A New Graph Node Classification Benchmark: Learning Structure from Histology Cell Graphs

Claudia Vanea[1, 2], Jonathan Campbell[1, 2, 3], Omri Dodi[4], Liis Salumäe[5], Karen Meir[4], Drorith Hochner-Celnikier[4], Hagit Hochner[6], Triin Laisk[7], Linda M. Ernst,[8] Cecilia M. Lindgren[*2], Christoffer Nellåker[*1, 2]
claudia.vanea@wrh.ox.ac.uk, jcampbell@robots.ox.ac.uk,
omri.dodi@mail.huji.ac.il, liis.salumae@kliinikum.ee,
karenm@hadassah.org.il, hochner@hadassah.org.il,
hagit.hochner@mail.huji.ac.il, triin.laisk@ut.ee,
lernst@northshore.org, cecilia.lindgren@bdi.ox.ac.uk,
christoffer.nellaker@bdi.ox.ac.uk
Nuffield Department of Women's & Reproductive Health, University of Oxford[1]
Big Data Institute, University of Oxford[2]
Visual Geometry Group, University of Oxford[3]
Faculty of Medicine, Hadassah Hebrew University Medical Center[4]
Department of Pathology, Tartu University Hospital[5]
Braun School of Public Health, Hebrew University of Jerusalem[6]
Institute of Genomics, University of Tartu[7]
Northshore University HealthSystem, University of Chicago[8]

Abstract

We introduce a new benchmark dataset, *Placenta*, for node classification in an underexplored domain: predicting microanatomical tissue structures from cell graphs in placenta histology whole slide images. This problem is uniquely challenging for graph learning for a few reasons. Cell graphs are large (>1 million nodes per image), node features are varied (64-dimensions of 11 types of cells), class labels are imbalanced (9 classes ranging from 0.21% of the data to 40.0%), and cellular communities cluster into heterogeneously distributed tissues of widely varying sizes (from 11 nodes to 44,671 nodes for a single structure). Here, we release a dataset consisting of two cell graphs from two placenta histology images totalling 2,395,747 nodes, 799,745 of which have ground truth labels. We present inductive benchmark results for 7 scalable models and show how the unique qualities of cell graphs can help drive the development of novel graph neural network architectures.

1 Introduction

The initial graph representation learning benchmarks facilitated the creation of the first graph neural network (GNN) architectures. *Cora*, *Citeseer*, *Pubmed* and *NELL* [29], [3] drove the development of Graph Convolutional Network (GCN) [16], *Reddit* was introduced alongside GraphSAGE [12], and Graph Attention Network (GAT) [26] achieved state-of-the-art on the protein-protein interaction (*PPI*) dataset [34]. However, it soon became apparent that these datasets were not sufficiently complex to challenge GNNs [13, 28]. The datasets had small graphs or a small number of graphs (the largest dataset, *Reddit*, had 232,965 nodes), they had potentially misleading train, validation and test splits [22], and they were so well suited to graph learning that randomly initialised, untrained models could

[*]Equal contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: Dataset characteristics for the intersection graph.

	Nodes	Edges	Labels	Homophily	Degree
Cell Graph 1	1,159,186	2,603,797	703,165	0.9868	5.4925
Cell Graph 2	1,236,561	2,882,292	96,580	0.9984	5.6618

perform nearly as well as trained models [27]. In recent years, there has been an influx of graph benchmark datasets which address these limitations. They are larger and cover more domains, have realistic train, validation and test splits, and are challenging enough to have not been 'solved' by current GNN architectures [13, 5, 24].

Cell graphs in histology whole slide images (WSIs) are a domain which has been largely unexplored for graph learning and present a challenge to the graph learning community [18, 33, 10, 4, 19]. Histology is the study of tissues and cells under a microscope. A histology WSI is a digitised, high-resolution scan of thin tissue samples showing individual cells, which can be upwards of 70GBs per image. All organs are composed of cells that collectively build up the structures and tissues comprising the body. Consequently, representing organs as cell graphs is a logical construction closely matching the true physical system. The value of cell graphs has been highlighted [30, 14, 23], but with today's computational power and automated aggregation methods, it can reach its full potential.

Here, we introduce *Placenta*, a cell graph from two WSIs as a new benchmark dataset to facilitate GNN creation. More specifically, we present a node classification task of predicting one of 9 placenta microanatomical tissue structures to which cells belong (Appendix A). This problem is challenging from a graph learning perspective for a number of reasons. Firstly, each WSI contains a cell graph of >1 million nodes, the node features are defined from an imbalanced distribution of cell classes, and the tissue labels also have a high, but expected, class imbalance, from 0.21% of the data to 40.0% (Appendix B). Secondly, unlike most graph datasets, node clusters are locally homophilic but spatially distributed throughout the graph, requiring models to understand same-type node clusters at different locations. Thirdly, as a WSI is only ever a small sample of the larger structure with arbitrary cut-offs in edges, out-of-plane cells, and optical and mechanical artefacts, the property of data incompleteness is inherently a feature of this data domain. Finally, placental tissue structures grow from one another in a tree-like manner [1, 7], so singular structures can contain multiple tissue classifications and these structures will vary in size (30-5,000 μm^2). Models must discriminate between small changes across small regions within node clusters but have a large enough receptive field to identify the largest of tissues.

Relevant to the real world domain, developing automated methods which can quantify the structural tissue units within histology images would be an invaluable tool for the scientific community and, in future, may have applications to clinical histopathologic diagnoses and patient care. Furthermore, this dataset provides a benchmark for developing novel, scalable GNNs that can adapt to node feature and class label imbalance, that can understand both large and small node communities distributed throughout the graph, and that are robust to data incompleteness. These models could be applied to new domains with similar qualities, such in modelling disease epidemiology.

2 Dataset Description

Overview At a high level, the learning task is to understand placental tissue structures as the composition of their cellular communities. Tissue structures in the placenta are precisely defined and distinguished from one another by their cellular composition [1]. In a WSI these manifest as spatially clustered groups of cells, with corresponding tissue structures comprising 10s-10,000s of cell nodes. Inductive message-passing GNNs, which can aggregate node neighbourhood information irrespective of graph shapes and sizes are, in theory, a natural fit for this task.

