration into clinical decision-making represents a cornerstone of modern pathology practice. It supports diagnostic algorithms by validating initial morphological findings and resolving differential diagnoses through confirmation of disease-specific molecular patterns. In treatment strategy development, IHC helps identify targetable pathways and predict treatment response, enabling more personalized therapeutic approaches.

A.1. CD Markers

CD markers (Cluster of Differentiation) are cell surface proteins that serve as essential identifiers in immunological analysis. Each marker identifies specific immune cell types, enabling detailed characterization of tissue immune responses.

- **CD20** is a B-lymphocyte-specific antigen expressed on the surface of pre-B and mature B cells. This marker is critically important in both diagnostic and therapeutic contexts, particularly in B-cell lymphomas and autoimmune disorders. CD20 serves as the target for rituximab and other monoclonal antibody therapies, making its detection crucial for treatment planning. In lymphoid tissue analysis, CD20 staining helps identify B-cell populations and assess their distribution within tissue architecture.
- **CD21** is predominantly expressed on mature B cells and follicular dendritic cells. It plays a crucial role in the formation and maintenance of germinal centers within lymphoid tissues. In diagnostic pathology, CD21 staining is particularly valuable for visualizing follicular dendritic cell networks and assessing lymphoid tissue organization. This marker is often used to evaluate lymphoid tissue architecture in conditions such as lymphomas and autoimmune disorders.
- **CD68** is a glycoprotein expressed primarily by macrophages and monocytes. In tissue analysis, CD68 serves as a reliable marker for identifying tissue-resident macrophages and assessing inflammatory responses. In autoimmune disease diagnostics, CD68 staining helps quantify macrophage infiltration and assess disease activity.
- **CD138** is a transmembrane heparan sulfate proteoglycan primarily expressed on plasma cells and some epithelial cells. In autoimmune disease diagnostics, CD138 helps evaluate plasma cell infiltration and potential antibody production sites within affected tissues.
- **CD3** is a fundamental marker of T lymphocytes, expressed throughout T-cell development and maintained on mature T cells. CD3 staining is crucial in diagnosing T-cell lymphomas, assessing T-cell-mediated immune responses. In the context of autoimmune diseases, CD3 staining helps characterize the T-cell component of inflammatory infiltrates.

These markers, when analyzed together, map the immune cell landscape within tissues, revealing patterns of immune response and inflammation that guide diagnosis and treatment decisions.

B. Dataset Characteristics

To provide a benchmark on autoimmune multistain datasets, we use two clinical datasets. One dataset derives from a clinical trial, where patients with difficult to treat RA were recruited for treatment with rituximab. The other dataset derives from WSIs gathered for research purposes with the purpose of examining differences between patients presenting with dry eyes and mouth (Sicca) and patients subsequently diagnosed with Sjogren's Disease. In Figs. 4 and 5, we present clear examples of RA pathotypes and Sicca versus Sjogren presentation. While these images highlight characteristic differences, they represent more extreme cases specifically selected for illustrative clarity. The actual dataset exhibits considerably more heterogeneity in presentation, with many cases showing more subtle differences. In Table 4, we give further information on the stains present in each dataset. Each dataset is composed of H&E slides, with approximately 3 IHC slides of different immune biomarkers per patient.

C. Hyperparameters

We trained using the AdamW optimizer set to $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$, with a learning rate $1e^{-3}$ and weight decay $L_2 = 0.01$. No learning scheduler was used. We show our model's hyperparameters in Table 5.

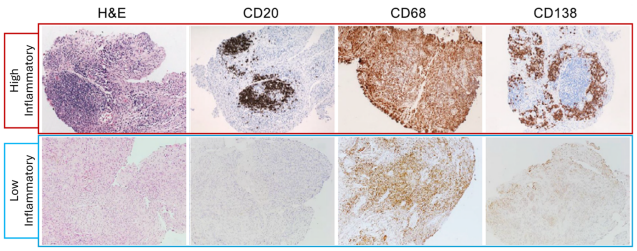

Figure 4. **Example of low inflammatory vs high inflammatory pathotype presentation in H&E and IHC stains for RA**: Rheumatoid Arthritis inflammatory pathotypes based on semi-quantitative analysis of synovial tissue biopsies stained with H&E, CD20+ B cells, CD68+ macrophages and IHC+ CD138 plasma cells.

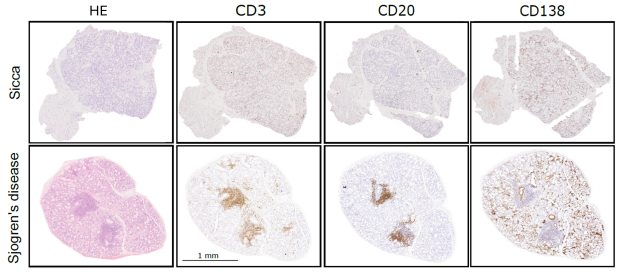

Figure 5. **Example of Sicca vs Sjogren presentation in H&E and IHC stains**: On top, a patient diagnosed with Sicca, on bottom a patient diagnosed with Sjogren. Here we show samples stained with IHC stains CD3+ T cells, CD20+ B cells and CD138+ plasma cells.

Table 4. **Metadata and dataset characteristics** for Sjogren and RA 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	**Detection**		**Subtyping**	
Labels	Negative	46	Low	66
	Positive	47	High	87

Table 5. **Our model hyperparameters**. We provide the hyperparameters used for each dataset to train our model.

Dataset	Seed	LR	# Layers	PE Dim	Pooling Ratio	Attention Heads	Dropout
RA	42	0.0001	4	20	0.7	2	0.2
Sjogren	42	0.0001	4	20	0.5	4	0.2

D. Memory Usage

Table 6 presents the RAM and VRAM utilization across all models compared against BioX-CPath. The varying RAM requirements stem from the distinct input representations each model processes: ABMIL/CLAM/TransMIL operate on embeddings, PatchGCN/GTP utilize region adjacency graphs, DeepGraphConv and MUSTANG work with feature space graphs, while BioX-CPath processes both feature and region adjacency graphs. VRAM consumption differences reflect the architectural complexity of each model. While simpler architectures like ABMIL [25] demonstrate minimal VRAM usage, our model's incorporation of GAT self-attention operations and an additional MHSA mechanism for interpretability results in higher peak VRAM consumption. We consider this increased memory footprint an acceptable trade-off given the model's superior performance and enhanced explainability. Future research could focus on developing a more memory-efficient architecture that maintains these characteristics, enabling translation to clinical practice.

