spAug: Understanding Complementary Augmentation Strategies for Spatial Transcriptomics

Abstract

Spatial transcriptomics analyses remain limited by scarce annotations, distribution shift, and small disease cohorts. Recent AI advances offer two complementary augmentation strategies: *feature-level* augmentation enriches existing observations with external representations, whereas *synthetic-observation* augmentation expands the empirical support by generating additional observations. Here we present spAug, a unified evaluation framework that compares these strategies under matched downstream protocols across clustering, spot-level prediction, cross-slice prediction, and sample-level disease prediction. Rather than identifying a universally superior strategy, spAug shows that the two routes address different statistical bottlenecks. Feature-level augmentation is particularly effective when external representations provide task-relevant information, whereas synthetic-observation augmentation benefits settings where additional observations improve downstream estimation. Neither route consistently resolves cross-slice distribution shift. These results suggest that augmentation should be chosen according to the dominant limitation of the downstream task rather than the richness of representations or the fidelity of synthetic data alone.

Keywords: spatial transcriptomics; data augmentation; synthetic observations; feature augmentation

Mathematics Subject Classification (2020): 62XXX

1 Introduction

Spatial transcriptomics technologies, including the original Spatial Transcriptomics assay (Ståhl et al., 2016), Slide-seq (Rodriques et al., 2019), high-definition spatial transcriptomics (Vickovic et al., 2019), and DNA nanoball-patterned spatial arrays (Chen et al., 2022), measure gene expression while preserving spatial coordinates within tissue sections. These technologies have enabled transcriptome-wide studies of tissue architecture, spatial domains, cellular neighborhoods, and region-specific molecular programs. Public resources, such as the human dorsolateral prefrontal cortex (DLPFC) dataset (Maynard et al., 2021), further allow computational methods to be evaluated against layer annotations and section-level variation.

Despite this progress, downstream spatial transcriptomics analyses remain constrained by several forms of data scarcity. Spot-level annotations are often limited or unevenly distributed across tissue regions. Tissue sections may exhibit substantial section-to-section variation due to biological heterogeneity, technical effects, or image acquisition differences. Disease-labeled cohorts are frequently small, making sample-level prediction unstable. These limitations affect a

wide range of downstream tasks, including unsupervised spatial domain recovery, within-section spot annotation, cross-section prediction, and sample-level phenotype classification. Data augmentation is therefore a natural strategy for improving spatial transcriptomics analysis. The effectiveness of augmentation depends not only on the quality of the added information, but also on what statistical object is augmented and what limitation the downstream task faces.

Existing augmentation strategies can be broadly viewed through two complementary routes. The first route is *feature-level* augmentation, which enriches each measured spot with additional information. For example, histology-aware methods and recent foundation-model-based approaches attach image, pathology, language-model, or other external embeddings to the observed gene expression profile. spEMO (Liu et al., 2025) is one representative example of this route, as it leverages multimodal foundation-model embeddings to improve spatial multi-omic representations. Schematically, feature-level augmentation changes the input representation from x_i to $[x_i, e_i]$, where e_i denotes an external embedding associated with the same measured spot.

The second route is *synthetic-observation* augmentation. Instead of enriching the features attached to existing spots, it expands the set of observations available to the downstream method by adding generated spatial transcriptomics spots. Schematically, this route changes the empirical training set from $\mathcal{R}$ to $\mathcal{R} \cup \mathcal{S}$, where $\mathcal{R}$ denotes real observations and $\mathcal{S}$ denotes generated observations.

Although both routes inject additional information into downstream analysis, they do so through fundamentally different mechanisms: one improves the quality of each observation, whereas the other increases the quantity of observations available for learning. This paper evaluates the two routes in parallel rather than treating feature augmentation as a secondary control for synthetic observations.

An intuitive way to distinguish the two routes is to ask what is missing from the downstream analysis. If the limitation is that each measured spot lacks informative covariates, then feature-level augmentation is natural: each existing spot is enriched with additional representations, such as histology or foundation-model embeddings. The set of spots remains unchanged, but each spot carries a richer feature vector. In contrast, if the limitation is that the downstream method sees limited observations, uneven spatial coverage, or insufficient examples from certain labels or tissue regions, then synthetic-observation augmentation is more appropriate. It adds generated spots to the training set and thereby changes the empirical distribution used by the downstream learner. In this sense, feature augmentation asks how to better represent the observed spots, whereas synthetic-observation augmentation asks whether adding plausible new spots can improve the statistical support available to the downstream task.

