xCG: Explainable Cell Graphs for Survival Prediction in Non-Small Cell Lung Cancer

Abstract

Understanding how deep learning models predict oncology patient risk can provide critical insights into disease progression, support clinical decision-making, and pave the way for trustworthy and data-driven precision medicine. Building on recent advances in the spatial modeling of the tumor microenvironment using graph neural networks, we present an explainable cell graph (xCG) approach for survival prediction. We validate our model on a public cohort of imaging mass cytometry (IMC) data for 416 cases of lung adenocarcinoma. We explain survival predictions in terms of known phenotypes on the cell level by computing risk attributions over cell graphs, for which we propose an efficient grid-based layer-wise relevance propagation (LRP) method. Our ablation studies highlight the importance of incorporating the cancer stage and model ensembling to improve the quality of risk estimates. Our xCG method, together with the IMC data, is made publicly available to support further research.

Keywords: Cell Graphs, Explainable AI, Graph Neural Networks, Survival Analysis

Data and Code Availability We publish our PyTorch implementation of xCG at https://anonymous.4open.science/r/explainable-cell-graphs. Our method is validated on a publicly available[1] data cohort (Sorin et al., 2023).

Institutional Review Board (IRB) Relevant ethics approval information will be provided if the paper is accepted.

1. Introduction

Lung cancer remains the leading cause of cancer-related death, accounting for over 20% of all cancer cases (Siegel et al., 2021). To improve patient outcomes, it is crucial to further advance our understanding of the disease mechanisms and identify more precise risk factors.

1. https://doi.org/10.5281/zenodo.7760826

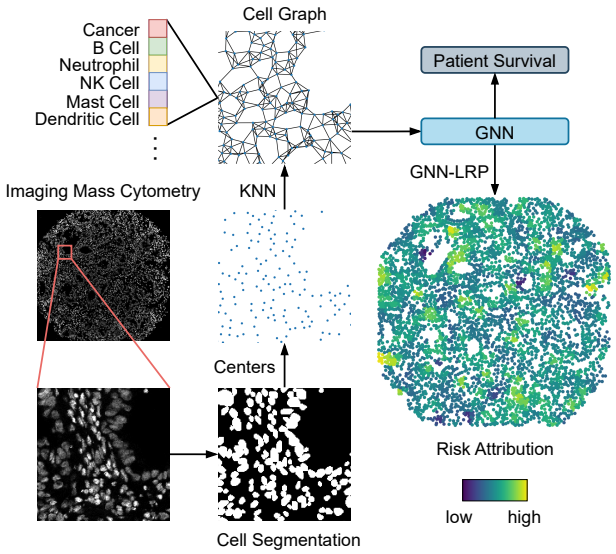


Figure 1: Overview of the xCG framework, including preprocessing steps, GNN for survival prediction, and GNN-LRP risk attribution.

Established risk factors such as the UICC8's TNM classification of malignant tumors (Brierley et al., 2017) represent the standard of care but lack granularity for personalized treatment decisions, e.g. not covering immune system, molecular, or metabolic parameters, often resulting in unnecessary side effects and rendering treatment insufficient.

Advances in spatially resolved single-cell technologies now allow us to explore the tumor microenvironment (TME) in unprecedented detail (Sorin et al., 2023). Leveraging these technologies, graph neural networks (GNNs) have shown promise in modeling the TME in several cancer types, including lung cancer (Zhou et al., 2019; Wang et al., 2022; Nakhli et al., 2023; Zhang et al., 2024). However, while recent studies have begun to incorporate explainability into graph-based models (Sureka et al., 2020; Jaume

et al., 2020; Hu et al., 2024; Zhang et al., 2024), explaining risk factors remains a significant challenge in the medical field due to the large scale of relevant graphs.

In this paper, we present the following key contributions to address this challenge:

- **Modality-Agnostic Survival Prediction**: We propose a versatile GNN framework that can (a) handle multiple tissue samples and graphs per patient, (b) incorporate multiple cell-level feature domains such as marker expression, tumor region-segmentation and patient-level clinical metadata, and (c) is capable of survival regression and classification. Our implementation in PyTorch is publicly available.
- **Scalable XAI for Cell Graphs**: We introduce a novel efficient grid-based GNN-LRP method for cell graphs that enables high-resolution risk attribution at the cell level.
- **Enhanced Risk Assessment**: Our ablation studies show that combining cancer stage fusion and model ensembling significantly improves the accuracy and reliability of risk assessments.

