Hyper-AdaC: Adaptive clustering-based hypergraph representation of whole slide images for survival analysis

Hakim Benkirane HAKIM.BENKIRANE@CENTRALESUPELEC.FR
CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, 91190
Oncostat U1018, Inserm, Université Paris-Saclay, Équipe Labellisée Ligue Contre le Cancer, CESP, Villejuif, 94805

Maria Vakalopoulou MARIA.VAKALOPOULOU@CENTRALESUPELEC.FR
Stergios Christodoulidis STERGIOS.CHRISTODOULIDIS@CENTRALESUPELEC.FR
CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, 91190

Ingrid-Judith Garberis INGRID-JUDITH.GARBERIS@GUSTAVEROUSSY.FR
Inserm UMR981, Gustave Roussy Cancer Campus, Villejuif, France
Université Paris-Saclay, 94270 Le Kremlin-Bicêtre, France

Stefan Michiels STEFAN.MICHIELS@GUSTAVEROUSSY.FR
Oncostat U1018, Inserm, Université Paris-Saclay, Équipe Labellisée Ligue Contre le Cancer, CESP, Villejuif, 94805
Bureau de Biostatistique et d'Épidémiologie, Gustave Roussy, Université Paris-Saclay, Villejuif, 94805

Paul-Henry Cournède PAUL-HENRY.COURNEDE@CENTRALESUPELEC.FR
CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, 91190

Abstract

The emergence of deep learning in the medical field has popularized the development of models to predict survival outcomes from histopathology images in precision oncology. Graph-based formalism has opened interesting perspectives for generating informative representations, as they can be context-aware and model local and global topological structures in the tumor's microenvironment. However, the critical issue in using graph representations lies in their generalizability. They can suffer from overfitting due to their large sizes or high discrepancies between nodes due to random sampling from WSI. In addition, standard graph formulations are limited to pairwise interactions, which can sometimes fail to represent the reality observed in histopathology and hinder the interpretability of those interactions. In this work, we present Hyper-AdaC, an adaptive clustering-based hypergraph representation to model high-order correlations among different regions of the WSIs while being compact enough to help graph neural networks generalize in the case of survival prediction. We evaluate our approach on 5 different public available cancer datasets. Our method outperforms most state-of-the-art graph-based methods for survival prediction with WSIs, creating a more efficient and robust alternative to other graph representations. Moreover, due to our formulation, attention maps are depicted at different resolutions depending on the tissue characteristics of each WSI. The code is available at: https://github.com/HakimBenkirane/Hyper-adaC.

Keywords: Histopathology, Hypergraphs, Survival Analysis, Representation Learning, Interpretability.

1. Introduction

Computational Pathology has rapidly developed over the past decade due to the development of whole-slide image (WSI) scanners that digitize histopathology, immunohistochemistry, or cytology slides into high-resolution images (Zhang, 2019; Lin, 2019). Indeed, their use for cancer diagnosis and prognosis has increased, relying on massive progress in gigapixel image analysis with statistical learning. In this perspective, WSIs have been used for numerous prediction tasks, one of the most challenging being survival prediction (Zhu et al., 2016, 2017), which models the survival function until the occurrence of a particular event (e.g., death, relapse). For this purpose, multiple approaches were adopted in the literature to deal with the challenge of processing large images to obtain survival models.

One of the most popular methods, Multiple Instance Learning (MIL), performs weakly-supervised learning on WSIs by extracting small image patches as independent instances and aggregating them in bags of unordered instances (Sudharshan et al., 2019). However, even if this approach has performed well for some tasks like cancer grading (Zhou et al., 2019) and subtyping (Anand et al., 2020), its adaptation to survival prediction is not straightforward as it should rely on local, as well as global-level characteristics of the WSI. Standard MIL approaches only consider bags of instances as independent and thus do not incorporate context information, failing to learn general associations in the tumor or its environment to assess patient mortality risk (Saltz et al., 2018). To alleviate this issue, graph representations have known a growing interest as they can embed global interactions between patches in a network that allows communication between them (Adnan et al., 2020; Li et al., 2018; Chen et al., 2021). However, existing works on this subject either consider huge graphs that can hinder Graph Neural Networks' (GNNs) generalizability (Yehudai et al., 2021) or involve sampling, which covers only part of the WSI and neglects lots of pathological tissues. Moreover, graph representations being limited to pairwise associations can sometimes fail to model local structures when there are significant discrepancies between instances (Garg et al., 2020).

In this work, we propose a novel hypergraph representation (Hyper-AdaC) based on adaptive clustering (Müllner, 2011) for accurate survival prediction. The contribution of this model to the representation of gigapixel WSIs is threefold. First, we deal with the limitations of the graph size by using hierarchical clustering based on both morphological similarity and spatial proximity to summarize WSIs information efficiently. This method is easy to adapt and does not rely on constraining hypotheses, like the number of clusters to consider. This method can also be seen as a way to efficiently bypass the limitations of random patch sampling as it filters the most relevant patches from the WSI, resulting in less loss of information. Secondly, we overcome the constraints induced by the local structures thanks to our hypergraph representations of those clustered instances depending on morphological and spatial features. Finally, our method succeeds at generating high-resolution attention maps that adapt to the morphology of the tissue thanks to agglomerative clustering, providing more insights into specific elements of the WSI like immunological response and directly linking it to survival. To benchmark the performance of our method, we quantitatively evaluate it on 5 different cancer datasets from The Cancer Genome Atlas (TCGA) and compare it to several state-of-the-art methods for survival outcome prediction, illustrating its better performance.

