SI-MIL: Taming Deep MIL for Self-Interpretability in Gigapixel Histopathology

Supplementary Material

In this supplementary material, details are provided on the following:

- Dataset details (additional) (8)
- Implementation details (additional) (9)
- SI-MIL additional results (10)
- SI-MIL ablation studies: hyperparameter sensitivity (11)
- SI-MIL ablation studies: model components (12)
- Dataset contribution details (additional) (13)
- Local interpretability analysis (additional) (14)
- Global interpretability analysis (additional) (15)
- Top-K comparative analysis (16)
- Hand-crafted PathExpert feature extraction (17)

8. Dataset details (additional)

We benchmark our SI-MIL on three WSI datasets, namely **TCGA-BRCA**, **TCGA-Lung**, and **TCGA-CRC**. SI-MIL necessitates PathExpert features for interpretable prediction. We use HoVer-Net for segmenting and classifying the nuclei, and afterwards computed hand-crafted PathExpert features. Since HoVer-Net is trained exclusively on 40× magnification patches, our analysis is confined to WSIs having 40× magnification. This ensures accurate nuclei prediction and thereby meaningful PathExpert feature extraction.

TCGA-BRCA is split into 825 training (653 IDC, 172 ILC) and 85 testing (67 IDC, 18 ILC) WSIs following [4]. **TCGA-Lung** dataset is split into 744 training (388 LUAD, 356 LUSC) and 192 testing (96 LUAD, 96 LUSC) WSIs following DSMIL [9]. For **TCGA-CRC**, following [2, 10], we use the first three folds for training, *i.e.*, 241 WSIs (38 hypermutated, 203 not) and the fourth for testing, *i.e.*, 79 slides (12 hypermutated, 67 not) with 40× filtering. The patch extraction process implemented in our study follows the methodology outlined in the aforementioned DSMIL repository [9].

9. Implementation details (additional)

9.1. Deep feature extractors

We compared SI-MIL against different baselines that include training Additive ABMIL using different types of patch features. Details of the patch feature extractors are presented as follows:

IN ViT-S: We train the Additive ABMIL using features extracted by a popular ImageNet-supervised model, specifically ViT-Small [6] model pre-trained using ImageNet dataset [5]. The model extracts a feature embedding of size $D = 384$ for each WSI patch.

RetCCL: We adopted a state-of-the-art feature extractor [15] pre-trained using pathology images. This benchmarks our trained feature extractor, described in Sec. 4 (main paper). This model extracts a feature embedding of size $D = 2048$.

CTransPath: Similar to RetCCL, we benchmarked against a Transformer-based feature extractor pre-trained using pathology images [14]. The resulting patch embeddings are of size $D = 768$.

It is important to note that CTransPath and RetCCL were pre-trained on the pan-TCGA [1] dataset, and our evaluated datasets are subset of this dataset. Therefore, these models were pre-trained using the WSIs in our test dataset, which can potentially result in inflated performances during classification. Though benchmarked, these models may not be suitable for reliable comparisons in our study.

DINO ViT-S: For a reliable comparison, we used DINO [3] to pre-train ViT-Small models for each dataset (TCGA-BRCA, TCGA-Lung, and TCGA-CRC), using the dataset-specific training splits as provided in the Supp Sec. 8 For pre-training, we used the default hyperparameter values of DINO [3], while using only two global crops. These pre-trained models extract a feature embedding of size $D = 384$ for each WSI patch. One RTX 8000 GPU is utilized for pre-training the ViT-S with a batch size of 256.

9.2. SI-MIL

Hyperparameter tuning is performed with a range of learning rates $\in \{1e^{-3}, 2e^{-3}, 1e^{-4}, 2e^{-4}\}$ and weight decays $\in \{1e^{-2}, 5e^{-3}\}$. By default, #PF-Mixer layers =4, $\lambda = 20$, $K = 20$, $\gamma = 0.75$, and $t = 3$. Additionally, $d = 246$ except for in TCGA-Lung where $d = 203$ as annotations for only 4 (instead of 5) cell types are available for HoVer-Net classification in the Lung dataset. For both the predictors $L(\cdot)$ and $C(\cdot)$, we use the sigmoid activation (ψ), since our tasks involve binary classification. Note that $\mathcal{L}_{KD}$ is utilized with stop-gradient since the goal is to align the performance of the *Self-Interpretable* branch to be close to high performing conventional MIL branch in SI-MIL. All MIL experiments are performed on one RTX 8000 GPU.

9.3. Interpretability analysis setup

For interpretability analysis, we compare the separability of the top K patches in the PathExpert feature space between the conventional MIL and SI-MIL (refer to Figure 4 (main paper)). To ensure a fair comparison, we select WSIs from the held-out test set where both MIL methods result in correct predictions.

Note that, we employ 5-fold cross-validation on the training split and held out the test set. We chose the best-performing fold for both local (visualization and pathologist relevancy score experiment) and global (visualization) interpretability analysis for the MIL methods. However, for multivariate class-separability scores (refer to Figure 4 (main paper)), we report the median and standard deviation from all 5-folds. Similarly, we report the median and standard deviation of Jensen-Shannon (JS) divergence across all 5-folds in Figure 4 (main paper).

9.4. SI-MIL complexity analysis

The mitigation of the trade-off between performance and interpretability by SI-MIL can be attributed to the choice of PathExpert features and the SI-MIL design choices, instead of merely an increase in the number of model parameters. It can be justified by comparing the size and performance of SI-MIL with the competing baselines. The number of model parameters in SI-MIL is 625K, while those in conventional MIL with DINO/IN ViT-S, CTransPath and ReTCCL are 345K, 985K, and 5.25M, respectively. Despite the differences in model sizes, SI-MIL results in comparable performance with respect to the competing baselines, as shown in Table 1 (main paper).

10. SI-MIL additional results

Here we provide the mean and standard deviations for the main experiments (refer to Table 1 (main paper)) in Table 4.

11. SI-MIL ablation studies: hyperparameters

In this section, we provide studies of SI-MIL hyperparameters on TCGA-BRCA dataset. Particularly, these ablations demonstrate the effect of varying K in the PAG Top-K module, the number of PF-Mixer layers, and the percentile and temperature for scaling β.

Effect of varying K in the PAG Top-K module: In Figure 5, we illustrate the impact of varying K on SI-MIL performance. We observe that a larger value of K leads to a significant drop in performance compared to the default $K = 20$. This decrease may be attributed to an increase in irrelevant noisy patches, which makes it difficult for the model to classify WSIs in the PathExpert feature space.

Effect of varying number of PF-Mixer layers: SI-MIL's performance is generally robust across various values of the number of PF-Mixer layers, but experiences a performance drop for very high values, *e.g.*, #PF-Mixer layers = 6 (Figure 6). This decline can be attributed to potential overfitting induced as a result of higher number of layers.

Effect of percentile and temperature for scaling β: In Figure 7, we show the variation in performance of SI-MIL with respect to the percentile value (Pr_γ) and temperature (t) for scaling the feature attention values β in eq. 5 (main paper). "None" in Figure 7 refers to the absence of percentile and standard deviation scaling.

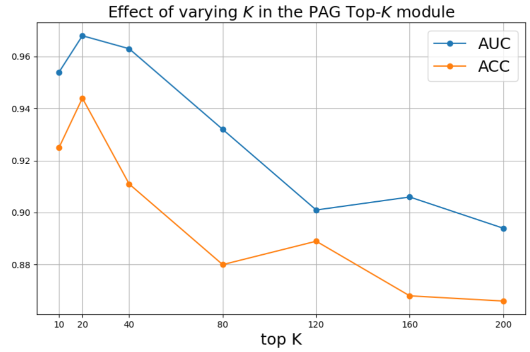


Figure 5. PAG Top-K module ablation

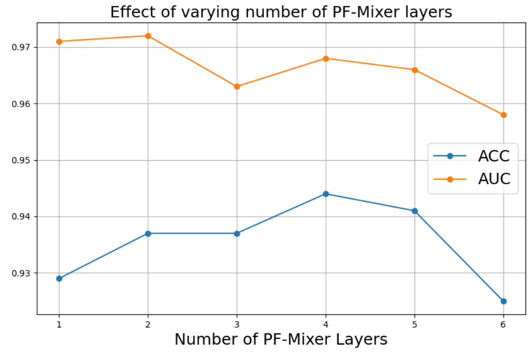


Figure 6. PF-Mixer module layers ablation

Pr_γ controls the percentage of features (d) that have a positive value before being fed to the sigmoid activation in eq. 5 (main paper). The temperature parameter (t) determines the sharpness of this curve, with a high value indicating that most values deviate from zero before being fed to the sigmoid. Thus, having high Pr_γ and high t leads to a very sparse selection of features. Since our goal is to interpret the prediction of WSI, it is beneficial to explain the prediction in terms of the contribution of "few" most discriminative features. Note that the absence of Pr_γ scaling and/or low temperature allows the model to use a large number of features for its prediction; thus making it harder to interpret the predictions. Therefore, the main goal is to have higher values of Pr_γ and t, while maintaining a good SI-MIL performance.

