Spatial Omics Driven Crossmodal Pretraining Applied to Graph-based Deep Learning for Cancer Pathology Analysis

Zarif L. Azher
Thomas Jefferson High School for Science and Technology
Alexandria, VA 22312, USA
Email: 2024zazher@tjhsst.edu

Michael Fatemi
University of Virginia, Department of Computer Science
Charlottesville, VA 22904, USA
Email: myfatemi04@gmail.com

Yunrui Lu, Gokul Srinivasan, Alos B. Diallo
EDIT, Department of Pathology and Laboratory Medicine, Dartmouth Hitchcock Medical Center
Lebanon, NH 03756, USA
Email: yunrui.lu@dartmouth.edu, gokulsrin@gmail.com, alos.b.diallo.gr@dartmouth.edu

Brock C. Christensen, Lucas A. Salas
Department of Epidemiology, Geisel School of Medicine at Dartmouth
Lebanon, NH 03756, USA
Email: brock.c.christensen@dartmouth.edu , lucas.a.salas@dartmouth.edu

Fred W. Kolling IV, Laurent Perreard
Genomics Shared Resource, Dartmouth Cancer Center
Lebanon, NH 03756, USA
Email: fred.w.kolling.iv@dartmouth.edu, laurent.perreard@dartmouth.edu

Scott M. Palisoul, Louis J. Vaickus, Joshua J. Levy[*†]
EDIT, Department of Pathology and Laboratory Medicine, Dartmouth Hitchcock Medical Center
Lebanon, NH 03756, USA
Email: scott.m.palisoul@hitchcock.org, louis.j.vaickus@hitchcock.org, joshua.j.levy@dartmouth.edu

Graph-based deep learning has shown great promise in cancer histopathology image analysis by contextualizing complex morphology and structure across whole slide images to make high quality downstream outcome predictions (ex: prognostication). These methods rely on informative representations (i.e., embeddings) of image patches comprising larger slides, which are used as node attributes in slide graphs. Spatial omics data, including spatial transcriptomics, is a novel paradigm offering a wealth of detailed information. Pairing this data with corresponding histological imaging

[*] To whom correspondence should be addressed.
[†] Work supported by grants P20GM130454, P20GM104416 to JL and K08CA267096 to LV.

localized at 50-micron resolution, may facilitate the development of algorithms which better appreciate the morphological and molecular underpinnings of carcinogenesis. Here, we explore the utility of leveraging spatial transcriptomics data with a contrastive crossmodal pretraining mechanism to generate deep learning models that can extract molecular and histological information for graph-based learning tasks. Performance on cancer staging, lymph node metastasis prediction, survival prediction, and tissue clustering analyses indicate that the proposed methods bring improvement to graph based deep learning models for histopathological slides compared to leveraging histological information from existing schemes, demonstrating the promise of mining spatial omics data to enhance deep learning for pathology workflows.

Keywords: spatial omics, transcriptomics, deep learning, graphs, cancer, colon cancer.

1. Introduction

1.1. *Deep Learning for Pathology*

In recent years, countless studies have demonstrated the potential for deep learning algorithms to solve challenging biomedical tasks, thereby improving risk stratification and alleviating the potential for clinical burnout by making tedious and unreliable tasks faster and more quantitative, potentially leading to improved patient health outcomes [1]. These algorithms are formulated on computational heuristics – specifically, machine learning -- which can make sense of many complex data types through the dynamic derivation of relevant patterns and features [2–4]. Analysis of pathology data, including whole slide imaging (WSI) – microscopic images of patient tissue – is an emerging application in this space, as WSIs are routinely collected and used for patient monitoring, diagnosis, and prognostication. Existing works have shown that specially designed deep learning algorithms, inspired by processes of the central nervous system, may be able to automate or assist in these tasks [5]. Most deep neural networks study small micromorphological changes given the enormity of these gigapixel images. Graph convolutional networks (GCNs), however, are a promising method in this domain, as they can effectively model macro and micro architectural features present across WSI in a human-interpretable manner [6]. Generally, these methods split WSI into patches (i.e., more manageable subimages), extract numeric representations (i.e., "embeddings") from each patch using a predetermined algorithm, and construct a graph where the nodes are given patch embeddings and edges are formed based on spatial adjacency [7–9]. Such methods have been applied for tumor stage prediction [9], survival analysis [8], and derive numerical representations of WSI that can be combined with other omics and imaging modalities [7].

The optimal algorithm used to extract node features is an area of ongoing research, though many works presently use a ResNet convolutional neural network (CNN) pretrained on the ImageNet database [10] for this task [8,11,12]. It has become increasingly common to additionally train these CNNs on various image tasks orthogonal to the task at hand to prepopulate an information registry of features which will ultimately improve predictive performance in other settings; these techniques are known as pretraining. Recently, self-supervised techniques have emerged as promising pretraining methodologies, where images are compared from several different vantage points without being explicitly labeled. Cross-modal pretraining has recently been highlighted as a common self-supervised method by leveraging complementary "paired" information across multiple input data types (e.g., images and text) which can improve the representation of all involved modalities. Here, we investigate the utility of using spatial omics data, which is paired at 50-micron

resolution to the histological information, to pretrain an encoder model for these patches, to demonstrate the power of leveraging spatial omics for deep learning-based pathology methods which are particularly suited for analysis using graph neural networks (GNNs).

