Exploring Augmentation-Driven Invariances for Graph Self-supervised Learning in Spatial Omics

Lovro Rabuzin* **Michel Tarnow*** **Valentina Boeva**
Department of Computer Science, ETH Zurich
{lrabuzin, mtarnow, vboeva}@ethz.ch

Abstract

Spatial omics technologies provide rich insights into biological processes by jointly capturing molecular profiles and the spatial organization of cells. The resulting high-dimensional data can be naturally represented as graphs, where Graph Neural Networks (GNNs) offer an effective framework to model interactions in the tissue. Self-supervised pretraining methods such as Bootstrapped Graph Latents (BGRL) and GRACE leverage graph augmentations to build invariances without costly labels. Yet, the design of augmentation strategies remains underexplored, particularly in the context of spatial omics. In this work, we systematically investigate how different graph augmentations affect embedding quality and downstream performance in spatial omics. We evaluate a suite of existing and novel augmentations, including transformations tailored to biological variation, across two representative tasks: unsupervised domain identification in healthy tissue and supervised phenotype prediction in cancer tissue. Our results show that carefully chosen augmentations substantially improve performance, whereas poorly aligned or overly complex augmentations may fail to help or even degrade performance. These findings highlight the central role of augmentation design in enforcing meaningful invariances for graph contrastive pretraining in spatial omics.

1 Introduction

Spatial omics technologies measure molecular profiles, such as RNA or protein expression, while preserving the spatial context of cells in their natural environment. This modality provides a more comprehensive view of cellular behavior, biological processes, disease mechanisms, and therapeutic responses compared to non-spatial single-cell methods [1]. Spatial transcriptomics platforms like multiplexed error-robust fluorescence in situ hybridization (MERFISH) [2], spatially-resolved transcript amplicon readout mapping (STARmap) [3], Xenium [4], and barcode in situ targeted sequencing (BaristaSeq) [5] use microscopy or in situ sequencing to generate spatial maps of RNA expression. Complementary proteomics methods such as imaging mass cytometry (IMC) [6] and co-detection by indexing (CODEX) [7] measure protein abundances with spatial resolution.

The complex and high-dimensional data produced by these technologies can be naturally represented as graphs, where nodes correspond to cells and edges encode spatial proximity or molecular similarity [8]. To exploit all available information from spatial omics data, graph-based methods like graph neural networks (GNNs) often exhibit superior characteristics compared to traditional analysis methods not taking spatial dependencies in the data into account [9, 10]. GNNs are well-suited to analyze spatial omics data, as they explicitly model relationships between cells through graph structures using a message-passing mechanism [9].

*Equal contribution.

This work was accepted at the NeurIPS 2025 Workhop on Unifying Representations in Neural Models (archival, PMLR) and the NeurIPS 2025 Workshop on New Perspectives in Advancing Graph Machine Learning (non-archival).

Pretraining enables GNNs to learn generalizable patterns from data before fine-tuning them for specific tasks. Self-supervised or unsupervised pretraining methods are especially valuable in biological contexts, where labeled data can be scarce and expensive [11]. Moreover, these approaches can introduce inductive biases, for instance via graph augmentations, that help models prioritize biologically relevant features and improve robustness [12].

A central principle of contrastive self-supervision is that it enforces invariance to augmentations: two different views of the same input are trained to have similar embeddings. Early works such as SimCLR [13] and its follow-ups demonstrated the effectiveness of this paradigm in computer vision by treating augmented images as positives and enforcing representation consistency. Subsequent methods like BYOL [14] removed the need for explicit negatives while still relying on augmented views to build invariances. The choice of augmentations defines the invariances learned by the model, in line with the *InfoMin principle* that views should remove nuisance factors but preserve task-relevant information [15]. Recent work has shown that in graph domains, augmentations explicitly inject desired invariances, such as robustness to node/edge perturbations or feature corruption [12, 16].

Several pretraining frameworks have operationalized these ideas in the graph setting. Deep Graph Contrastive Representation Learning (GRACE) [16] builds invariance to structural and feature perturbations via an InfoNCE-based contrastive loss. Bootstrapped Graph Latents (BGRL) [17] achieves similar invariances without negative samples, relying instead on online–target encoder consistency. Both methods demonstrate that invariances induced by carefully chosen augmentations significantly enhance representation quality and downstream performance. While recent benchmarks such as scSSL-Bench [18] have evaluated self-supervised learning in a biological context across diverse single-cell omics modalities, spatial omics remains underexplored. Most existing applications of GNNs to spatial omics adopt generic augmentations from other domains or do not leverage augmentation at all [8, 19, 20].

This project explores how different graph augmentation strategies affect the quality of node and graph embeddings in spatial omics data. We investigate both existing graph augmentations, identified through a review of methods for general and spatial omics graphs, as well as newly designed augmentations that encode biologically meaningful inductive biases. Their effectiveness is evaluated on two representative downstream tasks in spatial omics: unsupervised domain identification on healthy tissue and supervised phenotype prediction in cancer samples. These tasks differ not only in supervision regime but also in biological complexity: domain identification on healthy tissue emphasizes stable spatial compartments, while phenotype prediction on cancer tissue must contend with tissue heterogeneity and noisy clinical labels [21–23]. By systematically benchmarking augmentation strategies within contrastive pretraining frameworks, we aim to quantify how different invariances influence downstream performance. To our knowledge, this is the first systematic investigation of graph augmentations in spatial omics, introducing novel biologically motivated transformations that explicitly encode inductive biases such as cellular plasticity and spatial measurement variability.

2 Methods

Figure 1 provides a schematic overview of our study design. Graph augmentations (baseline and advanced) are applied to input graphs, models are pretrained using BGRL and GRACE, and evaluated on two downstream tasks: domain identification and phenotype prediction. The following subsections describe each component in detail.

Notations A graph is denoted by $\mathbf{G} = (\mathbf{X}, \mathbf{A})$, where $\mathbf{X} \in \mathbb{R}^{N \times F}$ is the node feature matrix with N nodes and F features per node, and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the binary adjacency matrix. Graph augmentations generate a new view $\tilde{\mathbf{G}} = (\tilde{\mathbf{X}}, \tilde{\mathbf{A}})$ by modifying $\mathbf{X}$, $\mathbf{A}$, or both. Optionally, node positions can be encoded in a spatial matrix $\mathbf{P} \in \mathbb{R}^{N \times d}$, typically with $d = 2$, yielding $\tilde{\mathbf{P}}$ after augmentation.

2.1 Baseline augmentations

Two baseline augmentations were used: **DropFeatures** and **DropEdges**. Models trained with these augmentations served as baselines for performance comparisons. All augmentation hyperparameters were tuned on validation sets over fixed ranges (see Section A.7 in the Appendix).

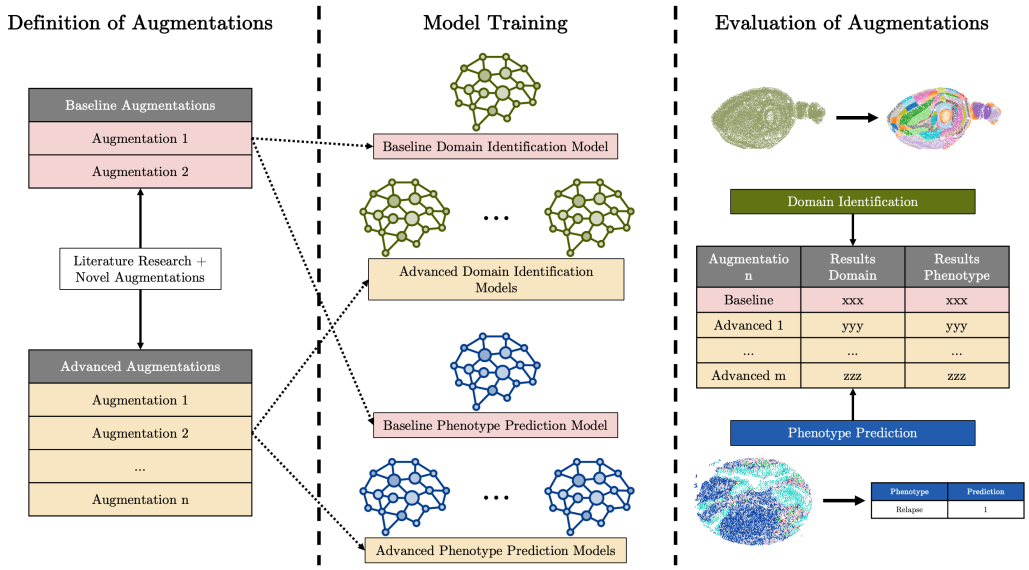

Figure 1: **Overview of the evaluation procedure.** Graph augmentations (baseline and advanced) are applied to input graphs. Models are pretrained using the BGRL and GRACE frameworks, then evaluated on two downstream tasks: domain identification and phenotype prediction.