2. Related Work

Several methods have been developed for survival analysis in computational pathology, mainly using MIL approaches (Mobadersany et al., 2018; Lu et al., 2021; Yao et al., 2020). Those methods rely on sampling a limited number of patches to deal with the large size of WSIs, suffering from coverage and generalization issues, as shown in multiple studies (Ciga et al., 2021; Di et al., 2022). To overcome those limitations, multiple approaches have been proposed, in which patches are grouped using clustering algorithms such as K-Means algorithm before sampling (Zhu et al., 2017; Yao et al., 2020) to identify morphological phenotypes in WSIs and reduce the dimensionality.

Recently, different studies (Chen et al., 2021; Shao et al., 2021) started taking an interest in correlations between small instances of gigapixel images, which is neglected by the initial hypothesis of the MIL approach (Carbonneau et al., 2018). Following this idea, graph-based representations have become an excellent alternative for robust context-aware representations (Li et al., 2018; Zheng et al., 2021). To alleviate the issue of limited sampling, (Chen et al., 2021) proposed a way to model interactions between features of adjacent patches using a k-nearest neighbors (k-nn) graph. As classical graph representations can only model pairwise interactions between image patches, new methods are considering broader representations by trying to lift the i.i.d. hypothesis from standard MIL (Shao et al., 2021), or by switching to hypergraph representations (Di et al., 2020, 2022). Contrary to these methods, Hyper-AdaC relies on hypergraphs to capture interesting spatial and morphological features from WSIs, harvesting informative global and local WSI dependencies for survival models.

3. Method

Within the scope of this study, we design, implement and evaluate a hypergraph-based survival network for survival outcome prediction. For $1 \leq i \leq N$, let us denote by W_i, the WSI of a patient, T_i its event time, and C_i its censoring status. The goal of this study is to build and train a survival neural network $\mathcal{S}$ and to determine a function ϕ that maps the WSI into a hypergraph representation, such that $\mathcal{S}(\phi(W_i), \mathbf{\Theta}) = r_i$, with $\mathbf{\Theta}$ a set of trainable parameters and r_i the hazard rate of the time-to-event outcome of interest.

3.1. Hypergraph Construction

We denote by G_i a hypergraph representation of W_i such that $\phi(W_i) = G_i$. Before the construction of the hypergraph, we first performed automatic tissue and background separation using (Lu et al., 2021). We then extract non-overlapping 256×256 patches at $20\times$ magnification that are fed to a ResNet-18 trained using the same contrastive learning strategy (SimCLR) as in (Ciga et al., 2022) that represents a 1024-dimensional feature vector $h \in \mathbb{R}^{1024}$ each patch. The set of $(h_j)_{1 \leq j \leq n_p}$ associated to a W_i with n_p patches will be stacked into a feature matrix $\mathbf{X}_i \in \mathbb{R}^{n_p \times 1024}$. Each patch x_j is characterized by its ResNet-18 feature representation h_j that embeds the morphological properties of the patch and a set of coordinates $g_j = (g_{x,j}, g_{y,j})$ that represents the spatial position of the center of the patch. Since the hypergraph should not be too large for the generalizability of the GNN (Yehudai et al., 2021), we perform a first step of Adaptive Agglomerative Clustering on the different patches. For that, we compute two similarity matrices $K_h \in \mathbb{R}^{n_p \times n_p}$ and $K_g \in \mathbb{R}^{n_p \times n_p}$ such that $K_h = (\kappa_h(x_i, x_j))_{1 \leq i,j \leq n_p}$ and $K_g = (\kappa_g(x_i, x_j))_{1 \leq i,j \leq n_p}$ where $\kappa_h(x_i, x_j) = e^{-\lambda_h ||h_i - h_j||^2}$ is a morphological similarity metric and $\kappa_g(x_i, x_j) = e^{-\lambda_g ||g_i - g_j||^2}$ is a spa-