Existing generators range from statistical simulators to modern deep generative models, including count-based simulators (e.g., Baruzzo et al. (2020)), GANs (e.g., Marouf et al. (2020)), and diffusion models (e.g., Luo et al. (2024)). However, high generator fidelity alone does not determine downstream usefulness. This gap motivates spAug. spAug provides a unified evaluation framework for studying two complementary augmentation routes under matched downstream protocols. Rather than asking only whether a feature source is informative or whether a generator produces realistic synthetic data, spAug asks which augmentation route and

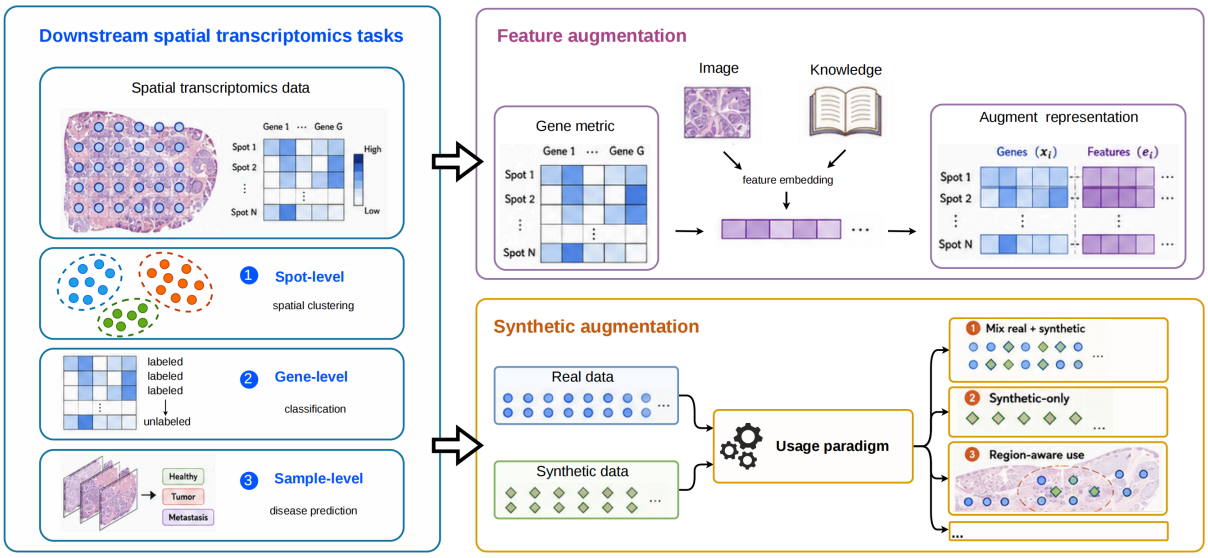


Figure 1: Conceptual overview of spAug. Feature-level augmentation adds external representations to measured spots, whereas synthetic-observation augmentation adds generated spatial transcriptomics observations to the empirical training support. spAug evaluates these augmentation routes under matched downstream protocols and compares strategies for incorporating synthetic pools, including Global Real+Synthetic, Synthetic-only, Local Real+Synthetic, Weighted ERM, and Weighted Aggregation. Utility is assessed relative to real-only baselines across tasks, datasets, feature sources or generators, parameters, and downstream models.

downstream configuration improve the target analysis. The resulting utility map is intentionally task dependent: a route that improves low-label spot prediction may not improve cross-slice transfer or sample-level prediction, and a route that helps under one backend may fail under another. This conceptual separation is summarized in Figure 1, which positions feature-level augmentation and synthetic-observation augmentation as two complementary ways of increasing information available to downstream spatial transcriptomics analyses.

The contributions of this manuscript are fourfold. First, we present a conceptual framework for spatial transcriptomics augmentation by distinguishing two complementary augmentation routes: feature-level augmentation, which enriches the representation of existing observations, and synthetic-observation augmentation, which expands the observation set through generated spots. This perspective emphasizes that the two routes modify different statistical objects and should therefore be viewed as complementary rather than competing approaches. Second, we develop spAug, a unified evaluation framework that enables matched comparisons between augmentation routes. The framework separates the roles of augmentation route, feature source or synthetic generator, augmentation strategy, downstream task, and evaluation protocol, allowing each component to be assessed under consistent experimental conditions. Third, we provide a systematic empirical characterization of augmentation across representative downstream spatial transcriptomics tasks, including spatial clustering, within-slice low-label prediction, cross-slice generalization, and sample-level disease prediction. Rather than searching for a universally superior augmentation method, spAug identifies the conditions under which each augmentation route succeeds or fails. Finally, we derive practical guidance for augmentation in spatial transcriptomics. Our results show that augmentation utility depends on the dominant statistical limitation of the downstream task. Feature-level augmentation is most effective when

richer representations provide task-relevant information, whereas synthetic-observation augmentation is beneficial when additional observations improve downstream estimation. Neither route consistently resolves cross-slice distribution shift, highlighting the need to select augmentation strategies according to the underlying statistical bottleneck rather than representation richness or generator fidelity alone.

The remainder of the paper is organized as follows. Section 2 reviews related work on downstream evaluation, feature-level augmentation, and synthetic-observation generation. Section 3 introduces the spAug framework, including the unified evaluation design and augmentation strategies. Section 4 presents task-oriented evaluations together with mechanistic analyses of augmentation utility. Section 5 concludes with a discussion of the complementary roles of the two augmentation routes and directions for future work.

