HAGE: Hierarchical Alignment Gene-Enhanced Pathology Representation Learning with Spatial Transcriptomics

Thao M. Dang, Haiqing Li, Yuzhi Guo, Hehuan Ma, Feng Jiang, Yuwei Miao, Qifeng Zhou, Jean Gao, and Junzhou Huang*

Department of Computer Science and Engineering, The University of Texas at Arlington, Arlington, TX 76019, USA
jzhuang@uta.edu

Abstract. Histopathology images capture tissue morphology, while spatial transcriptomics (ST) provides spatially resolved gene expression, offering complementary molecular insights. However, acquiring ST data is costly and time-consuming, limiting its practical use. To address this, we propose HAGE (**H**ierarchical **A**lignment **G**ene-**E**nhanced), a framework that enhances pathology representation learning by predicting gene expression directly from histological images and integrating molecular context into the pathology model. HAGE leverages gene-type embeddings, which encode relationships among genes, guiding the model in learning biologically meaningful expression patterns. To further improve alignment between histology and gene expression, we introduce a hierarchical clustering strategy that groups image patches based on molecular and visual similarity, capturing both local and global dependencies. HAGE consistently outperforms existing methods across six datasets. In particular, on the HER2+ breast cancer cohort, it significantly improves the Pearson correlation coefficient by 8.0% and achieves substantial reductions in mean squared error and mean absolute error by 18.1% and 38.0%, respectively. Beyond gene expression prediction, HAGE improves downstream tasks, such as patch-level cancer classification and whole-slide image diagnostics, demonstrating its broader applicability. To the best of our knowledge, HAGE is the first framework to integrate gene co-expression as prior knowledge into a pathology image encoder via a cross-attention mechanism, enabling more biologically informed and accurate pathology representations. https://github.com/uta-smile/gene_expression

Keywords: spatial transcriptomics · pathology images · multimodal contrastive learning · hierarchical alignment.

1 Introduction

Histopathology images offer rich morphological insights into tissue architecture and cellular features, while spatial transcriptomics (ST) provides location-specific gene expression data that can illuminate disease progression and patient

outcomes. However, acquiring ST data is costly and resource-intensive, whereas H&E-stained histology images are routinely generated in clinical settings [2]. This disparity motivates the development of methods to infer gene expression directly from histopathology. By leveraging paired pathology images and spatial expression spots, deep learning systems can extract visual features predictive of various molecular biomarkers, enabling more efficient molecular profiling without the need for expensive ST sequencing.

In typical ST data, each tissue spot corresponds to a patch in a whole slide image (WSI), providing expression measurements for thousands of genes. Deep learning has been explored to map histopathology to transcriptomic data. For instance, ST-Net [1] employs a CNN-based encoder for direct prediction, while HisToGene [2] and Hist2ST [3] enhance ST-Net by incorporating spatial relationships through Vision Transformers and graph-based models. THItoGene [4] further utilizes graph neural networks and dynamic convolutional networks. However, these methods rely on predefined spatial assumptions, which may not generalize well to heterogeneous tissues such as cancer, where gene expression patterns are highly variable [5]. Another research direction explores contrastive learning to align histology and gene expression representations in a shared latent space, such as BLEEP [5] and mclSTExp [6]. While these methods improve generalization, they often face data limitations when trained from scratch. Moreover, most existing methods treat gene targets as independent outputs, focusing solely on image-expression alignment and neglecting critical gene-gene relationships. Although these methods represent progress, a more biologically grounded framework is needed to fully integrate histology and transcriptomics.

Motivated by these challenges, we adopt the UNI [8] foundation model and introduce a hierarchical clustering strategy to efficiently align patches sharing morphological and molecular similarities. The use of foundation models for extracting image representations has been widely adopted in histology [9]. To move beyond purely visual features, we further strengthen our approach by integrating different modalities, as fusion has shown significant potential across a range of tasks and domains in bioinformatics [10–15]. Since co-expression correlations indicate functional associations, as genes in the same biological process are often co-regulated [16], we incorporate gene embeddings from Gene2Vec [17], trained on co-expression data, into the image encoder through cross-attention, capturing co-regulation patterns that better align histological features with gene expression. This enriched biological context helps capture "invisible" signals where microenvironments induce significant gene expression variability.

