Transitive Representation Learning Enhances Histopathology Annotation

Anonymous Authors[1]

Abstract

AI-driven disease characterization in histopathology promises to assist in clinical decision making, but its performance is limited by the scarcity of detailed annotations. In contrast, single-cell gene expression provides expressive and interpretable labels that compensate this scarcity, but assays are costly and rarely acquired in clinical workflows. To overcome this gap, we propose to bridge these data sources using a trimodal contrastive learning framework that aligns histopathology images, gene expression profiles, and natural-language descriptions. Our training data combines atlas-scale datasets of (i) spatially-resolved gene expression paired with histopathology images, and (ii) single-cell gene expression with curated annotations. Together, these data induce an alignment between image and text modalities, which we leverage for zero-shot image annotation tasks, such as the identification of immune cells. We present a sufficient condition under which this transfer can succeed and assess the performance of our approach against established baselines. We predict cell types at 15.4% improved relative AUROC over leading pathology vision language models. Our method also exhibits significant gains across diverse prediction tasks in low-data regimes, when combining training data from all three modality pairs. Our work thus establishes *transitive representation learning* as an effective strategy to enhance histopathology interpretation.

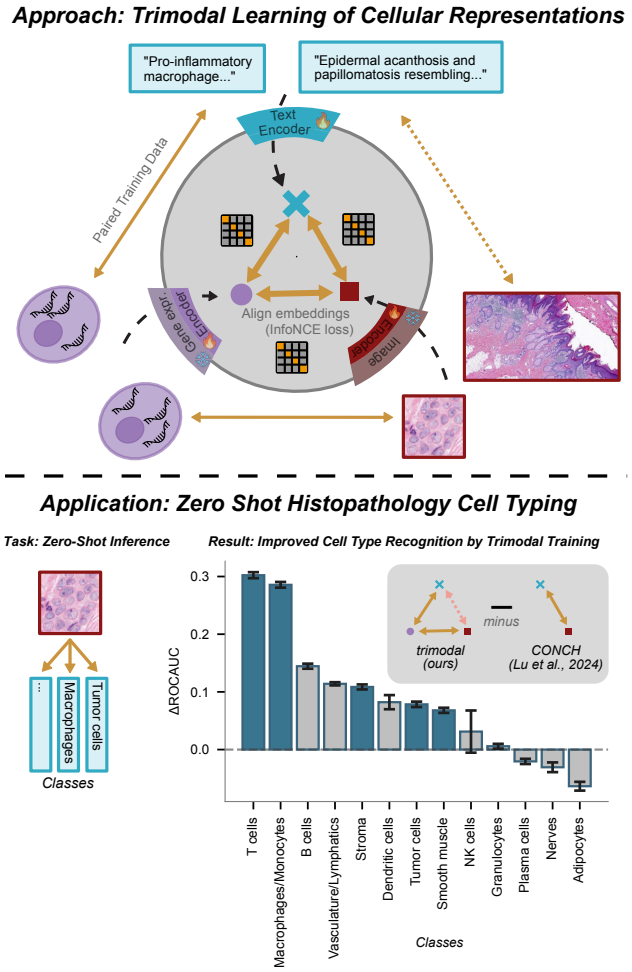

Figure 1. **Top:** Overview of the proposed trimodal training setup. We train on *disjoint* paired datasets that cover the edges *image* ↔ *gene expression* , *gene expression* ↔*text* , and (optionally) *image* ↔ *text*. **Bottom:** Per-class $AUROC$ difference for zero-shot cell type prediction in histopathology image patches, comparing a trimodal model trained on *image*↔*gene expression* and *gene expression* ↔ *text* versus a gold-standard pathology vision-language model (Lu et al., 2024). Blue bars indicate significant ($p < 0.05$) differences across n=109 sample images in the benchmark dataset.

1. Introduction

Histopathology is a cornerstone of clinical diagnosis and biomedical discovery – stained tissue biopsies are imaged under a microscope and examined to detect cancer, and characterize inflammatory and degenerative disease. Recent pathology foundation models and vision-language models, trained on large corpora of images and accompanying reports, promise to support this workflow through scalable retrieval, summarization, and zero-shot recognition of tissue patterns (Huang et al., 2023; Lu et al., 2024; Xiang et al.,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2025).

A central bottleneck in training these models is the limited availability of supervision for cellular interpretation. Most labels and captions are assigned to large regions that contain hundreds to thousands of cells (Ikezogwo et al., 2023; Huang et al., 2023). Yet, clinically relevant phenomena, such as immune infiltration, stromal remodeling, and tumor microenvironmental niches, rely on the characterization of few or even individual cells.

Single-cell gene expression profiling provides a complementary perspective. These data provide mechanistic biological interpretability beyond what image annotations typically afford. Large-scale annotation efforts furthermore facilitate their interpretation to derive cell types, activation states, and biological contexts. However, gene expression profiling assays are substantially more expensive and complex than routine H&E microscopy, limiting their widespread clinical deployment.

Transferring these gene expression-based annotations to histopathology analysis could facilitate cell-level interpretation, which could drastically elevate the predictive value of this routinely obtained data modality.

In this work, we demonstrate this annotation transfer, leveraging spatial transcriptomics as a bridge. Spatial transcriptomics jointly captures gene expression ($\mathcal{G}$) with microscopy images ($\mathcal{I}$), yielding paired ($\mathcal{I} \leftrightarrow \mathcal{G}$ samples at near-single-cell resolution. Although textual annotations ($\mathcal{T}$) for spatial transcriptomics data are rare, single-cell resources provide strong $\mathcal{G} \leftrightarrow \mathcal{T}$ supervision. Combined, these two large paired data resources share the gene expression modality, suggesting a transitive route to learning image$\mathcal{I} \leftrightarrow \mathcal{T}$ relationships through indirect supervision (Fig. 1).

Our goal is to learn an embedding $\mathcal{I} \leftrightarrow \mathcal{T}$ that supports zero-shot cellular annotation of images. As direct supervision for $\mathcal{I} \leftrightarrow \mathcal{T}$ is often coarse, we leverage the fine-grained signal provided through $\mathcal{I} \leftrightarrow \mathcal{G}$ and $\mathcal{G} \leftrightarrow \mathcal{T}$ datasets. The challenge lies in using these modality-adjacent datasets to regularize the latent space such that the fine-grained semantic information associated with $\mathcal{G}$ propagates to $\mathcal{I} \leftrightarrow \mathcal{T}$.

We address this with an InfoNCE-based trimodal contrastive objective that jointly aligns $\mathcal{I} \leftrightarrow \mathcal{G}$ and $\mathcal{G} \leftrightarrow \mathcal{T}$ (and optionally $\mathcal{I} \leftrightarrow \mathcal{T}$) in a shared space which we coin as *transitive representation learning*. We provide a theoretical analysis of the InfoNCE loss that confirms the expected information transfer across modalities: as the margins separating contrastive positives from negatives improve on the observed $\mathcal{I} \leftrightarrow \mathcal{G}$ and $\mathcal{G} \leftrightarrow \mathcal{T}$ pairs, an upper bound on the unobserved $\mathcal{I} \leftrightarrow \mathcal{T}$ loss tightens.

Empirically, we train a trimodal model on large, disjoint paired datasets spanning all three edges. Our model outperforms gold-standard pathology vision-language models on histopathology cell typing across 10 out of 13 cell types (max. AUROC 0.734 versus 0.670, see Table 2 and Fig. 1). In addition, we observe that when task-matching bimodal data is scarce, additional transitive representation learning improves downstream performance. Our results demonstrate a way to enhance shared representation learning in biomedical settings where annotations are often limited.

Contributions.

- We describe and assess a trimodal contrastive learning framework that is trained on disjoint paired datasets to transfer gene expression-based textual annotations to histopathology.
- We provide an analysis of the InfoNCE loss that gives sufficient conditions for transitive representation learning across modalities. This also highlights failure modes, such as the lack of overlap in the shared modalities, which we demonstrate empirically through text harmonization.
- We demonstrate improved zero-shot cell type prediction in histopathology and consistent gains in low-data regimes across established multimodal benchmarks.

We will make source code and model weights publicly available upon publication.