The dataset we introduce for this task, *Placenta*, consists of two combined cell graphs constructed across two healthy, term placenta hematoxylin and eosin stained WSIs from two institutes (Table 1). Ground truth labels were created by the primary author and validated by four practising perinatal pathologists (details in Vanea et al. [25]). See Figure 1 for a high level visualisation of all cell graph nodes and labelled ground truth nodes and Figure 2 for a representative region. The use of these

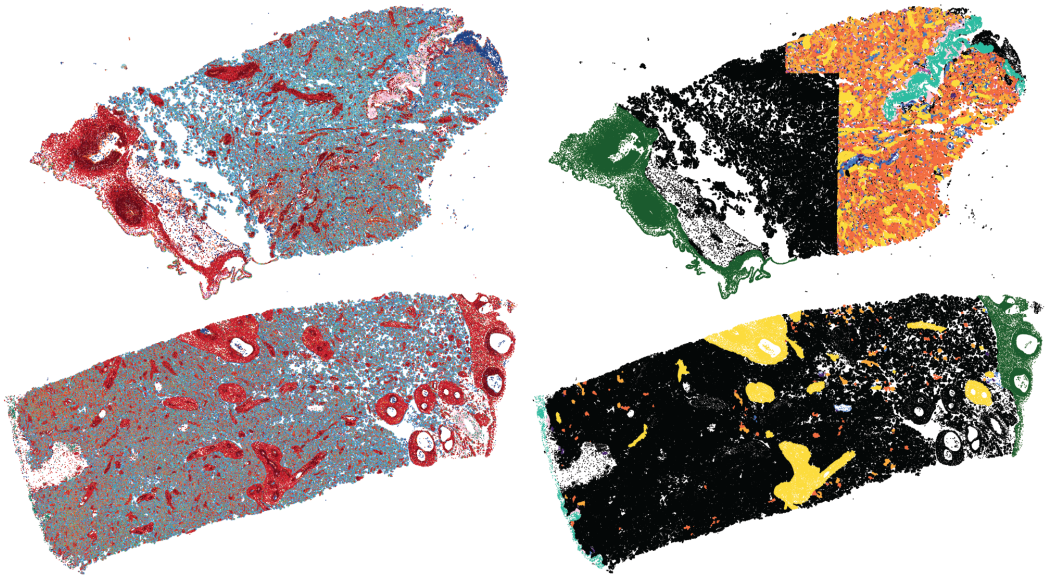

Figure 1: Cell graph nodes across from whole slide images (left) with corresponding labelled ground truth tissue nodes (right). Black tissue nodes are unlabelled.

images was approved and the requirement for consent was waived by local ethics committees at both institutes (approval 0735-18-HMO and 289/T-5).

We recommend a train, validation, test split (Appendix C) which ensures tissue types have similar distributions across splits. The type of tissue in the placenta can be influenced by macroscopic location (for example, proximity to the chorionic or basal plate), so these dataset split regions are chosen to cover a range of locations and to avoid neighbourhood information leakage.

Nodes and Node Features We use a prior two-stage deep learning pipeline to identify the centre point of all nuclei across the WSI and classify those nuclei as belonging to one of 11 cell types. The nucleus coordinates define the node coordinates in the graph and the 64-dimension embeddings of the penultimate layer of the cell classifier model are used for the node features. Full details of this pipeline can be found in Vanea et al. [25].

Edges and Edge Features We explore three edge-building algorithms which act as a proxy for cellular interaction: KNN [6], Delaunay Triangulation [11], and the intersection of the two (Appendix D). Unless stated otherwise, we use the intersection graph with k=5 for all experiments.

Edge construction in the context of cell graphs are representations of the approximate biological likelihood that nearby cells interact as part of the same structure. The Euclidean distance-based edge construction we have explored is one method for this, but alternative approaches for constructing the network are worth considering. We leave this as an open question for the community.

3 Experiments

Experimental Setup We first construct a cell graph across all cells in both WSIs. Nodes are masked as either unlabelled, training, validation, or test nodes and only nodes in the training set are used for back propagation. Both cell graphs are combined into one disjointed graph, allowing the sampler to treat it as a dataset of one graph. We restrict our experiments to scalable GNNs as the full graph cannot fit into GPU memory. During training, for **GraphSAGE** [12], **ClusterGCN** [5], **GraphSAINT** [31], **ShaDow** [32] and **SIGN** [9], node embeddings are sampled or precomputed according to original implementations. For models introduced without a sampling technique (**GAT** [26] and **GATv2** [2]), we use ClusterGCN subgraph sampling (ClusterGAT and ClusterGATv2). During inference, all neighbourhood nodes are used for aggregation. We additionally compare against two baselines, an **MLP on node embeddings** and a **randomly initialised GraphSAGE**.

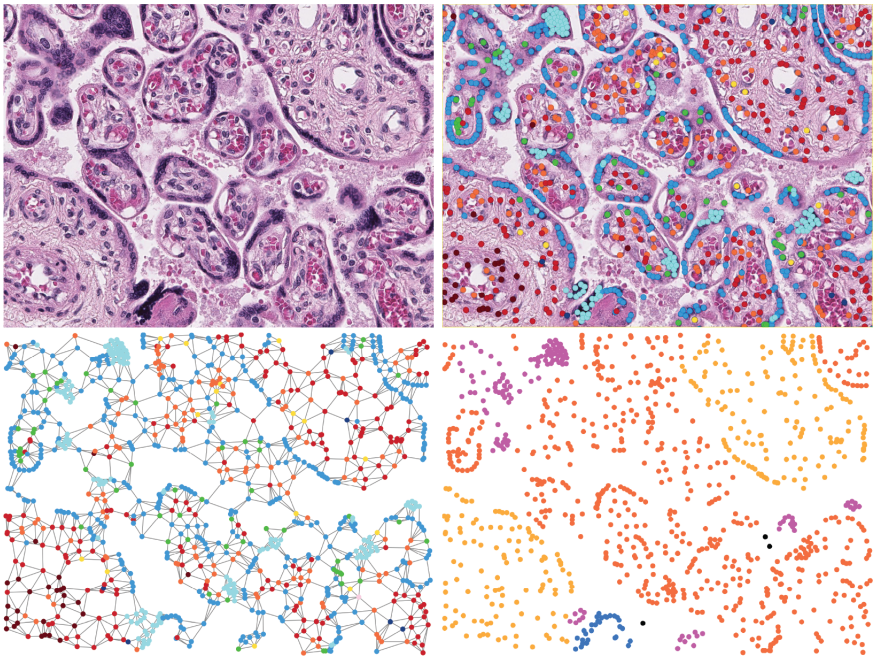

Figure 2: A 1600x1200 pixel region showing (top left) the raw histology image, (top right) overlaid cell nodes, (bottom left) the constructed intersection cell graph, and (bottom right) corresponding ground truth tissue nodes.

Hyperparameters Models are trained for inductive learning fully-batched to 1000 epochs with early stopping on validation data (evaluated every 100 epochs), weighted cross-entropy loss, Adam optimiser [15], and a 0.001 learning rate. Models have 256 hidden units per layer, a batch normalisation layer following each graph convolutional layer (except for ShaDow), and at most 16 graph convolutional layers each with a 0.5 dropout and relu activations. For GraphSage, ClusterGAT, and ClusterGATv2, 16 layer models could not fit into GPU memory even with reduced sampling and batch sizes and, as such, fewer layers were used. Further hyperparameter details can be found in the supplement (Appendix E).