DropFeatures randomly masks features by setting entries in $\mathbf{X}$ to zero with probability p, resulting in $\tilde{\mathbf{X}}$ while keeping $\mathbf{A}$ unchanged. If $\mathbf{X}$ contains a cell type feature, it is masked by setting entries to the numeric code of the "unassigned" type.

DropEdges randomly removes edges from $\mathbf{A}$ with probability p (Bernoulli sampling), resulting in $\tilde{\mathbf{A}}$ while keeping $\mathbf{X}$ unchanged.

2.2 Advanced augmentations

Advanced augmentations include both published and novel methods. These were tested individually and in combination to assess their effect on downstream tasks relative to baseline augmentations.

DropImportance masks node features and removes edges based on importance scores. Inspired by prior work [24, 25], it is controlled by dropout rate μ and threshold λ_p. Node importance is derived from log-degree centrality:

$$I_i^{(n)} = \frac{\log(1+\deg_i) - \bar{d}}{\max_j \log(1+\deg_j) - \bar{d}}, \tag{1}$$

where $\deg_i$ is the degree of node i and $\bar{d}$ is the mean log-degree. The node feature drop probability is

$$p_i = \min((1 - I_i^{(n)}) \cdot \mu, \lambda_p). \tag{2}$$

Edges are ranked by the mean importance of their endpoints,

$$I_{ij}^{(e)} = \tfrac{1}{2}(I_i^{(n)} + I_j^{(n)}), \tag{3}$$

normalized, and dropped with probability

$$p_{ij} = \min((1 - I_{ij}^{(e)}) \cdot \mu, \lambda_p). \tag{4}$$

This encourages invariance to the removal of less informative node features and edges.

SpatialNoise adds Gaussian noise to spatial positions:

$$\tilde{\mathbf{p}}_i = \mathbf{p}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{5}$$

This models experimental imprecision in cell localization and enforces invariance to small spatial perturbations. This augementation is applicable only to tasks with spatial coordinates (domain identification).

FeatureNoise adds Gaussian noise to node features:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{6}$$

This simulates variability in molecular readouts and enforces robustness to minor fluctuations in expression.

SmoothFeatures applies a convex combination of each node's features with the mean of its neighbors:

$$\tilde{\mathbf{x}}_i = (1-\alpha)\mathbf{x}_i + \alpha \cdot \tfrac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \mathbf{x}_j, \tag{7}$$

where $\alpha \in [0, 1]$ controls the smoothing strength. This simulates transcript leakage [26] and enforces invariance to local feature diffusion. This augementation is used only for domain identification.

PhenotypeShift randomly mutates discrete cell-type features c_i according to a transition map $\mathcal{M}$:

$$\tilde{c}_i = \begin{cases} c_i & \text{with probability } 1-p, \\ \text{sample}(\mathcal{M}[c_i]) & \text{with probability } p, \end{cases} \tag{8}$$

where $\mathcal{M}[c_i] \subseteq \mathcal{C}$ contains plausible phenotype alternatives. This models both plasticity (cell-type switching) and misclassification noise, training robustness to annotation uncertainty. This augementation is used only for phenotype prediction. Details of $\mathcal{M}$ are dataset-specific.

2.3 The task of domain identification

The first task employed to evaluate augmentations is unsupervised *Domain Identification*. The objective is to detect and segment spatially coherent regions within healthy tissue based on molecular data (e.g., gene expression) and spatial data (e.g., spatial relationships). These regions, or domains, ideally reflect biologically relevant structures such as tissue compartments or functional zones.

2.3.1 Data

We used three spatial transcriptomics datasets with expert domain annotations, obtained via the benchmarking study of Schaub *et al.* (2025) [27]. Dataset details are summarized in Table 1.

Table 1: **Datasets used for the domain identification task.**

Dataset	Technology	Samples	Cells
1	MERFISH	5	28,317
2	STARmap	4	4,397
3	BaristaSeq	3	5,257

Dataset 1 profiles 5 mouse brain samples via MERFISH [28]. Dataset 2 contains STARmap data from mouse cortex [3], with expert annotations by Li and Zhou (2022) [29]. Dataset 3 comprises BaristaSeq samples of mouse cortex tissue [30]. All datasets are publicly available [31].

2.3.2 Pipeline and model

An overview of the domain identification pipeline is shown in Figure 2. Each sample is preprocessed into a graph, passed through a GCN encoder pretrained with BGRL or GRACE with spatial regularization, and clustered into domains using the Leiden algorithm [32].

Data preprocessing and graph construction Each sample is first preprocessed using a sequence of filtering and normalization steps. Genes are filtered based on the number of cells they are detected in, and cells are filtered based on the number of genes they express. Cells lacking domain annotations are removed. Raw gene expression values are then normalized to a target sum of 10^5 counts per cell, log-transformed, and scaled to unit variance and zero mean. Principal Component Analysis (PCA) is subsequently applied to the processed expression matrix.

Following preprocessing, one spatial omics graph is constructed per sample. Each cell is represented as a node, with the top 50 principal components (PCs) serving as node features. Nodes are connected to their k nearest neighbors in Euclidean space, with edge weights uniformly set to 1. The number of neighbors k is optimized during hyperparameter search.

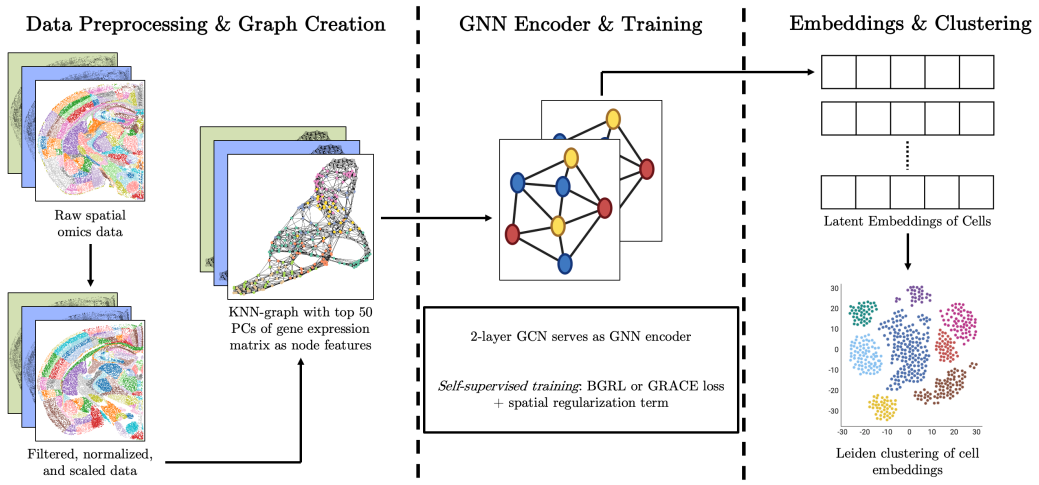

Figure 2: **Domain identification pipeline.** Each sample is preprocessed into a graph, passed through a GCN encoder pretrained with BGRL or GRACE and spatial regularization, and clustered into domains via the Leiden clustering method.

Model and training A two-layer Graph Convolutional Network (GCN) is used to compute node embeddings for each sample-specific graph. The network is trained in a self-supervised manner using both the BGRL and GRACE frameworks.

To encourage spatial coherence in the learned representations, a spatial regularization term is added. It penalizes high similarity in the embedding space for nodes that are spatially distant. This discourages long-range spurious similarities. The resulting overall loss function is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SSL}} + \gamma_{\text{spatial}} \cdot \frac{1}{N^2} \sum_{i,j} \mathbf{D}^{(s)}_{i,j} \cdot (1 - \mathbf{D}^{(z)}_{i,j}), \tag{9}$$

where $\mathbf{D}^{(s)}_{i,j}$ denotes the normalized Euclidean distance between cells i and j, and $\mathbf{D}^{(z)}_{i,j}$ denotes the normalized distance between their embeddings in the latent space. The regularization strength is controlled by the hyperparameter γ_{spatial}.