Our contributions are threefold: 1) We propose a novel patch-level framework that integrates gene co-expression into the pathology image encoder, leveraging gene co-activation patterns as biologically grounded prior knowledge to guide representation learning. This enables the encoder to attend to visual regions with awareness of gene-gene coordination, improving biological interpretability. 2) We introduce a hierarchical clustering strategy that aligns image-expression pairs across local and global levels, enhancing efficiency and predictive accuracy. 3) Comprehensive evaluations on six datasets across gene expression prediction,

patch-level cancer classification, and WSI-level cancer classification demonstrate state-of-the-art performance and more biologically grounded representations.

2 Methodology

2.1 Preliminaries

Our learning pipeline is illustrated in Figure 1. We denote the set of gene embeddings as $\mathbf{G} = \{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_{|G|}\}$ with $\mathbf{G} \in \mathbb{R}^{|G| \times d_g}$, retrieved from the Gene2Vec [17] library. Given histopathology image patch with features $\mathbf{x} \in \mathbb{R}^{d_i}$ extracted by a frozen UNI model, our Gene-informed Image Encoder (GE), $f_{GE}(\cdot)$, fuses the image features with $\mathbf{G}$ via cross-attention to yield gene-guided image embeddings: $\mathbf{z}^i \leftarrow f_{GE}(\mathbf{x}, \mathbf{G}) \in \mathbb{R}^d$. In parallel, the raw gene expression vector $\mathbf{y} \in \mathbb{R}^{d_e}$ is normalized and fed into a three-layer MLP Expression Encoder, $f_{EE}(\cdot)$, to produce expression embeddings: $\mathbf{z}^e \leftarrow f_{EE}(\mathbf{y}) \in \mathbb{R}^d$.

As stated in Section 1, our goal is to infer gene expression from images. Therefore, improving the alignment between modalities is crucial for accuracy. To this end, we align the paired embeddings $\{\mathbf{z}_i^i, \mathbf{z}_i^e\}_{i=1}^N$ in the local view. Specifically, we enforce cross-modality consistency using the CyCLIP [18] loss, denoted as $\mathcal{L}_{Cy}$, and ensure robust alignment using the well-known CLIP [19] loss. Due to space limitations, we only provide the formulation for the CyCLIP loss, and refer the reader to [19] for details on the CLIP loss. The CyCLIP loss is computed as:

$$\mathcal{L}_{Cy} = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\langle \mathbf{z}_j^i, \mathbf{z}_k^e\rangle - \langle \mathbf{z}_k^i, \mathbf{z}_j^e\rangle\right)^2 + \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\langle \mathbf{z}_j^i, \mathbf{z}_k^i\rangle - \langle \mathbf{z}_j^e, \mathbf{z}_k^e\rangle\right)^2, \quad (1)$$

with $\langle \cdot, \cdot \rangle$ represents the inner product.

In the global view, we apply k-means to raw expression $\mathbf{y}$ to form k_1 clusters. Within each cluster, we further cluster the corresponding $\mathbf{x}$ into k_2 subclusters. Let $\{\mathbf{c}_i^e, \mathbf{c}_i^i\}_{i=1}^{k_1 \times k_2}$ denote the centroids in the expression and image spaces, respectively; these centroid pairs are aligned using the CLIP loss [19], $\mathcal{L}_{CLGlobal}$.

A three-layer MLP Expression Predictor, $f_{Pred}(\cdot)$, then maps the gene-guided image embeddings to predicted expression vectors: $\hat{\mathbf{y}} \leftarrow f_{Pred}(\mathbf{z}^i)$. We train $f_{Pred}(\cdot)$ with a Pearson Correlation Coefficient loss:

$$\mathcal{L}_{PCC} = 1 - \frac{1}{|G|}\sum_{j=1}^{|G|}\frac{\sum_{i=1}^{N}\left(\hat{y}_{ij} - \overline{\hat{y}}_j\right)\left(y_{ij} - \overline{y}_j\right)}{\sqrt{\sum_{i=1}^{N}\left(\hat{y}_{ij} - \overline{\hat{y}}_j\right)^2 + \epsilon}\sqrt{\sum_{i=1}^{N}\left(y_{ij} - \overline{y}_j\right)^2 + \epsilon}}, \quad (2)$$

where $\overline{\hat{y}}_j$ and $\overline{y}_j$ are the mean of predicted expression and target expression for gene j across all N samples, respectively. While ϵ is a small constant for numerical stability. The total loss is defined as:

$$\mathcal{L} = \lambda_1\mathcal{L}_{Cy} + \lambda_2\mathcal{L}_{CLLocal} + \lambda_3\mathcal{L}_{CLGlobal} + \lambda_4\mathcal{L}_{PCC}. \quad (3)$$

At inference, given a patch features $\mathbf{x}$, we compute: $\hat{\mathbf{y}} = f_{Pred}(f_{GE}(\mathbf{x}, \mathbf{G}))$, thereby predicting the gene expression profile for each patch.

