

# Multi-Stain Self-Attention Graph Multiple Instance Learning Pipeline for Histopathology Whole Slide Images

Amaya Gallagher-Syed<sup>1, 2</sup>

a.r.syed@qmul.ac.uk

Luca Rossi<sup>3</sup>

luca.rossi@polyu.edu.hk

Felice Rivellese<sup>2</sup>

f.rivellese@qmul.ac.uk

Costantino Pitzalis<sup>2</sup>

c.pitzalis@qmul.ac.uk

Myles Lewis<sup>2</sup>

myles.lewis@qmul.ac.uk

Michael Barnes<sup>1, 2</sup>

m.r.barnes@qmul.ac.uk

Gregory Slabaugh<sup>2</sup>

g.slabaugh@qmul.ac.uk

<sup>1</sup> Digital Environment Research Institute  
Queen Mary University of London  
London, UK

<sup>2</sup> William Harvey Research Institute  
Queen Mary University of London  
London, UK

<sup>3</sup> Department of Electronic and  
Information Engineering  
The Hong Kong Polytechnic University  
Hong Kong

## Abstract

Whole Slide Images (WSIs) present a challenging computer vision task due to their gigapixel size and presence of numerous artefacts. Yet they are a valuable resource for patient diagnosis and stratification, often representing the gold standard for diagnostic tasks. Real-world clinical datasets tend to come as sets of heterogeneous WSIs with labels present at the patient-level, with poor to no annotations. Weakly supervised attention-based multiple instance learning approaches have been developed in recent years to address these challenges, but can fail to resolve both long and short-range dependencies. Here we propose an end-to-end multi-stain self-attention graph (MUSTANG) multiple instance learning pipeline, which is designed to solve a weakly-supervised gigapixel multi-image classification task, where the label is assigned at the patient-level, but no slide-level labels or region annotations are available. The pipeline uses a self-attention based approach by restricting the operations to a highly sparse  $k$ -Nearest Neighbour Graph of embedded WSI patches based on the Euclidean distance. We show this approach achieves a state-of-the-art F1-score/AUC of 0.89/0.92, outperforming the widely used CLAM model [24]. Our approach is highly modular and can easily be modified to suit different clinical datasets, as it only requires a patient-level label without annotations and accepts WSI sets of different sizes, as the graphs can be of varying sizes and structures. The source code can be found at <https://github.com/AmayaGS/MUSTANG>.

# 1 Introduction

In recent years, deep learning techniques have become the preferred methodology for analysing medical images, particularly in histology image classification [1]. Histopathology slide analysis is a time-consuming task that requires a qualified histopathologist. However, the digitisation of histopathology slides into Whole Slide Images (WSIs) has made it possible to automate analysis using deep learning techniques for routine workflows like subtyping, grading, and localising Regions of Interest (ROI). This automation accelerates analysis, enhances inter-observer concordance by offering a consistent benchmark, and can help overcome limitations due to restricted access to qualified professionals [2].

WSIs are large gigapixel multi-resolution image files, which present several challenges such as their large size, high memory requirements, and heterogeneity of artefacts (variations in staining intensity, scanner used, pen marks, etc). Since WSIs can be as large as  $100k \times 100k$  pixels, relevant information may be localised in small regions of the image (such as micro-tumours or Ectopic Lymphoid Structures (ELS)), but can also depend on interactions between far away parts (macro-tissue architecture) [3]. Labels are often only present at the slide or patient-level (represented by a set of WSIs), which can make it very challenging to develop classification models which accurately capture both micro and macro information.

# 2 Related Work

**Multiple Instance Learning.** Given the gigapixel size and heterogeneity of WSIs, they present a challenging computer vision task, with many successful deep learning methods rendered computationally intractable. Most approaches use variations on the weakly supervised Multiple Instance Learning (MIL) algorithm, where the gigapixel image is divided into a set of smaller patches (e.g.  $224 \times 224$  pixels), inheriting noisy slide/patient labels. Patches are then embedded into a feature vector and classified at the slide/patient level using some form of non-trainable global (e.g. max or mean) pooling on the set of instances [4]. This weakly supervised learning approach has yielded clinical grade performance, despite low instance level accuracy [5, 6, 7]. Furthermore, these approaches fail at capturing long-range dependencies and in learning which regions are most relevant to the final classification.

**Attention.** Recent methods have introduced linear attention pooling layers [8] and clustering-constrained attention pooling [9], which replaces non-trainable global pooling by a trainable weighted average aggregation layer where weights are given by a two-layered neural network [10]. This approach has demonstrated high performance on publicly available datasets and crucially provides information on key instances, permitting heatmap visualisation of the attention weights associated with each image patch [9].