Self-supervised training is conducted across all data samples. For model selection and evaluation, the dataset is split into 40% validation and 60% test samples. Hyperparameters such as learning rate and spatial regularization strength are optimized using a validation-based hyperparameter search. The model is trained using the Adam optimizer with a cosine annealing learning rate scheduler.

All experiments were run on a cluster with NVIDIA RTX 4090 GPUs (24GB) and 6-core CPUs, with up to 16GB RAM per job. Each training and evaluation run had a runtime of up to 30 minutes.

Clustering To obtain the final domain assignments for each node, the learned node embeddings are clustered using the Leiden algorithm [32]. The resolution parameter of the Leiden clustering is dynamically adjusted to match the number of ground truth domains in each sample. The resulting predicted domain labels are then evaluated against the ground truth annotations using clustering quality metrics (see Section A.5 in the Appendix). Metrics are calculated per sample and averaged across the validation or test set to report the overall performance. All reported means and standard deviations are computed over 5 independent runs with different random seeds.

2.4 The task of phenotype prediction

The second task used to evaluate augmentations is supervised *Phenotype Prediction* in human non-small cell lung cancer (NSCLC) tissue. The objective is to predict biological or clinical phenotypes directly from spatially resolved molecular data. Here, we predict cancer relapse after treatment. A detailed description of the data, model, and evaluation strategy for this task is provided below.

2.4.1 Data

The data used for phenotype prediction consists of one non-small cell lung cancer (NSCLC) spatial proteomics dataset obtained by imaging mass cytometry [33]. Marker expression was quantified with 45 metal-labeled antibodies in 1071 patients with at least 15 years follow-up, resulting in 1868 cancer samples. Each sample includes clinical annotations, for instance smoking status, cancer stage, relapse, clinical outcome, or cancer subtype. The raw data can be downloaded from the resource provided by Cords *et al.* (2024) [33].

2.4.2 Pipeline and model

The phenotype prediction pipeline is based on SPACE-GM [8]. An overview of the pipeline is shown in Figure 3.

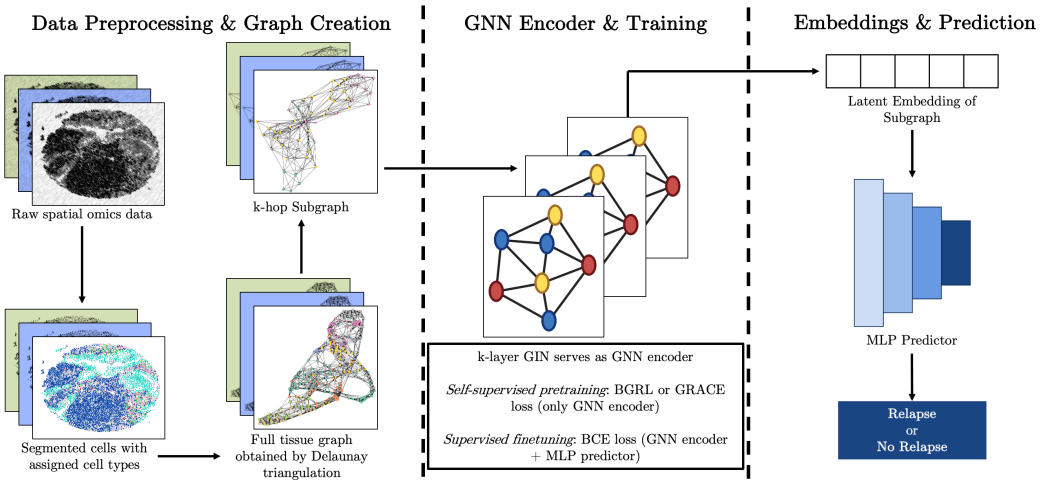

Figure 3: **Phenotype prediction pipeline.** Tissue graphs are built from omics data, subgraphs extracted, passed through a GIN encoder pretrained with BGRL or GRACE, and classified with an MLP.

Data preprocessing and graph creation Graphs were constructed from segmented cells using Delaunay triangulation. Node features included cell type (integer-encoded) and cell size. Edge features consist of a binary "near/distant" category based on centroid-to-centroid distance, using a threshold of 20 μm, reflecting the typical size of human cells. From each tissue graph, h-hop subgraphs were extracted ($h = 3$ by default).

Model and training An L-layer Graph Isomorphism Network (GIN) with edge-feature extension [8] was used, where messages are defined as

$$m_{vu}^{(\ell)} = h_u^{(\ell-1)} + e_{vu}^{(\ell)}, \tag{10}$$

with $e_{vu}^{(\ell)}$ mapped via an embedding lookup. Subgraph embeddings were obtained by max-pooling over final-layer node embeddings. The encoder was pretrained with BGRL and GRACE. For classification, a 3-layer MLP was added and jointly fine-tuned with a weighted BCE loss.

All experiments were run on a cluster with NVIDIA RTX 4090 GPUs (24GB) and 16-core CPUs, with up to 96GB RAM per job. Each training and evaluation run had a runtime of up to 4 hours.

Splits and optimization Pretraining used all samples without labels. For supervised fine-tuning, 1492 samples were used for training and 376 for evaluation, with evaluation split 50% validation / 50% test. Splits were stratified by relapse, and all samples from a patient were assigned to the same fold. The performance was evaluated against the ground truth patient labels using standard classification quality metrics (see Section A.4 in the Appendix). Hyperparameters were tuned on the validation set. All reported means and standard deviations are computed on the test set over 5 independent runs with different random seeds.

3 Results

3.1 Unsupervised domain identification in healthy mouse brain tissue

We first evaluated the effect of augmentations on the task of identifying structurally or functionally distinct domains in healthy mouse brain tissue. Baseline models were trained with *DropFeatures* and *DropEdges*, and compared against models with advanced augmentations or their combinations. The *Noise* augmentation denotes the joint application of *SpatialNoise* and *FeatureNoise*. Performance was measured using normalized mutual information (NMI), homogeneity (HOM), and completeness (COM).

Results with BGRL and GRACE are shown in Tables 2 and 3. Under BGRL, *DropImportance* improved NMI from 0.61 (baseline) to 0.66, with the next-best performance achieved by combining all augmentations (0.65). Under GRACE, *DropImportance* again achieved the best result (0.66 compared to 0.65 baseline), and was the only augmentation regime that substantially improved over the baseline. Across both frameworks, *DropImportance* provided the most consistent gains. With BGRL, nearly all augmentation regimes improved upon the baseline, whereas in GRACE the gains were smaller because the baseline was already comparatively strong.

Table 2: **Performance on domain identification task using BGRL.** Clustering performance on healthy mouse brain tissue using different augmentation strategies. Reported as mean ± standard deviation across 5 random seeds. The best and second-best results by mean are highlighted.

Augmentations	NMI	HOM	COM
Baseline	0.6145 ± 0.0195	0.6188 ± 0.0234	0.6121 ± 0.0175
Baseline + Noise	0.6488 ± 0.0083	0.6419 ± 0.0093	0.6576 ± 0.0074
DropImportance	**0.6585 ± 0.0033**	**0.6552 ± 0.0065**	**0.6635 ± 0.0008**
DropImportance + Noise	0.6488 ± 0.0166	0.6507 ± 0.0135	0.6498 ± 0.0217
SmoothFeatures	0.6497 ± 0.0065	0.6465 ± 0.0097	0.6538 ± 0.0061
DropImp. + Noise + SmoothFeat.	0.6540 ± 0.0104	0.6507 ± 0.0110	0.6579 ± 0.0103

Table 3: **Performance on domain identification task using GRACE.** Clustering performance on healthy mouse brain tissue using different augmentation strategies. Reported as mean ± standard deviation across 5 random seeds. The best and second-best results by mean are highlighted.

