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

Supplementary Material

1. 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.

1.1. 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.

2. 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. 1 and 2, 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 1, 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.

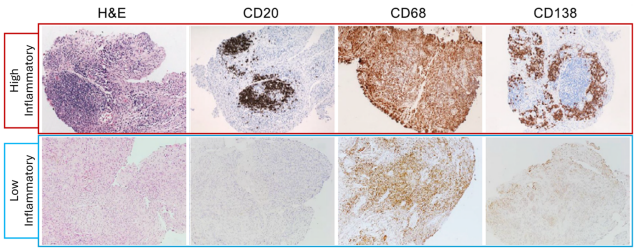

Figure 1. **Example of low inflammatory vs high inflammatory pathotype presentation in H&E and IHC stains for RA**: Rheumatoid Arthritis inflammatory pathotypes based on semi-quantitative analysis of synovial tissue biopsies stained with H&E, CD20+ B cells, CD68+ macrophages and IHC+ CD138 plasma cells.

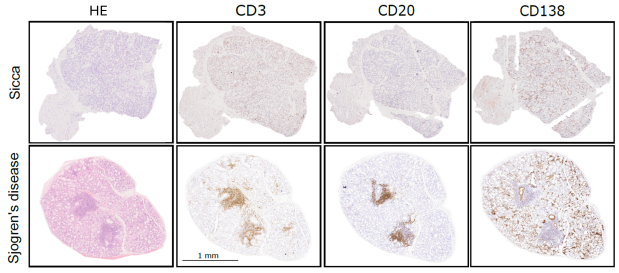

Figure 2. **Example of Sicca vs Sjogren presentation in H&E and IHC stains**: On top, a patient diagnosed with Sicca, on bottom a patient diagnosed with Sjogren. Here we show samples stained with IHC stains CD3+ T cells, CD20+ B cells and CD138+ plasma cells.

Table 1. **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

3. 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 2.

Table 2. **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

4. Memory Usage

Table 3 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 [3] 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 3. **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 [3]	38.11	**0.09**	32.93	**0.09**
CLAM-SB [7]	44.38	0.14	45.03	0.10
TransMIL [8]	**35.87**	1.47	**29.39**	0.79
DeepGraphConv [6]	55.65	1.31	45.10	0.68
Patch-GCN [1]	41.03	7.42	41.99	4.37
GTP [9]	47.11	2.40	48.15	1.97
MUSTANG [2]	36.00	6.18	36.25	3.52
BioX-CPath (ours)	41.30	11.25	36.61	9.19

5. 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 patient 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 6. 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.

6. 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 3 and 4.

6.1. 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 [5].

6.2. 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 [4]. 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.