**Self-attention.** In [11] and [12] the attention pooling layer measures the importance of a patch embedding given the whole sequence of patch embeddings. In contrast, self-attention measures the importance of a given instance compared to all other pairwise instances, which can provide a better understanding of long-range dependencies [13]. However, applying self-attention to the full sequence of embedded patches is computationally infeasible as the complexity grows quadratically to sequence length  $O(N^2)$ , rendering both runtime and memory

usage problematic [65]. Several approaches have been developed to improve the computational complexity of self-attention, such as relying on restricting the number of self-attention operations by inducing sparsity or by reformulating the problem via matrix factorisation for example [8, 24, 83, 85]. Sparse attention methods reduce the complexity by only considering a subset of the computations in the  $N \times N$  self-attention matrix  $P$ . In matrix factorisation approaches,  $P$  is assumed to be of low rank, meaning not all elements of the matrix are linearly independent of each other [85]. Intuitively, the idea that WSIs patches are not independent of each other is reasonable, suggesting that restricting self-attention operations via induced sparsity could achieve good results. The question then becomes how to restrict operations to subsets of relevant patches, while incurring minimal information loss. For example in [24] finding the nearest neighbours of tokens in high-dimensional space is achieved via locality-sensitive hashing.

**Graph Neural Networks.** Graph Neural Networks (GNNs) are capable of learning hierarchical representations of graphs by propagating node features through a series of message-passing and aggregation operations. 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 neighbours of  $u$ , i.e.

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

where the neighbourhood  $\mathcal{N}(u)$  is defined as the set of nodes that share an edge with  $u$ , UPDATE and AGGREGATE are arbitrary differentiable functions, and  $\mathbf{m}_{\mathcal{N}(u)}$  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  $\mathcal{N}(u)$  [18].

**Graphs in Histopathology.** Although Convolutional Neural Networks (CNNs) have shown impressive performance in histopathology analysis, they are less able to capture complex neighbourhood information as they analyse local areas determined by the size of the convolutional kernel [9]. GNNs can better exploit these irregular relationships by preserving neighbouring information and are thus ideally suited to representing relational information despite a graph structure not being explicitly present in the data [8, 19, 28, 37]. As presented in [9], applications in histopathology can be divided into cell, patch or tissue-level graphs, with both node and graph classification approaches being employed. In particular, the MIL problem can be reformulated as a patch-graph with graph-level classification, where GNN layers are combined with pooling and readout layers to produce an end-to-end framework. Patch-graphs can be constructed using features extracted from a WSI or a set of WSIs, and then connected via edges. This has been done by connecting selected or spatially adjacent patches, as well as “super-pixels” [11, 9, 37, 83]. However, in real-world settings, information across multiple WSIs is integrated by the pathologist for the purpose of disease diagnosis and subtyping [10]. Thus, combining the relevant information from multiple images more accurately reproduces the pathology pipeline. This only complexifies the MIL problem, as the labels are at the patient and not the slide level. Furthermore, large bodies of WSIs coming from clinical trials have few to no annotations, yet still present a valuable source of information. Histopathology deep learning frameworks therefore need to extend to real-world datasets with little curation and very noisy labels.

## 2.1 Contributions

- Here we propose a novel end-to-end Multi-STain self-AtteNtion Graph (MUSTANG) multiple instance learning pipeline. MUSTANG solves a weakly-supervised gigapixel multi-image classification task, where the label is assigned at the patient-level (across the multiple images), but no slide-level labels or region annotations are available.
- The pipeline introduces a self-attention based approach on multiple full gigapixel WSIs by restricting the attention operations to a highly sparse  $k$ -Nearest Neighbour Graph ( $k$ -NNG) of embedded WSI patches based on Euclidean distance.
- Importantly, our approach does not require registration of WSIs, preprocessing or labelling of ROIs, nor any feature engineering for the embedded feature vectors, making it straightforward and flexible to apply to real-world clinical datasets.

## 3 Methods

### 3.1 Dataset

**Rheumatoid Arthritis.** MUSTANG is designed for the Rheumatoid Arthritis R4RA clinical trial dataset [20, 53]. 20 European centres recruited a total of 164 patients who underwent ultrasound-guided synovial biopsy of a clinically active joint. The synovial tissue samples were then stained with Hematoxylin & Eosin (H&E) and Immunochemistry (IHC) stains. H&E provides information on overall tissue architecture and composition, while IHC identifies specific immune cells, such as CD20+ B cells, CD68+ macrophages and CD138+ plasma cells [20]. Each dye contains complimentary information about the underlying disease process. Pathologists then semi-quantitatively assigned patients to one of three groups: Fibroid, Myeloid, and Lymphoid, each corresponding to a disease subtype linked to drug response and patient trajectory [10, 27]. The different stain types and disease pathotypes can be seen in Figure 1. For the purpose of this study, the pathotypes Fibroid and Myeloid are aggregated and compared to the Lymphoid pathotype, as the latter more substantially differs in disease presentation, trajectory, and drug response. Samples were scanned into WSIs with .ndpi format with Hamamatsu digital scanners under 20x objectives. The dataset has a total of 651 WSIs, with a variable number of WSIs per patient.

