How Graph Neural Networks Enhance Convolutional Neural Networks Towards Mining the Topological Structures from Histology

Yiqing Shen [*1] **Bingxin Zhou** [*2 3] **Xinye Xiong** [1] **Ruitian Gao** [4] **Yu Guang Wang** [1 3]

Abstract

Deep learning methods play an increasingly significant role as feature extractors. Existing solutions heavily rely on convolutional neural networks (CNNs) for global pixel-level analysis, leaving the underlying local geometric structure such as the interaction between cells unexplored. The topological structure in medical images, as proven to be closely related to tumor evolution, can be well characterized by graphs. To obtain a more comprehensive representation for downstream oncology tasks, we propose a fusion framework for enhancing the global image-level representation captured by CNNs with the geometry of cell-level spatial information learned by graph neural networks (GNN). The fusion layer optimizes an integration between collaborative features of global images and cell graphs. Two fusion strategies have been developed: one with MLP which is simple but turns out efficient through fine-tuning, and the other with TRANSFORMER gains a champion in fusing multiple networks. We evaluate our fusion strategies on histology datasets curated from large patient cohorts of colorectal and gastric cancers for three biomarker prediction tasks. Both two models outperform plain CNNs or GNNs, reaching a consistent AUC improvement of more than 5% on various network backbones. The experimental results yield the necessity for combining image-level morphological features with cell spatial relations in medical image analysis. Code is available here.

[*]Equal contribution [1]Institute of Natural Sciences, School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, 200240, China [2]The University of Sydney Business School, The University of Sydney, Camperdown, NSW 2006, Australia. [3]Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai, 200240, China [4]Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai, 200240, China. Correspondence to: Yu Guang Wang <yuguang.wang@sjtu.edu.cn>.

The 2022 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2022.

1. Introduction

Histology provides a wealth of complex patterns and morphological features for deep learning algorithms to mine. Existing approaches routinely employ end-to-end convolutional neural networks (CNNs) frameworks, by taking the morphological and textural image features as input. Numerous practices with CNNs have been made in diagnostic and prognostic tasks, such as lesion detection, gene mutation identification, molecular biomarker classification, and patient survival analysis from Hematoxylin and Eosin (H&E) stained histology whole-slide images (WSIs) (Shaban et al., 2019; Fu et al., 2020; Liao et al., 2020; Calderaro & Kather, 2021; Echle et al., 2021). Determined by the kernel in convolutional layers, which are initially targeted to analyze fixed connectivity between local areas (*i.e.*, pixel grids), CNNs focus on extracting image-level feature representations. However, no guidance has been imposed explicitly on CNNs to exploit the underlying topology from histology, *e.g.*, the cell-cell interaction and the spatial distribution of cells, which have been clinically proven to be closely related to tumor evolution and biomarker expression (Galon et al., 2006; Feichtenbeiner et al., 2014; Barua et al., 2018; Noble et al., 2022). As a result, the recognition of the cell dispersal manner and their mutual interactions are essential for training robust and interpretable deep learning models (Gunduz et al., 2004; Yener, 2016; Wang et al., 2021).

Mathematically, the topological structures and cell relationships are formulated by graphs. By its definition, a graph can characterize the relationship between nodes, *e.g.*, superpixels in natural images, or the cells in histological images. Following the establishment of graphs, graph neural networks (GNNs) were proposed to learn the geometric information (Bronstein et al., 2017; Wu et al., 2020; Zhang et al., 2020). While CNNs are capable of learning global image representation, GNNs can provide machinery for the local topological features. Both global and local features serve as significant representation in learning the mapping of histological image space to clinical meaningful biomarkers. One strategy is to make use of the own merits of CNN and GNN models. Some recent attempts at combining GNNs with CNNs have achieved satisfactory performance boost in natural image classification tasks, such as remote sensing

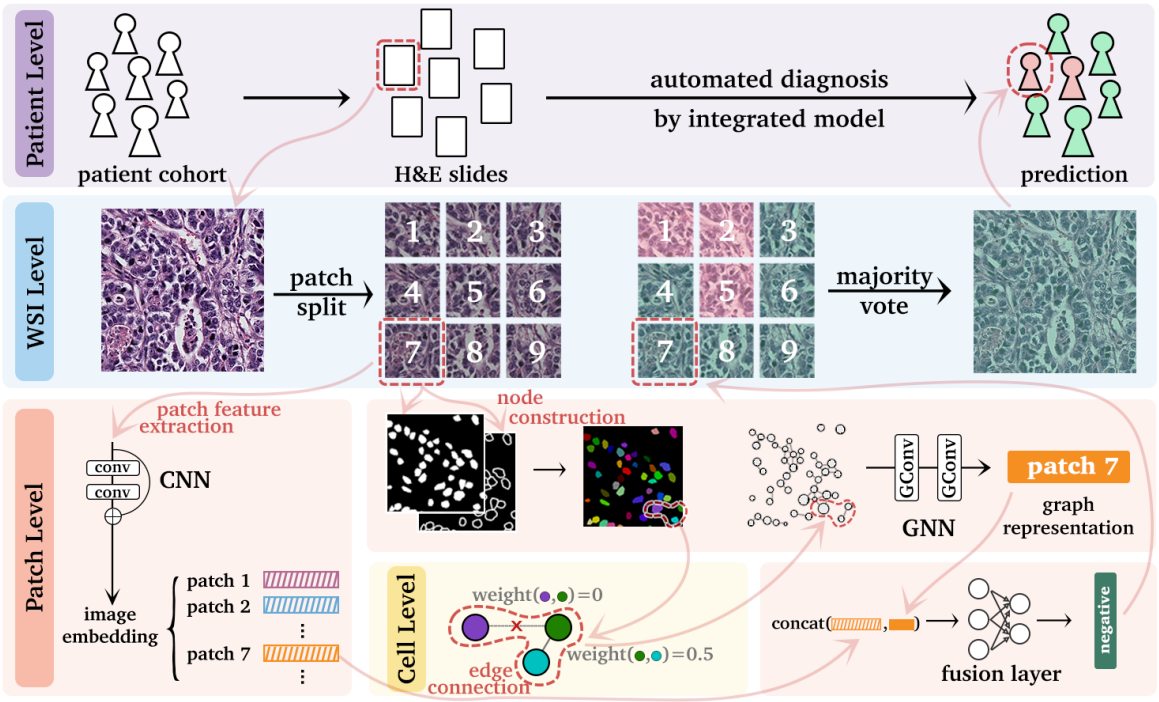

Figure 1. The overall workflow of the proposed fusion scheme for GNN and CNN in the patient-level diagnostic task. For each image patch tessellated from WSI, a cell graph is first generated to characterize the topological structure by segmenting the nuclei region with CA$^{2.5}$-Net as graph nodes, and extracting the pre-defined pathomics feature. In the training and inference stage, the global image-level representation and geometric representation discovered by CNN and GNN are integrated by a fusion layer to obtain a more comprehensive feature representation. The patient-level prediction is finally determined by a majority vote from all patch predictions.

scene recognition (Liang et al., 2020; Peng et al., 2022) and hyper-spectral image prediction (Dong et al., 2022).In the medical imaging domain, Wei et al. (2022) predicted isocitrate dehydrogenase gene mutation with a collaborative learning framework that aligns a CNN for tumor MRI with a GNN for tumor geometric shape analysis. To the best of our knowledge, a study of the interplay between CNNs and GNNs for histology is still absent.