2. Methods

2.1. Data and Preprocessing

Data We use a dataset published in Sorin et al. (2023) consisting of single tissue spots of 1.0 mm^2, obtained from 416 patients with adenocarcinoma of the lung. Spots are stained with a 35-plex imaging mass cytometry (IMC) panel, from which 17 distinct cell phenotypes were derived.

The cohort includes clinical metadata like overall survival and the UICC8 cancer stage, categorizing patients into early (I-II) and late (III-IV) stages. For survival classification, we adapt the two categories proposed by Sorin et al. (2023): short-term ($\leq$ 36 months) and long-term ($>$ 36 months) survival. Patients with a survival time of less than 36 months but without registered death events were excluded, as they could have died in either period.

Preprocessing As shown in Figure 1, the centers of mass of the cell segmentation masks are used to determine cell positions. For each tissue sample, we construct a cell graph where each cell is represented as a node. Each node is characterized by a one-hot encoded vector representing the cell phenotype. Biologically resembling mutual interactions between proximal cells, edges are established by k-nearest neighbors (KNN) fit with $k = 3$.

2.2. Models and Training

Survival Regression Our proposed cell graph encoder architecture (Figure 3) builds on the sparse hierarchical graph classifier framework presented in Cangea et al. (2018). To enable this architecture to input multiple tissue samples per patient, we incorporate attention-based MIL pooling as described in Ilse et al. (2018), so the input becomes a set of cell graphs, $\mathcal{X} = \{\mathcal{G}_1, ..., \mathcal{G}_K\}$. We made this generalization of the architecture since the use case of multiple spots per patient is often given, as was the case for a proprietary study cohort of ours.

In the first stage, the model computes an embedding for each graph individually $\mathcal{H} = \{\boldsymbol{h}_1, ..., \boldsymbol{h}_K\}$ by alternating between Graph Isomorphism Network (GIN) layers (Xu et al., 2019) and top-k pooling layers (Cangea et al., 2018). To integrate clinical metadata for enhanced risk assessment, we fuse the UICC8 cancer stage to the cell graph representations by addition. In the second stage, to produce a patient-level representation $\boldsymbol{h}_{\mathcal{X}}$, attention-based MIL-pooling is applied, which allows the model to prioritize the most relevant graphs when creating the overall patient representation. Finally, the aggregated patient representation is passed through a feed-forward network that is responsible for predicting survival risk.

Survival Classification To further simplify the explainability of the model (see Section 2.3), we also reformulate the problem as a binary survival classification task and distinguish between short-term ($\leq$ 36 months) and long-term ($>$ 36 months) survival.

To implement this classification task, we use a 3-layer graph isomorphism network (GIN) without graph pooling, adapted from the implementation by Schnake et al. (2022). Instead of using attention-based multiple instance learning (MIL), we average the GIN logits across the multiple cell graphs per patient to simplify relevance propagation. Future work may consider implementing xMIL-LRP (Hense et al., 2024). Additionally, we omit the cancer stage fusion for our explainability analysis, as its relevance would be non-localized and not contribute to the spatial explanation.

Training We train both our survival regression and classification models using stratified nested cross-

validation over five folds, training for 50 epochs. Hyperparameter optimization is performed on inner validation folds, with the learning rate chosen from $\gamma \in \{5e{-}5, 1e{-}5, 5e{-}6\}$. To improve the robustness of our risk predictions, particularly towards variability caused by different model initialization, we ensemble the risk predictions of five separate survival regression models. Each model is trained with a different random seed, and we compute the final prediction by taking the mean of the outputs.

2.3. Explainability

Ranking-based survival models, such as Cox regression, are widely used in medical research to predict patient outcomes (Jóźwiak et al., 2024). However, interpreting how these rank-based models make predictions is difficult, due to their implicit handling of survival times.

To illustrate, layer-wise relevance propagation (LRP) propagates the predicted risk score backwards through the network, assigning relevance (attributions) to input features. One challenge is that these attributions can be both positive and negative, depending on how we define the 'zero point' in the model's target range. As Letzgus et al. (2022) discusses, the point we set as the baseline (or zero point) determines whether a feature is seen as increasing or decreasing the predicted risk.