2. Related Work

Multimodal contrastive learning has enabled scalable alignment of heterogeneous data sources by learning a shared embedding space from paired observations (Jia et al., 2021; Radford et al., 2021). In biomedicine, this paradigm has been applied to connect histopathology with text reports (Huang et al., 2023; Lu et al., 2024; Xiang et al., 2025) and with spatial transcriptomics (Wang et al., 2025; Xie et al., 2023; Zeng et al., 2022), and to interface single-cell RNA-seq with textual annotations for zero-shot cellular analysis (Heimberg et al., 2025; Schaefer et al., 2025b; Zhao et al., 2024). Our work builds on these advances and studies how integration of a third modality can serve as a supervision-rich bridge that improves over task-matched paired data.

In the subsections below, we contextualize related work on the integration of more than two datasets.

2.1. Simultaneous Alignment of Comprehensive Measurements

When all modalities co-occur for the same sample, a natural approach is to jointly align them altogether. Recent methods

propose geometric regularization or multi-view matching objectives to improve cross-modal structure (Cicchetti et al., 2025; Piran et al., 2024). However, in cellular biology and clinical pathology, it is uncommon to observe images, gene expressions, and rich textual annotations simultaneously at scale. In this work, we acknowledge that paired data are observed more commonly and ask how to use them to infer missing relationships.

2.2. Hub-based Alignment with a Single Shared Modality

Complementary work has integrated more than two modalities using pair-wise data, learning a shared embedding space with images as a shared hub modality (Girdhar et al., 2023). This demonstrates a form of transitive representation learning, with images serving as the bridge modality, and shows that training-unseen relationships can be learned.

We build on this concept in the context of cellular biology. We provide a theoretical analysis of the conditions and failure modes for transitive learning to align representations, and support our findings with empirical results on diverse benchmark tasks.

2.3. Biomedical Frameworks Beyond Two Modalities

Integration of more than two modalities has been explored in biomedical contexts. Wang et al. (2023) freeze unimodal backbones and learn transformations between diverse modalities such as drugs and phenotypes from a knowledge-graph. Chen et al. (2025) combine diverse patient-centric modalities using both contrastive learning and task-specific objectives and a "mixture-of-modality-experts" architecture.

In contrast to these works, we learn a focused embedding space of cellular states and assess the effect of transitive representation learning on downstream task performance.

3. Method

3.1. Trimodal Contrastive Learning Framework

We learn a shared embedding space in which any two of the three modalities, histopathology images ($\mathcal{I}$), gene expression profiles ($\mathcal{G}$), and natural language ($\mathcal{T}$), can be compared via cosine similarity. The shared space should support (i) cross-modal retrieval (e.g., retrieve gene expressions or captions most compatible with an image patch), and (ii) prompt-based, zero-shot prediction (e.g., score a patch against a vocabulary of cell types). We build on paired data contrastive learning (Radford et al., 2021) and use modality-specific encoders followed by lightweight projection heads that map into a common d-dimensional space (Zhai et al., 2022).

Encoders: Let $\phi_{\mathcal{I}}, \phi_{\mathcal{G}}, \phi_{\mathcal{T}}$ denote the modality-specific encoders. Given inputs $x_i^{\mathcal{I}}, x_i^{\mathcal{G}}, x_i^{\mathcal{T}}$, our model produces unit-normalized embeddings $z_i^{\mathcal{X}} \in \mathbb{R}^{2048}$ in a shared space: $z_i^{\mathcal{I}} := \phi_{\mathcal{I}}(x_i^{\mathcal{I}})$, $z_i^{\mathcal{T}} := \phi_{\mathcal{T}}(x_i^{\mathcal{T}})$, and $z_i^{\mathcal{G}} := \phi_{\mathcal{G}}(x_i^{\mathcal{G}})$. These encoders are implemented through modality-specific backbone models stacked with two-layer MLP projection heads. We chose the following pretrained backbones in this study, which also define the input format for each modality

- **Image encoder:** *UNI 2*, a pretrained vision transformer for histopathology patches of size 224×224 pixels (typically comprising $\sim$10–50 cells per patch) (Chen et al., 2024).
- **Text encoder:** *BioBERT*, which embeds short natural-language descriptions of cellular phenotypes and tissue features (Lee et al., 2020).
- **Gene expression encoder:** *Geneformer*, which represents gene expression profiles as ranked gene tokens (Theodoris et al., 2023).

We use the inner product $\langle \cdot, \cdot \rangle$ as similarity. Under unit normalization this equals cosine similarity.

Zero-shot scoring. Once trained, the model supports zero-shot prediction by comparing embeddings across modalities. For example, to predict a cell type for an image patch, we encode the patch into $z^{\mathcal{I}}$ and encode a set of candidate cell types as text prompts $\{x_c^{\mathcal{T}}\}$ into $\{z_c^{\mathcal{T}}\}$. We then score each candidate by cosine similarity $\langle z^{\mathcal{I}}, z_c^{\mathcal{T}} \rangle$ and choose the highest-scoring label. Analogous retrieval is possible across all modality pairs.

3.2. Training

We optimize a composite contrastive objective that aligns each observed modality pair. Following (Schaefer et al., 2025b; Zhai et al., 2022), we freeze the image and gene expression backbones and train the text encoder and all projection heads, keeping computation manageable while adapting the shared space to the paired datasets. Let $\lambda_{\mathcal{I}\leftrightarrow\mathcal{T}} = \lambda_{\mathcal{I}\leftrightarrow\mathcal{G}} = \lambda_{\mathcal{T}\leftrightarrow\mathcal{G}} = 1.0$ be scalar weights (set to 1.0 in this study) and τ a learnable temperature. In this work, we define

$$\mathcal{L} = \lambda_{\mathcal{I}\leftrightarrow\mathcal{T}}\,\mathcal{L}_{\mathcal{I}\leftrightarrow\mathcal{T}} + \lambda_{\mathcal{I}\leftrightarrow\mathcal{G}}\,\mathcal{L}_{\mathcal{I}\leftrightarrow\mathcal{G}} + \lambda_{\mathcal{T}\leftrightarrow\mathcal{G}}\,\mathcal{L}_{\mathcal{T}\leftrightarrow\mathcal{G}}. \quad (1)$$

Each pairwise term $\mathcal{L}_{\mathcal{X}\leftrightarrow\mathcal{Y}}$ for $(\mathcal{X}, \mathcal{Y}) \in \{(\mathcal{I}, \mathcal{T}), (\mathcal{I}, \mathcal{G}), (\mathcal{T}, \mathcal{G})\}$ is an InfoNCE loss (van den Oord et al., 2018). For a positive pair $(z_{\mathcal{X}}, z_{\mathcal{Y}})$ in a minibatch, we treat the paired sample as the unique positive and all other samples in the batch as negatives:

$$\mathcal{L}_{\mathcal{X}\leftrightarrow\mathcal{Y}} = -\log \frac{\exp(\langle z^{\mathcal{X}}, z^{\mathcal{Y}} \rangle / \tau)}{\sum_k \exp(\langle z^{\mathcal{X}}, z_k^{\Updownarrow\parallel} \rangle / \tau)}. \quad (2)$$

Table 1. Datasets used for trimodal training. Each dataset provides paired data for two of the three modalities.

Name	Description	Num. Pairs
HEST-1K (Subset) (Jaume et al., 2024)	Gene expression ↔ Image: Histopathology patches paired with spatially resolved gene expression profiles.	921,154 (384 samples)
CellWhisperer Dataset (Schaefer et al., 2025b)	Gene expression ↔ Text: Gene expression profiles paired with textual annotations.	1,082,413
QUILT-1M (Subset) (Ikezogwo et al., 2023)	Image ↔ Text: Histopathology images with matched textual descriptions. (Not used for cell type prediction task)	104,362

We train across the different datasets in an interleaved manner, such that minibatches contain data points reflecting different modality pairs, providing a rich contrastive signal.

Training is performed with a cosine learning schedule with initial warmup for the first 3% of training steps and a learning rate of 10^{-5}. Training batch size was 512.

If not indicated otherwise, we freeze the backbones for the image and gene expression encoders, while keeping the text encoder backbone and all projection layers unfrozen, and trained for 4 epochs.