Note that the goal is not to find the best performing architecture for this dataset, but instead to show the problems existing architectures have on histology cell graphs. As such, we explore hyperparameters to the extent that we believe a fair comparison can be made between models but we do not perform an exhaustive search. This has the added benefit of avoiding overfitting hyperparameters to these two placenta WSIs, which only represent a small facet of the histology domain.

Hardware, Software, and Libraries Experiments were run on one NVIDIA RTX8000 on an internal cluster. Code was written in Python (v3.7.2) using PyTorch (v1.9.0)[20] and PyTorch Geometric (v2.0.4)[8]. WSI visualisation and ground truth annotations were made using QuPath (v0.3.1)[21]. Code, cell graph dataset, data splits, and experiment instructions may be found at (`https://github.com/Nellaker-group/placenta`) under a MIT licence.

4 Results

Table 2 shows the test mean accuracies and standard deviations of models over 5 random weight initializations. The randomly initialised baseline performs equivalent to random chance (ROC AUC 0.5) and GNN models outperform the MLP baseline. We found ClusterGAT and ClusterGATv2 were unstable to train with more layers and this might contribute to their lower performance score. All scalable GNN architectures: GraphSAGE, ClusterGCN, GraphSAINT, ShaDow, and SIGN are within 2% mean accuracy of each other, with no models surpassing 65% accuracy. These results highlight the need for new GNN approaches for cell graphs in histology.

Table 2: Mean model performance and standard deviations over 5 random weight initialisation. **Bold** indicates the best performing model for that metric.

	Accuracy	Top 2 Accuracy	ROC AUC
Random Baseline	5.56±6.24	14.88±11.35	0.459±0.057
MLP Baseline	47.98±0.79	75.22±0.92	0.750±0.003
GraphSAGE-mean	**64.88±0.43**	**88.94±0.38**	0.883±0.005
ClusterGCN	64.24±1.21	88.26±0.82	0.882±0.006
GraphSAINT-rw	63.94±0.23	87.86±0.15	**0.895±0.002**
SIGN	64.77±0.43	88.32±0.42	0.886±0.002
ShaDow	63.04±0.77	86.88±0.74	0.863±0.008
ClusterGAT	58.07±0.61	83.43±0.96	0.851±0.002
ClusterGATv2	57.07±0.65	83.21±0.55	0.854±0.005

In the supplement, we describe features which the models struggled with and we take a fast, high-performing model (ClusterGCN) to explore some of these in more detail (Appendix F, G, H). We show how class imbalance affects performance, how model size leads to overfitting, and the importance of a large enough receptive field. We additionally compare model performance using all edge-building algorithms (Appendix D), using one hot encoded cell type node features (Appendix I), and using random node data splits (Appendix C).

5 Limitations

Our cell graph dataset represents a small facet of the histology domain; two placenta WSIs from two institutes. WSIs from more placentas and other laboratories will be needed to validate the generalisability of models designed using this data. Likewise, the dataset is placenta-specific, so whilst there is a benefit to using the weights of placenta models for pretraining, the development of GNN tissue classifiers for other organs will require organ-specific cell and tissue ground truth data. Cross- or multi-organ self-supervised approaches could be developed from our dataset but it is not something we have explored. Regarding model comparison, we have purposely limited the scope of our hyperparameter search. It is possible that a different combination of hyperparameters and/or additional sampling, normalisation, and regularisation, would have yielded better performance. We release all data necessary for building cell graphs across our WSIs but not the raw images or pipeline needed to derive the cell features from scratch as this was out of scope for this work. Finally, it should be noted that uncertainty in ground truth annotations may limit model performance on the dataset, however label uncertainty is inherent to the biological problem and unavoidable in this domain.

6 Conclusion

We introduce a new inductive node classification benchmark dataset, classifying microanatomical tissue structures from cell graphs across placenta histology WSIs. We show how GNN architectures perform on this task and highlight the qualities unique to this data domain which present a challenge for existing models. We hope that this dataset facilitates the creation of new model architectures to overcome these challenges.

Acknowledgments and Disclosure of Funding

Claudia Vanea and Jonathan Campbell are supported by the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1). Cecilia M. Lindgren is supported by the Li Ka Shing Foundation, NIHR Oxford Biomedical Research Centre, Oxford, NIH (1P50HD104224-01), Gates Foundation (INV-024200), and a Wellcome Trust Investigator Award (221782/Z/20/Z). Triin Laisk is funded by the European Regional Development Fund and the programme Mobilitas Pluss (MOBTP155) and the Estonian Research Council grant PSG776. The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

[1] K. Benirschke, G. J. Burton, and R. N. Baergen. Architecture of Normal Villous Trees. In K. Benirschke, G. J. Burton, and R. N. Baergen, editors, *Pathology of the Human Placenta*, pages 101–144. Springer, Berlin, Heidelberg, 2012. ISBN 978-3-642-23941-0. doi: 10.1007/978-3-642-23941-0_7. URL https://doi.org/10.1007/978-3-642-23941-0_7.

[2] S. Brody, U. Alon, and E. Yahav. How Attentive are Graph Attention Networks?, Jan. 2022. URL http://arxiv.org/abs/2105.14491. arXiv:2105.14491 [cs].

[3] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr, and T. M. Mitchell. Toward an Architecture for Never-Ending Language Learning. *AAAI Press*, page 8, 2010.

[4] R. J. Chen, M. Y. Lu, J. Wang, D. F. K. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *IEEE transactions on medical imaging*, PP, Sept. 2020. ISSN 1558-254X. doi: 10.1109/TMI.2020.3021387.

[5] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh. Cluster-GCN: An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks. In *KDD'19*, May 2019. doi: 10.1145/3292500.3330925. URL https://arxiv.org/abs/1905.07953v2.

[6] D. Eppstein, M. S. Paterson, and F. F. Yao. On Nearest-Neighbor Graphs. *Discrete & Computational Geometry*, 17(3):263–282, Apr. 1997. ISSN 1432-0444. doi: 10.1007/PL00009293. URL https://doi.org/10.1007/PL00009293.

[7] L. M. Ernst and C. K. Carreon. Placenta. In L. M. Ernst, E. D. Ruchelli, C. K. Carreon, and D. S. Huff, editors, *Color Atlas of Human Fetal and Neonatal Histology*, pages 399–424. Springer International Publishing, Cham, 2019. ISBN 978-3-030-11425-1. doi: 10.1007/978-3-030-11425-1_36. URL https://doi.org/10.1007/978-3-030-11425-1_36.