1.2. *Spatial Omics*

Omics data – such as gene expression quantification and DNA methylation – have traditionally been collected on a bulk scale where measurements are taken across an entire sample or tissue section. Recent advancements in technology have allowed for collection on a more granular scale, such as the single cell level, or across specific spots/regions in a slide sample [13]. Prior studies have demonstrated that deep learning through specialized architectures like GCNs can mine spatial omics data to build a more comprehensive understanding of spatial cellular heterogeneity, especially as it pertains to how the tumor microenvironment can facilitate/inhibit further disease progression [14,15]. Notably, this type of data is not yet commonly available at large scale due to the prohibitive cost of these assays as well as batch effects and selection of limited slide area, meaning that methods which can learn from spatial omics data and effectively transfer this knowledge to improve other tasks may be valuable. Zeng et al [15] previously developed a model which utilized contrastive learning to mine a shared representation between image patches and corresponding spatial transcriptomics; however, their investigation centered on driving improved understanding on gene domains, rather than attempting to leverage the method to enhance downstream clinical outcome modeling in situations where only WSI – and no ST data – is available.

1.3. *Contributions*

We hypothesize that additional biological information can be learnt from spatially resolved transcriptomics data that may prove relevant for enhancing prediction models across a range of histological analyses. Existing works applying GCNs for WSI analysis have not yet leveraged spatial omics data to enhance modeling across orthogonal tasks. In part, this is because the quality of histological slides for spatially co-registered omics data has been limited as the standard Visium spatial transcriptomics (ST) workflow featured manual staining and low-resolution imaging– this information does not readily transfer to prediction models on higher resolution histological slides. Now, with the development of assays such as the CytAssist which permit the use of sophisticated laboratory processing (i.e., autostaining and 40X imaging prior to Visium profiling), the quality of slides has remarkably increased and allows for training image models that may more readily transfer to related domains. Here, we assess the ability of spatially resolved omics data to enhance predictions on a range of different histological assessment tasks by presenting an initial evaluation of a crossmodal pretraining mechanism using matched WSI and spatial omics measurements as means to encode biological information within WSI graphs to apply in scenarios where spatial omics data is not available. We compare this method against other common pretraining schemes on downstream predictive analyses (staging, lymph node metastasis, survival prognostication) of WSI, as well as explore generated image patch embeddings. Accurate methods for these downstream predictive tasks may enable more personalized patient treatments. In this study, we expect developed models which can mine for spatial molecular information to outperform the compared approaches on these tasks. We aim to demonstrate the potential benefits of utilizing spatial omics – spatial transcriptomics, in particular – methods to enhance deep learning-driven pathology analysis.

2. Methods

2.1. *Data Collection and Preprocessing*

Visium spatial transcriptomics data matched with WSI was collected from four colorectal cancer patients from the Dartmouth Hitchcock Medical Center, to serve as a training dataset for the crossmodal patch embedding method. This process was conducted through the 10x Genomics Visium spatial transcriptomics workflow, featuring H&E staining, followed by mRNA profiling and whole slide imaging. Spatial transcriptomics data were filtered to include the top 1000 most variable genes across slides identified by SpatialDE [16]. Separately, 708 WSIs were collected from colorectal cancer patients from the Dartmouth Hitchcock Medical Center, for whom, histological stage annotations were available. Finally, WSIs were obtained for a cohort of 350 colorectal cancer patients from The Cancer Genome Atlas (TCGA) for whom survival information and lymph node metastasis information was available. All WSIs were stain normalized using the Macenko [17] method. Collected WSIs were split into non overlapping 224 x 224 patches via the PathflowAI Python package [18], whose embeddings served as node attributes in a graph. We compared several methods described below to encode information for these patches, which is the main focus of this study. Nodes were connected with edges based on spatial adjacency using the *knn_graph* (k-nearest neighbor) method from the torch_cluster Python package, with k=16. Patients from the in-house dataset and TCGA were separately partitioned into training, validation, and testing sets using a random 80/10/10 split. The collected datasets and the downstream tasks they were used on, are summarized below:

1. **Visium spatial transcriptomics slides (n=4; 20,000 spots/patches; Co-Registered Spatial Transcriptomics, H&E WSI):** to pretrain contrastive crossmodal model
2. **Dartmouth Hitchcock Medical Center (n=708 H&E WSI):** used for histological stage prediction and clustering analysis
3. **TCGA Cohort (n=350 H&E WSI):** used for lymph node metastasis prediction, survival prognostication, and tumor infiltrating lymphocyte (TIL) alignment analysis

All analyses were conducted on a machine using a single Nvidia Tesla v100 GPU with 32 gigabytes of VRAM, and 100 gigabytes of RAM.