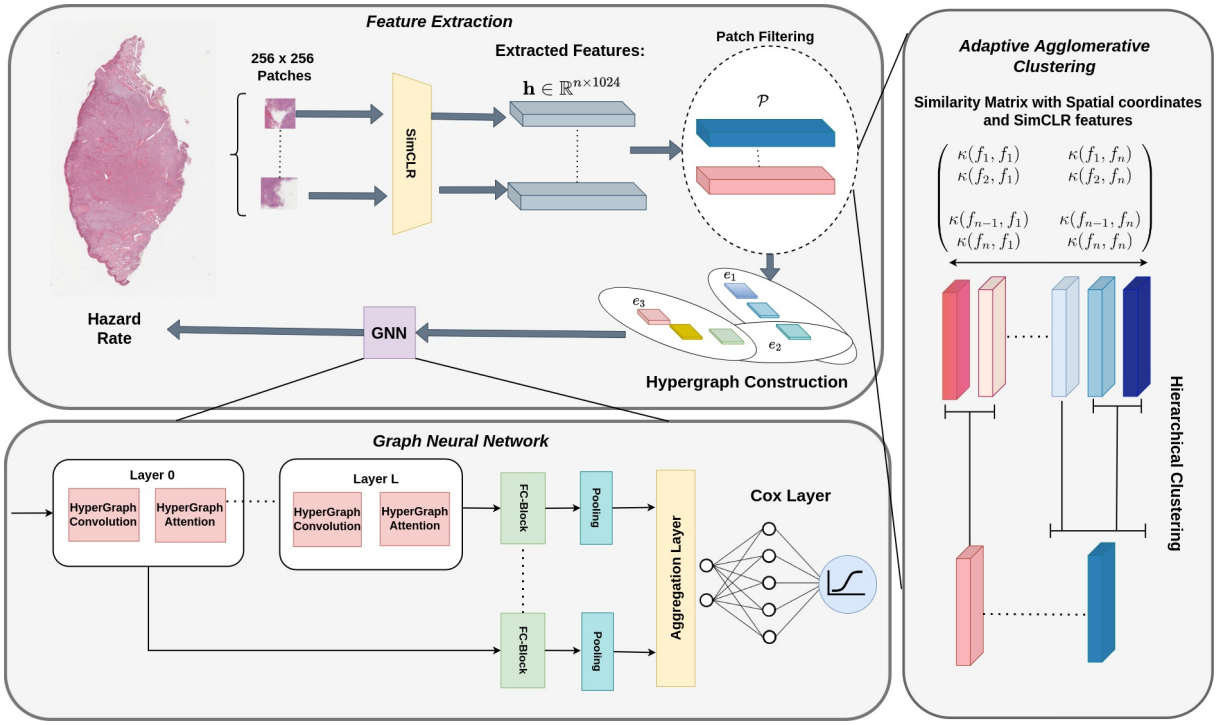

Figure 1: Overview of the Hyper-AdaC pipeline. We first perform a feature extraction step using SimCLR trained on TCGA images. Then the features are processed into a clustering step that performs agglomerative clustering based on a similarity metric κ. The clustered features serve as nodes to construct a hypergraph fed to a Graph Neural Network (GNN) optimized by hazard-based loss function. The GNN is composed of multiple hypergraph convolutions and attention modules, followed by an FC-block (Fully-Connected block) and a global pooling layer.

tial proximity metric. Following the ideas presented in (Lu et al., 2022), we use the kernel $\kappa(x_i, x_j) = \kappa_h(h_i, h_j)\kappa_g(g_i, g_j)$ as a similarity kernel for agglomerative clustering. This kernel will be computed for each pair of patches from the same WSI. All patches for which similarity will be greater than a threshold δ will be considered to belong to the same cluster C_k and merged hierarchically into a single patch representation $p_k = (\tilde{h}_k, \tilde{g}_k)$ where $\tilde{h}_k = \frac{1}{|C_k|}\sum_{j\in C_k} h_j$ and $\tilde{g}_k = \frac{1}{|C_k|}\sum_{j\in C_k} g_j$.

Now that we have a reduced set of points $\mathcal{P}_i$, a hypergraph denoted by $G_i =< V_i, E_i, \mathbf{X}_i >$ is constructed. For a single WSI, we consider each clustered patch as a vertex of the hypergraph such that $V_i = [p_j]_{j\in\mathcal{P}_i}$. Each hyperedge is associated to the neighbourhood of each node V_i. This neighborhood is defined as $\gamma(p_j) = \{p_k \in \mathcal{P}_i; \kappa_h(p_k, p_j) \geq \delta_h\}$, where δ_h is a threshold value to fine-tune. Those hyperedges are indicated by an incidence matrix $\mathbf{H} \in \mathbb{R}^{|\mathcal{P}_i|\times|E_i|}$ such that,

$$h(k,j) = \begin{cases} 1 & \text{if } p_j \in \gamma(p_k) \\ 0 & \text{else} \end{cases} \quad (1)$$

The interesting aspect about a hypergraph compared to a regular graph is that the

neighborhood of each node is depicted as a single hyperedge. This allows us to train our model with fewer parameters and thus decrease the time complexity of the convolution. In addition to this, it creates a community effect that gives more importance to bigger hyperedges, which will represent denser regions of our WSI.

3.2. Construction of the Graph Neural Network

Network's Architecture: The GNN we propose (Fig. 1) consists of a series of hypergraph convolutions and attentions as defined in (Bai et al., 2021), with each layer using a multi-layer perceptron to generate embedding of nodes based on the features of the node itself and its neighbors. Each layer consists of batch normalization and dropout layers to avoid instability during training. We also use the idea introduced in (Lu et al., 2022) of accumulating the feature representations of the convolution layers in the GNN. Those node-level representations are then pooled to generate a graph-level representation. This representation is then fed to a survival network composed of multi-layer perceptrons that predict the hazard rate used for survival outcome prediction.

