Dual-Path Knowledge-Augmented Contrastive Alignment Network for Spatially Resolved Transcriptomics

Wei Zhang*, **Jiajun Chu***, **Xinci Liu, Chen Tong, Xinyue Li**†

Department of Data Science, College of Computing, City University of Hong Kong, Hong Kong SAR, China
xinyueli@cityu.edu.hk

Abstract

Spatial Transcriptomics (ST) is a technology that measures gene expression profiles within tissue sections while retaining spatial context. It reveals localized gene expression patterns and tissue heterogeneity, both of which are essential for understanding disease etiology. However, its high cost has driven efforts to predict spatial gene expression from whole slide images. Despite recent advancements, current methods still face significant limitations, such as under-exploitation of high-level biological context, over-reliance on exemplar retrievals, and inadequate alignment of heterogeneous modalities. To address these challenges, we propose **DKAN**, a novel **D**ual-path **K**nowledge-**A**ugmented contrastive alignment **N**etwork that predicts spatially resolved gene expression by integrating histopathological images and gene expression profiles through a biologically informed approach. Specifically, we introduce an effective gene semantic representation module that leverages the external gene database to provide additional biological insights, thereby enhancing gene expression prediction. Further, we adopt a unified, one-stage contrastive learning paradigm, seamlessly combining contrastive learning and supervised learning to eliminate reliance on exemplars, complemented with an adaptive weighting mechanism. Additionally, we propose a dual-path contrastive alignment module that employs gene semantic features as dynamic cross-modal coordinators to enable effective heterogeneous feature integration. Through extensive experiments across three public ST datasets, DKAN demonstrates superior performance over state-of-the-art models, establishing a new benchmark for spatial gene expression prediction and offering a powerful tool for advancing biological and clinical research.

Code — https://github.com/coffeeNtv/DKAN
Extended version — https://arxiv.org/abs/2511.17685

Introduction

Spatial Transcriptomics (ST) is an advanced technology that measures gene expression profiles within tissue sections while preserving their spatial context, often by integrating data with Whole Slide Images (WSIs) (Ståhl et al. 2016).

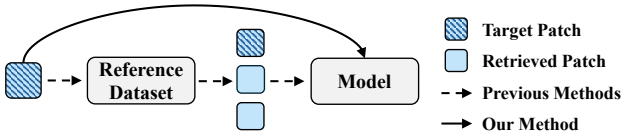

Figure 1: Pipeline comparison. Existing contrastive learning and exemplar-guided methods require constructing a reference dataset and retrieving similar patches as intermediate steps, whereas our one-stage contrastive learning method operates through a straightforward pipeline.

This spatially resolved perspective is pivotal in revealing the heterogeneity of gene expression across tissue microenvironments, offering critical insights into developmental processes, disease progression, and cell-cell interactions.

Despite the transformative potential of ST techniques, they still face limitations, including relatively low resolution, typically at the multicellular level, and high technical costs, which hinder their broader adoption (Moses and Pachter 2022). In contrast, Hematoxylin and Eosin (H&E) stained WSIs, as the gold standard in pathology, offer a cost-effective and widely accessible alternative. Their widespread availability and low costs make them suitable for supporting numerous downstream tasks such as survival prediction (Zhang et al. 2024a, 2025b), stain transfer (Li et al. 2023; Zhang et al. 2024b) and especially spatial transcriptomics (Lin et al. 2024; Zhang et al. 2025a). The potential of WSIs to predict spatially resolved gene expression has been successfully demonstrated (He et al. 2020; Schmauch et al. 2020), utilizing morphological and spatial details. More recently, several models have expanded on this foundation, further improving prediction accuracy through innovative approaches to leverage the rich tissue information in WSIs (Chung et al. 2024; Wang et al. 2024).

Current approaches for spatial gene expression prediction predominantly exploit the rich spatial information embedded in WSIs, extracting image features at various levels, including local (Xie et al. 2023; Mejia et al. 2023), global (Pang, Su, and Li 2021; Zeng et al. 2022), and multi-scale representations (Wang et al. 2024; Lin et al. 2024). Furthermore, several models incorporate multimodal contrastive learning to align imaging data with gene expression

*These authors contributed equally.
†Corresponding author.

profiles within a shared low-dimensional embedding space (Xie et al. 2023; Min et al. 2024). This alignment enables the models to effectively capture the intricate relationships between image-derived features and gene expression patterns.

Despite recent advancements, several challenges persist. First, many models rely heavily on image features derived from pixel intensity (e.g., color distribution) and cellular structure (e.g., shape and texture) (He et al. 2020; Pang, Su, and Li 2021). While these low-level visual cues are informative, they often fail to capture high-level semantic information, such as gene functions, biological pathways, or disease associations, limiting the depth of biological interpretation. Second, one notable challenge lies in the inclusion of additional and potentially redundant steps in models based on contrastive learning and exemplar-guided strategies. As shown in Figure 1, these pipelines typically involve constructing a reference dataset from all patches in the training set, retrieving similar patches, and feeding both the retrieved patches and the target patch into the model. While effective, this multi-step process introduces complexity that may not always be necessary (Xie et al. 2023; Min et al. 2024; Lin et al. 2024). Streamlining such workflows into a more cohesive approach, particularly in constrained settings or with limited datasets, remains challenging. Lastly, while existing methods leverage multi-scale image features (Zeng et al. 2022; Chung et al. 2024) or incorporate auxiliary modalities (Yang et al. 2024a) to address modality-specific semantics, their fusion strategies often fail to adequately preserve biologically relevant interactions. This limitation constrains performance, a gap further exacerbated by the absence of frameworks that explicitly incorporate gene functional semantics into multimodal alignment.