7. GNN Heatmaps

In Fig. 5, 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

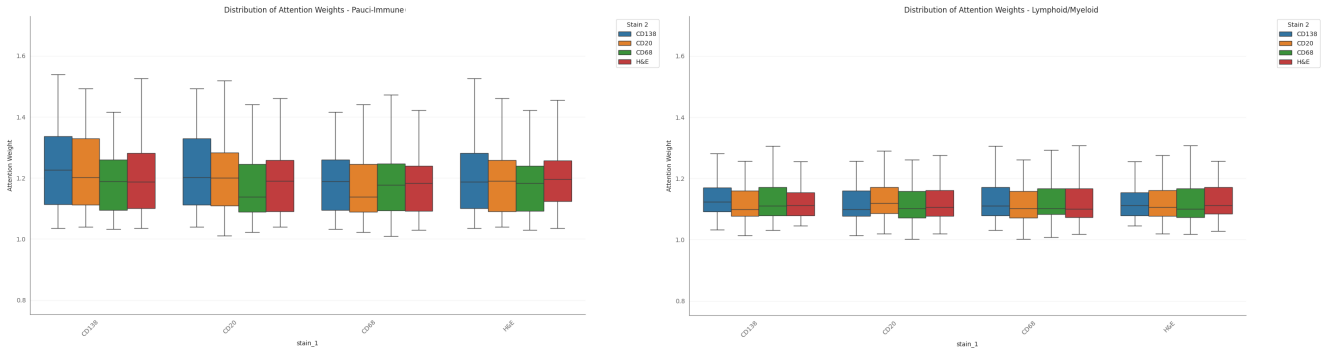

Figure 3. **Distribution of stain-to-stain interaction** scores for Pauci-Immune (Label 0, left) and Lymphoid/Myeloid (Label 1, right) cases. Each subplot shows the distribution of the average stain-stain attention scores for each stain pair (CD138, CD20, CD68, and H&E) interact with each other. For each source stain (x-axis), the box plots represent the distribution of interaction scores given to each target stain (colored boxes).

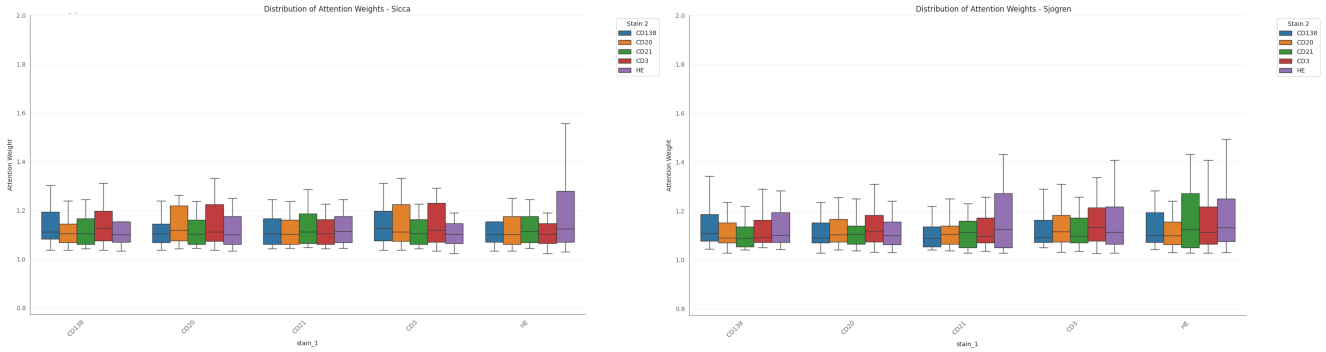

Figure 4. **Distribution of stain-to-stain interaction** scores for Sicca (Label 0, left) and Sjogren (Label 1, right) cases. Each subplot shows how different stains (CD138, CD20, CD21, CD3, and HE) interact with each other. For each source stain (x-axis), the box plots represent the distribution of interaction scores given to each target stain (colored boxes).