2.2. *Patch Level Pretraining Methods*

Three embedding production methods were compared for the 224x224 patches used as nodes of the graphs representing WSI.

2.2.1. *ImageNet-Pretrained ResNet18*

A ResNet18 CNN model pre trained on the ImageNet dataset (commonly used for embedding histopathology patches) was accessed using the *torchvision* Python package (https://github.com/pytorch/vision). The model was truncated through the penultimate layer, to extract length 512 vectors/embeddings for each input patch.

2.2.2. *Ciga Self Supervised Histopathology Pretrained ResNet18*

A separate ResNet18 CNN model pretrained using a self-supervised learning (SSL) SimCLR [19] contrastive procedure on histpathological imaging datasets was similarly accessed and truncated

through the penultimate layer to extract length 512 embeddings for all patches. In summary, SimCLR employs an objective function that encourages similarity between embeddings from augmented (i.e., "corrupted") views of the same image, while penalizing based on dissimilarity between views from different images. This model was made publicly available by Ciga et al [20], and has been previously shown to outperform the aforementioned ImageNet-pretrained model on a variety of downstream modeling tasks.

2.2.3. *Spatial Omics-driven Crossmodal Pretrained Encoder*

A contrastive cross-modal model encoding image patches and spatial transcriptomic profiles was created, similar to the model implemented by Zeng et al [21]. Input images patches of size 224x224 were encoded into embeddings of size 512 units, using the feature extraction portion of a CNN initialized with weights initialized from the ResNet model trained by Ciga et al. Spatial transcriptomics profiles containing expression of the most spatially variable 1000 genes across Visium slides, selected to avoid overfitting on genes with imprecise expression, were encoded with three standard fully connected (FC) layers of size 512. The embeddings from co-registered patches from each modality (ST, WSI) were passed through a common projection layer of size 512, to output a single embedding per modality (ie; one vector of length 512 which describes an image patch, and one of length 512 which describes the corresponding gene expression). Crossmodal and unimodal contrastive penalties are applied using the SimCLR loss function [19]; during training, several augmentation strategies were applied to both the image patches and corresponding transcriptomic profiles to generate "corrupted" representations of each data type as means for comparison. Transcriptomic profiles were randomly masked and corrupted with noise with 30% probability. Images were augmented using a series of random flips, color jitter transforms, random grayscaling, random rotation, and random image solarization. Both the original and augmented image patches and transcriptomics profiles were encoded using the aforementioned neural network layers. The loss mechanism penalizes the model based on the difference between the embeddings from the original and augmented data from each modality. A crossmodal loss is used to maximize the similarity between the corrupted image and transcriptomic embeddings from the same patch. These three loss functions (augmented image to image, augmented transcriptomics to transcriptomics, augmented image to augmented transcriptomics) were summed to optimize the crossmodal contrastive model.

This model was trained for 150 epochs with a batch size of 8 and a learning rate of 0.00001. Visium sections corresponding to six patients were partitioned into the training set, and tissue sections from two patients were partitioned to the validation set. Validation set loss was used to inform selection of the top model, following training. The RELU activation function was applied to outputs of every layer. The image encoder pretrained using the spatially co-registered transcriptomics information and the subsequent projection head were retained for subsequent analysis, and were used to embed image patches which GNN models were to operate on. The remaining layers of this pretrained model were not utilized. The usage of this image encoder derived using this training protocol for other ancillary tasks is the primary focus of this study, compared to the other image encoders (weights from ImageNet, Ciga et al.). This model is further described along with data collection procedure, in **Figure 1**.