To address these challenges, we propose DKAN, a novel **D**ual-path **K**nowledge-**A**ugmented contrastive alignment **N**etwork for spatial gene expression prediction. Unlike previous methods, DKAN integrates gene functional semantics into contrastive learning, enabling the biologically grounded fusion of histopathological images and expression profiles through a unified one-stage paradigm. Our major contributions are summarized as follows:

- We propose a novel paradigm for spatial gene expression prediction by incorporating gene functional semantics into contrastive learning, enabling the model to capture high-level biological context beyond low-level image features and align predictions with established genomic knowledge.
- We develop a unified one-stage contrastive learning framework that integrates supervised and contrastive objectives via adaptive weighting, simplifying the pipeline by removing exemplar dependence and eliminating separate storage or retrieval steps.
- We introduce a dual-path contrastive alignment module that processes image and gene expression features separately, avoiding forced direct alignment of heterogeneous modalities. Leveraging gene semantics enables precise multimodal integration into a shared embedding space, overcoming limitations of prior fusion strategies.
- We conduct extensive experiments on three public ST datasets, demonstrating DKAN consistently outperforms State-Of-The-Art (SOTA) models across benchmarks.

Related Work

Spatial Gene Expression Prediction

Spatial gene expression prediction seeks to model gene activity from WSIs by capturing both visual and spatial features. Existing methods fall into three main categories:

Local Methods. Local approaches focus on the target patch and its immediate surroundings (Yang et al. 2023, 2024b,a; Xie et al. 2023; Min et al. 2024). ST-Net (He et al. 2020) uses a pretrained DenseNet-121 (Huang et al. 2017) to extract patch-level features for prediction. EGN (Yang et al. 2023) enhances patch representations through image reconstruction and exemplar retrieval. EGGN (Yang et al. 2024b) extends EGN by applying graph convolutional networks to model relationships between the patch and its exemplars. SEPAL (Mejia et al. 2023) constructs a neighborhood graph and applies a graph neural network to capture local dependencies. These methods prioritize localized information and may not account for the broader spatial context.

Global Methods. Global methods incorporate positional and contextual information across the entire WSI (Jia et al. 2023; Yang et al. 2024a). HisToGene (Pang, Su, and Li 2021) uses vision transformers (Dosovitskiy et al. 2021) with positional encoding to model inter-patch relationships. HE2RNA (Schmauch et al. 2020) clusters patches into supertiles and aggregates them to form global contextual features. THItoGene (Jia et al. 2023) extracts deep molecular features using dynamic convolution and capsule modules, integrates them with positional data via ViT, and refines predictions using a graph attention network. Unlike local methods, global models leverage the full image context for more informed predictions.

Multi-Scale Methods. Multi-scale methods capture biological patterns at various resolutions (Zeng et al. 2022; Chung et al. 2024; Wang et al. 2024; Lin et al. 2024). TRIPLEX (Chung et al. 2024) combines features from multiple views. M2OST (Wang et al. 2024) decouples intra- and inter-scale feature extraction for many-to-one spatial prediction. ST-Align (Lin et al. 2024) clusters patches into niche-level groups to integrate both local and regional contexts. These methods aim to balance the fine granularity of local models with the broader context provided by global ones.

Contrastive Representation Learning

Contrastive learning is a self-supervised method that learns discriminative representations by pulling similar pairs together and pushing dissimilar pairs apart (Oord, Li, and Vinyals 2018). In spatial gene expression tasks, contrastive learning aligns visual and transcriptomic modalities in a shared embedding space (Xie et al. 2023; Min et al. 2024; Lin et al. 2024). BLEEP (Xie et al. 2023), inspired by CLIP (Radford et al. 2021), embeds images and gene profiles jointly to enable retrieval-based inference. mclSTExp (Min et al. 2024) refines this by encoding gene expression with learnable position embeddings for better spatial integration. ST-Align (Lin et al. 2024) advances contrastive

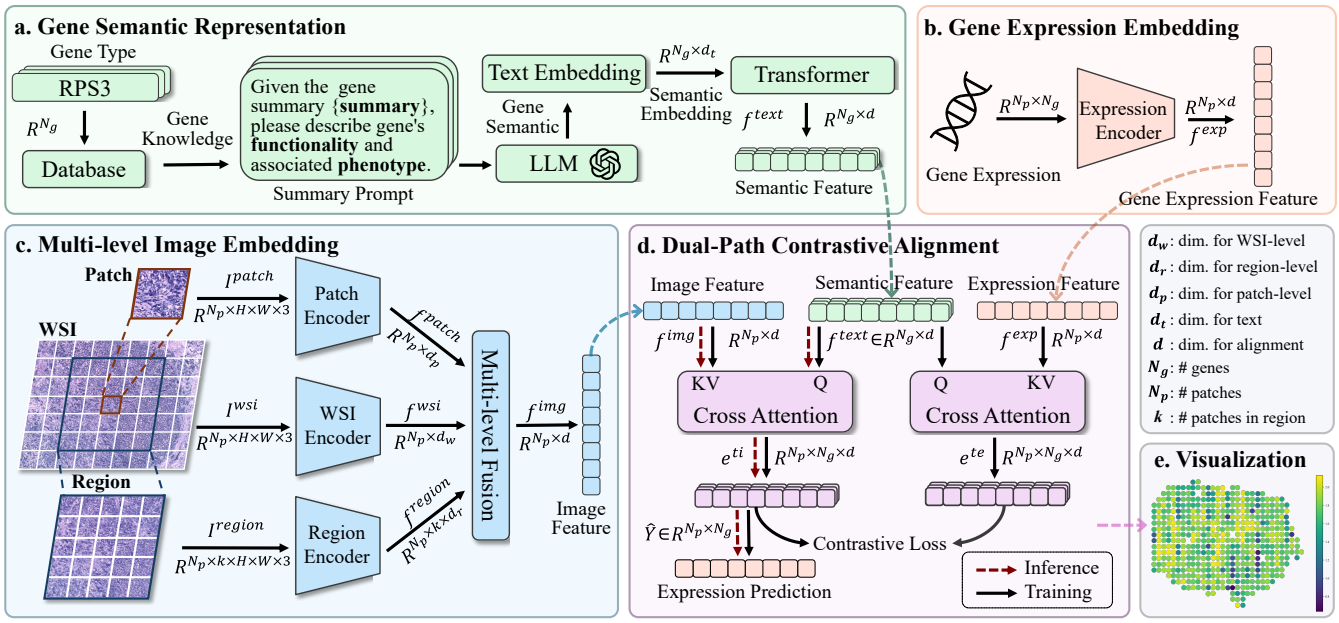

Figure 2: Overview of DKAN framework. (a) Gene semantic feature representation module. (b) Gene expression feature embedding module. (c) Multi-level image embedding module. (d) Dual-path contrastive alignment module. (e) Example spatial gene expression map for FN1 in the STNET dataset.

learning by curating 1.3 million image-gene pairs and introducing a multi-scale feature extractor with a three-target alignment strategy. This enhances the model's ability to capture complex structural patterns in spatial transcriptomics.

