Graph Neural Networks Ameliorate Potential Impacts of Imprecise Large-Scale Autonomous Immunofluorescence Labeling of Immune Cells on Whole Slide Images

Ramya Reddy[*] RAMYA.REDDY3189@GMAIL.COM and
Ram Reddy[*] RAM.N.REDDY15@GMAIL.COM
Thomas Jefferson High School, Alexandria, VA 22312, USA
Cyril Sharma SHARMA.CYRIL@GMAIL.COM
Purdue University, West Lafayette, IN 47907, USA
Dr. Christopher Jackson CJACKSON21@MGH.HARVARD.EDU
Massachusetts General Hospital, Boston, MA 02114, USA
Scott Palisoul SCOTT.M.PALISOUL@HITCHCOCK.ORG
Rachael Barney RACHAEL.E.BARNEY@HITCHCOCK.ORG
Emerging Diagnostic and Investigative Technologies, Department of Pathology and Laboratory Medicine, Lebanon, NH 03756, USA
Dr. Fred Kolling FRED.W.KOLLING.IV@DARTMOUTH.EDU
Dartmouth Cancer Center, Lebanon, NH 03756, USA
Dr. Lucas Salas LUCAS.A.SALAS@DARTMOUTH.EDU
Dr. Brock Christensen BROCK.C.CHRISTENSEN@DARTMOUTH.EDU
Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756, USA
Dr. Gabriel Brooks GABRIEL.ARNE.BROOKS@DARTMOUTH.EDU
Department of Medicine, Section of Oncology, Dartmouth Hitchcock Medical Center, Lebanon, NH 03756, USA
Dr. Gregory Tsongalis GREGORY.J.TSONGALIS@HITCHCOCK.ORG
Dr. Louis Vaickus LOUIS.J.VAICKUS@HITCHCOCK.ORG
Emerging Diagnostic and Investigative Technologies, Department of Pathology and Laboratory Medicine, Lebanon, NH 03756, USA
Dr. Joshua Levy[†] JOSHUA.J.LEVY@DARTMOUTH.EDU
Emerging Diagnostic and Investigative Technologies, Departments of Pathology and Laboratory Medicine, and Dermatology, Dartmouth Hitchcock Medical Center, Lebanon, NH 03756, USA

Abstract

The characteristics of tumor-infiltrating lymphocytes (TILs) are essential for cancer prognostication and treatment through the ability to indicate the tumor's capacity to evade the immune system (e.g., as evidenced by nodal involvement). In general, presence of TILs indicates a favorable prognosis. Machine learning technologies have demonstrated remarkable success for localizing TILs, though these methods require extensive curation of manual annotations or restaining procedures that can degrade tissue quality, resulting in imprecise annotation. In this study, we co-registered tissue slides stained for both hematoxylin and eosin (H&E) and immunofluorescence (IF) as means to rapidly perform large-scale

[*] Denotes equal contribution
[†] To whom correspondence should be addressed

annotation of nuclei. We integrated the following approaches to improve the prediction of TILs: 1) minimized tissue degradation on same-section tissue restaining, 2) developed a scoring algorithm to improve the selection of patches for machine learning modeling and 3) utilized a graph neural network deep learning approach to identify relevant contextual features for lymphocyte prediction. Our graph neural network approach accounts for surrounding contextual micro/macro-architecture tissue features to facilitate interpretation of registered IF. The graph neural network compares favorably (F1-score=0.9235, AUROC=0.9462) to two alternative modeling approaches. This study brings insight to the importance of contextual information leveraged from within and around neighboring cells in a nuclei classification workflow, as well as elucidate approaches which enable the rapid generation of large-scale annotations of lymphocytes for machine learning approaches for immune phenotyping. Such approaches can help further interrogate the spatial biology of colorectal cancer tumors and tumor metastasis.

Keywords: tumor immune microenvironment, colon cancer, immune phenotyping, deep learning, graph neural networks

1. Introduction

Numerous studies have demonstrated that immune cell infiltrates play a crucial role in the adaptive immune response for specific bacterial and viral infections and various types of cancers (Aoshi et al., 2011; Galon et al., 2006). In the context of cancer, the presence, type, and location of tumor-infiltrating lymphocytes (TILs) play a crucial role in prognostication as this can be indicative of the tumor's capacity to evade or suppress the immune system (Morrison et al., 2022; Whiteside, 2022). While presence of TILs at the primary site is generally related to favorable prognosis (i.e., inverse correlation between presence of TILs and lymph node metastasis), it is believed that somatic alterations (e.g., transcriptional changes) within TILs at the primary site may indicate concurrent or future activity at the regional lymph node (Caziuc et al., 2019). As such, localizing TILs at the primary site at the time of resection may obviate the need for additional therapies. Alternatively, if molecular alterations within TILs suggests a metastatic phenotype, this may indicate that a less invasive secondary treatment is required (e.g., adjuvant chemotherapy, radiotherapy).

Many methods for determining the presence of TILs lack spatial resolution. However, hematoxylin and eosin (H&E) or immunohistochemical (IHC) stained tissue slides allow for spatial assessment. H&E provides a morphological and cytological examination, and IHC allows for multiplexing of protein markers that can disaggregate distinct cellular populations. For instance, immunoscore is a digital pathology technology that can assess the density of CD8+ (cytotoxic) and CD3+ (co-receptor which activates cytotoxic T cells) T cells inside the tumor and at the invasive margin and is highly effective for prognostication (Kwak et al., 2016). Increased density is associated with favorable prognosis through induced anti-tumoral cytotoxicity, though alternatively in select cases have also been associated with lower progression-free survival as the presence of these cells can signal immune exhaustion (Bruni et al., 2020; Ido et al., 2020). Presence of TILs can also represent an adaptive response to mismatch repair deficiency tumors in colorectal cancer patients (Jimenez-Rodriguez et al.). Studying similar effects using routine H&E stained slides through accurate localization of these immune cells is an emerging study area. Inferring locations of TILs is particularly challenging because it either requires large-scale annotation or wash and restain procedures to tag H&E stained tissue with various pro-

tein markers which deforms and degrades tissue. However, several emerging applications for careful restaining have demonstrated success in tagging millions of cells with molecular information with relatively little effort (Jackson et al., 2020b). Tagging with IHC can also be done using serial sections, but is suboptimal as there is no microarchitectural alignment due to the 5-micron separation between adjacent levels. Immunofluorescence (IF) staining can label several antigens in the same slide through the emission of relatively discrete imaging spectra (which has higher multiplexing potential compared to IHC). Furthermore, for tagging multiple markers in addition to routine staining, the application of IF after H&E is arguably less destructive than destaining H&E then staining with IHC on the same section. Several methods have been proposed to infer IF computationally, though, these generally rely on serial section staining (Burlingame et al., 2018). For the application of H&E after IF, registration can still complicate analyses with imprecise alignment though it is preferred to serial staining as it maintains significantly better microarchitectural alignment.

Machine learning algorithms, in particular, deep learning through the use of artificial neural networks (ANN), have demonstrated remarkable performance across a wide variety of image classification and detection tasks, all of which are relevant for isolating lymphocytes for further analysis (e.g., assessment of molecular alterations). Convolutional neural networks (CNN), a deep learning approach that accounts for spatial dependencies in images have gained increased attention over the past few years for TILs-specific inference tasks. Notable methods include segmentation (e.g., U-Net) (Jackson et al., 2020a; Saltz et al., 2018; Turkki et al., 2016), detection (e.g., panoptic segmentation, Fast R-CNN, Panoptic FPN; as popularized by the Detectron2 library (Wu et al., 2019)), and generative adversarial network approaches (Burlingame et al., 2018). Relatively unexplored is the use of detection networks (e.g., Detectron2) for these prediction tasks. Furthermore, surrounding spatial information can provide additional context.