2 Related Work

2.1 Downstream evaluations in spatial transcriptomics

Spatial transcriptomics methods are commonly assessed through downstream biological or predictive tasks rather than intrinsic representation quality. Spatial domain identification methods, including BayesSpace (Zhao et al., 2021), SpaGCN (Hu et al., 2021), STAGATE (Dong and Zhang, 2022), and GraphST (Long et al., 2023), evaluate agreement between inferred domains and tissue annotations or histological structures. Cell-type mapping and deconvolution methods, such as Tangram (Biancalani et al., 2021), RCTD (Cable et al., 2022), cell2location (Kleshchevnikov et al., 2022), and SPOTlight (Elosua-Bayes et al., 2021), measure the transfer of cell-type information from single-cell references to spatial observations. Gene-level methods, including SpatialDE (Svensson et al., 2018), SPARK (Sun et al., 2020), and SPARK-X (Zhu et al., 2021), identify spatially variable genes, while analysis platforms, such as Giotto (Dries et al., 2021) and Squidpy (Palla et al., 2022), provide standardized workflows for neighborhood analysis, spatial statistics, image integration, and visualization.

These studies establish a task-oriented evaluation paradigm in spatial transcriptomics. Accordingly, augmentation strategies should be evaluated according to their downstream impact rather than surrogate measures such as generation fidelity or representation quality alone. Following this philosophy, spAug evaluates augmentation according to downstream performance. Unlike previous comparative studies that focus on downstream algorithms themselves, however, our objective is to understand how different augmentation strategies affect downstream analyses under matched evaluation protocols.

2.2 Feature-level augmentation

Recent advances in artificial intelligence have enabled feature-level augmentation through increasingly powerful pretrained representation models. Existing approaches enrich each measured spot by incorporating complementary information from paired histology images, neighboring spatial context, or pretrained multimodal representations. Histology-aware methods such as stLearn (Pham et al., 2020) and ST-Net (He et al., 2020) extract morphological features from

H&E images, while graph-based methods further integrate local spatial context for representation learning. More recently, pathology foundation models (PFMs), including GPFM (Ma et al., 2024), UNI (Chen et al., 2024), and GigaPath (Xu et al., 2024), have enabled transferable image representations that can be directly applied to spatial transcriptomics. Representative frameworks such as spEMO (Liu et al., 2025) further combine foundation-model embeddings with molecular measurements for multimodal analysis.

Despite differences in architecture and pretraining objectives, these methods share a common statistical mechanism: they enrich the representation of existing observations without altering the observation set. Formally, each measured spot is transformed from x_i to $[x_i, e_i]$, where e_i denotes external representations extracted from complementary modalities. Consequently, feature-level augmentation improves the information available for each observation while leaving the empirical observation set unchanged.

2.3 Synthetic-observation augmentation

Synthetic transcriptomics generation has evolved from statistical simulators to modern deep generative models. Statistical simulators, such as Splatter (Zappia et al., 2017), SPARSim (Baruzzo et al., 2020), and SRTsim (Zhu et al., 2023), reproduce count distributions and spatial expression characteristics, whereas recent approaches including scGAN (Marouf et al., 2020) and scDiffusion (Luo et al., 2024) employ adversarial or diffusion-based models to generate increasingly realistic transcriptomic observations. Broader simulation frameworks, such as scDesign3 (Song et al., 2024) and multi-task simulation benchmarks (Liang et al., 2024), have further emphasized the importance of evaluating synthetic data across multiple criteria. In spAug, synthetic observations are generated using the native implementations of representative generators and subsequently harmonized into a unified spatial transcriptomics schema containing expression profiles, spatial coordinates, slice or sample identifiers, and task-specific labels when applicable.

Most existing studies emphasize the fidelity of generated data or evaluate a single downstream application using one augmentation protocol. By contrast, comparatively less attention has been paid to how generated observations should be incorporated into downstream analyses. Outside spatial transcriptomics, recent work has begun to treat synthetic data as a controllable statistical resource rather than a fixed replacement for real data. Synthetic volume expansion (Shen et al., 2023) studies how predictive performance changes with increasing amounts of synthetic data, CoDSA (Tian and Shen, 2025) investigates targeted allocation of synthetic samples to under-represented regions, and scaling-law analyses (Jain et al., 2024) characterize trade-offs between real and synthetic observations under weighted empirical risk minimization. Collectively, these studies suggest that the effectiveness of synthetic observations depends not only on generator fidelity, but also on how generated observations are allocated, weighted, and coupled to downstream tasks.

2.4 From data generation to augmentation

Although feature-level augmentation and synthetic-observation augmentation have developed largely as separate research directions, both aim to improve downstream spatial transcriptomics analyses by introducing additional information. The former enriches the representation of ex-

isting observations, whereas the latter expands the available observation set. Existing studies typically evaluate these augmentation routes independently, making it difficult to disentangle the effects of augmentation route, feature source or synthetic generator, augmentation strategy, and downstream task. Consequently, it remains unclear when each augmentation route is most appropriate and whether their relative effectiveness depends on the statistical limitation of the downstream problem.

spAug addresses this gap by providing a unified evaluation framework that compares feature-level and synthetic-observation augmentation under matched downstream protocols. Rather than identifying a universally superior augmentation strategy, the goal is to characterize the complementary roles of these augmentation routes across representative spatial transcriptomics tasks.