Table 6. **Training and inference memory usage**. The table shows both RAM and VRAM peak usage during training and inference for the benchmark models shown in the main results table. We present results for the Sjogren dataset. Lower is **better**.

Model	Training		Inference	
	RAM (GB ↓)	VRAM (GB↓)	RAM (GB↓)	VRAM (GB↓)
ABMIL [25]	38.11	**0.09**	32.93	**0.09**
CLAM-SB [39]	44.38	0.14	45.03	0.10
TransMIL [51]	**35.87**	1.47	**29.39**	0.79
DeepGraphConv [37]	55.65	1.31	45.10	0.68
Patch-GCN [10]	41.03	7.42	41.99	4.37
GTP [63]	47.11	2.40	48.15	1.97
MUSTANG [17]	36.00	6.18	36.25	3.52
BioX-CPath (ours)	41.30	11.25	36.61	9.19

E. Technical clarification

The feature matrix is obtained through a hierarchical data loading architecture: (1) A slide-level DataLoader processes each stain-specific WSI, extracting patches and associated metadata (stain type, spatial coordinates, patient ID); (2) A patient-level loader stacks the stain-specific embeddings through vertical concatenation; (3) graphs are constructed using patch embeddings as node features with dual-criteria edge connectivity (feature and spatial proximity). The preprocessed patient graphs are then stored, loaded & batched with PyTorch Geometric DataLoader. We keep track of node and edge attributes, stored as categorical labels, through each layer of our model by mapping and storing their IDs after each pooling operation. When nodes are removed, edges are systematically pruned where either the source or target node was dropped, updating the edge list accordingly. While this can lead to disconnected components, the high initial connectivity of the pa-

tient graphs means these components emerge only in deeper layers of the encoder, where they exhibit "specialized" attention patterns focusing on specific stain or tissue regions. We exemplify this with a layer-wise graph WSIs overlay shown in Figure 9. The max (OR) operator was chosen over min (AND), based on graph connectivity patterns: using AND overly restricts edges ($\sim$10%), limiting message passing and cross-stain interactions. In contrast, OR preserves local and global connectivity, allowing the SAAP module to dynamically prioritize relevant edges. These design choices are all aimed at optimizing computational resources and information flow, under minimal supervision requirements (patient-level labels and stain-type slide annotations), while ensuring interpretable biologically-aligned results.

F. Stain-Stain Interactions

The stain-stain interaction patterns highlight key insights into model decision dynamics, which further deepen our understanding of model behavior and can be linked back to biological mechanisms. These attention-based interactions quantify how the model integrates information across different stain types when making classifications. We present the distribution of stain-stain interactions for both RA and Sjogren's in Figure 6 and 7.

F.1. RA

The stain-stain attention analysis reveals a consistent decrease in all self-interactions (CD138-CD138: -7.5%, CD20-CD20: -4.7%, H&E-H&E: -5.5%, CD68-CD68: -4.5%) in Lymphoid/Myeloid compared to Pauci-Immune pathotypes, suggesting a shift from examining intra-stain features toward integrated cross-stain attention patterns, which aligns with the higher entropy scores observed in Lymphoid/Myeloid and the known diffuse inflammatory infiltrates characteristic of this pathotype. The most pronounced changes in cross-stain interactions occur between lymphocyte markers and other stains (CD138-CD20: -7.4%, CD138-H&E: -5.3%, CD20-H&E: -5.3%), reflecting the disruption of normal tissue architecture by immune infiltrates in Lymphoid/Myeloid disease. In contrast, macrophage-related interactions (CD68-H&E: -4.4%, CD138-CD68: -4.2%, CD20-CD68: -4.0%) show more modest changes, suggesting a more consistent role for macrophages across pathotypes. The overall higher and more variable attention weights in Pauci-Immune samples compared to the more uniform, lower weights in Lymphoid/Myeloid indicate that Pauci-Immune classification relies on stronger, more specific feature relationships. Lymphoid/Myeloid requires broader integration of multiple signals, which is consistent with its more complex, heterogeneous inflammatory profile [35].

F.2. Sjogren

We see a systematic decrease in self-interactions (CD20-CD20: -6.0%, CD3-CD3: -2.2%, CD21-CD21: -2.1%, CD138-CD138: -1.3%), which suggests a shift from paying attention more broadly to the overall context in each single stain, and more toward integrated localized attention spanning across stain types, which aligns with the lower entropy scores obtained for Sjogren stains and the known pathology of more structured lymphoid organization in Sjogren [31]. We also note differences in the structural-immune interactions between Sjogren vs Sicca, with an increase in stain-stain attention between HE-CD21 ($+3.8\%$), HE-CD138 ($+1.9\%$) and HE-CD3 ($+1.2\%$) and a decrease in attention between HE-CD20 (-4.5%). On the other hand, changes in immune-immune interactions (CD138-CD3: -2.9%, CD138-CD20: -2.2%, CD20-CD3: -2.2%), taken in the context of the balanced stain attention scores obtained for these markers, also suggests a balanced model that integrates information across immune markers.

G. GNN Heatmaps

In Fig. 8, we show an example of the multistain stack of WSIs (CD138, CD3, CD20, C21, and HE) for one Sjogren positive patient, with the obtained cumulative node attention heatmap for each input stains. The stack of multistain WSIs is the input to our model, and the obtained GNN node heatmaps correspond to the direct mapping of the node attention scores to their original spatial location. We note that our proposed GNN heatmap accurately picks up on the presence of inflammatory aggregates in CD3, CD20, and H&E, as well as on more disperse attention patterns in CD138 and CD21. CD18 plasma cells are always present throughout the tissue, but will become over-activated and more prevalent in the inflamed tissue, leading to a more diffuse attention pattern. CD21 also accurately focuses on areas with presence of inflammatory aggregates, however also shows a more disperse attention pattern, potentially due to the smaller and fainter aggregates, compared to CD3/CD20 and H&E.

To illustrate cross-stack stain-stain interaction and the graph sparsification process through our model, Figure 9 shows G_{FRA} overlaid on the WSIs stack. The layer 1 graph is initially dense with two edge types: region-adjacent edges (red) connecting both across different stains and between spatial neighbors within each WSI, and feature-space edges (blue) linking semantically similar patches regardless of their location. As the graph progresses through the layers, it undergoes progressive sparsification. The transition shows a shift from more homogeneous distributions toward targeted cross-stain interactions, aligning with our quantitative findings of decreased self-attention and enhanced cross-stain integration. By layer 4, the preserved connections highlight important structural-immune relationships between tissue architecture (HE) and immune markers (CD3, CD20, CD21). This progressive refinement demonstrates how the model identifies the organized, integrated nature of immune infiltrates in Sjogren's, capturing diagnostically relevant cross-stain relationships rather than analyzing markers in isolation.