In this paper, we develop an efficient strategy that is able to integrate the structure feature from GNNs with the image feature of CNNs for H&E slides analysis. The fusion scheme partitions a WSI into non-overlapped patches and generates a cell graph for each patch by linking associated cells (see Section 2) to model the cell interactions. Then, a GNN is employed to distill geometric representation. To fuse the graph-level representation learning with image-level embedding, we train the GNN together with the CNN in parallel. The integration takes place in a learnable fusion layer which incorporates the morphology feature of the whole image with the geometric representation of cell graphs. In this way, insights into spatial structure is gained for a specific staining image, such as the distribution of cells, interaction of cancer and healthy cells, and tumor microenvironment.

In practice, we can simply connect a learnable fusion layer using MLP or TRANSFORMER next to the outputs of GNN and CNN modules. The simple amalgamation can produce a model which outperforms a sole GNN or CNN model on real histological image datasets (two public and one private). The key to performance improvement of the fusion model lies in that the local geometry of the cell graphs of patches which can only be perceived by GNNs tops up the global image feature of CNNs. We release the codes and constructed graph datasets, which can serve as a benchmark for future research in image-graph bimodal domain.

2. Integrating CNN with GNN

Method Overview. The complete pipeline of using the proposed GNN and CNN fusion scheme for downstream patient-level prediction is illustrated in Figure. 1. First, we partition a WSI into disjoint patches of the same size *e.g.*, 224×224 pixels in this research. Then, a cell graph is constructed for each patch to describe the topological structure of the image, where the nodes are defined as the cells segmented by a nuclei segmentation network, *i.e.*, CA$^{2.5}$-Net (Huang et al., 2021). Subsequently, GNN and CNN extract the geometric representation from the cell graph of a patch and the global image-level presentation respectively. Finally, the output image and graph embeddings are fused by

a learnable layer with MLP or TRANSFORMER (Figure 2).

Cell Graph. For each patch, we can establish a cell graph. The graph nodes (v_i, with its subscript i representing the node index) in a cell graph are biologically determined as the nuclei regions. The cell graph can represent the cell-cell interaction and the collection of cell graphs for all patches provide a precise characterization of tumor microenvironment. With only the availability of raw image patches, we leverage the nuclei region segmented by a well-tuned CA$^{2.5}$-Net (Huang et al., 2021) to extract the node features of each single nuclei node (See Appendix B). As the morphological signals are believed relative to cell-cell interplay, the cell-specific features $\boldsymbol{X}$, which include the nuclei coordination, optical, textual representations, then characterize the cell-level morphological behavior.

We then calculate the pair-wise Euclidean distance between nuclei centroids to establish edges of a cell graph (Wang et al., 2021) to quantify the interplay between cells in a patch. To be precise, for arbitrary two nuclei nodes v_i and v_j, with their associated centroid Cartesian coordinates (x_i, y_i) and (x_j, y_j), the edge weight w_{ij} for the interaction between two nodes reads

$$w_{ij} := \begin{cases} d_c/d(v_i, v_j), & d(v_i, v_j) \leq d_c \text{ pixels}, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $d(v_i, v_j)$ regards the Euclidean distance between v_i and v_j. From the clinical observations, two cells do not exert mutual influence with their centroid distance exceeding d_c (Barua et al., 2018). Thus, the critical distance d_c depicts the range where a cell can interact with another. Note that the precise value of d_c depends on the tissue structure, image category, and magnification of the WSI. An edge e_{ij} exists between v_i and v_j if and only if the weight $w_{ij} > 0$.

Geometric Feature Representation. For notation simplicity, we denote the generated cell graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \boldsymbol{X})$, where graph nodes in the collection $\mathcal{V}$ cover all nuclei regions, $\mathcal{E}$ is the set of all edges e_{ij} with the corresponding attribute w_{ij} describes the pair-wise cell interaction. The weighted edge is recorded in an adjacency matrix $\boldsymbol{A}$ with its element $\boldsymbol{A}_{ij} = w_{ij}$. When a patch is acquired as a cell graph $\mathcal{G}$, its geometric feature representation can be gradually learned by a graph neural network. The ℓth layer of the GNN finds hidden representation of the cell graph by

$$\boldsymbol{H}_{\mathcal{G}}^{\ell} = \texttt{ReLU}\big(\texttt{GraphConv}\left(\boldsymbol{A}, \boldsymbol{H}^{\ell-1}\right)\big). \tag{2}$$

The representation $\boldsymbol{H}_{\mathcal{G}}^{\ell}$ embeds spatial topological structures of the underlying graph, which is then sent to a readout layer, for example, a fully-connected (FC) layer before eventually being fed into the fusion layer. This FC layer helps to align the feature dimensions of the GNN with the parallel CNN output.

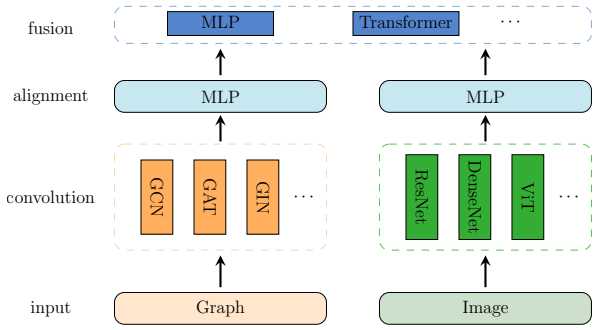

Figure 2. A schematic illustration for fusing the representations of local geometry from GNN and global image features from CNN.