3 Methods

3.1 Evaluation design

spAug evaluates two complementary augmentation routes: feature-level enrichment of measured spots and synthetic-observation (*Syn-obs*) expansion of the empirical support. The feature-level route extracts PFM embeddings from paired histology image patches and integrates them with gene expression representations through feature concatenation or graph-based architectures. The Syn-obs route generates synthetic pools from real data and constructs augmented training or analysis inputs using route-specific strategies. Each experimental configuration is defined by a unique combination of augmentation route, route-specific strategy, and downstream model, as summarized in Table 1. Synthetic pools are generated once and reused across multiple Syn-obs strategies.

Table 1: Evaluation design for the two augmentation routes in spAug.

Task	Dataset	Route	Input / strategy	Backend / model
Spatial clustering	DLPFC	Feature-level	PFM Concatenation	PCA-spatial Leiden
			PFM Graph Integration	SpaGCN
		Syn-obs	Global Real+Synthetic	PCA-spatial Leiden
			Local Real+Synthetic	PCA-spatial Leiden
Low-label spot prediction	DLPFC	Feature-level	PFM Concatenation	LR, MLP, XGB
		Syn-obs	Global Real+Synthetic	
			Synthetic-only	
			Local Real+Synthetic	
			Weighted ERM	
Cross-slice spot prediction	DLPFC	Feature-level	PFM Concatenation	LR, MLP, XGB
		Syn-obs	Global Real+Synthetic	
			Synthetic-only	
			Local Real+Synthetic	
			Weighted ERM	
Sample-level prediction	Kidney, Bowel, Brain	Feature-level	PFM Sample Aggregation	LR, MLP, XGB
		Syn-obs	Global Real+Synthetic	
			Synthetic-only	
			Weighted Aggregation	

To describe the two augmentation routes within a common framework, we first introduce a unified representation for both real and generated observations. This representation is shared across all downstream tasks and augmentation routes. We write a generic observation as

$$z_i = (x_i, p_i, r_i, c_i, y_i),$$

where $x_i \in \mathbb{R}^G$ is the expression vector on G genes, $p_i \in \mathbb{R}^2$ is the spatial coordinate, r_i is a spatial-region index when a local region partition is used, c_i records section-, sample-, fold-, or cohort-level metadata, and y_i is a spot- or sample-level label when available. Unsupervised generator inputs exclude evaluation labels. Supervised synthetic pools carry labels only when generated from label-wise training data. Cross-slice synthetic pools are restricted to training sections for each split, and sample-level disease models use real spots only for test samples.

3.2 Datasets and task inputs

DLPFC experiments use 12 human dorsolateral prefrontal cortex sections (Maynard et al., 2021), 47,280 real spots for spatial clustering, and a shared 3,000-gene highly variable gene space after filtering and normalization. Labels are excluded from unsupervised model inputs and used only for evaluation. The final spatial clustering task uses PCA-spatial Leiden clustering with seed 42. For each real-only or augmented input, expression profiles are projected into PCA space, spatial coordinates are scaled, and the two representations are concatenated to construct a joint expression–spatial neighbor graph. Leiden clustering (Traag et al., 2019) is then performed on this graph. When synthetic spots are included, clustering is performed on the combined real–synthetic input, but evaluation metrics are computed only on the real spots after removing synthetic observations. In the within-slice low-label task, each section is treated as an independent classification problem: within each slice–label group, 25%, 50%, or 75% of real spots are sampled for training under seeds 42, 43, and 44, and the remaining real spots in the same section are used for testing. In the cross-slice task, three peer 8-train/4-test split seeds are evaluated to assess split stability. Synthetic pools may be pre-generated for all DLPFC sections, but each split only permits synthetic observations from the training sections.

Sample-level disease prediction uses three cohort-level spatial transcriptomics benchmarks. The Kidney cohort contains 36 samples, 102,375 spots, and 2,000 genes, with 24 squamous cell carcinoma and renal cell carcinoma (SCCRCC) and 12 Healthy samples. The Bowel cohort contains 66 samples, 203,238 spots, and 1,063 genes, with 55 Cancer and 11 Healthy samples. The Brain cohort contains 26 samples, 80,895 spots, and 2,000 genes, and is constructed from extramedullary plasmacytoma (EPM) tumor samples and spatialLIBD healthy full-transcriptome samples, yielding 14 EPM and 12 Healthy samples. For disease prediction, the Brain labels are treated as a binary Cancer-versus-Healthy task. All three sample-level tasks are evaluated with five-fold stratified sample splits. When patient identifiers are available, folds are constructed with StratifiedGroupKFold using patient identity as the group; samples without patient identifiers are grouped by sample ID. If a group-stratified split cannot keep both disease labels in each train and test partition, the data-preparation pipeline falls back to StratifiedKFold and validates label coverage in every fold. Within each fold, synthetic spots are generated and used

only for training samples, whereas test samples are represented exclusively by real spots. The source data are treated as cohort-specific evaluation tasks, with HEST cited when the curated HEST-derived files are used (Jaume et al., 2024).