In Figure 7, we can observe that having no Pr_γ scaling generally results in the best performance, whereas a very high value, such as $Pr_\gamma = 0.9$, performs poorly in most cases. We find that having a slightly lower $Pr_\gamma = 0.75$ and $t = 3$ establishes an optimal balance, by enforcing adequate sparsity while still performing efficiently.

Table 4. Results indicate the mean and standard deviation of 5-fold cross-validation on test set. All methods are trained with Additive ABMIL as base MIL. Int. denotes self-interpretability of a method.

		Lung		BRCA		CRC	
	Int.	Acc.	AUC	Acc.	AUC	Acc.	AUC
IN ViT-S	✗	0.859 ± 0.014	0.919 ± 0.004	0.929 ± 0.011	0.967 ± 0.005	0.891 ± 0.013	0.898 ± 0.018
RetCCL	✗	0.860 ± 0.008	0.935 ± 0.003	0.929 ± 0.011	**0.976 ± 0.001**	0.889 ± 0.015	0.891 ± 0.047
CTransPath	✗	**0.904 ± 0.003**	**0.967 ± 0.002**	0.920 ± 0.023	0.974 ± 0.002	**0.906 ± 0.010**	**0.897 ± 0.023**
DINO ViT-S	✗	0.896 ± 0.003	0.957 ± 0.003	**0.937 ± 0.012**	0.974 ± 0.005	0.904 ± 0.006	**0.897 ± 0.014**
PathFeat	✗	0.830 ± 0.015	0.888 ± 0.009	0.885 ± 0.014	0.950 ± 0.005	**0.886 ± 0.016**	0.818 ± 0.031
PathFeat w/o $H(\cdot)$	✓	0.767 ± 0.018	0.837 ± 0.016	0.889 ± 0.012	0.914 ± 0.003	0.853 ± 0.013	0.720 ± 0.044
2-stage training	✓	0.865 ± 0.007	0.932 ± 0.009	0.908 ± 0.017	0.924 ± 0.019	0.876 ± 0.020	0.862 ± 0.036
SI-MIL (ours)	✓	**0.884 ±0.018**	**0.941 ±0.009**	**0.944 ±0.028**	**0.968 ±0.012**	0.884 ±0.017	**0.910 ±0.018**
Ablation study of SI-MIL components							
w/o PAG Top-K	✓	0.859 ±0.009	0.936 ±0.011	0.915 ±0.023	0.922 ±0.026	0.876 ±0.022	0.869 ±0.024
w/o KD	✓	0.853 ±0.010	0.915 ±0.007	0.932 ±0.016	0.951 ±0.024	0.878 ±0.024	0.830 ±0.039
w/o PAG Top-K & KD	✓	0.857 ±0.005	0.924 ±0.005	0.915 ±0.009	0.899 ±0.013	0.879 ±0.022	0.858 ±0.036

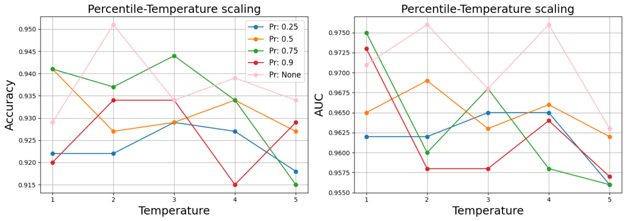


Figure 7. β scaling ablation

12. SI-MIL ablation studies: components

In Table 1 (main paper), we demonstrate the variations in performance when ablating different components of SI-MIL. Similarly, in Table 5, we present additional experiments on the impact of different SI-MIL components.

We observed that omitting the *Feature Attention module* $A^f(\cdot)$ results in better performance compared to using it without the *PF-Mixer* network $Mix(\cdot)$, though both scenarios underperform relative to the proposed SI-MIL. This indicates that $A^f(\cdot)$, which softly selects features, requires contextualization among the patches and features before highlighting or attenuating specific features within this module. Without appropriate contextualization, processing each feature row in matrix M^T independently leads to suboptimal decisions by $A^f(\cdot)$ and reduces performance.

We further investigate the necessity of deep features in SI-MIL. For this purpose, we substituted deep features with PathExpert features in the conventional MIL branch, thereby using the same PathExpert features in both SI-MIL branches. As shown in Table 5, the performance declines with or without $\mathcal{L}_{KD}$ when replacing deep features, underscoring the importance of employing potent deep features to guide the *Self-Interpretable* branch in SI-MIL.

Table 5. Results indicate the mean of 5-fold cross-validation on test set. All methods are trained with Additive ABMIL as the base MIL. Int. denotes self-interpretability of a method.

		Lung		BRCA		CRC	
	Int.	Acc.	AUC	Acc.	AUC	Acc.	AUC
PathFeat	✗	0.830	0.888	0.885	0.950	**0.886**	0.818
PathFeat w/o $H(\cdot)$	✓	0.767	0.837	0.889	0.914	0.853	0.720
2-stage training	✓	0.865	0.932	0.908	0.924	0.876	0.862
SI-MIL (ours)	✓	**0.884**	**0.941**	**0.944**	0.968	0.884	**0.910**
Ablation study of SI-MIL components							
w/o $A^f(\cdot)$	✓	0.853	0.935	0.939	**0.981**	0.871	0.857
w/o $Mix(\cdot)$	✓	0.838	0.915	0.925	0.953	0.866	0.863
w/ PathFeat only	✓	0.863	0.936	0.911	0.942	0.876	0.836
w/ PathFeat only & w/o KD	✓	0.847	0.911	0.911	0.945	0.853	0.781

13. Dataset contribution details (additional)

We contribute a unique comprehensive dataset aimed at enhancing interpretability and reproducibility in MIL research. The dataset encompasses nuclei maps and PathExpert features for over 2,200 WSIs. SI-MIL-generated patch-feature importance reports will also be made available for representative slides. It covers multiple organs and cancer types, including Lung (lung adenocarcinoma vs. lung squamous cell carcinoma), Breast (invasive ductal vs. invasive lobular carcinoma), and Colon (low vs. high mutation). This diverse collection facilitates in-depth studies across various cancer types, providing a valuable resource for advancements in interpretable MIL methodologies.

Successful translation of AI tools to the clinic hinges upon the interpretability and trustworthiness of the tools. This dataset will serve as a critical asset for both the medical vision and digital pathology communities, facilitating the exploration of new research directions in the development of interpretable AI techniques for computational pathology. A significant obstacle in digital pathology research has been the intensive resource requirements for extracting features that possess clear geometric and physical significance, and which are interpretable by pathologists. The dataset creation involved analyzing gigapixel WSIs at 40× magnification, leveraging HoVer-Net [7] for cell segmentation and classification, followed by extracting PathExpert features and feature importance scores detailed in Sec. 3.3 (main paper) and Sec. 5 (main paper), respectively. Processing each WSI required $\sim$2 hours, divided between GPU-based cell map prediction and CPU-based PathExpert feature extraction. Employing three RTX 8000 GPUs and a 40-core CPU with 500GB RAM, the total processing amounted to $\sim$4400 hours ($\approx$60 days). We provide the comprehensive set to enable further research.

In view of $\sim$2 TB memory foorprint of HoVer-Net nuclei maps and the processed PathExpert features, we intend to host this dataset on TCIA Analysis Results, akin to other popular preprocessed datasets [8, 13].

The dataset will be released under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0). It permits the sharing, copying, and redistribution in any medium or format, as well as adaptation, remixing, transforming, and building upon the material for non-commercial purposes. Appropriate credit must be given, a link to the license must be provided, and any changes made should be indicated.