In this study, we aim to explore algorithmic methods which, when used in conjunction with IF staining, can predict the presence of TILs while remaining sensitive to the imprecision in H&E cell-tagging from microarchitectural registration. We hypothesize that nascent graph neural network deep learning methods for cell type inference based on neighboring cells and micro/macro-architecture, can ameliorate lymphocyte inference challenges associated with IF tagging. Furthermore, we attempt to improve selected datasets for prototyping our algorithm through a patch-wise registration scoring algorithm.

Here, we investigate the effectiveness of graph neural networks (GNN) in identifying lymphocytes from H&E that were tagged through imprecisely registered IF.

2. Methods

2.1. Methods Overview

We performed: 1) large-scale annotation of lymphocytes using IF stains from the same section as the H&E, 2) developed a scoring algorithm to improve the selection of patches for algorithmic prototyping, and 3) utilized a GNN to mine contextual features relevant for IF-guided lymphocyte prediction. In brief, our method is as follows (See **Appendix A**):

1. Acquire H&E and IF stained whole slide images (WSI) from the same tissue section from 36 stage pT3 (pathological T-stage 3) matched colorectal cancer patients at

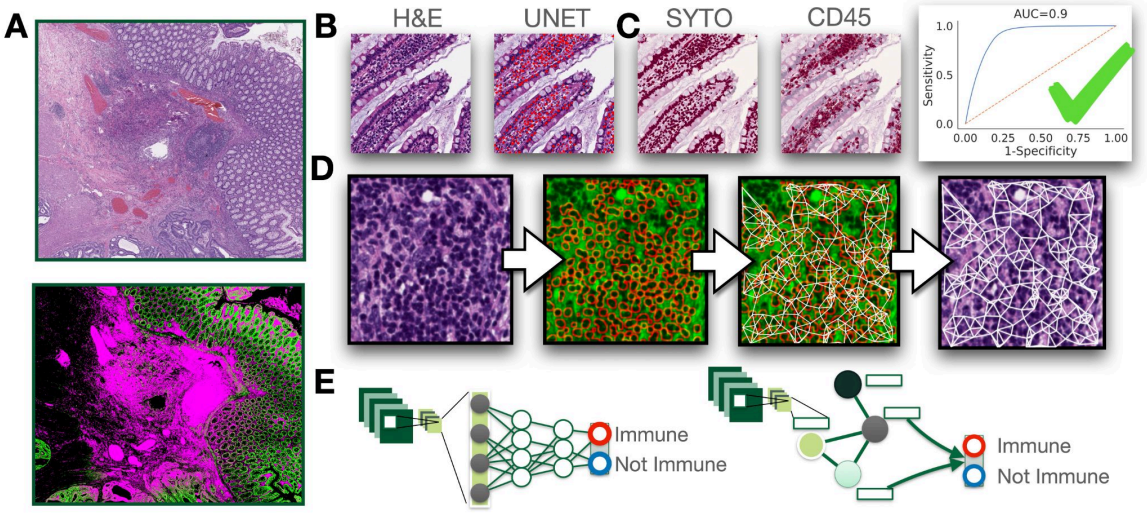

Figure 1: Workflow for cell dataset generation for lymphocyte prediction models: A) H&E and IF stains are collected and coregistered; B) UNET trained to predict nuclei to screen out C) slides based on concordance with SYTO13 stain via sensitivity analysis; D) graphical representation of detectron prediction and cell graph generation; E) application of CNN and GNN modeling approaches for immune cell prediction

Dartmouth Hitchcock Medical Center **(Data Collection, Section 2.2, Figure 1A)**.

2. Train U-net and detection neural network models on pathologist annotations from a combination of external public and private datasets to infer pixel-wise presence of nuclei and localize nuclei instances respectively **(Nuclei Detection, Section 2.3, Figure 1B)**.
3. Perform patch-wise registration of IF patches to H&E sections by aligning SYTO13 (fluorescent dye that binds to amino acids, with high fluorescent yield in nuclei) (Ullal et al., 2010) and Hematoxylin (nuclear) stains **(Stain Registration, Section 2.4)**.
4. Simultaneously score registration quality for specific patches and identify the optimal image intensity of the CD45 stain, to tag immune cells, on the H&E using a sensitivity analysis comparing the SYTO13 stain to the U-Net results **(Alignment Screening, Section 2.4, Figure 1C)**.
5. Use the Detectron2 nuclei detection model to create an annotated cell dataset from the stained WSIs, containing information on where the cell was located and whether it was an immune cell **(Cell Tagging, Section 2.5, Figure 1D)**.
6. Using the tagged cells, train and compare Detectron2, CNN, and GNN models for their ability to detect lymphocytes based on the annotated cell dataset **(Model Training, Section 2.6, Figure 1E)**.

2.2. Data Collection

The primary dataset utilized in this study was acquired from 36 Stage-pT3 matched (pTNM system; pT refers to invasion depth at primary site) colorectal cancer patients at Dartmouth-

Hitchcock Medical Center (DHMC), determined through a retrospective review of pathology reports from 2016 to 2019 following IRB review and approval. Half of the patients had concurrent tumor metastasis and were otherwise matched on age, sex, tumor grade, tissue size, mismatch repair status, and tumor site using iterative patient resampling with t-tests for continuous variables and fisher's exact tests for categorical variables. Tissue blocks were sectioned into 5-micron thick layers. Sections were stained with fluorescent-labeled, IF, antibodies for the following markers: 1) tumor/epithelial (PanCK), immune cells (CD45), and nuclei (SYTO13). These IF stains were initially acquired for a previously published study on spatial immune markers of metastasis, which utilized the GeoMX Digital Spatial Profiler (DSP, Nanostring Technologies, Seattle, WA) for image scanning into 16-bit unsigned color (one channel per stain) TIFF format images (Levy et al., 2022). After IF staining, the same sections were stained for H&E (without requiring destaining as the chemical reagents of the H&E minimally interacted with the fluorophores) and scanned into WSI using the Aperio AT2 scanner at 20x (8-bit unsigned color). The DHMC in-house dataset consisted of 36 WSIs, divided into 6,654 subarrays. Each of the subarrays were 768 pixels in each spatial dimension for patch-wise alignment and were further divided into nine square subarrays of side length 256-pixels without overlap, resulting in a total of 59,886 subarray images for cell identification.

Separately, we assembled an in-house dataset of 2,155 pathologist-annotated nuclei and a publicly available dataset of 30,837 pathologist-annotated nuclei to develop initial nuclei segmentation and detection approaches (Kumar et al., 2017, 2020).

2.3. Initial Nuclei Segmentation and Detection Models for Cell Localization

First, using the assembled nuclei detection dataset, a U-Net model was trained to detect the hematoxylin-stained nuclei on a pixel-wise basis for alignment scoring. The Detectron2 nuclei detection model was also trained on the same dataset which was more sensitive to adjacent cell boundaries through the adoption of panoptic segmentation methods and better allowed for cell counting (Wu et al., 2019).

The nuclei detection model was pre-trained with a 3x schedule that was available through the public Detectron2 Model Zoo, and it was then trained on the in-house data for a maximum of 5,000 epochs. Training was stopped when overfitting occurred (i.e., area under receiver operator curve (AUROC) on the validation set was maximal). The base learning rate was set to 0.0125 and 5 images were used per iteration. The model used a Mask R-CNN architecture that has a Residual Network+Feature Pyramid Network (ResNet+FPN) backbone based on the ResNet-101 model. It was tested using a detection threshold of 0.05 and a non-maximum suppression (NMS) threshold of 0.25. All hyperparameters were set to the Detectron2 config defaults if they were not otherwise specified after a coarse hyperparameter searchg.