3.3 Synthetic-pool generation

Five native generators are evaluated: SRTsim (Zhu et al., 2023), Splatter (Zappia et al., 2017), SPARSim (Baruzzo et al., 2020), scGAN (Marouf et al., 2020), and scDiffusion (Luo et al., 2024). In spAug, a generator is used to construct a task-matched synthetic pool rather than a downstream input directly. Synthetic generation is stratified by the downstream task: spatial-clustering pools are generated slice-wise; within-slice low-label pools are generated within slice and label strata; cross-slice pools are generated at the section level but are used only when the section belongs to the training split; and sample-level pools are generated within each sample, with cross-validation folds determining whether a sample's synthetic spots are eligible for training.

Let $\mathcal{R}_a^{\mathrm{tr}}$ denote the real training observations available for stratum a, where a stratum may be a slice, label, sample, region, or a task-specific combination of these factors. Each generator m is fitted only on task-allowed observations, using the components required by its native model, such as expression profiles alone for expression generators and spatial information when the generator uses spatial structure:

$$\widehat{P}_{m,a} = \mathrm{Fit}_m \left(\{(x_i, p_i, r_i, c_i, y_i) : (x_i, p_i, r_i, c_i, y_i) \in \mathcal{R}_a^{\mathrm{tr}}\} \right),$$

and then used to sample synthetic expression profiles

$$(\tilde{x}_j, \tilde{p}_j, \tilde{r}_j, \tilde{c}_j, \tilde{y}_j) \sim \widehat{P}_{m,a}.$$

The five generators differ in how $\widehat{P}_{m,a}$ is obtained. SRTsim uses a reference spatial transcriptomics section and preserves spatial expression ordering when simulating expression, making it naturally aligned with slice-level spatial structure. Splatter and SPARSim are count-generation models; in this evaluation they are fitted within task-defined strata so that their simulated expression profiles can be paired with coordinates from the corresponding slice, label, sample, or local region. This stratum-wise adaptation allows expression-matrix simulators to be used as spatial transcriptomics generators while preserving the downstream split constraints. scGAN and scDiffusion learn deep generative distributions over expression profiles within the permitted training material. Because these deep generators do not natively define spot coordinates, spAug assigns their generated profiles to coordinates through task-aware spatial resampling or label-aware coordinate perturbation from the matched stratum. Thus, the evaluation assesses generators through a common synthetic-pool interface while keeping the construction of the pool tied to the spatial and supervised structure of each task.

3.4 Synthetic-observation augmentation strategies

spAug treats a synthetic-observation augmentation strategy as a rule for converting a fixed real dataset and a fixed synthetic pool into the object consumed by a downstream algorithm. This

distinction is important because the generator is not evaluated only through its marginal fidelity. The same pool may be appended to real observations, used alone, sampled in a spatially local manner, or assigned a separate contribution in a supervised objective.

The strategies are task-compatible. Spatial clustering has no supervised loss, so the formal clustering matrix evaluates Global Real+Synthetic and Local Real+Synthetic, not Weighted empirical risk minimization (ERM). Synthetic-only training is used as a diagnostic for label-supervised spot prediction and sample-level disease prediction, where a classifier can be fitted from labeled synthetic observations and evaluated on held-out real observations. Weighted ERM is used for supervised spot prediction. For sample-level disease prediction, the same weighting idea is adapted to sample-level aggregation as Weighted Aggregation: real and synthetic spots within a training sample receive separate total mass before the sample is summarized. Local Real+Synthetic is not included in the sample-level matrix because stable, comparable local regions are not defined across heterogeneous cohorts. Figure 2 summarizes how a fixed synthetic pool is converted into task-specific augmented inputs.

For generator m, the synthetic pool matched to the fixed unit is

$$\mathcal{G}_m = \{(\tilde{x}_j, \tilde{p}_j, \tilde{r}_j, \tilde{c}_j, \tilde{y}_j)\}_{j=1}^{N_m}.$$

Synthetic observations share the same gene index, coordinate convention, and metadata schema as the real observations. The pool is generated only from the real material allowed by the task split. In cross-slice prediction, for example, synthetic spots are produced from training sections and are never generated from held-out sections. A synthetic-observation strategy with parameters θ is therefore defined as

$$\Phi_{\theta,m} : (\mathcal{R}^{\mathrm{tr}}, \mathcal{G}_m) \mapsto (\mathcal{D}^{\mathrm{tr}}_{m,\theta}, \omega_{m,\theta}),$$

where $\mathcal{D}^{\mathrm{tr}}_{m,\theta}$ is the augmented training or analysis input and $\omega_{m,\theta}$ is an optional observation-weight vector. A downstream procedure $\mathcal{F}$ then produces an output $\hat{o}_{m,\theta}$, which may be a classifier, a sample-level predictor, or a clustering assignment:

$$\hat{o}_{m,\theta} = \mathcal{F}(\mathcal{D}^{\mathrm{tr}}_{m,\theta}, \omega_{m,\theta}), \qquad \Delta M_{m,\theta} = M(\hat{o}_{m,\theta}, \mathcal{R}^{\mathrm{ev}}) - M(\hat{o}^{\mathrm{real}}, \mathcal{R}^{\mathrm{ev}}),$$

where M is the task metric, $\mathcal{R}^{\mathrm{ev}}$ is the held-out evaluation dataset, and $\hat{o}^{\mathrm{real}}$ is the matched real-only baseline output for the same fixed unit. All reported primary metrics are oriented so that larger values indicate better performance. When results are summarized across sections, seeds, splits, or folds, this unit-level quantity is aggregated over the corresponding set of evaluation units.