H. Layer Importance

We previously mentioned we chose to maintain a MHSA layer before the classification head in our model architecture, despite seeing a marginal drop in performance. This is because we considered it was a good trade-off with the additional insight obtained into model decision mechanics, providing another aspect to the explainability of our model with layer importance scores. Briefly, we concatenate the fixed size readouts obtained from each layer of our hierarchical graph patient encoder. This concatenated readout vector is the input to the MHSA. Because we know the size of each layer readout, we can now take the simple step of summing the corresponding attention weights. The rational is this will give

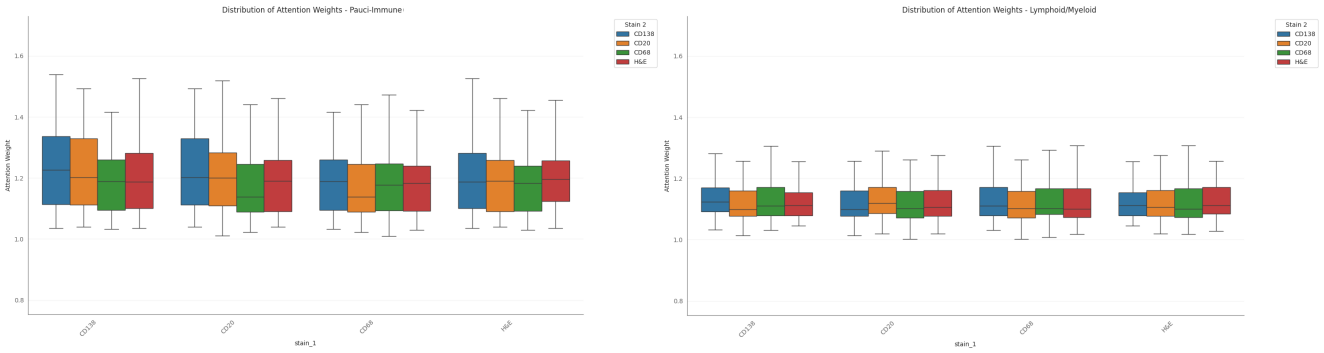

Figure 6. **Distribution of stain-to-stain interaction** scores for Pauci-Immune (Label 0, left) and Lymphoid/Myeloid (Label 1, right) cases. Each subplot shows the distribution of the average stain-stain attention scores for each stain pair (CD138, CD20, CD68, and H&E) interact with each other. For each source stain (x-axis), the box plots represent the distribution of interaction scores given to each target stain (colored boxes).

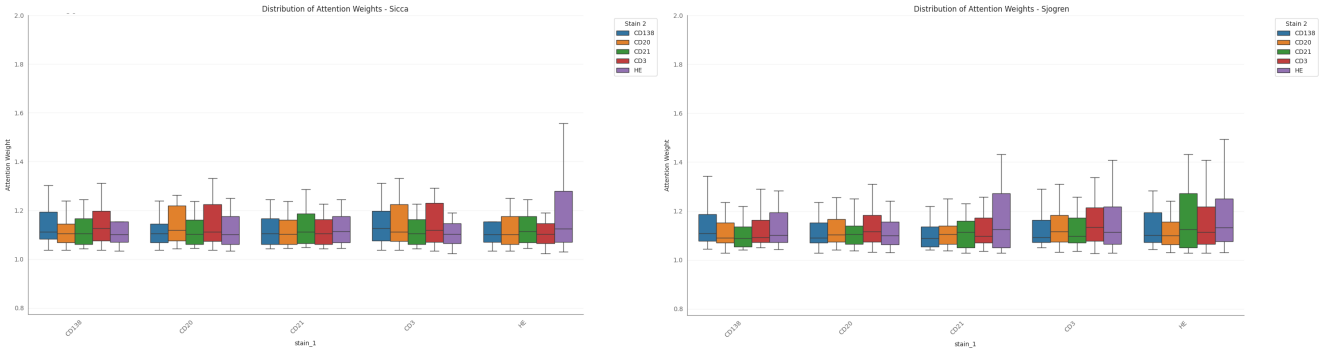

Figure 7. **Distribution of stain-to-stain interaction** scores for Sicca (Label 0, left) and Sjogren (Label 1, right) cases. Each subplot shows how different stains (CD138, CD20, CD21, CD3, and HE) interact with each other. For each source stain (x-axis), the box plots represent the distribution of interaction scores given to each target stain (colored boxes).

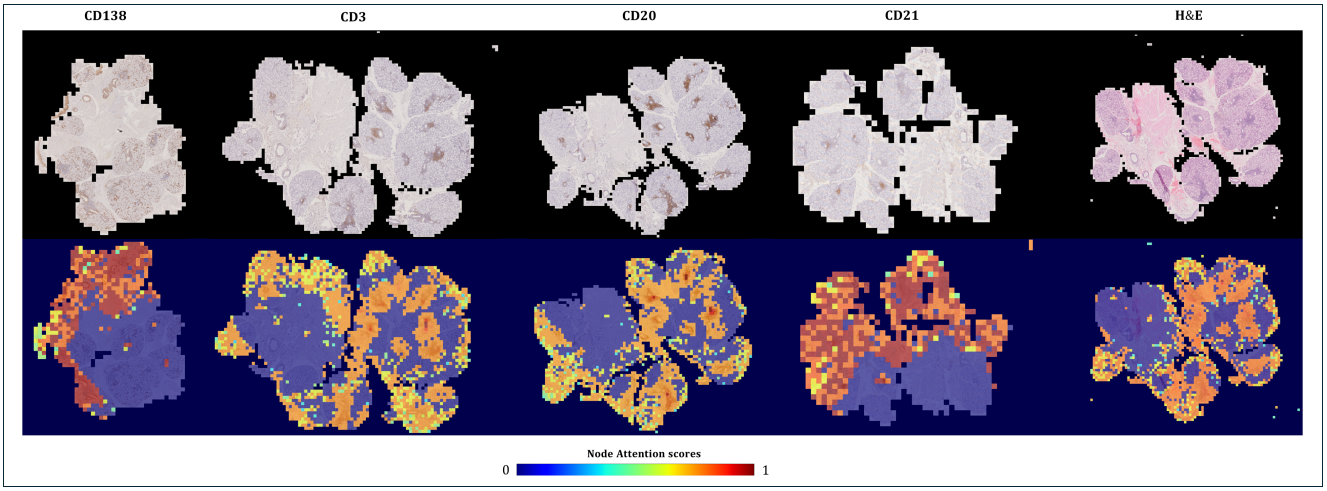

Figure 8. **Cumulative GNN node attention heatmap** obtained for a Sjogren positive patient with a stack of WSIs consisting of staining for CD138, CD20, C21, CD3 and H&E, where the red edges connecting across and correspond to region adjacency connectivity and the blue edges to the feature space connectivity. This stack is the input to our model and the obtained GNN heatmap corresponds to the direct mapping of the node attention scores back to their original spatial location.