Image-level Feature Representation. The image-level feature representation is directly extracted from histology patches by CNNs. For instance, denote $\{\boldsymbol{H}_{\mathcal{I}}^{1}, \ldots, \boldsymbol{H}_{\mathcal{I}}^{\ell-1}\}$ the output of the first $\ell - 1$ blocks after convolution layers. We can use different convolutional module for the CNN. For example, a DENSENET (Huang et al., 2017) defines

$$\boldsymbol{H}_{\mathcal{I}}^{\ell} = \texttt{ReLU}\big(\texttt{Conv}(\texttt{concat}[\boldsymbol{H}_{\mathcal{I}}^{1}, \ldots, \boldsymbol{H}_{\mathcal{I}}^{\ell-1}])\big). \tag{3}$$

Alternatively, RESNET (He et al., 2016) finds $\boldsymbol{H}_{\mathcal{I}}^{\ell}$ by

$$\boldsymbol{H}_{\mathcal{I}}^{\ell} = \texttt{ReLU}\big(\texttt{Conv}(\boldsymbol{H}^{\ell-1}) + \boldsymbol{H}^{\ell-1}\big) \tag{4}$$

with some activated convolutional layers $\texttt{conv}(\cdot)$. The residual connection in the second design can reduce the computational cost of deep CNNs and circumvent gradient diminishing. In the empirical study, a lightweight architecture namely MOBILENETV3 (Howard et al., 2019) is considered, where efficient depth-wise separable convolutions replaces traditional convolution layers. For all CNN blocks, we assign the input feature $\boldsymbol{H}_{\mathcal{I}}^{0}$ by staining normalized histology image patches. In the same fashion as geometric feature representation, the final image representation is fed to a learnable fully-connected layer to adjust the embedding feature dimensions.

Learnable Feature Fusion Layer. Denote the output image and graph representation for an arbitrary patch by $\boldsymbol{H}_{\mathcal{I}}$ and $\boldsymbol{H}_{\mathcal{G}}$. We then train the fusion layer to learn the optimal integration between them. In particular, we consider two candidates of MLP and TRANSFORMER (Vaswani et al., 2017) for fusing. The former approaches fused representation $\boldsymbol{H}_o$ by

$$\boldsymbol{H}_o = \texttt{Linear}(\texttt{MLPBlock}(...(\texttt{MLPBlock}(\boldsymbol{H}_c)))),$$

where $\boldsymbol{H}_c = \texttt{concat}\big[\boldsymbol{H}_{\mathcal{I}}, \boldsymbol{H}_{\mathcal{G}}\big]$, and

$$\texttt{MLPBlock}(\boldsymbol{H}) = \texttt{Dropout}(\texttt{ReLU}(\texttt{Linear}(\boldsymbol{H}))).$$

Table 1. Test ACC and AUC comparisons on three benchmarks. We compute the mean and standard deviation over seven random runs.

	GIST-PDL1			CRC-MSI			STAD-MSI		
Model	ACC	AUC	AUC*patient*	ACC	AUC	AUC*patient*	ACC	AUC	AUC*patient*
GCN2	68.52±1.71	73.34±1.33	58.23±2.54	66.78±2.12	56.52±1.48	51.10±6.52	69.33±6.75	55.89±1.31	62.91±1.91
GIN2	71.94±1.37	77.18±0.81	62.86±1.43	66.61±1.79	57.01±0.91	44.60±3.39	71.33±3.52	60.31±1.22	66.75±4.09
MOBILENETV3	73.62±2.14	86.79±1.24	83.92±4.08	72.97±0.75	66.29±0.55	64.65±3.09	75.29±1.26	66.90±2.11	72.77±1.15
MOBILENETV3-GCN2-MLP	77.18±0.68	88.74±0.51	91.98±0.36	72.72±0.34	**73.46±0.60**	77.51±2.82	76.08±0.42	71.87±0.86	73.12±0.99
MOBILENETV3-GIN2-MLP	74.95±1.23	89.38±0.36	94.24±2.28	73.41±0.20	69.08±3.80	78.12±3.98	75.82±0.37	69.80±1.01	73.24±0.79
MOBILENETV3-GCN2-TRANS	**77.89±1.17**	90.47±0.86	**96.74±1.02**	73.16±0.44	71.04±0.76	77.81±3.70	**76.31±0.41**	**73.63±0.70**	74.32±0.64
MOBILENETV3-GIN2-TRANS	76.18±1.37	**90.94±0.86**	94.43±1.80	**73.48±0.37**	70.53±1.33	**79.53±2.40**	76.25±0.68	73.37±1.44	**74.42±1.51**
DENSENET121	71.18±1.42	82.50±3.25	89.02±4.29	74.16±0.28	69.98±0.91	66.66±4.83	74.94±1.68	65.54±1.08	74.84±0.03
DENSENET121-GCN2-MLP	76.53±0.90	88.84±0.78	95.90±3.50	**75.20±0.29**	70.04±1.08	68.21±2.22	76.61±0.40	74.50±0.99	75.88±1.24
DENSENET121-GIN2-MLP	76.71±0.71	88.01±0.67	94.62±1.19	74.62±0.24	70.14±0.88	71.55±1.32	**77.01±0.33**	**74.80±1.13**	75.42±0.74
DENSENET121-GCN2-TRANS	**79.63±0.76**	**89.79±1.08**	**97.49±1.57**	74.93±0.36	73.99±0.59	**83.20±2.05**	76.71±0.79	73.58±0.71	**76.28±1.06**
DENSENET121-GIN2-TRANS	75.82±1.51	87.70±1.09	96.01±2.00	74.82±0.54	**74.62±0.63**	75.69±3.17	76.84±0.58	74.36±1.00	75.82±0.95
RESNET18	70.65±2.09	82.06±1.53	86.26±1.65	73.53±0.40	65.31±3.95	61.66±4.97	73.75±1.51	72.56±0.95	74.13±0.82
RESNET18-GCN2-MLP	**81.90±3.46**	**92.56±1.52**	94.07±3.43	74.15±0.43	**75.16±0.85**	83.15±1.26	76.01±0.63	73.06±0.79	74.79±0.75
RESNET18-GIN2-MLP	76.26±2.02	87.58±2.48	91.31±2.57	74.52±0.70	69.85±2.19	82.99±1.67	**76.18±0.42**	**74.69±0.89**	75.62±2.17
RESNET18-GCN2-TRANS	76.04±1.78	86.40±2.68	93.66±3.64	**74.79±0.32**	73.64±1.28	**84.70±2.11**	76.10±0.80	72.66±0.45	75.56±1.23
RESNET18-GIN2-TRANS	76.81±1.08	92.05±0.57	**95.53±0.94**	74.61±0.47	73.43±0.77	83.78±1.65	76.17±0.41	74.56±0.24	**75.84±1.71**

The TRANSFORMER fusion scheme formulates $\boldsymbol{H}_o$ by

$$\boldsymbol{H}_o = \texttt{Linear}(\texttt{TransBlock}(\cdots(\texttt{TransBlock}(\boldsymbol{H}_c)))),$$

where $\boldsymbol{H}_c = \texttt{stack}\left[\boldsymbol{H}_{\mathcal{I}}, \boldsymbol{H}_{\mathcal{G}}\right]$ and `TransBlock` writes for the PreNorm variant of TRANSFORMER (Wang et al., 2019). The `stack` operation requires an identical dimension of $\boldsymbol{H}_{\mathcal{I}}$ and $\boldsymbol{H}_{\mathcal{G}}$.