2.4. Slide registration and screening imprecise alignments through sensitivity analysis

H&E and IF WSIs were registered through patch-wise alignment algorithms applied to the nuclear stains, Hematoxylin (determined using the Macenko stain deconvolution method), and SYTO13 staining intensities respectively. As the tissue was minimally deformed during H&E staining, the H&E and IF sections from the same tissue specimen were co-registered

through low-resolution rigid transformations. Then, both the H&E and IF WSI were divided into 768-pixel patches for more precise microarchitectural alignment (**Appendix A.1**). After registering the IF and H&E nuclear stains, the CD45 stain was overlaid by leveraging the same displacement field as the SYTO13 stain to tag immune cells.

We employed several mechanisms to screen out poorly aligned tissue patches. First, we calculated the pixel-wise difference in normalized staining intensities between the nuclear stains. The patch was removed from the set if this difference exceeded a specific threshold (mean squared error of 80). We also applied the trained U-Net model on the H&E patches to establish nuclei annotations. For each patch, we used a sensitivity analysis to calculate a C-statistic to provide an overall measure of agreement between the IF nuclear stain intensity and the predicted nuclei mask across a range of intensity thresholds. Patches with an overall agreement C-statistic of at least 0.85 were included in the set **(Figure S1)**. The sensitivity analysis was also used to identify a staining intensity threshold used to establish an IF nuclear mask based on maximum fidelity to the nuclei mask predictions. The same intensity threshold was applied to the CD45 stain to establish an immune cell mask, which was confirmed through visual inspection with collaborating pathologists.

2.5. Lymphocyte Prediction Dataset

There were no initial pre-existing annotations of immune and non-immune cells for the lymphocyte prediction model. By leveraging a highly accurate nuclei detection model (**Appendix A.2**) and registered immune cell masks, we were able to detect and label 5,377,681 nuclei, after filtering false positives. Nuclei were algorithmically annotated by overlaying the immune cell masks (**Appendix A.1**). This dataset contained 953,274 immune cells and 4,424,407 non-immune cells in the training/validation set and 19,408 immune cells and 90,231 non-immune cells in the test set.

2.6. Lymphocyte Prediction Modeling Approaches

We compared the following model approaches for the prediction of immune cells across our newly annotated dataset: 1) a cell detection model (Detectron2 framework) which outputs two classes (immune/non-immune), 2) a CNN-trained off of small images extracted around the bounding boxes, and 3) a GNN model trained on embeddings extracted from the CNN. Details of each model specification can be found below (**Appendix A.3, A.4**).

Detectron2, unlike other methods, can automatically detect the lymphocyte cells from the original image instead of just classifying subimages like the CNN and GNN.

We trained a convolutional neural network (i.e., ResNet18) on 64-pixel patches extracted around the initially predicted nuclei as means to more precisely model cellular morphology. Embeddings of the cells were extracted from the penultimate layer of the CNN, as well as the cells' positional x-y coordinates. These coordinates were used to generate graph datasets using both a k-nearest neighbor and radius neighbors graph for the GNN to train on. An ablation study was carried out to identify the optimal number of neighbors. Cells represent nodes of the graph and were connected by local proximity. Node attributes were captured using CNN embeddings.

A GNN model was used to integrate macroarchitectural cues for the inference of immune and non-immune cells (Ahmedt-Aristizabal et al., 2022; Fey and Lenssen, 2019). Such a model could potentially help overcome imprecise labeling from the automated registration,

Model	Accuracy	F1-Score	AUROC	IOU
Detectron2	82.11	0.8175	0.6961	0.6737
CNN	90.35	0.8995	0.9297	N/A
GNN	92.45	0.9235	0.9462	N/A

Table 1: Summary of model classification metrics

filtering, and application of the co-registered immune cell mask. The model outputs a vector representing the likelihood of observing an immune cell **(Figure S2)** (**Appendix A.3,A.4**). Several GNN architectures were compared as discussed in **Appendix A.4, B**.

All hyperparameters were determined through a coarse grid search (e.g., number of neighbors, learning rate, batch size; see **Appendix B**), with evaluation on the internal validation set. Optimal hyperparameters and number of neighbors were selected based on this internal evaluation.

2.7. Model evaluations

The classification performances of the lymphocyte prediction models were determined through reports of the accuracy, F1-score, AUROC, and intersection over union (IOU). A F1-score is equal to the harmonic mean of the precision and recall, and was calculated through the Scikit-learn Python library. Due to the inherent class imbalance between immune and non-immune cells in the data, we considered the weighted F1-score, with thresholds calculated using Youden's index (Ruopp et al., 2008). We depicted the number of true positives, true negatives, false positives, and false negatives through a confusion matrix, where immune cells were considered positive classifications and non-immune cells were considered negative classifications for calculation of sensitivity and specificity statistics. All of the previously described methods were used to compare the Detectron2, CNN, and GNN models. IOU was only used for the Detectron2 model, as it was the only model that had an output of bounding boxes, and showed how the predicted classifications compared to the annotated classifications. The CNN and GNN models were also interpreted through the generation of Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) embeddings of the nuclei, which allowed for visual assessment on the ability of the neural networks to delineate cell types.

3. Results

All three modeling approaches achieved an F1-score above 0.8 (**Appendix D**). Both the CNN and the GNN models outperformed the cell detection model for their ability to delineate immune cells. Notably, the GNN obtained optimal performance after taking into account Youden's threshold, obtaining an F1-score of 0.92 **(Table 1, Figure S3)**.

We visualized the UMAP projection of the extracted CNN and GNN embeddings of cells from the WSI **(Figure S4)**. The GNN model was able to learn contextual features and delineate cell types based on co-localized cells that gave this model a competitive advantage over the CNN model. We also visually compared the output from all three

modeling approaches versus the ground truth on a few randomly selected patches which corroborated with the aforementioned findings **(Figure 2)**.

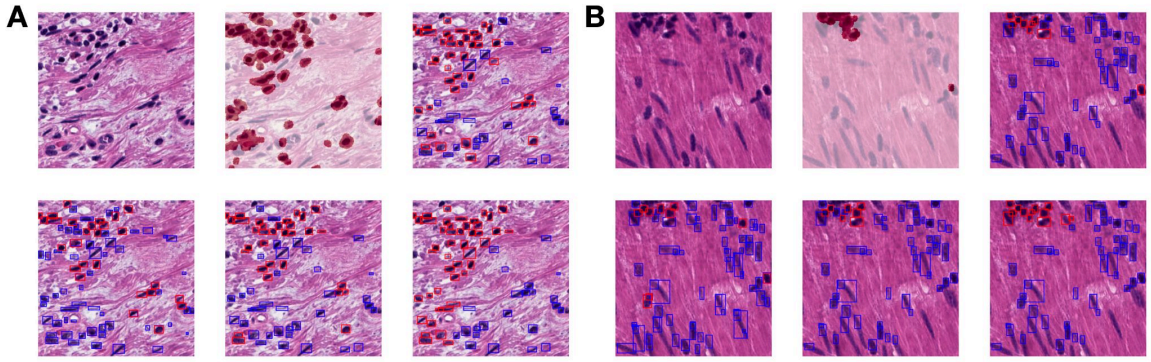

Figure 2: WSI patch, immune mask, ground truth images on upper row with predictions from Dectron2, CNN, and GNN models on bottom row for each example image subarray (A-B)

4. Discussion

The tumor immune microenvironment is an amalgamation of immune cells, chemokines, cytokines, and other immune modulators and plays a crucial role in coordinating the immune response to processes governing tumorigenesis and metastasis. As such, understanding spatial biology at the primary site is crucial for informing timely and relevant disease management options. Thus, the localization and quantification of distinct immune cell lineages may help inform the development of new spatial biomarkers. Informatics methods are still being developed to make sense of the data from this nascent field. Inference from morphological findings from an H&E tissue slide is an attractive approach because H&E staining is routinely done and inexpensive. However, optimal means of data collection and annotation are presently quite onerous and are an active area of exploration. As pathologists may incorrectly localize immune cells, IF staining to tag cells may present a viable alternative for labeling at scale, though it is expected that detection networks may struggle to make use of the antigenically tagged information.