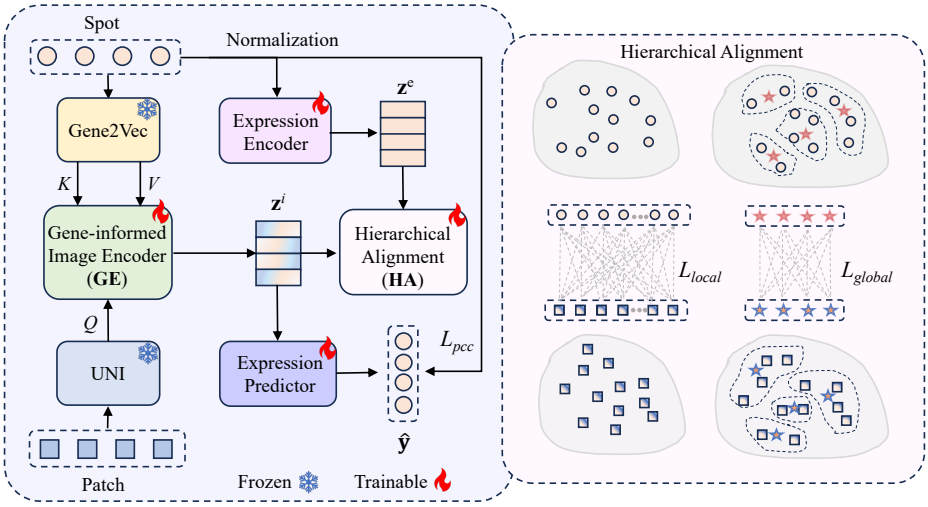

Fig. 1. HAGE predicts gene expression from pathology images by aligning expression profiles and image patches in a shared embedding space. 1) Given a list of interesting gene names, the corresponding gene embeddings retrieved from Gene2Vec are fused with image features from a frozen foundation model via GE. 2) Expression features are encoded through an MLP-based Expression Encoder. 3) HA aligns gene-guided image and expression embeddings at both local and global levels using contrastive learning. 4) The Expression Predictor predicts gene expression from gene-informed image features. At inference, given a WSI, the model outputs the predicted gene expression, while GE remains applicable to other downstream tasks.

2.2 Gene-informed Image Encoder

We design a gene-informed image encoder (**GE**) that merges patch-level image embeddings with gene embeddings derived from Gene2Vec. Traditional image encoders that lack explicit gene information merely regress expression values, without capturing gene identity or correlations. By integrating Gene2Vec embeddings, our encoder learns both individual gene features and gene–gene relationships, allowing it to align visual features with biological context.

We project the image features $\{\mathbf{x}_i\}_{i=1}^{N} \in \mathbb{R}^{d_i}$ and the gene embeddings $\{\mathbf{g}_j\}_{j=1}^{|G|} \in \mathbb{R}^{d_g}$ into a common space of dimension d using learnable projection matrices. Treating the visual features as queries and the gene features as keys and values, we employ a cross-attention mechanism [20] to fuse the two modalities. Standard residual connections and a two-layer MLP block are then applied to refine the fused representations. The overall transformation is given by: $\mathbf{z}_i^i \leftarrow f(\mathbf{x}_i) = f_{MLP}\Big(f_{Attn}(W_x\mathbf{x}_i, \{W_g\mathbf{g}_j\}_{j=1}^{|G|}) + W_x\mathbf{x}_i\Big)$, yielding gene-informed image embeddings that are enriched with biologically grounded information. Notably, the trained GE module is reusable in other downstream tasks beyond gene expression prediction, enabling broader applicability in biomedical applications (see Section 3.2 for details).

2.3 Hierarchical Alignment Module

The data for paired image patches and gene expression profiles is scarce, which makes it difficult for contrastive learning methods to achieve robust alignment from limited pairs alone. The proposed Hierarchical Alignment (**HA**) module addresses this issue by gathering additional pairs in two stages of clustering.