14. Local interpretability analysis (additional)

Here, we present additional predictions (refer to Sec. 5 (main paper)) for WSIs from other datasets, *i.e.*, TCGA-Lung and TCGA-CRC. Please note that the predictions for all WSIs in the evaluated datasets will be released as part of the contributed dataset. Qualitative patch-feature importance reports for TCGA-Lung and TCGA-CRC are illustrated in the upper and lower half of Figure 8, respectively.

15. Global interpretability analysis (addition)

Here, we present global interpretability analysis (refer to Sec. 5 (main paper)) for patches from the test set WSIs. We include only the WSIs that were correctly predicted by both the conventional MIL and SI-MIL, to ensures a fair comparison, as described in Sec. 9. Cohort-level interpretation for TCGA-Lung and TCGA-CRC are illustrated in the upper and lower half of Figure 9, respectively.

16. Top-K comparative analysis

In this section, we demonstrate how the *Self-Interpretable* branch of SI-MIL tames the patch attention map of conventional deep MIL. Specifically, Figures 10 and 11 compare the spatial attention maps generated after the training of conventional MIL (*i.e.*, without PathExpert features) and SI-MIL, which integrates both the conventional MIL and *Self-Interpretable* branches. We proceed to visualize the top $K = 20$ patches from both MIL methods, categorizing them into groups based on whether they are common or distinct between the methods.

In Figure 10, we contrast the top $K = 20$ patch selection of our Self-Interpretable MIL (SI-MIL) with conventional MIL in analyzing invasive ductal carcinoma (IDC) WSIs. SI-MIL and conventional MIL share 6 out of 20 patches, but differ in the remaining 14. While conventional MIL often chooses patches near the dermis, featuring IDC with smooth connective areas and occasionally normal glands, SI-MIL targets patches indicative of malignancy, marked by malignant cancerous ducts with large, distorted nuclei. This difference, especially evident in the unique patches of SI-MIL, underscores its focus on diagnostically relevant areas like malignant glands with compressed lumens and hyperchromatic nuclei, contrasted against the tissue highlighted by conventional MIL. SI-MIL's emphasis is on patches comprising 70-80% of malignant features, including dense pink-colored cancer-associated stroma, aligning with its goal of accurate diagnosis.

In the context of invasive lobular carcinoma (ILC), early detection is crucial due to its rapid spread and poor long-term survival outcomes, necessitating clear differentiation from invasive ductal carcinoma (IDC). In Figure 11, we observe an absence of common patches between the top $K = 20$ attended patches of both MIL methods. Our method, in contrast to conventional MIL, distinctively focuses on invasive single file chains, often found at the periphery of the tumor bulk or the invasive front, which are more characteristic of ILC. This is in contrast to the conventional MIL's emphasis on patches with high cellularity within the tumor bulk. The rapid spread of lobular cancer is evident in its infiltration through various tissues, and unlike IDC, which often presents as glandular structures with clear separations between tumorous and connective nuclei, ILC is characterized by discohesive arrangements, leading to single file patterns with a notable mixing of tumor nuclei with connective nuclei.