3.3. Datasets & Benchmarks

We train on different combinations of three large bimodal datasets (Table 1), covering the edges $\mathcal{I} \leftrightarrow \mathcal{G}$, $\mathcal{G} \leftrightarrow \mathcal{T}$, and $\mathcal{I} \leftrightarrow \mathcal{T}$. Specifically, we use (i) HEST-1K (Jaume et al., 2024), which pairs histopathology image patches with Visium spatial transcriptomics readouts (921,154 pairs across 384 samples); (ii) the CellWhisperer dataset (Schaefer et al., 2025b), which pairs gene expressions with textual annotations (1,082,413 pairs); and (iii) QUILT-1M (Ikezogwo et al., 2023), which pairs histopathology images with captions (104,362 pairs after filtering).

All three datasets span diverse biological and pathological contexts. For compatibility with the $\mathcal{T} \leftrightarrow \mathcal{G}$ data, we filter the $\mathcal{G} \leftrightarrow \mathcal{I}$ HEST-1K data to samples that provide expression profiles across the whole genome. At the spatial level, such data typically resolve close to the single-cell scale with individual *patches* representing on the order of ∼10 cells. For the $\mathcal{T} \leftrightarrow \mathcal{I}$ QUILT-1M dataset, which contains heterogeneous image sources and formatting, we follow the metadata provided by the authors to filter for high-magnification histological images while excluding non-histological content.

We evaluate on benchmark tasks that cover each modality pairs, including both retrieval-style settings (e.g., retrieve the correct gene expression profile for an image patch) and classification-style settings that can be expressed through natural-language prompts and underline the zero-shot capabilities of our approach. Our first experiment (Fig. 1) evaluates 88,014 image patches from 109 annotated colorectal cancer tissue images (35 patients). Ground truth cell type labels were derived through a proteomics-based assay performed in parallel to the histopathology imaging (Koreuber et al., 2025; Schürch et al., 2020). Table 3 provides details on the benchmark used in our second experiment (Fig. 6).

3.4. Analysis of Transitive Representation Learning

Here, we study when and how aligning the observed modality pairs $\mathcal{I} \leftrightarrow \mathcal{G}$ and $\mathcal{G} \leftrightarrow \mathcal{T}$ induces good alignment for the unobserved pair $\mathcal{I} \leftrightarrow \mathcal{T}$. Let z_i^m denote the unit-normalized embedding of sample i in modality $m \in \{\mathcal{I}, \mathcal{G}, \mathcal{T}\}$.

First, we show why training the encoders on the observed pairs also reduces the loss on the unobserved pairs. For simplicity, we assume perfect overlaps in the shared modality and we quantify the quality of encoder training via a uniform margin that measures how well the encoders separate positive from negative pairs. We then discuss implications on training with datasets that span all three edges. Finally, we analyze the impact of imperfect overlaps in the shared modality, which is an expected scenario in real-world data. We complement our analyses with empirical insights in Section 4.

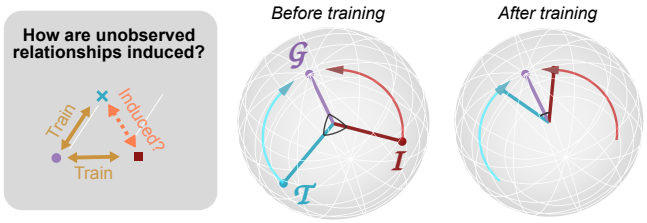

Figure 2. Intuition for transitive representation learning in a schematic 3-dimensional shared embedding space. As the InfoNCE loss shrinks the angular gaps on the observed pairs, the unobserved modalities become closer as well.

Why transitive representation learning works. As illustrated in Fig. 2, the composite objective intuitively aligns the unobserved pair by shrinking positive-vs-negative angular gaps on the observed pairs. In the positive case, if $z_\mathcal{T}$ is similar to $z_\mathcal{G}$ and $z_\mathcal{G}$ is similar to $z_\mathcal{I}$, then $z_\mathcal{T}$ and $z_\mathcal{I}$ will likely be similar as well. The caveat is that to achieve a

low InfoNCE, we also need to implicitly shrink the angular distance for the negative pairs without training on the unobserved pairs. Below we make this intuition precise and show when transfer is expected to succeed. Rather than deriving a strict guarantee of transfer under potentially unrealistic assumptions, we use a simplified setting to verify the intuition and guide the design of our experiments.

Encoder quality is captured by separation margins (ϵ, η) between positive and negative pairs, for i and $j \neq i$,

$$\langle z_i^{\mathcal{I}}, z_i^{\mathcal{G}} \rangle \geq 1-\epsilon, \qquad \langle z_i^{\mathcal{G}}, z_i^{\mathcal{T}} \rangle \geq 1-\epsilon, \qquad \text{(P)}$$

$$\langle z_i^{\mathcal{I}}, z_j^{\mathcal{G}} \rangle \leq \eta, \qquad \langle z_i^{\mathcal{G}}, z_j^{\mathcal{T}} \rangle \leq \eta. \qquad \text{(N)}$$

Thus, true pairs are close (similarity at least $1-\epsilon$), while mismatched pairs are bounded by η.

Under these margins, we bound the per-sample InfoNCE loss for the unobserved $\mathcal{I} \to \mathcal{T}$ query in terms of (ϵ, η), making the intuition precise and capturing a sufficient condition for transfer. Here, we use the gene expression encoder as a semantic reference that links the other two modalities; this choice depends on the available data.

Lemma 3.1 (Per-sample InfoNCE bound for $\mathcal{I} \to \mathcal{T}$)**.** *Given a query $x_i^{\mathcal{I}}$, its matched (but unobserved) $x_i^{\mathcal{T}}$, and N unpaired $x_{j_1}^{\mathcal{T}}, \ldots, x_{j_N}^{\mathcal{T}}$ contrastive negatives, define*

$$s^+ = \langle z_i^{\mathcal{I}}, z_i^{\mathcal{T}} \rangle, \qquad s_k^- = \langle z_i^{\mathcal{I}}, z_{j_k}^{\mathcal{T}} \rangle.$$

Let $\tau > 0$ be the temperature. If (P) and (N) hold with margins (ϵ, η), then the loss ℓ_i across modalities $\mathcal{I}$ and $\mathcal{T}$ is bounded by ϵ and η:

$$\ell_i = \log\Big(1 + \sum_{k=1}^{N} e^{(s_k^- - s^+)/\tau}\Big) \leq \log\Big(1 + N\, e^{r(\epsilon,\eta)/\tau}\Big),$$

where $r(\epsilon, \eta) = q(\epsilon, \eta) - p(\epsilon)$ with $p(\epsilon) = 2(1-\epsilon)^2 - 1$ and $q(\epsilon, \eta) = \max\{\eta, (1-\epsilon)\eta\} + \sqrt{2\epsilon - \epsilon^2}$.

Holding τ and N fixed, the bound decreases as observed positives tighten ($\epsilon \downarrow$) and observed negatives are pushed away ($\eta \downarrow$). As $\epsilon \to 0$ and $\eta \to -1$, it approaches the optimal loss $\log(1 + Ne^{-2/\tau})$, which tends to 0 as $\tau \to 0$. Proof details are in Appendix A; conceptually, we decompose each embedding along the $\mathcal{G}$ *reference direction* and bound positive and negative projections and residuals.

We empirically probe these conditions in Section 4.

Implications for training on all three edges. Under our trimodal objective, training can be performed jointly for datasets than span all three paired combinations. In this situation, any given data pair receives supervision directly, as well as transitively, through the combination of the other two pairs. Our analysis illustrates how these two signals complement each other. Imagine this as a sequential learning protocol, where optimization on the transitive pairs $\mathcal{I} \leftrightarrow \mathcal{G}$ and $\mathcal{G} \leftrightarrow \mathcal{T}$ tightens an upper bound on the loss of the third pair $\mathcal{I} \to \mathcal{T}$, effectively pre-conditions the shared space. Including data for $\mathcal{I} \leftrightarrow \mathcal{T}$ then provides direct signal that further refines alignment. As a result, combined training requires less $\mathcal{I} \leftrightarrow \mathcal{T}$ supervision to achieve strong alignment (Fig. 3).

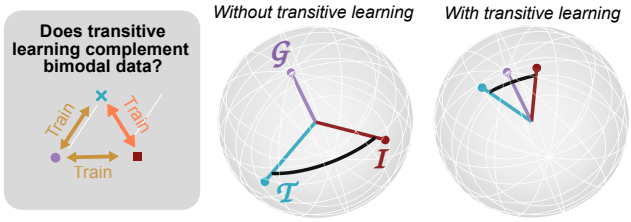

Figure 3. Leveraging transitive pretraining: transitive learning preconditions the unseen modality pair to be close, easing subsequent fine-tuning on the third edge. Black arc between $\mathcal{I} \leftrightarrow \mathcal{T}$ indicates angular distance to be refined by training.