Furthermore, the magnitude of these attributions changes with the scale of the target range. Without a clear definition of both the baseline and the scale, it becomes difficult to interpret the model's explanations. This is particularly problematic for ranking-based models, as they are trained using an implicit ranking loss that does not allow us to set an explicit reference point and scale. Therefore, we only generate explanations for the survival classification model discussed in Section 2.2, where these reference points are made explicit. For more details on interpreting model explantions see Section B.2

We use sGNN-LRP (Xiong et al., 2022), an optimized variant of GNN-LRP (Schnake et al., 2022) tailored to subgraph attribution. This method reduces the computational complexity from exponential $\mathcal{O}(|\mathcal{S}|^L)$ to linear $\mathcal{O}(L|\mathcal{S}|^2)$, where L is the number of layers in the network and the number of nodes in the subgraph. While the graphs considered in Xiong et al. (2022) only span at most hundreds of nodes, our cell graphs can reach tens of thousands of nodes. We significantly reduce the memory requirements by exploiting the sparse connectivity of the KNN-based graph adjacency matrix and rewrite sGNN-LRP with sparse matrix multiplications using PyTorch Sparse[2].

To further reduce memory and compute costs of high-resolution risk attributions, we split the global attribution task into local subtasks utilizing a shifted-grid approximation approach. The cell graph is partitioned into a square grid and the subgraph relevance is calculated for each tile and normalized by the number of cells, mitigating over-weighting of densely populated areas. The grid is then repeatedly shifted in the x- and y-directions by a stride s, such that the average over strides results in a smooth heatmap of cell-level risk attributions.

3. Results and Discussion

3.1. Survival Regression

We evaluate the performance of our cell graph survival regression model by reporting the C-index (Harrell et al., 1982), a metric of a model's ability to correctly rank survival times, for three different scenarios: the unmodified standard UICC8 clinical baseline, our GNN-based survival regression model, and the mean risk ensemble of our model, each averaged over five test folds. The results presented in Table 1 show an improvement in C-index from 0.568 to 0.593 when using our GNN ensemble, highlighting the model's ability to capture complex interactions within the TME and effectively advancing the clinical baseline. The improved performance of the ensemble over the single GIN model shows that model ensembling can mitigate the variance found in individual model predictions and produce more robust, generalizing survival estimates.

Removing cancer-stage fusion leads to a significant decrease in C-index. This result highlights that the integration of clinical staging information remains a crucial context for accurate survival prediction.

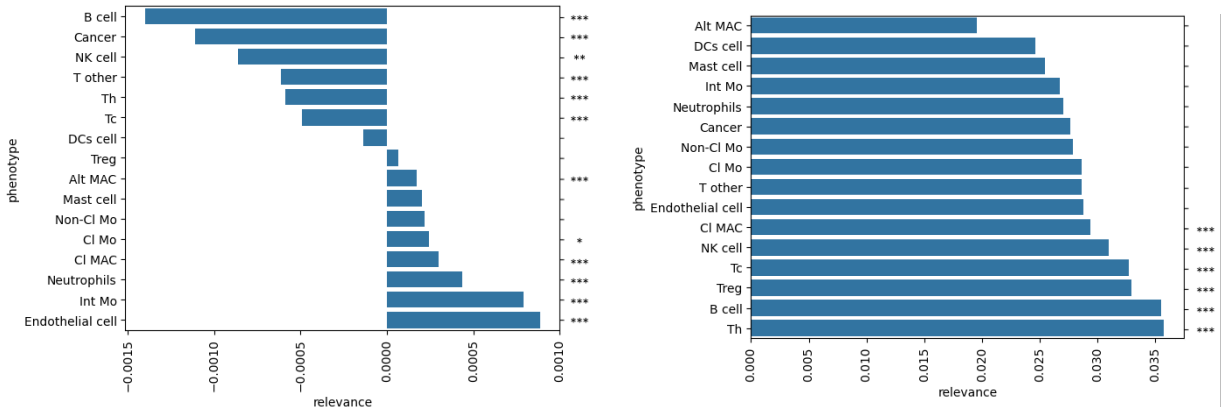
3.2. Survival Classification

We train a survival classification model to generate risk attribution heatmaps for cell graphs following the methodology described in Section 2.3. Our model was trained on five different seeds and evaluated across five test folds, resulting in a binary AUROC of 0.700 $\pm$ 0.028 without cancer stage fusion.

2. https://github.com/rusty1s/pytorch_sparse

Table 1: Ablation of UICC8 cancer stage fusion and model ensembling for our survival regression model. We report the C-index and its standard deviation over five seeds, averaged over five test folds.

UICC8 Fusion	UICC8 Baseline	GIN	Ensemble
✓	0.568 ± 0.005	0.559 ± 0.019	0.593 ± 0.033
-	-	0.507 ± 0.022	0.518 ± 0.043



(*a*) Short survival.

(*b*) Long survival.

Figure 2: Median LRP relevance over cases, localized by cell phenotype for (*a*) short ($\leq$ 36 months) and (*b*) long ($>$ 36 months) survival. Phenotypes are sorted by their median LRP relevance. We perform a permuation test over 1,000 iterations (*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$).