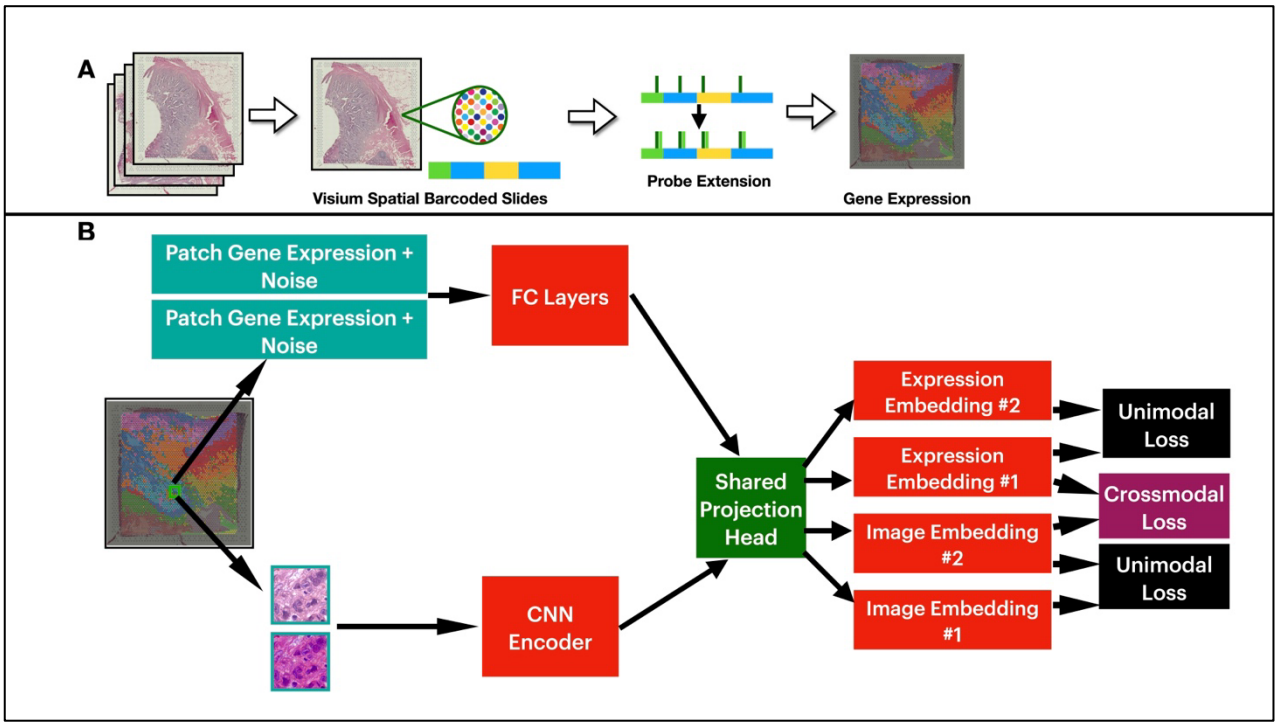

Figure 1: A) Data collection protocol for Visium spatial transcriptomics slide. **B)** Training protocol for spatial omics-driven crossmodal contrastive model; two views are generated per modality, per patch; each view is passed through the corresponding branch of the crossmodal model; embeddings are transformed using a shared projection head; unimodal and crossmodal contrastive losses are applied to output embeddings.

2.3. *Downstream Outcome Prediction*

We sought to understand whether CNN encoders, pretrained on co-registered spatial transcriptomics data, could enhance the predictions on a range of different GCN tasks. A graph convolutional network was constructed to take an input graph of nodes represented by length 512 embeddings, followed by three GCNConv graph convolutional layers [22] to contextualize and aggregate embeddings into length 128, with SAGEPooling pooling [23] layers (ie: 30% of patches retained, for subsequent layers; SAGEPooling stochastically samples higher-order neighborhoods of patches) placed after each convolutional layer. These pooling layers learn to downsample graphs, to push the model to learn focused information relevant to the training task. Graph embeddings were aggregated using global mean pooling after each SAGEPooling layer. These embeddings were combined using the JumpingKnowledge mechanism, resulting in a single vector of length 128 to represent the entire input graph/WSI. Finally, two fully connected layers were applied to this embedding, followed by a single output layer. The model (**Figure 2**) was applied to the following prognostication-focused experiments/outputs to assess patch encoding mechanisms:

2.3.1. *Histological Stage Prediction*

The in-house dataset was used to train and assess model capability to predict dichotomized tumor histological stage (T-stage; signifies depth of invasion) - low (stage 0, stage 1, stage 2) or high (stage 3, stage 4). A sigmoid function was applied to the output of the final layer in the GCN, and model training was supervised using a binary crossentropy loss function.

2.3.2. *Survival Prognostication*

The TCGA dataset was used to train and evaluate GCNs to assess for time to death using hazard predictions, indicating the real-time risk of death. Model training was supervised using a standard Cox loss, which considers the predicted risk, patient censor status, and duration (either days to death or days to last follow up). This setup entails the proportional hazards assumption, that predictors have a constant hazard ratio (i.e., relative risk between two patient groups) over time.

All GCN models were trained for up to 30 epochs, using a learning rate of 0.001 and batch size 8. Top model checkpoints were selected for evaluation following training, based on validation set loss. GCN models were implemented using the Pytorch Geometric [24] Python package. Three separate GCN models were trained for each prediction task - one for each patch embedding mechanism. Stage prediction and lymph node metastasis models were evaluated on held-out test sets using F1-score and area under the curve (AUC), while C-index was used to evaluate prognostication models. These metrics are reported using 95% confidence interval derived from 1000 sample non-parametric bootstrapping procedures.

