



## Masked hypergraph learning for weakly supervised histopathology whole slide image classification



Jun Shi<sup>a</sup>, Tong Shu<sup>b</sup>, Kun Wu<sup>c</sup>, Zhiguo Jiang<sup>c,d</sup>, Liping Zheng<sup>a</sup>, Wei Wang<sup>e,f</sup>, Haibo Wu<sup>e,f</sup>, Yushan Zheng<sup>g,\*</sup>

<sup>a</sup> School of Software, Hefei University of Technology, Hefei, 230601, Anhui Province, China

<sup>b</sup> School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230601, Anhui Province, China

<sup>c</sup> Image Processing Center, School of Astronautics, Beihang University, Beijing, 102206, China

<sup>d</sup> Tianmushan Laboratory, Hangzhou, 311115, Zhejiang Province, China

<sup>e</sup> Department of Pathology, the First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, 230036, Anhui Province, China

<sup>f</sup> Intelligent Pathology Institute, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, 230036, Anhui Province, China

<sup>g</sup> School of Engineering Medicine, Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing, 100191, China

### ARTICLE INFO

#### Keywords:

Computer-aided diagnosis

Computational pathology

Hypergraph learning

Weak supervision

Whole slide image classification

### ABSTRACT

**Background and objectives:** Graph neural network (GNN) has been extensively used in histopathology whole slide image (WSI) analysis due to the efficiency and flexibility in modelling relationships among entities. However, most existing GNN-based WSI analysis methods only consider the pairwise correlation of patches from one single perspective (e.g. spatial affinity or embedding similarity) yet ignore the intrinsic non-pairwise relationships present in gigapixel WSI, which are likely to contribute to feature learning and downstream tasks. The objective of this study is therefore to explore the non-pairwise relationships in histopathology WSI and exploit them to guide the learning of slide-level representations for better classification performance.

**Methods:** In this paper, we propose a novel Masked HyperGraph Learning (MaskHGL) framework for weakly supervised histopathology WSI classification. Compared with most GNN-based WSI classification methods, MaskHGL exploits the non-pairwise correlations between patches with hypergraph and global message passing conducted by hypergraph convolution. Concretely, multi-perspective hypergraphs are first built for each WSI, then hypergraph attention is introduced into the jointed hypergraph to propagate the non-pairwise relationships and thus yield more discriminative node representation. More importantly, a masked hypergraph reconstruction module is devised to guide the hypergraph learning which can generate more powerful robustness and generalization than the method only using hypergraph modelling. Additionally, a self-attention-based node aggregator is also applied to explore the global correlation of patches in WSI and produce the slide-level representation for classification.

**Results:** The proposed method is evaluated on two public TCGA benchmark datasets and one in-house dataset. On the public TCGA-LUNG (1494 WSIs) and TCGA-EGFR (696 WSIs) test set, the area under receiver operating characteristic (ROC) curve (AUC) were  $0.9752 \pm 0.0024$  and  $0.7421 \pm 0.0380$ , respectively. On the USTC-EGFR (754 WSIs) dataset, MaskHGL achieved significantly better performance with an AUC of  $0.8745 \pm 0.0100$ , which surpassed the second-best state-of-the-art method SlideGraph+ 2.64%.

**Conclusions:** MaskHGL shows a great improvement, brought by considering the intrinsic non-pairwise relationships within WSI, in multiple downstream WSI classification tasks. In particular, the designed masked hypergraph reconstruction module promisingly alleviates the data scarcity and greatly enhances the robustness and classification ability of our MaskHGL. Notably, it has shown great potential in cancer subtyping and fine-grained lung cancer gene mutation prediction from hematoxylin and eosin (H&E) stained WSIs.

\* Corresponding author.

E-mail address: [ysheng@buaa.edu.cn](mailto:ysheng@buaa.edu.cn) (Y. Zheng).

## 1. Introduction

Histopathology whole slide images (WSIs) provide rich biological information and are the gold standard for cancer diagnosis. Traditional pathological diagnosis relies mainly on pathologists to observe and judge the morphological features of tissues under the microscope with human eyes, which is highly professional, subjective and inefficient. Thanks to the development of digital WSI scanning technology, it has become possible to use computer-aided diagnosis (CAD) systems to assist pathologists in analysing histopathology images in an efficient, accurate and reproducible manner. Deep learning based histopathology image analysis has shown promising performance in cancer screening [1–5], tumour grading [6–8], prognosis analysis [9–12], gene mutation prediction [13–15], etc.

However, owing to the special properties of WSIs, deep learning based WSI classification methods face two main challenges: (1) the large size of WSIs makes it impossible to directly process gigapixel WSIs with conventional deep neural networks under hardware limitations; (2) fine-grained labelling is quite lacking and difficult to obtain, since pathological diagnosis is highly professional and time-consuming. To address these problems, most studies pre-divide WSI into massive patches as the smallest unit in WSI analysis and utilize multiple instance learning (MIL) to deal with the problem of annotation scarcity. MIL paradigm that takes patches as instances and WSIs as bags, tries to aggregate the final predictions [16] or intermediate representations [17,18] of the instances into bag-level representation via weakly supervised learning, and eventually makes slide-level and patch-level predictions with only coarse-grained WSI label supervision.

Graph-based WSI analysis is another important branch of the WSI classification method. Graph neural network (GNN) [19] could concurrently achieve local and global message passing, according to the specific definitions of nodes and edges. Unlike convolutional neural networks (CNNs) [20] that explore local information in Euclidean space and Transformer [21,22] that exploits global information based on sequences, GNNs provide higher flexibility in analytical perspectives and better interpretability by focusing on the graph topology. However, existing graph-based WSI classification methods still have three limitations. (1) The constructed graphs are usually based on the regular graph, which uses one edge to connect two adjacent nodes. However, the data relationships in practical applications are generally non-pairwise, especially in the gigapixel WSI that contains rich contextual information. Therefore, the constraint on pairwise correlation among patches may be unreasonable. (2) The adjacency relationship is measured from one single perspective, such as spatial coordinate [9,18] in Euclidean space or embedding similarity [23]. The graph established from a single perspective may lose flexibility and generalization for different downstream tasks. (3) Random sampling [9] or cluster-based sampling [24] is usually performed on the split patches for producing the instances and thus the computation complexity can be reduced. However, in the cases of cancer subtyping and gene mutation prediction, it is clinically uncertain whether the uninvolved patches have no significant association with the tumour's heterogeneous structure or gene mutation sites. Therefore, the model interpretability and performance of downstream tasks will be influenced.

To address these limitations, we propose a novel masked hypergraph learning framework (MaskHGL) for weakly supervised WSI classification that leverages hypergraph to learn non-pairwise relationships among the patches without sampling from different perspectives by defining multiple types of hyperedges and ultimately obtains slide-level representation for classification through a self-attention mechanism based aggregation module. The entire workflow is shown in Fig. 1. Particularly, a Masked Hypergraph ReConstruction (MHRC) module is proposed to enhance the robustness and classification performance of MaskHGL, which implicitly conducts slide-level augmentation by masking a certain proportion of nodes with learnable MASK token and reconstructing hypergraph architecture in the subsequent decoder.

Compared with previous hypergraph-related methods [9,25] in WSI analysis, MaskHGL involves further considerations in WSI modelling and hypergraph learning, including the hypergraph construction strategy, the hypergraph attention, the mask reconstruction auxiliary tasks. In the end, we evaluate the proposed MaskHGL on two public TCGA benchmark datasets and one in-house dataset and compare it with 6 state-of-the-art WSI classification methods. Experimental results have demonstrated the proposed MaskHGL is more effective in the tasks of histopathology WSI cancer subtyping and gene mutation prediction.

The main contributions of this paper are summarized below:

- We propose a novel masked hypergraph learning framework (MaskHGL) for weakly supervised WSI classification, utilizing hypergraph convolutional neural networks. Hypergraphs from multi-perspective (i.e. embedding space and spatial space) are constructed and hypergraph attention is introduced into hypergraph learning. Consequently, the non-pairwise relationships in WSI are well-learned, and more discriminative node representation can be gained. Besides, an efficient self-attention-based aggregator is applied to explore the global dependencies of the patches and generate refined patch-level embeddings for instance-level weak supervision as well as the slide-level representation for downstream classification.

- We further design a masked hypergraph reconstruction module for the proposed MaskHGL. It masks a certain proportion of nodes in hypergraph and reconstructs the masked node embeddings with a trainable decoder. This module potentially enriches the WSI samples seen by the encoder, which alleviates data scarcity and significantly boosts the robustness against instance-level noise. Therefore, the classification performance and generalization of MaskHGL are greatly improved.