3. Experiments and Results

Dataset. We leverage three H&E stained histology benchmarks, termed as **CRC-MSI**, **STAD-MSI**, and **GIST-PDL1**, for evaluations. The first two datasets are constructed for binary microsatellite instability (MSI) status classification (Kather et al., 2019), where we follow the original train and test split for a fair comparison. We also evaluate the performance of the model on a binary Programmed Death-Ligand 1 (PD-L1) status binary classification dataset, which was curated from 129 well-annotated WSIs of gastric cancer patients. We supplement further details for data collection and descriptions in Appendix A.

Model Configurations. We evaluate the performance gain of our proposed fusion scheme with a comprehensive comparison against three CNN backbones of different scales: MOBILENETV3, DENSENET, and RESNET. We stack two graph convolution layers for graph representation learning. Two candidates of graph convolution GCN (Kipf & Welling, 2017) and GIN (Xu et al., 2018) are taken into account, following a 2-layer TOPK (Cătălina et al., 2018) graph pooling scheme. The graph convolution plays the critical role in extracting the geometric feature of the patch. For the fusion layer, both a 1-layer MLP and TRANSFORMER are validated. We name the models in Table 1 with the adopted model architectures and modules. For instance, RESNET18-GCN2-MLP indicates a 18-layer RESNET for image embedding, 2-layer GCN plus TOPK for graph representation learning, and MLP with one `MLPBlock` for features fusion. Details of the model configurations and training hyper-parameters are elaborated in Appendix E.

Results. Image-level performance is evaluated with two metrics, namely test accuracy (ACC) and area-under-curve (AUC). Additionally, we evaluate patient-level prediction with AUC (AUC*patient*). As shown in Table 1, the fused learning schemes achieve more than 5% performance gain over plain CNNs. The improvement is more significant at the patient level at up to 23%. The additional performance boost suggests that our design of the integrated scheme has better potential to overcome the disturbance of heterogeneous patches for patient-level overall diagnosis. The main takeaways include: 1) An individual GNN fails to achieve satisfactory performance. But as a parallel layer, GNNs can enhance the learning capability of CNN by a learnable fusion layer. 2) MLP, though simple, serves as a good fusion layer. 3) Generally speaking, the TRANSFORMER integrator outperforms the simple MLP scheme. However, one can not tell whether MLP or TRANSFORMER a universally better fusion solution. 4) All the integrated models outperform the plain CNNs or GNNs. 5) For the choice of a GNN module, GCN and GIN do not present a significant advantage one over the other. More empirical investigations are supplemented in Appendix F, including training cost, performance improvement rate, as well as the performance ranking.

4. Discussion

This work proposes a fusion framework of CNN and GNN for biomarker prediction from histology slides. On top of

CNNs which extracts a whole-slide image feature, we integrate GNNs to add a local geometric representation for cell-graph patches. The CNNs and GNNs are trained in parallel and their output features are integrated in a fusion layer. This is important as the fusion scheme addresses the expression of the tumor microenvironment by supplementing topology inside local patches in network training. We validate the framework using different combinations of CNN, GNN and fusion modules on real H&E stained histology datasets, which surpasses the plain CNN or GNN methods to a significant margin.

Acknowledgements

YW acknowledges support from the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102) and SJTU and Huawei ExploreX Funding (SD6040004/034). We are also grateful to the anonymous reviewers for their feedback.

References

Barua, S., Fang, P., Sharma, A., et al. Spatial interaction of tumor cells and regulatory t cells correlates with survival in non-small cell lung cancer. *Lung Cancer*, 117:73–79, 2018.

Bronstein, M. M., Bruna, J., LeCun, Y., et al. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Calderaro, J. and Kather, J. N. Artificial intelligence-based pathology for gastrointestinal and hepatobiliary cancers. *Gut*, 70(6):1183–1193, 2021.

Cătălina, C., Veličković, P., Jovanović, N., et al. Towards sparse hierarchical graph classifiers. In *NeurIPS*, 2018.

Dong, Y., Liu, Q., Du, B., et al. Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. *IEEE Transactions on Image Processing*, 2022.

Echle, A., Laleh, N. G., Schrammen, P. L., et al. Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: a systematic literature review. *ImmunoInformatics*, pp. 100008, 2021.

Feichtenbeiner, A., Haas, M., Büttner, M., Grabenbauer, G. G., Fietkau, R., et al. Critical role of spatial interaction between cd8+ and foxp3+ cells in human gastric cancer: the distance matters. *Cancer Immunology, Immunotherapy*, 63(2):111–119, 2014.

Fu, Y., Jung, A. W., Torne, R. V., et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, 1(8):800–810, 2020.

Galon, J., Costes, A., Sanchez-Cabo, F., et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*, 313(5795): 1960–1964, 2006.

Gunduz, C., Yener, B., and Gultekin, S. H. The cell graphs of cancer. *Bioinformatics*, 20(suppl_1):i145–i151, 2004.

He, K., Zhang, X., Ren, S., et al. Deep residual learning for image recognition. In *CVPR*, 2016.

Howard, A., Sandler, M., Chu, G., et al. Searching for mobilenetv3. In *ICCV*, 2019.

Huang, G., Liu, Z., Van Der Maaten, L., et al. Densely connected convolutional networks. In *CVPR*, 2017.

Huang, J., Shen, Y., Shen, D., et al. Ca 2.5-net nuclei segmentation framework with a microscopy cell benchmark collection. In *MICCAI*, 2021.

Kather, J. N., Pearson, A. T., Halama, N., et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine*, 25 (7):1054–1056, 2019.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

Liang, J., Deng, Y., and Zeng, D. A deep neural network combined cnn and gcn for remote sensing scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:4325–4338, 2020.

Liao, H., Long, Y., Han, R., et al. Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. *Clinical and Translational Medicine*, 10(2), 2020.

Noble, R., Burri, D., Le Sueur, C., et al. Spatial structure governs the mode of tumour evolution. *Nature Ecology & Evolution*, 6(2):207–217, 2022.

Peng, F., Lu, W., Tan, W., et al. Multi-output network combining gnn and cnn for remote sensing scene classification. *Remote Sensing*, 14(6):1478, 2022.