First, we apply k-means with $k = k_1$ to cluster raw gene expression vectors $\{\mathbf{y}_i\}_{i=1}^{N} \in \mathbb{R}^{d_e}$ into coarse clusters $\{C_p\}_{p=1}^{k_1}$, such that $C_p = \{i \,:\, \mathbf{y}_i \sim \text{cluster } p\}$. This step prevents the problem of grouping patches by appearance alone, since patches may share morphological features yet differ substantially in gene expression. For each coarse cluster C_p, we further cluster the corresponding image features $\{\mathbf{x}_i\}_{i \in C_p}$ using k-means with $k = k_2$, yielding subclusters $\{S_{p,q}\}_{q=1}^{k_2}$. For each subcluster $S_{p,q}$, we compute the centroids in both expression and image domains: $\mathbf{c}_{p,q}^{e} = \frac{1}{|S_{p,q}|} \sum_{j \in S_{p,q}} \mathbf{z}_j^{e}$ and $\mathbf{c}_{p,q}^{i} = \frac{1}{|S_{p,q}|} \sum_{j \in S_{p,q}} \mathbf{z}_j^{i}$. These centroid pairs $(\mathbf{c}_{p,q}^{e}, \mathbf{c}_{p,q}^{i})$ are used as additional alignment pairs. By first clustering based on gene expression, and then on morphology, this two-tier approach provides a global view of the data and adds more training pairs.

3 Experiments

3.1 Implementation Details

Datasets. For the gene expression prediction task, following the setting of related works [3, 4, 6], we select 32 breast samples and 12 skin samples from HER2+ [21] and cSCC [22] datasets, respectively. To assess the quality of the learned embeddings, we evaluate patch-level classification on PCAM [23] and SkinCancer [24] datasets. For WSI-level diagnosis, we use TCGA-BRCA [25] for cancer subtype classification (e.g., IDC and ILC) and the SLN-Breast [26] for positive or negative carcinoma prediction. Totally, six tasks are conducted.

Gene list selection. We apply a standard gene selection pipeline as described in [6]. We compile a list of the top 1,000 highly variable genes for each WSI and intersect these lists to identify common genes across all WSIs. We then refine the gene list by intersecting it with the Gene2Vec dictionary, yielding a final set of 771 genes for the breast dataset and 168 genes for the skin dataset.

Baselines and Evaluation metrics. For the gene expression prediction task, we compare our approach against six SOTA methods: ST-Net [1], HisToGene [2], Hist2ST [3], THItoGene [4], BLEEP [5], and mclSTExp [6]. To ensure fairness, we adopt the leave-one-out protocol used by previous studies [1, 6] and perform five experimental runs with different training and validation splits for each left-out subject. Following related works [3–6], performance is evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and Pearson Correlation Coefficient (PCC) metrics. In particular, we compute the PCC for all genes in the gene list as well as for the top 50 highly expressed genes (HEG).

Other downstream tasks. To ensure a fair comparison, we do not evaluate against the six baselines mentioned earlier. Instead, we report the performance of our gene-informed image encoder and UNIv1 [8]. Notably, none of the datasets used in our study overlap with the testing datasets, eliminating potential data leakage. We employ 5-fold cross-validation and assess performance using Accuracy, AUC, and Recall, reporting the mean and standard deviation.

3.2 Results

Gene expression prediction. After leaving one subject out, we split the remaining data into training and validation sets at a 4:1 ratio and run five trials per subject. Training lasts 15 epochs for HER2+ and 100 for cSCC, with early stopping after 5 epochs. Both breast and skin cancer experiments use same hyperparameters: $\lambda_1 = \lambda_4 = 10$, $\lambda_2 = \lambda_3 = 1$, $k_1 = 30$, $k_2 = 2$, and $d = 1024$. CLIP loss temperature is fixed at $\tau = 0.07$. We optimize with Adam (learning rate 10^{-4}, weight decay 10^{-5}) and StepLR (step size 10, decay factor $\gamma = 0.5$).

Table 1. Gene expression prediction performance of various methods.