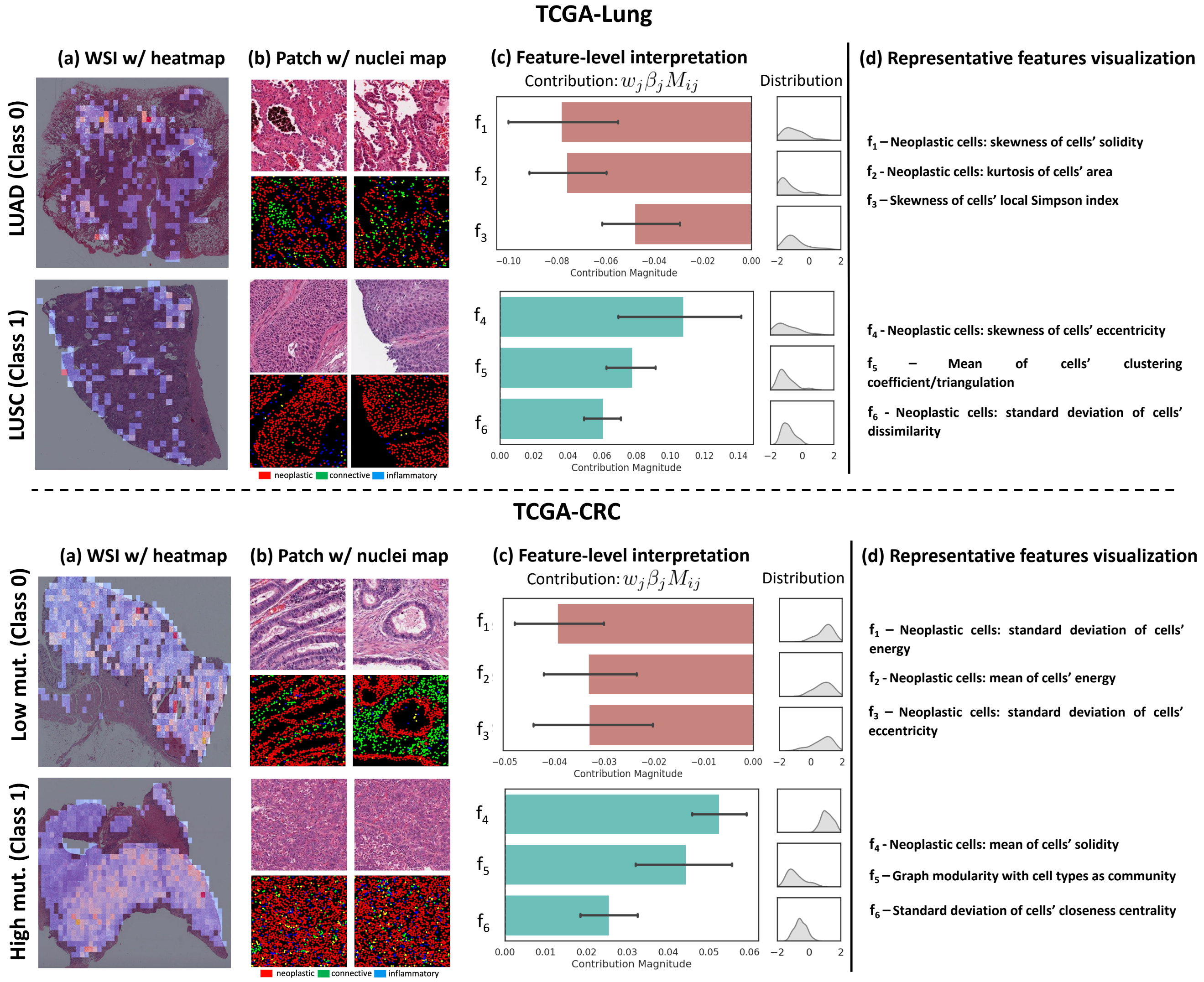


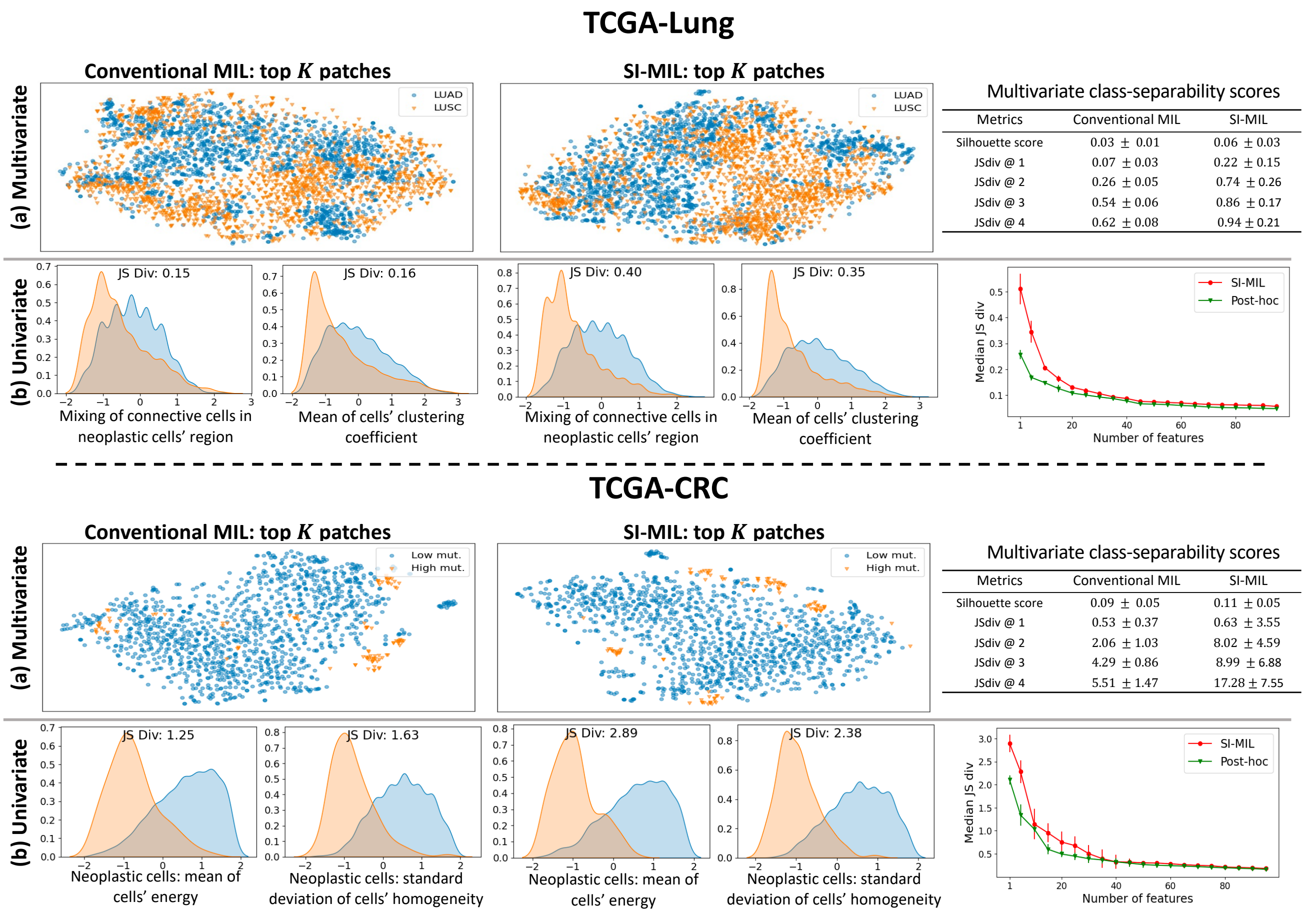
Figure 8. **Qualitative Patch-Feature importance report:** In (a) and (b), we present WSIs with overlaid attention heatmaps and the top two patches, along with their nuclei maps. In (c), we demonstrate the mean contribution magnitude of select representative features across the top K patches employed in the Self-Interpretable branch. Additionally, we display a feature density plot that quantifies the distribution of features within the K patches. For brevity, we omit the y-axis. Given that these features are normalized, a curve leaning towards the right indicates higher/positive values, while one towards the left signifies lower/negative values, depending on the feature. Finally, in (d), we illustrate, the description of representative features in (c).

17. Hand-crafted PathExpert feature extraction

In Sec. 3.3 (main paper), we briefly discussed the categories of handcrafted PathExpert features, such as *Morphometric* and *Spatial distribution* properties. This section provides a detailed description of these features, accompanied by visualizations to elucidate the significance of their geometrical and physical meaning in computational pathology. We further refine these features into three categories, based on the studies from which they were adopted: *Morphometric properties* (from FLocK [11]), *Graph-based Social Network Analysis* [16], and *Spatial heterogeneity properties* [12].

17.1. Feature categories

We employ HoVer-Net [7] to segment and classify nuclei in each WSI patch p_i, using the model trained on PanNuke. The classified nuclei include Neoplastic epithelial, Connective, Inflammatory, Necrosis, and Non-neoplastic epithelial classes. Subsequently, image-processing tools are used to quantify the properties and spatial distribution of nuclei in



Multivariate class-separability scores (TCGA-Lung)

Metrics	Conventional MIL	SI-MIL
Silhouette score	0.03 ± 0.01	0.06 ± 0.03
JSdiv @ 1	0.07 ± 0.03	0.22 ± 0.15
JSdiv @ 2	0.26 ± 0.05	0.74 ± 0.26
JSdiv @ 3	0.54 ± 0.06	0.86 ± 0.17
JSdiv @ 4	0.62 ± 0.08	0.94 ± 0.21

Multivariate class-separability scores (TCGA-CRC)

Metrics	Conventional MIL	SI-MIL
Silhouette score	0.09 ± 0.05	0.11 ± 0.05
JSdiv @ 1	0.53 ± 0.37	0.63 ± 3.55
JSdiv @ 2	2.06 ± 1.03	8.02 ± 4.59
JSdiv @ 3	4.29 ± 0.86	8.99 ± 6.88
JSdiv @ 4	5.51 ± 1.47	17.28 ± 7.55

Figure 9. **Cohort-level Interpretation:** Separability of top K patches of WSIs across classes in the PathExpert feature space. Multivariate and Univariate analyses depict that the top K patches selected by SI-MIL and their PathExpert features are more separable.

each patch. Next we provide a detailed description of each feature within these categories.

17.1.1 Morphometric properties

In a patch p_i, we extract 10 morphometric properties for each segmented nucleus as outlined in Table 6. To represent the entire patch, these nuclei-level features are aggregated using 4 statistical properties: mean, standard deviation, skewness, and kurtosis. This aggregation is performed separately for each of the 5 nuclei classes identified by HoVer-Net. Additionally, the number of nuclei in each class is included as a feature.

Consequently, this results in a total of 205 patch-level aggregated morphometric properties: 10 × 4 × 5 for the morphometric properties aggregated across the four statistics and five nuclei classes, plus five for the count of nuclei in each class.

Group	Feature
Shape	Area Eccentricity Roundness Orientation
Morphology	Mean of Intensity Standard Deviation of Intensity Contrast of Texture Dissimilarity of Texture Homogeneity of Texture Energy of Texture

Table 6. Description of extracted morphometric properties for each segmented nuclei

In computational pathology, these 205 morphometric properties from each patch p_i provide a holistic tissue profile. These features encapsulate key morphological characteristics of nuclei, crucial for pathological assessment. By employing statistics like mean, standard deviation, skewness, and kurtosis, we gain insights into the variability, asymmetry, and tailedness of the nuclei's morphological properties within each patch. Separately analyzing these features for each of the five nuclei classes as identified by HoVer-Net enriches the model's understanding of the tissue heterogeneity and cellular composition. Additionally, counting nuclei per class quantifies cellular composition, further enriching the diagnostic value in computational pathology.

17.1.2 Graph-based Social Network Analysis

In a patch p_i, we construct a graph based on the centroids of nuclei and quantify the properties of this network. Drawing inspiration from [16], we initially create a k-nearest neighbor graph (with $k = 6$) using the centroid locations of each segmented nucleus, irrespective of their classes. Subsequently, we extract 4 traditional social network analysis properties for each nucleus, as detailed in Table 7. This is followed by statistical aggregation to patch-level using mean, standard deviation, skew, kurtosis, and max. This results in total 20 aggregated Social Network features.

Feature
Degree
Degree centrality
Clustering coefficient
Closeness centrality

Table 7. Description of extracted social network analysis properties for each nuclei

The Degree and Degree Centrality metric in our study provides insight into the number of direct connections a nucleus has within the tissue network, illuminating its level of interaction. This is pivotal in understanding nuclei communication and behavior in various pathological states. The Clustering Coefficient is another key measure, offering insights into the extent of interconnectivity among a nucleus's neighbors. This can reveal localized nuclei clusters, a feature often observed in certain pathological conditions. Lastly, Closeness Centrality assesses the average shortest distance from a nucleus to all others, aiding in identifying nuclei that are central or isolated within the tissue architecture. This comprehensive analysis of nuclei organization and interaction patterns through these SNA features is crucial for an in-depth understanding of the tissue's pathology.

17.1.3 Spatial Heterogeneity properties

This feature group goes beyond analyzing just the centroids of nuclei; it also incorporates their classes to assess the spatial heterogeneity of various nuclei communities within a patch [12]. Heterogeneity is quantified at both global and local levels in each patch.

Global level: A range of entropy-based descriptors and k-function metrics are utilized, examining all segmented nuclei to evaluate the uniformity versus randomness in their spatial distribution. These global heterogeneity descriptors are listed in Table 8.