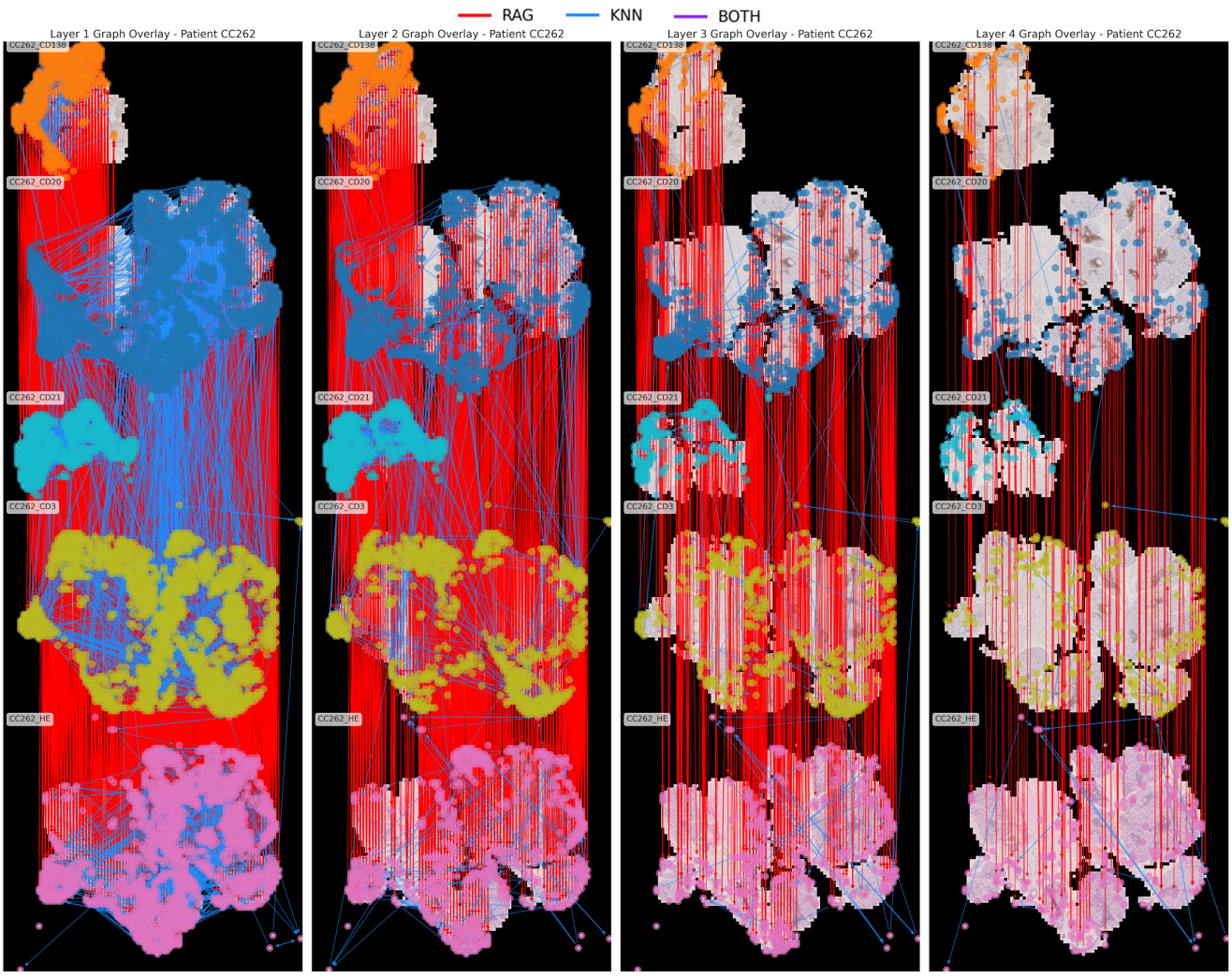

Figure 9. **Sparsification of input G_{FRA} through the GNN layers.** We plot the multistain patient input graph G_{FRA} as a spatial overlay on the stack of WSIs, to exemplify the connectivity both across and in the images. Edges connect nearest neighbors in both feature (blue) and region adjacent (red) space, with edges which are both feature and region nearest neighbors shown as purple.

us further insight into the role played by each layer in the model decision process and can potentially highlight inherent characteristics on the input data. We present these results in Figures 10 and 12.

H.1. RA

The layer attention results reveal distinct patterns between pathotypes. Pauci-Immune samples show balanced attention across Layers 2-4 ($\mu = 0.38$, $\mu = 0.31$, $\mu = 0.32$), suggesting reliance on features at multiple abstraction levels. In contrast, Lymphoid/Myeloid samples demonstrate strong preference for Layer 2 ($\mu = 0.47$, $\sigma = 0.08$), indicating mid-level features are particularly diagnostic. This aligns with our stain-stain interaction findings, where Lymphoid/Myeloid showed decreased self-attention and likely depends more on cross-stain integrations occurring at intermediate layers. Both pathotypes assign minimal attention to Layer 1 ($\mu = 0.00$), indicating here the raw features have limited classification value without higher-level processing. The higher variance in Layer 2 attention for Lymphoid/Myeloid ($\sigma = 0.08$ vs $\sigma = 0.02$) suggests greater patient-to-patient variability, consistent with its more heterogeneous inflammatory profile.