The sampling formulas below use strata to preserve task-relevant structure. Let $\mathcal{C}$ be the set of matching strata. In unsupervised clustering, this can be a slice-level stratum; in supervised spot prediction, it is usually label or label–region; in sample-level prediction it is sample or sample–label. For a stratum $a \in \mathcal{C}$, write $\mathcal{R}^{\mathrm{tr}}_a$ and $\mathcal{G}_{m,a}$ for the real and synthetic subsets in that stratum, and let $\mathrm{Sample}(B, q)$ denote sampling without replacement $\min\{|B|, q\}$ observations from B. Table 2 summarizes the augmented input induced by each synthetic-observation strategy.

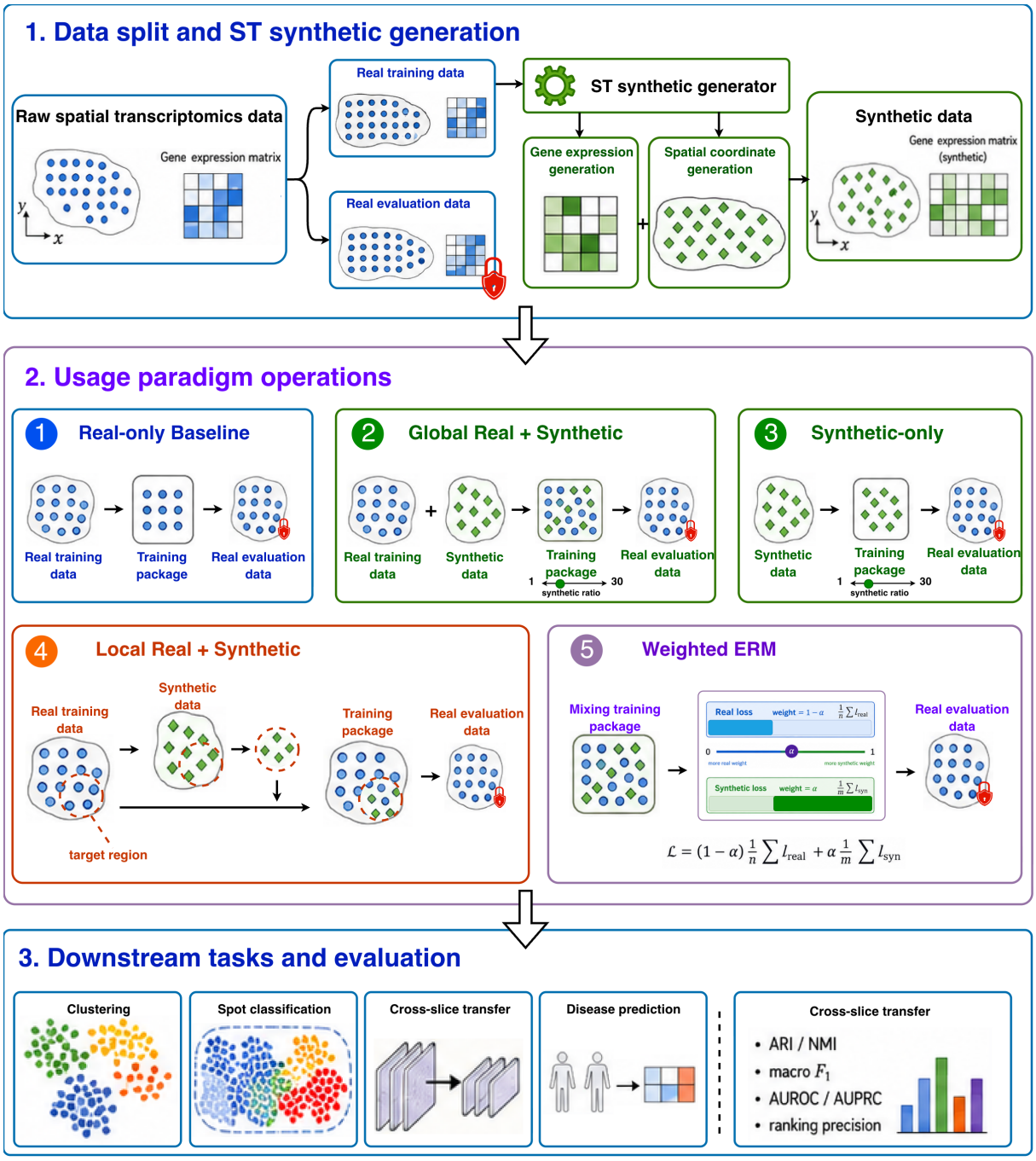


Figure 2: Synthetic-observation augmentation strategies in spAug. A real-only input provides the matched baseline. Global Real+Synthetic appends a stratum-matched synthetic subset; Synthetic-only removes real training observations and tests synthetic substitutability; Local Real+Synthetic samples synthetic observations by spatial region under an adaptive local cap; and Weighted ERM or Weighted Aggregation separates synthetic count from synthetic contribution through α. Each downstream task activates only the strategies that are meaningful for its prediction or clustering objective.