In this study, we detailed an approach for the rapid and accurate immune cell annotation of nuclei based on the registration of IF which requires no tissue destaining. We applied geometric deep learning methods to potentially ameliorate inexact cell tagging by explicitly leveraging morphological and architectural information from neighboring cells. Our preliminary analysis suggests that GNNs, when combined with IF tagging of nuclei can accurately localize and tag immune cells in WSIs. We plan to further investigate the clinical utility of this technique and will further iterate and improve this method for downstream approaches for large-scale phenotyping in the context of studying tumor metastasis.

There were some limitations in the methodology of this study. We assumed the initial nuclei detection model achieved sufficient accuracy as judged by visual inspections from our practicing pathologists. While there may have been some inaccuracies in the initial cell localization, this is not outside of what is expected from other similar studies which leverage these datasets and further exploration is outside of the study scope (Mahmood et al., 2020).

Furthermore, the accuracy of our approach may have been impacted by manual staining processes for IF and batch effects. For instance, some slides were coverslipped for a few days, which may have impacted the specificity of the IF stain but was a pragmatic consideration when planning our experiment due to technical staffing shortages. Staining for H&E, by contrast, used automated staining protocols. Staining and cell tagging inaccuracies may have also been introduced in the registration process though we attempted to control for this through the sensitivity analysis. Future iterations of this model will attempt to more tightly control experimental preplanning through additional workflow automation.

Although the emphasis of this work was the application of methodology to this crucial clinical challenge, improvements in feature extraction methods and comparison of model architectures could result in more accurate detection models. There exists an exhaustive list of CNN and GNN model architectures from which to choose from as means to accomplish this aim. For instance, the CNN and GNN approaches featured in this work are not rotationally invariant or equivariant, despite application of rotation-based data augmentation. Leveraging rotationally invariant/equivariant methods such as Capsule Neural Networks, Equivariant Graph Neural Networks, PointNet-based neural networks, etc. may improve the external applicability of our approach (Chidester et al., 2019; Freyre et al., 2021; Keriven and Peyré, 2019; Mazzia et al., 2021; Satorras et al., 2022; Yao et al., 2021). Data preprocessing methods that oversegment images such as simple linear iterative clustering (SLIC) can decompose images into constituent superpixels. Building graphs from embeddings extracted from these superpixels may offer a more flexible modeling framework than featured in this work and will be something we plan to explore in future works (Dwivedi et al., 2022; He et al., 2022; Jaume et al., 2021; Sornapudi et al., 2018). We also did not perform an in-depth comparison between graph-based topological augmentation techniques (e.g., DropEdge, DropNode, etc.) and contrastive self-supervised learning methods, which warrants further assessment in this context (Hu et al., 2020, 2019; Qiu et al., 2020; Zhao et al., 2021; Zhu et al., 2021). While we developed nuclei detection algorithms using the Detectron2 framework, other object detection frameworks such as MMDetection offer a greater variety of architectures and modeling objectives and could be worth comparing to in future iterations of this work (Chen et al., 2019).

In the future, we plan to explore end-to-end training of these cell-graph neural networks, which jointly optimize both the CNN encoder and GNN prediction layers and compare to our two-stage approach (i.e, separately training CNN and GNN). It should be noted that there remains outstanding debate on optimal feature extraction methods for CNN and GNNs, specifically for inferring immune cell types, which are outside of the study scope. While we employed feature extraction methods across the cells, this component has not been well explored and could shed light on what morphological and architectural information is relevant for cell typing (e.g., nuclear morphology, cytoplasm, surrounding architecture). As an example, applying GNN to large patches extracted around nuclei as opposed to nuclear morphology would support the hypothesis that context matters. In the future, we would want to experiment with more complex GNN and CNN feature extraction models for the GNN, rather than simply use a pretrained CNN model, as it could yield more nuanced information for cell type classification.

Despite these limitations, our proposed lymphocyte prediction tool is valuable for researchers aiming to study spatial biology as it allows for the easy creation of robust IF

tagged dataset on millions of cells even in small-scale clinical feasibility studies. When pairing with macroarchitectural annotations (e.g., within the tumor, at the tumor-immune interface, etc.) identifying immune cells in these regions and concomitant molecular alterations can help infer the impact of immune cells in these regions for outcome such as metastasis, recurrence and survival. Furthermore, the presence of different cell types can confound or reduce the power of molecular association studies on microdissected tumor (Aran et al., 2015). Several recent studies have explored machine learning-based inference of highly multiplexed protein and RNA markers inferred on the cellular/subcellular level (He et al., 2020; Moses and Pachter, 2022; Zeng et al., 2022). Integrating histological information with predicted cell types through deconvolution approaches can more precisely identify canonical cellular populations (e.g., FOXP3+ T regulatory cells) which could further inform the coordinated response. In the future, we aim to investigate the interplay between histomorphology and protein/RNA expression localized to distinct locations on the slide through the adoption of highly multiplexed spatial assays including the GeoMX DSP and 10x Genomics Visium Spatial Transcriptomics.

5. Conclusion

In this work, we demonstrated the first application of GNN methods to H&E slides that were tagged through co-registered IF in the context of studying TILs. Our study suggests that contextual information leveraged from neighboring cells are important for nuclei classification and this workflow, as a whole, can be effective for generating large-scale immune phenotype data for studying the spatial biology of colorectal cancer tumors and tumor metastasis. We plan to further standardize this process and employ measures to explore the downstream implications of these findings with high precision.

Acknowledgments

This work was supported by the Dartmouth Hitchcock Medical Center Department of Pathology and Laboratory Sciences through the Emerging Diagnostic and Investigative Technologies program. JL is supported through NIH subawards P20GM104416 and P20GM130454.

References

David Ahmedt-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. A survey on graph-based deep learning for computational histopathology. *Computerized Medical Imaging and Graphics*, 95:102027, January 2022. ISSN 0895-6111. doi: 10.1016/j.compmedimag.2021.102027.

Taiki Aoshi, Shohei Koyama, Kouji Kobiyama, Shizuo Akira, and Ken J Ishii. Innate and adaptive immune responses to viral infection and vaccination. *Current Opinion in Virology*, 1(4):226–232, 2011. ISSN 1879-6257. doi: https://doi.org/10.1016/j.coviro.2011.07.002. URL https://www.sciencedirect.com/science/article/pii/S1879625711000538. Vaccines/Viral genomics.

Dvir Aran, Marina Sirota, and Atul J. Butte. Systematic pan-cancer analysis of tumour purity. *Nature Communications*, 6(1):8971, December 2015. ISSN 2041-1723. doi: 10.1038/ncomms9971.