[8] M. Fey and J. E. Lenssen. Fast Graph Representation Learning with PyTorch Geometric. *arXiv:1903.02428 [cs, stat]*, Apr. 2019. URL http://arxiv.org/abs/1903.02428. arXiv: 1903.02428.

[9] F. Frasca, E. Rossi, D. Eynard, B. Chamberlain, M. Bronstein, and F. Monti. SIGN: Scalable Inception Graph Neural Networks. In *ICML 2020*, Apr. 2020. URL https://arxiv.org/abs/2004.11198v3.

[10] S. Gadiya, D. Anand, and A. Sethi. Histographs: Graphs in Histopathology. *arXiv:1908.05020 [cs, eess]*, Aug. 2019. URL http://arxiv.org/abs/1908.05020. arXiv: 1908.05020.

[11] L. J. Guibas, D. E. Knuth, and M. Sharir. Randomized incremental construction of Delaunay and Voronoi diagrams. *Algorithmica*, 7(1):381–413, June 1992. ISSN 1432-0541. doi: 10.1007/BF01758770. URL https://doi.org/10.1007/BF01758770.

[12] W. L. Hamilton, R. Ying, and J. Leskovec. Inductive Representation Learning on Large Graphs. June 2017. URL https://arxiv.org/abs/1706.02216v4.

[13] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *arXiv:2005.00687 [cs, stat]*, Feb. 2021. URL http://arxiv.org/abs/2005.00687. arXiv: 2005.00687.

[14] S. Javed, A. Mahmood, M. M. Fraz, N. A. Koohbanani, K. Benes, Y.-W. Tsang, K. Hewitt, D. Epstein, D. Snead, and N. Rajpoot. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Medical Image Analysis*, 63:101696, July 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101696. URL https://www.sciencedirect.com/science/article/pii/S136184152030061X.

[15] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. Dec. 2014. URL https://arxiv.org/abs/1412.6980v9.

[16] T. N. Kipf and M. Welling. Semi-Supervised Classification with Graph Convolutional Networks, Feb. 2017. URL http://arxiv.org/abs/1609.02907. arXiv:1609.02907 [cs, stat].

[17] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres. Quantifying the Carbon Emissions of Machine Learning, Nov. 2019. URL http://arxiv.org/abs/1910.09700. arXiv:1910.09700 [cs].

[18] W. Lu, S. Graham, M. Bilal, N. Rajpoot, and F. Minhas. Capturing Cellular Topology in Multi-Gigapixel Pathology Images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1049–1058, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72819-360-1. doi: 10.1109/CVPRW50498.2020.00138. URL https://ieeexplore.ieee.org/document/9150693/.

[19] W. Lu, M. Toss, M. Dawood, E. Rakha, N. Rajpoot, and F. Minhas. SlideGraph+: Whole slide image level graphs to predict HER2 status in breast cancer. *Medical Image Analysis*, 80:102486, Aug. 2022. ISSN 1361-8415. doi: 10.1016/j.media.2022.102486. URL https://www.sciencedirect.com/science/article/pii/S1361841522001335.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, Dec. 2019. URL http://arxiv.org/abs/1912.01703. arXiv:1912.01703 [cs, stat].

[21] Peter Bankhead. QuPath: Open source software for digital pathology image analysis | Scientific Reports, 2017. URL https://www.nature.com/articles/s41598-017-17204-5.

[22] O. Shchur, M. Mumme, A. Bojchevski, and S. Günnemann. Pitfalls of Graph Neural Network Evaluation. *arXiv:1811.05868 [cs, stat]*, June 2019. URL http://arxiv.org/abs/1811.05868. arXiv: 1811.05868.

[23] K. Sirinukunwattana, D. Snead, D. Epstein, Z. Aftab, I. Mujeeb, Y. W. Tsang, I. Cree, and N. Rajpoot. Novel digital signatures of tissue phenotypes for predicting distant metastasis in colorectal cancer. *Scientific Reports*, 8(1):13692, Sept. 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-31799-3. URL https://www.nature.com/articles/s41598-018-31799-3.

[24] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. Mering. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, Jan. 2019. ISSN 0305-1048. doi: 10.1093/nar/gky1131. URL https://doi.org/10.1093/nar/gky1131.

[25] C. Vanea, J. Džigurski, V. Rukins, O. Dodi, S. Siigur, L. Salumäe, K. Meir, W. T. Parks, D. Hochner-Celnikier, A. Fraser, H. Hochner, T. Laisk, L. M. Ernst, C. M. Lindgren, and C. Nellåker. A pipeline for whole slide placenta histology analysis building cell-to-tissue graphs using deep learning. *In Preparation*.

[26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph Attention Networks, Feb. 2018. URL http://arxiv.org/abs/1710.10903. arXiv:1710.10903 [cs, stat].

[27] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm. Deep Graph Infomax, Dec. 2018. URL http://arxiv.org/abs/1809.10341. arXiv:1809.10341 [cs, math, stat].

[28] F. Wu, T. Zhang, A. H. d. Souza Jr., C. Fifty, T. Yu, and K. Q. Weinberger. Simplifying Graph Convolutional Networks, June 2019. URL http://arxiv.org/abs/1902.07153. arXiv:1902.07153 [cs, stat].

[29] Z. Yang, W. W. Cohen, and R. Salakhutdinov. Revisiting Semi-Supervised Learning with Graph Embeddings, May 2016. URL http://arxiv.org/abs/1603.08861. arXiv:1603.08861 [cs].

[30] B. Yener. Cell-Graphs: Image-Driven Modeling of Structure-Function Relationship. *Communications of the ACM*, 60(1):74–84, 2017. URL https://cacm.acm.org/magazines/2017/1/211111-cell-graphs/fulltext.

[31] H. Zeng, H. Zhou, A. Srivastava, R. Kannan, and V. Prasanna. GraphSAINT: Graph Sampling Based Inductive Learning Method. In *arXiv:1907.04931 [cs, stat]*, Feb. 2020. URL http://arxiv.org/abs/1907.04931. arXiv: 1907.04931.

[32] H. Zeng, M. Zhang, Y. Xia, A. Srivastava, A. Malevich, R. Kannan, V. Prasanna, L. Jin, and R. Chen. Decoupling the Depth and Scope of Graph Neural Networks, Jan. 2022. URL http://arxiv.org/abs/2201.07858. arXiv:2201.07858 [cs].

[33] Y. Zhou, S. Graham, N. A. Koohbanani, M. Shaban, P.-A. Heng, and N. Rajpoot. CGC-Net: Cell Graph Convolutional Network for Grading of Colorectal Cancer Histology Images. *arXiv:1909.01068 [cs, eess]*, Sept. 2019. URL http://arxiv.org/abs/1909.01068. arXiv: 1909.01068.