Augmentations	NMI	HOM	COM
Baseline	0.6470 ± 0.0081	0.6475 ± 0.0081	0.6484 ± 0.0110
Baseline + Noise	0.6405 ± 0.0221	0.6390 ± 0.0157	0.6438 ± 0.0271
DropImportance	**0.6639 ± 0.0056**	**0.6569 ± 0.0082**	**0.6726 ± 0.0046**
DropImportance + Noise	0.6477 ± 0.0125	0.6409 ± 0.0120	0.6557 ± 0.0127
SmoothFeatures	0.6460 ± 0.0100	0.6423 ± 0.0115	0.6509 ± 0.0085
DropImp. + Noise + SmoothFeat.	0.6412 ± 0.0058	0.6336 ± 0.0050	0.6502 ± 0.0080

Qualitative results of the models trained using BGRL are shown in Figure 4 for a representative MERFISH sample. Different augmentation strategies produce visibly different domain segmentations, broadly consistent with the quantitative metrics.

Overall, domain identification highlights how augmentations can improve unsupervised discovery of spatial structure in healthy tissue. To complement this, we next evaluate phenotype prediction, a supervised task in noisy and heterogeneous human cancer tissue. Together, these tasks represent distinct regimes, unsupervised and supervised, as well as healthy and cancerous tissue, which help reveal when particular augmentations are most beneficial.

3.2 Supervised phenotype prediction in human cancer tissue

We evaluated augmentation strategies on relapse prediction in NSCLC samples, with performance measured by F1 score and AUROC (Tables 4 and 5). Models were trained with baseline augmentations (*DropFeatures* and *DropEdges*) and compared against advanced augmentations. In *PhenotypeShift*,

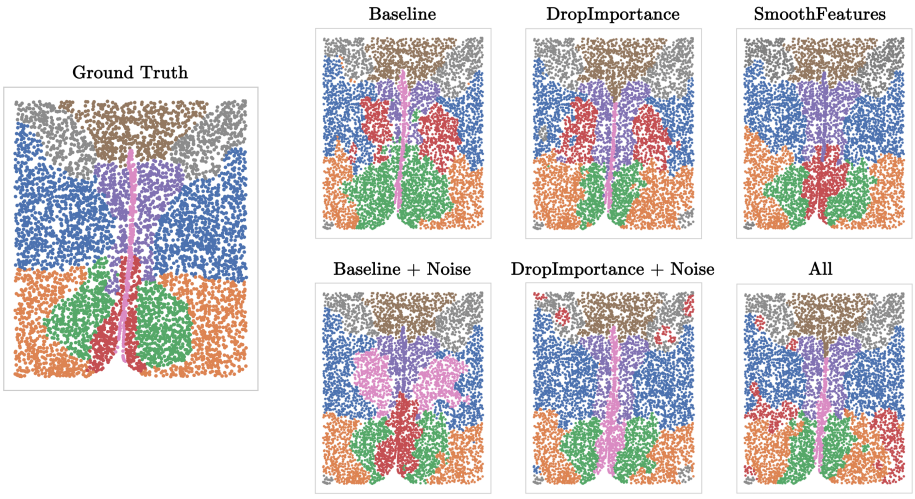

Figure 4: **Predicted and ground-truth domains in MERFISH tissue.** Visualization of a representative mouse brain sample. The left-most panel shows expert-annotated ground truth; remaining panels display predicted domains under different augmentation strategies. Augmentations strongly influence segmentation quality, broadly consistent with the quantitative results.

we incorporated biologically motivated cell state transitions, including tumor adaptation to hypoxia, fibroblast subtype plasticity, and T cell differentiation into regulatory, exhausted, or proliferative states, while also accounting for myeloid–neutrophil annotation noise (see Section A.6 in the Appendix).

Under BGRL, the baseline achieved F1 = 0.59 and AUROC = 0.60. *FeatureNoise* improved AUROC to 0.61, while *DropImportance* alone decreased performance. The best F1 was obtained with *PhenotypeShift* (0.64), while the best AUROC was achieved by *FeatureNoise* (0.61). Combining all augmentations yielded intermediate gains (F1 = 0.62, AUROC = 0.60).

Under GRACE, the results showed similar patterns. The best F1 score was obtained with *DropImportance* + *FeatureNoise* (0.63), while the best AUROC was achieved with *Baseline* + *FeatureNoise* (0.59). The second best scores were achieved using *PhenotypeShift* both in terms of F1 score (0.63) and AUROC (0.59). Adding all augmentations together did not improve over single strategies. Overall, these results indicate that for phenotype prediction in cancer, both noise-based and biologically motivated augmentations may improve performance, but combining them provides only limited additional benefit, if at all.

Table 4: **Performance on phenotype prediction task using BGRL.** Relapse prediction in NSCLC samples using BGRL pretraining with different augmentation strategies. Reported as mean ± standard deviation across 5 random seeds. The best and second-best results by mean are highlighted.

Augmentations	F1 Score	AUROC
Baseline	0.5896 ± 0.0213	0.5986 ± 0.0142
Baseline + FeatureNoise	0.6265 ± 0.0155	**0.6084 ± 0.0097**
DropImportance	0.6171 ± 0.0245	0.5848 ± 0.0031
DropImportance + FeatureNoise	0.6277 ± 0.0011	0.5665 ± 0.0106
PhenotypeShift	**0.6375 ± 0.0090**	0.6006 ± 0.0098
DropImp. + FeatNoise + PhenotypeShift	0.6218 ± 0.0291	0.6030 ± 0.0100

4 Discussion

We systematically evaluated the role of graph augmentations in self-supervised GNN pretraining for spatial omics, using both BGRL [17] and GRACE [16] across two tasks: domain identification and phenotype prediction. The results show that augmentation choice has a decisive impact on

Table 5: **Performance on phenotype prediction task using GRACE.** Relapse prediction in NSCLC samples using GRACE pretraining with different augmentation strategies. Reported as mean ± standard deviation across 5 random seeds. The best and second-best results by mean are highlighted.

Augmentations	F1 Score	AUROC
Baseline	0.6157 ± 0.0125	0.5759 ± 0.0194
Baseline + FeatureNoise	0.6208 ± 0.0142	**0.5932 ± 0.0090**
DropImportance	0.6137 ± 0.0355	0.5707 ± 0.0074
DropImportance + FeatureNoise	**0.6338 ± 0.0052**	0.5545 ± 0.0122
PhenotypeShift	0.6318 ± 0.0096	0.5897 ± 0.0167
DropImp. + FeatNoise + PhenotypeShift	0.6070 ± 0.0409	0.5870 ± 0.0088

downstream performance. In line with prior contrastive learning work [13–15], we find that well-aligned augmentations can enhance representations by encoding task-relevant invariances, whereas overly strong or misaligned transformations can degrade performance.

Domain identification benefited most from structural perturbations, with *DropImportance* improving performance by removing structurally redundant nodes and edges. In contrast, phenotype prediction showed limited gains from structural perturbations and instead improved with noise-based and biologically motivated augmentations such as *FeatureNoise* and *PhenotypeShift*. Composing these augmentations, however, provided only limited additional benefit or even hurt performance. A likely explanation is that the combined perturbations either dilute informative signal or exceed the capacity of the model to leverage additional invariances in this noisy, small-sample setting. These divergent outcomes highlight the distinct demands of the two tasks: in healthy tissue, domain identification profits from invariance to redundant structure, as tissue compartments are relatively stable, whereas in heterogeneous cancer tissue, phenotype prediction must contend with biological variability and label uncertainty [21–23].

Our findings parallel those in image and graph contrastive learning. SimCLR [13] demonstrated that augmentation design largely determines representation quality, while BYOL [14] and BGRL [17] highlighted that not all perturbations are beneficial. GRACE [16] also emphasized the role of augmentation strategies for graphs. Our results extend these insights to spatial omics: invariances induced by augmentations must be carefully matched to biological and experimental noise characteristics to improve downstream generalization. This has direct implications for the use of spatial omics in translational research, where robust embeddings can support tissue diagnostics and patient stratification.

This study was limited to two downstream tasks and a curated set of augmentations. Domain identification was based on a small number of annotated healthy tissue samples, constraining statistical power, while phenotype prediction was restricted to a single cancer cohort. Moreover, the two tasks differed both in supervision regime and biological complexity, making it difficult to disentangle whether augmentation effectiveness depends primarily on task type (unsupervised vs. supervised) or tissue context (healthy vs. cancerous). Future work could address this by including supervised tasks on healthy tissue and unsupervised tasks on cancer tissue. Beyond these design choices, broader evaluation on additional omics technologies and clinical endpoints will be important. More extensive augmentation design could directly support translational use cases such as patient stratification or treatment response prediction, where robustness to biological noise is critical.