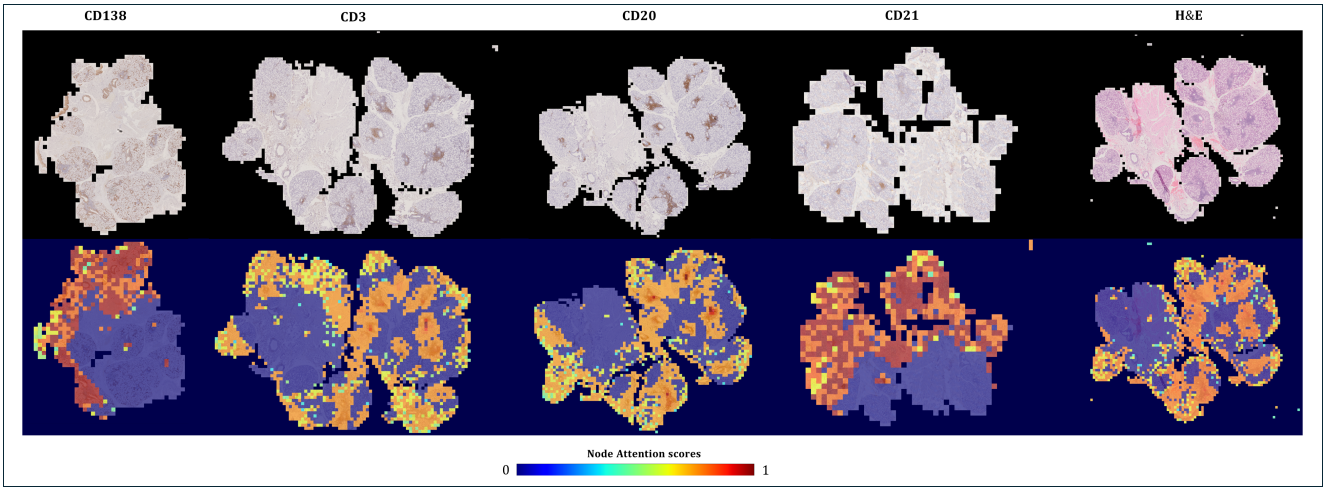

Figure 5. **Cumulative GNN node attention heatmap** obtained for a Sjogren positive patient with a stack of WSIs consisting of staining for CD138, CD20, C21, CD3 and H&E, where the red edges connecting across and correspond to region adjacency connectivity and the blue edges to the feature space connectivity. 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.