Methodology

Problem Formulation

Spatial gene expression prediction is framed as a regression task. Given N_p image patches extracted from a WSI, represented as $X \in \mathbb{R}^{N_p \times H \times W \times 3}$, where H and W are the height and width of each patch, the goal is to predict gene expression levels. A learnable mapping function f is applied to produce predictions $\hat{Y} = f(X) \in \mathbb{R}^{N_p \times N_g}$, where N_g denotes the number of target genes.

Overview

As illustrated in Figure 2, our framework integrates high-level gene semantics by retrieving information from an external gene database (Sayers et al. 2024) and leveraging prompts to tap into the summarization capabilities and domain knowledge of large language models (LLMs). To extract informative visual features from WSIs, we adopt a multi-scale strategy that captures representations at the patch, region, and whole-slide levels. These features are then fused to form a comprehensive visual embedding. To effectively align gene expression, image, and textual modalities, we introduce a dual-path knowledge-augmented contrastive alignment module, which employs two distinct contrastive pathways for robust multimodal integration.

Gene Semantic Representation

As shown in Figure 2(a), for N_g genes of interest, we designed a workflow to extract semantic features f^{text}. We retrieved gene-related knowledge from a well-established gene database NCBI (Sayers et al. 2024). However, the retrieved gene knowledge lacks structural uniformity, with some information being redundant or incomplete. To address this issue, we leveraged the summarization capabilities and embedded knowledge of LLM (GPT-4o) to generate accurate and efficient gene semantic texts.

Specifically, we embedded the gene knowledge into a prompt, which includes role definitions, task requirements, and output specifications, before feeding it into the LLM to produce the gene semantic text. The prompt is provided in the supplementary materials. Subsequently, we employed BioBERT (Lee et al. 2019) as our text embedding model to extract textual features, generating semantic embeddings of dimensionality $d_t = 1024$. This model, pre-trained on the extensive biomedical semantic corpus, excels at capturing domain-specific contextual representations effectively. The semantic features are processed by a standard transformer module, preceded by a linear projection to ensure dimensional alignment for multimodal fusion. The transformer module efficiently captures global dependencies and commonalities among semantic embeddings, ultimately yielding the final semantic features f^{text}.

Gene Expression Embedding

As illustrated in Figure 2(b), the gene expression with shape $N_p \times N_g$ is processed by the gene expression encoder to generate gene expression features $f^{exp} \in \mathbb{R}^{N_p \times d}$, ensuring feature dimension consistency between gene semantic

and image features. Specifically, the expression encoder first projects the input to a d-dimensional space via a linear layer, applies GELU activation, and then processes features through a second linear layer with dropout. To stabilize gradient flow, we employ residual connections: the initial linear projection's output is directly added to the final dropout output, followed by layer normalization for feature standardization. This design mitigates gradient vanishing while maintaining feature discriminability.

Multi-level Image Embedding

Given the large size of WSIs, relying solely on either WSI-level images (I^{wsi}) or patch-level images (I^{patch}) is insufficient to fully capture their morphological complexity. While WSIs offer rich global context, a significant gap remains between the global view and the localized detail at the patch level. To bridge this gap, as illustrated in Figure 2(c), we introduce a region-level representation (I^{region}) by selecting the k nearest neighboring patches around each target patch.

Our model extracts image patches from the WSI at these three hierarchical levels and processes them using dedicated encoders. Specifically, for the WSI-level features f^{wsi} and the region-level features f^{region}, we utilize UNI (Chen et al. 2024), a general-purpose foundation model pre-trained on extensive WSI datasets for computational pathology. Due to the scale constraints and computational demands of WSIs and region-level images, UNI serves as a fixed feature extractor without updating its weights during training. To enhance feature adaptability, we append a multi-head transformer after each UNI encoder. For the patch-level feature f^{patch}, we employ ResNet18 (Ciga, Xu, and Martel 2022) as the encoder. To adapt it to feature extraction, we remove the final pooling and fully connected layers, retaining only the activations of the last hidden layer as the output. Notably, the parameters of ResNet18 remain trainable.

To effectively integrate multi-scale features, we employ two cross-attention mechanisms: one fuses the WSI and the region-level images, while the other combines the WSI and the patch-level images, with WSI-level features serving as the query in both cases. The resulting fused features from these two groups are then summed to produce the final multi-scale feature of the image f^{img}.

Dual-Path Contrastive Alignment

After extracting the image, semantic, and expression features, we propose a novel dual-path knowledge-augmented contrastive alignment paradigm for multimodal alignment as shown in Figure 2(d). Our approach leverages gene semantic features as dynamic cross-modal coordinators, operating through two parallel pathways. In the image pathway, gene semantic knowledge serves as a "functional query instruction" to filter morphology-related regions from image features. Similarly, in the expression pathway, gene semantic knowledge acts as a "distribution correction factor" to constrain the predicted gene expression features, ensuring alignment with the established biological pathway logic. In the implementation, we employ a cross-attention module, using the semantic feature as the query. Each semantic feature independently queries the image and expression features, ultimately generating gene knowledge-augmented representations, denoted as e^{ti} and e^{te} respectively.

Inspired by CLIP (Radford et al. 2021), we apply contrastive learning to align e^{ti} and e^{te} in the latent embedding space. One distinctive aspect of this method is that, instead of forcing the alignment of the heterogeneous image and gene expression modalities directly, each modality interacts independently with the gene semantic knowledge, achieving implicit alignment through knowledge-guided queries. The decoupling of the image and gene expression modules enhances flexibility and reduces inter-modal dependencies.