Network's Loss Function: The entire network is trained using the Cox proportional hazard loss introduced in (Ching et al., 2018), it uses the partial log-likelihood as the cost function, defined as follows:

$$pl(\mathbf{\Theta}) = \frac{1}{|\{i : C_i = 1\}|} \sum_{i:C_i=1} [\mathcal{S}(\phi(W_i), \mathbf{\Theta})) - log \sum_{T_i \geq T_j} exp(\mathcal{S}(\phi(W_j), \mathbf{\Theta}))] \quad (2)$$

where ϕ is a neural network modelling the hazard ratio and $\mathbf{\Theta}$ are the network's parameters. The cost function to train the model is therefore defined by:

$$\mathcal{L}(\mathbf{\Theta}) = pl(\mathbf{\Theta}) + \lambda ||\mathbf{\Theta}||_2^2 \quad (3)$$

4. Experimental Setup

4.1. Dataset

For this study, we performed extensive experiments using five different cohorts from The Cancer Genome Atlas (TCGA) detailed in Table 1. We chose those five datasets based on size and censoring rate. On average, each WSI contains approximately 12691 patches at 20× magnification that are then reduced by hierarchical clustering to around 3147 points.

4.2. Implementation Details

The architecture of the GNN is constructed using three hypergraph convolution layers of 256 neurons each followed by a three layers survival network of respectively 256, 128, and 64 neurons with ReLU activation that outputs the hazard ratio using a sigmoid activation function in the output layer. The entire architecture is built using fully-connected blocks. For each layer, we use a batch normalization layer to address the problem of internal covariate shift. Also, to avoid overfitting problems, we use dropout with a rate of 0.2. For the graph construction, we select a similarity threshold of 80% with $\lambda_h = 3\lambda_g$ to give more importance to morphological features during the clustering. This choice of hyperparameters has been validated with the experiments presented in Appendix A. To train Hyper-AdaC, we used Adam optimization with a learning rate of 10^{-3} with an exponential scheduler, a weight decay of 10^{-5}, and 20 epochs. All models were trained using an Nvidia Tesla V100S with 32 GB of memory.

Table 1: A detailed description of the cohorts used for the study. The table includes the different cancer types, as well as the number of patients and WSIs per type.

Cancer Type	# of Patients	# of WSIs
Bladder Urothelial Carcinoma (BLCA)	437	457
Breast Invasive Carcinoma (BRCA)	1022	1133
Glioblastoma & Lower Grade Glioma (GBMLGG)	1011	1704
Lung Adenocarcinoma (LUAD)	515	541
Ulterine Corpus Endometrial Carcinoma (UCEC)	538	566

4.3. Evaluation

To evaluate Hyper-AdaC, we perform 5-fold cross-validation for each cancer type. We compute the concordance index (C-index) (Uno et al., 2011) across all the validation folds to measure the predictive performance of the method. We also compare our proposed method to multiple other state-of-the-art methods for the same task. For all our experiments and for a fair comparison, we used the same survival loss function, the exact SimCLR feature embeddings, and training hyperparameters for all methods. The basis of comparison we consider is the following:

- **DeepAttnMISL** (Yao et al., 2020): Performs standard Multiple-Instance Learning by first applying K-Means algorithm to cluster instance-level features and then process each cluster using Siamese networks.
- **DeepGraphSurv** (Li et al., 2018): A graph-based representation over sampled patches, which uses spectral graph convolution (Chung, 1997) to consider the topological relationships between them. We also integrate K-Means before sampling in another setup, presented as C.DeepGraphSurv on the result session.
- **Patch-GCN** (Chen et al., 2021): Current state-of-the-art for GNN for the survival task. It performs graph multiple instance learning by considering the WSI as a 2D-point cloud, building a k-nearest neighbors graph.
- **knn-hypergraph** (Di et al., 2020): k-nearest neighbors hypergraph construction using sampling of patches. We use the same pipeline as Hyper-AdaC.

5. Results & Discussion

When comparing our approach to other methods, we note that Hyper-AdaC outperforms most of these in terms of C-index (Table 2 and Figure 2). In general, our approach outperforms by at least 1.6% the overall C-index on all datasets and, more specifically, in most of the individual datasets (except for BLCA and GBMLGG). When comparing with the results of DeepGraphSurv, we can immediately identify the limitations of sampling patches from WSIs, as this method is the weakest in these comparisons. It only covers around 20% of the WSI and fails to train GNNs due to significant discrepancies between sampled patches. We also witness a clear improvement by adding context information, as almost all the graph representations outperform the multiple-instance learning method DeepAttnMISL.

Except for the superiority in performance, our method reports better robustness, high-

Table 2: Survival prediction of state-of-the-art methods using the concordance index (C-index) on 5 TCGA cohorts: Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Glioblastoma & Lower Grade Glioma (GBMLGG), Lung Adenocarcinoma (LUAD) and Ulterine Corpus Endometrial Carcinoma (UCEC).