Table 2: Task-compatible synthetic-observation augmentation strategies in spAug. The same synthetic pool can lead to different augmented inputs depending on the strategy, task, and parameters.

Strategy	Augmented input	Parameter	Task scope
Real-only Baseline	$\mathcal{R}^{\mathrm{tr}}$	none	all tasks
Global Real+Synthetic	$\mathcal{R}^{\mathrm{tr}} \cup \mathcal{S}_m^{\mathrm{global}}(\rho)$	ratio ρ	all tasks
Synthetic-only	$\mathcal{S}_m^{\mathrm{global}}(\rho)$	ratio ρ	supervised spot, sample-level
Local Real+Synthetic	$\mathcal{R}^{\mathrm{tr}} \cup \mathcal{S}_m^{\mathrm{local}}(\theta)$	$\rho_{\max}$	DLPFC spot tasks
Weighted ERM	$\mathcal{R}^{\mathrm{tr}} \cup \mathcal{S}_m^{\mathrm{global}}(\rho)$ with observation weights	ρ, α	supervised spot, sample-level

The real baseline is $\mathcal{D}_{\mathrm{real}}^{\mathrm{tr}} = \mathcal{R}^{\mathrm{tr}}$. When a downstream learner accepts observation weights, the real baseline uses the uniform optional weight vector $\omega_i^{\mathrm{real}} = 1/n$. Global Real+Synthetic augmentation appends synthetic observations sampled at a ratio ρ relative to the real observations in each stratum:

$$\mathcal{S}_m^{\mathrm{global}}(\rho) = \bigcup_{a \in \mathcal{C}} \mathrm{Sample}\left(\mathcal{G}_{m,a}, \lfloor \rho |\mathcal{R}_a^{\mathrm{tr}}| \rceil\right), \qquad \mathcal{D}_{m,\rho}^{\mathrm{tr}} = \mathcal{R}^{\mathrm{tr}} \cup \mathcal{S}_m^{\mathrm{global}}(\rho),$$

where $\lfloor x \rceil$ denotes the integer nearest to x. Thus $\rho = 1$ aims to add approximately one synthetic observation for each matched real observation, whereas larger ρ values expand the empirical support more aggressively, subject to the finite pool size.

Synthetic-only training uses the same sampled synthetic subset but excludes real training observations:

$$\mathcal{D}_{m,\rho,\mathrm{synonly}}^{\mathrm{tr}} = \mathcal{S}_m^{\mathrm{global}}(\rho).$$

Evaluation remains on real held-out spots, sections, or samples. This strategy is therefore not an intended replacement for real-data analysis; it is a diagnostic of whether the generated observations are sufficiently aligned with the task labels and decision boundaries to stand alone.

Local Real+Synthetic further conditions sampling on spatial regions. Let $\mathcal{K} = \{\kappa(p_i)\}$ be the set of slice-aware spatial regions, implemented as coordinate-grid or task-specific local regions; each $k \in \mathcal{K}$ indexes a possible value of the region label r_i. A local policy assigns each region k a multiplier $\lambda_k(\theta)$ and samples

$$\mathcal{S}_m^{\mathrm{local}}(\theta) = \bigcup_{a \in \mathcal{C}} \bigcup_{k \in \mathcal{K}} \mathrm{Sample}\left(\mathcal{G}_{m,a,k}, \lfloor \lambda_k(\theta) |\mathcal{R}_{a,k}^{\mathrm{tr}}| \rceil\right).$$

For Local Real+Synthetic, $\theta = \rho_{\max}$, where $\rho_{\max}$ is a local cap (i.e., an upper limit within each spatial region) on the synthetic-to-real sampling ratio. It is not a global augmentation ratio: instead, it controls how strongly sparse or selected regions may be enriched relative to the number of real observations available in the same region. Let $\mathcal{R}_k^{\mathrm{tr}}$ denote all real training observations in region k after aggregating over the matching strata used by the task. Region multipliers satisfy

$$\lambda_k(\rho_{\max}) \in \{1, \ldots, \rho_{\max}\}, \qquad |\mathcal{R}_k^{\mathrm{tr}}| \geq |\mathcal{R}_l^{\mathrm{tr}}| \Rightarrow \lambda_k(\rho_{\max}) \leq \lambda_l(\rho_{\max}).$$

Dense regions therefore receive lower ratios and sparse regions receive higher ratios. Candidate synthetic spots are drawn from the same region when available. In supervised spot prediction,

matching is label-aware and region-aware; if an exact label–region subset is unavailable, the sampler falls back to label-matched or nearest local candidates to keep the selected synthetic subset spatially constrained.

The role of $\rho_{\max}$ is to make local augmentation adaptive but bounded. Without such a cap, a small or sparse region could be dominated by synthetic observations, or the sampler could request more matched local synthetic candidates than are available in the synthetic pool. The cap therefore preserves the intended spatial reallocation: dense regions receive little enrichment, sparse regions can receive stronger enrichment, and every request remains constrained by the finite matched pool.