Daniela Bruni, Helen K Angell, and Jérôme Galon. The immune contexture and immunoscore in cancer prognosis and therapeutic efficacy. *Nature reviews. Cancer*, 20(11): 662—680, November 2020. ISSN 1474-175X. doi: 10.1038/s41568-020-0285-7. URL https://doi.org/10.1038/s41568-020-0285-7.

Erik A. Burlingame, Adam A. Margolin, Joe W. Gray, and Young Hwan Chang. SHIFT: speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks. In John E. Tomaszewski and Metin N. Gurcan, editors, *Medical Imaging 2018: Digital Pathology*, volume 10581, page 1058105. International Society for Optics and Photonics, SPIE, 2018. doi: 10.1117/12.2293249. URL https://doi.org/10.1117/12.2293249.

Alexandra Caziuc, Diana Schlanger, Giorgiana Amarinei, and George Calin Dindelegan. Can tumor-infiltrating lymphocytes (tils) be a predictive factor for lymph nodes status in both early stage and locally advanced breast cancer? *Journal of Clinical Medicine*, 8 (4), 2019. ISSN 2077-0383. URL https://www.mdpi.com/2077-0383/8/4/545.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, and Jiarui Xu. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Benjamin Chidester, Tianming Zhou, Minh N Do, and Jian Ma. Rotation equivariant and invariant neural networks for microscopy image analysis. *Bioinformatics*, 35(14): i530–i537, July 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz353.

Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking Graph Neural Networks, May 2022.

Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric. *arXiv:1903.02428 [cs, stat]*, April 2019.

Christophe A. C. Freyre, Stephan Spiegel, Caroline Gubser Keller, Marc Vandemeulebroecke, Holger Hoefling, Valerie Dubost, Emre Cörek, Pierre Moulin, and Imtiaz Hossain. Biomarker-Based Classification and Localization of Renal Lesions Using Learned Representations of Histology—A Machine Learning Approach to Histopathology. *Toxicologic Pathology*, 49(4):798–814, June 2021. ISSN 0192-6233. doi: 10.1177/0192623320987202.

Jérôme Galon, Anne Costes, Fatima Sanchez-Cabo, Amos Kirilovsky, Bernhard Mlecnik, Christine Lagorce-Pagès, Marie Tosolini, Matthieu Camus, Anne Berger, Philippe Wind, Franck Zinzindohoué, Patrick Bruneval, Paul-Henri Cugnenc, Zlatko Trajanoski, Wolf-Herman Fridman, and Franck Pagès. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*, 313(5795):1960–1964, 2006. doi: 10.1126/science.1129139. URL https://www.science.org/doi/abs/10.1126/science.1129139.

Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4(8):827–834, August 2020. ISSN 2157-846X. doi: 10.1038/s41551-020-0578-x.

Fuyun He, M. A. Parvez Mahmud, Abbas Z. Kouzani, Adnan Anwar, Frank Jiang, and Sai Ho Ling. An Improved SLIC Algorithm for Segmentation of Microscopic Cell Images. *Biomedical Signal Processing and Control*, 73:103464, March 2022. ISSN 1746-8094. doi: 10.1016/j.bspc.2021.103464.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open Graph Benchmark: Datasets for Machine Learning on Graphs. In *Advances in Neural Information Processing Systems*, volume 33, pages 22118–22133. Curran Associates, Inc., 2020.

Gregory E. Idos, Janet Kwok, Nirupama Bonthala, Lynn Kysh, Stephen B. Gruber, and Chenxu Qu. The Prognostic Implications of Tumor Infiltrating Lymphocytes in Colorectal Cancer: A Systematic Review and Meta-Analysis. *Scientific Reports*, 10(1):3360, February 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-60255-4.

Christopher Jackson, Aravindhan Sriharan, and Louis Vaickus. A machine learning algorithm for simulating immunohistochemistry: development of sox10 virtual ihc and evaluation on primarily melanocytic neoplasms. *Modern Pathology*, 33, 04 2020a. doi: 10.1038/s41379-020-0526-z.

Christopher R. Jackson, Aravindhan Sriharan, and Louis J. Vaickus. A machine learning algorithm for simulating immunohistochemistry: Development of SOX10 virtual IHC and evaluation on primarily melanocytic neoplasms. *Modern Pathology*, pages 1–11, April 2020b. ISSN 1530-0285. doi: 10.1038/s41379-020-0526-z.

Guillaume Jaume, Pushpak Pati, Valentin Anklin, Antonio Foncubierta, and Maria Gabrani. HistoCartography: A toolkit for graph analytics in digital pathology. In *MICCAI Workshop on Computational Pathology*, pages 117–128. PMLR, 2021.

Rosa M. Jimenez-Rodriguez, Sujata Patil, Ajaratu Keshinro, Jinru Shia, Efsevia Vakiani, Zsofia Stadler, Neil H. Segal, Rona Yaeger, Tsuyoshi Konishi, Yoshifumi Shimada, Maria Widmar, Iris Wei, Emmanouil Pappou, J. Joshua Smith, Garrett Nash, Philip Paty, Julio Garcia-Aguilar, and Martin R. Weiser. Quantitative assessment of tumor-infiltrating lymphocytes in mismatch repair proficient colon cancer. *Oncoimmunology*, 9(1):1841948. ISSN 2162-4011. doi: 10.1080/2162402X.2020.1841948.

Nicolas Keriven and Gabriel Peyré. Universal Invariant and Equivariant Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7):1550–1560, 2017. doi: 10.1109/TMI.2017.2677499.

Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng-Ann Heng, Jiahui Li, Zhiqiang Hu, Yunzhi Wang, Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajeddin, Ali Gooya, Nasir Rajpoot, Xuhua Ren, Sihang Zhou, Qian Wang, Dinggang Shen, Cheng-Kun Yang, Chi-Hung Weng, Wei-Hsiang Yu, Chao-Yuan Yeh, Shuang Yang, Shuoyu Xu, Pak Hei Yeung, Peng Sun, Amirreza Mahbod, Gerald Schaefer, Isabella Ellinger, Rupert Ecker, Orjan Smedby, Chunliang Wang, Benjamin Chidester, That-Vinh Ton, Minh-Triet Tran, Jian Ma, Minh N. Do, Simon Graham, Quoc Dang Vu, Jin Tae Kwak, Akshaykumar Gunda, Raviteja Chunduri, Corey Hu, Xiaoyang Zhou, Dariush Lotfi, Reza Safdari, Antanas Kascenas, Alison O'Neil, Dennis Eschweiler, Johannes Stegmaier, Yanping Cui, Baocai Yin, Kailin Chen, Xinmei Tian, Philipp Gruening, Erhardt Barth, Elad Arbel, Itay Remer, Amir Ben-Dor, Ekaterina Sirazitdinova, Matthias Kohl, Stefan Braunewell, Yuexiang Li, Xinpeng Xie, Linlin Shen, Jun Ma, Krishanu Das Baksi, Mohammad Azam Khan, Jaegul Choo, Adrián Colomer, Valery Naranjo, Linmin Pei, Khan M. Iftekharuddin, Kaushiki Roy, Debotosh Bhattacharjee, Anibal Pedraza, Maria Gloria Bueno, Sabarinathan Devanathan, Saravanan Radhakrishnan, Praveen Koduganty, Zihan Wu, Guanyu Cai, Xiaojie Liu, Yuqin Wang, and Amit Sethi. A multi-organ nucleus segmentation challenge. *IEEE Transactions on Medical Imaging*, 39(5):1380–1391, 2020. doi: 10.1109/TMI.2019.2947628.