To eliminate the dependency on exemplars and streamline the workflow, we adopt a unified one-stage framework for contrastive learning and seamlessly integrate it with supervised training. In the training phase, all modalities are utilized, whereas during inference, only the image and semantic modalities are used. Consequently, the loss function combines a contrastive loss and a supervised loss. The contrastive loss is indicated in Equation 1, where positive samples are representations of the same gene paired together, and negative samples are drawn from representations of different genes. Here $sim(\cdot,\cdot)$ denotes the cosine similarity function that measures the alignment between feature vectors, and τ is a temperature parameter that controls the sharpness of the similarity distribution.

$$\mathcal{L}_{cont} = -\sum_{i} \log \frac{\exp(sim(e^i_{ti}, e^i_{te})/\tau)}{\sum_j \exp(sim(e^i_{ti}, e^j_{te})/\tau)}. \quad (1)$$

For the supervised loss, we calculate the mean squared error (MSE) between the predicted gene expression $\hat{Y}$ and the ground truth gene expression Y. This can be further enhanced by knowledge distillation (Chung et al. 2024), which improves prediction consistency and generalization by aligning intermediate representations with the final output. These intermediate predictions, $\hat{Y}_{img}$, $\hat{Y}_{patch}$, $\hat{Y}_{wsi}$, and $\hat{Y}_{region}$, are obtained through linear transformations of the model features f^{img}, f^{patch}, f^{wsi}, and f^{region}. To enforce both accuracy and coherence, we compute the MSE between these intermediate predictions and two targets: the ground truth Y and the final predicted output $\hat{Y}$, with their contributions balanced by a hyperparameter λ.

The distillation-aware supervised loss for each intermediate prediction $d \in \mathcal{D}$ is defined in Equation 2:

$$\mathcal{L}_d = \lambda \|\hat{Y}_d - \hat{Y}\|^2 + (1-\lambda)\|\hat{Y}_d - Y\|^2, \quad (2)$$

where $\mathcal{D} = \{\text{img}, \text{patch}, \text{wsi}, \text{region}\}$. The total supervised loss is then aggregated across all intermediate predictions, combined with the MSE between the ground truth Y and the final predicted output $\hat{Y}$:

$$\mathcal{L}_{sup} = \sum_{d \in \mathcal{D}} \mathcal{L}_d + \|\hat{Y} - Y\|^2. \quad (3)$$

To ensure balanced optimization between the supervised loss ($\mathcal{L}_{sup}$) and the contrastive loss ($\mathcal{L}_{cont}$) which exhibit different numerical scales and convergence characteristics, we propose an adaptive weighting scheme. The weights are

Comparison Settings			Error		PCC			
	Type	Model	MAE↓	MSE↓	ALL↑	HPG↑	HEG↑	HVG↑
HER2+ Dataset	Local	ST-Net	0.432 ± 0.05	0.311 ± 0.07	0.150 ± 0.13	0.287 ± 0.19	0.115 ± 0.11	0.090 ± 0.08
		BLEEP	0.401 ± 0.03	0.277 ± 0.05	0.151 ± 0.11	0.277 ± 0.16	0.246 ± 0.09	0.261 ± 0.07
		EGN	0.366 ± 0.04	0.229 ± 0.05	0.204 ± 0.12	0.364 ± 0.16	0.152 ± 0.09	0.120 ± 0.05
		mclSTExp	0.398 ± 0.04	0.272 ± 0.05	0.163 ± 0.11	0.289 ± 0.16	0.114 ± 0.08	0.091 ± 0.06
	Global	HisToGene	0.388 ± 0.06	0.253 ± 0.07	0.150 ± 0.09	0.295 ± 0.15	0.099 ± 0.07	0.079 ± 0.05
		THItoGene	0.424 ± 0.05	0.291 ± 0.06	0.051 ± 0.05	0.118 ± 0.08	0.045 ± 0.05	0.030 ± 0.03
		SGN	0.734 ± 0.20	0.749 ± 0.38	0.035 ± 0.03	0.065 ± 0.05	0.022 ± 0.03	0.017 ± 0.03
	Multi-view	Hist2ST	0.417 ± 0.07	0.293 ± 0.08	0.193 ± 0.10	0.360 ± 0.17	0.126 ± 0.07	0.109 ± 0.03
		TRIPLEX	0.364 ± 0.05	0.234 ± 0.06	0.304 ± 0.14	0.491 ± 0.18	0.271 ± 0.10	0.260 ± 0.06
		M2OST	0.446 ± 0.10	0.340 ± 0.15	0.147 ± 0.12	0.313 ± 0.19	0.098 ± 0.09	0.090 ± 0.06
		DKAN(Ours)	$\mathbf{0.361 \pm 0.04}$	$\mathbf{0.224 \pm 0.06}$	$\mathbf{0.330 \pm 0.13}$	$\mathbf{0.531 \pm 0.15}$	$\mathbf{0.317 \pm 0.09}$	$\mathbf{0.304 \pm 0.07}$
STNET Dataset	Local	ST-Net	0.357 ± 0.04	0.222 ± 0.05	0.081 ± 0.05	0.192 ± 0.09	0.026 ± 0.05	0.096 ± 0.04
		BLEEP	0.369 ± 0.02	0.235 ± 0.02	0.095 ± 0.05	0.193 ± 0.10	0.063 ± 0.03	0.111 ± 0.05
		EGN	0.354 ± 0.02	0.214 ± 0.03	0.107 ± 0.05	0.207 ± 0.09	0.089 ± 0.04	0.108 ± 0.04
		mclSTExp	0.350 ± 0.02	0.210 ± 0.02	0.095 ± 0.05	0.202 ± 0.09	0.052 ± 0.04	0.088 ± 0.03
	Global	HisToGene	0.326 ± 0.02	0.180 ± 0.02	0.103 ± 0.04	0.217 ± 0.11	0.060 ± 0.02	0.074 ± 0.03
		THItoGene	0.347 ± 0.03	0.200 ± 0.04	0.040 ± 0.02	0.092 ± 0.02	0.025 ± 0.02	0.028 ± 0.02
		SGN	1.180 ± 1.78	4.952 ± 12.00	0.027 ± 0.01	0.048 ± 0.02	0.014 ± 0.01	0.027 ± 0.01
	Multi-view	Hist2ST	0.352 ± 0.02	0.208 ± 0.03	0.142 ± 0.04	0.268 ± 0.09	0.094 ± 0.03	0.122 ± 0.03
		TRIPLEX	0.342 ± 0.02	0.200 ± 0.02	0.194 ± 0.07	0.344 ± 0.10	0.160 ± 0.06	0.224 ± 0.07
		M2OST	0.369 ± 0.04	0.226 ± 0.04	0.022 ± 0.02	0.081 ± 0.04	0.008 ± 0.03	-0.001 ± 0.03
		DKAN(Ours)	$\mathbf{0.322 \pm 0.02}$	$\mathbf{0.179 \pm 0.02}$	$\mathbf{0.219 \pm 0.07}$	$\mathbf{0.387 \pm 0.09}$	$\mathbf{0.200 \pm 0.06}$	$\mathbf{0.244 \pm 0.07}$
cSCC Dataset	Local	ST-Net	0.410 ± 0.05	0.262 ± 0.07	0.170 ± 0.08	0.289 ± 0.09	0.079 ± 0.05	0.065 ± 0.02
		BLEEP	0.430 ± 0.04	0.297 ± 0.05	0.269 ± 0.07	0.396 ± 0.08	0.266 ± 0.09	0.250 ± 0.10
		EGN	0.438 ± 0.05	0.303 ± 0.06	0.278 ± 0.06	0.388 ± 0.06	0.194 ± 0.06	0.180 ± 0.06
		mclSTExp	0.445 ± 0.05	0.311 ± 0.06	0.168 ± 0.04	0.291 ± 0.08	0.098 ± 0.03	0.096 ± 0.05
	Global	HisToGene	0.441 ± 0.06	0.297 ± 0.07	0.178 ± 0.10	0.319 ± 0.14	0.099 ± 0.05	0.094 ± 0.04
		THItoGene	0.495 ± 0.11	0.380 ± 0.16	0.040 ± 0.05	0.101 ± 0.06	0.014 ± 0.03	0.020 ± 0.03
		SGN	0.832 ± 0.13	0.897 ± 0.27	0.059 ± 0.02	0.086 ± 0.03	0.048 ± 0.02	0.044 ± 0.01
	Multi-view	Hist2ST	0.468 ± 0.11	0.338 ± 0.16	0.185 ± 0.14	0.261 ± 0.15	0.133 ± 0.09	0.110 ± 0.07
		TRIPLEX	0.415 ± 0.06	0.278 ± 0.08	0.363 ± 0.07	0.476 ± 0.07	0.276 ± 0.07	0.272 ± 0.06
		M2OST	0.443 ± 0.07	0.313 ± 0.09	-0.018 ± 0.04	0.126 ± 0.09	0.012 ± 0.07	-0.041 ± 0.04
		DKAN(Ours)	$\mathbf{0.383 \pm 0.05}$	$\mathbf{0.239 \pm 0.06}$	$\mathbf{0.407 \pm 0.08}$	$\mathbf{0.508 \pm 0.08}$	$\mathbf{0.346 \pm 0.09}$	$\mathbf{0.321 \pm 0.08}$