Model	BLCA	BRCA	GBMLGG	LUAD	UCEC
DeepAttnMISL (Yao et al., 2020)	0.514 ± 0.052	0.564 ± 0.050	0.781 ± 0.037	0.558 ± 0.060	0.595 ± 0.067
DeepGraphSurv (Li et al., 2018)	0.495 ± 0.045	0.551 ± 0.077	0.816 ± 0.031	0.563 ± 0.050	0.614 ± 0.052
C.DeepGraphSurv (Li et al., 2018)	0.504 ± 0.042	0.564 ± 0.043	0.787 ± 0.028	0.559 ± 0.036	0.625 ± 0.057
Patch-GCN (Chen et al., 2021)	0.561 ± 0.042	0.587 ± 0.043	$\mathbf{0.834 \pm 0.029}$	0.570 ± 0.050	0.632 ± 0.059
k-nn Hypergraph (Di et al., 2020)	$\mathbf{0.611 \pm 0.049}$	0.545 ± 0.071	0.805 ± 0.044	0.584 ± 0.061	0.615 ± 0.020
Hyper-AdaC (ours)	0.564 ± 0.034	$\mathbf{0.592 \pm 0.025}$	0.778 ± 0.024	$\mathbf{0.595 \pm 0.012}$	$\mathbf{0.667 \pm 0.022}$

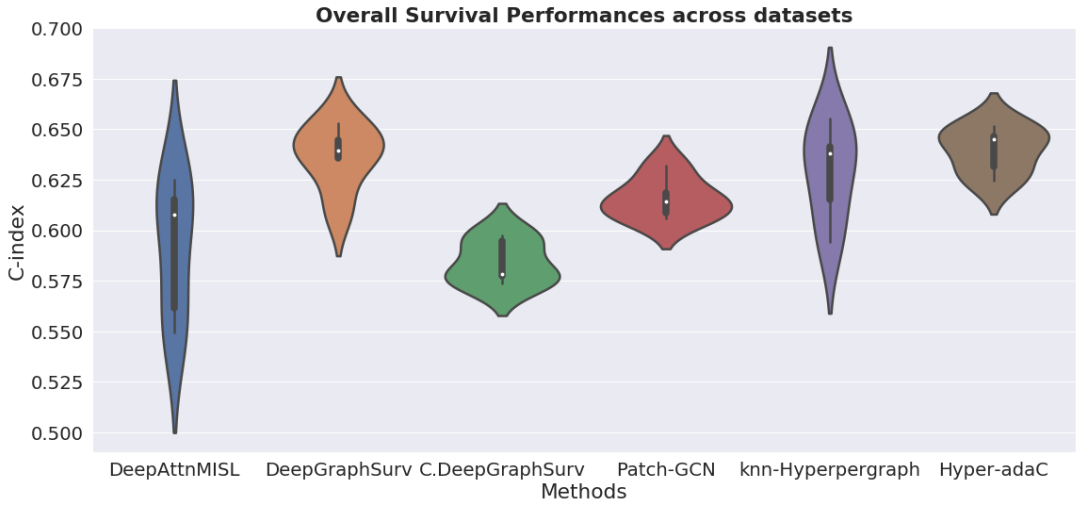

Figure 2: Survival prediction performances across all datasets. They are computed by taking the C-indices on all folds of the evaluation, for all datasets.

lighted by the standard deviation between the C-index values across the 5 folds. One can observe that Hyper-AdaC reports the lowest standard deviation, suggesting a more robust model due to the compact form of its representation. Moreover, as the representation is smaller on Hyper-AdaC, the computing time is lower than considering the entire WSI graph since the graph convolution has in the worst-case complexity of $O(n^3)$ where n is the number of nodes. However, this reduction comes with a trade-off since the graph construction part is heavier due to the hierarchical clustering step that comes with the additional complexity of $\mathcal{O}(kn^2)$, where k is the final number of clusters and n is the initial number of patches. In practice, our method is about 30% slower than graphs constructed using the whole WSI like Patch-GCN or random sampling like DeepGraphSurv. On the other hand, we are almost 20% faster during training due to more compact