Yoonjin Kwak, Jiwon Koh, Duck Woo Kim, Sung Bum Kang, Woo Ho Kim, and Hye Seung Lee. Immunoscore encompassing cd3+ and cd8+ t cell densities in distant metastasis is a robust prognostic marker for advanced colorectal cancer. *Oncotarget*, 7(49):81778–81790, 2016. ISSN 1949-2553. doi: https://doi.org/10.18632/oncotarget.13207. URL https://www.oncotarget.com/article/13207/.

Joshua J. Levy, Carly A. Bobak, Mustafa Nasir-Moin, Eren M. Veziroglu, Scott M. Palisoul, Rachael E. Barney, Lucas A. Salas, Brock C. Christensen, Gregory J. Tsongalis, and Louis J. Vaickus. Mixed Effects Machine Learning Models for Colon Cancer Metastasis Prediction using Spatially Localized Immuno-Oncology Markers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 27:175–186, 2022. ISSN 2335-6936.

Faisal Mahmood, Daniel Borders, Richard J. Chen, Gregory N. Mckay, Kevan J. Salimian, Alexander Baras, and Nicholas J. Durr. Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images. *IEEE transactions on medical imaging*, 39(11): 3257–3267, November 2020. ISSN 0278-0062. doi: 10.1109/TMI.2019.2927182.

Vittorio Mazzia, Francesco Salvetti, and Marcello Chiaberge. Efficient-CapsNet: Capsule network with self-attention routing. *Scientific Reports*, 11(1):14634, July 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-93977-0.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, September 2018. ISSN 2475-9066. doi: 10.21105/joss.00861.

Steven L. Morrison, Gang Han, Faith Elenwa, John T. Vetto, Graham Fowler, Stanley P. Leong, Mohammed Kashani-Sabet, Barbara A. Pockaj, Heidi E. Kosiorek, Jonathan S. Zager, Vernon K. Sondak, Jane L. Messina, Nicola Mozzillo, Schlomo Schneebaum, Dale Han, and for the Sentinel Lymph Node Working Group. Is the presence of tumor-infiltrating lymphocytes predictive of outcomes in patients with melanoma? *Cancer*, 128(7):1418–1428, 2022. doi: https://doi.org/10.1002/cncr.34013. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/cncr.34013.

Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19 (5):534–546, May 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01409-2.

Jiezhong Qiu, Qibin Chen, Yuxiao Dong, Jing Zhang, Hongxia Yang, Ming Ding, Kuansan Wang, and Jie Tang. Gcc: Graph contrastive coding for graph neural network pre-training. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1150–1160, 2020.

Marcus D. Ruopp, Neil J. Perkins, Brian W. Whitcomb, and Enrique F. Schisterman. Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection. *Biometrical journal. Biometrische Zeitschrift*, 50(3):419–430, June 2008. ISSN 0323-3847. doi: 10.1002/bimj.200710415.

Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R. Shroyer, Tianhao Zhao, Rebecca Batiste, and John Van Arnam. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Reports*, 23(1):181–193.e7, 2018. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2018.03.086. URL https://www.sciencedirect.com/science/article/pii/S2211124718304479.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) Equivariant Graph Neural Networks, February 2022.

Sudhir Sornapudi, Ronald Joe Stanley, William V. Stoecker, Haidar Almubarak, Rodney Long, Sameer Antani, George Thoma, Rosemary Zuna, and Shelliane R. Frazier. Deep Learning Nuclei Detection in Digitized Histology Images by Superpixels. *Journal of Pathology Informatics*, 9(1):5, January 2018. ISSN 2153-3539. doi: 10.4103/jpi.jpi_74_17.

Riku Turkki, Nina Linder, Panu E. Kovanen, Teijo Pellinen, and Johan Lundin. Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *Journal of Pathology Informatics*, 7 (1):38, 2016. ISSN 2153-3539. doi: https://doi.org/10.4103/2153-3539.189703. URL https://www.sciencedirect.com/science/article/pii/S2153353922005569.

Anirudh J. Ullal, David S. Pisetsky, and Charles F. Reich. Use of SYTO 13, a fluorescent dye binding nucleic acids, for the detection of microparticles in in vitro systems. *Cytometry.*

Part A : the journal of the International Society for Analytical Cytology, 77(3):294–301, March 2010. ISSN 1552-4922. doi: 10.1002/cyto.a.20833.

Theresa L. Whiteside. *Tumor-Infiltrating Lymphocytes and Their Role in Solid Tumor Progression*, pages 89–106. Springer International Publishing, Cham, 2022. ISBN 978-3-030-91311-3. doi: 10.1007/978-3-030-91311-3_3. URL https://doi.org/10.1007/978-3-030-91311-3_3.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Kai Yao, Kaizhu Huang, Jie Sun, Amir Hussain, and Curran Jude. PointNu-Net: Simultaneous Multi-tissue Histology Nuclei Segmentation and Classification in the Clinical Wild, November 2021.

Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks. *Briefings in Bioinformatics*, page bbac297, July 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac297.

Tong Zhao, Yozen Liu, Leonardo Neves, Oliver Woodford, Meng Jiang, and Neil Shah. Data augmentation for graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11015–11023, 2021.

Yanqiao Zhu, Yichen Xu, Qiang Liu, and Shu Wu. An empirical study of graph contrastive learning. *arXiv preprint arXiv:2109.01116*, 2021.

Appendix A. Supplementary Methods

A.1. Microarchitectural Alignment and Additional Cell Tagging Details

For each pair of nuclear stains, akaze features were extracted from each nucleus stain image from matching patches, and a k-nearest neighbors and radius neighbors brute force feature matcher were used to identify matching local features between the images. Matched features between the H&E and IF nuclear stains were used to compute a perspective transformation for the final registration.

The immune cells were tagged by calculating the percentage overlap between the predicted nuclei instance mask/bounding box by the cell detection model and the immune cell mask. If at least 25% of the nucleus instance mask was labeled pixel-wise as an immune cell, the nucleus was tagged as such and was not labeled as an immune cell if it failed to surpass this threshold.

A.2. Detection Dataset Format

These datasets were prepared in the Microsoft COCO (MS COCO) format. The COCO format is commonly used for machine learning and computer vision projects and can be used for object detection, segmentation, and captioning. It uses the JavaScript object notation (JSON) format and includes information about the categories the images were

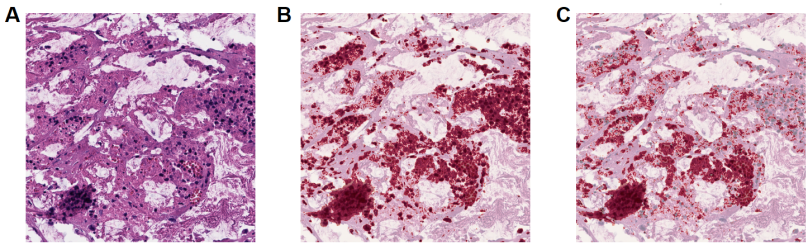

Figure S1: Comparison of (A) H&E stain on a WSI with (B) an additional nuclei mask (SYTO13) and (C) immune mask (CD45)

classified into, raw image information such as filename and image size, and a list of all object annotations for every image in the dataset. Training and validation datasets were generated which contained nuclei annotations. For annotations, nuclei areas were calculated using the contour_area method of the OpenCV-Python package based on manually segmented splines placed by the pathologists. The bounding box format for the annotations also followed the default boxmode of absolute minimum-X and minimum-Y coordinates, width, and height.