[34] M. Zitnik and J. Leskovec. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14):i190–i198, July 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx252. URL https://doi.org/10.1093/bioinformatics/btx252.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We claim to introduce a new dataset benchmark in histology cell graphs and show how existing GNN models perform on this data.
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix K.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 3 and at (`https://github.com/Nellaker-group/placenta`).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 2, 3 and Appendix C, E.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 4. Experiments were run with 5 random seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 3 and Appendix L.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We existing GNN models and code from PyTorch Geometric and we cite the creators
   (b) Did you mention the license of the assets? [Yes] See Section 3.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 2 and at (`https://github.com/Nellaker-group/placenta`).
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 2. Data was obtained without consent with approval from local ethics committees for research use.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 2 and 5.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Placenta Cell and Tissue Types

11 placenta cells used in node features:

- SYN: Syncytiotrophoblast
- CYT: Cytotrophoblast
- KNT: Syncytial Knot
- EVT: Extravillus Trophoblast
- FIB: Fibroblast
- HOF: Hofbauer Cell
- VEN: Vascular Endothelial
- VMY: Vascular Myocyte
- MES: Mesenchymal Cell
- MAT: Maternal Decidua
- WBC: Leukocyte

9 placenta tissue types used as node labels:

- Sprout: Villus Sprout
- TVilli: Terminal Villi
- MIVilli: Mature Intermediate Villi
- SVilli: Stem Villi
- AVilli: Anchoring Villi
- Chorion: Chorionic Plate
- Maternal: Basal Plate and Septum
- Fibrin: Fibrin
- Avacular: Avascular Villi

B Cell Types in Node Features and Ground Truth Tissue Distributions

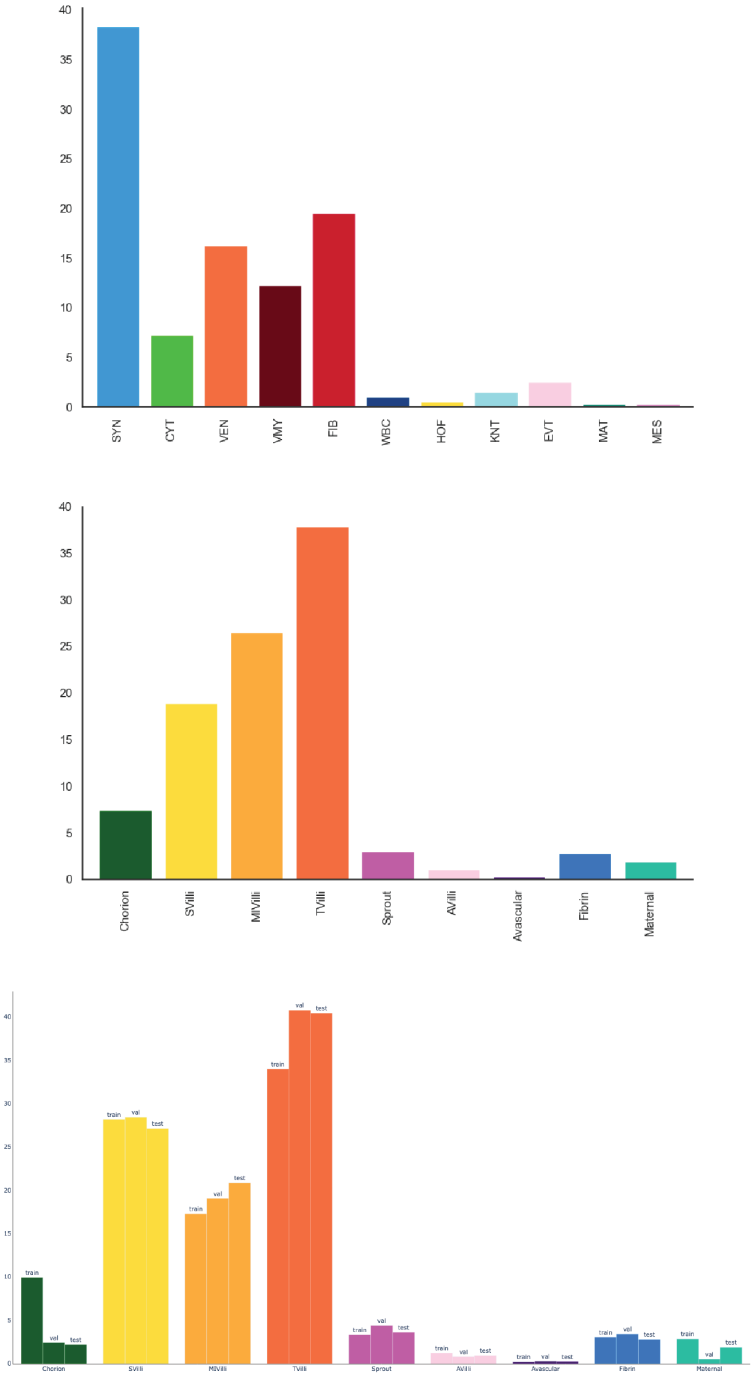

Figure 3: Cell type proportions in node features (top), tissue label proportions in all labelled data (middle), and distribution of tissue classes within dataset splits (bottom).

C Data Splits and the Effect of Random Node Splits

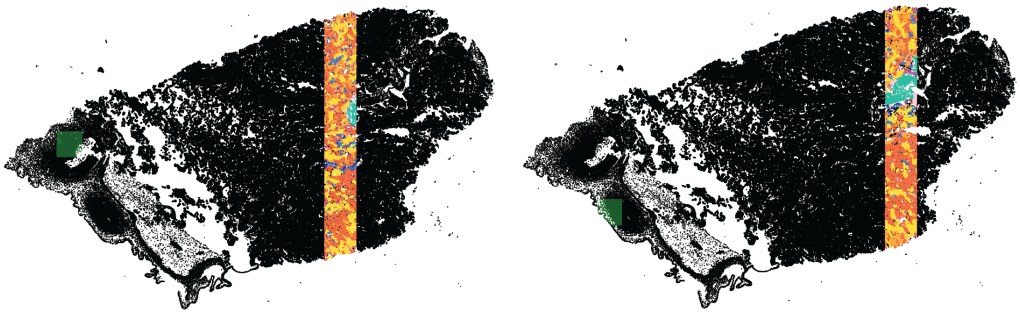

Figure 4: Validation region (left) and test region (right).

Table 3: Distribution of class labels across dataset splits.