Local level: Following the methodology in [12], we construct a k-nearest neighbor graph (with $k = 6$) using the nuclei. For each nucleus, entropy and interaction-based properties are extracted, focusing on immediate neighbors. A local interaction-score is then aggregated at the patch level, as per the process in [12]. Additionally, skewness of entropy property distribution across nuclei is computed. This skewness metric discerns whether most nuclei have lower, medium, or higher entropy values, thus offering a detailed view of cellular interactions and complexity. This local-level approach highlights the intermixing of different nuclei communities, taking into account their spatial relationships, an aspect overlooked by global entropy-based descriptors. These local-level features are enumerated in Table 9.

This results in total 21 Spatial Heterogeneity features (9 Global and 12 Local). For an in-depth explanation and visualization of these features, we direct readers to the seminal work by [12], which extensively explores these methodologies and their implications.

Group	**Feature**
Global Entropy	Global Shannon index Global Simpson index Global max entropy Global Richness (number of cell-types present) Graph modularity with cell types as community
k-function	Neoplastic cells: k-function at radius 224 pixels Neoplastic cells: k-function at radius 448 pixels Neoplastic cells: k-function at radius 672 pixels Neoplastic cells: k-function at radius 896 pixels

Table 8. Global Spatial Heterogeneity Descriptors

Group	**Feature**
Local Entropy [12]	Skewness of cells' local Shannon index Skewness of cells' local Simpson index Skewness of cells' local max-entropy Skewness of cells' local richness
Local Interaction score [12]	Mixing of neoplastic cells in inflammatory cells' region Mixing of neoplastic cells in connective cells' region Mixing of neoplastic cells in necrosis cells' region Mixing of neoplastic cells in non-neoplastic epithelial cells' region Mixing of inflammatory cells in neoplastic cells' region Mixing of connective cells in neoplastic cells' region Mixing of necrosis cells in neoplastic cells' region Mixing of non-neoplastic epithelial cells in neoplastic cells' region

Table 9. Local Spatial Heterogeneity Descriptors

For instance, clustered arrangements of different nuclei communities typically result in lower local-level entropy, as the neighboring nuclei are mainly from the same class. Conversely, intermixed arrangements lead to higher local-level entropy due to the diversity of neighboring nuclei classes. However, at the global level, these differing arrangements may yield similar entropy values if the overall count of each nuclei class remains constant, despite their distinct spatial distributions. This highlights the importance of analyzing spatial heterogeneity at both local and global levels to capture the full complexity of cellular arrangements in tissue pathology.

17.2. Normalization

In this study, we implemented a two-step normalization process for all handcrafted PathExpert features. The first step addresses potential inaccuracies in nuclei segmentation and classification by HoVer-Net, using a binning operation. Each feature within a patch is assessed based on its percentile relative to other patches in the training split of a WSI task. These percentiles are then categorized into 10 discrete bins, ranging from the 0-10th to the 90-100th percentile, effectively shifting the scale of features from absolute values to a relative range from very-low to very-high. This approach transforms feature values into a robust and interpretable format across different patches. The second step involves mean and standard deviation normalization, once again using the training split data. This step centers the data around zero, optimizing it for effective processing by neural networks.

We emphasize that the normalization process alters the scale of the features. For instance, the skewness properties listed in Table 9 would typically be near 0, negative, or positive in their absolute scale. However, after mean-standard deviation and binning normalization, the scale of skewness may shift, with a 0 skew potentially appearing on either the positive or negative side. Hence, for a more accurate interpretation of our predictions in the local slide-level interpretable predictions (refer to Figure 3 (main paper)), it is crucial to consider this scaling effect. Readers should interpret the features as being generally in the lower or higher range and then conceptually approximate these back to their absolute scale. This approach ensures a more nuanced understanding of the predictions post-normalization.

17.3. Feature Visualization

In the following figures (Figures 12 to 16), we present a visual exploration of some representative features by showcasing patches with low and high values of these features. Each figure is accompanied by a detailed caption that elucidates the feature in the context of the patches, providing insights into what constitutes low and high values with respect to that specific feature. This visual representation aids in understanding the impact of these features on the tissue's pathology and offers a deeper perspective on how they manifest in different patches.

Note that the terms 'cell' and 'nucleus' are used interchangeably. However, since the imaging modality is H&E, all the features actually pertain to nuclei.

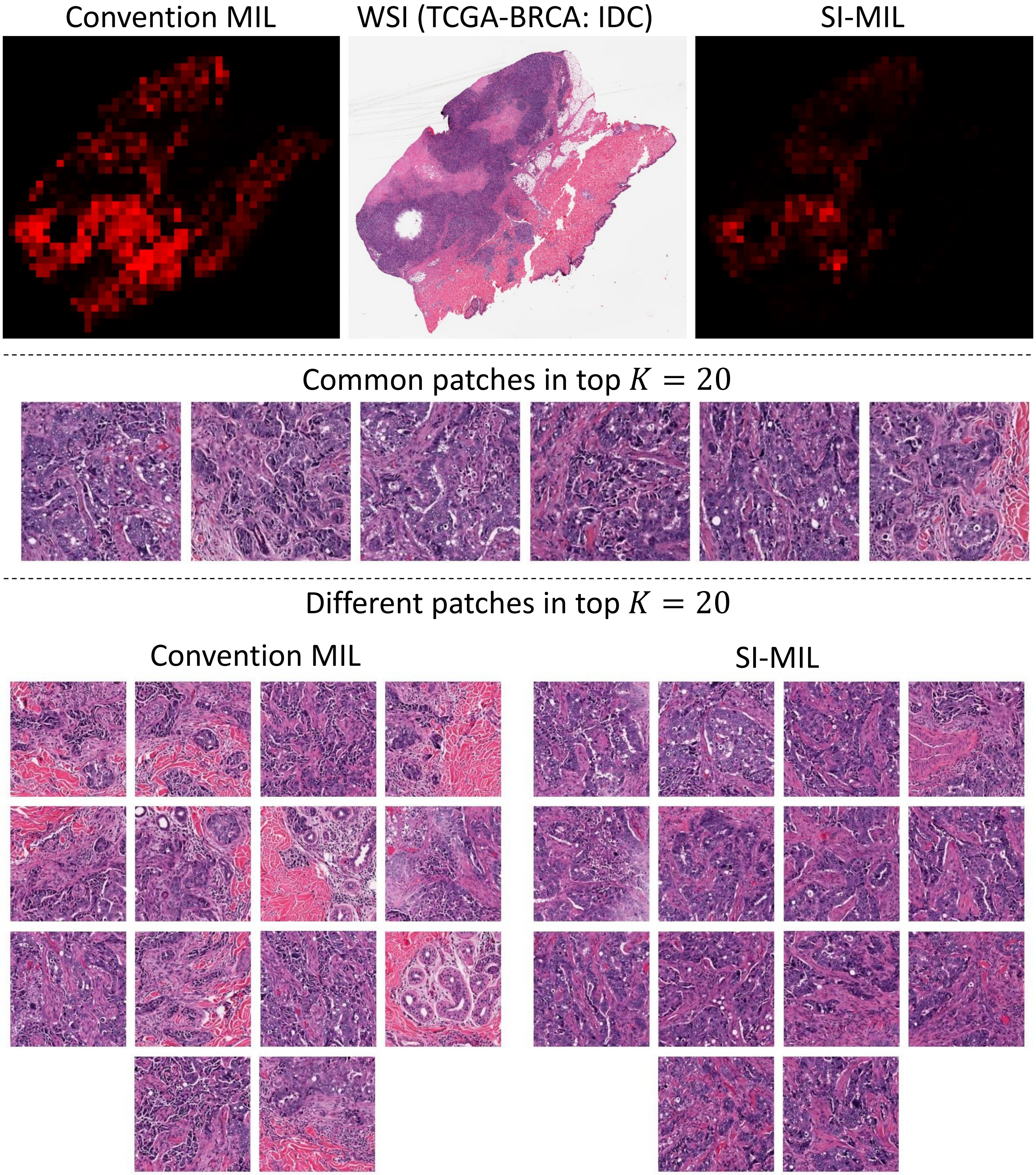

Convention MIL
WSI (TCGA-BRCA: IDC)
SI-MIL
Common patches in top $K = 20$
Different patches in top $K = 20$
Convention MIL
SI-MIL

Figure 10. TCGA-BRCA Invasive Ductal Carcinoma (IDC) sample. Refinement of the patch attention map by the Self-Interpretable branch, transitioning from conventional MIL to SI-MIL.