In summary, augmentation design is a critical factor in self-supervised learning on spatial omics graphs. Effective augmentations encode biologically plausible invariances, improving model robustness and downstream accuracy, while misaligned ones can add cost without benefit. Our results reinforce the view that augmentation choice is not incidental but a central design decision in graph contrastive learning.

References

[1] Dario Bressan, Giorgia Battistoni, and Gregory J. Hannon. The dawn of spatial omics. *Science*, 381(6657):4964, 2023. doi: 10.1126/science.abq4964. URL `https://www.science.org/doi/abs/10.1126/science.abq4964`.

[2] Kok Hao Chen, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233): 6090, 2015. doi: 10.1126/science.aaa6090. URL https://www.science.org/doi/abs/10.1126/science.aaa6090.

[3] Xiao Wang, William E. Allen, Matthew A. Wright, Emily L. Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P. Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science*, 361(6400):5691, 2018. doi: 10.1126/science.aat5691. URL https://www.science.org/doi/abs/10.1126/science.aat5691.

[4] Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Stephen R. Williams, Morgane Rouault, Ghezal Beliakoff, Carolyn A. Morrison, Michelli F. Oliveira, Jordan T. Sicherman, Andrew Kohlway, Jawad Abousoud, Tingsheng Yu Drennon, Seayar H. Mohabbat, Sarah E. B. Taylor, and 10x Development Teams. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14(1):8353, 2023. doi: 10.1038/s41467-023-43458-x. URL https://www.nature.com/articles/s41467-023-43458-x.

[5] Xiaoyin Chen, Yu-Chi Sun, George M. Church, Je Hyuk Lee, and Anthony M. Zador. Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Research*, 46(4): e22, 2018. doi: 10.1093/nar/gkx1206. URL https://academic.oup.com/nar/article/46/4/e22/4668654.

[6] Charlotte Giesen, Hao A. O. Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J. Schüffler, Daniel Grolimund, Joachim M. Buhmann, Simone Brandt, Zsuzsanna Varga, Peter J. Wild, Detlef Günther, and Bernd Bodenmiller. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11 (4):417–422, 2014. doi: 10.1038/nmeth.2869. URL https://www.nature.com/articles/nmeth.2869.

[7] Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P. Nolan. Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell*, 174(4):968–981, 2018. doi: 10.1016/j.cell.2018.07.010. URL https://www.cell.com/cell/fulltext/S0092-8674(18)30904-8.

[8] Zhenqin Wu, Alexandro E. Trevino, Eric Wu, Kyle Swanson, Honesty J. Kim, H. Blaize D'Angio, Ryan Preska, Gregory W. Charville, Piero D. Dalerba, Ann Marie Egloff, Ravindra Uppaluri, Umamaheswar Duvvuri, Aaron T. Mayer, and James Zou. Graph deep learning for the characterization of tumour microenvironments from spatial protein profiles in tissue specimens. *Nature Biomedical Engineering*, 6(12):1435–1448, November 2022. ISSN 2157-846X. doi: 10.1038/s41551-022-00951-w. URL http://dx.doi.org/10.1038/s41551-022-00951-w.

[9] Bharti Khemani, Shruti Patil, Ketan Kotecha, and Sudeep Tanwar. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data*, 11(1):18, 2024. doi: 10.1186/s40537-023-00876-4. URL https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00876-4.

[10] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. *arXiv e-prints*, art. arXiv:1704.01212, April 2017. doi: 10.48550/arXiv.1704.01212.

[11] Raphael Schäfer, Till Nicke, Henning Höfener, Annkristin Lange, Dorit Merhof, Friedrich Feuerhake, Volkmar Schulz, Johannes Lotz, and Fabian Kiessling. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nature Computational Science*, 4(7):495–509, July 2024. ISSN 2662-8457. doi: 10.1038/s43588-024-00662-z. URL http://dx.doi.org/10.1038/s43588-024-00662-z.

[12] Yue You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *arXiv preprint arXiv:2010.13902*, 2020. URL https://arxiv.org/abs/2010.13902.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv e-prints*, art. arXiv:2002.05709, February 2020. doi: 10.48550/arXiv.2002.05709.

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised Learning. *arXiv e-prints*, art. arXiv:2006.07733, June 2020. doi: 10.48550/arXiv.2006.07733.

[15] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What Makes for Good Views for Contrastive Learning? *arXiv e-prints*, art. arXiv:2005.10243, May 2020. doi: 10.48550/arXiv.2005.10243.

[16] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep Graph Contrastive Representation Learning. *arXiv e-prints*, art. arXiv:2006.04131, June 2020. doi: 10.48550/arXiv.2006.04131.

[17] Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Rémi Munos, Petar Veličković, and Michal Valko. Large-scale representation learning on graphs via bootstrapping. *arXiv preprint arXiv:2102.06514*, 2021. URL `https://arxiv.org/abs/2102.06514`.

[18] Olga Ovcharenko, Florian Barkmann, Philip Toma, Imant Daunhawer, Julia Vogt, Sebastian Schelter, and Valentina Boeva. scSSL-Bench: Benchmarking Self-Supervised Learning for Single-Cell Data. *arXiv e-prints*, art. arXiv:2506.10031, June 2025. doi: 10.48550/arXiv.2506.10031.

[19] Yuxuan Hu, Jiazhen Rong, Yafei Xu, Runzhi Xie, Jacqueline Peng, Lin Gao, and Kai Tan. Unsupervised and supervised discovery of tissue cellular neighborhoods from cell phenotypes. *Nature Methods*, 21(2):267–278, January 2024. ISSN 1548-7105. doi: 10.1038/s41592-023-02124-2. URL `http://dx.doi.org/10.1038/s41592-023-02124-2`.

[20] Shay Shimonov, Joseph M Cunningham, Ronen Talmon, Lilach Aizenbud, Shruti J Desai, David Rimm, Kurt Schalper, Harriet Kluger, and Yuval Kluger. Sorbet: Automated cell-neighborhood analysis of spatial transcriptomics or proteomics for interpretable sample classification via gnn. January 2024. doi: 10.1101/2023.12.30.573739. URL `http://dx.doi.org/10.1101/2023.12.30.573739`.

[21] Cyril Neftel, Julie Laffy, Mariella G. Filbin, Toshiro Hara, Marni E. Shore, Gilbert J. Rahme, Alyssa R. Richman, Dana Silverbush, McKenzie L. Shaw, Christine M. Hebert, John Dewitt, Simon Gritsch, Elizabeth M. Perez, L. Nicolas Gonzalez Castro, Xiaoyang Lan, Nicholas Druck, Christopher Rodman, Danielle Dionne, Alexander Kaplan, Mia S. Bertalan, Julia Small, Kristine Pelton, Sarah Becker, Dennis Bonal, Quang-De Nguyen, Rachel L. Servis, Jeremy M. Fung, Ravindra Mylvaganam, Lisa Mayr, Johannes Gojo, Christine Haberler, Rene Geyeregger, Thomas Czech, Irene Slavc, Brian V. Nahed, William T. Curry, Bob S. Carter, Hiroaki Wakimoto, Priscilla K. Brastianos, Tracy T. Batchelor, Anat Stemmer-Rachamimov, Maria Martinez-Lage, Matthew P. Frosch, Ivan Stamenkovic, Nicolo Riggi, Esther Rheinbay, Michelle Monje, Orit Rozenblatt-Rosen, Daniel P. Cahill, Anoop P. Patel, Tony Hunter, Inder M. Verma, Keith L. Ligon, David N. Louis, Aviv Regev, Bradley E. Bernstein, Itay Tirosh, and Mario L. Suvà. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell*, 178 (4):835–849.e21, August 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.06.024. URL `http://dx.doi.org/10.1016/j.cell.2019.06.024`.