Class Label	Training	Validation	Test
Sprout	15,545 (50%)	9,066 (29%)	6,441 (21%)
TVilli	159,464 (50%)	84,052 (27%)	72,402 (23%)
MIVilli	80,954 (51%)	39,296 (25%)	37,332 (24%)
SVilli	132,138 (55%)	58,658 (25%)	48,568 (20%)
AVilli	5,769 (64%)	1,634 (18%)	1,654 (18%)
Chorion	46,584 (84%)	4,942 (9%)	3,937 (7%)
Maternal	13,329 (75%)	1,029 (6%)	3,359 (19%)
Fibrin	14,195 (54%)	7,047 (27%)	5,010 (19%)
Avacular	891 (49%)	526 (29%)	392 (22%)

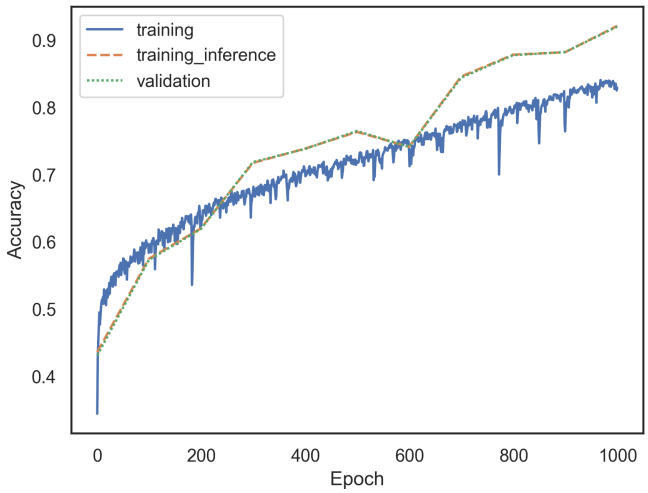

Figure 5: Accuracy curves during training using random node splits of 70%/15%/15%. Validation performance is identical to training performance using the same inference sampling method, showing information leakage across splits.

D Edge Construction and Graph Differences

Table 4: Graph Characteristics of KNN, Delaunay, and intersection graphs.

	Nodes	Edges	Labels	Homophily	Degree
Intersection Graph 1	1,159,186	2,603,797	703,165	0.9868	5.4925
Intersection Graph 2	1,236,561	2,882,292	96,580	0.9984	5.6618
KNN Graph 1	1,159,186	8,398,932	703,165	0.9854	8.2455
KNN Graph 2	1,236,561	8,870,066	96,580	0.7286	8.1732
Delaunay Graph 1	1,159,186	3,477,526	703,165	0.9522	6.9999
Delaunay Graph 2	1,236,561	3,709,652	96,580	0.7217	6.9999

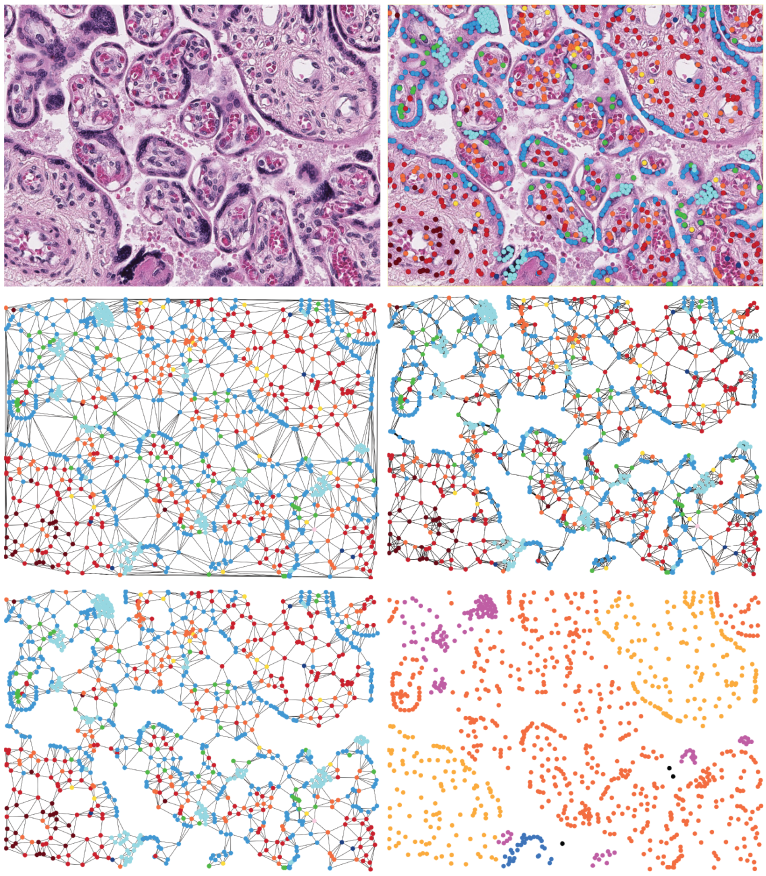

Figure 6: From top left to bottom right, the raw histology, histology with cell overlay, Delaunay Triangulation edges, KNN edges, intersection edges, and ground truth tissue labels.

Table 5: ClusterGCN performance with a KNN graph where k=5 and a Delaunay graph.

	Accuracy	Top 2 Accuracy	ROC AUC
KNN	0.654	0.894	0.903
Delaunay	0.676	0.898	0.916

E Hyperparameter Details for Each Model Architecture

Weighted cross entropy weights: [1, 0.85, 0.9, 10.5, 0.8, 1.3, 5.6, 3, 77]

Additional hyperparameter details and any changes from the default hyperparameters reported in Section 3 are reported below.

MLP 16 Linear layers, 51,200 batch size.

GraphSAGE-mean 12 SAGEConv layers due to GPU out of memory error with larger models, 32,000 batch size, 10 neighbours per hop, mean aggregation.

ClusterGCN 16 SAGEConv layers, 200 batch size, and 400 max subgraph size.

GraphSAINT-rw 16 GraphConv layers, in-node normalisation, 32,000 batch size, 30 steps, 500 sample coverage, random walk sampler.

SIGN 16 Linear layers, 51,200 batch size, SIGN transform with 16 hops.

ShaDow 8 SAGEConv layers due to performance drop with larger models, no batch norm layers for the same reason, 4000 batch size, 5 neighbours per hop, hop depth of 6. This smaller receptive field is due to a RAM memory error with a higher depth or more neighbours.

ClusterGAT 2 GATConv layers due to GPU out of memory error and performance drop with larger models, 400 batch size, 400 max subgraph size, 4 attention heads with 0.25 dropout.

ClusterGATv2 2 GATv2Conv layers due to GPU out of memory error and performance drop with larger models, 200 batch size, 400 max subgraph size, 4 attention heads with 0.25 dropout.