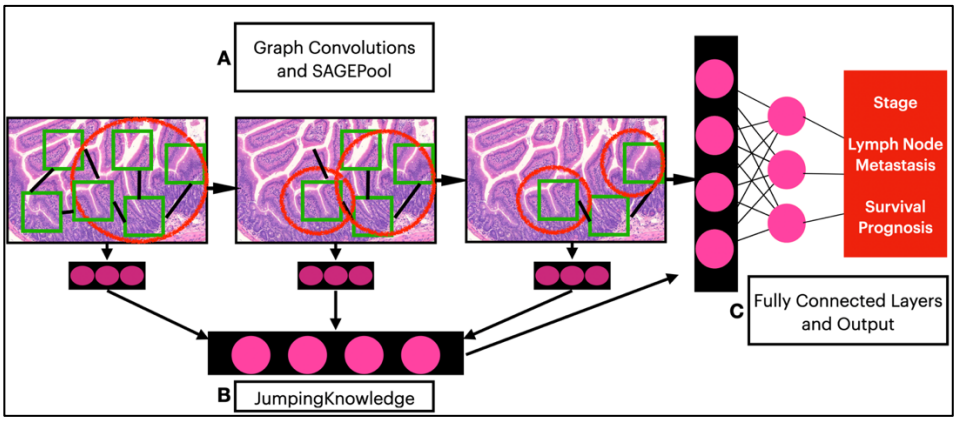

Figure 2: Overview of generalized GCN for downstream outcome modeling; initial patch embeddings vary across experimentation. A) Graph convolution layers contextualize each node embedding; after each such layer, SAGEPool operators aggregate nodes/patch embeddings, removing up to 70% of them, to only retain informative ones. B) A JumpingKnowledge scheme aggregates embeddings across graphs to create a single embedding for the image. C) The image embedding is used to make downstream predictions.

2.4. *Embedding Clustering Quality Analysis*

The ability of patch embeddings to capture morphological and molecular heterogeneity across slides was assessed across embedding methods, using an unsupervised clustering approach and the in-house dataset. For each WSI in the dataset, KMeans clustering (k=5; chosen via coarse optimization to ensure stability when run numerous times) was applied to the patch embeddings derived by each pretraining method (standard ResNet, Ciga et al, spatial pretrained) to elucidate sub-groups of patches implicitly captured by the representations. Clusters were plotted across slides to visually ensure that they represented different morphologies and structures within slides. Subsequently, the Calinski-Harabanz (CH) index [25] and the Davies-Bouldin (DB) index [26] were computed for the clustering result for each pretraining strategy. The ANOVA-based CH score assesses the density and separation of clusters, with a higher value indicating greater density within clusters and separation among different clusters. Similarly, the DB index measures the ratio between within-

cluster and cross-cluster separation. Thus, superior patch embeddings should result in a relatively high CH index and low DB index. The per-WSI scores were used to calculate average CH index and DB score at a 95% confidence interval, for each pretraining method.

2.5. *TIL-based Model Interpretation*

Previous research has demonstrated the importance of tumor infiltrating lymphocytes (TILs) and the tumor microenvironment on the progression of colon cancer [27]. We sought to demonstrate the interpretability of GCN models developed here using the TCGA dataset, by comparing regions of WSI given high attention with previously published predicted TIL maps [28] for corresponding slides. Patches deemed important by GCN models trained on lymph node metastasis prediction were determined by extracting patches remaining in WSI graphs following the final pooling layer; for a given patch, being left in its graph by a GCN model following three pooling layers, indicates its significance to the model. The coordinates of these patches were compared to those describing the locations of predicted TILs via Wald Wolfowitz testing [29], where the null hypothesis would indicate high overlap between these two sets of coordinates. Accordingly, Wald Wolfowitz testing was used to calculate a test statistic per slide per GCN model trained with each patch embedding method– negative values of this test statistics, W, represents the localization of TILs. Spearman's rank correlation coefficients (alpha p-value = 0.05) were calculated to evaluate the relationship between the test statistic (W), and predicted hazard. A negative correlation coefficient would suggest a statistically significant association between predicted hazard and TIL spatial localization, following biological knowledge holding that TILs help inhibit colon cancer proliferation and migration [30]. Test statistics were further dichotomized to indicate presence/lack of TIL localization, to compare these relationships across the GCN model using embeddings derived from the Ciga et al method, versus the model using spatially pretrained embeddings.

3. Results[†]

3.1. *Quantitative Predictive Analysis*

Held out testing-set performance for GCNs trained to predict stage, lymph node metastasis, and survival prognosis, are presented in **Table 1**; models which used patch embeddings derived from the spatial omics-driven mechanism outperformed those using the compared methods for all three experiments.

Table 1: Test set performance metrics (95% confidence interval) of GCNs trained using various patch embedding mechanisms, for binary stage prediction, lymph node metastasis prediction, and survival prognostication.

Task	Measure	ImageNet ResNet	Ciga et al ResNet	Spatial Pretrained
Stage Prediction	**AUC**	0.935 ± 0.003	0.948 ± 0.002	**0.981 ± 0.001**
	F1-Score	0.863 ± 0.004	0.858 ± 0.004	**0.878 ± 0.004**
Lymph Node Metastasis	**AUC**	0.651 ± 0.004	0.612 ± 0.004	**0.708 ± 0.003**
	F1-Score	0.560 ± 0.002	0.630 ± 0.003	**0.671 ± 0.005**
Survival Prognostication	**C-index**	0.597 ± 0.003	0.582 ± 0.002	**0.638 ± 0.002**

[†] Supplementary materials can be found at the following DOI: https://doi.org/10.5281/zenodo.8197573.

