1. Evaluation Details

We used concordance index (c-index) [5] and the p-value of the LongRank test [1] to measure the performance of our model and that of the baselines.

C-index is a metric to measure the quality of ranking between the predicted and the observed survival times. More specifically, it is defined as Eq. (1)

\[
c = \frac{1}{|\varepsilon|} \sum_{T_i \text{uncensored}} \sum_{T_j > T_i} \mathbb{1}_{f(x_i) < f(x_j)}, \tag{1}
\]

where \(1_{a<b} = 1\) if \(a < b\) and 0 otherwise, \(f(x_i)\) is the predicted survival time for \(x_i\), and \(\varepsilon\) is the ordered pairs of the data points [5].

The statistical test known as the LongRank test evaluates the validity of the null hypothesis, which states that there is no difference between the survival curves of two populations at any given period. Using the median survival time predicted by the model, we split the patients into two groups of low-risk (patients with a predicted survival time greater than the median) and high-risk (patients with a predicted survival time less than the median). The p-value of the LogRank test on the survival curves of these two cohorts demonstrates the separability of the curves (p-value < 0.05).

In all the experiments, at inference time, we censor the patients with a survival time greater than 10 and 7 years for the InUIT and MIBC datasets, respectively.

2. Implementation Details

2.1. AMIGO

All the experiments were performed on a single GeForce RTX 3090 using Pytorch and DGL packages. The output feature size of the GraphSAGE layers and the MLP layers in each branch were set to 128 and 32, respectively. Adam optimizer with a learning rate of 0.002, a cosine scheduler, a weight decay of 0.0001, and a batch size of 128 were used for the training of the models. The Transformer module of the cross-modal aggregator included 4 MHSAs, and the BCP was set to 0.1. We also used a sparsity ratio of 0.8, which was only applied at training time.

2.2. Baselines

For the implementation of the baselines, we used the official repository of Patch-GCN\(^1\), which also included the implementation for the DeepSet, Attention MIL, and DGC methods. The hyperparameters were set to the suggested values in the paper [3], and we used the NLL loss [6] as suggested. Similarly, for the Pathomic Fusion model, we also used the official implementation repositories with the suggested parameters.

For the HIPT model, the pre-trained weights and model implementation were both taken from the official repository. A loss function similar to the one we used for our model (Cox loss) was used to train an MLP on top of the representation produced by the pre-trained model. Additionally, we adopted the Adam optimizer with a learning rate of 0.001.

It is of note to mention that both the baselines and

\(^1\)https://github.com/mahmoodlab/Patch-GCN
AMIGO utilize the same data sources (all images regardless of stain type). Equivalent to AMIGO’s aggregator, all baseline models include attention pooling in the last layer to pick the most relevant data.

3. Heatmap Visualization

We also visualized the heatmaps of our model on the cellular graphs of a MIBC patient in Fig. S1. Interestingly, we realized that the model learns to pay more attention to the P16 stain, which is in line with previous studies in bladder cancer as they showed the importance of Ki67 in the outcomes of MIBC cases. [4].

4. Masking Visualization

The visualization of the cellular graph after the two masking operations can be found in Fig. S2. These two images are generated at two different training iterations from the same graph. We would like to highlight the variation in the sub-structural graphs within each cell graph during training, which leads to a strong augmentation and regularization for the model.

5. Full Ablation Study

The complete results for the ablation studies can be found in Tab. S3, where both c-index and p-value are reported in each case. Although all of the results support our design decisions, we found that full weight sharing of the branches considerably impairs our model’s capacity to distinguish between cohorts at low and high risk on both datasets. This finding supports our assertion that a multi-modal design is crucial for capturing tissue heterogeneity.

6. Effect of Normalization Type

We also compared the performance of our model with different types of normalization layers applied after the instance attention (Tab. S4). Our results demonstrate that instance normalization has a superior performance across both datasets compared to the other types.
same input graph. The adopted augmentation operations include random dropping of nodes and edges of the graph, and the loss function is a cosine distance loss that brings the cosine similarity of the graph representations close to 1.

We tested this approach in two scenarios: 1) single-modal pre-training; 2) multi-modal pre-training. In the former, we only pre-train the encoders of each multi-modal branch using the BRGL framework. However, in the latter, we initialize the encoder weights with the pre-trained ones from the single-modal version and train the rest of our model (MLPs, instance attention, and the transformer model) using BGRL. Finally, in both scenarios, we fine-tuned the pre-trained model using the survival information. The results of our experiments demonstrate that the BGRL
DeepSet - ResNet50
Attention MIL - ResNet50
DGC - ResNet50

FLOPs (M)
Parameter Count (K)

Table S5. Comparison of the Cox and NLL loss functions for the baselines and our model.