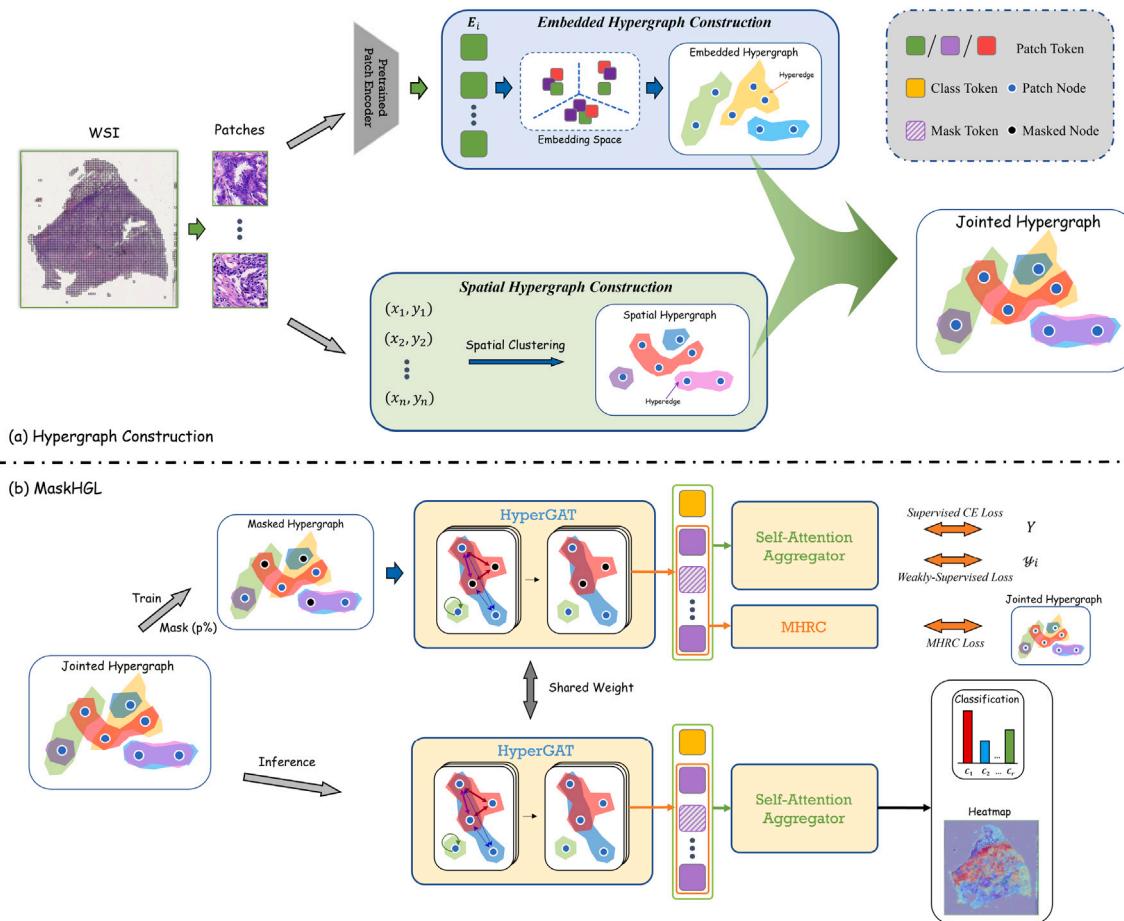
- We have evaluated our method and existing state-of-the-art WSI classification methods on two public TCGA benchmark datasets and one in-house clinical dataset. Experimental results demonstrate the effectiveness of our method in lung cancer subtyping and epidermal growth factor receptor (EGFR) mutation prediction of non-small cell lung cancer (NSCLC). More importantly, it has shown a more promising effect for EGFR gene mutation subtyping from hematoxylin and eosin (H&E) stained WSIs, compared with conventional costly sequencing technologies.

The rest of this paper is organized as follows: Section 2 reviews the related works, together with recent GNN-based WSI classification methods and especially recent hypergraph-related WSI analysis methods. Section 3 introduces the methodology of our proposed framework including the details about hypergraph construction and the detailed architecture. The experiments and corresponding analysis are shown in Section 4, and Section 5 summarizes the conclusions and future work.

## 2. Related works

### 2.1. GNN-based methods for WSI classification

Existing WSI analysis methods are generally patch-based. Since the relationships between patches and WSI are very similar to those between vertices and graphs, graph-based methods are extensively involved in WSI analysis and become an important branch. Chen et al. [18] proposed a context-aware, spatially-resolved patch-based graph convolutional network (Patch-GCN) for survival prediction in patients with multiple WSIs. It takes WSIs as point clouds and constructs simple graphs with the Euclidean space similarity between patches. Zhang et al. [23] designed a two-stage cervical cancer screening method based on graph attention network (GAT) and supervised contrastive learning, which selects top-K and bottom-K suspicious lesion patches to construct complete graphs separately and uses GAT to aggregate their features for WSI classification. Guan et al. [24] introduced a Node-Aligned GCN (NAGCN) based on a hierarchical global-to-local clustering strategy for WSI representation and classification. Lu et al. [14] proposed a GNN model termed SlideGraph+ to predict HER2 status in breast cancer, which uses the biomarker



**Fig. 1.** The paradigm of our proposed weakly supervised WSI analysis framework MaskHGL. (a) describes the overall data pre-processing and the hypergraph construction. Given WSI is first split into patches, then two types of hypergraphs are constructed from different perspectives and finally stacked into one jointed hypergraph which can capture the non-pairwise relationships among patches. (b) denotes the workflows of MaskHGL in the training and inference stages, separately. In the training stage, the inputted hypergraph has a certain probability of being masked for augmentation, and then the masked hypergraph is fed into the HyperGAT for learning the node representations. After that, node embeddings are inputted into a self-attention-based aggregator and MHRC module for generating discriminative slide-level representation and masked node reconstruction separately. In the inference stage, the mask operation and the related MHRC module are dropped. The inputted hypergraph is straightly sent into the trained HyperGAT and then the self-attention-based aggregator to produce slide-level classification prediction and concerned regions as a heatmap.

attributes and neural network embeddings to represent the complex organization of cells and the overall tissue micro-architecture.

The above methods directly adapt simple graphs established from a single perspective to explore inter-relationships among patches, thus lacking flexibility. Recent study [26] proposed a heterogeneous graph representation learning framework for WSI classification, which utilizes a heterogeneous graph with various pre-defined types of nodes to exploit the heterogeneity within WSI. It breaks free from the limitations of simple graphs yet still ignores the non-pairwise relations between nodes. Different from previous works, we apply hypergraphs with multiple types of hyperedges to model the non-pairwise inter-relationships among patches and finally achieve promising performance for different downstream tasks, e.g. cancer subtyping and gene mutation prediction.

## 2.2. Hypergraph learning in WSI analysis

A hypergraph is a generalization of a simple graph in which an edge can connect more than two nodes, and it is called hyperedge. Formally, a hypergraph is a pair  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V}$  is a set of vertices and  $\mathcal{E}$  is a set of subsets of  $\mathcal{V}$ . Each hyperedge contains a variable number of nodes  $(v_1, v_2, \dots, v_n) \in \mathcal{E}$ , where  $v_j \in \mathcal{V}$  and  $n$  indicate the number of nodes in this hyperedge ( $n$  can vary for different hyperedge). Hypergraph learning aims to learn valuable non-pairwise relationships from the given hypergraph structure. It has gained lots of attention owing to its flexibility and capability to represent complex data correlations.

Recently, researchers have introduced hypergraphs into deep learning and turn out to achieved successful performance improvements in various fields, such as computer vision [27–29], citation networks [30,31], social networks [32,33], and recommendation systems [34–36].

In the field of histopathology WSI analysis, hypergraph learning still has not been fully explored. Di et al. [9] proposed a big-Hypergraph Factorization Neural Network (b-HFGN) for processing the data with large-scale vertices (e.g. WSI) in an efficient way, and combining it with a multi-level ranking-based method for prediction survival hazard scores. Soon after, they further designed HGSurvNet [25] to generate high-order hypergraph-based representations of WSI for survival prediction by introducing topology information in hypergraph construction. Hou et al. [37] recently borrowed the concept of hypergraphs to capture intrinsic dependence between modalities for cancer survival prediction. Li et al. [38] proposed a hypergraph-guided retrieval module to explore high-order correlation in slide-level histology retrieval.

In these distinguished works mentioned above, HGSurvNet [25] partly shares similar thoughts with our work, specifically, in the WSI modelling strategy. We both tried to model WSI from embedding space as well as spatial space. However, there are still some significant differences. HGSurvNet proposed a topology-based sampling strategy which introduces a prior that the differentiable information is uniformly distributed along the tissue topology. Such prior may be ineffective

for the case of gene mutation prediction since there is still no evidence to indicate a clear connection between gene mutation sites and histopathological morphology. Besides, HGSurvNet simply performed the K-nearest neighbour (KNN) strategy on the instance embeddings of each WSI separately which ignores the semantic consistency of hyperedges across slides. We took the above considerations into account and proposed the novel MaskHGL with the refined hypergraph construction strategy.

### 3. Methodology

#### 3.1. Hypergraph construction

Since the background in WSI occupies a significant portion and contains no information related to the WSI diagnosis, we first adopt the OTSU thresholding algorithm [39] and sliding window strategy to crop each WSI into non-overlapping patches with fixed size at specific magnification (e.g., 20x) and remove the background regions without tissue. Each remaining patch represents a local tissue region in the WSI. Then, DINO [40] is adopted to pre-train a Vision Transformer-Small (ViT-S) [22] network for the patch-level feature extraction.

---

#### Algorithm 1: Hypergraph Construction

---

```
# input: Training set on embedding space  $D_e$ 
# input: Given WSI  $B = (E, C)$ 
# output: Hypergraph representation  $\mathcal{G} = (\mathbf{V}, \mathbf{H})$  for  $B$ 
# hyper-parameter: number of embedded hyperedges  $K$ 
# hyper-parameter: similarity weights  $\lambda_e$  and  $\lambda_s$ 
# hyper-parameter: distance threshold  $\rho_d$ 

# Conduct global clustering on embedding space
glo_cluster = KMeans(n_clusters=K).fit( $D_e$ )

# Assign patch embeddings  $E$  into cluster
embedded_labels = glo_cluster.predict( $E$ )

# Connect patches with the same embedded labels
H_embedded = index_connect(embedded_labels)

# Compute Gaussian kernel-based similarity
for i in range(n):
    for j in range(n):
        dis[i, j] = exp(- $\lambda_e * \text{torch.sum}(E[i] - E[j]) * \exp(-\lambda_s * \text{torch.sum}(C[i] - C[j]))$ 

# Conduct agglomerative clustering and connect patches in the same
# cluster
spa_labels = agglo_cluster.fit(dis)
H_spatial = index_connect(spa_labels)

# Concat embedded hyperedge and spatial hyperedge
H = hyperedge_concat(H_embedded, H_spatial)

return  $\mathcal{G} = (\mathbf{V}, \mathbf{H})$ , where  $\mathbf{V} = E$ 
```