Dataset	HER2+			
Methods	PCC (All) ↑	PCC (HEG) ↑	MSE ↓	MAE ↓
ST-Net [1]	0.0561 ± 0.017	0.0134 ± 0.013	0.5312 ± 0.008	0.6306 ± 0.011
HisToGene [2]	0.0842 ± 0.015	0.0711 ± 0.014	0.5202 ± 0.014	0.6422 ± 0.005
Hist2ST [3]	0.1443 ± 0.013	0.1849 ± 0.015	0.5135 ± 0.009	0.6087 ± 0.013
THItoGene [4]	0.1726 ± 0.018	0.2809 ± 0.013	0.5012 ± 0.011	0.5956 ± 0.009
BLEEP [5]	0.1873 ± 0.005	0.2909 ± 0.016	0.6015 ± 0.016	0.5824 ± 0.004
mclSTExp [6]	0.2304 ± 0.011	0.3866 ± 0.021	0.5897 ± 0.013	0.5813 ± 0.008
HAGE (ours)	**0.2489 ± 0.001**	**0.4458 ± 0.003**	**0.4830 ± 0.005**	**0.3606 ± 0.002**
Dataset	cSCC			
Methods	PCC (All) ↑	PCC (HEG) ↑	MSE ↓	MAE ↓
ST-Net [1]	0.0012 ± 0.022	0.0018 ± 0.015	0.6806 ± 0.006	0.6404 ± 0.003
HisToGene [2]	0.0771 ± 0.024	0.0919 ± 0.012	0.6805 ± 0.012	0.6234 ± 0.007
Hist2ST [3]	0.1838 ± 0.011	0.2175 ± 0.016	0.6748 ± 0.017	0.6107 ± 0.006
THItoGene [4]	0.2373 ± 0.009	0.2719 ± 0.012	0.6546 ± 0.006	0.6012 ± 0.019
BLEEP [5]	0.2449 ± 0.017	0.3122 ± 0.027	0.5163 ± 0.007	0.5399 ± 0.015
mclSTExp [6]	0.3235 ± 0.019	0.4261 ± 0.016	0.4302 ± 0.005	0.5208 ± 0.009
HAGE (ours)	**0.3397 ± 0.006**	**0.4607 ± 0.007**	**0.4248 ± 0.002**	**0.3296 ± 0.002**

Table 1 summarizes HAGE's performance against six SOTA methods on HER2+ and cSCC datasets. On HER2+, HAGE achieves a PCC of 0.2489 for all genes and 0.4458 for the top 50 HEG, improving by 8.0% and 15.3%, respectively, over the second-best method. It also attains an MSE of 0.4830 and MAE of 0.3606, reducing errors by 18.1% and 38.0% relative to mclSTExp.

On cSCC, HAGE achieves a PCC of 0.3397 (all genes) and 0.4607 (HEG), improving by roughly 5.0% and 8.1% over mclSTExp. While MSE improvement is modest, MAE decreases significantly by 36.7%. These results confirm HAGE's consistent advantage in expression prediction across cancer types.

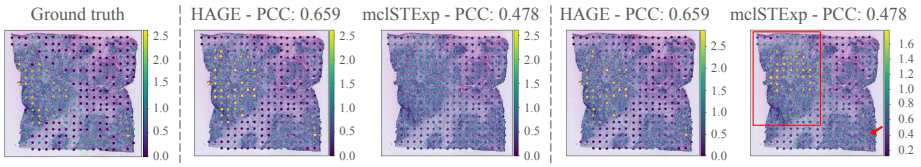

Fig. 2. Visualization of ITGB6 gene expression predictions. (Left) Ground truth for slide D2 in the HER2+ dataset, (middle) fixed scale, and (right) variable scale.

Figure 2 visualizes ITGB6 [27], a key diagnostic biomarker for breast cancer, particularly in the HER2+ dataset. High ITGB6 expression implies tumor presence; if the model can effectively localize such regions, it demonstrates potential for clinical interpretation. Our method achieves a PCC of 0.659, significantly outperforming the SOTA method, mclSTExp (0.478). In the variable scale setting, both HAGE and mclSTExp capture the overall trend, but the latter fails to accurately reflect expression levels at certain spots. In the fixed scale setting, which compares predictions to absolute ground truth values, mclSTExp struggles to identify tumor regions and distinguish signals from connective and invasive tumor tissues, whereas HAGE provides a clearer, more precise prediction.