F The Effect of Class imbalance on Performance

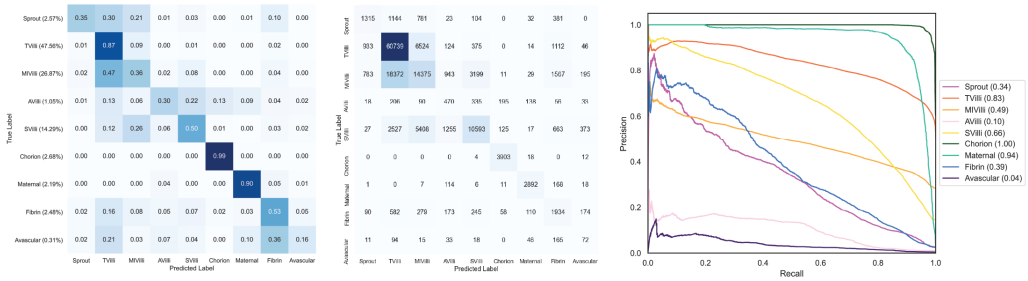

Figure 7: Confusion matrix with values proportional to class imbalance (left), confusion matrix per number of nodes (middle), and precision-recall curves with corresponding AUC-PR values (right). The dataset is highly imbalanced and models perform worse on many minority classes.

G The Effect of Model Size on Overfitting to Training Data

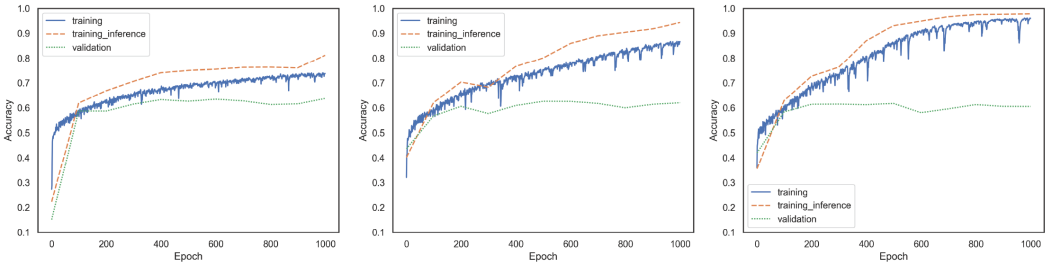

Figure 8: ClusterGCN with 128 hidden units (left), 256 hidden units (middle), and 512 hidden units (right). Too few hidden units and the model does not learn the training data, too many and the model overfits.

H The Effect of Receptive Field Sizes

Table 6: ClusterGCN performance with varied receptive fields. Receptive fields are varied by changing cluster subgraph sizes and the number of model layers.

	Accuracy	Top 2 Accuracy	ROC AUC
800 Subgraph 32 Layers	0.647	0.874	0.893
400 Subgraph 16 Layers	0.642	0.883	0.882
200 Subgraph 8 Layers	0.623	0.871	0.876
100 Subgraph 4 Layers	0.594	0.849	0.856

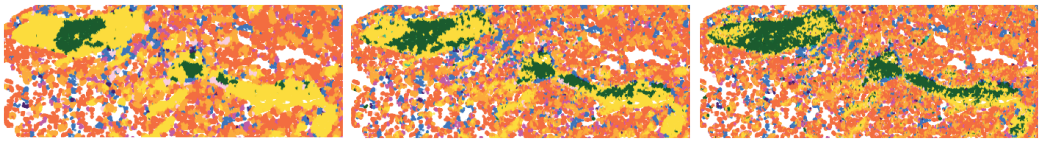

Figure 9: Reducing receptive field sizes (left to right) leads to confused predictions inside large structures. The yellow tissue structure (SVilli) in these images is incorrectly labelled as the green tissue structure (Chorion).

I The Effect of Cellular Node Features

Table 7: ClusterGCN performance when trained and evaluated with one hot encoded cell types. Performance is marginally worse than training with 64-dim feature vectors.