Table 1: Comparison with SOTA methods. The best results are highlighted in bold.

dynamically adjusted based on the real-time loss values to maintain appropriate gradient contributions from each objective. Specifically, the weighting coefficients are computed as the normalized reciprocals of the respective losses, to ensure that the loss with a smaller value could receive a higher weight, allowing the model to dynamically prioritize the more reliable objective during training. The final composite loss function is thus formulated as:

$$\mathcal{L} = w_{sup}\mathcal{L}_{sup} + w_{cont}\mathcal{L}_{cont}. \tag{4}$$

Experiments

Datasets

We evaluated our approach on three public ST datasets: two human breast cancer (BC) datasets and one cutaneous squamous cell carcinoma (cSCC) dataset. The HER2+ BC dataset (Andersson et al. 2021) includes 36 samples from 8 patients, with 13,620 spots and 14,873 genes profiled per spot. The STNET BC dataset (He et al. 2020) contains 68 WSIs from 23 patients, comprising 30,612 spots and 26,949 genes per spot. The cSCC dataset (Ji et al. 2020) consists of 12 samples from 4 patients, with 8,671 spots and 17,047 genes measured per spot.

Evaluation and Metrics

To ensure robust evaluation, we applied cross-validation strategies tailored to each dataset, ensuring no patient overlap between training and test sets. For STNET, we used 8-fold cross-validation. For the smaller HER2+ and cSCC datasets, we adopted leave-one-patient-out cross-validation, with 8 folds for HER2+ and 4 folds for cSCC, where each fold used one patient's samples for testing and the rest for training. This setup aligns with prior work (Chung et al. 2024) to ensure fair comparison.

We evaluated our model using six metrics to ensure comprehensive assessment and comparability with prior stud-

	Error		PCC			
Text Emb.	MAE↓	MSE↓	ALL↑	HPG↑	HEG↑	HVG↑
Conch	0.324	0.181	0.211	0.374	0.194	0.230
PLIP	0.323	**0.179**	0.217	0.381	0.194	**0.244**
BioGPT	0.327	0.183	0.217	0.385	0.193	0.239
BioBERT	**0.322**	**0.179**	**0.219**	**0.387**	**0.200**	**0.244**
Image Emb.	MAE↓	MSE↓	ALL↑	HPG↑	HEG↑	HVG↑
ResNet18	0.341	0.197	0.202	0.347	0.176	0.224
ResNet50	0.354	0.212	0.196	0.349	0.167	0.215
PLIP	0.333	0.188	0.190	0.333	0.157	0.211
Conch	0.324	**0.177**	0.209	0.369	0.179	0.231
UNI	**0.322**	0.179	**0.219**	**0.387**	**0.200**	**0.244**

Table 2: Ablation studies on text and image encoders.

ies (Yang et al. 2023; Xie et al. 2023; Chung et al. 2024). These include Mean Absolute Error (MAE), Mean Squared Error (MSE), and Pearson Correlation Coefficient (PCC) across: (1) all genes of interest, (2) the top 50 Highly Predictive Genes (HPG), (3) the top 50 Highly Expressed Genes (HEG), and (4) the top 50 Highly Variable Genes (HVG). PCC was computed per gene across all spots within each sample and averaged over all cross-validation folds.