---

Specifically, we define the train subset as  $D = \{B_1, B_2, \dots, B_N\}$ , where  $B$  denotes WSI and  $N$  is the number of WSIs. After pre-processing, each WSI can be denoted as  $B_i = (\mathbf{P}_i, \mathbf{C}_i)$ , where  $\mathbf{P}_i = \{\mathbf{p}_i^0, \mathbf{p}_i^1, \dots, \mathbf{p}_i^n\}$  indicates the image set of patches and  $\mathbf{C}_i = \{\mathbf{c}_i^0, \mathbf{c}_i^1, \dots, \mathbf{c}_i^n\}$  is the set of patch coordinates that  $\mathbf{c}_i^j$  denotes the coordinate of the  $j$ th patch of  $i$ th WSI.  $n$  is the number of patches ( $n$  can vary for different WSIs). It is worth mentioning that we conduct no sampling on patches to prevent potential gene-related information drops. Then a pre-trained ViT-S  $f(\cdot)$  is adopted to encode each patch  $p_i^j$  into a fixed dimensional embedding  $\mathbf{e}_i^j \in \mathbb{R}^L \leftarrow f(p_i^j)$ . For  $\mathbf{e}_i^j$  contains semantic information of  $p_i^j$ , we can transfer the dataset into embedding space  $D_e = \{(\mathbf{E}_1, \mathbf{C}_1), \dots, (\mathbf{E}_N, \mathbf{C}_N)\}$ , where  $\mathbf{E}_i = \{\mathbf{e}_i^1, \mathbf{e}_i^2, \dots, \mathbf{e}_i^n\} \in \mathbb{R}^{n \times L}$  indicates the  $i$ th WSI in embedding space. To maintain the correspondence

between the constructed hypergraph and WSI, we define patches as the nodes and patch features as node embeddings, thereby modelling WSI as a hypergraph. Formulaically, hypergraph vertices can be denoted as  $\mathbf{V} = \mathbf{E} \in \mathbb{R}^{n \times L}$ , where  $\mathbf{E}$  literally follows the definition of  $\mathbf{E}_i$  for simplification.

Following the definition of nodes, we define the hyperedge relationships from two different perspectives. Inspired by the node-aligned graph construction strategy in NAGCN [24], we adopt a similar global clustering but different hyperedge relationships to model and align structural and contextual information in each WSI for taking overall patch distribution and separate heterogeneous patterns into account. In the embedding space, we first perform global K-Means clustering on  $D_e$  to decompose the embedding space into  $K$  sub-spaces. Then, the patches distributed in the same sub-spaces are connected as an embedded hyperedge  $\mathbf{H}_e$  in each WSI. Moreover, at spatial space, we conduct adaptive spatial agglomerative clustering [14] in each WSI and each cluster is defined as a spatial hyperedge  $\mathbf{H}_s$ . Finally, we stack these two types of hyperedges into one jointed hypergraph  $\mathcal{G}$ :

$$\mathbf{H} = [\mathbf{H}_e || \mathbf{H}_s] \in \{0, 1\}^{n \times (K+M_s)}, \quad (1)$$

where  $\mathbf{H}_e \in \{0, 1\}^{n \times K}$  and  $\mathbf{H}_s \in \{0, 1\}^{n \times M_s}$  represent embedded hyperedge and spatial hyperedge in matrix form, respectively.  $M_s$  is the number of spatial hyperedges and is variable in different WSIs. Note that  $\mathbf{H}(i, j) = 1$  means the  $i$ th node belongs to the  $j$ th hyperedge. After that, we obtain the hypergraph  $\mathcal{G}$  for each WSI  $B$  as:

$$\mathcal{G} = (\mathbf{V}, \mathbf{H}). \quad (2)$$

The pseudo-code of hypergraph construction is shown in Algorithm 1.

#### 3.2. Masked HyperGraph learning framework

In this section, we will discuss the details of the proposed masked hypergraph learning framework (MaskHGL). The structure details of our MaskHGL is shown in Fig. 2. It includes four parts, namely Hypergraph Attention Network (HyperGAT), Self-attention-based Aggregator (SAA), Masked Hypergraph Reconstruction, and Weakly Supervised Patch-level Loss.

##### 3.2.1. Hypergraph attention network

Our HyperGAT is composed of hypergraph convolutional layers with attention mechanism [30]. We first provide the definition of hypergraph convolution and hypergraph attention.

A hypergraph with  $N$  vertices and  $M$  hyperedges is defined as  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ , where a diagonal matrix  $\mathbf{W} \in \mathbb{R}^{M \times M}$  stores the weights for each hyperedge. Different from a simple graph where an adjacency matrix is defined as  $\mathbf{A} \in \mathbb{R}^{N \times N}$ , the adjacency relationship in hypergraph  $\mathcal{G}$  can be characterized by an incidence matrix  $\mathbf{H} \in \mathbb{R}^{N \times M}$  in general. When a vertex  $v_i \in \mathcal{V}$  is connected by  $\epsilon$ ,  $\mathbf{H}_{i\epsilon} = 1$ , otherwise 0. It is worth mentioning that we sparsify  $\mathbf{H}$  during implementation, so the computation wouldn't be too costly even when the number of vertices is enormous (e.g. 10 k~20 k). Then, the vertex degree is defined as

$$\mathbf{D}_{ii} = \sum_{\epsilon=1}^M \mathbf{W}_{\epsilon\epsilon} \mathbf{H}_{i\epsilon} \quad (3)$$

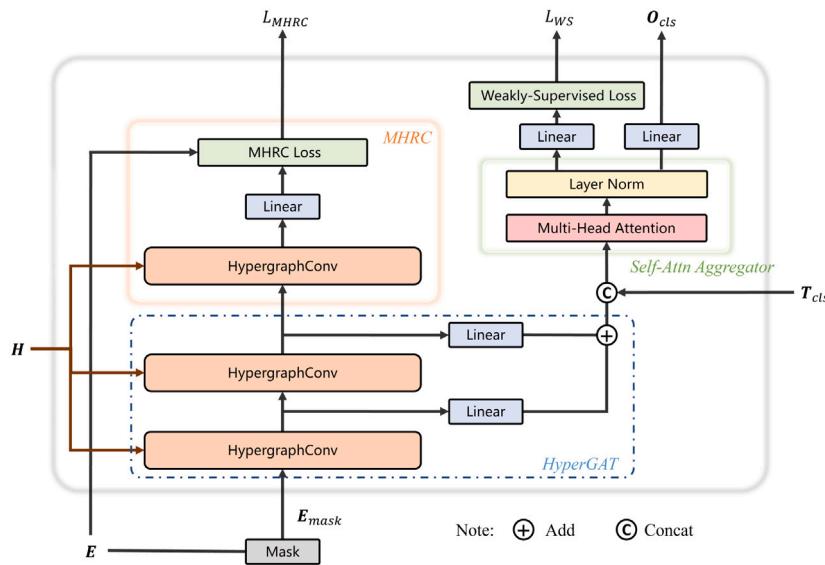
and the hyperedge degree is defined as

$$\mathbf{B}_{\epsilon\epsilon} = \sum_{i=1}^N \mathbf{H}_{i\epsilon}. \quad (4)$$

With the definition of vertex degree and hyperedge degree, the hypergraph convolution is defined as

$$\mathbf{X}^{(l+1)} = \sigma(\mathbf{D}^{-1/2} \mathbf{H} \mathbf{W} \mathbf{B}^{-1} \mathbf{H}^\top \mathbf{D}^{-1/2} \mathbf{X}^{(l)} \mathbf{P}), \quad (5)$$

where  $\mathbf{X}^{(l)} \in \mathbb{R}^{N \times F^{(l)}}$  indicates the input of the  $l$ th layer,  $\mathbf{X}^{(0)} = \mathbf{E}$ , and  $\mathbf{P} \in \mathbb{R}^{F^{(l)} \times F^{(l+1)}}$  is the learnable weight matrix between the  $l$ th and  $(l+1)$ th layer.  $\sigma(\cdot)$  is a non-linear activation function such as LeakyReLU.



**Fig. 2.** The detailed structure of MaskHGL, where  $H$  and  $E$  indicate hyperedge incidence matrix and node embeddings of input WSI respectively,  $E_{mask}$  denotes the node embeddings after the mask operation,  $T_{cls}$  represents the appended class token, and  $O_{cls}$  means the output logits for slide-level prediction.

Superscript  $F^{(l)}$  denotes the dimension of nodes embedding in the  $l$ th layer.

Hypergraph convolution has a sort of fundamental attention mechanism but which is not trainable for a given graph structure  $\mathbf{H}$ . The goal of hypergraph attention is to learn a dynamic incidence matrix for better revealing the intrinsic relationship between vertices [30]. For a given vertex  $x_i$  and its related hyperedge  $x_j$ , the attentional incidence matrix can be calculated as

$$H_{ij} = \frac{\exp(\sigma(sim(x_i\mathbf{P}, \bar{x}_j\mathbf{P})))}{\sum_{k \in \mathcal{N}_i} \exp(\sigma(sim(x_i\mathbf{P}, x_k\mathbf{P})))}, \quad (6)$$

where  $\bar{x}_j$  denotes the attribution of hyperedge  $x_j$ , which is implemented in MaskHGL as the mean average of the embeddings of covered nodes.  $\mathcal{N}_i$  indicates the neighbourhoods of  $x_i$ .  $sim(\cdot)$  computes the similarity between two vertices, defined as

$$sim(x_i, x_j) = \mathbf{a}^\top [x_i || x_j], \quad (7)$$

here  $[\cdot || \cdot]$  denotes concatenation and  $\mathbf{a}$  is a weight vector used to output a scalar similarity value.

Finally, hypergraph convolution described as Eq. (5) with the incidence matrix  $\mathbf{H}$  enriched by the above attention mechanism can efficiently learn the intermediate embeddings of vertices layer-by-layer. Based on that, we designed our HyperGAT as illustrated in Fig. 2 by utilizing two hypergraph convolution layers and a couple of linear layers.

### 3.2.2. Self-attention based aggregator

To generate discriminative graph/slides-level representation efficiently, we designed the SAA module to capture global information about WSI. Unlike conventional methods of Mean-pooling or Max-pooling, we introduce the attention mechanism that adaptively finds important parts of input and weights it in the following aggregation [21].