Tumor classification. *Patch-level.* We evaluate our learned representations using a linear probing protocol for patch-level tumor classification. The classifier is a single linear layer that projects the input embeddings onto the target classes. Both the breast (e.g., PCAM) and skin (e.g., SkinCancer) linear probing tasks are trained for 10 epochs using cross-entropy loss. We report the average performance of our five models and the UNIv1 model.

Slide-level. SLN-Breast followed the preprocessing steps and implementation settings described in [28], while TCGA-BRCA was preprocessed using the default settings of CLAM [29]. We use the pretrained gene-guided encoder with the best performance in PCAM classification as the image encoder and adopt ABMIL [30] to aggregate patch-level features into slide-level embeddings. Despite using standard mechanisms and lightweight modeling choices, the integration of gene co-expression enhances downstream performance. Figure 3 shows that our gene-informed encoders consistently outperform UNIv1 in both patch-level and slide-level classification tasks on breast and skin datasets.

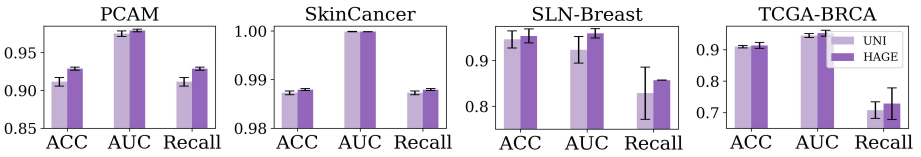

Fig. 3. Comparison of the gene-informed image encoder in HAGE and the UNIv1 image encoder on patch-level and slide-level classification across breast and skin datasets.

Table 2. Impact of each HAGE component on gene prediction on the HER2+ dataset.

Architecture	PCC (A) ↑	PCC (H) ↑	MSE ↓	MAE ↓
HAGE w/o both HA and GE (I)	0.2330	0.4131	0.5081	0.3547
HAGE w/o Gene-informed Encoder (II)	0.2338	0.4164	0.5076	0.3543
HAGE w/o Hierarchical Alignment (III)	0.2474	0.4410	0.4842	0.3613
HAGE with Single-layer cluster align. (IV)	0.2461	0.4390	0.4812	0.3603

3.3 Ablation Study

Impact of each component. Table 2 ablates two key components of HAGE: the gene-informed encoder (GE) and hierarchical alignment (HA). Removing both (I) significantly reduces performance, confirming that morphological features alone are insufficient for gene expression prediction. Retaining HA without GE (II), or GE without HA (III), improves performance over the baseline but still falls short of the full HAGE. While clustering reduces MSE and MAE for both HA and its variant, only HA consistently improves all metrics. Replacing the two-level hierarchical alignment with a single-layer cluster alignment (IV) degrades performance, suggesting that cluster formation impacts results more than simply adjusting their number (see Table 3 for details). These results emphasize the necessity of integrating gene-level knowledge into the image encoder and the effectiveness of hierarchical alignment in capturing molecular patterns.

Impact of cluster. We conduct a grid search over k_1 and k_2 to evaluate the robustness of the HA component. As shown in Table 3, performance remains relatively stable across a practical range of cluster settings (e.g., $k_1 \in \{10, 20, 30, 40\}, k_2 \in \{2, 3\}$), indicating that our method is not overly sensitive to moderate adjustments in these hyperparameters. While clustering facilitates the alignment of patches with similar molecular characteristics, robust performance across variations in cluster size is desirable, as large result changes under different configurations may suggest noise or biologically implausible groupings. In practice, we set $k_1 = 30$ and $k_2 = 2$ to balance performance and efficiency.

The primary factor influencing performance is how clusters are formed. We test an alternative approach that relies solely on gene expression, creating a single-layer clustering by setting $k_1 = 60$ and removing the subcluster step. Although this approach results in the same total number of centroids as our hierarchical structure, it yields lower performance (Table 2).

Table 3. Grid search on the impact of cluster settings on the HER2+ dataset.