[22] Karin Pelka, Matan Hofree, Jonathan H. Chen, Siranush Sarkizova, Joshua D. Pirl, Vjola Jorgji, Alborz Bejnood, Danielle Dionne, William H. Ge, Katherine H. Xu, Sherry X. Chao, Daniel R. Zollinger, David J. Lieb, Jason W. Reeves, Christopher A. Fuhrman, Margaret L. Hoang, Toni Delorey, Lan T. Nguyen, Julia Waldman, Max Klapholz, Isaac Wakiro, Ofir Cohen, Julian Albers, Christopher S. Smillie, Michael S. Cuoco, Jingyi Wu, Mei-ju Su, Jason Yeung, Brinda Vijaykumar, Angela M. Magnuson, Natasha Asinovski, Tabea Moll, Max N. Goder-Reiser, Anise S. Applebaum, Lauren K. Brais, Laura K. DelloStritto, Sarah L. Denning, Susannah T. Phillips, Emma K. Hill, Julia K. Meehan, Dennie T. Frederick, Tatyana Sharova, Abhay Kanodia, Ellen Z. Todres, Judit Jané-Valbuena, Moshe Biton, Benjamin Izar, Conner D. Lambden,

Thomas E. Clancy, Ronald Bleday, Nelya Melnitchouk, Jennifer Irani, Hiroko Kunitake, David L. Berger, Amitabh Srivastava, Jason L. Hornick, Shuji Ogino, Asaf Rotem, Sébastien Vigneau, Bruce E. Johnson, Ryan B. Corcoran, Arlene H. Sharpe, Vijay K. Kuchroo, Kimmie Ng, Marios Giannakis, Linda T. Nieman, Genevieve M. Boland, Andrew J. Aguirre, Ana C. Anderson, Orit Rozenblatt-Rosen, Aviv Regev, and Nir Hacohen. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell*, 184(18):4734–4752.e20, September 2021. ISSN 0092-8674. doi: 10.1016/j.cell.2021.08.003. URL http://dx.doi.org/10.1016/j.cell.2021.08.003.

[23] Andrew L. Ji, Adam J. Rubin, Kim Thrane, Sizun Jiang, David L. Reynolds, Robin M. Meyers, Margaret G. Guo, Benson M. George, Annelie Mollbrink, Joseph Bergenstråhle, Ludvig Larsson, Yunhao Bai, Bokai Zhu, Aparna Bhaduri, Jordan M. Meyers, Xavier Rovira-Clavé, S. Tyler Hollmig, Sumaira Z. Aasi, Garry P. Nolan, Joakim Lundeberg, and Paul A. Khavari. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2):497–514.e22, July 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.05.039. URL http://dx.doi.org/10.1016/j.cell.2020.05.039.

[24] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Graph Contrastive Learning with Adaptive Augmentation. *CoRR*, 2010.14945, 2020. URL https://arxiv.org/abs/2010.14945.

[25] Yichun Li, Jin Huang, Weihao Yu, and Tinghua Zhang. Neighborhood-enhanced contrast for pre-training graph neural networks. *Neural Computing and Applications*, 36(8):4195–4205, 2024. doi: 10.1007/s00521-023-09274-6. URL https://link.springer.com/article/10.1007/s00521-023-09274-6.

[26] Yue You, Yuting Fu, Lanxiang Li, Zhongmin Zhang, Shikai Jia, Shihong Lu, Wenle Ren, Yifang Liu, Yang Xu, Xiaojing Liu, Fuqing Jiang, Guangdun Peng, Abhishek Sampath Kumar, Matthew E. Ritchie, Xiaodong Liu, and Luyi Tian. Systematic comparison of sequencing-based spatial transcriptomic methods. *Nature Methods*, 21(9):1743–1754, July 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02325-3. URL http://dx.doi.org/10.1038/s41592-024-02325-3.

[27] Darius P. Schaub, Behnam Yousefi, Nico Kaiser, Robin Khatri, Victor G. Puelles, Christian F. Krebs, Ulf Panzer, and Stefan Bonn. PCA-based spatial domain identification with state-of-the-art performance. *Bioinformatics*, 41(1):5, 2025. doi: 10.1093/bioinformatics/btaf005. URL https://academic.oup.com/bioinformatics/article/41/1/btaf005/7945104.

[28] Jeffrey R. Moffitt, Devjanee Bambah-Mukku, Stephen W. Eichhorn, Eric Vaughn, Karthik Shekhar, Jonathan D. Perez, Nimrod D. Rubinstein, Junjie Hao, Aviv Regev, Catherine Dulac, and Xiaowei Zhuang. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416):eaat5324, 2018. doi: 10.1126/science.aau5324. URL https://www.science.org/doi/abs/10.1126/science.aau5324.

[29] Zheng Li and Xiang Zhou. BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome Biology*, 23(1):168, 2022. doi: 10.1186/s13059-022-02734-7. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02734-7.

[30] Brian Long, Jeremy Miller, and The SpaceTx Consortium. SpaceTx: A Roadmap for Benchmarking Spatial Transcriptomics Exploration of the Brain, 2023. URL https://arxiv.org/abs/2301.08436.

[31] Zhiyuan Yuan, Fangyuan Zhao, Senlin Lin, Yu Zhao, Jianhua Yao, Yan Cui, Xiao-Yong Zhang, and Yi Zhao. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods*, 21(4):712–722, 2024. doi: 10.1038/s41592-024-02215-8. URL https://www.nature.com/articles/s41592-024-02215-8.

[32] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019. doi: 10.1038/s41598-019-41695-z. URL https://www.nature.com/articles/s41598-019-41695-z.

[33] Lena Cords, Stefanie Engler, Martina Haberecker, Jan Hendrik Rüschoff, Holger Moch, Natalie de Souza, and Bernd Bodenmiller. Cancer-associated fibroblast phenotypes are associated with patient outcome in non-small cell lung cancer. *Cancer Cell*, 42(3):396–412, 2024. doi: 10.1016/j.ccell.2023.12.021. URL https://www.cell.com/cancer-cell/fulltext/S1535-6108(23)00449-X.

[34] Zhou Chen, Fangfang Han, Yan Du, Huaqing Shi, and Wence Zhou. Hypoxic microenvironment in cancer: molecular mechanisms and therapeutic interventions. *Signal Transduction and Targeted Therapy*, 8(1), February 2023. ISSN 2059-3635. doi: 10.1038/s41392-023-01332-8. URL http://dx.doi.org/10.1038/s41392-023-01332-8.

[35] Daniel Öhlund, Ela Elyada, and David Tuveson. Fibroblast heterogeneity in the cancer wound. *Journal of Experimental Medicine*, 211(8):1503–1523, July 2014. ISSN 0022-1007. doi: 10.1084/jem.20140692. URL http://dx.doi.org/10.1084/jem.20140692.

[36] Raghu Kalluri. The biology and function of fibroblasts in cancer. *Nature Reviews Cancer*, 16(9):582–598, August 2016. ISSN 1474-1768. doi: 10.1038/nrc.2016.73. URL http://dx.doi.org/10.1038/nrc.2016.73.

[37] Giulia Biffi and David A. Tuveson. Diversity and biology of cancer-associated fibroblasts. *Physiological Reviews*, 101(1):147–176, January 2021. ISSN 1522-1210. doi: 10.1152/physrev.00048.2019. URL http://dx.doi.org/10.1152/physrev.00048.2019.

[38] Jinfang Zhu and William E. Paul. Cd4 t cells: fates, functions, and faults. *Blood*, 112(5):1557–1569, September 2008. ISSN 1528-0020. doi: 10.1182/blood-2008-05-078154. URL http://dx.doi.org/10.1182/blood-2008-05-078154.

[39] E. John Wherry and Makoto Kurachi. Molecular and cellular insights into t cell exhaustion. *Nature Reviews Immunology*, 15(8):486–499, July 2015. ISSN 1474-1741. doi: 10.1038/nri3862. URL http://dx.doi.org/10.1038/nri3862.

[40] Susan M. Kaech and Weiguo Cui. Transcriptional control of effector and memory cd8+ t cell differentiation. *Nature Reviews Immunology*, 12(11):749–761, October 2012. ISSN 1474-1741. doi: 10.1038/nri3307. URL http://dx.doi.org/10.1038/nri3307.

[41] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1), January 2017. ISSN 2041-1723. doi: 10.1038/ncomms14049. URL http://dx.doi.org/10.1038/ncomms14049.

A Appendix / supplemental material

A.1 Bootstrapped Graph Latents (BGRL)

Bootstrapped Graph Latents (BGRL) [17] is a self-supervised graph representation learning method used in this project. It avoids labels and negative samples by predicting alternate augmentations of the same input graph.