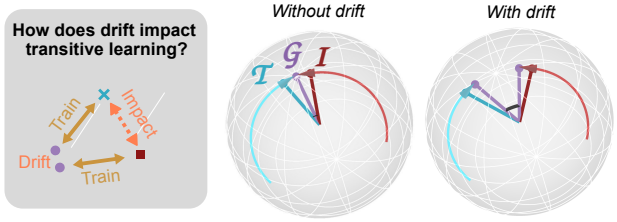

Figure 4. Impact of drift in the overlapping modality: imperfect overlap introduces an additive slack in the unobserved pair loss bound.

Effect of imperfect overlap in $\mathcal{G}$. Although above we assumed perfect overlaps in $\mathcal{G}$ across datasets, in reality this is rarely the case. If the $\mathcal{G}$ reference directions differ, an extra change-of-reference projection introduces a slack in the exponent gap r. In the high-overlap regime $\langle z_i^{\mathcal{G}}, z_j^{\mathcal{G}} \rangle \geq 1-\delta$, the degradation is $O(\sqrt{\delta})$ (with smaller $O(\delta)$ and mixed $O(\sqrt{\epsilon\delta})$ terms), vanishing as $\delta \to 0$.

Our defined degradation of the bound thus motivates data curation and diagnostics of overlapping data modalities, such as through examining the distribution similarity.

4. Experiments

Our approach allows for fine-grained analysis of histopathology images with free-text queries. Applying a trimodal model to annotate a lung cancer histopathology sample (Dawo et al., 2025; Schaefer et al., 2025a), we observe it can accurately detect cancer and immune cells by zero-shot querying for specific cell types (Fig. 5).

Here, we assess our trimodal approach systematically through three complementary experiments that directly

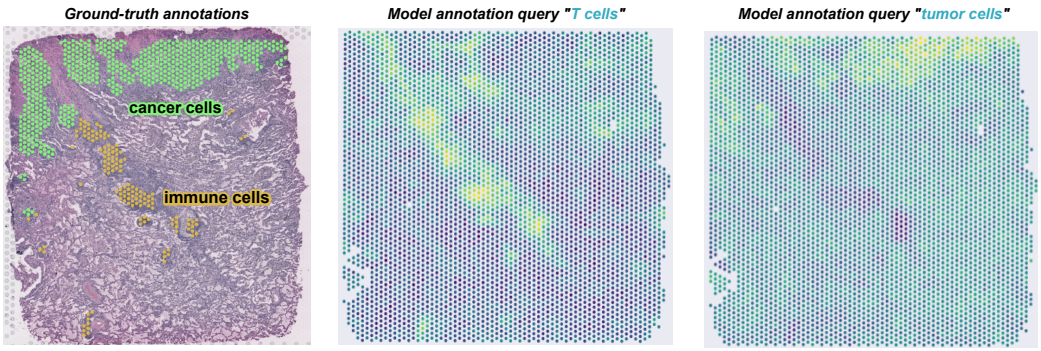

Figure 5. Zero-shot cell type prediction on a held-out lung cancer tissue sample using free-text queries. Inference is performed on patch-level and colormaps show cosine similarities between encoded image patches and each of the two encoded queries. The shown sample was not part of our training data. Ground truth annotations were provided by an expert histopathologist. Original label for immune cells (yellow) was 'tertiary lymphoid structure', a niche characterized by dense immune cell infiltration.

probe the analyses in Section 3.4. First we train on two datasets and rigorously measure its transitively learned cell type detection performance through a benchmark of annotated histopathology images. We compare this model against established and self-trained pathology vision-language models. Second, we assess how transitive pretraining complements with task-matched paired data. We examine this with a focus on training data scarcity, as high-quality data-annotation pairs are often rare. Third, we test the impact of shared data modality overlap on information transfer. We harmonize annotation style in the text modalities across datasets and measure its effect on cross-modal retrieval.

Unless stated otherwise, all evaluations are zero-shot: We query the learned shared embedding space with modality-specific encoders and perform nearest-neighbor retrieval or prompt-based scoring without task-specific fine-tuning.

4.1. Transitive Representation Learning Improves Cell Type Prediction in Histopathology

To assess whether gene expression-mediated supervision improves *cell-level* recognition in histopathology data, we performed zero-shot cell type inference in samples of colorectal cancer biopsies with ground truth cell type annotations (Koreuber et al., 2025; Schürch et al., 2020). Individual patch embeddings were scored against textual embeddings of 13 cell types in the shared embedding space, followed by a comparison to the patch-level cell type labels. We compared our trimodal model to leading pathology vision–language models CONCH and PLIP (Huang et al., 2023; Lu et al., 2024), as they match our model's capability for zero-shot inference of histopathology image patches. To isolate the performance of gene expression-mediated transitive representation learning, we trained our model solely on the $\mathcal{I} \leftrightarrow \mathcal{G}$ and $\mathcal{G} \leftrightarrow \mathcal{T}$ datasets.

Trimodal training improved performance for most cell types over our baselines, with particularly strong gains for several immune and vasculature populations, which are well-captured by gene expression profiling assays. Overall, our model achieves the best AUROC on 10 of 13 cell types and improves the mean AUROC by 15.4% relative to the second-best model (Fig. 1 and Table 2). The highest observed prediction performance is 0.734 (AUROC), demonstrating the challenging nature of this benchmark.

Table 2. Zero-shot cell type prediction performances (mean AUROC scores across $n = 109$ image samples). Best per row is **bold**; second-best is underlined.

Cell Type	*Trimodal*	CONCH	PLIP
Adipocytes	0.473	**0.545**	0.467
B cells	**0.726**	0.584	0.538
Dendritic cells	**0.611**	0.525	0.485
Granulocytes	**0.636**	0.631	0.428
Macrophages/ Monocytes	**0.712**	0.426	0.406
NK cells	**0.457**	0.437	0.395
Nerves	0.541	**0.567**	0.547
Plasma cells	0.639	**0.660**	0.496
Smooth muscle	**0.734**	0.670	0.630
Stroma	**0.633**	0.515	0.363
T cells	**0.734**	0.426	0.502
Tumor cells	**0.575**	0.497	0.506
Vasculature/ Lymphatics	**0.720**	0.611	0.579
Mean	**0.630**	0.546	0.488

Cell types with poor recognition performance coincide with known measurement limitations. Adipocytes exhibit few molecules and are therefore poorly captured through gene expression profiling. Abbreviated labels (e.g., "NK cells") can reduce alignment as training data predominantly contains full spelling (e.g., "natural killer cells").

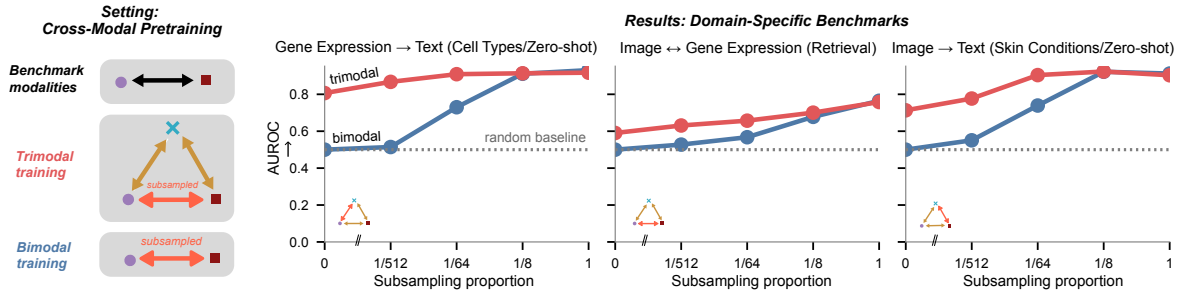

Figure 6. Effect of task-matching data with (red line; trimodal model) and without (blue line; bimodal baseline) transitive data. Task-matching data is provided in different amounts, to simulate low-data scenarios. Modality-matching tasks described in Table 3. Scores are macro-averaged across classes. Baseline performance at x=0 is set to 0.5 (corresponding to fully random AUROC performance).