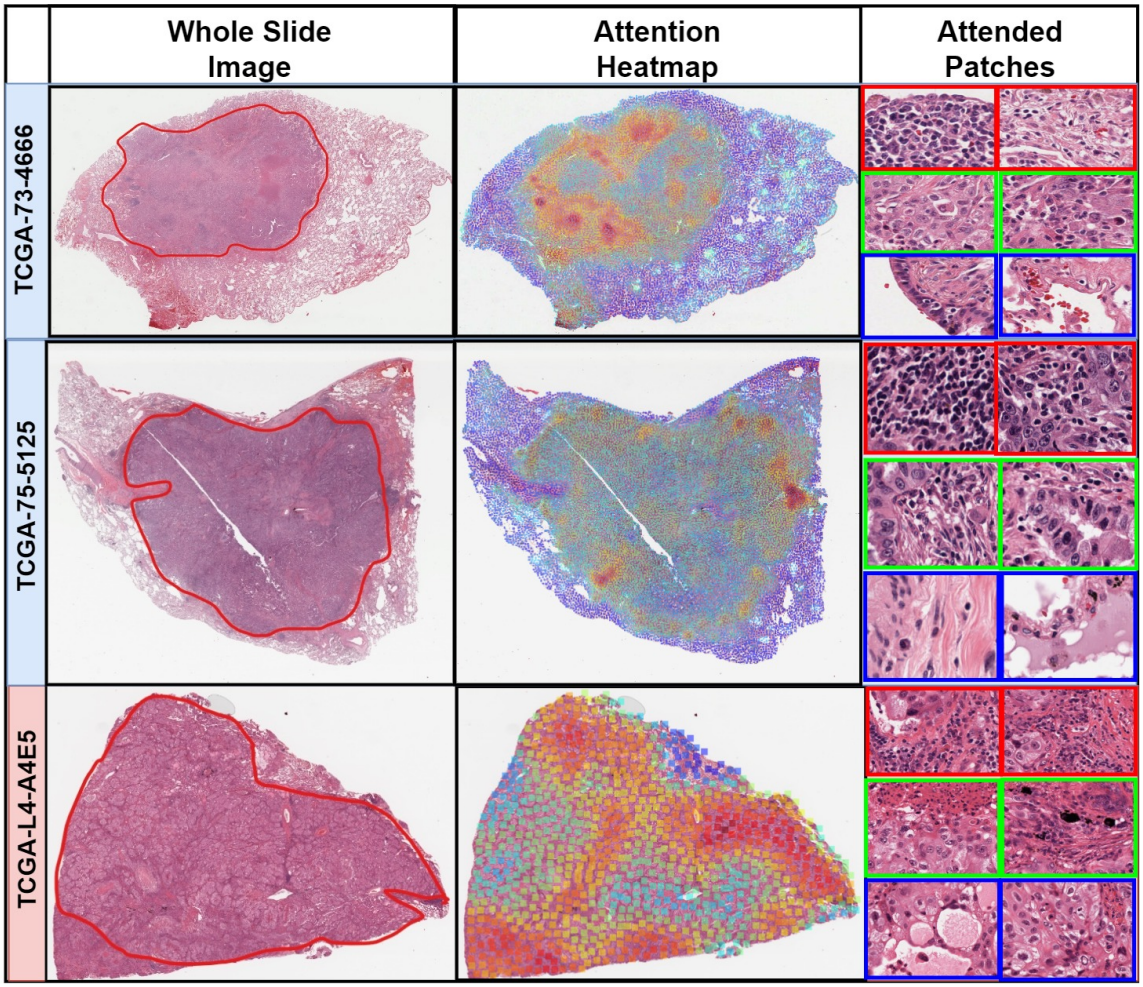

Figure 3: Comparison between the model attention heatmaps and manual annotations of tumor regions for three different patients from the TCGA-LUAD dataset (blue for low-risk patients and red for high-risk patients). First column: annotations of tumor regions (in red) are superimposed in the WSI. Second column: attention heatmaps. Third column: sampled patches from 3 different attention regions; high attention (red border), medium attention (green border), and low attention (blue Border).

representations and fewer parameters. Finally, when we compare the adaptive clustering to k-means through C.DeepGraphSurv, we observe that the adaptive property of the hierarchical clustering compared to K-means provides us with more information as it sums up quite well the discrepancies in the tissue without having to include the number of clusters as it can be adapted one slide to another, depending on the depicting texture.

Our experiments indicate lower performances on BLCA and GBMLGG datasets.

To analyze this point more, we performed some additional experiments detailed in Appendix B. In fact, for the BLCA the number of elements conserved after the agglomerative clustering is still too high, resulting in a bigger graph and, therefore weaker performances. This reasoning can be inverted for GBMLGG for which agglomerative clustering conserves only very little information, meaning that the morphological structure of this particular cancer is more homogeneous than others, and we lose a lot of information as this clustering disregards local variability. To alleviate these issues, dataset-specific hyperparameter tuning can be performed (while we originally preferred common hyperparameters for all datasets to enhance the generalizability of our model). In practice, we add more constraints on the graph construction for BLCA dataset by setting the similarity threshold δ to 85% and relax them on the GBMLGG dataset where δ was set to 70%. We also set $\lambda_h = 2\lambda_g$ for the GBMLGG dataset to focus less on morphological properties since the tissue is generally highly homogeneous and the clustering will be more uniform across the WSI. By doing this, we can witness a spike in performance as the C-index for our method in the BLCA dataset gets to 0.619 ± 0.037 and to 0.812 ± 0.025 for the GBMLGG dataset, similar to the state-of-the-art results.

Examples of WSIs annotated by a pathologist and the corresponding model attention heatmaps are presented in Figure 3. We can observe that our model succeeds in discriminating zones based on their morphological and spatial features. Agglomerative clustering, by being able to adapt the number of clusters to the WSI, enables us to output attention maps that adapt well to the morphology of the slide, focusing on more relevant information and thus providing more precise information on important local regions. Moreover, the tumoral zone indicated by the pathologist in the first column of Figure 3 matches the regions where attention is at its highest. In addition, the model focused on dense inflammatory cell regions for patients with low predicted risk, which are signs of good immunity response. The multiple purple dots highlight those inflammatory cell regions in high-attention regions for the two low-risk patients (third column of Figure 3 showing a zoom of the attended patches). For high-risk patients, regions of tumor cells contain more attention due to their density. This is where the hypergraph construction presents its advantage: it creates a community behavior with hyperedges. It can assess the density of small regions through their weights. Thanks to message passing between hyperedges, areas with more significant communities have a more decisive influence on survival prediction.