Plotting the risk attributions computed via our grid-based GNN-LRP approach on a per-cell basis, we achieve high-resolution heatmaps as depicted in Figure 6. We visualize min-max normalized relevance and compare risk attribution heatmaps between spots with short and long survival. We observe strongly localized regions of high (yellow) and low (blue) risk attribution, which can be traced back to known cell phenotypes.

Tying Risk Attributions to Disease Biology We calculate the median relevance attributed to the cells of a phenotype for long and short survivors separately and perform a permutation test over 1,000 iterations. The comparison of the median LRP relevance across cases (Figure 2) according to cell phenotype reveals a notable difference between long and short survivors. Among the long survivors, several immune cell phenotypes such as T helper cells, B cells, regulatory T cells, or cytotoxic T cells are particularly relevant ($p < 0.001$), aligning with the findings of previous studies in the literature (Debatin et al., 2024; Denkert et al., 2018; Galon et al., 2006; Laumont et al., 2022; Wieland et al., 2021; Hu et al., 2021). In contrast, the particular relevance of neutrophils in short-term survivors ($p < 0.001$) is consistent with the established association between increased neutrophil counts and poorer prognosis in NSCLC patients (Ilie et al., 2012).

4. Conclusion

In this work, we presented a framework for explaining large-scale cell graphs using high-resolution cell-level risk attributions. We further showed that cancer stage fusion and model ensembling can improve survival prediction based on cell graphs. Moreover, we see an indication that cell graphs are capable of learning spatial TME features that are beyond the classical understanding of what is relevant for prognosis in the clinic.

References

James D Brierley, Mary K Gospodarowicz, and Christian Wittekind. *TNM classification of malignant tumours*. John Wiley & Sons, 2017.

Catalina Cangea, Petar Veličković, Nikola Jovanovic, Thomas Kipf, and Pietro Liò. Towards Sparse Hierarchical Graph Classifiers. *arXiv preprint arXiv:1811.01287*, 2018. URL https://arxiv.org/abs/1811.01287.

Nicolaus F Debatin, Elena Bady, Tim Mandelkow, Zhihao Huang, Magalie CJ Lurati, Jonas B Raedler, Jan H Müller, Eik Vettorazzi, Henning Plage, Henrik Samtleben, et al. Prognostic Impact and Spatial Interplay of Immune Cells in Urothelial Cancer. *European Urology*, 2024.

Carsten Denkert, Gunter von Minckwitz, Silvia Darb-Esfahani, Bianca Lederer, Barbara I Heppner, Karsten E Weber, Jan Budczies, Jens Huober, Frederick Klauschen, Jenny Furlanetto, et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *The lancet oncology*, 19(1):40–50, 2018.

Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric. In *International Conference on Learning Representations (ICLR), RLGM Workshop*, 2019. URL https://arxiv.org/abs/1903.02428.

Jérôme Galon, Anne Costes, Fatima Sanchez-Cabo, Amos Kirilovsky, Bernhard Mlecnik, Christine Lagorce-Pagès, Marie Tosolini, Matthieu Camus, Anne Berger, Philippe Wind, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science*, 313 (5795):1960–1964, 2006.

Jr Harrell, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982. ISSN 0098-7484. doi: 10.1001/jama.1982.03320430047030. URL https://jamanetwork.com/journals/jama/article-abstract/372568.

Julius Hense, Mina Jamshidi Idaji, Oliver Eberle, Thomas Schnake, Jonas Dippel, Laure Ciernik, Oliver Buchstab, Andreas Mock, Frederick Klauschen, and Klaus-Robert Müller. xMIL: Insightful Explanations for Multiple Instance Learning in Histopathology. *arXiv preprint arXiv:2406.04280*, 2024. URL https://arxiv.org/abs/2406.04280.

Qingtao Hu, Yu Hong, Pan Qi, Guangqing Lu, Xueying Mai, Sheng Xu, Xiaoying He, Yu Guo, Linlin Gao, Zhiyi Jing, et al. Atlas of breast cancer infiltrated B-lymphocytes revealed by paired single-cell RNA-sequencing and antigen receptor profiling. *Nature communications*, 12(1):2186, 2021.

Thomas Hu, Mayar Allam, Vikram Kaushik, Steven L. Goudy, Qin Xu, Pamela Mudd, Kalpana Manthiram, and Ahmet F. Coskun. Spatial Morphoproteomic Features Predict Uniqueness of Immune Microarchitectures and Responses in Lymphoid Follicles. *bioRxiv*, 2024. doi: 10.1101/2024.01.05.574186. URL https://www.biorxiv.org/content/early/2024/01/07/2024.01.05.574186.