For the classification experiments, models using embeddings derived from the spatial omics-driven mechanism outperformed those which used embeddings from the ImageNet-trained ResNet18 CNN by an average of 6.98% measured by AUC, and outperformed models using embeddings derived from the ResNet18 pretrained by Ciga et al, by average of 9.47%. GCNs using spatial omics-driven embeddings (C-index 0.638) also outperformed ImageNet-trained ResNet18 embeddings (C-index 0.597) and embeddings derived from the model trained by Ciga et al (C-index 0.582).

3.2. *Clustering Evaluation*

A KMeans clustering approach paired with CH index and DB index calculation was employed to compare the abilities of these different pretraining approaches to elucidate molecular and morphological heterogeneity across slides; the results of this analysis are presented in **Table 2**. An example visualization including regions of a slide assigned to clusters indicating by different coloring, is presented in **Figure 3**; additional examples are available in Supplementary Figures S2 and S3.

Embeddings from the contrastive crossmodal spatial model resulted in a significantly higher CH index and lower DB index, versus both the ImageNet-pretrained ResNet and the ResNet trained on histopathology datasets via self-supervised learning by Ciga et al.

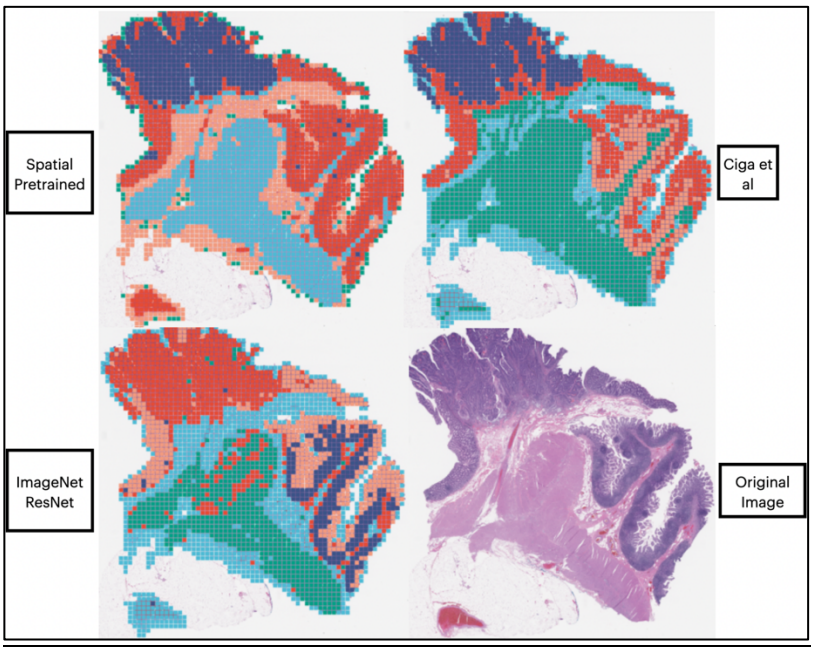

Figure 3: Example visualization of clustering of embeddings derived using various methods, for a single WSI.

Table 2: Clustering quality metrics calculated across embedding methods.

Measure	ImageNet ResNet	Ciga et al ResNet	Spatial Pretrained
Calinski-Harabanz Index	643.76 ± 17.51	786.70 ± 20.40	**2605.68 ± 70.66**
Davies-Bouldin Index	1.90 ± 0.01	1.719 ± 0.01	**0.975 ± 0.01**

3.3. *Model Interpretation*

Spearman's correlation coefficient values testing the relationship between lymph node metastasis risk predicted by GCN models using various patch embedding mechanisms and TIL localization elucidated via Wald Wolfowitz testing, are presented in **Table 3** along with corresponding p-values, suggesting both the Ciga and spatial pretrained models were able to derive TIL-associated embeddings related to instantaneous hazards. Boxplot visualizations comparison of predicted model risk and dichotomized TIL alignment are presented in **Supplementary Figure S4**.

Table 3: Spearman's correlation coefficient values for TIL localization versus predicted lymph node metastasis risk, across GCN models using various patch embedding methods.

	ImageNet ResNet	**Ciga et al ResNet**	**Spatial Pretrained**
Spearman's Coefficient	-0.061	-0.426	-0.218
Spearman's P-value	0.2693	2.2e-16	7.74e-5

4. Discussion and Conclusion

This is the first study which aims to determine whether leveraging spatial omics data to pretrain image patch encoders using a cross modal contrastive mechanism can improve downstream performance in graph convolutional networks, which may improve automated cancer patient analysis. While most prior research leveraged a GCN to integrate spatially localized omics with imaging for spot-level spatial transcriptomics enhancement or histological feature extraction tied to bulk transcriptional characteristics, our approach discerns spatial transcriptomics features from standalone slides. Recognizing the inaccessibility of spatial transcriptomics data, we employed transfer learning to apply extracted spatial transcriptomics features to a diverse range of subsequent tasks. We compared spatial omics-driven embeddings against those extracted from a standard ResNet18 CNN pretrained on the ImageNet dataset, and a ResNet18 pretrained using self-supervised learning on histopathology datasets. GCN models trained and evaluated using the spatially enhanced embeddings outperformed those using the baseline embedding methods on three downstream tasks – stage prediction, lymph node metastasis prediction, and prognostication. This suggests that incorporating spatial transcriptomics information into the pretraining process of image patch encoders, enhances the quality of learned representations, beyond what is extracted from state-of-the-art techniques which use solely images for patch encoding pretraining.