Shaban, M., Khurram, S. A., Fraz, M. M., et al. A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma. *Scientific Reports*, 9(1):1–13, 2019.

Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is all you need. In *NeurIPS*, 2017.

Wang, Q., Li, B., Xiao, T., et al. Learning deep transformer models for machine translation. In *the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Wang, Y., Wang, Y., Hu, C., et al. Cell graph neural networks enable digital staging of tumour microenvironment and precisely predict patient survival in gastric cancer. *medRxiv*, 2021.

Wei, Y., Li, C., Chen, X., et al. Collaborative learning of images and geometrics for predicting isocitrate dehydrogenase status of glioma. *arXiv:2201.05530*, 2022.

Wu, Z., Pan, S., Chen, F., et al. A comprehensive survey on graph neural networks. *IEEE TNNLS*, 32(1):4–24, 2020.

Xu, K., Hu, W., Leskovec, J., et al. How powerful are graph neural networks? In *ICLR*, 2018.

Yener, B. Cell-graphs: image-driven modeling of structure-function relationship. *Communications of the ACM*, 60 (1):74–84, 2016.

Zhang, Z., Cui, P., and Zhu, W. Deep learning on graphs: a survey. *IEEE TKDE*, 2020.

A. Dataset Descriptions

This section reveals detail for three benchmark datasets. We start by reviewing the two public datasets, *i.e.*, **CRC-MSI** of colorectal cancer patients and **STAD-MSI** of gastric adenocarcinoma cancer patients, as well as our privately curated gastric cancer dataset **GIST-PDL1**. In Table 2, we brief the three datasets with numerical summary statistics.

A.1. CRC-MSI and STAD-MSI

The two public datasets focus on the prediction of distinguishing the microsatellite instability (MSI) from microsatellite stability (MSS) in H&E stained histology. Notably, MSI is a crucial clinical indicator for oncology workflow in the determination of whether a cancer patient responds well to immunotherapy. It is not until very recently that researches have shown the promising performance of deep learning methods in MSI prediction. With the lack of an abundant number of annotated histology, MSI prediction is still very challenging. Thus, it is required to incorporate prior knowledge such as geometric representation for MSI prediction.

In the experiment, the two datasets classify images patches to either *MSS* (microsatellite stable) or *MSIMUT* (microsatellite instable or highly mutated). We treat *MSIMUT* as the positive label and *MSS* as the negative label in computing the AUC. The original whole-slide images (WSIs) are derived from diagnostic slides with formalin-fixed paraffin-embedded (FFPE) processing. In particular, **CRC-MSI** contains H&E stained histology slides of 315 colorectal cancer patients, and **STAD-MSI** includes H&E slides of 360 gastric cancer patients. For both datasets, a WSI with respect to a patient is tessellated into non-overlapped patches/images with a resolution of 224×224 pixels at the magnification of $20\times$. The patches from 70% patients are used for training and the remaining patches from 30% patients are left for validation. Note that each patient is associated with only one WSI. Yet, the number of generated image patches from one WSI varies from each other. Consequently, the ratios of training and test image samples depicted in Table 2 for **CRC-MSI** and **STAD-MSI** are not $70\% : 30\%$ as its patient-level ratio.

A.2. GIST-PDL1

The privately collected **GIST-PDL1** predicts programmed death-ligand 1 (PD-L1) status from gastric cancer histology slides. PD-L1 is a type of immune-checkpoint protein from tumor cells that disturbs the body's immune system through binding programmed death 1 (PD-1) on T cells. The PD-L1 expression is one of the only established biomarkers that determine the efficacy of immunotherapy in gastric and esophageal cancer in advanced stages (Smyth et al., 2021).

This dataset collects 129 well-annotated H&E stained histology slides of gastric cancer patients between the year 2020 to the year 2021 from [anonymous] hospital. Each whole-slide image (WSI) corresponds to one patient, which is labeled as either *positive* (CPS≥ 5) or *negative* (CPS< 5) determined by its PD-L1 combined positive score (CPS) tested from the immunohistochemistry (IHC) test. The patch-level annotation inherits the associated WSI-level label. The resolution of a WSI is around $10,000 \times 10,000$ pixels, which is split into non-overlapping images (patches) of 512×512 pixels at the magnification $20\times$, and afterward resized to 224×224 to get aligned with the two public datasets. Background patches are excluded from downstream analysis. In the pre-processing, the remaining patches are subsequently stain normalized to reduce the data heterogeneity. Each patch comprises approximately 200 cells *i.e.*, nodes. Different from **CRC-MSI** and **STAD-MSI**, we conduct down-sampling on the number of image patches from each WSI to balance the ration between *positive* and *negative* samples. Consequently, we achieve a balanced image-level sample ratio that is close to $50\% : 50\%$.

A.3. Data Availability

The two public datasets for MSI classification with their annotations freely available at

`https://doi.org/10.5281/zenodo.2530834.`

The private dataset is available for limited usage by contacting the authors. The extracted graphs for **CRC-MSI** and **STAD-MSI** are available at `https://github.com/yiqings/HEGnnEnhanceCnn`.

B. Nuclei Segmentation

In the node construction process, we employ CA$^{2.5}$-Net (Huang et al., 2021) as the backbone for nuclei segmentation, due to its outstanding performance in challenging clustered edges segmentation tasks, which frequently occurs in histology analysis.

Table 2. Summary of the three datasets.

	Dataset	GIST-PDL1	CRC-MSI	STAD-MSI
IMAGE	# Patients	129	315	360
	# Training Images	7,676	93,408	100,570
	Training Positive Rate	41.10%	50.0%	50.0%
	# Test Images	2,471	99,904	118,008
	Test Positive Rate	47.71%	29.4%	23.6%
	Magnification	20×	20×	20×
	Original Patch Size	512 × 512	224 × 224	224 × 224
GRAPH	Min # Nodes	50	1	1
	Max # Nodes	621	103	120
	Median # Nodes	199	40	51
	Avg # Nodes	206	40	50
	Avg # Edges	3,402	163	246

We use the implementation at `https://github.com/JH-415/CA2.5-net`. Specifically, CA$^{2.5}$-Net formulates nuclei segmentation task in a multi-task learning paradigm that uses edge and cluster edge segmentation to provide extra supervision signals. To be more concretely, the decoder in CA$^{2.5}$-Net comprises three output branches that learn the nuclei semantic segmentation, normal-edge segmentation (*i.e.*, non-clustered edges), and clustered-edge segmentation respectively. A proportion of the convolutional layers and upsampling layers in the CA$^{2.5}$-Net is shared to learn common morphological features. We follow the original settings (Huang et al., 2021) by using the IoU loss for the segmentation path of the nuclei semantic ($\mathcal{L}_{sem}$) and the smooth truncated loss for segmentation paths of normal-edges ($\mathcal{L}_{nor}$) and clustered-edges ($\mathcal{L}_{clu}$). Formally, the overall loss thus takes a weighted average over the three terms of segmentation losses, *i.e.*,

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{sem} + \beta \cdot \mathcal{L}_{nor} + \gamma \cdot \mathcal{L}_{clu}. \tag{5}$$

In particular, we applies the balancing coefficients $\alpha = 0.7$, $\beta = 0.2$, and $\gamma = 0.1$. We trained CA$^{2.5}$-Net with ADAM optimizer for 50 epochs, with an initial learning rate of 1×10^{-4} that decayed by 0.95 for every other epochs. At the inference stage of nuclei locations, we adopt the nuclei segmentation path to derive the prediction result.