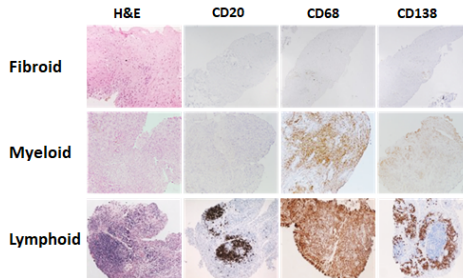


Figure 1: Rheumatoid Arthritis inflammatory pathotypes Fibroid, Myeloid & Lymphoid based on semi-quantitative analysis of synovial tissue biopsies stained with H&E, CD20+ B cells, CD68+ macrophages and IHC+ CD138 plasma cells [20].

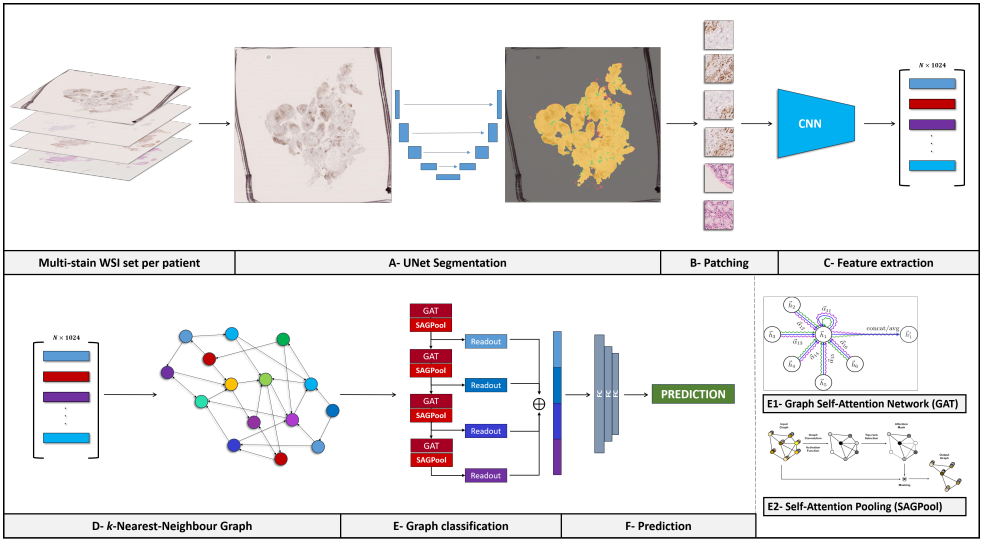


Figure 2: MUSTANG pipeline. Schemas in E1 and E2 are reproduced from [76, 34]

### 3.2 MUSTANG

The MUSTANG pipeline, which is graphically represented in Figure 2 is composed of:

- **A - Segmentation:** A automated segmentation step, where UNet is used to segment tissue areas on the WSIs. The user can use the trained weights provided on our GitHub repository or use their own.
- **B - Patching:** After segmentation, the tissue area is divided into patches at a size chosen by the user, which can be overlapping or non-overlapping.
- **C - Feature extraction:** Each image patch is passed through a VGG16 CNN feature extractor and embedded into a  $[1 \times 1024]$  feature vector. All feature vectors from a given patient are aggregated into a matrix. The number of rows in the matrix will vary as each patient has a variable set of WSIs, each with their own dimensions.
- **D -  $k$ -Nearest-Neighbour Graph:** The matrix of feature vectors of each patient is used to create a sparse directed  $k$ -NNG using the Euclidean distance metric, with a default of  $k = 5$ . The attribute of each node corresponds to a  $[1 \times 1024]$  feature vector. This graph is used as input to the GNN.
- **E - Graph classification:** The  $k$ -NNG is successively passed through four Graph Attention Network layers (GAT) [34] and SAGPooling layers [76]. The SAGPooling readouts from each layer are concatenated and passed through three MLP layers and finally classified.
- **F - Prediction:** A pathology or diagnosis prediction is obtained at the patient-level.

**Segmentation and patch extraction.** From the 651 WSIs, a total 309,248 non-overlapping  $224 \times 224$  pixel patches were extracted at 10x magnification from the tissue area segmented by UNet [13]. The 10x magnification was chosen based on our domain knowledge of RA as a compromise to show both the macro/micro-architecture of the tissue and to reduce the number of patches for storage and computation [13].