6. Conclusion

Computational Pathology has made tremendous progress when dealing with WSI global representations. However, many approaches still suffer from generalizability problems and do not properly model the whole tumor's microenvironment. In this work, we have introduced a compact hypergraph representation, Hyper-AdaC, that solves the size issue of graphs for GNNs without losing important and patient-specific information from whole-slide images. We showed through our experimentation that Hyper-AdaC creates an efficient and robust representation for training GNNs and allows broader associations between patches. In the future, we aim to explore the efficiency of this representation in the promising context of multi-modal learning for survival outcome prediction, combining WSIs with multi-omics and clinical data.

Acknowledgments

The project is supported by the Public Health graduate school of Paris-Saclay University and by the Prism project, funded by the Agence Nationale de la Recherche under grant number ANR-18-IBHU-0002. The Data has been made available by the TCGA research network.

References

Mohammed Adnan, Shivam Kalra, and Hamid R Tizhoosh. Representation learning of histopathology images using graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 988–989, 2020.

Deepak Anand, Shrey Gadiya, and Amit Sethi. Histographs: graphs in histopathology. In *Medical Imaging 2020: Digital Pathology*, volume 11320, pages 150–155. SPIE, 2020.

Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110: 107637, 2021.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–349. Springer, 2021.

Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.

Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

Ozan Ciga, Tony Xu, Sharon Nofech-Mozes, Shawna Noy, Fang-I Lu, and Anne L Martel. Overcoming the limitations of patch-based learning to detect cancer in whole slide images. *Scientific Reports*, 11(1):1–10, 2021.

Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

Donglin Di, Shengrui Li, Jun Zhang, and Yue Gao. Ranking-based survival prediction on histopathological whole-slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 428–438. Springer, 2020.

Donglin Di, Jun Zhang, Fuqiang Lei, Qi Tian, and Yue Gao. Big-hypergraph factorization neural network for survival prediction from whole slide image. *IEEE Transactions on Image Processing*, 31: 1149–1160, 2022.

Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *International Conference on Machine Learning*, pages 3419–3430. PMLR, 2020.

Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International*

Conference on Medical Image Computing and Computer-Assisted Intervention, pages 174–182. Springer, 2018.

et al. Lin, Huangjing. Fast scannet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection. *IEEE Transactions on Medical Imaging*, 38:1948–58, August 2019.

Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Wenqi Lu, Michael Toss, Muhammad Dawood, Emad Rakha, Nasir Rajpoot, and Fayyaz Minhas. Slidegraph+: Whole slide image level graphs to predict her2 status in breast cancer. *Medical Image Analysis*, page 102486, 2022.

Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

PJ Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, and Paul Honeine. Multiple instance learning for histopathological breast cancer image classification. *Expert Systems with Applications*, 117:103–111, 2019.

Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

Gilad Yehudai, Ethan Fetaya, Eli Meirom, Gal Chechik, and Haggai Maron. From local structures to size generalization in graph neural networks. In *International Conference on Machine Learning*, pages 11975–11986. PMLR, 2021.

et al. Zhang, Zizhao. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1:236–45, May 2019.

Yi Zheng, Rushin Gindra, Margrit Betke, Jennifer Beane, and Vijaya B Kolachalama. A deep learning based graph-transformer for whole slide image classification. *medRxiv*, 2021.

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 *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019.

Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547. IEEE, 2016.

Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7234–7242, 2017.

Appendix A. Patch Clustering

We compute the average number of elements remaining after the hierarchical clustering step for each dataset separately, the results along with the ratio between initial and filtered patches are represented in Table 3. We observe that, in general, approximately 14% of the WSI is used (see 3), and as shown in Figure 3, those elements are well spread across the WSI. However, we can see that both BLCA and GBMLGG datasets behave differently from the others. For BLCA, the ratio of remaining elements over the total number of patches is higher than all the other datasets, whereas for GBMLGG it is the opposite. Our method does not perform well for those particular test cases.

Appendix B. Ablation Studies

We perform an ablation study on the different graph hyperparameters to justify our construction choices. In Figure 4, we can see the effect of the similarity threshold δ_h on the survival performances. The stricter the constraint, the better the performance, indicating that larger graphs fail at learning generalizable properties. This idea is also supported by the standard deviation across the 5-folds that decreases, suggesting that the model is less robust with larger graphs. A similarity threshold of 80% achieves the peak performance; past that point, the performances start to decrease again because we tend to oversimplify the WSI and start neglecting information.