Table 3. Benchmark datasets used for evaluating performance in low-n data scenarios (Fig. 6)

Name	Task	Description	Ref.
Tabula Sapiens	Gene expression → Text (zero-shot)	Cell type annotation from a large-scale scRNA-seq atlas covering 177 cell types.	(Consortium* et al., 2022)
HEST-1K Benchmark	Gene expression ↔ Image (retrieval)	Gene expression retrieval from histopathology images, covering 8 human cancers and 9 organs.	(Jaume et al., 2024)
Skin Conditions	Image → Text (zero-shot)	16 malignant and nonmalignant skin conditions.	(Kriegsmann et al., 2022), adopted from (Xiang et al., 2025)

We observe similar trends for the F_1 and other metrics and discuss those in Appendix C. We also evaluated alternative query formulations, which led us to use bare cell type labels as those performed the best in our baselines (see Appendix B).

4.2. Task-Matching and Transitive Data Jointly Boost Performance in Low-N Data Regimes

We next asked whether transitive representation learning can compensate for low-n paired data in the modality pair of interest. This question is particularly relevant in biomedical applications, where high-quality paired annotations are often expensive and therefore small in scale.

For this purpose we collected representative benchmark tasks covering each of the three modality pairs (Table 3). We then constructed low-N settings by randomly subsampling the paired target data to fractions $\{1, 1/8, 1/64, 1/512\}$, along with a 0-data control (Fig. 6) and trained two sets of models. First a bimodal model using only the subsampled target pairs, and second a trimodal model that additionally leverages the other two paired datasets through our trimodal objective.

Across all three benchmark tasks, incorporating auxiliary paired supervision via the shared modality (i.e., transitive representation learning) improved performance over the bimodal baseline, with the largest gains appearing in the smallest-data regimes (Fig. 6). At full provided data, the benefits vanish, which we attribute to a combination of saturated benchmarks and complementary information in the transitive and task-matching data modalities.

Overall, our results demonstrate that trimodal learning acts as a strong inductive bias: when direct paired supervision is weak, auxiliary pairs that overlap through the shared modality provide informative structure that improves generalization. Notably, we did not observe this positive effect for the cell type prediction task. We attribute this to a lack of task-relevant information in the $\mathcal{I} \leftrightarrow \mathcal{T}$ QUILT-1M dataset (see Appendix B.

4.3. Transitive Representation Learning Requires Overlapping Modalities

As highlighted in Section 3.4, transfer depends on an overlap in the shared modality: if the text descriptions associated with images and gene expressions deviate, the model does not reliably align the corresponding image and gene expression modalities. In practice, this overlap can be weakened by stylistic inconsistencies (e.g., different vocabularies, levels of specificity, and biological focus) even when the underlying semantics match.

To probe this effect, we curated the labels in the $\mathcal{I} \leftrightarrow \mathcal{T}$ QUILT-1M dataset to better match the labels' style in the $\mathcal{G} \leftrightarrow \mathcal{T}$ dataset. We used an LLM to rewrite labels to mimic

the annotation style of 20 hand-picked $\mathcal{G} \leftrightarrow \mathcal{T}$ examples (Appendix D). We then evaluated how well the curated versus original labels aligned to their corresponding images by measuring their $\mathcal{I} \leftrightarrow \mathcal{T}$ similarity under a model trained only on the other two modality pairs.

Consistent with our expectations, text harmonization yielded significantly higher $\mathcal{I} \leftrightarrow \mathcal{T}$ similarity scores (Fig. 7), which was also reflected in higher retrieval scoring (AUROC 0.695 vs 0.645; scores were computed on a subset of 20,000 pairs for performance reasons).

These results underline the importance of overlapping modalities for transitive representation learning. Targeted data curation can thus function as a practical lever for strengthening the alignment of unobserved modality pairs.

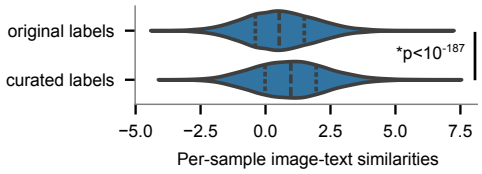

Figure 7. Comparison of cross-modal similarity for curated and original $\mathcal{I} \leftrightarrow \mathcal{T}$ data. Shown is the distribution of temperature-scaled cosine similarities across samples in (Ikezogwo et al., 2023) computed using a trimodal model trained with $\mathcal{I} \leftrightarrow \mathcal{G}$ and $\mathcal{G} \leftrightarrow \mathcal{T}$ data. Black lines indicate distribution quartiles. Statistics: Two-sided Mann-Whitney U

5. Discussion

Gene expression profiling provides rich, mechanistically interpretable labels that can inject cellular semantics into histopathology. By training on paired $\mathcal{I} \leftrightarrow \mathcal{G}$ and $\mathcal{G} \leftrightarrow \mathcal{T}$ data, our model learns an implicit $\mathcal{I} \leftrightarrow \mathcal{T}$ mapping that enables zero-shot cellular annotation from routine H&E using natural-language queries. This establishes a practical route to molecularly grounded interpretation of tissue morphology, linking morphological findings to curated molecular concepts.

Transitive representation learning also improves data efficiency. In low-data regimes, which are common in biomedical settings due to the cost of high-quality labels, leveraging auxiliary paired datasets acts as a strong regularizer. Across 3 benchmarks, gains are most pronounced when target pairs are scarce, which motivates a pretraining strategy in which broadly available paired corpora, for example spatial transcriptomics and atlas annotations, are used to strengthen task-specific bimodal performance in problems with low data availability.

Our analysis and experiments provide insights into when transfer succeeds, or fails. Strong positive margins and low negative similarities on the observed pairs tighten an upper bound on the unobserved $\mathcal{I} \leftrightarrow \mathcal{T}$ InfoNCE loss, explaining how the composite objective aligns training-unseen relationships. Effective transfer further depends on substantive overlap in the shared modality across datasets: when overlap is imperfect, the bound degrades predictably with the mismatch, which we observe empirically. These observations motivate practical remedies such as style harmonization, and protocol-/batch-aware adapters that better cope with heterogeneity in transcriptomic measurements. Simple diagnostics, such as empirical margin estimates and hub-overlap scores in gene expression, can be computed during training to anticipate transitive potential and to guide curation.

While we focus on histopathology, transcriptomics, and text, the trimodal perspective extends naturally to additional measurements. Proteomics or epigenomics could enrich the shared modality with complementary aspects of cell state, and longitudinal or perturbation datasets could introduce causal structure that is not available from observational data alone. We expect that multimodal foundation models for biomedicine will increasingly rely on compositional training setups, where large but incomplete paired datasets are combined to approximate richer supervision than any single dataset can provide, with careful attention to shared-modality overlap, data informativeness, and domain shift.

5.1. Limitations & Future Work

While our results demonstrate that molecularly informed supervision can transfer effectively across modalities, several limitations point to clear opportunities for future work. First, because our image inputs are patch-level, each example aggregates signals from multiple cells, which limits cell-resolved supervision. As spatial data at sub-cellular resolution is increasingly becoming available, we expect our approach to further improve trimodal model perfomance by aligning true single-cell data across all modalities.

Second, we chose state-of-the-art modality backbones and focused this work on a systematic comparison of training regimes. Future work can explore alternative encoders, parameter freezing and adapter mechanisms to better cope with heterogeneity in the data (e.g., bulk, single-cell, and spatial transcriptomics) and potentially improve alignment.

Third, although we evaluate on established benchmarks covering representative modality pairs, the clinical and biological space is far broader. Expanding evaluation to additional tissues and tasks, and out-of-distribution testing across scanners and gene expression profiling assays will be necessary to foster wider data compatibility of our model.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3): 850–862, 2024.

Chen, W., Zhao, Z., Yao, J., Zhang, Y., Bu, J., and Wang, H. Multi-modal medical diagnosis via large-small model collaboration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30763–30773, 2025.

Cicchetti, G., Grassucci, E., and Comminiello, D. A TRIANGLE enables multimodal alignment beyond cosine similarity. *arXiv [cs.LG]*, September 2025. doi: 10.48550/arXiv.2509.24734. URL http://arxiv.org/abs/2509.24734.

Consortium*, T. T. S., Jones, R. C., Karkanias, J., Krasnow, M. A., Pisco, A. O., Quake, S. R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., et al. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.