	$k_2 = 2$				$k_2 = 3$			
Metrics	PCC (A)	PCC (H)	MSE	MAE	PCC (A)	PCC (H)	MSE	MAE
$k_1 = 10$	0.2510	0.4499	0.4856	0.3606	0.2480	0.4433	0.4832	0.3611
$k_1 = 20$	0.2480	0.4422	0.4817	0.3605	0.2485	0.4447	0.4862	0.3567
$k_1 = 30$	0.2503	0.4490	0.4823	0.3604	0.2463	0.4405	0.4833	0.3612
$k_1 = 40$	0.2493	0.4468	0.4841	0.3603	0.2501	0.4504	0.4826	0.3606

4 Conclusion

We introduce HAGE (Hierarchical Alignment Gene-Enhanced), a framework that integrates gene features and hierarchical clustering to infer spatial transcriptomic signals from histopathology images. By incorporating gene-specific information into the image encoder via cross-attention, HAGE learns representations that are both visually and biologically informative. Experiments on six datasets demonstrate its effectiveness, showing stronger correlations, lower error metrics, and adaptability across diverse tissues and clinical contexts. Overall, HAGE offers a robust, gene-aware pathology representation that enhances both expression inference and downstream task performance.

Acknowledgments. This work was partially supported by US National Science Foundation IIS-2412195, CCF-2400785, the Cancer Prevention and Research Institute of Texas (CPRIT) award (RP230363), the National Institutes of Health (NIH) R01 award (1R01AI190103-01) and Microsoft Accelerate Foundation Models Research (2024).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. B. He, L. Bergenstråhle, L. Stenbeck, A. Abid, A. Andersson, A. Borg, J. Maaskola, J. Lundeberg, J. Zou.: Integrating spatial gene expression and breast tumour morphology via deep learning. Nat Biomed Eng 4, 827–834 (2020). https://doi.org/10.1038/s41551-020-0578-x
2. Pang M, Kenong S, Li M.: Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. bioRxiv 2021.11.28.470212. doi: https://doi.org/10.1101/2021.11.28.470212
3. Y. Zeng, Z. Wei, W. Yu, R. Yin, Y. Yuan, B. Li, Z. Tang, Y. Lu, Y. Yang: Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. Brief Bioinform, https://doi.org/10.1093/bib/bbac297
4. Y. Jia, J. Liu, L. Chen, T. Zhao, Y. Wang: THItoGene: a deep learning method for predicting spatial transcriptomics from histological images. Briefings in Bioinformatics, vol 25, 2024, https://doi.org/10.1093/bib/bbad464
5. R. Xie, K. Pang, S. W. Chung, C. T. Perciani, S. A. MacParland, B. Wang, G. D. Bader: Spatially resolved gene expression prediction from histology images via bimodal contrastive learning. In: Advances in Neural Information Processing Systems
6. W. Min, Z. Shi, J. Zhang, J. Wan, C. Wang: Multimodal contrastive learning for spatial gene expression prediction using histology images. Briefings in Bioinformatics, vol 25, 2024, https://doi.org/10.1093/bib/bbae551
7. Y. Yang, Md. Z. Hossain, E. A. Stone, S. Rahman: Exemplar Guided Deep Neural Network for Spatial Transcriptomics Analysis of Gene Expression Prediction. In Proceedings of the Winter Conference on Applications of Computer Vision, 2023, 10.1109/WACV56688.2023.00501
8. Chen, R.J., Ding, T., Lu, M.Y. et al. Towards a general-purpose foundation model for computational pathology. Nat Med 30, 850–862 (2024). https://doi.org/10.1038/s41591-024-02857-3