Additional quantitative analysis from clustering patch embeddings indicates that the models leveraging spatially-pretrained embeddings were superior at capturing distinct heterogeneities across slides, versus models using patch embeddings from existing strategies. Thus, we expect future applications of the developed spatial pretraining method for patch embeddings, to improve the performance of workflows aiming to capture tissue heterogeneity, including tumor subcompartment segmentation.

Furthermore, Wald Wolfowitz testing paired with Spearman's correlation coefficients, suggests that GCN models using embeddings from the spatial pretraining method and the Ciga et al method, learned to highlight TILs to contextualize prognostic assessment of cancerous tissue when considering lymph node metastatic potential, particularly in patients whom the models understood to be at lower risk. The Spearman's coefficient value for the GCN model using ImageNet ResNet patch representations was markedly closer to 0 versus the other two methods, indicating far weaker correlation in this relationship. Interestingly, the magnitude of the coefficient for the GCN model using the Ciga et al embeddings was nearly double that of the spatially pretrained embeddings,

indicating that the Ciga et al method may induce greater tendency to turn to TILs for understanding patient profiles. Though this does not indicate greater predictive power among models, that such nuances can be extrapolated related to model reasoning, demonstrates the interpretability of graph-based modeling for cancer histopathology, and further emphasizes the importance of enhancing the ability of such methods.

Overall, our results indicate that spatial omics data can be effectively mined in a crossmodal fashion, to improve existing image-based deep learning workflows to analyze cancer histopathology; this also adds to the growing body of literature [31–33] which reflects the importance of enhancing pretraining mechanisms as a basis of improving deep learning models for cancer histopathology. Notably, ours is the first study to mine spatial omics data in the pretraining process to enhance the capability of such image-based models, while others have focused on mechanisms which use solely imaging. Several AI methods also exist to integrate spatial transcriptomics with histology through contrastive learning to improve the identification of spatial domains. This work differs from prior approaches as it aims to improve the extraction of imaging information on held-out tissue slides from which Visium spatial transcriptomics assaying has not been done, training with paired imaging and spatial expression data to enhance this capability.

A key limitation of this study is the relatively small dataset used to pretrain the spatially-enhanced crossmodal contrastive model; spatial transcriptomics data was only generated for 4 total slides due to high resource and time costs and the limited size of the tissue placement area on Visium slides. Furthermore, coarse hyperparameter search was used to select GCN architecture parameters, as a detailed experiment here was beyond the scope of this study. It should be noted that optimization of the convolutional neural network and GCN parameters can be done end-to-end, i.e., simultaneously, which can improve predictive results– as will incorporating additional varied histologies and tumor characteristics, improved specimen processing/imaging using the CytAssist and commensurate hardware to fit larger models. Future works will seek to use larger cohorts to pretrain the spatial model to improve quality of extracted embeddings. Additionally, the embeddings from the spatially enhanced model can be evaluated for use in applications other than GCNs, such as Transformer networks – which have become popular in cancer histopathology in recent years [34,35] – histology image search, and multimodal data integration.

5. Acknowledgements and Location of Supplementary Material

The results published here are in part based on data generated by the TCGA Research Network: https://cancer.gov/tcga. The authors acknowledge the support of the Center for Clinical Genomics and Advanced Technology in the Department of Pathology and Laboratory Medicine of the Dartmouth Hitchcock Health System which includes the Pathology Shared Resource, at the Dartmouth Cancer Center with NCI Cancer Center Support Grant 5P30 CA023108-37. Spatial transcriptomics assays were carried out in the Genomics and Molecular Biology Shared Resource (GMBSR) at Dartmouth which is supported by NCI Cancer Center Support Grant 5P30CA023108 and NIH S10 (1S10OD030242) awards. Spatial studies were conducted through the Dartmouth Center for Quantitative Biology in collaboration with the GMBSR with support from NIGMS (P20GM130454) and NIH S10 (S10OD025235) awards. Supplementary materials can be found at the following DOI: https://doi.org/10.5281/zenodo.8197573. Code for primary model implementation can be found at the following Github repository: https://github.com/zarif101/histopath_spatial_omics_pretrain