**$k$ -Nearest Neighbour Graph.** The  $k$ -NNG is a directed graph (digraph) in which node  $p$  is connected by a directed edge to node  $q$ , if  $q$  is among the  $k$  closest nodes to  $p$ . Here we measure the distance between two nodes in terms of the Euclidean distance between the corresponding embedded feature vectors. Notice that not all connections are reciprocal because if node  $p$  is  $q$ 's closest neighbour, it does not follow that  $q$  is also  $p$ 's closest neighbour. A digraph is weakly connected if there exists a path between every pair of nodes in the underlying undirected graph.

**Graph Attention Network.** Graph Attention Networks (GATs), based on the self-attention mechanism [4, 63], incorporate masked self-attention layers into graph convolutions and use attention weights to define a weighted sum of the neighbours:

$$\mathbf{m}_{\mathcal{N}(u)} = \sum_{v \in \mathcal{N}(u)} \alpha_{u,v} \mathbf{h}_v, \quad (2)$$

where  $\alpha_{u,v}$  denotes the attention on neighbour  $v \in \mathcal{N}(u)$  when aggregating information at node  $u$ . In the original GAT paper, the attention weights are defined as:

$$\alpha_{u,v} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^\top [\mathbf{W}\mathbf{h}_u \oplus \mathbf{W}\mathbf{h}_v]))}{\sum_{v' \in \mathcal{N}(u)} \exp(\text{LeakyReLU}(\mathbf{a}^\top [\mathbf{W}\mathbf{h}_u \oplus \mathbf{W}\mathbf{h}_{v'}]))}, \quad (3)$$

where  $\mathbf{a}$  is a trainable attention vector,  $\mathbf{W}$  is a trainable matrix and  $\oplus$  denotes the concatenation operation. To stabilise the learning process multi-head attention can be used, where  $m$  different attention heads are applied to compute mutually independent features in parallel, and their features are then averaged.

**Graph Attention Pooling.** Graph pooling is used to downsample a graph, reducing its size while seeking to preserve topological information. Global pooling methods use max or mean aggregation layers to pool all the representations of nodes in each layer, which enables graphs of different structures to be processed but tends to lose topological information. Similarly, hierarchical methods such as DiffPool [36] or gPool [45], which use a learnable vector to calculate projection scores and select the top-ranked nodes, do not fully take into account graph topology [9, 14, 26]. SAGPool [26] uses the GCN defined in [23] to calculate the self-attention scores  $Z \in \mathbb{R}^{N \times 1}$  as follows:

$$Z = \sigma\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} X \Theta_{att}\right), \quad (4)$$

where  $\sigma$  is the activation function,  $\tilde{A} \in \mathbb{R}^{N \times N}$  is the adjacency matrix with self-connections,  $\tilde{D} \in \mathbb{R}^{N \times N}$  is the degree matrix of  $\tilde{A}$ ,  $X \in \mathbb{R}^{N \times F}$  is the matrix of input features of the graph with  $N$  nodes and  $F$ -dimensional features, and  $\Theta_{att} \in \mathbb{R}^{F \times 1}$  is the only parameter of the SAGPool layer [26]. By utilizing graph convolution to obtain self-attention scores, the result of the pooling is based on both graph and topological features. The node selection method follows [9, 14, 25] by retaining a portion of nodes of the input graph even when graphs of varying sizes and structures are input. The pooling ratio  $k \in (0, 1]$  is a hyperparameter that determines the number of nodes to keep. The top  $\lceil kN \rceil$  nodes are selected based on the value of  $Z$ , i.e.

$$\text{idx} = \text{top-rank}(Z, \lceil kN \rceil), \quad Z_{\text{mask}} = Z_{\text{idx}}, \quad (5)$$

where  $\text{top} - \text{rank}$  is the function that returns the indices of the top  $\lceil kN \rceil$  values,  $\text{idx}$  is an indexing operation, and  $Z_{\text{mask}}$  is the feature attention mask. Finally, the readout layer, adopted from [9], aggregates node features to make a fixed size representation. The summarized output feature of the readout layer is

$$s = \frac{1}{N} \sum_{i=1}^N x_i \oplus \max_{i=1}^N x_i, \quad (6)$$

where  $N$  is the number of nodes,  $x_i$  is the feature vector of the  $i$ -th node, and  $\oplus$  denotes concatenation [9].

**Benchmarking and ablation studies.** We benchmark our method against CLAM [24], which is a clustering-constrained gated attention based MIL method, widely used within the histopathology community. We modify the framework slightly to accommodate the patient WSI set’s DataLoader and use VGG16 instead of ResNet50 for feature extraction. We also try several different GNN architectures using combinations of GAT or GCN, with SAGPooling or TopKPooling (GCN\_SAG, CGN\_TopK, GAT\_SAG), to assess the importance of each component. Finally, we look at F1-scores for different values of  $k$  in the  $k$ -NNG.