Dawo, S., Nonchev, K., and Silina, K. 10x visium spatial transcriptomics dataset: Kidney (3) and lung (5) cancer with tertiary lymphoid structures, 2025. URL http://dx.doi.org/10.5281/zenodo.14620362.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Heimberg, G., Kuo, T., DePianto, D. J., Salem, O., Heigl, T., Diamant, N., Scalia, G., Biancalani, T., Turley, S. J., Rock, J. R., et al. A cell atlas foundation model for scalable search of similar human cells. *Nature*, 638(8052): 1085–1094, 2025.

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J., and Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023.

Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P. K., Krishna, R., and Shapiro, L. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017, 2023.

Jaume, G., Doucet, P., Song, A., Lu, M. Y., Almagro Pérez, C., Wagner, S., Vaidya, A., Chen, R., Williamson, D., Kim, A., et al. Hest-1k: A dataset for spatial transcriptomics and histology image analysis. *Advances in Neural Information Processing Systems*, 37:53798–53833, 2024.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Koreuber, J., , Franzen, N., , Reith, J., , , Winklmayr, F. H., Baumann, C., , Schürch, E., , , Kainmüller, C. M., Rumberger, D., , and Lorenz, J. PathoCellBench: A comprehensive benchmark for cell phenotyping. https://papers.miccai.org/miccai-2025/0670-Paper4441, September 2025. Accessed: 2026-1-18.

Kriegsmann, K., Lobers, F., Zgorzelski, C., Kriegsmann, J., Janssen, C., Meliß, R. R., Muley, T., Sack, U., Steinbuss, G., and Kriegsmann, M. Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology*, 12:1022967, 2022.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, February 2020. ISSN 1367-4803,1367-4811. doi: 10.1093/bioinformatics/btz682. URL http://dx.doi.org/10.1093/bioinformatics/btz682.

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. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

OpenAI, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., Chen, M., Cheung, E., Clark, A., Cook, D., Dukhan, M., Dvorak, C., Fives, K., Fomenko, V., Garipov, T., Georgiev, K., Glaese, M., Gogineni, T., Goucher, A., Gross, L., Guzman, K. G., Hallman, J., Hehir, J., Heidecke, J., Helyar, A., Hu, H., Huet, R., Huh, J., Jain, S., Johnson, Z., Koch, C., Kofman, I., Kundel, D., Kwon, J., Kyrylov, V., Le, E. Y., Leclerc, G., Lennon, J. P., Lessans, S., Lezcano-Casado, M., Li, Y., Li, Z., Lin, J., Liss, J., Lily, Liu, Liu, J., Lu, K., Lu, C., Martinovic, Z., McCallum,

L., McGrath, J., McKinney, S., McLaughlin, A., Mei, S., Mostovoy, S., Mu, T., Myles, G., Neitz, A., Nichol, A., Pachocki, J., Paino, A., Palmie, D., Pantuliano, A., Parascandolo, G., Park, J., Pathak, L., Paz, C., Peran, L., Pimenov, D., Pokrass, M., Proehl, E., Qiu, H., Raila, G., Raso, F., Ren, H., Richardson, K., Robinson, D., Rotsted, B., Salman, H., Sanjeev, S., Schwarzer, M., Sculley, D., Sikchi, H., Simon, K., Singhal, K., Song, Y., Stuckey, D., Sun, Z., Tillet, P., Toizer, S., Tsimpourlas, F., Vyas, N., Wallace, E., Wang, X., Wang, M., Watkins, O., Weil, K., Wendling, A., Whinnery, K., Whitney, C., Wong, H., Yang, L., Yang, Y., Yasunaga, M., Ying, K., Zaremba, W., Zhan, W., Zhang, C., Zhang, B., Zhang, E., and Zhao, S. gpt-oss-120b & gpt-oss-20b model card. *arXiv [cs.CL]*, August 2025. doi: 10.48550/arXiv.2508.10925. URL http://arxiv.org/abs/2508.10925.

Piran, Z., Klein, M., Thornton, J., and Cuturi, M. Contrasting multiple representations with the multi-marginal matching gap. *arXiv preprint arXiv:2405.19532*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Schaefer, M., Nonchev, K., Awasthi, A., Burton, J., Koelzer, V. H., Ratsch, G., and Bock, C. Molecularly informed analysis of histopathology images using natural language. *bioRxiv*, 2025a. doi: 10.1101/2025.07.14.664402. URL https://www.biorxiv.org/content/early/2025/07/18/2025.07.14.664402.

Schaefer, M., Peneder, P., Malzl, D., Lombardo, S. D., Peycheva, M., Burton, J., Hakobyan, A., Sharma, V., Krausgruber, T., Sin, C., Menche, J., Tomazou, E. M., and Bock, C. Multimodal learning enables chat-based exploration of single-cell data. *Nature Biotechnology*, 2025b. doi: 10.1038/s41587-025-02857-9.

Schürch, C. M., Bhate, S. S., Barlow, G. L., Phillips, D. J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D. R., Kinoshita, S., Samusik, N., Goltsev, Y., and Nolan, G. P. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*, 182(5):1341–1359.e19, September 2020. ISSN 1097-4172,0092-8674. doi: 10.1016/j.cell.2020.07.005. URL http://dx.doi.org/10.1016/j.cell.2020.07.005.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., and Ellinor, P. T. Transfer learning enables predictions in network biology. *Nature*, 618 (7965):616–624, June 2023. ISSN 0028-0836,1476-4687. doi: 10.1038/s41586-023-06139-9. URL http://dx.doi.org/10.1038/s41586-023-06139-9.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv [cs.LG]*, July 2018. URL http://arxiv.org/abs/1807.03748.

Wang, C., Chan, A. S., Fu, X., Ghazanfar, S., Kim, J., Patrick, E., and Yang, J. Y. Benchmarking the translational potential of spatial gene expression prediction from histology. *Nature Communications*, 16(1):1544, 2025.

Wang, Z., Wang, Z., Srinivasan, B., Ioannidis, V. N., Rangwala, H., and Anubhai, R. Biobridge: Bridging biomedical foundation models via knowledge graphs. *arXiv preprint arXiv:2310.03320*, 2023.

Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

Xie, R., Pang, K., Chung, S., Perciani, C., MacParland, S., Wang, B., and Bader, G. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 70626–70637. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/df656d6ed77b565e8dcdfbf568aead0a-Paper-Conference.pdf.

Zeng, Y., Wei, Z., Yu, W., Yin, R., Yuan, Y., Li, B., Tang, Z., Lu, Y., and Yang, Y. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Briefings in Bioinformatics*, 23(5): bbac297, September 2022. ISSN 1467-5463,1477-4054. doi: 10.1093/bib/bbac297. URL https://dx.doi.org/10.1093/bib/bbac297.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18102–18112, 2022. doi: 10.1109/CVPR52688.2022.01759.

Zhao, S., Zhang, J., Wu, Y., Luo, Y., and Nie, Z. Langcell: Language-cell pre-training for cell identity understanding. *arXiv preprint arXiv:2405.06708*, 2024.

A. Proof & Discussion of the Lemma

Intuition: We declare the reference modality as an *angular ruler* to decompose each embedding into its component along the reference and an orthogonal residual. The positive constraints make both projections large and the residual term small (by Cauchy–Schwarz), yielding a uniform lower bound on the positive score, while the negative constraint caps the negative projection and the same residual control yields a uniform upper bound on each negative score. Substituting these two bounds into the InfoNCE and upper-bounding the log-sum by N times the worst-case gap gives the stated per-sample bound, which tightens as $\epsilon \downarrow$ and $\eta \downarrow$.

Analogously to modalities $\mathcal{I}, \mathcal{G}, \mathcal{T}$ in the main text, here we use A, B, C.

Lemma 3.1 (Per-sample InfoNCE bound for $\mathcal{I} \to \mathcal{T}$)**.** *Given a query $x_i^{\mathcal{I}}$, its matched (but unobserved) $x_i^{\mathcal{T}}$, and N unpaired $x_{j_1}^{\mathcal{T}}, \dots, x_{j_N}^{\mathcal{T}}$ contrastive negatives, define*

$$s^+ \;=\; \langle z_i^{\mathcal{I}},\, z_i^{\mathcal{T}} \rangle, \qquad s_k^- \;=\; \langle z_i^{\mathcal{I}},\, z_{j_k}^{\mathcal{T}} \rangle.$$

Let $\tau > 0$ be the temperature. If (P) and (N) hold with margins (ϵ, η), then the loss ℓ_i across modalities $\mathcal{I}$ and $\mathcal{T}$ is bounded by ϵ and η:

$$\ell_i \;=\; \log\Big(1 + \sum_{k=1}^{N} e^{(s_k^- - s^+)/\tau}\Big) \;\le\; \log\Big(1 + N\, e^{r(\epsilon,\eta)/\tau}\Big),$$

where $r(\epsilon, \eta) = q(\epsilon, \eta) - p(\epsilon)$ with $p(\epsilon) \;=\; 2(1-\epsilon)^2 - 1$ and $q(\epsilon, \eta) \;=\; \max\{\eta, (1-\epsilon)\eta\} + \sqrt{2\epsilon - \epsilon^2}$.