References

1. Egger, J. *et al.* Medical deep learning—A systematic meta-review. *Computer Methods and Programs in Biomedicine* **221**, 106874 (2022).
2. Cao, C. *et al.* Deep Learning and Its Applications in Biomedicine. *Genomics, Proteomics & Bioinformatics* **16**, 17–32 (2018).
3. Shamshirband, S., Fathi, M., Dehzangi, A., Chronopoulos, A. T. & Alinejad-Rokny, H. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics* **113**, 103627 (2021).
4. Van Der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat Med* **27**, 775–784 (2021).
5. Dimitriou, N., Arandjelović, O. & Caie, P. D. Deep Learning for Whole Slide Image Analysis: An Overview. *Front. Med.* **6**, 264 (2019).
6. Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C. & Petersson, L. A survey on graph-based deep learning for computational histopathology. *Computerized Medical Imaging and Graphics* **95**, 102027 (2022).
7. Azher, Z. L., Vaickus, L. J., Salas, L. A., Christensen, B. C. & Levy, J. J. Development of biologically interpretable multimodal deep learning model for cancer prognosis prediction. in *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing* 636–644 (ACM, 2022). doi:10.1145/3477314.3507032.
8. Chen, R. J. *et al.* 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* (eds. De Bruijne, M. et al.) vol. 12908 339–349 (Springer International Publishing, 2021).
9. Levy, J., Haudenschild, C., Barwick, C., Christensen, B. & Vaickus, L. Topological Feature Extraction and Visualization of Whole Slide Images using Graph Neural Networks. *Pac Symp Biocomput* **26**, 285–296 (2021).
10. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009). doi:10.1109/CVPR.2009.5206848.
11. Guan, Y. *et al.* Node-aligned Graph Convolutional Network for Whole-slide Image Representation and Classification. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 18791–18801 (IEEE, 2022). doi:10.1109/CVPR52688.2022.01825.
12. Wu, W., Liu, X., Hamilton, R. B., Suriawinata, A. A. & Hassanpour, S. Graph Convolutional Neural Networks for Histologic Classification of Pancreatic Cancer. *Archives of Pathology & Laboratory Medicine* (2023) doi:10.5858/arpa.2022-0035-OA.
13. Wu, Y., Cheng, Y., Wang, X., Fan, J. & Gao, Q. Spatial omics: Navigating to the golden era of cancer research. *Clinical & Translational Med* **12**, (2022).
14. Biancalani, T. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nat Methods* **18**, 1352–1362 (2021).
15. Zeng, Z., Li, Y., Li, Y. & Luo, Y. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biol* **23**, 83 (2022).
16. Svensson, V., Teichmann, S. A. & Stegle, O. SpatialDE: identification of spatially variable genes. *Nat Methods* **15**, 343–346 (2018).

17. Macenko, M. *et al.* A method for normalizing histology slides for quantitative analysis. in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 1107–1110 (IEEE, 2009). doi:10.1109/ISBI.2009.5193250.
18. Levy, J. J., Salas, L. A., Christensen, B. C., Sriharan, A. & Vaickus, L. J. PathFlowAI: A High-Throughput Workflow for Preprocessing, Deep Learning and Interpretation in Digital Pathology. *Pac Symp Biocomput* **25**, 403–414 (2020).
19. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. (2020) doi:10.48550/ARXIV.2002.05709.
20. Ciga, O., Xu, T. & Martel, A. L. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications* **7**, 100198 (2022).
21. Zeng, Y. *et al.* Identifying spatial domain by adapting transcriptomics with histology through contrastive learning. *Briefings in Bioinformatics* **24**, bbad048 (2023).
22. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. (2016) doi:10.48550/ARXIV.1609.02907.
23. Lee, J., Lee, I. & Kang, J. Self-Attention Graph Pooling. (2019) doi:10.48550/ARXIV.1904.08082.
24. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. (2019) doi:10.48550/ARXIV.1903.02428.
25. Calinski, T. & Harabasz, J. A dendrite method for cluster analysis. *Comm. in Stats. - Theory & Methods* **3**, 1–27 (1974).
26. Davies, D. L. & Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**, 224–227 (1979).
27. Bai, Z. *et al.* Tumor-Infiltrating Lymphocytes in Colorectal Cancer: The Fundamental Indication and Application on Immunotherapy. *Front. Immunol.* **12**, 808964 (2022).
28. Saltz, J. *et al.* Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep* **23**, 181-193.e7 (2018).
29. Magel, R. C. & Wibowo, S. H. Comparing the Powers of the Wald-Wolfowitz and Kolmogorov-Smirnov Tests. *Biom. J.* **39**, 665–675 (1997).
30. Idos, G. E. *et al.* The Prognostic Implications of Tumor Infiltrating Lymphocytes in Colorectal Cancer: A Systematic Review and Meta-Analysis. *Sci Rep* **10**, 3360 (2020).
31. Azher, Z. L. *et al.* Assessment of emerging pretraining strategies in interpretable multimodal deep learning for cancer prognostication. *BioData Mining* **16**, 23 (2023).
32. Cheerla, A. & Gevaert, O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* **35**, i446–i454 (2019).
33. Schirris, Y., Gavves, E., Nederlof, I., Horlings, H. M. & Teuwen, J. DeepSMILE: Contrastive self-supervised pre-training benefits MSI and HRD classification directly from H&E whole-slide images in colorectal and breast cancer. *Medical Image Analysis* **79**, 102464 (2022).
34. Chen, R. J. & Krishnan, R. G. Self-Supervised Vision Transformers Learn Visual Concepts in Histopathology. (2022) doi:10.48550/ARXIV.2203.00585.
35. Li, Z. *et al.* Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors. *iScience* **26**, 105872 (2023).