**Training schedule.** This is a binary classification task with weakly-supervised labels. To obtain our results we use a 70/30 train/test split and train with the Adam optimizer  $\beta_1 = 0.9$ ,  $\beta_2 = 0.98$  and  $\epsilon = 10^{-9}$ . We use the default parameters recommended for CLAM [24]:  $lr_{\text{CLAM}} = 0.0001$ , with no dropout. We train MUSTANG with  $lr_{\text{MUSTANG}} = 0.0001$ ,  $\text{pooling ratio}=0.8$ ,  $\text{attention heads}=2$ . We train each method for 50 epochs and keep the best F1-score, whilst checking the loss is stable (i.e. that we are not underfitting). The training was prototyped locally on a commercial workstation with NVidia GPU RTX3080 and trained with QMUL’s Apocrita HPC facility on an NVidia A100 GPU supported by QMUL Research-IT [22].

## 4 Experimental Results

**Multi-stain WSIs.** In Table 1 we present the results obtained by both MUSTANG and CLAM on the multi-stain R4RA Rheumatoid Arthritis test set. MUSTANG outperforms the benchmark by 5 percentage points for both F1-score and AUC and runtime on inference is substantially similar to that of CLAM, despite having a larger number of total parameters and relying on self-attention operations. In Figure 3, we show MUSTANG performs well at identifying both correct and true positives (Sensitivity=0.93), an important consideration for healthcare.

|                | F1-score    | AUC         | Sens        | Spec | Params [M] | Test runtime [min] |
|----------------|-------------|-------------|-------------|------|------------|--------------------|
| CLAM           | 0.84        | 0.88        | 0.86        | 0.82 | 0.47       | 10                 |
| MUSTANG (ours) | <b>0.89</b> | <b>0.92</b> | <b>0.93</b> | 0.82 | 3.29       | 11                 |

Table 1: Multi-stain F1-score, AUC, Sensitivity (Sens) and Specificity (Spec) results, total parameter number, and test runtime results.

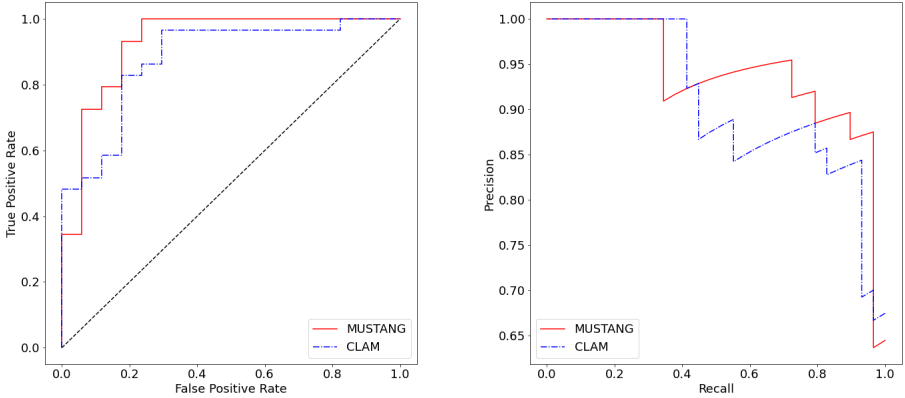


Figure 3: Area Under the ROC and Precision-Recall curves for both MUSTANG (full line) and CLAM (dotted line).

|  |                | Single Stain |             |             |             |
|--|----------------|--------------|-------------|-------------|-------------|
|  |                | CD138        | CD68        | CD20        | HE          |
|  | CLAM           | 0.85         | 0.87        | <b>0.88</b> | 0.76        |
|  | MUSTANG (ours) | <b>0.89</b>  | <b>0.89</b> | 0.87        | <b>0.78</b> |

Table 2: Single-stain F1-score results

**Single-stain WSIs.** To understand if our model is able to outperform single-stain F1-score, we propagate the patient WSIs set label to each slide and rerun the method on the single-stain. The results are presented in Table 2. MUSTANG outperforms CLAM, except for CD20+ staining, and obtains similar accuracy results to the multi-stain problem with 89% F1-score for CD138 and CD68. H&E is the worst performing stain in both cases, suggesting any IHC staining performed in the clinic would be beneficial for patient pathology assignment. CLAM performs better on the single-stain problem than on the multi-stain one, confirming it does not fully capture long-range dependencies as the patient matrix size increases. MUSTANG does not show increased performance compared to the single-stain problem, but successfully integrates information across stains, identifying complex relations and spatial arrangements pertaining to disease subtyping. This is valuable because we do not know a priori which stains contain the most information about the disease pathology, as most clinical datasets have not received curation.

**GNN model ablation.** In Table 3 we check which component parts of the GNN model provide the most accuracy gain. From the high accuracy of the GAT/GCN + SAGPool models, it is clear that the self-attention pooling topology preserving method is crucial in aggregating graph topological information which preserves the relations between nodes.