Apart from the conventional attention mechanism and the gated attention used in [41], self-attention is more effective in processing variant-length data. It can be formally expressed as the following formula:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \quad (8)$$

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \mathbf{K} = \mathbf{X}\mathbf{W}_K, \mathbf{V} = \mathbf{X}\mathbf{W}_V, \quad (9)$$

where  $\mathbf{X}$  is the input sequence,  $\mathbf{W}_Q$ ,  $\mathbf{W}_K$ , and  $\mathbf{W}_V$  are learnable transformation matrices,  $d$  is the dimension of vector in  $\mathbf{Q}$ .

To aggregate patch-level features with low computational complexity, we directly employ the Nyström Attention [42], a self-attention variant with linear complexity followed by layer normalization operation as our aggregation module. Note that, a class token will be appended to the patch token sequence before feeding it to the aggregator for learning the slide-level representation.

After information passing and aggregation, the class token can be gained as slide-level embedding and further inputted into a linear classifier for slide-level prediction. Supervised Cross-Entropy loss will be calculated as below:

$$L_{SCE} = \frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C Y_i^c \log(p_i^c), \quad (10)$$

here  $C$  is the number of classes,  $Y_i^c$  and  $p_i^c$  indicate the true label and predict probability of each sample  $i$  for class  $c$ .

### 3.2.3. Masked hypergraph reconstruction

For fine-granularity-focused downstream tasks, like gene mutation prediction, sufficient and low-noised slide samples are crucial. Compared with collecting enormous samples with high quality, augmenting the available data and improving model robustness are more feasible. Inspired by the idea of masked autoencoders [43,44], we considered the mask reconstruction as an auxiliary task and bound it with our hypergraph attention network, then proposed the Masked Hypergraph ReConstruction (MHRC) module to tackle above problem. With the mask reconstruction, the network attaches importance to the shallow information that can be used for node reconstruction rather than blindly mining the deep representations for downstream tasks, which will make the network more receptive to the instance-level noise and significantly improve the learning ability of the encoder at the instance level. The implementation is described as follows:

During training, the given WSI has a probability  $p_{mask}$  of being augmented with the MHRC module. When applying MHRC, we first mask a certain ratio of nodes in the hypergraph by replacing them with a learnable MASK Node  $e_{mask}$  without modifying the hypergraph structure. The mask operation can be formatted as below:

$$\mathcal{G}_{mask} = (\mathbf{E}_{mask}, \mathbf{H}) \leftarrow \mathcal{G} = (\mathbf{E}, \mathbf{H}) \quad (11)$$

$$\mathbf{E}_{mask} = \{e_1^1, \dots, e_n^1\} \leftarrow \mathbf{E} = \{e^1, \dots, e^n\} \quad (12)$$

$$\mathbf{e}_\text{mask}^i = \begin{cases} \mathbf{e}_\text{masked}, & \text{if } i \in \text{RandIdx}(n * r_\text{mask}, L) \\ \mathbf{e}^i, & \text{else.} \end{cases}, \quad (13)$$

where  $\text{RandIdx}(k, N)$  is a function for generating  $|k|$  random indexes from 0 to  $N - 1$  and  $r_\text{mask}$  is the mask ratio hyperparameter which controls the proportion of masked nodes.

Then, the masked hypergraph  $G_\text{mask}$  is input into the target encoder HyperGAT. The corresponding output is delivered to our MHRC module which includes a hypergraph convolutional decoder for predicting the masked node embeddings.

Finally, the mean squared error (MSE) between the reconstructed and original node embeddings will be calculated.

$$L_{MHRC} = \frac{\sum_{i \in \text{RandIdx}(n * r_\text{mask}, L)} \|\mathbf{e}^i - \mathbf{e}_\text{mask}^i\|_1^2}{[n * r_\text{mask}]}, \quad (14)$$

where  $\|\cdot\|_1$  denotes the  $L_1$  norm.

From the perspective of hypergraph representation of WSI, the MHRC module is able to effectively augment available WSIs by randomly replacing a portion of nodes with **MASK Node**. It can be analogized with the Mosaic augmentation strategy [45] from the graph view. From the perspective of instance (node/vertex), the learnable **MASK Node** can capture category-related instance-level information for further improving the authenticity of the augmented hypergraph.

### 3.2.4. Weakly supervised patch-level loss

As defined by the weakly supervised learning paradigm, we wish to give patch-level predictions in the absence of fine-grained annotations. We directly take the patch token output from the SAA as patch-level embeddings and train a patch-level classifier under the supervision of pseudo labels. Pseudo labels of patches  $y_i^c$  are determined by the true label  $Y_i^c$  of its belonging WSI.

$$L_{WS} = \frac{1}{n} \sum_{i=1}^n \sum_{c=1}^C y_i^c \log(p_i^c), \text{ where } y_i^c = Y_i^c \quad (15)$$

It is worth mentioning that we have noticed such intuition that hard-label supervision will introduce much noise. We entrusted the task of alleviating the effect of instance-level noise to the MHRC module. The shallow information enhanced by the MHRC could strongly alleviate the impact of such noise and make promising improvements to downstream tasks. More analysis can be found in Section 4.

Finally, the entire loss function can be formally expressed in the following forms:

$$L = L_{SCE} + \lambda_1 L_{MHRC} + \lambda_2 L_{WS} \quad (16)$$

where  $\lambda_1$  and  $\lambda_2$  are weights for each part of our loss function.

## 4. Experiments

### 4.1. Datasets

The proposed method is evaluated on two public TCGA datasets (*TCGA-LUNG* and *TCGA-EGFR*)<sup>1</sup> and one in-house clinical dataset (*USTC-EGFR*)<sup>2</sup>. *TCGA-LUNG* dataset is used for lung cancer subtyping, *TCGA-EGFR* and *USTC-EGFR* datasets are used for EGFR gene mutation prediction which is the most critical driver mutation in NSCLC [47]. Compared with *TCGA-EGFR*, which only includes coarse-grained labels (Mutant and Wild), *USTC-EGFR* is introduced with fine-grained mutation status labels and thus it can evaluate the potential of experimental

**Table 1**

The WSI Distribution of the Experimental Datasets.

TCGA-LUNG	Neg	LUAD	LUSC		
Train	385	326	333		
Test	165	141	144		
TCGA-EGFR	Wild	Mutant			
Train	415	74			
Test	175	32			
USTC-EGFR	Neg	L858R	19del	Wild	Others
Train	117	80	137	99	98
Test	48	38	47	47	43

methods in clinical practice. Note that we aim to directly predict the EGFR gene mutation subtypes through histopathology WSIs without polymerase chain reaction (PCR) or next-generation sequencing (NGS).

The profiles of these datasets are presented below and the detailed distribution summary is provided in Table 1.

- *TCGA-LUNG* contains 1345 cases of lung histopathology collected from The Cancer Genome Atlas (TCGA) program of the National Cancer Institute (NCI). All WSIs are categorized into 3 subtypes including cancer-free tissue (Normal, 553 WSIs), lung adenocarcinoma (LUAD, 1257 WSIs) and lung squamous cell carcinoma (LUSC, 1254 WSIs). To alleviate the class imbalance, we sampled 1494 WSIs for experiments and each WSI contains an average of 7992 patch images.
- *TCGA-EGFR* contains 696 WSIs from 364 patients for EGFR mutation detection and also collected from the TCGA program. These WSIs are categorized into 2 subtypes of wild type (Wild) and EGFR-mutant (Mutant). Each WSI contains an average of 6652 patch images.
- *USTC-EGFR* contains 754 WSIs of lung histopathology collected from 521 patients by The First Affiliated Hospital of USTC (University of Science and Technology of China). Unlike the *TCGA-EGFR* dataset, we further consider the two clinically common EGFR mutation types which account for approximately 90% of mutations of NSCLC [48], i.e. in-frame deletions in exon 21 (19del) and a missense mutation in exon 21 (L858R). Therefore, it is of great importance to classify these two subtypes. Besides, we collect three other subtypes including negative (Neg), wild type (Wild) and other driver gene mutations (Others). All labels have been confirmed by pathologists. Each WSI contains an average of 10503 patch images.

In each dataset, the WSIs were randomly separated into train and test sets by the proportion of 7:3. Five-fold cross-validation was conducted within the train set where the validation part in each fold was used to guide optimal model selection and hyper-parameter determination.

### 4.2. Implementation details

The proposed framework is implemented in Python with the *PyTorch* and *PyTorch Geometric (PyG)* libraries. Before training, the window sliding strategy was applied on the WSIs under 20× lenses (the resolution is 0.48 μm/pixel). The window size was 256 × 256. Then all the patches were fed into the self-supervised framework DINO [40] for pre-training and each patch was represented by a 384 dimensional embedding with pre-trained ViT-S. During the training stage, the Adam optimizer [49] and OneCycleLR [50] scheduler were employed with an initial learning rate of 2e-4 and weight decay of 1e-5. The size of the mini-batch is 1. In the inference stage, the MHRC module is off. All experiments are done with an NVIDIA RTX 3090. The code is available at <https://github.com/HFUT-miaLab/MaskHGL>.

The average accuracy (ACC), macro-average F1 score (F1), the macro-average area under the receiver operating characteristic curve (AUC) are calculated for evaluating the classification performance.

<sup>1</sup> *TCGA-LUNG* and *TCGA-EGFR* were obtained from The Cancer Genome Atlas portal (<https://portal.gdc.cancer.gov/>) [46].

<sup>2</sup> The study was approved by the Medical Research Ethics Committee of the First Affiliated Hospital of the University of Science and Technology of China (Anhui Provincial Hospital) under the protocol No. 2022-RE-454.

**Table 2**

Comparison of the State-of-the-art Methods, where ACC, AUC, and F1 denote the accuracy, macro-average area under roc curve, and macro-average F1 score, respectively.