To exemplify this process, in Figure 11 we show the GNN node attention heatmaps obtained for each layer of the model for a WSI with CD18 staining of a RA patient with Lymphoid/Myeloid subtype. We can see a progressive refinement of attention across the layers, with Layer 1 showing broad, diffuse attention across the tissue, while Layers 2-4 reveal increasingly focused attention on specific regions. Layer 2 demonstrates the most pronounced attention patterns, concentrating on areas with visible cellular infiltrates, which aligns with our finding that this layer receives the highest attention weight ($\mu = 0.47$) for Lymphoid/Myeloid patients. Layers 3 and 4 further refine this attention, focusing on

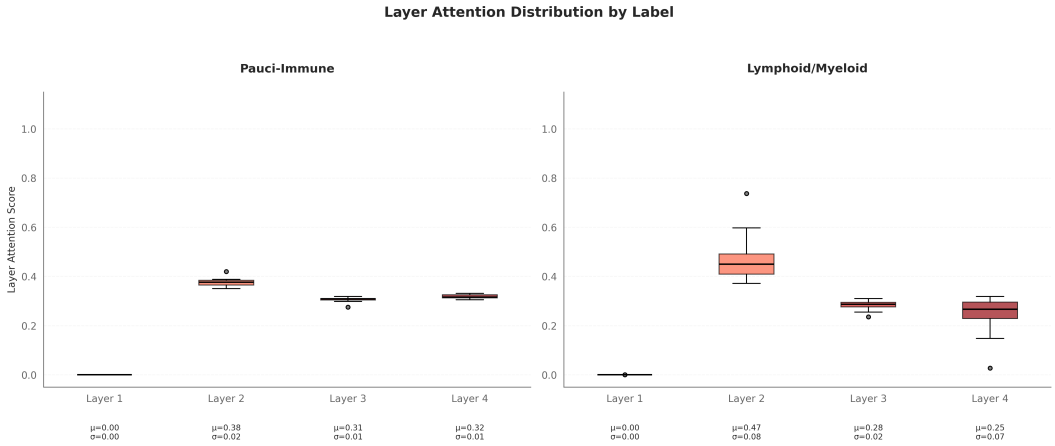

Figure 10. **Layer-wise attention patterns by label** in the hierarchical graph patient encoder, showing the distribution of attention scores across layers (1-4) for Pauci-Immune and Lymphoid/Myeloid cases, with corresponding mean (μ) and standard deviation (σ) values.

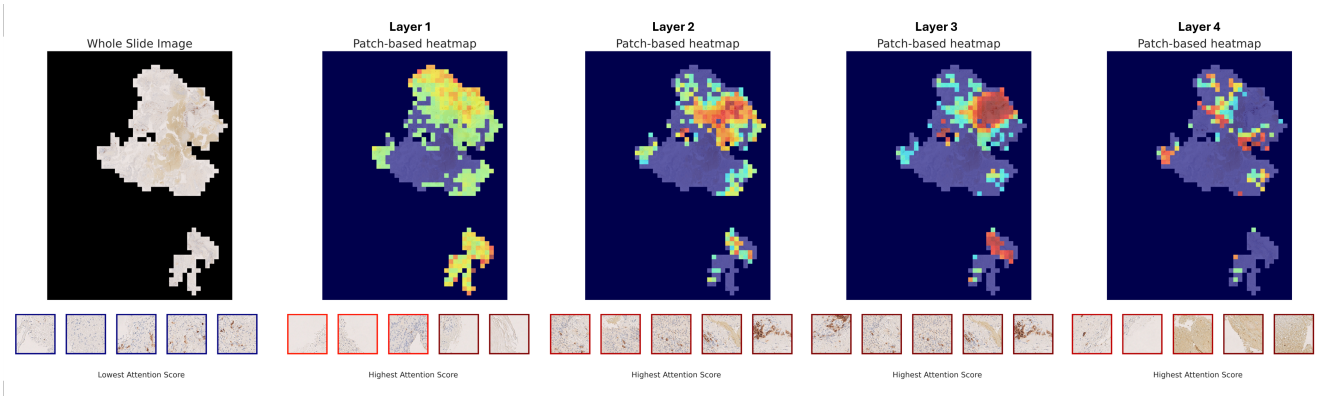

Figure 11. **Layer-wise attention visualization for a CD18-stained WSI Lymphoid/Myeloid RA patient.** The heatmaps show progression from broad attention in Layer 1 to increasingly focused attention in subsequent layers, with Layer 2 exhibiting the strongest patterns, consistent with quantitative attention scores. Bottom panels show highest and lowest attention patches, revealing cellular infiltrates in high-attention regions.

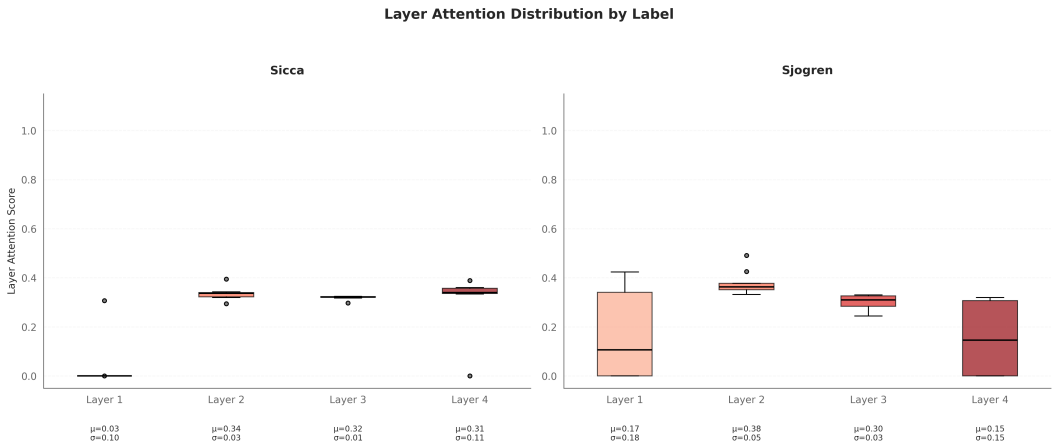

Figure 12. **Layer-wise attention patterns by label** in the hierarchical graph patient encoder, showing the distribution of attention scores across layers (1-4) for Sicca and Sjogren cases, with corresponding mean (μ) and standard deviation (σ) values.

smaller, more specific regions that likely represent areas with distinctive immune cell aggregates. This visualization supports our quantitative findings and illustrates how the model progressively builds its understanding of the pathotype from general tissue architecture to specific inflammatory aggregates characteristic of Lymphoid/Myeloid disease.