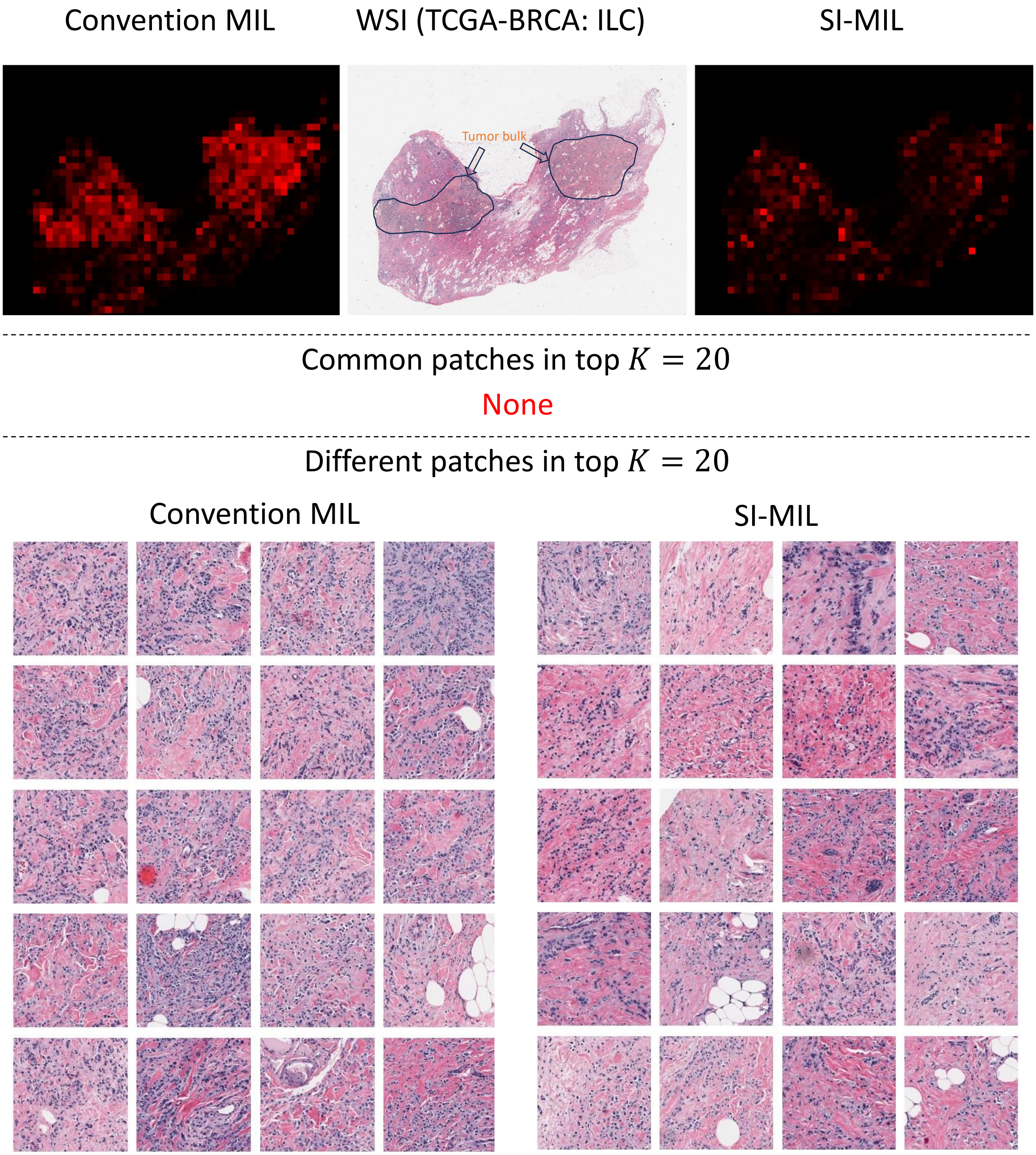


Figure 11. TCGA-BRCA Invasive Lobular Carcinoma (ILC) sample. Refinement of the patch attention map by the Self-Interpretable branch, transitioning from conventional MIL to SI-MIL.

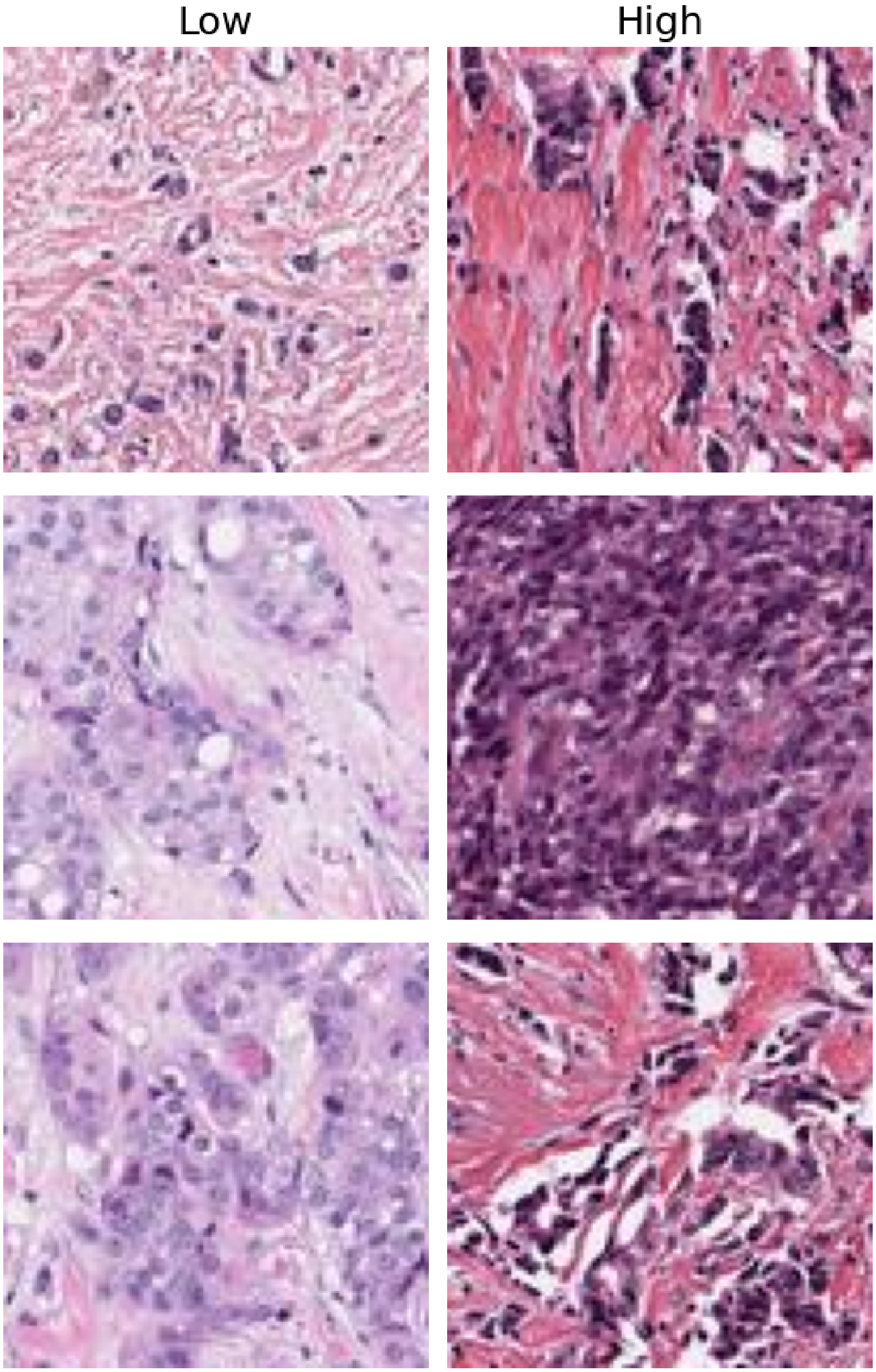


Figure 12. **Neoplastic Cells: Mean of Eccentricity.** As illustrated, in the patches under the column named ‘Low’, there are round cancer cells (the larger ones), whereas on the right side, under the column named ‘High’, elliptical cells are present, indicating a higher mean of eccentricity. In histopathology, this feature refers to the average deviation of cancer cells from a perfect circular shape. A higher mean eccentricity, as observed in the ‘High’ column, suggests more elliptical cells, often associated with more aggressive or advanced cancer forms.