Method	TCGA-LUNG ( <i>cls</i> =3)			TCGA-EGFR ( <i>cls</i> =2)			USTC-EGFR ( <i>cls</i> =5)		
	ACC(%)	AUC(%)	F1(%)	ACC(%)	AUC(%)	F1(%)	ACC(%)	AUC(%)	F1(%)
CLAM-SB	89.56 ± 1.16	96.84 ± 0.64 <sup>†</sup>	88.99 ± 1.30	79.13 ± 2.10 <sup>‡</sup>	63.84 ± 7.10 <sup>†</sup>	53.19 ± 2.70	45.20 ± 2.88 <sup>‡</sup>	73.53 ± 0.70 <sup>‡</sup>	43.56 ± 3.80 <sup>‡</sup>
CLAM-MB	89.42 ± 1.06	96.90 ± 0.23 <sup>‡</sup>	88.84 ± 1.10	80.29 ± 1.41	65.79 ± 6.70	54.83 ± 3.00	50.22 ± 2.96 <sup>‡</sup>	78.05 ± 1.50 <sup>‡</sup>	49.78 ± 3.30 <sup>‡</sup>
TransMIL	88.04 ± 0.85 <sup>‡</sup>	96.48 ± 0.23 <sup>‡</sup>	87.42 ± 0.93 <sup>‡</sup>	82.22 ± 1.79	59.05 ± 5.00 <sup>‡</sup>	55.05 ± 2.30	45.83 ± 1.54 <sup>‡</sup>	74.67 ± 0.77 <sup>‡</sup>	43.86 ± 0.47 <sup>‡</sup>
Patch-GCN	89.51 ± 1.07	96.46 ± 0.32 <sup>‡</sup>	88.97 ± 1.10	76.91 ± 3.31 <sup>†</sup>	67.11 ± 5.00	55.87 ± 4.40	41.35 ± 1.15 <sup>‡</sup>	69.62 ± 0.96 <sup>‡</sup>	35.36 ± 1.80 <sup>‡</sup>
NAGCN	79.02 ± 0.90 <sup>‡</sup>	91.22 ± 0.54 <sup>‡</sup>	78.06 ± 0.95 <sup>‡</sup>	81.55 ± 1.35	67.24 ± 4.90 <sup>†</sup>	50.69 ± 3.40	53.54 ± 3.31	83.07 ± 1.00 <sup>‡</sup>	52.42 ± 2.70 <sup>‡</sup>
SlideGraph+	88.09 ± 1.16 <sup>‡</sup>	96.69 ± 0.33 <sup>‡</sup>	87.46 ± 1.20 <sup>†</sup>	82.32 ± 0.78	66.48 ± 4.10 <sup>‡</sup>	53.75 ± 6.00	56.41 ± 0.96 <sup>†</sup>	84.81 ± 0.61 <sup>‡</sup>	55.48 ± 0.74 <sup>‡</sup>
MaskHGL	<b>90.49 ± 0.74</b>	<b>97.52 ± 0.24</b>	<b>89.98 ± 0.77</b>	<b>82.55 ± 2.44</b>	<b>74.21 ± 3.80</b>	<b>59.71 ± 6.00</b>	<b>60.63 ± 2.65</b>	<b>87.45 ± 1.00</b>	<b>60.44 ± 2.80</b>

<sup>†</sup> denotes P-Value < 0.05.

<sup>‡</sup> denotes P-Value < 0.01.

#### 4.3. Results and analysis

We conduct a comprehensive comparison of MaskHGL with the state-of-the-art WSI analysis approaches, including (1–2) CLAM [17] with single-branch (CLAM-SB) and multi-branch (CLAM-MB), (3) TransMIL [51], (4) Patch-GCN [18], (5) NAGCN [24], (6) SlideGraph+ [14]. For fairness of comparison, we use the same patch encoder for all the methods. All approaches are evaluated using the same five-fold cross-validation splits and have undergone significance test [52]. The mean and standard deviation results on the test set are presented in Table 2, where the special indicator <sup>†</sup> and <sup>‡</sup> denote that there is a statistically significant difference (P-Value < 0.05 or P-Value < 0.01).

Overall, the proposed MaskHGL achieves the best performance with an AUC of 87.45% on the USTC-EGFR dataset, 97.52% on the TCGA-LUNG dataset and 74.21% on the TCGA-EGFR dataset. The significance test has assisted in proving that our method is significantly superior to other comparative methods in most metrics and experimental settings.

Among these comparison methods, CLAM (both SB and MB) are classical MIL-based methods, TransMIL is a Transformer-based method, and Patch-GCN, NAGCN, and SlideGraph+ are recent GNN-based methods. It can be observed that the overall performance of GNNs-based methods is superior. The results show that the correlation information learned from the graph structure can generally improve the performance of classification, especially on the class-imbalanced gene mutation prediction dataset TCGA-EGFR. Specifically, Patch-GCN, NAGCN, SlideGraph+, and MaskHGL achieve AUC gains of 1.32%, 1.45%, 0.69%, and 8.42% compared with CLAM-MB, respectively, on the TCGA-EGFR dataset. It has demonstrated the innate robustness of GNNs-based methods against class imbalance.

Compared with Patch-GCN, the other GNNs-based methods and proposed MaskHGL achieve significant AUC gains of 13.45%, 15.19%, and 17.83% on the fine-grained gene mutation prediction dataset USTC-EGFR, respectively. The results have indicated that the patch similarity on embedding space is critical for fine-grained gene mutation prediction, and causes general gaps. More importantly, MaskHGL provides a more flexible strategy for multi-perspective consideration by hypergraph attention mechanism, which adaptively weights each type of hyperedges and avoids multi-perspective coupling from the very beginning.

As shown in Table 2, MaskHGL surpasses the overall second-best method SlideGraph+ over 2.64% AUC on USTC-EGFR, 0.83% on TCGA-LUNG and 7.73% on class-imbalanced dataset TCGA-EGFR. Compared with NAGCN and SlideGraph+, MaskHGL has the advantages of hypergraph-based non-pairwise relationships modelling and an efficient representation learning strategy guided by hypergraph mask reconstruction auxiliary task. The quantitative results clearly demonstrate the effectiveness of MaskHGL.

Additionally, we conduct t-SNE visualization to demonstrate the distributions of slide-level embeddings generated by the proposed MaskHGL and two other comparison methods in Fig. 4. We choose the overall second-best method SlideGraph+ (Fig. 4a) and primitive

**Table 3**

Effects of different modules in MaskHGL.

	Modules			USTC-EGFR ( <i>cls</i> = 5)	
	SAA	MHRC	WS	ACC(%)	AUC(%)
(A)	✗	✗	✗	50.76 ± 3.42	82.46 ± 1.60
(B)	✓	✗	✗	53.63 ± 3.23	83.57 ± 0.95
(C)	✓	✓	✗	53.90 ± 4.05	83.47 ± 1.50
(D)	✓	✗	✓	56.41 ± 2.91	84.58 ± 0.86
MaskHGL	✓	✓	✓	60.63 ± 2.65	87.45 ± 1.00

HyperGAT without further modification (Fig. 4b) as the comparison methods. It can be observed that compared with the WSI representation distribution generated by SlideGraph+ and primitive HyperGAT, our method has relatively better intra-class compactness and inter-class separability. Especially for the clinical-concerned categories (Neg, L858R, and 19del), homogeneous WSI representations are closer and heterogeneous WSI representations are as far away as possible (marked in coloured circles shown in Fig. 4c). It intuitively indicates that MaskHGL is able to generate more discriminative slide-level representations and proves the effectiveness of the proposed modules. Notably, for the other two categories (Wild and Others), all these three methods have not gained a satisfying distribution. We consider it is due to the intrinsic complexity of other driver genes and the histopathological patterns for Wild and Others share some similarity, which is hard to distinguish accurately with limited cases.

#### 4.4. Ablation studies and analysis

Our proposed framework can be divided into one main backbone HyperGAT and three additional components: (1) Self-Attention based Aggregator (SAA), (2) Masked Hypergraph ReConstruction module (MHRC) and (3) Patch-level Weak Supervision (WS). To evaluate the effectiveness of each component, we performed ablation studies on these three modules.

To verify the effectiveness of the SAA module, we perform a comparison with the Gated-Attention mechanism used in [41]. As shown in Table 3 (A)&(B), (A) indicates HyperGAT with Gated-Attention for slide-level representation aggregation, and (B) represents the ablation versions of MaskHGL w/o MHRC-WS. Note that (B) surpassed (A) by 1.11% AUC and 2.87% ACC. It is worth mentioning that Gated-Attention was originally implemented on Top-K sampled patches in [41]. The results have indicated that the SAA module has the ability to exploit complementary information in WSI when facing unsampled variable-length instance sequences than the classical Gated-Attention mechanism.

For the other two additional components MHRC and WS, Table 3 (C) indicates MaskHGL w/o WS and Table 3 (D) indicates MaskHGL w/o MHRC. Compared with the full version of MaskHGL, an improvement of 3.98% AUC and 2.87% AUC is brought by WS and MHRC, respectively, which proves the effectiveness of the WS and MHRC modules. Note

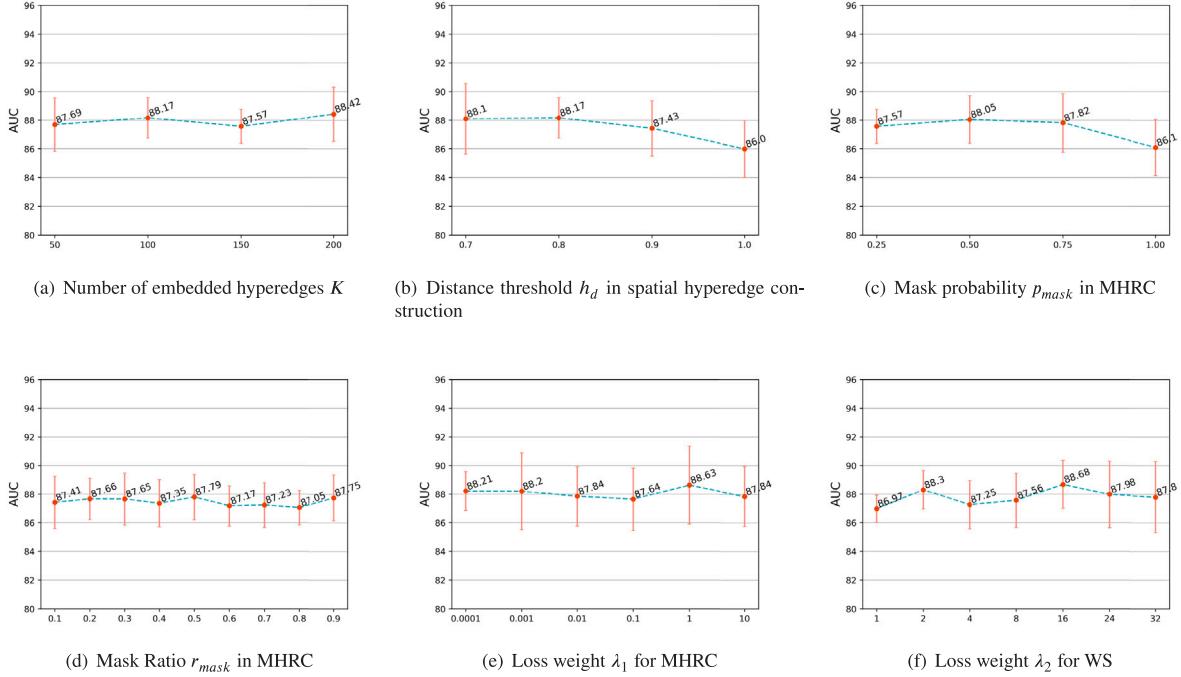


Fig. 3. Performance curves on the validation dataset in the five-fold cross-validation, where the error bar indicates the standard deviation of the AUCs.

that, when using the MHRC module alone (B), the improvement does not appear as significant as when combined with WS. It has been suggested the MHRC module can be more effective after the WS module further enhances the instance-level embeddings.