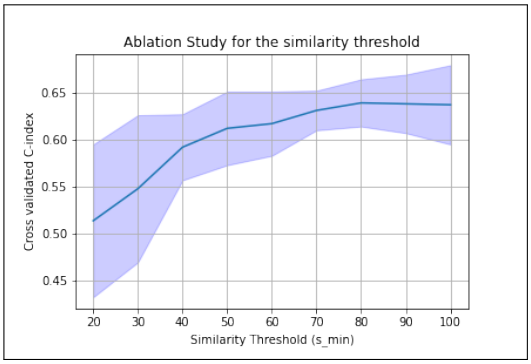

Figure 4: Ablation Study for the similarity threshold δ used in the hierarchical clustering step. We evaluate for each hyperparameter the 5-fold cross-validated C-index on the overall 5 TCGA datasets used in this study.

Figure 5 highlights the relationship between morphological features and geographical properties with respect to the survival prediction performance. We see that, in general, focusing on morphological properties is more beneficial to the performances than spatial properties as they hold more information about the structure of the tissue (including, to a certain extent, spatial information because similar patches tend to be close). However, focusing too much on mor-

Table 3: Average number of nodes after the hierarchical clustering step for each dataset. Due to our selection criteria, the GBMLGG dataset had a significantly lower number of nodes (as the ratio is also lower, it may indicate higher homogeneity among tissues), which may explain the lower performance with respect to the other cancer types. In this study, we selected the same hyperparameters for all the cancer types to prove the generalizability of our method, outperforming the other state-of-the-art methods. Some specific hyperparameters tuning for the GBMLGG and BLCA may resolve this issue.

Cancer Type	**# of patches**	**# of nodes**	$\frac{\text{\# of nodes}}{\text{\# of patches}}$
Bladder Urothelial Carcinoma (BLCA)	58586	9187	0.16
Breast Invasive Carcinoma (BRCA)	38107	5304	0.14
Glioblastoma & Lower Grade Glioma (GBMLGG)	15855	961	0.06
Lung Adenocarcinoma (LUAD)	43445	6003	0.14
Ulterine Corpus Endometrial Carcinoma (UCEC)	56162	7748	0.14

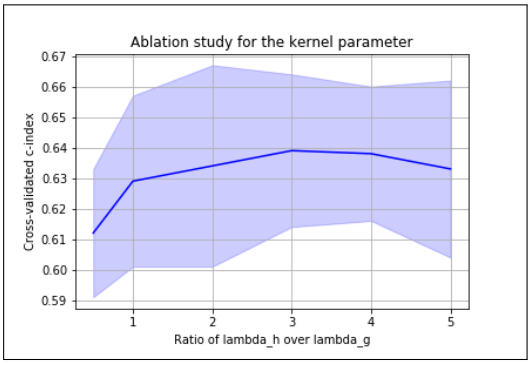

Figure 5: Ablation study for $\frac{\lambda_h}{\lambda_g}$ used in the hierarchical clustering step. We evaluate for each hyperparameter the 5-fold cross-validated C-index on the overall 5 TCGA datasets.

phological features can hinder the accuracy of our survival predictions, as sometimes the homogeneity of specific tissues can make the filtering biased and overlook chunks of WSIs that may hold vital information.

Appendix C. About Hypergraphs

A Hypergraph is a generalization of the graph structure that extends the interaction between instances to a higher level. To describe this complex relationship where an edge can connect to more than two nodes, we define a hypergraph $\mathcal{G} = (V, E)$ as a hypergraph with M vertices and N hyperedges. The hypergraph can then be generated using an incidence matrix $\mathbf{H} \in \mathbb{R}^{N \times M}$. For each vertex i, the vertex degree is defined as $D_{ii} = \sum_{e \in E} H_{ie}$ and the hyperedge degree will be $B_{ee} = \sum_{i \in V} H_{ie}$.

C.1. Hypergraph Convolution

This hypergraph can be associated to a feature matrix $\mathbf{X} \in \mathbb{R}^{N \times F}$ where F is the feature dimension of one node. In the context of our study, this node feature will represent the aggregated Resnet-18 features of one cluster. A step of this convolution is defined in (Bai et al., 2021) as follows:

$$\mathbf{X}^{(l+1)} = \sigma(\mathbf{D}^{-\frac{1}{2}}\mathbf{H}\mathbf{W}\mathbf{B}^{-1}\mathbf{H}^T\mathbf{D}^{-\frac{1}{2}}\mathbf{X}^{(l)}\mathbf{P}) \tag{4}$$

where $\mathbf{W}$ is the weight matrix, σ a non-linear transformation and $\mathbf{P}$ is the weight matrix between layer l and l+1.

C.2. Hypergraph Attention

To build the attention visualization, we use an attention mechanism for hypergraphs described in Bai et al. (2021) as:

$$\alpha_{ij} = \frac{exp(\sigma(sim(x_i\mathbf{P}, x_j\mathbf{P})))}{\sum_{k\in\mathcal{N}_i} exp(\sigma(sim(x_i\mathbf{P}, x_k\mathbf{P})))} \quad (5)$$

where the similarity function computes similarity between two vertices as follows:

$$sim(x_i, x_j) = \mathbf{a}^T[x_i||x_j] \quad (6)$$

where $\mathbf{a}$ is a weight vector and $[.||.]$ denotes concatenation.