Marius Ilie, Véronique Hofman, Cécile Ortholan, Christelle Bonnetaud, Céline Coëlle, Jérôme Mouroux, and Paul Hofman. Predictive clinical outcome of the intratumoral CD66b-positive neutrophil-to-CD8-positive T-cell ratio in patients with resectable nonsmall cell lung cancer. *Cancer*, 118(6):1726–1737, 2012.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 7 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Guillaume Jaume, Pushpak Pati, Antonio Foncubierta-Rodriguez, Florinda Feroce, Giosue Scognamiglio, Anna Maria Anniciello, Jean-Philippe Thiran, Orcun Goksel, and Maria Gabrani. Towards Explainable Graph Representations in Digital Pathology. *arXiv*, 2020. URL https://arxiv.org/abs/2007.00311.

K Jóźwiak, VH Nguyen, L Sollfrank, SC Linn, and M Hauptmann. Cox proportional hazards regression in small studies of predictive biomarkers. *Scientific Reports*, 14(1):14232, 2024. URL https://pubmed.ncbi.nlm.nih.gov/38902269/.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-Event Prediction with Neural Networks and Cox Regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019. URL http://jmlr.org/papers/v20/18-424.html.

Céline M Laumont, Allyson C Banville, Mara Gilardi, Daniel P Hollern, and Brad H Nelson. Tumour-infiltrating B cells: immunological mechanisms, clinical impact and therapeutic opportunities. *Nature Reviews Cancer*, 22(7):414–430, 2022.

Simon Letzgus, Patrick Wagner, Jonas Lederer, Wojciech Samek, Klaus-Robert Müller, and Grégoire Montavon. Toward Explainable Artificial Intelligence for Regression Models: A methodological perspective. *IEEE Signal Processing Magazine*, 39 (4):40–58, 2022. doi: 10.1109/MSP.2022.3153277. URL https://ieeexplore.ieee.org/document/9810062.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1711.05101.

R. Nakhli, P. Moghadam, H. Mi, H. Farahani, A. Baras, B. Gilks, and A. Bashashati. Sparse Multi-Modal Graph Transformer with Shared-Context Processing for Representation Learning of Giga-pixel Images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11547–11557, Los Alamitos, CA, USA, 6 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.01111. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01111.

Thomas Schnake, Oliver Eberle, Jonas Lederer, Shinichi Nakajima, Kristof T. Schütt, Klaus-Robert Müller, and Grégoire Montavon. Higher-Order Explanations of Graph Neural Networks via Relevant Walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7581–7596, 2022. doi: 10.1109/tpami.2021.3115452. URL https://ieeexplore.ieee.org/document/9547794.

Rebecca L. Siegel, Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. Cancer Statistics, 2021. *CA: A Cancer Journal for Clinicians*, 71 (1):7–33, 2021. doi: 10.3322/caac.21654. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21654.

Ryan Soklaski, Justin Goodwin, Olivia Brown, Michael Yee, and Jason Matterer. Tools and Practices for Responsible AI Engineering. *arXiv preprint arXiv:2201.05647*, 2022. URL https://arxiv.org/abs/2201.05647.

Mark Sorin, Morteza Rezanejad, Elham Karimi, Benoit Fiset, Lysanne Desharnais, Lucas J.M. Perus, Simon Milette, Miranda W. Yu, Sarah M. Maritan, Samuel Doré, Émilie Pichette, William Enlow, Andréanne Gagné, Yuhong Wei, Michele Orain, Venkata S.K. Manem, Roni Rayes, Peter M. Siegel, Sophie Camilleri-Broët, Pierre Olivier Fiset, Patrice Desmeules, Jonathan D. Spicer, Daniela F. Quail, Philippe Joubert, and Logan A. Walsh. Single-cell spatial landscapes of the lung tumour immune microenvironment. *Nature 2023 614:7948*, 614:548–554, 2 2023. ISSN 1476-4687. doi: 10.1038/s41586-022-05672-3. URL https://www.nature.com/articles/s41586-022-05672-3.

Mookund Sureka, Abhijeet Patil, Deepak Anand, and Amit Sethi. Visualization for Histopathology Images using Graph Convolutional Neural Networks. In *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 331–335, 2020. doi: 10.1109/BIBE50027.2020.00060. URL https://arxiv.org/abs/2006.09464.