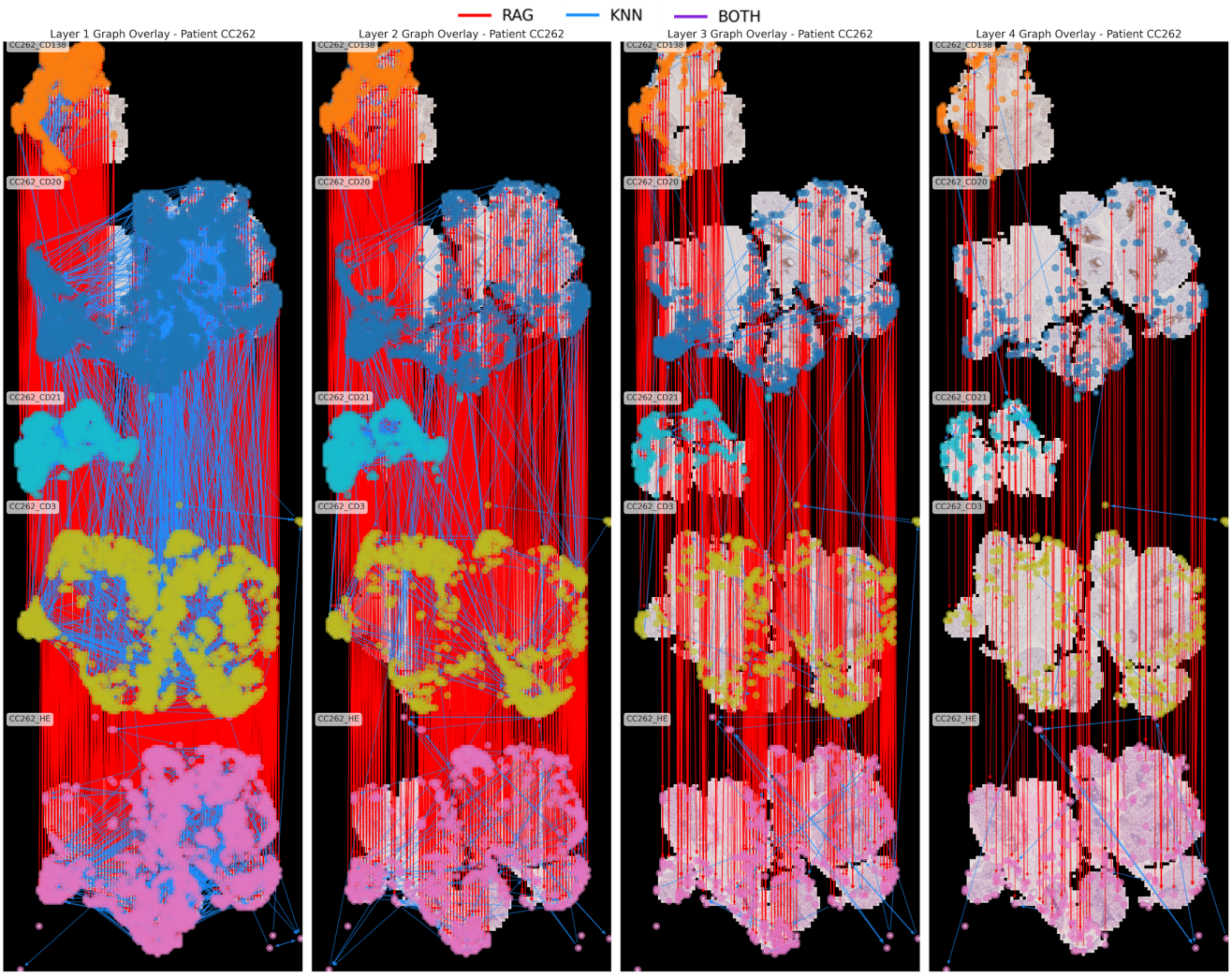

Figure 6. **Sparsification of input G_{FRA} through the GNN layers.** We plot the multistain patient input graph G_{FRA} as a spatial overlay on the stack of WSIs, to exemplify the connectivity both across and in the images. Edges connect nearest neighbors in both feature (blue) and region adjacent (red) space, with edges which are both feature and region nearest neighbors shown as purple.

graph sparsification process through our model, Figure 6 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.

8. Layer Importance

We previously mentioned we chose to maintain a MHSA layer before the classification head in our model architecture, despite seeing a marginal performance drop in. This is because we considered it was a good trade-off with obtained additional insight into the 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

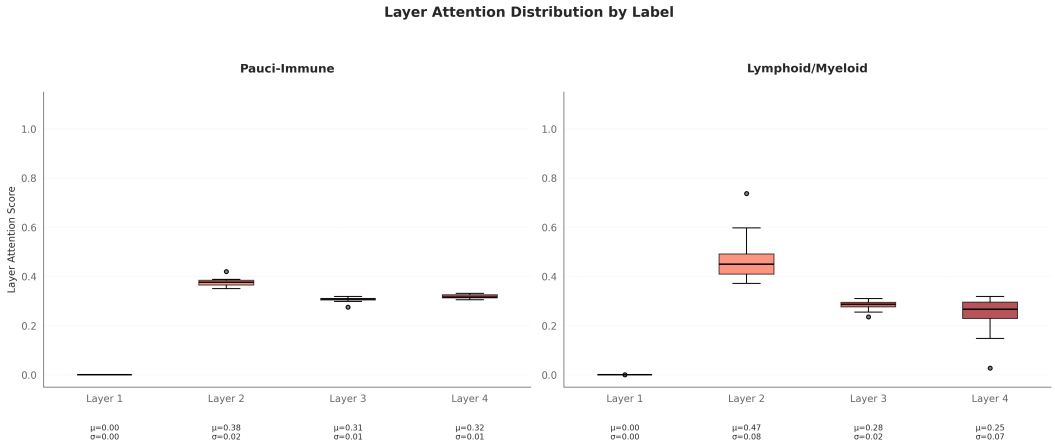

Figure 7. **Layer-wise attention patterns by label** in the hierarchical graph patient encoder, showing the distribution of attention scores across layers (1-4) for Pauci-Immune and Lymphoid/Myeloid cases, with corresponding mean (μ) and standard deviation (σ) values.