Moreover, refers to Table 3 (D) and the full version of MaskHGL, a great improvement (4.22% on ACC and 3.98% on AUC) is observed when the MHRC module is cooperating with WS, which indicates the ability of MHRC in improving generalization. We consider it is likely to be due to the fact that the learnable **MASK Node** can learn discriminative representation related to slide-level hard labels thus alleviating the noise brought by the WS module. At the same time, the WS module refines instance-level embeddings which makes the reconstruction process more instructive.

#### 4.5. Hyper-parameters verification

Now, we systematically explore the effect of hyper-parameters in MaskHGL on the *USTC-EGFR* dataset. We will discuss the impact of the following parameters: (1) number of embedded hyperedges  $K$ , (2) distance threshold  $h_d$  in spatial hyperedge construction, (3) mask probability  $p_{mask}$  and (4) mask ratio  $r_{mask}$  in MHRC and (5–6) the final loss weights  $\lambda_1$ ,  $\lambda_2$ .

(1) *Number of embedded hyperedges  $K$* :  $K$  equals the number of embedded hyperedges as described in Section III.B. This parameter is an important factor affecting the computational complexity of MaskHGL. We tuned  $K$  in the range of [50, 200] with a step of 50. The curve in Fig. 3a shows that the classification performance shakes in a range of less than 1%. The insensitivity comes from the global clustering step. As long as the  $K$  is not extremely small (e.g. <10), the global clusterer should be able to capture valuable phenotype information from enormous patches. In the end, we choose  $K = 100$  as the optimal parameter for gaining high accuracy and low computational complexity.

(2) *Distance threshold in spatial hyperedge construction  $h_d$* :  $h_d$  introduced in Algorithm 1 controls the number of spatial hyperedges. The higher it is, the larger each cluster will be, and the smaller the number of spatial hyperedges. When it comes to  $h_d = 1$ , there will be only one spatial hyperedge that contains all nodes. When it is too

small, the agglomerative clustering will fail. Experimentally, we tuned  $h_d \in [0.7, 0.8, 0.9, 1.0]$  for verification. From Fig. 3b, it can be clearly seen that the model performance will decrease when  $h_d$  exceeds a certain threshold (e.g. 0.9). That is because as the clusters grow too large, they will lose the characteristics that help to distinguish their phenotype patterns and eventually result in a significant performance decline. Thus, it is important to keep the number of spatial hyperedges at a fitting level for capturing spatial information. In the end, Fig. 3b indicates that  $h_d = 0.8$  is an appropriate value to achieve the best performance.

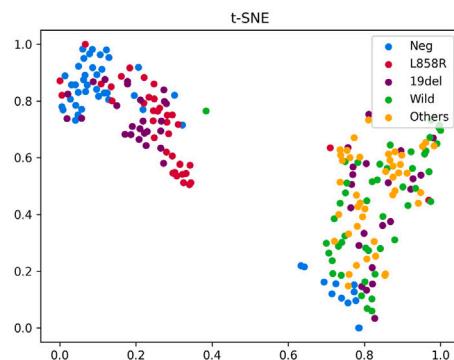
(3) *Mask probability  $p_{mask}$* : When we consider the MHRC module as a type of data augmentation, it is natural to consider the probability of augmentation. Following the results shown in Fig. 3c, model performance is quite stable until every WSI is masked ( $p_{mask}$  reaches 1.0) and performs better when  $p_{mask}$  is relatively high (0.50 or 0.75 compared with 0.25 in [44]), it is another evidence that MHRC module has the ability to alleviate the instance-level noise since negative patches generally dominate even in positive slide. We set a relatively high  $p_{mask} = 0.75$  as default.

(4) *Mask ratio  $r_{mask}$* :  $r_{mask}$  controls the proportion of masked tokens when MHRC is on during training. We tested it in the range of [0.1, 0.9] with a step of 0.1. From Fig. 3d, we noticed that the results are not sensitive to the ratio. The key is whether mask augmentation and reconstruction have been performed. As long as the **Mask Node** can eventually converge during training, the information contained in it is sufficient to improve the robustness of the network. Experimentally, we choose  $r_{mask} = 0.9$  as the default setting.

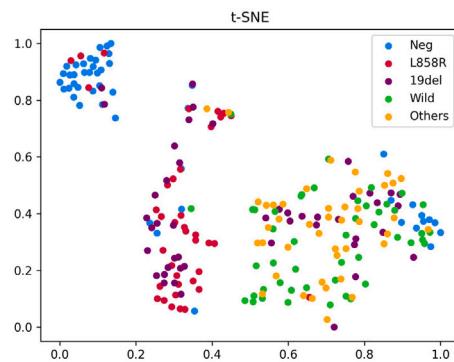
(5–6) *Loss weights for MHRC and WS, respectively denoted as  $\lambda_1$  and  $\lambda_2$* :  $\lambda_1$  and  $\lambda_2$  controls the contribution from each module. As shown in Fig. 3e and 3f,  $\lambda_2$  is more sensitive than  $\lambda_1$ . Since the WS module is the main contributor to the noise, it is important to balance its impact with the benefits of the MHRC module. Empirically, we set  $\lambda_1 = 1$  and  $\lambda_2 = 16$ .

#### 4.6. Interpretability and visualization

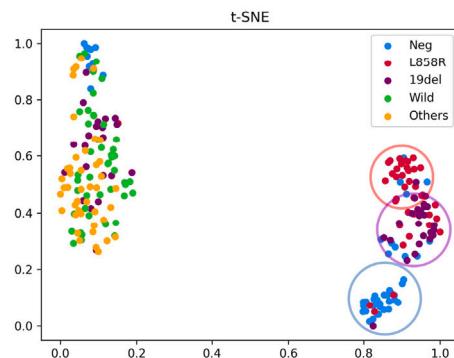
In this section, we will further investigate the interpretability of the proposed MaskHGL. As shown in Fig. 5, each column shows a typical



(a) t-SNE Visualization of SlideGraph+



(b) t-SNE Visualization of HyperGAT



(c) t-SNE Visualization of MaskHGL

**Fig. 4.** T-SNE visualization for three representative graph-based methods on the USTC-EGFR test set.

visualization example for each positive class (L858R, 19del, Wild, Others) in the *USTC-EGFR* dataset. Note that it is difficult to obtain the mutation-related tissue regions directly from H&E-stained WSI. Therefore, we exhibit the lesion tissue annotated by pathologists and try to investigate the model interpretability for the potential mutation-related sites within the lesion tissue. The first two rows display the thumbnail and lesion tissue mask for each slide, and the last row shows the visualization heatmaps of MaskHGL. More visualization examples can be found in Supplementary material.

The proposed MaskHGL provides attention-based visualization, the attention scores generated by the SAA module reflect the concerned regions for downstream tasks. From Fig. 5, the consistency between

**Table 4**

Comparison of computational costs and inferential efficiency.

Method	Inference time (ms/WSI)	Param (M)	FLOPs (G)
CLAM-SB	6.728	0.001	9.008
CLAM-MB	9.765	0.001	9.008
TransMIL	14.257	2.345	47.686
Patch-GCN	20.069	1.105	16.490
NAGCN	17.233	1.643	0.004
SlideGraph+	57.625	0.008	3.240
<i>MaskHGL</i>	36.606	1.976	34.164

fine-grained annotation areas and visualized heatmaps illustrates that our proposed MaskHGL is able to accurately distinguish suspicious lesion regions at the instance level and also indicates the proposed MaskHGL can predict the lesion regions through weakly supervised learning and also infer the gene mutation status and possible sites directly from H&E-stained WSI. It is more promising and less costly for clinical practice compared with conventional technology (e.g. PCR or NGS).

Other than the above visualization of attention scores, we also demonstrated the patches that belong to the same embedded hyperedge in Fig. 6, which indicates a clear uniformity on phenotype patterns from the same embedded hyperedge. Go a step further, according to Fig. 5, Fig. 6 and other visualization examples in Supplementary material, we found that the 19del mutation and L858R mutation have respective uniform visual patterns, which might lead to the finding of potential biomarkers. The tumours with the 19del mutation typically exhibit glandular and cribriform patterns with mucin secretion. The high response areas are surrounded by the formation of papillary while the low response areas overlap the micropapillary structures alongside mucin lakes. Additionally, for tumours with the L858R mutation, high-attention areas overlap the predominant papillary and micropapillary structures, accompanied by pronounced stromal reactions and scattered lymphocytic infiltration.