	Accuracy	Top 2 Accuracy	ROC AUC
One Hot Cell Types	0.628	0.871	0.833

J Visualisation of Model Predictions

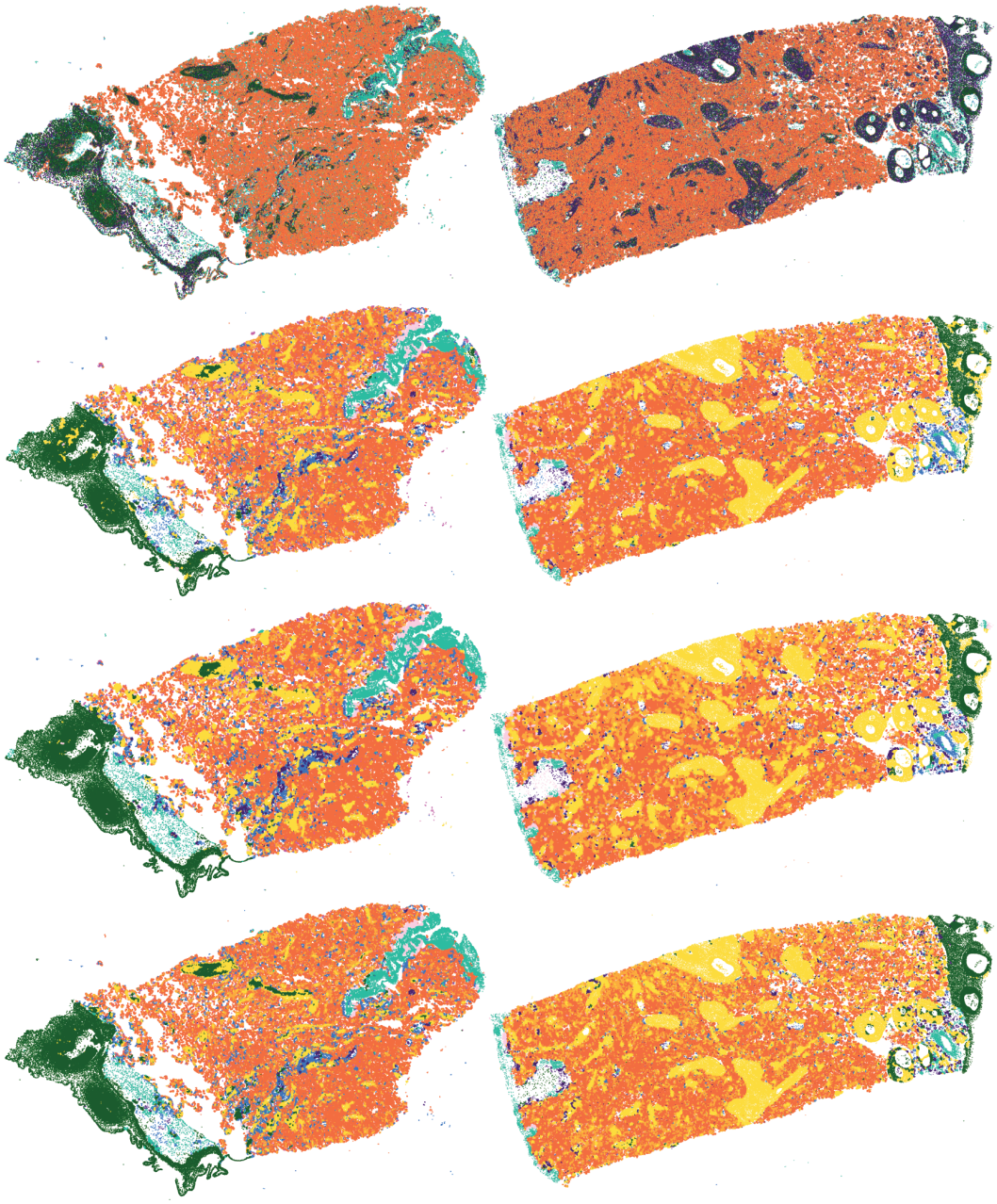

Figure 10: From top to bottom, MLP, GraphSAGE-mean, ClusterGCN, and GraphSAINT-rw tissue predictions across both WSIs.

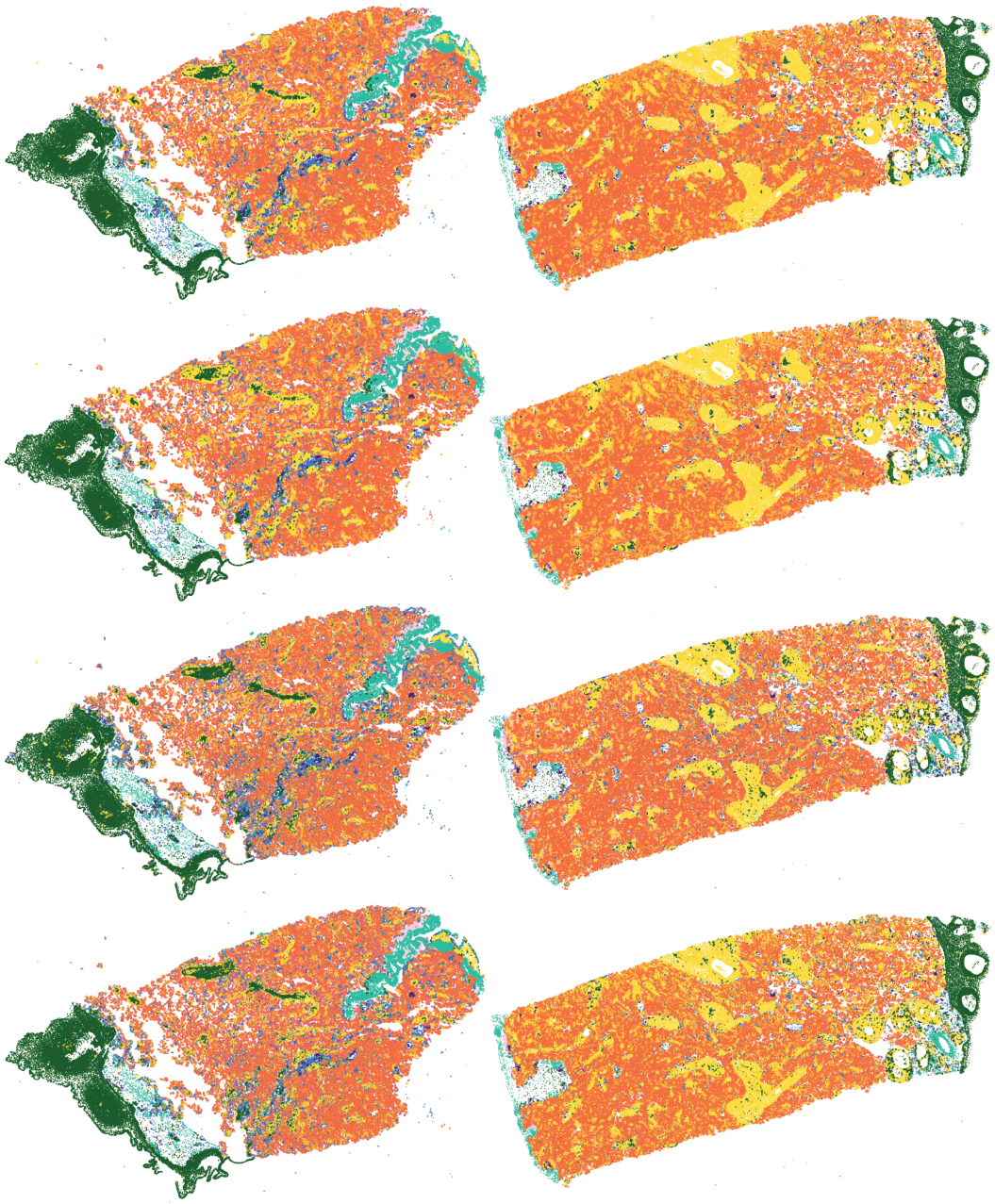

Figure 11: From top to bottom, SIGN, ShaDow, ClusterGAT, and ClusterGATv2 tissue predictions across both WSIs.

K Potential Negative Societal Impacts

The development of similar models in the context of histological analyses could, if in future fully realised and implemented across healthcare service settings, negatively impact the work of pathologists who do not have access to the models: "AI won't replace radiologists, but radiologists who use AI will replace radiologists who don't," (Curtis Langlotz, 2019, Nature, Innovations In). As with the uptake of any new technology, there would be a 'growing pains' period which could increase the time between tissue sample collection and diagnosis. Slides will have to be digitised and model inference will have to be integrated, causing a shift in existing clinical and histological preparation workflows. We must also consider that any models developed for histology analysis could cause serious patient harm if, in the extreme case, they are incorrectly validated and quality controlled, or they have an undetected bias which influences diagnoses. If we assume models perform as well or better than clinicians, but their outputs are not clearly expressed in clinical language and/or there is a lack of trust in these automated, uninterpretable methods, then there may be resulting uncertainty in diagnosis and consequently higher referral rates. Conversely, if automated histology analysis becomes popular, for organs unlike the placenta which require biopsies to be sampled, the rate of biopsies could harmfully increase. Finally, the publication of this dataset could facilitate the development of new GNN architectures that themselves could confer negative impacts in other domains.

L CO2 Emission Related to Experiments

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 $kgCO_2eq/kWh$. An approximate cumulative of 200 hours of computation was performed on hardware of type RTX 8000 (TDP of 260W).

Total emissions are estimated to be 22.46 $kgCO_2eq$ of which 0 percents were directly offset.

Estimations were conducted using the MachineLearning Impact calculator presented in [17].