**$k$ -NNG ablation on  $k$ .** In Figure 4 we show the  $k$ -NNG layout for different values of  $k$ . Up to  $k = 4$ , the graph is not a weakly connected digraph, meaning that it has isolated nodes and



| GNN Model   | GAT + SAGPool | GAT + TopK | GCN + SAGPool | GCN + TopK |
|-------------|---------------|------------|---------------|------------|
| Multi Stain | <b>0.89</b>   | 0.67       | <b>0.82</b>   | 0.76       |

Table 3: GNN model ablation F1-score results

subgraphs. For message-passing purposes, we reason that the initial graph should be weakly-connected in order to obtain better results. We check this assumption running MUSTANG for several values of  $k$  and present the results in Table 4. The  $k$ -NNG with  $k = 5$  obtains the highest score, with  $k = 1$  the lowest.  $k = 2$  already shows good accuracy, evidencing even highly sparse poorly connected graphs suffice to extract relevant signals. Higher values of  $k$  also obtain good results, but a denser graphs comes at the cost of higher memory requirements and the potential to over-smooth the signal, suggesting that using a minimally weakly connected graph is a good strategy. In Figure 5 we show how the  $k$ -NNG graph connectivity structure changes after each GAT + SAGPool layer in the GNN model: after each layer the graph loses structure, restricting message passing to increasingly many small subgraphs. We posit the early layers aggregate macro-tissue topological information while the later layers concentrate on micro-tissue information, suggesting that graph disaggregation is not a phenomenon to avoid.

Figure 4:  $k$ -NNG layout for different values of  $k$ . From left to right,  $k = 1$ ,  $k = 2$ ,  $k = 3$ ,  $k = 4$  and  $k = 5$ .Figure 5:  $k$ -NNG graph connectivity layout after each GAT + SAGPool layer, with pooling ratio of 0.5

| $k$      | 1    | 2    | 3    | 4    | 5           | 10   | 20   | 50   | 100  |
|----------|------|------|------|------|-------------|------|------|------|------|
| Accuracy | 0.71 | 0.85 | 0.87 | 0.87 | <b>0.89</b> | 0.83 | 0.87 | 0.85 | 0.87 |

Table 4:  $k$ -NNG model  $k$  ablation F1-score results

**Graph connectivity.** We empirically check graph connectivity to see what each node is connecting to in Figure 6. On the left, we show a set of WSIs for a single patient and on the right the corresponding  $k$ -NNG ( $k = 5$ ) graph structure plotted using Networkx spring-layout [14], with nodes coloured in function of their provenance. The force-directed spring layout uses the Fruchterman-Reingold algorithm, where edges act as “springs” and nodes

repel each other, hence resulting in closely connected nodes clustering together [12]. As expected nodes tend to connect to other nodes in the same WSI, but there is a good degree of mixing between WSIs, indicating information can flow between them.

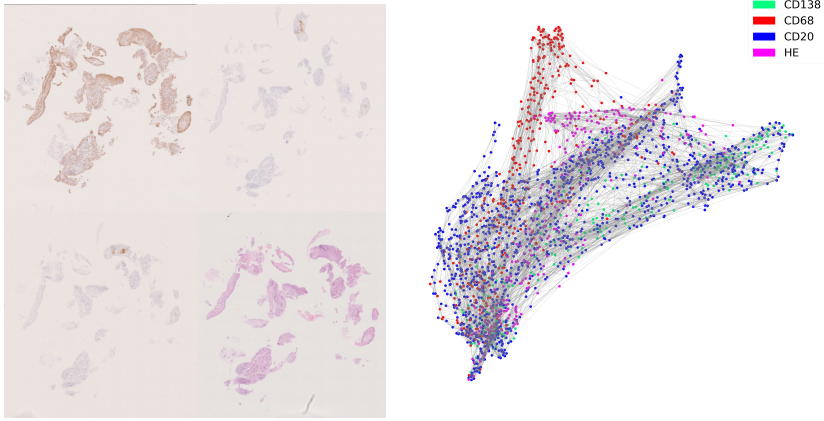


Figure 6: On left a set of WSI corresponding to a patient and on right its corresponding  $k$ -NNG ( $k = 5$ ) plotted using Networkx’s spring-layout.

**Limitations.** Ideally we would apply MUSTANG to publicly available datasets for comparison purposes. However, we are not aware of a public dataset with multi-stain WSI sets labelled at the patient-level, despite this being a common histopathology pipeline. We address this limitation by comparing our method against CLAM, which has strong performance on a variety of public datasets. Another limitation is that we have not yet developed a visualisation of heatmap overlays indicating the regions of interest that the model attends to. Visualisation is a crucial step for the translation of these tools into clinical practice and which remains as future work.