Three well-experienced pathologists annotated a number of 132 image patches from **GIST-PDL1** for training the CA$^{2.5}$-Net, where we use 100 images for training and the remaining 32 for validation. To increase the data variations, we adopt offline augmentation by randomly flipping and rotating for 90 degrees. Eventually, we come to a total number of 400 training samples. For illustration purposes, we pick one annotated sample and show it in Figure 3. The pixel-level instance annotations were conducted with '`labelme`' (`https://github.com/wkentaro/labelme`), where the semantic masks can be generated directly from the instance segmentation (Huang et al., 2021).

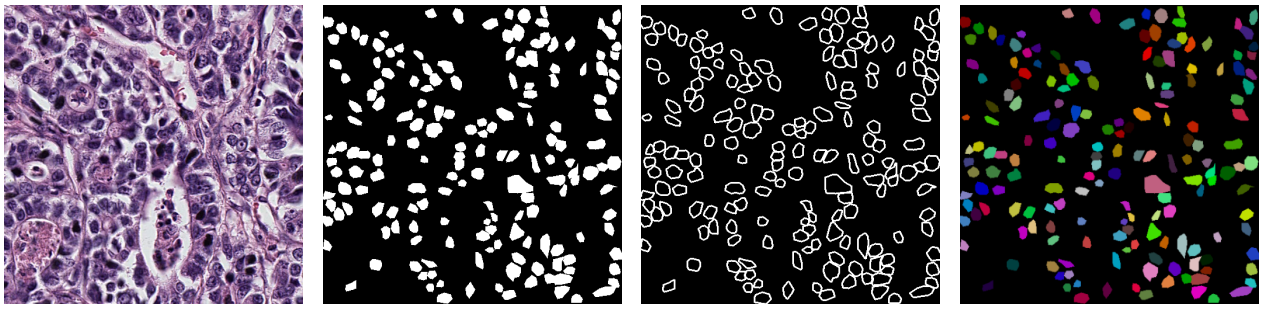

Figure 3. An illustrative example of annotated histology patch from **GIST-PDL1** for training the nuclei segmentation network. The four subgraphs from left to right are the raw patch image, the generated semantic nuclei masks, the generated semantic nuclei edge, and the annotated instance nuclei mask (ground truth).

Table 3. The pre-defined node attributes computed for each nuclei area.

GLCM	**GLDM**	**GLRLM**
autocorrelation	dependence entropy	gray-level non-uniformity
cluster prominence	dependence non-uniformity	gray-level non-uniformity normalized
cluster shade	dependence non-uniformity normalized	gray-level variance
cluster tendency	dependence variance	high gray-level run emphasis
contrast	gray-level non-uniformity	long-run emphasis
correlation	gray-level variance	long-run high gray-level emphasis
difference average	high gray-level emphasis	long-run low gray-level emphasis
difference entropy	large-dependence emphasis	low gray-level run emphasis
difference variance	large-dependence high gray-level emphasis	run entropy
inverse difference	large-dependence low gray-level emphasis	run length non-uniformity
inverse difference moment	low gray-level emphasis	run length non-uniformity normalized
inverse difference moment normalized	small-dependence emphasis	run percentage
inverse difference normalized	small-dependence high gray-level emphasis	run variance
informational measure of correlation 1	small-dependence low gray-level emphasis	short-run emphasis
informational measure of correlation 2		short-run high gray-level emphasis
Inverse variance		short-run low gray-level emphasis
joint average	**FIRST-ORDER**	
joint energy	10 percentile	
joint entropy	90 percentile	**GLSZM**
maximal correlation coefficient	energy	gray-level non-uniformity
maximum probability	entropy	gray-level non-uniformity normalized
sum average	inter quartile range	gray-level variance
sum entropy	kurtosis	high gray-level zone emphasis
sum squares	maximum	large area emphasis
	mean absolute deviation	large area high gray-level emphasis
LOCATION	mean	large area low gray-level emphasis
center of mass-x	median	low gray-level zone emphasis
center of mass-y	minimum	size zone non-uniformity
	range	size zone non-uniformity normalized
NGTDM	robust mean absolute deviation	small area emphasis
busyness	root mean squared	small area high gray-level emphasis
coarseness	skewness	small area low gray-level emphasis
complexity	total energy	zone Entropy
contrast	uniformity	zone Percentage
strength	variance	zone variance

C. Node Feature Extraction

This section details the essential pre-processing of the raw histology input (*i.e.*, images) to extract morphological features as node attributes in the construction of cell graphs. The same procedure applies to all three datasets. The segmentation results of CA$^{2.5}$-Net on slide patches generate nodes of graphs. For an arbitrary patch, a graph is generated where nodes represent cells and the weighted edges reveal the Euclidean distance between nodes.

Next, we select 94 features from pathomics, *i.e.*, a pre-defined feature library for medical image analysis that describe the location, first-order statistics, and the gray-level textural features of each segmented cell. To be specific, the five dimensions of the spatial distribution include gray-level co-occurrence (GLCM), gray-level distance-zone (GLDM), gray-level run-length (GLRLM), gray-level size-zone (GLSZM), and neighborhood gray tone difference (NGTDM). In total, there are 2 coordinates of the cell location, 18 values of the first-order statistics, 24 GLCM, 14 GLDM, 16 GLRLM, 16 GLSZM, and 5 NGTDM. We give the name of all 94 features in Table 3 for a better understanding. For a detailed calculation of each attribute, we refer interested readers to the work by Lambin et al. (2017) .

D. Alignment Layer

For the representation, which embeds the spatial topological structures of the underlying graph, is usually sent to a readout layer, such as a linear layer, before eventually being fed into the fusion layer. We term this linear layer as the *alignment layer*, which helps to align the feature dimensions of the GNN with the parallel CNN output.