Yanan Wang, Yu Guang Wang, Changyuan Hu, Ming Li, Yanan Fan, Nina Otter, Ikuan Sam, Hongquan Gou, Yiqun Hu, Terry Kwok, John Zalcberg, Alex Boussioutas, Roger J. Daly, Guido Montúfar, Pietro Liò, Dakang Xu, Geoffrey I. Webb, and Jiangning Song. Cell graph neural networks enable the precise prediction of patient survival in gastric cancer. *npj Precision Oncology 2022 6:1*, 6:1–12, 6 2022. ISSN 2397-768X. doi: 10.1038/s41698-022-00285-5. URL https://www.nature.com/articles/s41698-022-00285-5.

Andreas Wieland, Mihir R Patel, Maria A Cardenas, Christiane S Eberhardt, William H Hudson, Rebecca C Obeng, Christopher C Griffith, Xu Wang, Zhuo G Chen, Haydn T Kissick, et al. Defining HPV-specific B cell responses in patients with head and neck cancer. *Nature*, 597(7875):274–278, 2021.

Ping Xiong, Thomas Schnake, Grégoire Montavon, Klaus-Robert Müller, and Shinichi Nakajima. Efficient Computation of Higher-Order Subgraph

Attribution via Message Passing. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24478–24495. PMLR, 7 2022. URL https://proceedings.mlr.press/v162/xiong22a.html.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1810.00826.

Baoyi Zhang, Chenyang Li, Jia Wu, Jianjun Zhang, and Chao Cheng. DeepCG: A cell graph model for predicting prognosis in lung adenocarcinoma. *International Journal of Cancer*, 154(12):2151–2161, 2024. ISSN 1097-0215. doi: 10.1002/ijc.34901. URL https://pubmed.ncbi.nlm.nih.gov/38429627/.

Yanning Zhou, Simon Graham, Navid Alemi Koohbanani, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. CGC-Net: Cell Graph Convolutional Network for Grading of Colorectal Cancer Histology Images. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 388–398, 2019. doi: 10.1109/ICCVW.2019.00050. URL https://arxiv.org/abs/1909.01068.

Table 2: Hyperparameters used for survival regression and classification. Due to implementation constraints, we use online learning for survival classification.

Hyperparameter	Value(s)
Learning rate	{5e−5, 1e−5, 5e−6}
Batch size	16 (1)
Hidden dimension	64
Message passing layers	3
Number of epochs	50

Appendix A. Methods

A.1. Model and Training

We use the AdamW optimizer (Loshchilov and Hutter, 2019) with the default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1\text{e}{-8}$ to train our models. Table 2 lists other hyperparameters used for training our survival regression and classification models. During training, the learning rate is reduced by a cosine annealing schedule. Our training setup is implemented using `hydra-zen` (Soklaski et al., 2022) to be easily configurable and reproducible. We use PyTorch Geometric (Fey and Lenssen, 2019) for the implementation of our GNN-based survival regression model.

Our survival regression models are trained using the Cox negative partial log-likelihood loss (Kvamme et al., 2019):

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N_{\delta=1}} \sum_{i\,:\,\delta_i=1} \log \left(\sum_{j \in \mathcal{R}_i} \exp \left[\hat{h}_{\boldsymbol{\theta}}(\boldsymbol{x}_j) - \hat{h}_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \right] \right)$$

All models were trained on an NVIDIA L4 GPU with 24 GB of memory.

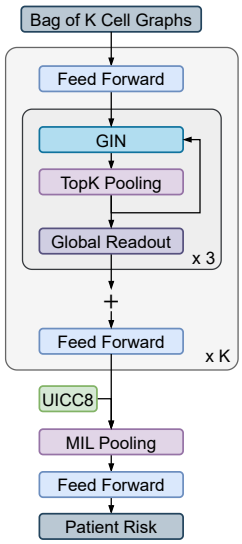


Figure 3: Illustration of our proposed bag of cell graphs GNN architecture for survival regression. The cell graphs are constructed from the KNN neighborhood of the individual cell of patients. The UICC8 cancer stage is fused before the MIL pooling. MIL pooling is either implemented with attention-based pooling (Ilse et al., 2018) for better performance or mean pooling, for simpler interpretation.