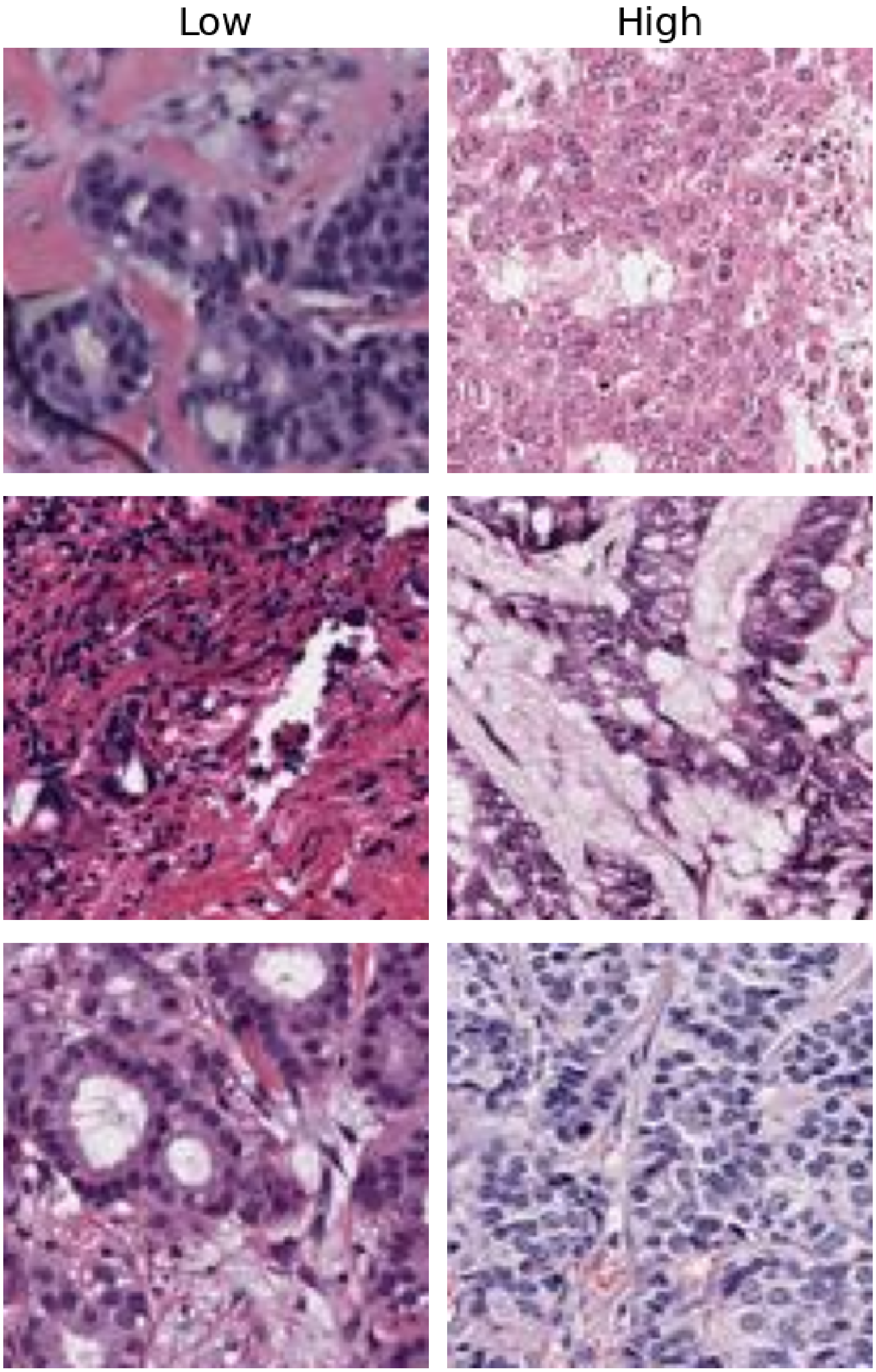


Figure 13. **Neoplastic Cells: Mean of Intensity Standard Deviation.** As illustrated, in the patches under the column named ‘Low’, there are cancer cells (the larger ones) with uniform intensity, thus the standard deviation is low for each cell, leading to a low mean intensity standard deviation. Whereas on the right side, under the column named ‘High’, the cells exhibit anisochromasia, indicating a higher mean intensity standard deviation. In histopathology, this feature refers to the average deviation of cancer cells from homogeneous intensity. A higher value of this feature, as observed in the ‘High’ column, suggests more anisochromasia, often associated with more aggressive or advanced cancer forms.

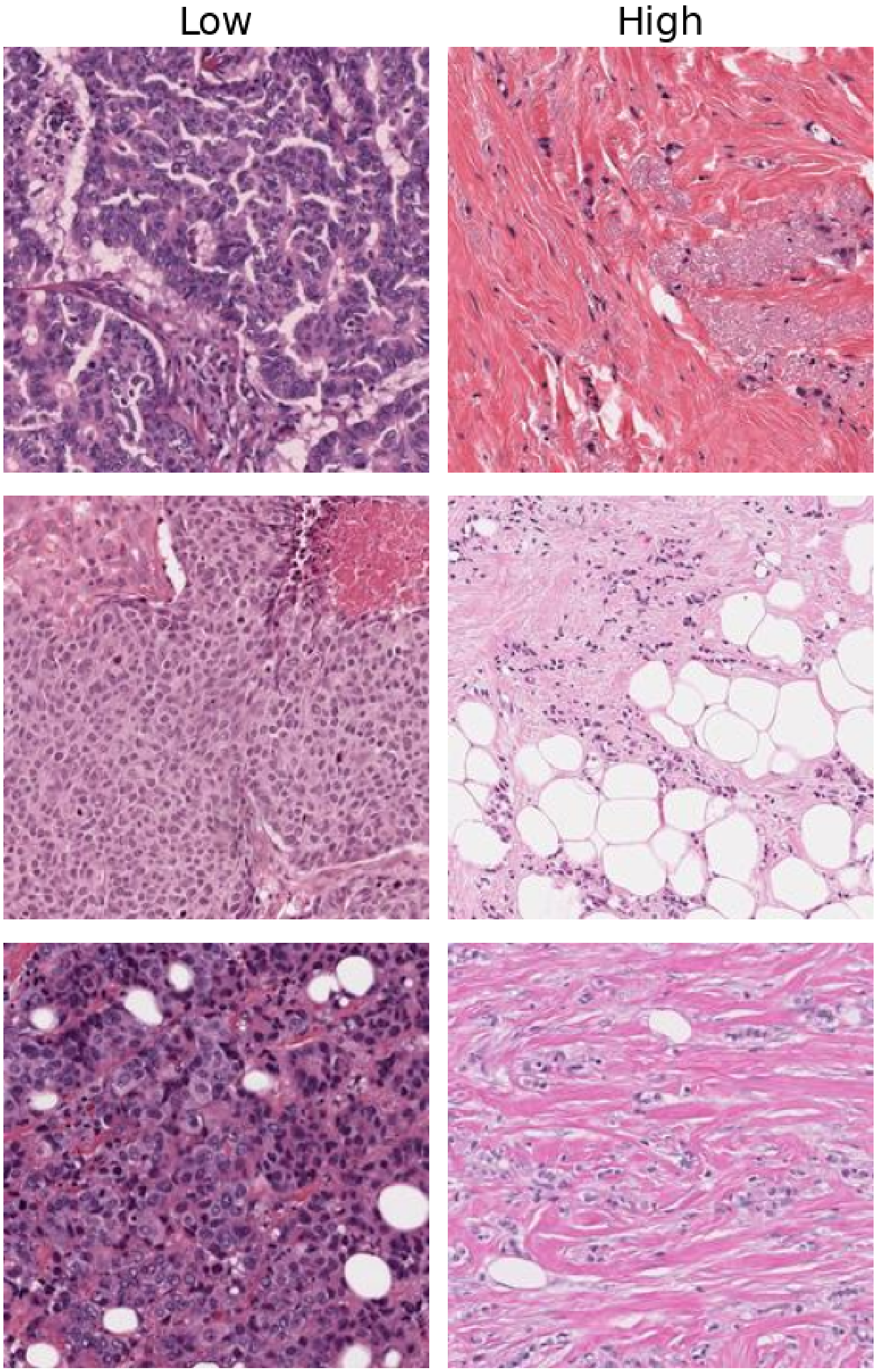


Figure 14. **Standard Deviation of Cells' Degree.** As illustrated, in the patches under the column named 'Low', there is a homogeneous distribution of cells of all types. For this feature, we first construct a k-nearest neighbor graph from cells' centroid and then calculate the cell's degree for each cell. Therefore, a homogeneous distribution leads to each cell having a similar degree, resulting in a lower value of standard deviation. Whereas on the right side, under the column named 'High', the cells are much more randomly distributed (disorganized), with grouping in some areas and sparse cells in others. This leads to some cells having a higher degree and others lower, resulting in a high standard deviation of cells' degree in a patch. In histopathology, this feature loosely refers to cohesive versus non-cohesive or homogeneous versus heterogeneous distribution in a spatial context.