H.2. Sjogren

The layer attention distributions reveal distinct hierarchical processing patterns between Sicca and Sjogren's. For Sicca, attention is negligible in Layer 1 ($\mu = 0.03$, $\sigma = 0.10$) but distributes relatively uniformly across Layers 2-4 ($\mu = 0.34$, $\mu = 0.32$, $\mu = 0.31$ respectively). In contrast, Sjogren's shows substantial Layer 1 attention ($\mu = 0.17$, $\sigma = 0.18$) followed by peak attention at Layer 2 ($\mu = 0.38$, $\sigma = 0.05$) and then progressive decline through Layers 3-4 ($\mu = 0.30$, $\mu = 0.15$), with higher variance observed for Layers 1 and 4. The higher early-layer attention in Sjogren's suggests the model identifies organized immune structures in initial processing stages, corresponding to the decreased self-attention and increased cross-stain integration observed in Sjogren's stain-stain interaction scores. The declining attention pattern in deeper layers for Sjogren's, compared to sustained attention in Sicca, indicates different processing requirements: Sjogren's features are captured earlier through identification of organized lymphoid structures, while Sicca requires more distributed processing across abstraction levels, consistent with its more homogeneous, less structured immune distributions (reflected in higher entropy values).

References

[1] Ralph Abboud, Radoslav Dimitrov, and İsmail İlkan Ceylan. Shortest Path Networks for Graph Property Prediction. Technical report, 2023. arXiv:2206.01003 [cs] type: article. 3

[2] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. 2

[3] Mohammed Adnan, Shivam Kalra, and Hamid R. Tizhoosh. Representation Learning of Histopathology Images using Graph Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4254–4261, 2020. ISSN: 2160-7516. 2

[4] Uri Alon and Eran Yahav. On the Bottleneck of Graph Neural Networks and its Practical Implications. Technical report, 2021. arXiv:2006.05205 [cs, stat] type: article. 3

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. Technical report, 2016. arXiv:1409.0473 [cs, stat] type: article. 3

[6] Michele Bombardieri, Myles Lewis, and Costantino Pitzalis. Ectopic lymphoid neogenesis in rheumatic autoimmune diseases. *Nature Reviews Rheumatology*, 13(3):141–154, 2017. 8

[7] Cătălina Cangea, Petar Veličković, Nikola Jovanović, Thomas Kipf, and Pietro Liò. Towards Sparse Hierarchical Graph Classifiers. Technical report, 2018. arXiv:1811.01287 [cs, stat] type: article. 3, 4

[8] Tsai Hor Chan, Fernando Julio Cendra, Lan Ma, Guosheng Yin, and Lequan Yu. Histopathology whole slide image analysis with heterogeneous graph representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15661–15670, 2023. 2, 3, 4

[9] Richard J. Chen, Ming Y. Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Whole Slide Images are 2D Point Clouds: Context-Aware Survival Prediction using Patch-based Graph Convolutional Networks. Technical report, 2021. arXiv:2107.13048 [cs, eess, q-bio] type: article. 2, 3, 6

[10] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16134, 2022. ISSN: 2575-7075. 8, 2

[11] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 4, 5

[12] Glynn Dennis, Cécile TJ Holweg, Sarah K Kummerfeld, David F Choy, A Francesca Setiadi, Jason A Hackney, Peter M Haverty, Houston Gilbert, Wei Yu Lin, Lauri Diehl, S Fischer, An Song, David Musselman, Micki Klearman, Cem Gabay, Arthur Kavanaugh, Judith Endres, David A Fox, Flavius Martin, and Michael J Townsend. Synovial phenotypes in rheumatoid arthritis correlate with response to biologic therapeutics. *Arthritis Research & Therapy*, 16(2):R90, 2014. 7

[13] Chaitanya Dwivedi, Shima Nofallah, Maryam Pouryahya, Janani Iyer, Kenneth Leidal, Chuhan Chung, Timothy Watkins, Andrew Billin, Robert Myers, John Abel, and Ali Behrooz. Multi Stain Graph Fusion for Multimodal Integration in Pathology. pages 1835–1845, 2022. 2, 4

[14] Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph Neural Networks with Learnable Structural and Positional Representations. Technical report, 2022. arXiv:2110.07875 [cs] type: article. 3, 4

[15] I. Keren Evangeline, J. Glory Precious, Natesan Pazhanivel, and S. P. Angeline Kirubha. Automatic detection and counting of lymphocytes from immunohistochemistry cancer images using deep learning. *Journal of Medical and Biological Engineering*, 40:735 – 747, 2020. 1

[16] Olga Fourkioti, Matt De Vries, and Chris Bakal. CAMIL:

Context-Aware Multiple Instance Learning for Cancer Detection and Subtyping in Whole Slide Images. 2023. 2

[17] Amaya Gallagher-Syed, Luca Rossi, Felice Rivellese, Costantino Pitzalis, Myles Lewis, Michael Barnes, and Gregory Slabaugh. 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. 2, 4, 6, 8

[18] Amaya Gallagher-Syed, Elena Pontarini, Myles J Lewis, Michael R Barnes, and Gregory Slabaugh. Going beyond h&e and oncology: How do histopathology foundation models perform for multi-stain ihc and immunology? *arXiv preprint arXiv:2410.21560*, 2024. 4

[19] Hongyang Gao and Shuiwang Ji. Graph U-Nets. pages 2083–2092. PMLR, 2019. 3

[20] Hongyang Gao, Yongjun Chen, and Shuiwang Ji. Learning Graph Pooling and Hybrid Convolutional Operations for Text Representations. In *The World Wide Web Conference*, pages 2743–2749, New York, NY, USA, 2019. Association for Computing Machinery. 3

[21] William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159. 3

[22] Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. Hˆ2-MIL: Exploring Hierarchical Representation with Heterogeneous Multiple Instance Learning for Whole Slide Image Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):933–941, 2022. 4

[23] Frances Humby, Myles Lewis, Nandhini Ramamoorthi, Jason A Hackney, Michael R Barnes, Michele Bombardieri, A Francesca Setiadi, Stephen Kelly, Fabiola Bene, Maria DiCicco, et al. Synovial cellular and molecular signatures stratify clinical response to csdmard therapy and predict radiographic progression in early rheumatoid arthritis patients. *Annals of the rheumatic diseases*, 78(6):761–772, 2019. 7