Proof. Proof idea. To bound ℓ_i we (i) lower-bound the positive score s^+ (the numerator) and (ii) upper-bound each negative score s_k^- (the denominator). We proceed with both bounds.

(i) Lower-bounding the numerator (positive A–C score). Assume $\langle \hat{a}_i, \hat{b}_i \rangle \ge 1 - \epsilon$ and $\langle \hat{b}_i, \hat{c}_i \rangle \ge 1 - \epsilon$ from (P). Decompose $\hat{a}_i$ and $\hat{c}_i$ into the part along $\hat{b}_i$ and the residual orthogonal to $\hat{b}_i$:

$$\hat{a}_i = \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i + \big(\hat{a}_i - \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i\big),$$
$$\hat{c}_i = \langle \hat{c}_i, \hat{b}_i \rangle\, \hat{b}_i + \big(\hat{c}_i - \langle \hat{c}_i, \hat{b}_i \rangle\, \hat{b}_i\big).$$

By bilinearity, with explicit multiplication and orthogonality indicated,

$$\begin{aligned}
\langle \hat{a}_i, \hat{c}_i \rangle &= \Big\langle \Big[\langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i + \underbrace{\big(\hat{a}_i - \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i\big)}_{\perp\, \hat{b}_i}\Big], \Big[\langle \hat{c}_i, \hat{b}_i \rangle\, \hat{b}_i + \underbrace{\big(\hat{c}_i - \langle \hat{c}_i, \hat{b}_i \rangle\, \hat{b}_i\big)}_{\perp\, \hat{b}_i}\Big] \Big\rangle \\
&= \langle \hat{a}_i, \hat{b}_i \rangle \langle \hat{c}_i, \hat{b}_i \rangle \underbrace{\langle \hat{b}_i, \hat{b}_i \rangle}_{=1} + \langle \hat{a}_i, \hat{b}_i \rangle \underbrace{\langle \hat{b}_i,\, \hat{c}_i - \langle \hat{c}_i, \hat{b}_i \rangle\, \hat{b}_i \rangle}_{=0} + \langle \hat{c}_i, \hat{b}_i \rangle \underbrace{\langle \hat{a}_i - \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i,\, \hat{b}_i \rangle}_{=0} \\
&\qquad + \big\langle \hat{a}_i - \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i,\; \hat{c}_i - \langle \hat{c}_i, \hat{b}_i \rangle\, \hat{b}_i \big\rangle \\
&= \langle \hat{a}_i, \hat{b}_i \rangle \langle \hat{c}_i, \hat{b}_i \rangle \;+\; \big\langle \hat{a}_i - \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i,\; \hat{c}_i - \langle \hat{c}_i, \hat{b}_i \rangle\, \hat{b}_i \big\rangle.
\end{aligned}$$

Apply Cauchy–Schwarz to the residual inner product:

$$\langle \hat{a}_i, \hat{c}_i \rangle \;\ge\; \langle \hat{a}_i, \hat{b}_i \rangle \langle \hat{c}_i, \hat{b}_i \rangle \;-\; \big\|\hat{a}_i - \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i\big\|\, \big\|\hat{c}_i - \langle \hat{c}_i, \hat{b}_i \rangle\, \hat{b}_i\big\|.$$

Since $\|\hat{a}_i\| = \|\hat{b}_i\| = \|\hat{c}_i\| = 1$,

$$\big\|\hat{a}_i - \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i\big\| = \sqrt{1 - \langle \hat{a}_i, \hat{b}_i \rangle^2}, \qquad \big\|\hat{c}_i - \langle \hat{c}_i, \hat{b}_i \rangle\, \hat{b}_i\big\| = \sqrt{1 - \langle \hat{c}_i, \hat{b}_i \rangle^2}.$$

Let $\alpha := \langle \hat{a}_i, \hat{b}_i \rangle$ and $\beta := \langle \hat{c}_i, \hat{b}_i \rangle$; by (P), $\alpha, \beta \in [1 - \epsilon, 1]$. Thus

$$\langle \hat{a}_i, \hat{c}_i \rangle \;\ge\; \alpha\beta - \sqrt{1 - \alpha^2}\, \sqrt{1 - \beta^2}.$$

The right-hand side is minimized over $\alpha, \beta \in [1-\epsilon, 1]$ at $\alpha = \beta = 1-\epsilon$, giving

$$\langle \hat{a}_i, \hat{c}_i \rangle \;\geq\; (1-\epsilon)^2 - \big(1-(1-\epsilon)^2\big) \;=\; 2(1-\epsilon)^2 - 1 \;=\; p(\epsilon).$$

(ii) Upper-bounding the denominator (negative A–C scores). Let $j \neq i$. Decompose $\hat{a}_i$ and $\hat{c}_j$ along $\hat{b}_i$:

$$\hat{a}_i = \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i + \big(\hat{a}_i - \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i\big), \qquad \hat{c}_j = \langle \hat{c}_j, \hat{b}_i \rangle\, \hat{b}_i + \big(\hat{c}_j - \langle \hat{c}_j, \hat{b}_i \rangle\, \hat{b}_i\big).$$

As above, cross terms vanish:

$$\langle \hat{a}_i, \hat{c}_j \rangle = \langle \hat{a}_i, \hat{b}_i \rangle \langle \hat{c}_j, \hat{b}_i \rangle + \big\langle \hat{a}_i - \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i,\; \hat{c}_j - \langle \hat{c}_j, \hat{b}_i \rangle\, \hat{b}_i \big\rangle.$$

Pivot term: $\langle \hat{a}_i, \hat{b}_i \rangle \in [1-\epsilon, 1]$ and $\langle \hat{c}_j, \hat{b}_i \rangle \leq \eta$, hence

$$\langle \hat{a}_i, \hat{b}_i \rangle \langle \hat{c}_j, \hat{b}_i \rangle \;\leq\; \max\{\eta, (1-\epsilon)\eta\}.$$

Orthogonal term: $\|\hat{a}_i - \langle \hat{a}_i, \hat{b}_i \rangle\, \hat{b}_i\| \leq \sqrt{1-(1-\epsilon)^2} = \sqrt{2\epsilon - \epsilon^2}$ and $\|\hat{c}_j - \langle \hat{c}_j, \hat{b}_i \rangle\, \hat{b}_i\| \leq 1$, hence by Cauchy–Schwarz the residual inner product is at most $\sqrt{2\epsilon - \epsilon^2}$. Thus

$$\langle \hat{a}_i, \hat{c}_j \rangle \;\leq\; \max\{\eta, (1-\epsilon)\eta\} \;+\; \sqrt{2\epsilon - \epsilon^2} \;=\; q(\epsilon, \eta).$$

Combining the numerator and denominator bounds yields the stated inequality for ℓ_i. Monotonicity follows because $p(\epsilon)$ is strictly decreasing in $\epsilon \in [0, 1)$ while $q(\epsilon, \eta)$ is increasing in each of ϵ and η. For the limit: $p(\epsilon) \to 1$ as $\epsilon \to 0$, and $q(\epsilon, \eta) \to -1$ as $(\epsilon, \eta) \to (0, -1)$ since $\max\{\eta, (1-\epsilon)\eta\} \to -1$ and $\sqrt{2\epsilon - \epsilon^2} \to 0$. Therefore $q(\epsilon, \eta) - p(\epsilon) \to -2$ and $\ell_i \leq \log(1 + Ne^{-2/\tau})$, which tends to 0 only if additionally $\tau \downarrow 0$. □

A.1. Discussion of Lemma 3.1

Implicit assumptions. The proof treats (P)–(N) as *uniform* worst-case margins over the evaluation set, whereas InfoNCE optimizes an *average*; to instantiate the bounds in practice, one could choose ϵ and η from held-out/evaluation statistics, but this is not the main point of this bound. We assume each anchor b_i has a single intended positive c_i; if multiple c's legitimately match b_i, exclude those pairs from the negative set or use a multi-positive protocol, otherwise (N) is contradicted. The negative cap η must bound all mismatched pairs that appear at test time; if evaluation introduces harder negatives (larger corpus or domain shift), update η accordingly. Because cosine $\in [-1, 1]$, the per-sample bound tightens with better margins but does not vanish as $\epsilon \to 0$ and $\eta \to -1$ unless $\tau \to 0$ (cf. Lemma 3.1).