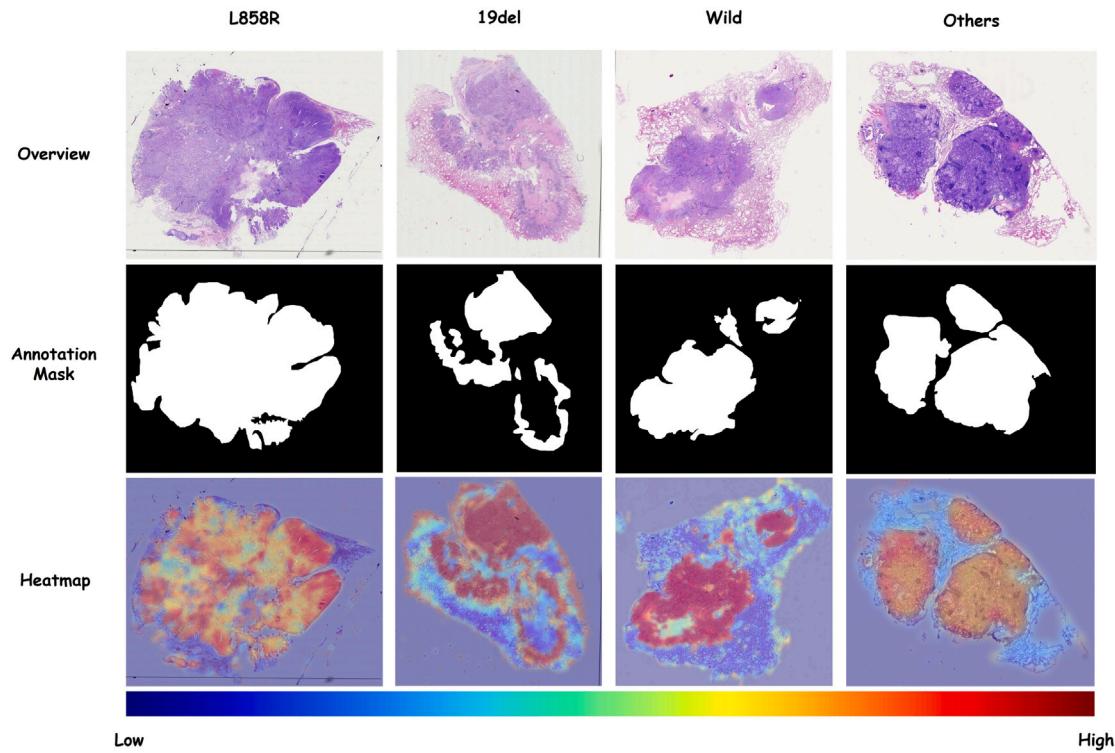
#### 4.7. Computational and inferential efficiency

Given that computational complexity and inferential efficiency can affect the flexibility of practical deployment, comparative experiments on computational complexity and inference time consumption with baselines are conducted. The results are shown in Table 4.

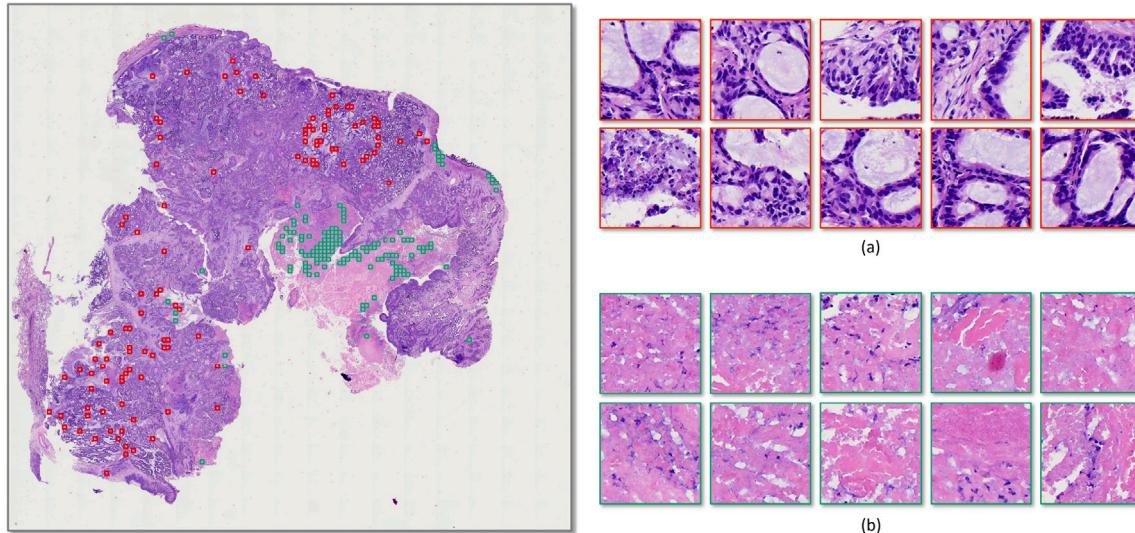
The experimental results indicate that the proposed MaskHGL requires relatively more computational resources yet does not exceed too much. The main computational workload comes from two aspects: (1) the SAA module involves the multi-head self-attention (MHSA) module, which requires a larger amount of computation during inference than other pooling methods. (2) the non-sampling strategy amplifies the effect brought by the self-attention mechanism, although our intention was to preserve visual representations potentially associated with mutation sites as much as possible. An opposite example is NAGCN which minimizes the number of input nodes by clustering, but also makes it difficult to achieve good performance on most tasks.

#### 5. Conclusion

In this paper, we present a novel Masked HyperGraph Learning framework, called MaskHGL, for weakly supervised WSI classification, which analyses WSI from multi-perspectives and exploits non-pairwise relationships among its patches through the hypergraph architecture built by the refined hypergraph construction strategy. More importantly, a masked hypergraph reconstruction (MHRC) module is proposed to alleviate data scarcity and improve the robustness of MaskHGL. Instance-level supervision cooperated with the MHRC module and eventually achieved a remarkable improvement. Besides, we further apply a simple yet efficient self-attention-based aggregator to



**Fig. 5.** Visualization for our proposed framework MaskHGL. Each column shows a typical visualization example for each class in the *USTC-EGFR* dataset. The first row shows the thumbnails for each slide, second row provides lesion tissue masks annotated by pathologists. The last row shows the heatmap visualization output from MaskHGL.



**Fig. 6.** Visualization of embedded hyperedge. We randomly selected 10 patches from two embedded hyperedge to visualize, where the patches with boxes of the same colour belong to the same hyperedge. The red markings (a) are more from the high response areas in the attentional heatmap, while the green markings (b) are the opposite.

explore the global dependence among patches and thus generate more discriminative slide-level representation and enhance the patch-level predictions. Extensive experiments on two public datasets and one in-house dataset validate the effectiveness of our proposed framework, and more importantly, demonstrate the potential of MaskHGL in the tasks including cancer subtyping and gene mutation prediction directly from H&E-stained histopathology WSIs in clinical practice.

Although the proposed MaskHGL achieved promising performance in various tasks, some limitations still exist. Due to the application of

the self-attention mechanism and the non-sampling strategy, MaskHGL requires more computational resources. Therefore, one of the future works is to lower the computational requirements of our work to improve the feasibility of deployment. In future work, we will further discuss the possibility of the proposed mask reconstruction based augmentation strategy becoming a general hypergraph augmentation method. Additionally, we will try to apply hypergraph learning for the mutation prediction of Anaplastic lymphoma kinase (ALK) and ROS oncogene 1 (ROS1) gene fusions in NSCLC from H&E-stained WSIs.

## Research data for this article

The approvals of the Ethics Committee were available for the private USTC-EGFR dataset. Since we do not have the necessary permissions on the USTC-EGFR dataset, the data is not shared. For the public datasets TCGA-LUNG and TCGA-EGFR, we have cited the required references.

## Ethics statement

Data used in this study includes the publicly available TCGA-LUNG and TCGA-EGFR datasets, which have been ethically approved for research purposes. All data from the USTC-EGFR dataset were anonymized and approved by the Medical Research Ethics Committee of the First Affiliated Hospital of the University of Science and Technology of China (Anhui Provincial Hospital) under protocol No. 2022-RE-454 for use in this study. No animals were used in this study, and all experiments involving human data were performed in compliance with the relevant ethical guidelines and regulations.

## CRediT authorship contribution statement

**Jun Shi:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization. **Tong Shu:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kun Wu:** Software, Data curation. **Zhiguo Jiang:** Supervision, Project administration, Funding acquisition. **Liping Zheng:** Supervision, Resources. **Wei Wang:** Validation, Resources, Data curation. **Haibo Wu:** Validation, Resources, Funding acquisition, Data curation. **Yushan Zheng:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

## Declaration of competing interest

No conflict of interest exists in the submission of this manuscript, and all the authors listed have approved the manuscript that is enclosed.

## Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (Grant No. 61906058, 61901018, 62171007 and 61771031), partly supported by Beijing Natural Science Foundation (Grant No. 7242270), partly supported by the Fundamental Research Funds for the Central Universities of China (Grant No. JZ2022HGTB0285), partly supported by Emergency Key Program of Guangzhou Laboratory (Grant No. EKPG21-32), partly supported by Joint Fund for Medical Artificial Intelligence (Grant No. MAI2023C014), partly supported by National Key Research and Development Program of China (Grant No. 2021YFF1201000) and partly supported by Anhui Provincial Health and Medical Research Project (Grant No. AHWJ2023A10143).

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.cmpb.2024.108237>.