[24] Frances Humby, Patrick Durez, Maya H. Buch, Myles J. Lewis, Hasan Rizvi, Felice Rivellese, Alessandra Nerviani, Giovanni Giorli, Arti Mahto, Carlomaurizio Montecucco, Bernard Lauwerys, Nora Ng, Pauline Ho, Michele Bombardieri, Vasco C. Romão, Patrick Verschueren, Stephen Kelly, Pier Paolo Sainaghi, Nagui Gendi, Bhaskar Dasgupta, Alberto Cauli, Piero Reynolds, Juan D. Cañete, Robert Moots, Peter C. Taylor, Christopher J. Edwards, John Isaacs, Peter Sasieni, Ernest Choy, Costantino Pitzalis, Charlotte Thompson, Serena Bugatti, Mattia Bellan, Mattia Congia, Christopher Holroyd, Arthur Pratt, João Eurico Cabral da Fonseca, Laura White, Louise Warren, Joanna Peel, Rebecca Hands, Liliane Fossati-Jimack, Gaye Hadfield, Georgina Thorborn, Julio Ramirez, and Raquel Celis. 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. *The Lancet*, 397(10271):305–317, 2021. 6

[25] M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *Proc. 35th ICML*, pages 2127–2136, 2018. 2, 6, 8

[26] Guillaume Jaume, Anurag Jayant Vaidya, Andrew Zhang, Andrew H Song, Richard J. Chen, Sharifa Sahai, Dandan Mo, Emilio Madrigal, Long Phi Le, and Mahmood Faisal. Multistain pretraining for slide representation learning in pathology. In *European Conference on Computer Vision*. Springer, 2024. 1, 2

[27] Jakob Nikolas Kather, Alexander T. Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H. Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, Heike I. Grabsch, Takaki Yoshikawa, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, Christian Trautwein, and Tom Luedde. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine*, 25(7):1054–1056, 2019. 2

[28] Thomas King, Simon Butcher, and Lukasz Zalewski. Apocrita - High Performance Computing Cluster for Queen Mary University of London. 2017. 8

[29] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. Technical report, 2017. arXiv:1609.02907 [cs, stat] type: article. 3

[30] Boris Knyazev, Graham W Taylor, and Mohamed Amer. Understanding Attention and Generalization in Graph Neural Networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[31] Frans G. M. Kroese, Erlin A. Haacke, and Michele Bombardieri. The role of salivary gland histopathology in primary Sjögren's syndrome: promises and pitfalls. *Clinical and Experimental Rheumatology*, 36 Suppl 112(3):222–233, 2018. 3

[32] Dongha Lee, Su Kim, Seonghyeon Lee, Chanyoung Park, and Hwanjo Yu. Learnable Structural Semantic Readout for Graph Classification. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1180–1185, 2021. ISSN: 2374-8486. 3

[33] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. 3, 4

[34] Yongju Lee, Jeong Hwan Park, Sohee Oh, Kyoungseob Shin, Jiyu Sun, Minsun Jung, Cheol Lee, Hyojin Kim, Jin-Haeng Chung, Kyung Chul Moon, and Sunghoon Kwon. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nature Biomedical Engineering*, pages 1–15, 2022. 2

[35] Myles J. Lewis, Michael R. Barnes, Kevin Blighe, Katriona Goldmann, Sharmila Rana, Jason A. Hackney, Nandhini Ramamoorthi, Christopher R. John, David S. Watson, Sarah K. Kummerfeld, Rebecca Hands, Sudeh Riahi, Vidalba Rocher-Ros, Felice Rivellese, Frances Humby, Stephen Kelly, Michele Bombardieri, Nora Ng, Maria DiCicco, Désirée van der Heijde, Robert Landewé, Annette van der Helm-van Mil, Alberto Cauli, Iain B. McInnes, Christopher D. Buckley, Ernest Choy, Peter C. Taylor, Michael J. Townsend, and Costantino Pitzalis. Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment

Response Phenotypes. *Cell Reports*, 28(9):2455–2470.e5, 2019. 7, 3

[36] Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. *Conference on Computer Vision and Pattern Recognition Workshops. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Workshops*, 2021:14318–14328, 2021. 2

[37] Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph CNN for Survival Analysis on Whole Slide Pathological Images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 174–182, Cham, 2018. Springer International Publishing. 2, 6, 8

[38] Sitong Liu, Kechun Liu, Samuel Margolis, Wenjun Wu, Stevan R. Knezevich, David E. Elder, Megan M. Eguchi, Joann G. Elmore, and Linda Shapiro. Generating seamless virtual immunohistochemical whole slide images with content and color consistency. 2024. arXiv:2410.01072. 2

[39] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng*, 5(6):555–570, 2021. 2, 6, 8

[40] Wenqi Lu, Simon Graham, Mohsin Bilal, Nasir Rajpoot, and Fayyaz Minhas. Capturing Cellular Topology in Multi-Gigapixel Pathology Images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1049–1058, 2020. ISSN: 2160-7516. 2

[41] Yao Ma, Suhang Wang, Charu C. Aggarwal, and Jiliang Tang. Graph Convolutional Networks with EigenPooling. 2019. arXiv:1904.13107. 3

[42] Shino Magaki, Seyed A. Hojat, Bowen Wei, Alexandra So, and William H. Yong. An Introduction to the Performance of Immunohistochemistry. *Methods in molecular biology (Clifton, N.J.)*, 1897:289, 2019. 1, 2

[43] Emmanuel Noutahi, Dominique Beaini, Julien Horwood, Sébastien Giguère, and Prudencio Tossou. Towards Interpretable Sparse Graph Representation Learning with Laplacian Pooling. 2020. arXiv:1905.11577. 3

[44] Costantino Pitzalis, Gareth W. Jones, Michele Bombardieri, and Simon A. Jones. Ectopic lymphoid-like structures in infection, cancer and autoimmunity. *Nature Reviews Immunology*, 14(7):447–462, 2014. 1

[45] Elena Pontarini, Elisabetta Sciacca, Farzana Chowdhury, Sofia Grigoriadou, Felice Rivellese, William J Murray-Brown, Davide Lucchesi, Liliante Fossati-Jimack, Alessandra Nerviani, Edyta Jaworska, et al. Serum and tissue biomarkers associated with composite of relevant endpoints for sjögren syndrome (cress) and sjögren tool for assessing response (star) to b cell–targeted therapy in the trial of anti–b cell therapy in patients with primary sjögren syndrome (tractiss). *Arthritis & Rheumatology*, 76(5):763–776, 2024. 8

[46] Daniel Reisenbüchler, Sophia J. Wagner, Melanie Boxberg, and Tingying Peng. Local attention graph-based transformer for multi-target genetic alteration prediction. In *Lecture Notes in Computer Science*, pages 377–386. Springer Nature Switzerland, 2022. 4