The Detectron2-derived lymphocyte prediction model dataset followed the same MS COCO dataset format as what was done for the nuclei detection model (e.g., similar area and boxmode annotations). However, there were two categories of images: immune and non-immune cells. The datasets used for the CNN and GNN lymphocyte prediction models were also adapted from the Detectron2 model datasets.

A.3. Addressing Class Imbalances

There was also a significant class imbalance of lymphocyte and non-lymphocyte nuclei, with a ratio of approximately one immune cell for every four non-immune cells. To address these challenges, we employed class balancing techniques, such as resampling (dynamic batch-wise undersampling), reweighting the model objective, and using evaluation metrics that were relatively robust to these differences. We found these methods did not significantly impact performance. It is expected that without reweighting, the decision threshold would be shifted to reflect this proportion of immune cells.

A.4. Additional Motivation/Training Details for Modeling Approaches

We used the Detectron2 library instead of other object detection methods like YOLO because many of the methods featured in the Detectron2 framework are easier to implement, more accurate than other methods, and had demonstrated to us favorable performance in underrepresented classes. Lymphocyte prediction relied heavily on the modeling results from our initial nuclei detection model, which generated training datasets for all three modeling approaches.

The Detectron2 lymphocyte prediction model also takes less time to train compared to the CNN and GNN models because it contains state-of-the-art prewritten libraries based on PyTorch; however, the study results indicate that the Detectron2 model requires a relatively extensive amount of data **(Table 1)**. Due to this, Detectron2 is able to make predictions more quickly during inference.

The Detectron2 lymphocyte prediction model used the same pre-trained model, the number of training images per step, and Mask R-CNN architecture as that of the nuclei detection model. The model was trained for 4,000 epochs at a base learning rate of 0.0125. Training was stopped after no significant change in validation set accuracy. Visual assessments of the model predictions were done at a detection threshold of 0.2 and NMS threshold of 0.3. All hyperparameters were set to the Detectron2 configuration defaults if not otherwise specified. The model output bounding boxes, category names, and measure of model certainty (%) of the category for each detected cell.

As a classification model, the CNN had an additional step, requiring the Detectron2 detected cell annotations to be used as an input, instead of just an image, in order to have an output of category classifications. The model capacity of the residual neural network is greater than cell class prediction layers after proposed regions of interest from Detectron2. The CNN model was trained using the PyTorch framework. A data loader was configured which loads detected cells into a Torch tensor format. Data augmentation was performed including horizontal and vertical flips, random rotations, and color jitter. The CNN lymphocyte prediction model was trained on the training dataset of nuclei for 20 epochs, a learning rate of 0.0125, and a batch size of 128 cells. The validation set F1-score was assessed after each epoch; ultimately, we saved the model at the epoch with the highest F1-score.

The GNN model had two extra steps; the first step is the same as that of the CNN model, where it requires the Detectron2 nuclei detection model to process an image, and a second step of extracting embeddings (or features) from the detected cells. The GNN model creates a more abstract representation of the WSI through the graph datasets and extracted embeddings, which can lead to an overall better generalization of the data. The model was trained for 200 epochs with a batch size of 32 and a learning rate of 1e-4 and similar to the CNN model, we kept track of the epoch which achieved the highest F1-score and saved the model when this occurred. The model was trained using the PyTorch-Geometric software framework, which takes as input a graph dataset and outputs a probability vector (Torch tensor). We additionally explored the usage of equivariant graph neural networks. However, we found in this simple use case that the vanilla GNNs were sufficient for demonstrating performance differences as compared to the CNN though this is an active area of research.

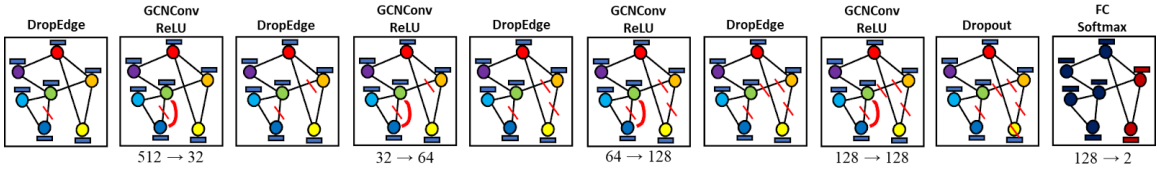

Figure S2: Graph Neural Network architecture

Appendix B. Selection of Hyperparameters and Ablation over Number of Neighbors

The following hyperparameters and number of neighbors (for the graph construction) were optimized over (selected hyperparameters/number of neighbors in bold):

- Batch size: 16, **32**, 64
- Model warmup: 0, **50**, 100
- Learning rate: 1e-2, 1e-3, **1e-4**
- Number of Epochs: 100, 150, **200**
- Number of Neighbors: 3, **8**, 10
- Usage of DropEdge: **Yes**, No

Appendix C. Code Availability

The python programming language (Version 3.8.8) was used in all coding aspects of this study. Code was prototyped using Jupyter notebook (version 6.4.11) and leveraged computing resources (Tesla v100s GPUs) housed at the Dartmouth College Discovery Research Computing cluster. Code is available upon reasonable request.

Appendix D. Supplementary Result Figures

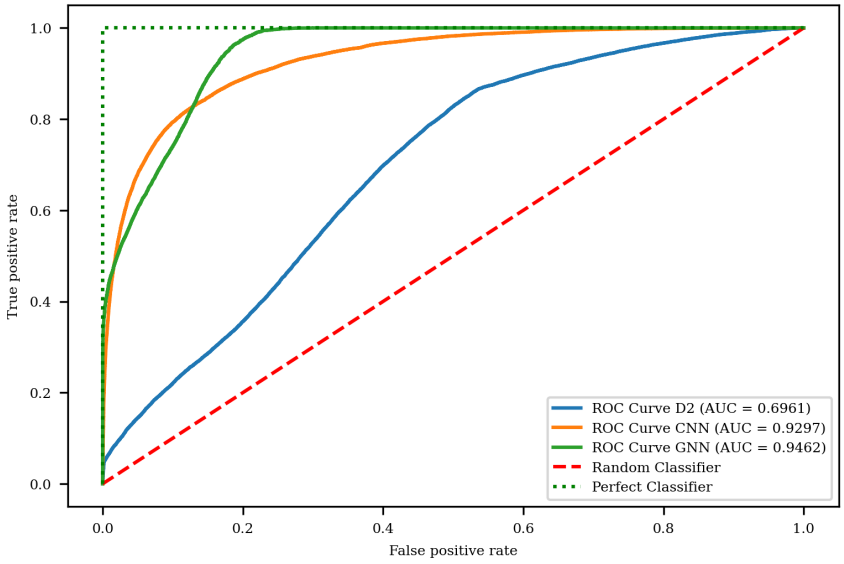

Figure S3: Graph comparing ROC curves of models

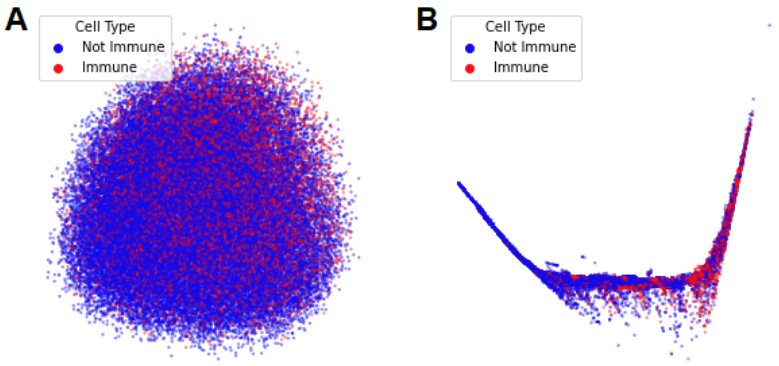

Figure S4: PCA projection of extracted embeddings from cell images using (A) CNN and (B) GNN models

List of Changes

Graph Neural Networks Ameliorate Potential Impacts of Imprecise Large-Scale Autonomous Immunofluorescence Labeling of Immune Cells on Whole Slide Images

Author Comments to Editor

Our team looks forward to presenting our work in Amsterdam. We have included description of point-by-point changes to the manuscript in the form of reviewer responses in the following pages.