**Supplementary material.** Our supplementary material has additional results and analysis.

## 5 Conclusion

We introduced a novel end-to-end MULTI-STain self-AtteNtion Graph Multiple Instance Learning pipeline, which we call MUSTANG. The pipeline employs a self-attention based approach on multiple full gigapixel Whole Slide Images by restricting the attention operations to a highly sparse  $k$ -NNG of embedded WSI patches based on Euclidean distance. We show this approach achieves state-of-the-art accuracy, outperforming the widely used CLAM model. This shows that when data is highly correlated, applying self-attention operations on a very sparse matrix is sufficient to integrate both long-range dependencies and local behaviour. This approach is highly modular and can easily be modified to suit different clinical datasets, as it only requires a patient-level label without any WSI annotations and accepts WSI sets of different sizes, as the graphs can be of varying sizes and structures.

## Acknowledgements

This work was supported with funding from the Wellcome Trust (grant no. 218584/Z/19/Z). The R4RA trial was funded by the Efficacy and Mechanism Evaluation (EME) Programme, a partnership between the Medical Research Council (MRC) and the National Institute for Health and Care Research (NIHR) (grant no. 11/100/76). This work acknowledges the support of the National Institute for Health Research Barts Biomedical Research Centre (NIHR203330).

## References

- [1] 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, November 2012. ISSN 1939-3539. doi: 10.1109/TPAMI.2012.120.
- [2] 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, June 2020. doi: 10.1109/CVPRW50498.2020.00502. ISSN: 2160-7516.
- [3] David Ahméd-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. A survey on graph-based deep learning for computational histopathology. *Computerized Medical Imaging and Graphics*, 95: 102027, January 2022. ISSN 0895-6111. doi: 10.1016/j.compmedimag.2021.102027. URL <https://www.sciencedirect.com/science/article/pii/S0895611121001762>.
- [4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. Technical report, May 2016. URL <http://arxiv.org/abs/1409.0473>. arXiv:1409.0473 [cs, stat] type: article.
- [5] Cătălina Cangea, Petar Veličković, Nikola Jovanović, Thomas Kipf, and Pietro Liò. Towards Sparse Hierarchical Graph Classifiers. Technical report, November 2018. URL <http://arxiv.org/abs/1811.01287>. arXiv:1811.01287 [cs, stat] type: article.
- [6] Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, Tiffany Y. Chen, Drew F.K. Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4025, October 2021.
- [7] 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, June 2022. doi: 10.1109/CVPR52688.2022.01567. ISSN: 2575-7075.

- [8] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating Long Sequences with Sparse Transformers. Technical report, April 2019. URL <http://arxiv.org/abs/1904.10509>. arXiv:1904.10509 [cs, stat] type: article.
- [9] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyő, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, October 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0177-5. URL <https://www.nature.com/articles/s41591-018-0177-5>.
- [10] 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. ISSN 1478-6354. doi: 10.1186/ar4555. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4060385/>.
- [11] 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. URL [https://openaccess.thecvf.com/content/CVPR2022W/CVMI/html/Dwivedi\\_Multi\\_Stain\\_Graph\\_Fusion\\_for\\_Multimodal\\_Integration\\_in\\_Pathology\\_CVPRW\\_2022\\_paper.html](https://openaccess.thecvf.com/content/CVPR2022W/CVMI/html/Dwivedi_Multi_Stain_Graph_Fusion_for_Multimodal_Integration_in_Pathology_CVPRW_2022_paper.html).
- [12] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991. ISSN 1097-024X. doi: 10.1002/spe.4380211102. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.4380211102>.
- [13] Amaya Gallagher-Syed, Abbas Khan, Felice Rivellese, Costantino Pitzalis, Myles J. Lewis, Gregory Slabaugh, and Michael R. Barnes. Automated segmentation of rheumatoid arthritis immunohistochemistry stained synovial tissue. Technical report, September 2023. URL <http://arxiv.org/abs/2309.07255>. arXiv:2309.07255 [cs, eess, q-bio] type: article.
- [14] Hongyang Gao and Shuiwang Ji. Graph U-Nets. pages 2083–2092. PMLR, May 2019. URL <https://proceedings.mlr.press/v97/gao19a.html>.
- [15] Hongyang Gao, Yongjun Chen, and Shuiwang Ji. Learning Graph Pooling and Hybrid Convolutional Operations for Text Representations. In *The World Wide Web Conference*, WWW ’19, pages 2743–2749, New York, NY, USA, May 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313395. URL <https://doi.org/10.1145/3308558.3313395>.
- [16] Narmin Ghaffari Laleh, Hannah Sophie Muti, Chiara Maria Lavinia Loeffler, Amelie Echle, Oliver Lester Saldanha, Faisal Mahmood, Ming Y. Lu, Christian Trautwein, Rupert Langer, Bastian Dislich, Roman D. Buelow, Heike Irmgard Grabsch, Hermann Brenner, Jenny Chang-Claude, Elizabeth Alwers, Titus J. Brinker, Firas