## References

- [1] M. Liu, L. Hu, Y. Tang, C. Wang, Y. He, C. Zeng, K. Lin, Z. He, W. Huo, A deep learning method for breast cancer classification in the pathology images, *IEEE J. Biomed. Health Inf.* 26 (10) (2022) 5025–5032.
- [2] J. Shi, R. Wang, Y. Zheng, Z. Jiang, H. Zhang, L. Yu, Cervical cell classification with graph convolutional network, *Comput. Methods Programs Biomed.* 198 (2021) 105807.
- [3] Y. Zheng, J. Li, J. Shi, F. Xie, J. Huai, M. Cao, Z. Jiang, Kernel attention transformer for histopathology whole slide image analysis and assistant cancer diagnosis, *IEEE Trans. Med. Imaging* (2023).
- [4] X. Sun, W. Li, B. Fu, Y. Peng, J. He, L. Wang, T. Yang, X. Meng, J. Li, J. Wang, et al., TGML: A hybrid multi-instance learning model based on the transformer and the graph attention network for whole-slide images classification of renal cell carcinoma, *Comput. Methods Programs Biomed.* 242 (2023) 107789.
- [5] M. Liang, Q. Chen, B. Li, L. Wang, Y. Wang, Y. Zhang, R. Wang, X. Jiang, C. Zhang, Interpretable classification of pathology whole-slide images using attention based context-aware graph convolutional neural network, *Comput. Methods Programs Biomed.* 229 (2023) 107268.
- [6] P. Huang, X. Tan, X. Zhou, S. Liu, F. Mercaldo, A. Santone, FABNet: fusion attention block and transfer learning for laryngeal cancer tumor grading in P63 IHC histopathology images, *IEEE J. Biomed. Health Inf.* 26 (4) (2021) 1696–1707.
- [7] P. Huang, P. He, S. Tian, M. Ma, P. Feng, H. Xiao, F. Mercaldo, A. Santone, J. Qin, A ViT-AMC network with adaptive model fusion and multiobjective optimization for interpretable laryngeal tumor grading from histopathological images, *IEEE Trans. Med. Imaging* 42 (1) (2022) 15–28.
- [8] H. Wang, G. Huang, Z. Zhao, L. Cheng, A. Juncker-Jensen, M.L. Nagy, X. Lu, X. Zhang, D.Z. Chen, CCF-GNN: A unified model aggregating appearance, microenvironment, and topology for pathology image classification, *IEEE Trans. Med. Imaging* (2023).
- [9] D. Di, J. Zhang, F. Lei, Q. Tian, Y. Gao, Big-hypergraph factorization neural network for survival prediction from whole slide image, *IEEE Trans. Image Process.* 31 (2022) 1149–1160.
- [10] W. Shao, J. Liu, Y. Zuo, S. Qi, H. Hong, J. Sheng, Q. Zhu, D. Zhang, FAM3L: Feature-aware multi-modal metric learning for integrative survival analysis of human cancers, *IEEE Trans. Med. Imaging* (2023).
- [11] L. Zhao, R. Hou, H. Teng, X. Fu, Y. Han, J. Zhao, CoADS: Cross attention based dual-space graph network for survival prediction of lung cancer using whole slide images, *Comput. Methods Programs Biomed.* 236 (2023) 107559.
- [12] P. Liu, L. Ji, F. Ye, B. Fu, GraphLSurv: A scalable survival prediction network with adaptive and sparse structure learning for histopathological whole-slide images, *Comput. Methods Programs Biomed.* 231 (2023) 107433.
- [13] R. Yamashita, J. Long, T. Longacre, L. Peng, G. Berry, B. Martin, J. Higgins, D.L. Rubin, J. Shen, Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study, *Lancet Oncol.* 22 (1) (2021) 132–141.
- [14] W. Lu, M. Toss, M. Dawood, E. Rakha, N. Rajpoot, F. Minhas, SlideGraph+: Whole slide image level graphs to predict HER2 status in breast cancer, *Med. Image Anal.* 80 (2022) 102486.
- [15] Y. Wei, X. Chen, L. Zhu, L. Zhang, C.-B. Schönlieb, S. Price, C. Li, Multi-modal learning for predicting the genotype of glioma, *IEEE Trans. Med. Imaging* (2023).
- [16] S.A. Javed, D. Juyal, H. Padigela, A. Taylor-Weiner, L. Yu, A. Prakash, Additive MIL: intrinsically interpretable multiple instance learning for pathology, *Adv. Neural Inf. Process. Syst.* 35 (2022) 20689–20702.
- [17] M.Y. Lu, D.F. Williamson, T.Y. Chen, R.J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, *Nat. Biomed. Eng.* 5 (6) (2021) 555–570.
- [18] R.J. Chen, M.Y. Lu, M. Shaban, C. Chen, T.Y. Chen, D.F. Williamson, F. Mahmood, Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks, in: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII* 24, Springer, 2021, pp. 339–349.
- [19] F. Scarselli, M. Gori, A.C. Tsai, M. Hagenbuchner, G. Monfardini, The graph neural network model, *IEEE Trans. Neural Netw.* 20 (1) (2008) 61–80.
- [20] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* 86 (11) (1998) 2278–2324.
- [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*, 2021, OpenReview.net.
- [23] X. Zhang, M. Cao, S. Wang, J. Sun, X. Fan, Q. Wang, L. Zhang, Whole slide cervical cancer screening using graph attention network and supervised contrastive learning, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, pp. 202–211.
- [24] Y. Guan, J. Zhang, K. Tian, S. Yang, P. Dong, J. Xiang, W. Yang, J. Huang, Y. Zhang, X. Han, Node-aligned graph convolutional network for whole-slide image representation and classification, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18813–18823.
- [25] D. Di, C. Zou, Y. Feng, H. Zhou, R. Ji, Q. Dai, Y. Gao, Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction, *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (5) (2022) 5800–5815.
- [26] T.H. Chan, F.J. Cendra, L. Ma, G. Yin, L. Yu, Histopathology whole slide image analysis with heterogeneous graph representation learning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15661–15670.

- [27] Y. Cai, Z. Zhang, Z. Cai, X. Liu, X. Jiang, Hypergraph-structured autoencoder for unsupervised and semisupervised classification of hyperspectral image, *IEEE Geosci. Remote Sens. Lett.* 19 (2022) 1–5.
- [28] J.H. Giraldo, V. Scarrica, A. Staiano, F. Camastrà, T. Bouwmans, Hypergraph convolutional networks for weakly-supervised semantic segmentation, in: 2022 IEEE International Conference on Image Processing, ICIP, IEEE, 2022, pp. 16–20.
- [29] Y. Gao, Y. Feng, S. Ji, R. Ji, HGNN+: General hypergraph neural networks, *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (3) (2022) 3181–3199.
- [30] S. Bai, F. Zhang, P.H. Torr, Hypergraph convolution and hypergraph attention, *Pattern Recognit.* 110 (2021) 107637.
- [31] Y. Feng, H. You, Z. Zhang, R. Ji, Y. Gao, Hypergraph neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3558–3565.
- [32] H. Gao, Y. Liu, S. Ji, Topology-aware graph pooling networks, *IEEE Trans. Pattern Anal. Mach. Intell.* 43 (12) (2021) 4512–4518.
- [33] H. Wu, M.K. Ng, Hypergraph convolution on nodes-hyperedges network for semi-supervised node classification, *ACM Trans. Knowl. Discov. Data (TKDD)* 16 (4) (2022) 1–19.
- [34] J. Wang, K. Ding, L. Hong, H. Liu, J. Caverlee, Next-item recommendation with sequential hypergraphs, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1101–1110.
- [35] X. Chen, K. Xiong, Y. Zhang, L. Xia, D. Yin, J.X. Huang, Neural feature-aware recommendation with signed hypergraph convolutional network, *ACM Trans. Inf. Syst. (TOIS)* 39 (1) (2020) 1–22.
- [36] H. Fan, F. Zhang, Y. Wei, Z. Li, C. Zou, Y. Gao, Q. Dai, Heterogeneous hypergraph variational autoencoder for link prediction, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (8) (2021) 4125–4138.
- [37] W. Hou, C. Lin, L. Yu, J. Qin, R. Yu, L. Wang, Hybrid graph convolutional network with online masked autoencoder for robust multimodal cancer survival prediction, *IEEE Trans. Med. Imaging* 42 (8) (2023) 2462–2473.
- [38] S. Li, Y. Zhao, J. Zhang, T. Yu, J. Zhang, Y. Gao, High-order correlation-guided slide-level histology retrieval with self-supervised hashing, *IEEE Trans. Pattern Anal. Mach. Intell.* (2023).
- [39] X. Xu, S. Xu, L. Jin, E. Song, Characteristic analysis of otsu threshold and its applications, *Pattern Recognit. Lett.* 32 (7) (2011) 956–961.
- [40] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650–9660.
- [41] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 2127–2136.
- [42] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, V. Singh, Nyströmformer: A nyström-based algorithm for approximating self-attention, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 14138–14148.
- [43] S. Alaparthi, M. Mishra, BERT: A sentiment analysis odyssey, *J. Mark. Anal.* 9 (2) (2021) 118–126.
- [44] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
- [45] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, Yolov4: Optimal speed and accuracy of object detection, 2020, arXiv preprint [arXiv:2004.10934](https://arxiv.org/abs/2004.10934).
- [46] D.A. Gutman, J. Cobb, D. Somanan, Y. Park, F. Wang, T. Kurc, J.H. Saltz, D.J. Brat, L.A. Cooper, J. Kong, Cancer digital slide archive: an informatics resource to support integrated in silico analysis of TCGA pathology data, *J. Am. Med. Inf. Assoc.* 20 (6) (2013) 1091–1098.
- [47] H. Shigematsu, L. Lin, T. Takahashi, M. Nomura, M. Suzuki, I.I. Wistuba, K.M. Fong, H. Lee, S. Toyooka, N. Shimizu, et al., Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers, *J. Natl. Cancer Inst.* 97 (5) (2005) 339–346.
- [48] W.-Q. Li, J.-W. Cui, Non-small cell lung cancer patients with ex19del or exon 21 L858R mutation: distinct mechanisms, different efficacies to treatments, *J. Cancer Res. Clin. Oncol.* 146 (9) (2020) 2329–2338.
- [49] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).
- [50] L.N. Smith, N. Topin, Super-convergence: Very fast training of neural networks using large learning rates, in: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, vol. 11006, SPIE, 2019, pp. 369–386.
- [51] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al., Transmil: Transformer based correlated multiple instance learning for whole slide image classification, *Adv. Neural Inf. Process. Syst.* 34 (2021) 2136–2147.
- [52] A. Ross, V.L. Willson, A. Ross, V.L. Willson, Paired samples T-test, in: Basic and Advanced Statistical Tests: Writing Results Sections and Creating Tables and Figures, Springer, 2017, pp. 17–19.