Implementations

To align with previous studies (He et al. 2020; Chung et al. 2024), all patches were segmented with dimensions of H=W=224 pixels, and regions were constructed using k=25 (a 5×5 patch grid). We select N_g=250 spatially variable genes for training to align with previous studies. We use the Adam optimizer with a learning rate of 0.0001 and a StepLR scheduler (step size=50, gamma=0.9). The temperature τ in the contrastive loss was set to 0.1 for HER2+ and STNET and 0.08 for cSCC. Image encoders included UNI for WSI and region levels (d_h=d_r=1024) and ResNet18 for patch level (d_p=512), while the text embedding model, BioBERT, produced embeddings with d_t=1024. All models were trained on a NVIDIA RTX A800 GPU with a batch size of 128.

Experimental Results and Analysis

We compared our proposed model, DKAN, against extensive SOTA baselines across three categories: (1) Local methods: ST-Net (He et al. 2020), BLEEP (Xie et al. 2023), EGN (Yang et al. 2023), and mclSTExp (Min et al. 2024); (2) Global methods: HisToGene (Pang, Su, and Li 2021), THItogene (Jia et al. 2023), and SGN (Yang et al. 2024a); and (3) Multi-scale methods: Hist2ST (Zeng et al. 2022), TRIPLEX (Chung et al. 2024), and M2OST (Wang et al. 2024). As shown in Table 1, DKAN consistently outperforms all baselines across datasets and evaluation metrics.

Take the HER2+ dataset as an example, DKAN achieves the lowest MAE (0.361) and MSE (0.224), along with the highest PCC values for all genes (0.330), HPG (0.531), HEG (0.317), and HVG (0.304). In comparison, the current SOTA method TRIPLEX reports 0.364 (MAE), 0.234 (MSE), and PCCs of 0.304 (all genes), 0.491 (HPG), 0.271 (HEG), and 0.260 (HVG). These results demonstrate DKAN's superior performance over TRIPLEX and other baselines. Notably, DKAN also consistently outperforms all methods on the cSCC and STNET datasets, highlighting its effectiveness and robust generalization across diverse ST datasets.

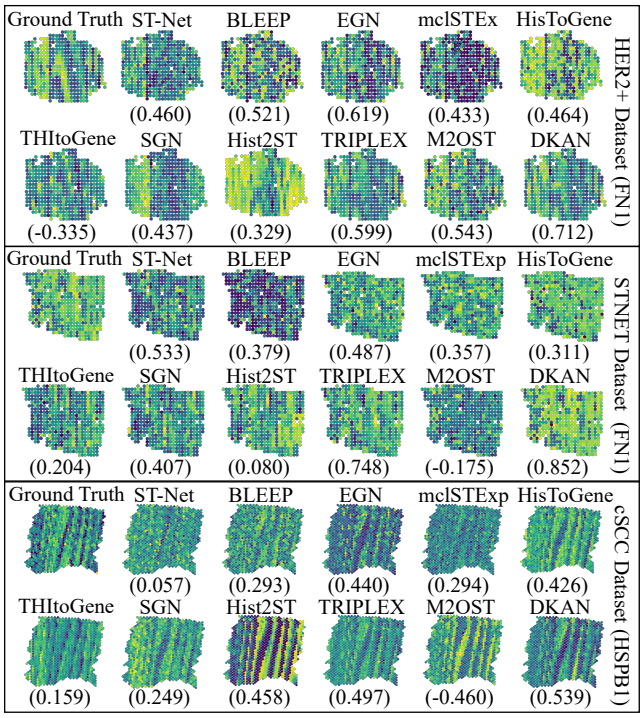

Figure 3: Visualization of expression patterns of cancer biomarker genes alongside PCC values for all datasets.

Visualization of Cancer Biomarker Genes

To evaluate the model's ability to capture spatial gene expression patterns both quantitatively and qualitatively, we visualize the log-normalized expression levels of two well-established cancer biomarkers in Figure 3. Specifically, FN1, frequently overexpressed in breast cancer (Zhang, Luo, and Wu 2022), and HSPB1, implicated in cancer progression (Liang et al. 2023), are highlighted. We also report their PCCs with the ground truth to assess spatial consistency and predictive accuracy. Additional visualizations are available in the supplementary materials.

Ablation Studies

To validate the effectiveness of our model design, we conducted ablation studies on several key components: the choice of text and image encoders, textual representations with different prompt strategies and LLMs, the contributions of individual modules, and the selection of fusion strategies and loss functions. For clarity, we present mean results on the STNET dataset in the main text. Consistent trends were also observed on the HER2+ and cSCC datasets, with detailed results provided in the supplementary materials.

(1) Text and Image Encoders. For the text encoder, we evaluated four embedding models: Conch (Lu et al. 2024)

	Error		PCC			
Prompt Strategy	MAE↓	MSE↓	ALL↑	HPG↑	HEG↑	HVG↑
w/o text constraint	0.340	0.198	0.206	0.351	0.177	0.227
w/o text summary	0.342	0.199	0.199	0.341	0.177	0.220
Ours	**0.322**	**0.179**	**0.219**	**0.387**	**0.200**	**0.244**
LLM Candidate	MAE↓	MSE↓	ALL↑	HPG↑	HEG↑	HVG↑
Deepseek-R1	0.343	0.199	0.198	0.345	0.174	0.226
Deepseek-v3	0.337	0.193	0.202	0.342	0.187	0.221
LLaMA 2	0.339	0.194	0.193	0.337	0.165	0.218
GPT-4o	**0.322**	**0.179**	**0.219**	**0.387**	**0.200**	**0.244**

Table 3: Ablation studies on prompt strategies and LLMs.

and PLIP (Huang et al. 2023), which are medical vision-language models, as well as BioBERT (Lee et al. 2019) and BioGPT (Luo et al. 2022), which are pretrained on large-scale biomedical corpora such as PubMed. For image encoders at both the region and WSI levels, we compared PLIP (Huang et al. 2023), Conch (Lu et al. 2024), UNI (Chen et al. 2024), ResNet50(He et al. 2016), and ResNet18(Ciga, Xu, and Martel 2022). As shown in Table 2, the performance of different text encoders was generally comparable, with BioBERT consistently achieving the best results. Among image encoders, UNI, a vision foundation model pretrained on histopathological images, achieved the highest overall performance, showing a notable improvement over other models.