9. T. M. Dang, Y. Guo, H. Ma, Q. Zhou, S. Na, J. Gao, J. Huang: MFMF: Multiple Foundation Model Fusion Networks for Whole Slide Image Classification. In Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '24)
10. H. Li, C. Wang, G. Zhao, Z. He, Y. Wang, Z. Sun: Sclera-TransFuse: Fusing Swin Transformer and CNN for Accurate Sclera Segmentation, 2023 IEEE International Joint Conference on Biometrics (IJCB), doi: 10.1109/IJCB57857.2023.10448814.
11. C. Wang, H. Li, Y. Zhang, G. Zhao, Y. Wang and Z. Sun: Sclera-TransFuse: Fusing Vision Transformer and CNN for Accurate Sclera Segmentation and Recognition, in IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 6, no. 4, pp. 575-590, Oct. 2024, doi: 10.1109/TBIOM.2024.3415484
12. T. M. Dang, T. D. Nguyen, T. Hoang, H. Kim, A. Beng Jin Teoh, D. Choi: AVET: A Novel Transform Function to Improve Cancellable Biometrics Security, in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 758-772, 2023, doi: 10.1109/TIFS.2022.3230212
13. F. Jiang, Y. Guo, H. Ma, S. Na, W. Zhong, Y. Han, T. Wang, J. Huang: GTE: a graph learning framework for prediction of T-cell receptors and epitopes binding specificity, Briefings in Bioinformatics, https://doi.org/10.1093/bib/bbae343
14. Wang, C., Li, H., Ma, W. et al. MetaScleraSeg: an effective meta-learning framework for generalized sclera segmentation. Neural Comput & Applic 35, 21797–21826 (2023). https://doi.org/10.1007/s00521-023-08937-8
15. F. Jiang, Y. Guo, H. Ma, S. Na, W. An, B. Song, Y. Han, J. Gao, T. Wang, J. Huang: AlphaEpi: Enhancing B Cell Epitope Prediction with AlphaFold 3. In Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '24), https://doi.org/10.1145/3698587.3701389
16. Miller, H.E., Bishop, A.J.R.: Correlation AnalyzeR: functional predictions from gene co-expression correlations. BMC Bioinformatics 22, 206 (2021). https://doi.org/10.1186/s12859-021-04130-7
17. Du, J., Jia, P., Dai, Y. et al. Gene2vec: distributed representation of genes based on co-expression. BMC Genomics 20. https://doi.org/10.1186/s12864-018-5370-x
18. Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, Aditya Grover: CYCLIP: Cyclic Contrastive Language-Image Pretraining. In: Advances in Neural Information Processing Systems, 2022
19. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the International Conference on Machine Learning, 2021
20. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I Polosukhin: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017
21. A. Andersson, L. Larsson, L. Stenbeck, F. Salmén, A. Ehinger, S. Z. Wu, G. Al-Eryani, D. Roden, A. Swarbrick, A. Borg, J. Frisén, C. Engblom, J. Lundeberg: Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. Nat Commun. https://doi.org/10.1038/s41467-021-26271-2
22. A. L. Ji, A. J. Rubin, K. Thrane, S. Jiang, D. L. Reynolds, R. M. Meyers, M. G. Guo, B. M. George, A. Mollbrink, J. Bergenstråhle, L. Larsson, Y. Bai, B. Zhu, A. Bhaduri, J. M. Meyers, X. Rovira-Clavé, S. T. Hollmig, S. Z. Aasi, G. P. Nolan, J. Lundeberg, P. A. Khavari: Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell, vol 182, pp. 497-514, 2020. https://doi.org/10.1016/j.cell.2020.05.039

23. B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, M. Welling.: Rotation Equivariant CNNs for Digital Pathology. arXiv:1806.03962 (2018)
24. Kriegsmann, K., Lobers, F., Zgorzelski, C., Kriegsmann, J., Meliß, Rolf R., Sack, U., Steinbuss, G., Kriegsmann, M.: Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections [Data set], (2023). https://doi.org/10.11588/DATA/7QCR8S, heiDATA, V1
25. Lingle, W., Erickson, B. J., Zuley, M. L., Jarosz, R., Bonaccio, E., Filippini, J., Net, J. M., Levi, L., Morris, E. A., Figler, G. G., Elnajjar, P., Kirk, S., Lee, Y., Giger, M., Gruszauskas, N. (2016). The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA) (Version 3) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.AB2NAZRP
26. Campanella, G., Hanna, M. G., Brogi, E., Fuchs, T. J.: Breast Metastases to Axillary Lymph Nodes [Data set]. The Cancer Imaging Archive (2019). https://doi.org/10.7937/tcia.2019.3xbn2jcc
27. Desai K, Nair MG, Prabhu JS, et al.: High expression of integrin $\beta 6$ in association with the Rho-Rac pathway identifies a poor prognostic subgroup within HER2 amplified breast cancers. Cancer Med. 2016;5(8):2000-2011. doi:10.1002/cam4.756
28. T. M. Dang, Q. Zhou, Y. Guo, H. Ma, S. Na, T. B. Dang, J. Gao, J. Huang: Abnormality-Aware Multimodal Learning for WSI Classification. Front. Med., vol 12, 2025. doi: 10.3389/fmed.2025.1546452
29. Lu, M.Y., Williamson, D.F.K., Chen, T.Y. et al. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 5, 555–570 (2021). https://doi.org/10.1038/s41551-020-00682-w
30. M. Ilse, J. M. Tomczak, M. Welling: Attention-based Deep Multiple Instance Learning. In: Proceedings of the International Conference on Machine Learning, 2018