E. Implementation Details

The code is available at:

`https://github.com/yiqings/HEGnnEnhanceCnn`

All the experiments are implemented in Python 3.8.12 environment on one NVIDIA ® Tesla A100 GPU device with 6,912 CUDA cores and 80GB HBM2 mounted on an HPC cluster. We implement GNNs on PyTorch-Geometric (version 2.0.3) and CNNs on PyTorch (version 1.10.2). All CNNs used **ImageNet** pre-trained weights.

E.1. Training Settings

All the model architectures follow the training scheme with the hyper-parameters listed in Table 4. We employ the standard cross-entropy as the loss function. The training stage continues until stopping improvements on the validation set after 8 consecutive epochs.

Table 4. Hyper-parameters for training the models.

Hyper-parameters	Value
Initial learning rate	5×10^{-4}
Minimum learning rate	5×10^{-6}
Scheduler	Cosine Annealing (T_max=10)
Optimizer	AdamW
Weight Decay	1×10^{-5}
num_workers	12
Batch size	256
Maximum epoch number	100

E.2. Model Configuration

Table 5 describes the configuration of MLP fusion layer, TRANSFORMER fusion layer, GCN and GIN used in this research. The model architectures for all three datasets apply the same configuration.

Table 5. Default configurations.

MLP	# `MLPBlock`	1
	Feature embedding size	128
	Activation	`Leaky ReLU`
	Dropout rate	0.1
TRANSFORMER	# `TransBlock`	1
	Feature embedding size	192
	Activation	`ReGLU`
	# Attention heads	4
	Dropout rate in `TransBlock`	0.1
GCN and GIN	# layers	2
	Feature embedding size	128
	Activation	`GeLU`

E.3. Number of Trainable Parameters Comparison

Table 6 reports the number of trainable parameters of all the models we evaluated in Table 1. The values are given with the image input size of $3 \times 224 \times 224$ and the node feature dimension of 94. The scale of the trained model is jointly determined by the choice of modules in CNN, GNN, and fusion layers, where we highlighted different options by color. The choice of colors aligns with the associated modules visualized in Figure 2. For instance, green colors include three selections of CNN modules, including MOBILENETV3, DENSENET, and RESNET. 'N/A' indicates an absence of such layers in the framework. The numbers are reported in millions (1×10^6). For instance, 13.1277 at the bottom-right of the table means it involves 13.1277 millions of learnable parameters when training an integrated model with the RESNET18-GIN2-TRANS architecture.

Table 6. Comparison on the number of trainable parameters (in millions).

		CNN			
GNN	Fusion	N/A	MOBILENETV3	DENSENET	RESNET
N/A	N/A	-	1.7865	7.2225	11.3110
GCN	MLP	0.0665	1.8506	7.2866	11.3751
	TRANSFORMER	-	8.4943	13.9303	13.1231
GIN	MLP	0.0712	1.8552	7.2912	11.3797
	TRANSFORMER	-	8.4989	13.9349	13.1277

F. Further Performance Analysis

Table 7 below reveals the absolute percentage improvement in the main evaluation tasks. The comparisons are made on basis of CNN baselines. To be specific, the absolute improvement score is calculated by

$$\Delta \text{ Score} = S_{\text{fused}} - S_{\text{CNN}},$$

where S_{CNN} denotes the performance score (*i.e.*, ACC, AUC or AUC*patient*) achieved by plain CNNs (*i.e.*, MOBILENETV3, DENSENET or RESNET), and S_{fused} is the associated score by the integrated models, which are listed in the first column at the very left. For instance, the 3.56 at the top-left of the table means the test accuracy of MOBILENETV3-GCN2-MLP is improved by 3.56% to MOBILENETV3.

Table 7. Test Acc and AUC improvements of the fusion model to pure CNN on three benchmarks.

	GIST-PDL1			CRC-MSI			STAD-MSI		
Model	ΔACC	ΔAUC	ΔAUC*patient*	ΔACC	ΔAUC	ΔAUC*patient*	ΔACC	ΔAUC	ΔAUC*patient*
MOBILENETV3-GCN2-MLP	3.56	1.96	8.06	−0.25	7.17	12.87	0.78	4.98	0.35
MOBILENETV3-GIN2-MLP	1.33	2.59	10.32	0.43	2.79	13.47	0.53	2.90	0.46
MOBILENETV3-GCN2-TRANS	4.27	3.69	12.82	0.19	4.75	13.16	1.01	6.73	1.53
MOBILENETV3-GIN2-TRANS	2.56	4.16	10.51	0.51	4.24	14.89	0.95	6.47	1.65
DENSENET121-GCN2-MLP	5.35	6.35	6.89	1.04	0.07	1.60	1.66	8.96	1.04
DENSENET121-GIN2-MLP	5.58	5.51	5.61	0.46	0.16	4.95	2.07	9.25	0.58
DENSENET121-GCN2-TRANS	8.45	7.29	8.47	0.81	4.01	16.60	1.77	8.04	1.44
DENSENET121-GIN2-TRANS	4.64	5.20	6.99	0.65	4.64	9.08	1.90	8.82	0.98
RESNET18-GCN2-MLP	11.25	10.50	7.81	0.62	9.84	21.50	2.26	0.50	0.66
RESNET18-GIN2-MLP	5.61	5.51	5.04	0.99	4.54	21.33	2.43	2.13	1.49
RESNET18-GCN2-TRANS	5.39	4.31	7.39	1.26	8.33	23.04	2.35	0.10	1.43
RESNET18-GIN2-TRANS	6.16	9.99	9.27	1.08	8.12	22.13	2.42	1.99	1.71

We also investigate the overall ranking of each model fusion configuration in Table 8. For each CNN architecture in one dataset, ranks are calculated by sorting the reported score, where we report the averaged ranking over three datasets. Generally speaking, the TRANSFORMER integrator outperforms the simple MLP scheme. For the choice of GNNs, GCN

and GIN do not present a significant advantage one over another. Nevertheless, all the integrated models outperform designs with plain CNNs.

Table 8. Performance ranking reports the average rank across three datasets. For simplicity, MV3,DENSE and RES denote MOBILENETV3, DENSENET-121 and RESNET-18 respectively. The '**Avg**' columns report the averaged ranking over three CNN architectures. The '**Overall**' reports the averaged ranking over three metrics.