(2) Gene Semantic Representation. To evaluate textual representations of genes, we compared model performance using three prompt strategies: (a) gene summaries without constraints on function or phenotype, (b) gene symbols without summaries, and (c) gene summaries enriched with specific constraints on function and phenotype. We also evaluated four widely used LLMs: DeepSeek-R1 (DeepSeek-AI 2025), DeepSeek-v3 (Liu et al. 2024), LLaMA2 (7B-chat-hf) (Touvron et al. 2023), and GPT-4o (OpenAI 2024). As shown in Table 3, our proposed prompt strategy (c) achieved the best performance, as it effectively captures more informative and concise gene semantics. Among the LLMs, GPT-4o consistently outperformed the others.

(3) Individual Modules. We evaluated the contributions of key components in our model, including multi-scale spatial context, gene semantic features, contrastive learning, and the use of text features as Key and Value (referred to as text as KV) in dual-path contrastive learning. As shown in Table 4, removing multi-scale spatial context, gene semantics, or contrastive learning led to a decline in performance in terms of PCC. Interestingly, removing contrastive learning resulted in a slight improvement in MAE, but our configuration achieved the best overall performance across all metrics. Additionally, using text features as Query rather than as Key/Value proved more effective for multimodal integration. These findings highlight the effectiveness of our architectural design and the importance of each module.

(4) Fusion Strategy and Loss Design. We investigated the impact of different fusion strategies and loss function designs, as shown in Table 5. Specifically, we evaluated the

	Error		PCC			
Module	MAE↓	MSE↓	ALL↑	HPG↑	HEG↑	HVG↑
w/o multi-scale	0.350	0.210	0.117	0.210	0.101	0.112
w/o text	0.343	0.201	0.210	0.372	0.177	0.233
w/o contrast.	**0.320**	**0.179**	0.209	0.380	0.187	0.231
Text as KV	0.333	0.186	0.216	0.379	0.182	0.242
Ours	0.322	**0.179**	**0.219**	**0.387**	**0.200**	**0.244**

Table 4: Ablation studies on individual modules.

	Error		PCC			
Fusion Strategy	MAE↓	MSE↓	ALL↑	HPG↑	HEG↑	HVG↑
Concat.	0.336	0.189	0.154	0.292	0.048	0.178
Concat.+Trans.	0.331	0.191	0.214	0.380	0.189	0.236
Sum.	0.326	**0.179**	0.151	0.282	0.053	0.171
Sum.+Trans.	0.329	0.188	**0.221**	0.383	0.198	**0.247**
Cross Atten.	**0.322**	**0.179**	0.219	**0.387**	**0.200**	0.244
Loss Design	MAE↓	MSE↓	ALL↑	HPG↑	HEG↑	HVG↑
Fixed weights	0.338	0.191	0.148	0.276	0.053	0.168
w/o distillation	0.336	0.195	0.217	0.382	**0.202**	**0.251**
Ours	**0.322**	**0.179**	**0.219**	**0.387**	0.200	0.244

Table 5: Ablation studies on fusion methods and weights.

cross-attention mechanism used in our contrastive alignment module against four alternatives: addition (Sum.), concatenation (Concat.), concatenation followed by a transformer layer (Concat. + Trans.), and addition followed by a transformer layer (Sum. + Trans.). For the loss functions, we ablated two key components, our dynamic weight balancing strategy and the knowledge distillation loss, to assess their contributions. The results show a consistent performance drop across metrics, particularly in MAE and MSE, when alternative fusion strategies or simplified loss designs are used. In contrast, our full model design achieves the best overall performance in PCC, while also maintaining the lowest MAE and MSE, demonstrating the effectiveness of both our fusion strategy and our loss function design.

Conclusion

In this study, we propose DKAN, a dual-path knowledge-augmented contrastive alignment framework that integrates high-level biological gene knowledge into multimodal feature alignment for spatial gene expression prediction. Comprehensive experiments demonstrate the superior performance of DKAN compared to existing state-of-the-art methods, highlighting the effectiveness of structured biological priors in enhancing cross-modal representation learning. This approach offers a practical pathway for linking histological morphology with spatial gene expression, supporting future discoveries in tissue microenvironments and biomarker identification.

Acknowledgements

This work was supported by the Institute of Digital Medicine, City University of Hong Kong, the Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA) and City University of Hong Kong internal grant 7005967.

References

Andersson, A.; Larsson, L.; Stenbeck, L.; Salmén, F.; Ehinger, A.; Wu, S. Z.; Al-Eryani, G.; Roden, D.; Swarbrick, A.; Borg, Å.; et al. 2021. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature communications*, 12(1): 6012.

Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F. K.; Jaume, G.; Song, A. H.; Chen, B.; Zhang, A.; Shao, D.; Shaban, M.; Williams, M.; Oldenburg, L.; Weishaupt, L. L.; Wang, J. J.; Vaidya, A.; Le, L. P.; Gerber, G.; Sahai, S.; Williams, W.; and Mahmood, F. 2024. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3): 850–862.

Chung, Y.; Ha, J. H.; Im, K. C.; and Lee, J. S. 2024. Accurate Spatial Gene Expression Prediction by Integrating Multi-Resolution Features . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11591–11600. Los Alamitos, CA, USA: IEEE Computer Society.

Ciga, O.; Xu, T.; and Martel, A. L. 2022. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7: 100198.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

He, B.; Bergenstråhle, L.; Stenbeck, L.; Abid, A.; Andersson, A.; Borg, Å.; Maaskola, J.; Lundeberg, J.; and Zou, J. 2020. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8): 827–834.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269. Los Alamitos, CA, USA: IEEE Computer Society.

Huang, Z.; Bianchi, F.; Yuksekgonul, M.; Montine, T. J.; and Zou, J. 2023. A visual–language foundation model for pathology image analysis using medical Twitter. *Nature Medicine*, 1–10.

Ji, A. L.; Rubin, A. J.; Thrane, K.; Jiang, S.; Reynolds, D. L.; Meyers, R. M.; Guo, M. G.; George, B. M.; Mollbrink, A.; Bergenstråhle, J.; et al. 2020. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *cell*, 182(2): 497–514.