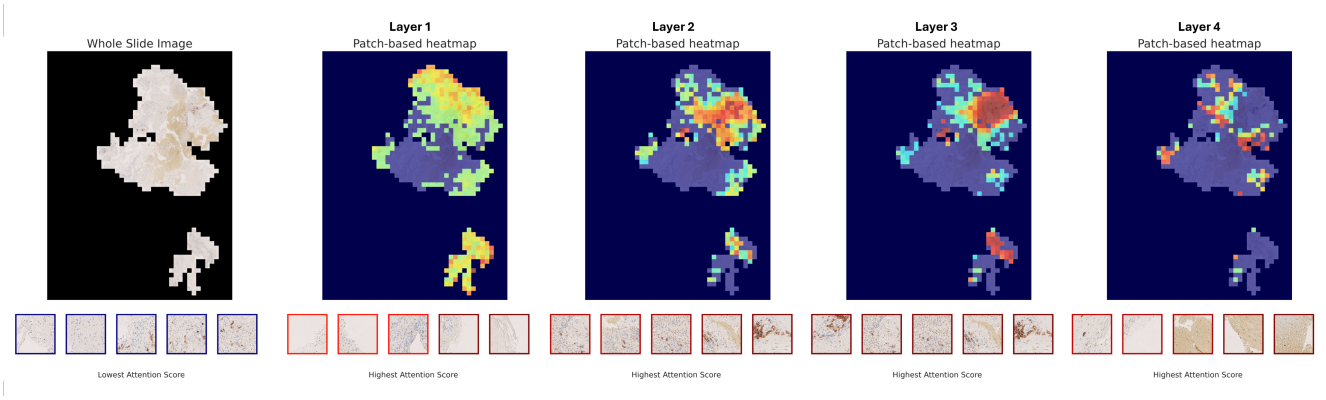

Figure 8. **Layer-wise attention visualization for a CD18-stained WSI Lymphoid/Myeloid RA patient.** The heatmaps show progression from broad attention in Layer 1 to increasingly focused attention in subsequent layers, with Layer 2 exhibiting the strongest patterns, consistent with quantitative attention scores. Bottom panels show highest and lowest attention patches, revealing cellular infiltrates in high-attention regions.

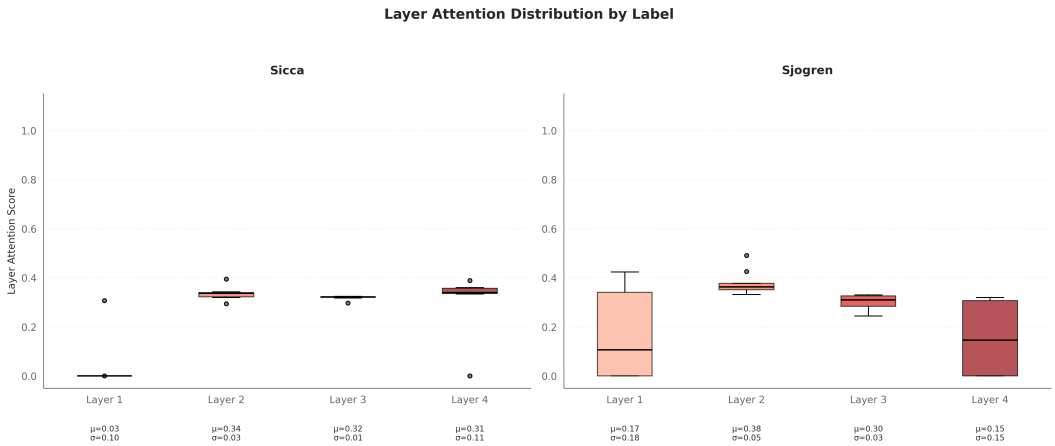

Figure 9. **Layer-wise attention patterns by label** in the hierarchical graph patient encoder, showing the distribution of attention scores across layers (1-4) for Sicca and Sjogren cases, with corresponding mean (μ) and standard deviation (σ) values.

graph patient encoder. This concatenated readout vector is the input to our 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 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 7 and 9.

8.1. 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 8 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 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.

8.2. 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] 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. 2

[2] 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

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

[4] 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

[5] 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. 3

[6] 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

[7] 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

[8] 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

[9] 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