[47] Abtin Riasatian, Morteza Babaie, Danial Maleki, Shivam Kalra, Mojtaba Valipour, Sobhan Hemati, Manit Zaveri, Amir Safarpoor, Sobhan Shafiei, Mehdi Afshari, Maral Rasoolijaberi, Milad Sikaroudi, Mohd Adnan, Sultaan Shah, Charles Choi, Savvas Damaskinos, Clinton JV Campbell, Phedias Diamandis, Liron Pantanowitz, Hany Kashani, Ali Ghodsi, and H. R. Tizhoosh. Fine-Tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. *Medical Image Analysis*, 70: 102032, 2021. 2

[48] Emanuelle M. Rizk, Robyn D. Gartrell, Luke W. Barker, Camden L. Esancy, Grace G. Finkel, Darius D. Bordbar, and Yvonne M. Saenger. Prognostic and predictive immunohistochemistry-based biomarkers in cancer and immunotherapy. *Hematology/oncology clinics of North America*, 33(2):291, 2019. 2

[49] J. Saltz, Rajarsi R. Gupta, Le Hou, Tahsin M. Kurç, Pankaj Kumar Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca C. Batiste, John S. Van Arnam, Ilya Shmulevich, Arvind U. K. Rao, Alexander J. Lazar, Ashish Sharma, and Vésteinn Thorsson. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23 1:181–193.e7, 2018. 2

[50] Lisa Schneckenreiter, Richard Freinschlag, Florian Sestak, Johannes Brandstetter, Günter Klambauer, and Andreas Mayr. GNN-VPA: A Variance-Preserving Aggregation Strategy for Graph Neural Networks. 2024. 3

[51] 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. 2, 6, 8

[52] Zaneta Swiderska-Chadaj, Hans Pinckaers, Mart van Rijthoven, Maschenka C. A. Balkenhol, Margarita Melnikova, Oscar G. F. Geessink, Quirine F. Manson, Mark E. Sherman, António Polónia, Jeremy Parry, Mustapha Abubakar, Geert J. S. Litjens, Jeroen van der Laak, and Francesco Ciompi. Learning to detect lymphocytes in immunohistochemistry with deep learning. *Medical image analysis*, 58:101547, 2019. 1

[53] Roger Trullo, Quoc-Anh Bui, Qi Tang, and Reza Olfati-Saber. Image translation based nuclei segmentation for immunohistochemistry images. 2022. arXiv:2208.06202. 1

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3

[55] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. Technical report, 2018. arXiv:1710.10903 [cs, stat] type: article. 3, 4, 5

[56] Philippe Weitz, Masi Valkonen, Leslie Solorzano, Circe Carr, Kimmo Kartasalo, Constance Boissin, Sonja Koivukoski, Aino Kuusela, Dusan Rasic, Yanbo Feng, Sandra Sinius Pouplier, Abhinav Sharma, Kajsa Ledesma

Eriksson, Stephanie Robertson, Christian Marzahl, Chandler D. Gatenbee, Alexander R. A. Anderson, Marek Wodzinski, Artur Jurgas, Niccolò Marini, Manfredo Atzori, Henning Müller, Daniel Budelmann, Nick Weiss, Stefan Heldmann, Johannes Lotz, Jelmer M. Wolterink, Bruno De Santi, Abhijeet Patil, Amit Sethi, Satoshi Kondo, Satoshi Kasai, Kousuke Hirasawa, Mahtab Farrokh, Neeraj Kumar, Russell Greiner, Leena Latonen, Anne-Vibeke Laenkholm, Johan Hartman, Pekka Ruusuvuori, and Mattias Rantalainen. The ACROBAT 2022 challenge: Automatic registration of breast cancer tissue. *Medical Image Analysis*, 97:103257, 2024. 2

[57] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14138–14148, 2021. 2

[58] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? Technical report, 2019. arXiv:1810.00826 [cs, stat] type: article. 3

[59] Zhaoyang Xu, Xingru Huang, Carlos Fernández Moro, Béla Bozóky, and Qianni Zhang. Gan-based virtual re-staining: A promising solution for whole slide image analysis. 2022. arXiv:1901.04059. 2

[60] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical Graph Representation Learning with Differentiable Pooling. Technical report, 2019. arXiv:1806.08804 [cs, stat] type: article. 3, 4

[61] Fan Zhang, Kevin Wei, Kamil Slowikowski, Chamith Y Fonseka, Deepak A Rao, Stephen Kelly, Susan M Goodman, Darren Tabechian, Laura B Hughes, Karen Salomon-Escoto, et al. Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nature immunology*, 20(7):928–942, 2019. 7

[62] Zhen Zhang, Jiajun Bu, Martin Ester, Jianfeng Zhang, Chengwei Yao, Zhi Yu, and Can Wang. Hierarchical Graph Pooling with Structure Learning. Technical report, 2019. arXiv:1911.05954 [cs, stat] type: article. 3

[63] Yi Zheng, Rushin H. Gindra, Emily J. Green, Eric J. Burks, Margrit Betke, Jennifer E. Beane, and Vijaya B. Kolachalama. A Graph-Transformer for Whole Slide Image Classification. *IEEE Transactions on Medical Imaging*, 41(11): 3003–3015, 2022. 2, 6, 8

[64] Yushan Zheng, Zhiguo Jiang, Jun Shi, Fengying Xie, Haopeng Zhang, Wei Luo, Dingyi Hu, Shujiao Sun, Zhongmin Jiang, and Chenghai Xue. Encoding histopathology whole slide images with location-aware graphs for diagnostically relevant regions retrieval. *Medical Image Analysis*, 76: 102308, 2022. 2

[65] Yue Zhou, Lei Tao, Jiahao Qiu, Jing Xu, Xinyu Yang, Yu Zhang, Xinyu Tian, Xinqi Guan, Xiaobo Cen, and Yinglan Zhao. Tumor biomarkers for diagnosis, prognosis and targeted therapy. *Signal Transduction and Targeted Therapy*, 9 (1):1–86, 2024. 1

[66] Dovile Zilenaite-Petrulaitiene, Allan Rasmusson, Justinas Besusparis, Ruta Barbora Valkiuniene, Renaldas Augulis, Aida Laurinaviciene, Benoit Plancoulaine, Linas Petkevicius, and Arvydas Laurinavicius. Intratumoral heterogeneity of ki67 proliferation index outperforms conventional immunohistochemistry prognostic factors in estrogen receptor-positive her2-negative breast cancer. *Virchows Archiv*, pages 1–12, 2024. 1