Appendix B. Evaluation

B.1. Scalability

Runtime Comparison To evaluate the scalability of our grid-based sGNN-LRP method to large cell graphs, we compare our approach to a naive GNN-LRP implementation (Schnake et al., 2022) computing relevances exhaustively. For this purpose, we synthetically generate cell graphs by sampling a given number of nodes uniformly in the unit circle and applying KNN as before. Figure 4 shows wall-clock run-

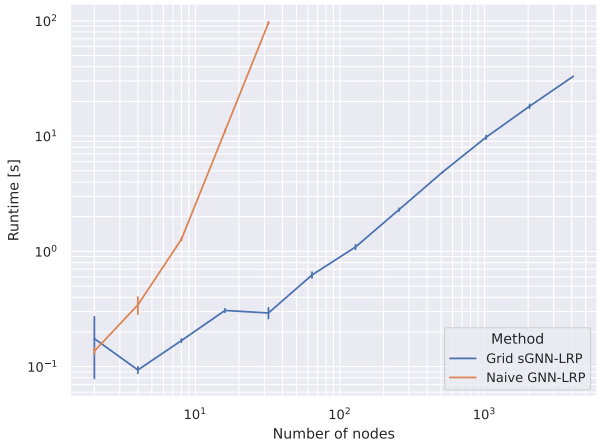


Figure 4: Comparison of wall-clock runtime between the naive GNN-LRP implementation and our grid-based sGNN-LRP method for synthetically generated cell graphs with different numbers of nodes. The runtimes are averaged over three repetitions, with the vertical lines indicating the standard deviation.

times averaged over three repetitions for increasing numbers of nodes. While the naive implementation is prohibitively slow, reaching a runtime of over one minute at 32 nodes, our grid-based method is capable of explaining graphs with more than a thousand nodes in seconds.

Memory Usage Storing the full adjacency matrix as a dense matrix results in a memory requirement of $\mathcal{O}(n^2)$, where n is the number of nodes. Due to the particular sparsity of our graphs, albeit with high numbers of nodes, we especially profit from the reduced memory requirements of the PyTorch Sparse implementation, which does not store the zero entries of the adjacency matrix.

B.2. Interpreting Model Explanations

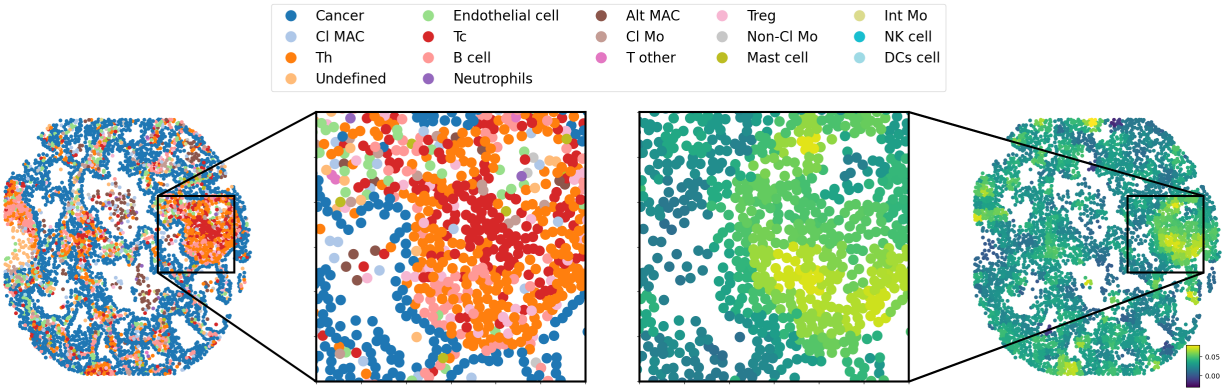
The layer-wise relevance propagation (LRP) input attribution heatmaps can be interpreted similarly to those produced by the input-times-gradient rule, as both methods reflect the sensitivity of the model to input perturbations. In our case, the input is a one-hot vector that can only activate or deactivate the attribution without affecting its sign, allowing an independent discussion of the role of the gradient.

To interpret the sign, consider a scenario where two output neurons represent the softmax probabilities for positive and negative predictions. A positive gradient in the neuron for the positive prediction indicates evidence for a positive outcome (e.g. long survival). Conversely, a positive gradient in the negative prediction neuron signals evidence for a negative outcome (e.g. short survival).

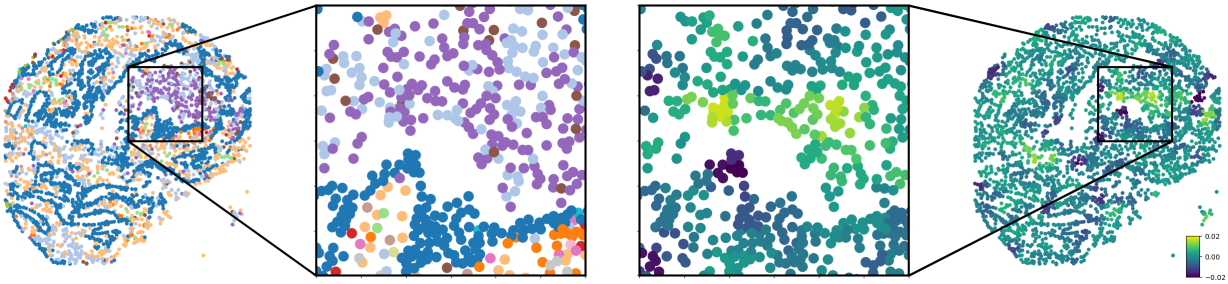
Therefore, the interpretation of positive attributions depends on the context of the specific classification result. In summary, these heatmaps should be seen as the evidence that the model uses to justify its prediction.