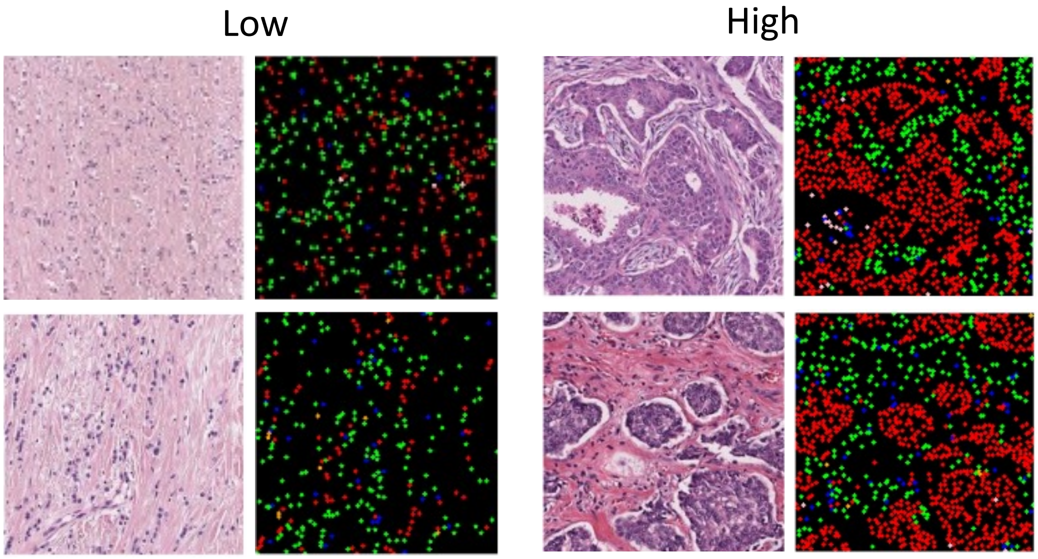


Figure 15. **Graph Modularity with Cell Types as Community.** As illustrated, in the patches under the column named 'Low', cancer cells (in red) co-occur in close spatial proximity with other cell types such as connective (in green) and inflammatory cells (in blue). This results in interconnections among different cell classes when constructing a graph for this feature, leading to low graph modularity. Whereas on the right side, under the column named 'High', cells of different classes/communities are more distinctly separated and grouped, resulting in more connections within the same community in the k-nearest neighbor graph, leading to higher graph modularity. In histopathology, this feature can serve as a proxy for distinguishing ductal versus single file line patterns in IDC versus ILC classification in TCGA-BRCA.

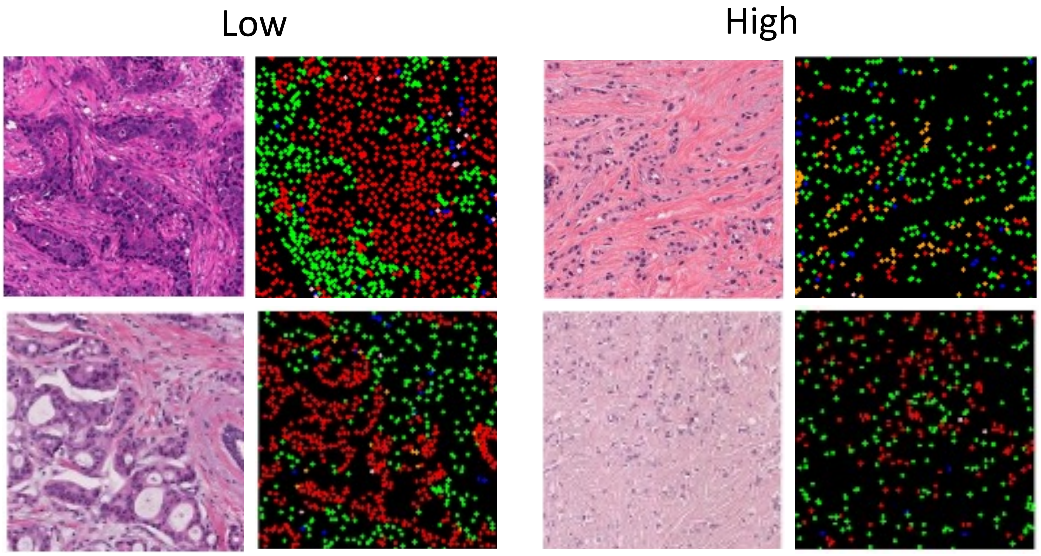


Figure 16. **Infiltration of Connective Cells in Neoplastic Cells' Region.** In contrast to the scenario presented in Figure 15, the 'High' column patches display cancer cells (colored in red) closely intermingled with connective cells (colored in green). This proximity results in more interactions between these cell types in the graph-based analysis of this feature, leading to a marked increase in the infiltration of connective cells within the neoplastic area. On the other hand, in the 'Low' column, there is a clearer segregation and clustering of the two cell classes, manifesting in reduced connectivity between them in the k-nearest neighbor graph, and consequently, lower levels of infiltration. In histopathological analysis, this characteristic can be instrumental in differentiating ductal cancers, which show minimal infiltration by other cell types, from invasive patterns characterized by a significant presence of connective cells within the neoplastic areas. This explanation is also applicable to features like the Infiltration of Inflammatory Cells in Neoplastic Cells' Region.

References

[1] The cancer genome atlas (tcga) research network. https://www.cancer.gov/tcga. 1

[2] Mohsin Bilal, Shan E Ahmed Raza, Ayesha Azam, Simon Graham, Mohammad Ilyas, Ian A Cree, David Snead, Fayyaz Minhas, and Nasir M Rajpoot. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *The Lancet Digital Health*, 3(12):e763–e772, 2021. 1

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[4] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[7] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019. 4, 5

[8] Jayashree Kalpathy-Cramer, Andrew Beers, Artem Mamonov, Erik Ziegler, Rob Lewis, Andre Botelho Almeida, Gordon Harris, Steve Pieper, David Clunie, Ashish Sharma, et al. Crowds cure cancer: Data collected at the rsna 2017 annual meeting. *The Cancer Imaging Archive. DOI*, 10:K9. 4

[9] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 1

[10] Yang Liu, Nilay S Sethi, Toshinori Hinoue, Barbara G Schneider, Andrew D Cherniack, Francisco Sanchez-Vega, Jose A Seoane, Farshad Farshidfar, Reanne Bowlby, Mirazul Islam, et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer cell*, 33(4):721–735, 2018. 1

[11] Cheng Lu, Can Koyuncu, German Corredor, Prateek Prasanna, Patrick Leo, XiangXue Wang, Andrew Janowczyk, Kaustav Bera, James Lewis Jr, Vamsidhar Velcheti, et al. Feature-driven local cell graph (flock): new computational pathology-based descriptors for prognosis of lung cancer and hpv status of oropharyngeal cancers. *Medical image analysis*, 68:101903, 2021. 5

[12] Adriano Luca Martinelli and Maria Anna Rapsomaniki. Athena: analysis of tumor heterogeneity from spatial omics measurements. *Bioinformatics*, 38(11):3151–3153, 2022. 5, 7

[13] J Saltz, R Gupta, L Hou, T Kurc, P Singh, V Nguyen, D Samaras, KR Shroyer, T Zhao, R Batiste, et al. Tumor-infiltrating lymphocytes maps from tcga h&e whole slide pathology images [data set]. *Cancer Imaging Arch*, 2018. 4

[14] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022. 1

[15] Xiyue Wang, Yuexi Du, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Retccl: clustering-guided contrastive learning for whole-slide image retrieval. *Medical image analysis*, 83: 102645, 2023. 1

[16] Neda Zamanitajeddin, Mostafa Jahanifar, and Nasir Rajpoot. Cells are actors: Social network analysis with classical ml for sota histology image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 288–298. Springer, 2021. 5, 7