A graph $\mathbf{G} = (\mathbf{X}, \mathbf{A})$ is first augmented into two alternate views $\mathbf{G}_1 = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{A}}_1)$ and $\mathbf{G}_2 = (\tilde{\mathbf{X}}_2, \tilde{\mathbf{A}}_2)$ via graph augmentation functions $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively. An online encoder $\mathcal{E}_\theta$ with parameters θ then produces an online representation from the first augmented view, $\tilde{\mathbf{H}}_1 := \mathcal{E}_\theta(\tilde{\mathbf{X}}_1, \tilde{\mathbf{A}}_1)$, and a target encoder $\mathcal{E}_\phi$ with parameters ϕ produces a target representation from the second augmented view, $\tilde{\mathbf{H}}_2 := \mathcal{E}_\phi(\tilde{\mathbf{X}}_2, \tilde{\mathbf{A}}_2)$. A prediction of the target representation, $\tilde{\mathbf{Z}}_1 := p_\theta(\tilde{\mathbf{H}}_1)$, is obtained by feeding the online representation into a node-level predictor p_θ.

To update the online encoder's parameters θ, the gradient of the cosine similarity of the predicted target representation $\tilde{\mathbf{Z}}_1$ and the true target representation $\tilde{\mathbf{H}}_2$ is computed with respect to θ:

$$l(\theta, \phi) = -\frac{2}{N} \sum_{i=0}^{N-1} \frac{\tilde{\mathbf{Z}}_{(1,i)} \tilde{\mathbf{H}}_{(2,i)}^{\mathsf{T}}}{\left\|\tilde{\mathbf{Z}}_{(1,i)}\right\| \left\|\tilde{\mathbf{H}}_{(2,i)}\right\|} \tag{11}$$

$$\theta \leftarrow \text{optimize}(\theta, \eta, \partial_\theta l(\theta, \phi)). \tag{12}$$

Here, η is the learning rate and in practice, the loss is symmetrized by also predicting the target representation of the first view with the online representation of the second view.

The target encoder's parameters ϕ are updated as an exponentially moving average with decay rate τ of the online encoder's parameters θ:

$$\phi \leftarrow \tau\phi + (1 - \tau)\theta. \tag{13}$$

A.2 Deep Graph Contrastive Representation Learning (GRACE)

Deep Graph Contrastive Representation Learning (GRACE) [16] is a self-supervised method for unsupervised graph representation learning. Unlike methods relying on global readouts, GRACE directly contrasts node-level embeddings across two randomly corrupted views of the same graph.

Formally, given a graph $\mathbf{G} = (\mathbf{X}, \mathbf{A})$, GRACE generates two augmented views $\mathbf{G}_1 = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{A}}_1)$ and $\mathbf{G}_2 = (\tilde{\mathbf{X}}_2, \tilde{\mathbf{A}}_2)$ by applying stochastic corruption functions $\mathcal{T}_1, \mathcal{T}_2$ to features and edges. Specifically, GRACE uses (i) *edge removal* with probability p_r and (ii) *feature masking* with probability p_m to generate diverse contexts.

A shared GNN encoder f_θ then computes node embeddings $\mathbf{U} = f_\theta(\tilde{\mathbf{X}}_1, \tilde{\mathbf{A}}_1)$ and $\mathbf{V} = f_\theta(\tilde{\mathbf{X}}_2, \tilde{\mathbf{A}}_2)$. For a node i, the embeddings $(\mathbf{u}_i, \mathbf{v}_i)$ from the two views form a positive pair, while embeddings from other nodes act as negatives. The similarity between two embeddings is estimated by a critic

$$\theta(\mathbf{u}, \mathbf{v}) = \frac{g(\mathbf{u})^\top g(\mathbf{v})}{\|g(\mathbf{u})\| \, \|g(\mathbf{v})\|}, \tag{14}$$

where $g(\cdot)$ is a two-layer projection head and the similarity is scaled by a temperature τ.

The contrastive loss for node i is defined as

$$\ell(\mathbf{u}_i, \mathbf{v}_i) = -\log \frac{\exp(\theta(\mathbf{u}_i, \mathbf{v}_i)/\tau)}{\exp(\theta(\mathbf{u}_i, \mathbf{v}_i)/\tau) + \sum_{k \neq i} \exp(\theta(\mathbf{u}_i, \mathbf{v}_k)/\tau) + \sum_{k \neq i} \exp(\theta(\mathbf{u}_i, \mathbf{u}_k)/\tau)}. \tag{15}$$

The final symmetric objective averages over all nodes:

$$J = \frac{1}{2N} \sum_{i=1}^{N} \left[\ell(\mathbf{u}_i, \mathbf{v}_i) + \ell(\mathbf{v}_i, \mathbf{u}_i)\right]. \tag{16}$$

A.3 Augmentation benchmark

To assess the computational costs associated with different augmentations and combinations of augmentations, they were applied to synthetic graphs of varying sizes while measuring runtime and memory usage.

For augmentations relevant to domain identification, synthetic graphs were generated to mimic the structure of real domain identification data. These graphs consisted of nodes with 50 numerical features, with feature similarities reflecting group structures, i.e., nodes within a group had more similar features than those in different groups. For phenotype prediction augmentations, graphs were designed to contain nodes annotated with a cell type feature and a cell size feature. Additionally, edges were annotated with a binary indicator distinguishing "near" from "distant" connections.

All individual augmentations applicable to either domain identification or phenotype prediction were tested on their respective synthetic graph types. Furthermore, combinations of augmentations, corresponding to those evaluated in the main experiments, were also benchmarked. Each augmentation or

combination was applied to synthetic graphs of increasing size, with each experiment repeated three times on a single GPU. For each run, both the runtime and peak GPU memory usage were recorded. The mean values across the three replicates were reported as the final result.

The results for domain identification augmentations are shown in Figure 5. Augmentation modes using *DropImportance* exhibit higher runtime compared to baseline augmentations (*DropFeatures* and *DropEdges*) and noise-based augmentations (*SpatialNoise* and *FeatureNoise*), though still running for 1 second or less for all graph sizes. Smoothing exhibits the highest memory usage of all the augmentations.

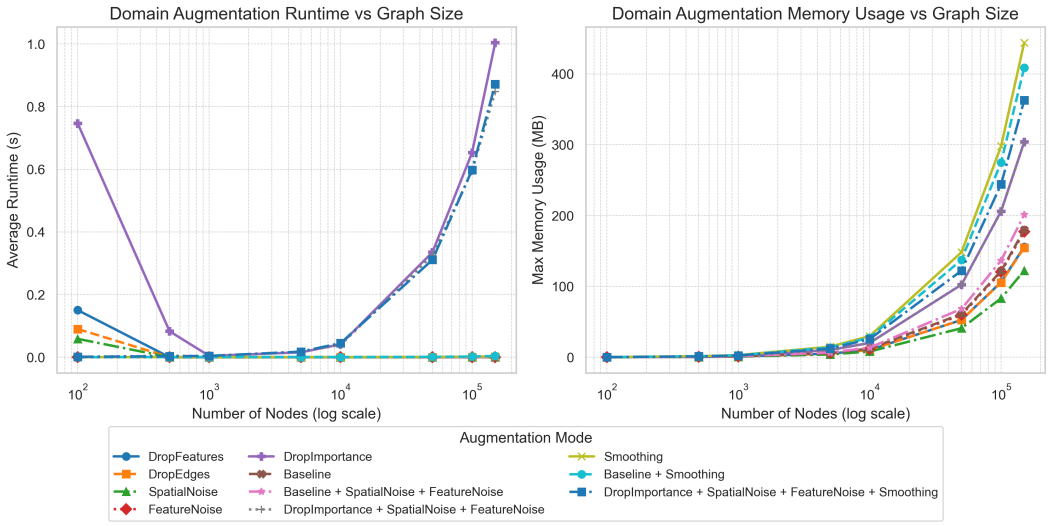

Figure 5: **Benchmark of domain identification augmentations.** Runtime (left) and peak GPU memory usage (right) for domain identification augmentations across increasing graph sizes. Each line represents either an individual augmentation or a combination of augmentations.

The results for phenotype prediction augmentations are shown in Figure 6. The runtime scaling trends are similar to those in the domain identification results. Augmentation modes using *DropImportance* scale worse than baseline and noise-based augmentations in both runtime and memory usage.