	ACC				AUC				AUC*patient*				
	MV3	DENSE	RES	**Avg**	MV3	DENSE	RES	**Avg**	MV3	DENSE	RES	**Avg**	**Overall**
CNN	4.7	5.0	5.0	4.9	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
+ GCN-MLP	3.3	2.7	3.0	3.0	2.3	2.7	1.7	2.2	4.0	3.0	3.0	3.3	2.9
+ GCN-TRANS	3.3	2.3	2.3	2.7	4.0	2.3	2.7	3.0	2.7	3.7	3.3	3.2	3.0
+ GIN-MLP	1.7	2.0	2.7	2.1	1.7	2.3	3.3	2.4	2.0	1.0	2.3	1.8	**2.1**
+ GIN-TRANS	2.0	3.0	2.0	2.3	2.0	2.7	2.3	2.3	1.3	2.3	1.3	1.7	**2.1**

G. Visualization of Nuclei Segmentation and Cell Graph

To better understand the learned graphs that are generated from histology images, Figure 4-6 investigate some random patch images from the three datasets and visualize the nuclei segmentation results and the associated graphs. In particular, the four subgraphs from left to right of each figure display the raw patch image, the segmented cells masks, the patch image with overlaid segmentation masks, and the generated graph.

H. Limitations and Future Works

In this research, we use GIN and GCN as the GNN backbones. Other variants such as GAT, DCNN, and GraphSAGE have not been experimented on, which is left to future work. Additionally, node features are morphological-based, which empirically does not perform well. Deep features by self-supervised learning are expected to be another candidate.

I. Additional Reference in Appendix

Smyth, E., Gambardella, V., Cervantes, A., and Fleitas, T. Checkpoint inhibitors for gastroesophageal cancers: dissecting heterogeneity to better understand their role in first-line and adjuvant therapy. *Annals of Oncology*, 32 (5):590–599, 2021.

Lambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., De Jong, E. E., Van Timmeren, J., Sanduleanu, S., Larue, R. T., Even, A. J., Jochems, A., et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, 14(12): 749–762, 2017.

Veličković P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv preprint arXiv:1710.10903. 2017 Oct 30.

Atwood, J. and Towsley, D., 2016. Diffusion-convolutional neural networks. *Advances in neural information processing systems*, 29.

Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

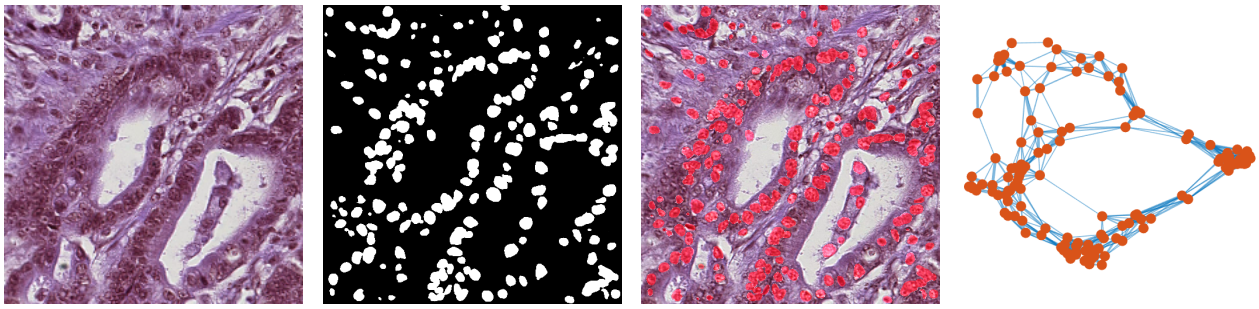

Figure 4. Visualization of the segmented cells and the generated graphs from an arbitrary patch sample of **GIST-PDL1**. The four subgraphs from left to right are the raw patch image, the segmented cells masks, the patch image with overlaid segmentation masks, and the generated graph.

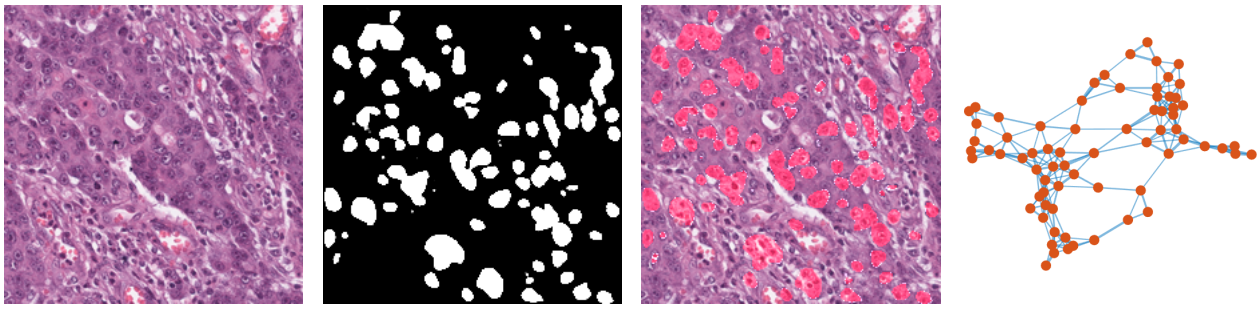

Figure 5. Visualization of the segmented cells and the generated graphs from an arbitrary patch sample of **CRC-MSI**. The four subgraphs from left to right are the raw patch image, the segmented cells masks, the patch image with overlaid segmentation masks, and the generated graph.

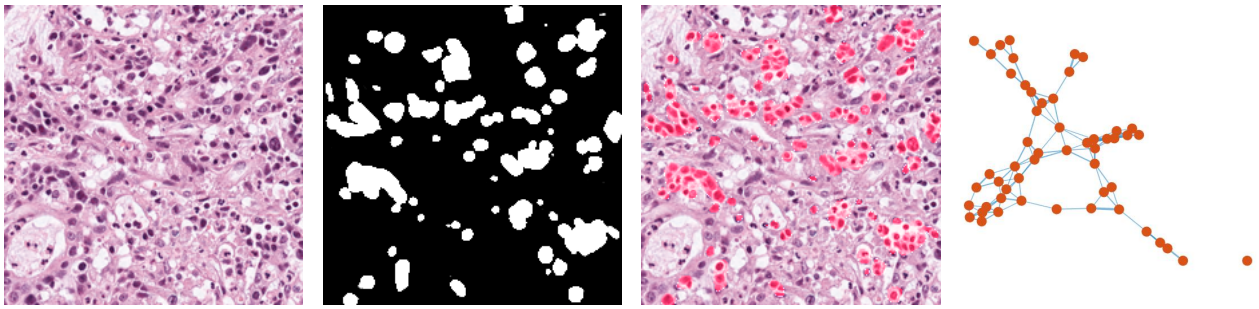

Figure 6. Visualization of the segmented cells and the generated graphs from an arbitrary patch sample of **STAD-MSI**. The four subgraphs from left to right are the raw patch image, the segmented cells masks, the patch image with overlaid segmentation masks, and the generated graph.