In addition, we have made the following changes to our work based on suggestions by one of our oncologist collaborators:

- Clarification in the introduction that generally TILs are related to favorable prognosis, *Introduction*, "…While presence of TILs at the primary site is generally related to favorable prognosis, it is believed that somatic alterations (e.g., transcriptional changes) within TILs…" and reporting of inverse correlation with overall presence of TILs and nodal metastasis.
- Additional supportive references for TILs, specifically for CRC versus breast cancer, *Introduction*, "…Presence of TILs are represents an adaptive response to mismatch repair deficiency tumors in colorectal cancer patients…"
- Clarification that stage 3 refers to histological staging based on invasion depth, stage pT3 according to the pTNM classification system, *Methods*, "…stage-pT3 matched (pTNM system; pT refers to invasion depth at primary site)…"
- Ensured consistent use of TILs over TIL as both are plural.

Reviewer 1

– Several works in different domains have shown the potential of using CNNs as embeddings extraction to construct a graph representation of medical data. However, in the current work, authors mentioned that they use the extracted embeddings along with the cell's potential coordinates. A usual protocol to obtain a more sparficied graph is to generate an oversegmentation (e.g., superpixels [*]) of the image and take it as a graph. Firstly, what is the effect on the number of neighbours (k) for the performance? Secondly, how the performance and efficiency is compared against other forms to create the graph, for example, get it from oversegmenting the images?

[*] Dwivedi, V. P., Joshi, C. K., Laurent, T., Bengio, Y., & Bresson, X. (2020). Benchmarking graph neural networks. arXiv preprint arXiv:2003.00982.

Response: We thank the reviewers for these suggestions. The reviewer has suggested performing an ablation study over the number of nearest neighbors. We have added two other nearest neighbor settings. We report that performance was optimal for the original set selected, which has been included in the manuscript, Appendix, "...The following hyperparameters and number of neighbor…". The author has also suggested a comparison to SLIC based methods. This is uncommon although not unheard of for applications of graph neural networks to histopathology. However, there are limited applications of both SLIC and GNNs simultaneously for nuclei detection and as such we feel is out of scope for the current article as it further complicates analysis and detracts from the central aims of this work, which is centered on application of methodology. This is something we will pursue in future work and we have added such discussion to the limitations section, Disscussion, "...Data preprocessing methods that oversegment images such as simple linear iterative clustering (SLIC) can decompose images into constituent superpixels. Building graphs from embeddings extracted from these superpixels may offer a more flexible modeling framework than featured in this work and will be something we plan to explore in future works...". We have included relevant citations for previous applications of SLIC.

– Whilst the application description is very well-taken, the technical side is limited in explanation. Although technical content is provided in the appendixes. There are open questions. For example in Figure S2 authors include DropEdge – what is the ration that you are using for the dropedge? What not using another type of feature or topological augmenters?

Response: DropEdge was used as means to penalize against spurious connections. DropEdge has been used in several works as a topological augmenter. The reviewer correctly pointed out that this does not substantiate its usage. We did experiment with DropEdge, as detailed in the Appendix, , "...The following hyperparameters and number of neighbor…". While the focus of this work was to apply previous methodology, not substantially exploring topological augmenters is a significant study limitation and something worth exploring in a future work, tailored for journal publication. We have acknowledged in limitations and discussed other operators, see discussion, "...We also did not perform an in-depth comparison between graph-based topological augmentation techniques (e.g., DropEdge, DropNode, etc.) and contrastive self-supervised learning methods, which warrants further assessment in this context...".

Moreover, experimental wise authors are limited to Detectron2 (which is a widely used model), CNN and GNN. What about other techniques beyond detectro2 such as MMDetection etc?

Response: The reviewer has suggested exploring architectures and modeling objectives featured in the MMDetection framework as opposed to the more limited set featured in Detectron2. This is a worthwhile

consideration and something we have now included as discussion in limitations along with the SLIC discussion, “…While we developed nuclei detection algorithms using the Detectron2 framework, other object detection frameworks such as MMDetection offer a greater variety of architectures and modeling objectives and could be worth comparing to in future iterations of this work…”.

Reviewer 2

- While the main paper is strong on details in of the biological type, much of the machine learning choices have been deferred to the appendix. Even there, I have been unable to find hyperparameter settings for the cutoff radius, or the number of nearest neighbours. Also, why was training of the Detectron2 stopped when accuracy exceeded 90%?

  Response: We have now included in the supplementary materials hyperparameters that were coarsely scanned over, *Section 2.6* "…All hyperparameters were determined through a coarse grid search (e.g., number of neighbors, radius, learning rate, batch size; see…", which includes information on nearest neighbors. We have clarified the model training stop criteria to be: "…area under receiver operator curve (AUROC) on the validation set was maximal…" (Methods).

- The CNN and GNN models are not rotation equivariant. While rotation augmentation is used—unclear to what extent—using invariant or equivariant models might prove more robust.

  Response: This is an excellent suggestion. While it is intractable to implement all state-of-the-art invariant or equivariant models, we have attempted to implement at least one approach *E(n)-Equivariant Graph Neural Networks* with citations added, and noted where other approaches, e.g., capsule networks for feature extraction, PointNets, etc would be appropriate. We mentioned in the supplemental methods section but wanted to focus this work on the main comparisons (CNN vs GNN). See discussion, "…Although the emphasis of this work was the application of methodology to this crucial clinical challenge, improvements in feature extraction methods and comparison of model architectures could result in more accurate detection models. There exists an exhaustive list of CNN and GNN model architectures from which to choose from as means to accomplish this aim. For instance, the CNN and GNN approaches featured in this work are not rotationally invariant or equivariant, despite application of rotation-based data augmentation. Leveraging rotationally invariant/equivariant methods such as Capsule Neural Networks, Equivariant Graph Neural Networks, PointNet-based neural networks, etc. may improve the external applicability of our approach…".

Feedback I believe this is an interesting work. The goal is clearly motivated and the results show promise. The combination of multiple scales of the CNN and the GNN is a promising research direction. I agree with the discussion that the success hinges on the Detectron2 module not missing too many candidates. I believe the biggest weaknesses are on the machine learning end. Some of the work will be hard to reproduce due to missing hyperparameters. It seems that the CNN and GNN were trained separately, while the combination of these two could have been trained end-to-end. I believe it would have been interesting to compare these two approaches.

Response: We concur with the reviewer about implementing an end-to-end approach, which is certainly possible and something we have been thinking about for a long time. We are planning to explore this approach as a future direction, *Discussion*, "…In the future, we plan to explore end-to-end training of these cell-graph neural networks, which jointly optimize both the CNN encoder and GNN prediction layers and compare to our two-stage approach (i.e, separately training CNN and GNN)….".

Typo's

- I cannot find the first occurrence of SYTO where the acronym is explained.

  Response: We have added clarification of SYTO13 in the Methods overview section: "…SYTO13 (fluorescent dye that binds to amino acids, with high fluorescent yield in nuclei)…" with a new citation.