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature Extractor</th>
<th>Loss</th>
<th>InUIT C-Index (↑)</th>
<th>P-value (↓)</th>
<th>MIBC C-Index (↑)</th>
<th>P-value (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepSet</td>
<td>ResNet34</td>
<td>NLL</td>
<td>0.50 ± 0.0</td>
<td>0.43</td>
<td>0.50 ± 0.001</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>ResNet34</td>
<td>Cox</td>
<td>0.50 ± 0.0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>NLL</td>
<td>0.53 ± 0.007</td>
<td>0.40</td>
<td>0.45 ± 0.004</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>Cox</td>
<td>0.50 ± 0.0</td>
<td>–</td>
<td>0.51 ± 0.001</td>
<td>–</td>
</tr>
<tr>
<td>Attention MIL</td>
<td>ResNet34</td>
<td>NLL</td>
<td>0.51 ± 0.004</td>
<td>0.62</td>
<td>0.59 ± 0.007</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>ResNet34</td>
<td>Cox</td>
<td>0.50 ± 0.002</td>
<td>0.22</td>
<td>0.50 ± 0.002</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>NLL</td>
<td>0.55 ± 0.004</td>
<td>0.65</td>
<td>0.55 ± 0.004</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>Cox</td>
<td>0.51 ± 0.003</td>
<td>0.57</td>
<td>0.47 ± 0.001</td>
<td>0.08</td>
</tr>
<tr>
<td>DGC</td>
<td>ResNet34</td>
<td>NLL</td>
<td>0.53 ± 0.007</td>
<td>0.46</td>
<td>0.58 ± 0.007</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>ResNet34</td>
<td>Cox</td>
<td>0.52 ± 0.003</td>
<td>0.69</td>
<td>0.46 ± 0.012</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>NLL</td>
<td>0.55 ± 0.005</td>
<td>0.31</td>
<td>0.54 ± 0.007</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>Cox</td>
<td>0.51 ± 0.001</td>
<td>0.75</td>
<td>0.50 ± 0.010</td>
<td>0.36</td>
</tr>
<tr>
<td>Patch-GCN</td>
<td>ResNet34</td>
<td>NLL</td>
<td>0.53 ± 0.008</td>
<td>0.45</td>
<td>0.50 ± 0.004</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>ResNet34</td>
<td>Cox</td>
<td>0.53 ± 0.002</td>
<td>0.41</td>
<td>0.47 ± 0.005</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>NLL</td>
<td>0.50 ± 0.004</td>
<td>0.25</td>
<td>0.46 ± 0.009</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>Cox</td>
<td>0.52 ± 0.002</td>
<td>0.80</td>
<td>0.52 ± 0.006</td>
<td>0.14</td>
</tr>
<tr>
<td>HIPT</td>
<td>Hierarchical ViT</td>
<td>NLL</td>
<td>0.53 ± 0.001</td>
<td>0.87</td>
<td>0.46 ± 0.003</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Hierarchical ViT</td>
<td>Cox</td>
<td>0.50 ± 0.002</td>
<td>0.18</td>
<td>0.53 ± 0.010</td>
<td>0.10</td>
</tr>
<tr>
<td>AMIGO (Ours)</td>
<td>ResNet34</td>
<td>NLL</td>
<td>0.57 ± 0.003</td>
<td>0.02</td>
<td>0.57 ± 0.010</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>AMIGO (Ours)</td>
<td>ResNet34</td>
<td>Cox</td>
<td>0.57 ± 0.002</td>
<td>0.01</td>
<td>0.61 ± 0.004</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Figure S7. Parameter vs. Flops comparison of our model with the baselines. The size of the points shows is relative to the multiplication of its parameter count and FLOPs.

Figure S8. Self-supervised Model

self-supervised pre-training does not help achieve a consistent improvement over both datasets compared to the supervised setting (Tab. S9). We believe this shows that the self-supervised learning for histopathology graphs requires special considerations that can be part of our future works.