Requirements for strong transfer The sufficient condition $p(\epsilon) > q(\epsilon, \eta)$ can fail–or the bound loosen–when (i) supervision is contradictory (true multi-matches treated as negatives), (ii) domain shift makes ϕ_B an unstable "ruler", (iii) evaluation negatives are harder than those used to set η, (iv) residual components orthogonal to $\hat{b}_i$ are *aligned* across A and C (thereby *increasing* q), (v) the negative set is very large (the loss scales like $\log(1 + Ne^{(q-p)/\tau})$), or (vi) evaluation uses a different similarity than cosine.

B. Extended Cell Type Benchmark Analysis

Baseline optimization: To strengthen the published pathology vision-language model baselines in our analysis, we assessed their performance on two sets of queries. First, the raw cell type labels, and second a full sentence constructed as *A sample of* {*label*}. The former version outperformed the latter, and so we performed evaluations for all models (including ours) with that label set. The comparison of label performance is shown in Table 4.

Combining Task-Matching and Transitive Data for Cell Type Prediction Benchmark: We assessed whether the cell type prediction task benefits from combined transitive representation learning and task modality-matching $\mathcal{I} \leftrightarrow \mathcal{T}$ data. Interestingly, we observed a slight deterioration of performance that scaled with increasing amounts of task-specific data (Fig. 8), which is in contrast to our main experiments' results (Section 4). We attribute this to a lack of task-relevant information in the $\mathcal{I} \leftrightarrow \mathcal{T}$ QUILT-1M dataset. While that dataset exhibits some levels of predictive performance ($AUROC > 0.5$), its information seems to be redundant, and even disrupting, in combination with the transitive data sources.

This is not entirely unexpected, as most QUILT-1M annotations are coarse-grained. The results underline the importance of task-relevant data, both in the transitive and the target-matched case.

Table 4. Comparison of Conch and PLIP Terms Performance. **Direct** corresponds to raw labels (cell type names). **Phrase** corresponds to a constructed sentence

Class Label	Conch Terms		PLIP Terms	
	Direct	Phrase	Direct	Phrase
Adipocytes	0.545	0.518	0.467	0.469
B cells	0.584	0.663	0.538	0.541
Dendritic cells	0.525	0.557	0.485	0.478
Granulocytes	0.631	0.431	0.428	0.413
Macrophages/Monocytes	0.426	0.327	0.406	0.372
NK cells	0.437	0.337	0.395	0.345
Nerves	0.567	0.525	0.547	0.575
Plasma cells	0.660	0.649	0.496	0.514
Smooth muscle	0.670	0.662	0.630	0.595
Stroma	0.515	0.494	0.363	0.338
T cells	0.426	0.459	0.502	0.476
Tumor cells	0.497	0.613	0.506	0.505
Vasculature/Lymphatics	0.611	0.595	0.579	0.510
Mean	0.546	0.525	0.488	0.472

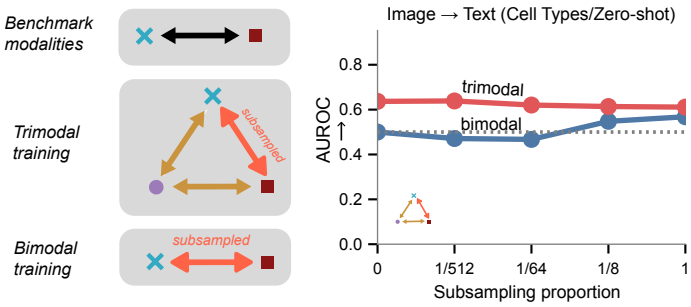

Figure 8. Effect of target-matching paired data in combination with transitive data on zero-shot performance on cell type prediction benchmark.

C. Discussion of metrics and comparison

The patch-level resolution of our model implies a soft-labeled prediction scenario where most patches contain cells of various types (i.e. classes). An intuitive metric to capture this would be the Kullback–Leibler divergence over the categorical distribution of cell types. Indeed, we observe strongly improved performance of our trimodal model compared to published baselines (Fig. 9).

For representing per-class (i.e. cell type) performances, which are biologically most interesting, we found AUROC to be most expressive, as it considers the model's continuous CLIP score for the most-represented class in a given patch.

Other scores, such as the commonly employed F_1, are less suited for the categorical label setting. F_1 rewards models predictions *only* in the case where the model assigns the highest score to the model with the highest abundance, neglecting any other reasonable outputs as equally wrong. In line with this, we observed less coherent and generally weaker results for F_1 scores across all models (see Table 5).

D. Effects of text annotation harmonization

As the textual annotations for images (Ikezogwo et al., 2023) and gene expression profiles (Schaefer et al., 2025b) have been generated independently, they differ in style and focus, hindering alignment and information transfer (see Section3.4).

To assess this effect in our textual data, we reformatted all annotations from (Ikezogwo et al., 2023) and assessed their capability to retrieve their matched images in the dataset through a model that was trained on the other two paired datasets (Schaefer et al., 2025b; Jaume et al., 2024). Here we provide the specifics on the data curation:

Reformatting prompt and processing We used GPT OSS 120B(OpenAI et al., 2025) with a prompt that instructed to rewrite "histopathological image descriptions into biological sample descriptions" with the goal to "transform descriptions that focus on histological features and pathological findings into descriptions that emphasize the biological sample, cell types, and experimental context", drawing from a list of 20 *bona fide* examples sampled from the transcriptome-text dataset (Schaefer et al., 2025b). The full prompt is provided in our code repository. The prompted LLM was provided the original label together with more extensive textual contexts that emerged from the original annotation processing (Ikezogwo et al., 2023).

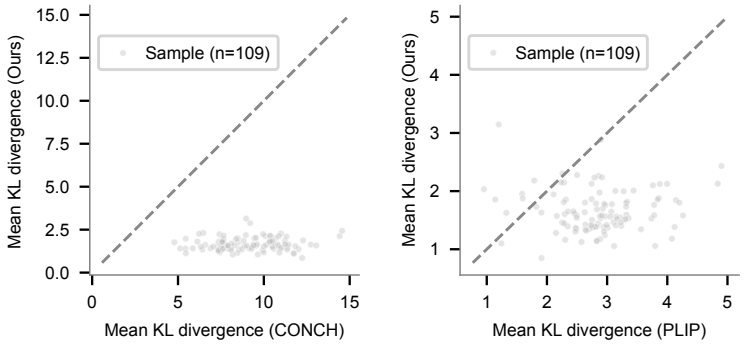

Figure 9. Kullback–Leibler divergence between patch-level predicted and ground truth label distribution. Shown are means across patches for each of 109 sample images in the PathoCellBench benchmark (Koreuber et al., 2025; Schürch et al., 2020).

Table 5. Zero-shot cell type prediction performances (F1 score). Best per row is **bold**; second-best is underlined.

Cell Type	*Trimodal*	CONCH	PLIP
Adipocytes	0.0289	0.0000	**0.0685**
B cells	**0.0690**	0.0000	0.0427
Dendritic cells	**0.0051**	0.0000	0.0000
Granulocytes	0.0363	**0.0898**	0.0000
Macrophages/ Monocytes	0.0653	**0.1319**	0.0000
NK cells	0.0000	0.0000	0.0000
Nerves	0.0000	**0.0053**	0.0016
Plasma cells	0.0599	**0.1086**	0.0000
Smooth muscle	**0.2966**	0.2058	0.0200
Stroma	**0.0535**	0.0000	0.0245
T cells	0.0169	0.0000	**0.0580**
Tumor cells	0.0982	**0.1074**	0.0010
Vasculature/ Lymphatics	**0.1009**	0.0000	0.0174
Mean	**0.0639**	0.0499	0.0180