- Khader, Daniel Truhn, Nadine T. Gaisa, Peter Boor, Michael Hoffmeister, Volkmar Schulz, and Jakob Nikolas Kather. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Medical Image Analysis*, 79:102474, July 2022. ISSN 1361-8415. doi: 10.1016/j.media.2022.102474. URL <https://www.sciencedirect.com/science/article/pii/S1361841522001219>.
- [17] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.
- [18] William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159.
- [19] Kai Han, Yunhe Wang, Jianyuan Guo, Yehui Tang, and Enhua Wu. Vision GNN: An image is worth graph of nodes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=htM1WJZVB2I>.
- [20] 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, January 2021. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(20)32341-2. URL [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(20\)32341-2/fulltext](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)32341-2/fulltext).
- [21] M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In *Proc. 35th ICML*, volume 80, pages 2127–2136, 2018. URL <https://proceedings.mlr.press/v80/ilse18a.html>.
- [22] Thomas King, Simon Butcher, and Lukasz Zalewski. Apocrita - High Performance Computing Cluster for Queen Mary University of London. March 2017. URL <https://zenodo.org/record/438045>.
- [23] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. Technical report, February 2017. URL <http://arxiv.org/abs/1609.02907>. arXiv:1609.02907 [cs, stat] type: article.
- [24] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The Efficient Transformer. Technical report, February 2020. URL <http://arxiv.org/abs/2001.04451>. arXiv:2001.04451 [cs, stat] type: article.

- [25] Boris Knyazev, Graham W Taylor, and Mohamed Amer. Understanding Attention and Generalization in Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper\\_files/paper/2019/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2019/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html).
- [26] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *Proceedings of the 36th International Conference on Machine Learning*, 09–15 Jun 2019.
- [27] 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, August 2019. ISSN 2211-1247. doi: 10.1016/j.celrep.2019.07.091. URL <https://www.sciencedirect.com/science/article/pii/S2211124719310071>.
- [28] Michelle M. Li, Kexin Huang, and Marinka Zitnik. Graph Representation Learning in Biomedicine. Technical report, June 2022. URL <http://arxiv.org/abs/2104.04883>. arXiv:2104.04883 [cs, q-bio] type: article.
- [29] 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, June 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w. URL <https://www.nature.com/articles/s41551-020-00682-w>.
- [30] D. Lucchesi, E. Pontarini, V. Donati, et al. The use of digital image analysis in the histological assessment of Sjögren’s syndrome salivary glands improves inter-rater agreement and facilitates multicentre data harmonisation. *Clin. and Exp. Rheumatology*, 38 Suppl 126(4):180–188, 2020. ISSN 0392-856X.
- [31] Felice Rivellese, Anna E. A. Surace, Katriona Goldmann, Elisabetta Sciacca, Cankut Çubuk, Giovanni Giorli, Christopher R. John, Alessandra Nerviani, Liliane Fossati-Jimack, Georgina Thorborn, Manzoor Ahmed, Edoardo Prediletto, Sarah E. Church, Briana M. Hudson, Sarah E. Warren, Paul M. McKeigue, Frances Humby, Michele Bombardieri, Michael R. Barnes, Myles J. Lewis, and Costantino Pitzalis. Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial. *Nature Medicine*, 28(6):1256–1268, June 2022. ISSN 1546-170X. doi: 10.1038/s41591-022-01789-0. URL <https://www.nature.com/articles/s41591-022-01789-0>.
- [32] Henry Senior, Gregory Slabaugh, Shanxin Yuan, and Luca Rossi. Graph neural networks in vision-language image understanding: A survey. *arXiv preprint arXiv:2303.03761*, 2023.

- [33] 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*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [34] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. Technical report, February 2018. URL <http://arxiv.org/abs/1710.10903>. arXiv:1710.10903 [cs, stat] type: article.
- [35] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-Attention with Linear Complexity. Technical report, June 2020. URL <http://arxiv.org/abs/2006.04768>. arXiv:2006.04768 [cs, stat] version: 1 type: article.
- [36] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical Graph Representation Learning with Differentiable Pooling. Technical report, February 2019. URL <http://arxiv.org/abs/1806.08804>. arXiv:1806.08804 [cs, stat] type: article.
- [37] 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, November 2022. ISSN 1558-254X. doi: 10.1109/TMI.2022.3176598.
- [38] 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, February 2022. ISSN 1361-8415. doi: 10.1016/j.media.2021.102308. URL <https://www.sciencedirect.com/science/article/pii/S1361841521003534>.