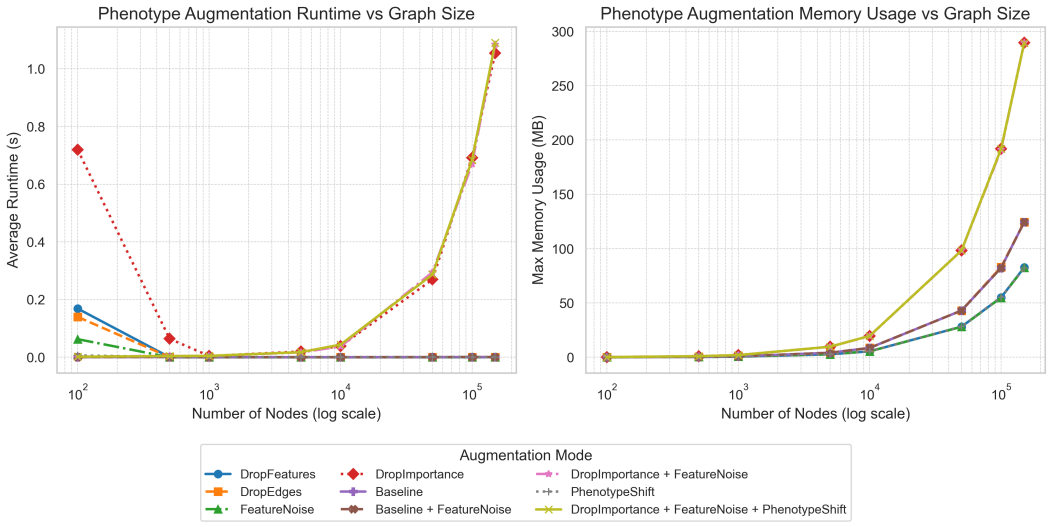

Figure 6: **Benchmark of phenotype prediction augmentations.** Runtime (left) and peak GPU memory usage (right) for phenotype prediction augmentations across increasing graph sizes. Each line represents either an individual augmentation or a combination of augmentations.

Overall, the benchmark highlights substantial variability in the computational efficiency of different augmentation strategies. Especially more complex augmentations, such as *DropImportance* and *Smoothing*, significantly increase runtime and memory consumption on large graphs, which also introduces considerable computational overhead during model training.

A.4 Classification metrics

To assess the performance of the phenotype prediction model, several binary classification metrics were used. These were computed from the predicted logits $\mathbf{z} \in \mathbb{R}^N$ and the ground truth binary labels $\mathbf{y} \in \{0,1\}^N$ for all N samples.

First, the predicted logits were transformed into probabilities using the sigmoid function:

$$\hat{\mathbf{p}} = \sigma(\mathbf{z}) = \frac{1}{1 + e^{-\mathbf{z}}} \tag{17}$$

A threshold $\tau \in [0,1]$ was applied to convert probabilities into binary predictions:

$$\hat{\mathbf{y}} = \mathbb{I}[\hat{\mathbf{p}} \geq \tau] \tag{18}$$

During validation, the threshold τ was chosen to maximize the F1 score across a set of candidate thresholds. Once the optimal threshold was selected, the following metrics were computed:

- **AUROC (Area Under the Receiver Operating Characteristic Curve)**: The AUROC quantifies the probability that a randomly chosen positive sample is ranked higher than a randomly chosen negative sample by the model's scoring function. Formally, if $s(x)$ denotes the prediction score, then

  $$\text{AUROC} = \mathbb{P}\big(s(x^+) > s(x^-)\big)\,,$$

  where x^+ and x^- are independent draws from the positive and negative classes, respectively. Equivalently, AUROC corresponds to the area under the curve tracing the true positive rate (TPR) against the false positive rate (FPR) as the classification threshold is varied:

  $$\text{TPR}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)}, \quad \text{FPR}(t) = \frac{\text{FP}(t)}{\text{FP}(t) + \text{TN}(t)},$$

  where TP, FP, TN, FN denote true/false positives/negatives at threshold t. A value of 0.5 corresponds to random guessing, while 1.0 indicates perfect class separability.
- **Precision**: Fraction of predicted positives that are correct:

  $$\text{Precision} = \frac{TP}{TP + FP} \tag{19}$$

- **Recall (Sensitivity)**: Fraction of actual positives that are correctly identified:

  $$\text{Recall} = \frac{TP}{TP + FN} \tag{20}$$

- **F1 Score**: Harmonic mean of precision and recall, balancing both metrics:

  $$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{21}$$

A.5 Clustering evaluation metrics

To evaluate the quality of clustering results obtained, three metrics were employed: Normalized Mutual Information (NMI), Homogeneity, and Completeness. These metrics assess how well the predicted clustering aligns with ground truth domain labels.

NMI measures the mutual dependence between the predicted clustering C and the ground truth labels Y, normalized by the entropy of both. It is defined as:

$$\text{NMI}(C, Y) = \frac{2 \cdot I(C;Y)}{H(C) + H(Y)} \tag{22}$$

where $I(C;Y)$ is the mutual information between C and Y, and $H(\cdot)$ denotes entropy. Mutual information is given by:

$$I(C;Y) = \sum_{c \in C} \sum_{y \in Y} P(c,y) \log \left(\frac{P(c,y)}{P(c)P(y)} \right) \tag{23}$$

Here, $P(c,y)$ is the joint probability of a sample being in cluster c and class y, while $P(c)$ and $P(y)$ are the marginal probabilities.

Homogeneity assesses whether each cluster contains only data points that belong to a single class. It is defined as:

$$\text{HOM}(C,Y) = \begin{cases} 1 & \text{if } H(Y|C) = 0 \\ 1 - \frac{H(Y|C)}{H(Y)} & \text{otherwise} \end{cases} \tag{24}$$

where $H(Y|C)$ is the conditional entropy of the ground truth labels given the cluster assignments, and $H(Y)$ is the entropy of the ground truth.

Completeness measures whether all members of a given class are assigned to the same cluster. It is defined as:

$$\text{COM}(C,Y) = \begin{cases} 1 & \text{if } H(C|Y) = 0 \\ 1 - \frac{H(C|Y)}{H(C)} & \text{otherwise} \end{cases} \tag{25}$$

where $H(C|Y)$ is the conditional entropy of the predicted cluster assignments given the true class labels.

A.6 Possible cell type transitions for the *PhenotypeShift* augmentation

We allow a restricted set of biologically motivated cell type transitions, reflecting known plasticity and differentiation processes in the tumor microenvironment:

- **Tumor adaptation:** Tumor cells (normal) can transition to hypoxic tumor states [34].
- **Fibroblast (CAF) plasticity:** Collagen CAFs may become myofibroblastic CAFs (mCAFs) or adapt to hypoxia; mCAFs can further switch into SMA^+ CAFs, PDPN^+ CAFs, vascular CAFs, or hypoxic CAFs; iCAFs can adopt PDPN^+ or IDO^+ states; IDO^+ CAFs can also adapt to hypoxia; tumor-promoting CAFs (tCAFs) can transition to hypoxic tCAFs [35–37].
- **CD4^+ T cell differentiation:** CD4 T cells can give rise to regulatory T cells (Tregs), PD1^+ exhausted cells, IDO^+ subsets, proliferative (Ki67^+) states, or TCF1/7^+ progenitor-like cells [38, 39].
- **CD8^+ T cell differentiation:** CD8 T cells can give rise to IDO^+ subsets, proliferative (Ki67^+) states, or TCF1/7^+ progenitor exhausted cells [39, 40].
- **Myeloid refinement:** Myeloid cells can be further refined into neutrophil identities, reflecting annotation resolution rather than a true biological transition [41].

A.7 Hyperparameter ranges used for tuning augmentations

Table 6: **Hyperparameter search ranges for graph augmentations.** For each augmentation, the tuned hyperparameters and their respective ranges are listed. Intervals denote uniform sampling from the specified range.

Augmentation	**Hyperparameter**	**Range**
DropEdges	p	$[0.1, 0.4]$
DropFeatures	p	$[0.1, 0.4]$
DropImportance	λ_p μ	$[0.4, 0.6]$ $[0.1, 0.4]$
SpatialNoise	σ_{spatial}	$[2.0, 30.0]$
FeatureNoise	σ_{feature}	$[0.05, 1.0]$
SmoothFeatures	α	$[0.0, 0.5]$
PhenotypeShift	p_{shift}	$[0.0, 0.3]$