Jia, Y.; Liu, J.; Chen, L.; Zhao, T.; and Wang, Y. 2023. THItoGene: a deep learning method for predicting spatial transcriptomics from histological images. *Briefings in Bioinformatics*, 25(1): bbad464.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, F.; Hu, Z.; Chen, W.; and Kak, A. 2023. Adaptive Supervised PatchNCE Loss for Learning H&E-to-IHC Stain Translation with Inconsistent Groundtruth Image Pairs. In *Medical Image Computing and Computer Assisted Intervention(MICCAI)*, 632–641.

Liang, Y.; Wang, Y.; Zhang, Y.; Ye, F.; Luo, D.; Li, Y.; Jin, Y.; Han, D.; Wang, Z.; Chen, B.; et al. 2023. HSPB1 facilitates chemoresistance through inhibiting ferroptotic cancer cell death and regulating NF-κB signaling pathway in breast cancer. *Cell Death & Disease*, 14(7): 434.

Lin, Y.; Luo, L.; Chen, Y.; Zhang, X.; Wang, Z.; Yang, W.; Tong, M.; and Yu, R. 2024. ST-Align: A Multimodal Foundation Model for Image-Gene Alignment in Spatial Transcriptomics. arXiv:2411.16793.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report.

Lu, M. Y.; Chen, B.; Williamson, D. F.; Chen, R. J.; Liang, I.; Ding, T.; Jaume, G.; Odintsov, I.; Le, L. P.; Gerber, G.; et al. 2024. A visual-language foundation model for computational pathology. *Nature Medicine*, 30: 863–874.

Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; and Liu, T.-Y. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6): bbac409.

Mejia, G.; Cárdenas, P.; Ruiz, D.; Castillo, A.; and Arbeláez, P. 2023. SEPAL: Spatial Gene Expression Prediction from Local Graphs. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2286–2295. Paris, France: IEEE. ISBN 979-8-3503-0744-3.

Min, W.; Shi, Z.; Zhang, J.; Wan, J.; and Wang, C. 2024. Multimodal contrastive learning for spatial gene expression prediction using histology images. *Briefings in Bioinformatics*, 25(6): bbae551.

Moses, L.; and Pachter, L. 2022. Museum of spatial transcriptomics. *Nature Methods*, 19(5): 534–546.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2024. GPT-4o. https://openai.com. Accessed: 2025-01-17.

Pang, M.; Su, K.; and Li, M. 2021. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Sayers, E. W.; Beck, J.; Bolton, E. E.; Brister, J. R.; Chan, J.; Connor, R.; Feldgarden, M.; Fine, A. M.; Funk, K.; Hoffman, J.; et al. 2024. Database resources of the National Center for Biotechnology Information in 2025. *Nucleic acids research*, 53(D1): D20.

Schmauch, B.; Romagnoni, A.; Pronier, E.; Saillard, C.; Maillé, P.; Calderaro, J.; Kamoun, A.; Sefta, M.; Toldo, S.; Zaslavskiy, M.; Clozel, T.; Moarii, M.; Courtiol, P.; and Wainrib, G. 2020. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nature Communications*, 11(1): 3877.

Ståhl, P. L.; Salmén, F.; Vickovic, S.; Lundmark, A.; Navarro, J. F.; Magnusson, J.; Giacomello, S.; Asp, M.; Westholm, J. O.; Huss, M.; Mollbrink, A.; Linnarsson, S.; Codeluppi, S.; Åke Borg; Pontén, F.; Costea, P. I.; Sahlén, P.; Mulder, J.; Bergmann, O.; Lundeberg, J.; and Frisén, J. 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294): 78–82.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Wang, H.; Du, X.; Liu, J.; Ouyang, S.; Chen, Y.-W.; and Lin, L. 2024. M2OST: Many-to-one Regression for Predicting Spatial Transcriptomics from Digital Pathology Images. arXiv:2409.15092.

Xie, R.; Pang, K.; Chung, S.; Perciani, C.; MacParland, S.; Wang, B.; and Bader, G. 2023. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Advances in Neural Information Processing Systems*, 36: 70626–70637.

Yang, Y.; Hossain, M. Z.; Li, X.; Rahman, S.; and Stone, E. 2024a. Spatial Transcriptomics Analysis of Zero-Shot Gene Expression Prediction. In Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; and Schnabel, J. A., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 492–502. Cham: Springer Nature Switzerland. ISBN 978-3-031-72083-3.

Yang, Y.; Hossain, M. Z.; Stone, E.; and Rahman, S. 2024b. Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. *Pattern Recogn.*, 145(C).

Yang, Y.; Hossain, M. Z.; Stone, E. A.; and Rahman, S. 2023. Exemplar Guided Deep Neural Network for Spatial Transcriptomics Analysis of Gene Expression Prediction . In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5028–5037. Los Alamitos, CA, USA: IEEE Computer Society.

Zeng, Y.; Wei, Z.; Yu, W.; Yin, R.; Yuan, Y.; Li, B.; Tang, Z.; Lu, Y.; and Yang, Y. 2022. Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5): bbac297.

Zhang, W.; Chen, T.; Xu, W.; and Li, X. 2024a. SAMamba: Integrating State Space Model for Enhanced Multi-modal Survival Analysis. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1334–1341. IEEE.

Zhang, W.; Hui, T. H.; Tse, P. Y.; Hill, F.; Lau, C.; and Li, X. 2024b. High-Resolution Medical Image Translation via Patch Alignment-Based Bidirectional Contrastive Learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 178–188. Springer.

Zhang, W.; Liu, X.; Chen, T.; Xu, W.; Sakal, C.; Nie, X.; Wang, L.; and Li, X. 2025a. Bridging Imaging and Genomics: Domain Knowledge Guided Spatial Transcriptomics Analysis. *Information Fusion*, 103746.

Zhang, W.; Xu, W.; Chen, T.; Sakal, C.; and Li, X. 2025b. Integrating images and genomics for multi-modal cancer survival analysis via mixture of experts. *Information Fusion*, 103521.

Zhang, X.-X.; Luo, J.-H.; and Wu, L.-Q. 2022. FN1 overexpression is correlated with unfavorable prognosis and immune infiltrates in breast cancer. *Frontiers in Genetics*, 13: 913659.