References
[3] Richard J Chen, Ming Y Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y Chen, Drew FK Williamson, and Faisal Mah-


<table>
<thead>
<tr>
<th>Method</th>
<th>Feature Extractor</th>
<th>BCP</th>
<th>InUIT C-Index (↑)</th>
<th>P-value (↓)</th>
<th>MIBC C-Index (↑)</th>
<th>P-value (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepSet</td>
<td>ResNet34</td>
<td>w/ BCP</td>
<td>0.50 ± 0.0</td>
<td>−</td>
<td>0.50 ± 0.001</td>
<td>−</td>
</tr>
<tr>
<td></td>
<td>w/o BCP</td>
<td></td>
<td>0.50 ± 0.0</td>
<td>−</td>
<td>0.50 ± 0.0</td>
<td>−</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>w/ BCP</td>
<td>0.53 ± 0.007</td>
<td>0.40</td>
<td>0.51 ± 0.001</td>
<td>−</td>
</tr>
<tr>
<td></td>
<td>w/o BCP</td>
<td></td>
<td>0.50 ± 0.0</td>
<td>−</td>
<td>0.51 ± 0.001</td>
<td>−</td>
</tr>
<tr>
<td>Attention MIL</td>
<td>ResNet34</td>
<td>w/ BCP</td>
<td>0.55 ± 0.002</td>
<td>0.84</td>
<td>0.52 ± 0.016</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>w/o BCP</td>
<td></td>
<td>0.50 ± 0.002</td>
<td>0.22</td>
<td>0.50 ± 0.002</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>w/ BCP</td>
<td>0.52 ± 0.001</td>
<td>0.44</td>
<td>0.53 ± 0.004</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>w/o BCP</td>
<td></td>
<td>0.51 ± 0.003</td>
<td>0.57</td>
<td>0.47 ± 0.001</td>
<td>0.08</td>
</tr>
<tr>
<td>DGC</td>
<td>ResNet34</td>
<td>w/ BCP</td>
<td>0.51 ± 0.002</td>
<td>0.47</td>
<td>0.51 ± 0.009</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>w/o BCP</td>
<td></td>
<td>0.52 ± 0.003</td>
<td>0.69</td>
<td>0.46 ± 0.012</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>w/ BCP</td>
<td>0.53 ± 0.003</td>
<td>0.92</td>
<td>0.50 ± 0.007</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>w/o BCP</td>
<td></td>
<td>0.51 ± 0.001</td>
<td>0.75</td>
<td>0.50 ± 0.010</td>
<td>0.36</td>
</tr>
<tr>
<td>Patch-GCN</td>
<td>ResNet34</td>
<td>w/ BCP</td>
<td>0.53 ± 0.002</td>
<td>0.44</td>
<td>0.55 ± 0.009</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>w/o BCP</td>
<td></td>
<td>0.53 ± 0.002</td>
<td>0.41</td>
<td>0.47 ± 0.005</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>w/ BCP</td>
<td>0.52 ± 0.003</td>
<td>0.77</td>
<td>0.58 ± 0.009</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>w/o BCP</td>
<td></td>
<td>0.52 ± 0.002</td>
<td>0.80</td>
<td>0.50 ± 0.006</td>
<td>0.14</td>
</tr>
<tr>
<td>HIPT</td>
<td>Hierarchical ViT</td>
<td>w/ BCP</td>
<td>0.49 ± 0.001</td>
<td>0.33</td>
<td>0.49 ± 0.001</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>w/o BCP</td>
<td></td>
<td>0.50 ± 0.002</td>
<td>0.18</td>
<td>0.53 ± 0.010</td>
<td>0.10</td>
</tr>
<tr>
<td>AMIGO (Ours)</td>
<td>ResNet34</td>
<td>w/ BCP</td>
<td>0.57 ± 0.002</td>
<td>0.01</td>
<td>0.61 ± 0.004</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>AMIGO (Ours)</td>
<td>w/o BCP</td>
<td></td>
<td>0.54 ± 0.001</td>
<td>0.07</td>
<td>0.58 ± 0.013</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Table S6. Comparison of the effect of BCP on the baseline models.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-training Type</th>
<th>InUIT C-Index (↑)</th>
<th>P-value (↓)</th>
<th>MIBC C-Index (↑)</th>
<th>P-value (↓)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-supervised (BGRL)</td>
<td>Single-Modal</td>
<td>0.56 ± 0.002</td>
<td>0.04</td>
<td>0.61 ± 0.008</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>Multi-Modal</td>
<td>0.55 ± 0.005</td>
<td>0.28</td>
<td>0.62 ± 0.014</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table S9. Comparison of self-supervised training.