B.3. Qualitative Evaluation of Risk Attributions

In addition to quantitatively evaluating risk attributions among phenotypes across cases described in Section 3.2, we perform a qualitative analysis by visualizing cell types and risk attribution heatmaps for selected spots. Figure 5 shows zoomed-in regions of interest for exemplary spots. For long survival, we see that high relevance is predominantly assigned to immune cells (T helper cells, B cells, regulatory T cells, and cytotoxic T cells), while the surrounding cancer tissue is assigned lower relevance. Looking at short survival, we see that high relevance is attributed to an area enriched in neutrophils and classical macrophages, while the surrounding cancer tissue is of lower relevance. These findings are in line with our quantitative analysis across cases in Section 3.2, showing our method can give meaningful insight into the TME for individual tissue spots, which is consistent with existing domain knowledge.

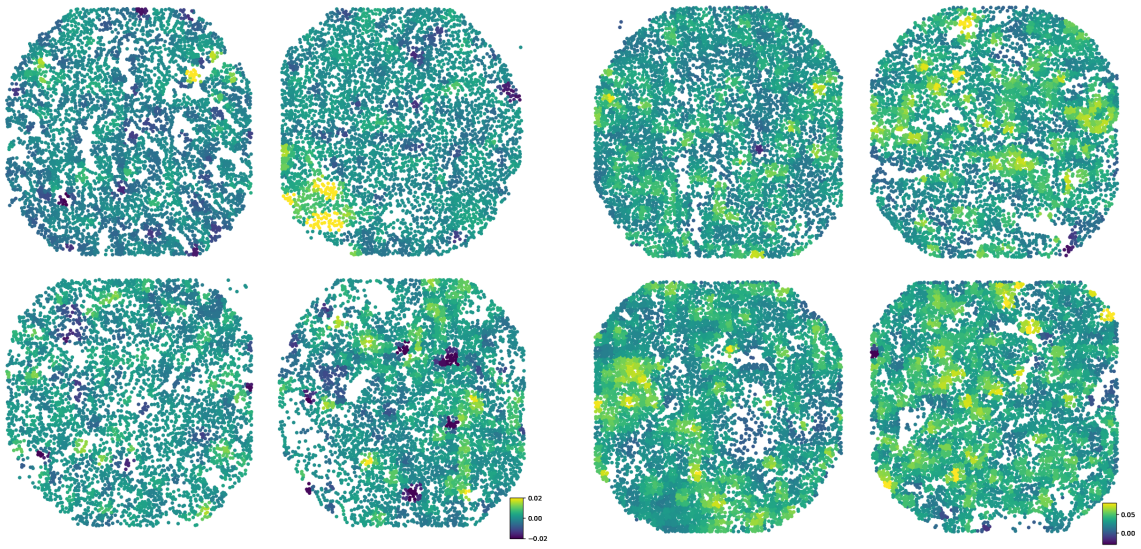


(a) Long survival.



(b) Short survival.

Figure 5: Regions of interest showing the spatial distribution of cell phenotypes next to the corresponding risk attribution heatmaps in (a) long and (b) short survival. High values of attribution indicate positive evidence for the respective model decision, for more details on interpreting explanation heatmaps see Section B.2.



(a) Short survival. (b) Long survival.

Figure 6: Exemplary risk attribution heatmaps generated by our grid-based GNN-LRP method for (a) short ($\leq$ 36 months) and (b) long ($>$ 36 months) survival. Attributions are computed with a tile size $t = 0.05$ mm and a stride $s = 0.025$ mm. To ensure comparability within survival classes, we min-max normalize risk attributions in the intervals $[-0.02, 0.02]$ and $[-0.03, 0.08]$ for short and long survival, respectively. High values of attribution indicate positive evidence for the respective model decision, for more details on interpreting explanation heatmaps, see Section B.2. For a side-by-side comparison to the annotated phenotypes of highly attributed cells, see Figure 5