In addition to simply appending synthetic observations as in Global/Local Real+Synthetic, weighted empirical risk minimization (ERM) provides an alternative strategy for supervised spot-level prediction with synthetic data (Jain et al., 2024). Rather than combining real and synthetic observations with equal weight, weighted ERM assigns separate loss weights to the real and synthetic subsets. For a non-empty selected synthetic subset, the objective is

$$\mathcal{L}_\alpha(f) = (1-\alpha)\frac{1}{|\mathcal{R}^{\mathrm{tr}}|}\sum_{(x_i,y_i)\in\mathcal{R}^{\mathrm{tr}}} \ell(f(x_i),y_i) + \alpha\frac{1}{|\mathcal{S}_m^{\mathrm{global}}(\rho)|}\sum_{(\tilde{x}_j,\tilde{y}_j)\in\mathcal{S}_m^{\mathrm{global}}(\rho)} \ell(f(\tilde{x}_j),\tilde{y}_j), \tag{1}$$

where ℓ denotes the supervised loss used by the downstream classifier and $\alpha \in [0,1]$ denotes the total synthetic contribution. If no synthetic observations are selected for a unit after task-compatible matching, the weighted augmented input reduces to the matched real-only input for that unit. This formulation decouples the number of sampled synthetic observations from their contribution to the objective: two variants can use the same ρ and therefore the same synthetic count, but different α values to test how strongly the downstream learner should trust the synthetic pool.

For sample-level disease prediction, Weighted Aggregation adapts the weighting idea to sample aggregation rather than a per-spot supervised loss. For a training sample s, let $\mathcal{B}_s^{\mathrm{real}}$ and $\mathcal{B}_s^{\mathrm{syn}}$ denote the real and selected synthetic spots associated with that sample, and let $\mathcal{B}_s = \mathcal{B}_s^{\mathrm{real}} \cup \mathcal{B}_s^{\mathrm{syn}}$. The aggregation weights are

$$w_i^{(s)}(\alpha) = \begin{cases} (1-\alpha)/|\mathcal{B}_s^{\mathrm{real}}|, & z_i \in \mathcal{B}_s^{\mathrm{real}}, \\ \alpha/|\mathcal{B}_s^{\mathrm{syn}}|, & z_i \in \mathcal{B}_s^{\mathrm{syn}}. \end{cases}$$

When $\mathcal{B}_s^{\mathrm{syn}}$ is empty, the sample reduces to the real-only representation. Let $\psi(x_i)$ be the PCA-transformed expression vector for spot i. The weighted sample representation is

$$b_s(\alpha) = \mathrm{Agg}\left(\{(\psi(x_i), p_i, w_i^{(s)}(\alpha)) : z_i \in \mathcal{B}_s\}\right),$$

where Agg computes the weighted PCA-score moments and quantiles together with coordinate summaries. The sample-level classifier g is then trained with the ordinary sample-level empirical risk

$$\mathcal{L}_{\mathrm{sample}}(g) = \frac{1}{|\mathcal{T}|}\sum_{s\in\mathcal{T}} \ell(g(b_s(\alpha)), Y_s),$$

where $\mathcal{T}$ is the set of training samples, Y_s is the disease label of sample s, and ℓ is the sample-

level classification loss. Thus α has the same interpretation as total synthetic contribution, but it acts through the construction of the sample representation rather than through a spot-level supervised loss.

The above framework is developed for synthetic-observation augmentation. Section 3.5 introduces the complementary feature-level route $x_i \to [x_i, e_i]$ instantiated with PFM embeddings. When the two routes use the same PCA-spatial or classifier backend, their Δ metrics in Sections 4.2.1–4.2.4 are computed against matched real-only baselines and can be compared within the same task protocol. SpaGCN-based PFM integration is reported separately as a graph-aware architectural reference because it changes the clustering backend.

3.5 Feature-level augmentation with pathology foundation models

Feature-level augmentation enriches each measured spot with external embeddings derived from paired histology images, a route that is structurally complementary to synthetic-observation augmentation. Whereas Section 3.4 develops five task-compatible strategy variants that expand the observation set $\mathcal{R} \to \mathcal{R} \cup \mathcal{S}$, here we instantiate the paired transformation $x_i \to [x_i, e_i]$ defined in Section 2.2 using pathology foundation model (PFM) embeddings as e_i. To evaluate this route, we incorporate representations extracted by multiple PFMs as additional input channels alongside the gene-expression PCA used in the main analysis. The common extraction pipeline crops each spot's corresponding histology patch h_i at 224×224 px resolution (centered on the spot coordinates with radius 112 px) and feeds it through a frozen pre-trained vision encoder to obtain a dense embedding vector.

Formally, let $\mathcal{R}^{\text{tr}}$ denote the set of measured training spots, where each spot $z_i = (x_i, p_i, r_i, c_i, y_i)$ has a gene-expression vector $x_i \in \mathbb{R}^G$ and is additionally paired with a histology patch h_i. Feature-level augmentation applies a feature-integration rule

$$\Psi_{\text{PFM}} : (x_i, h_i) \mapsto x_i^{\text{aug}} = [\psi_{\text{expr}}(x_i),\ \phi_{\text{PFM}}(h_i)],$$

where $\psi_{\text{expr}} : \mathbb{R}^G \to \mathbb{R}^{d_{\text{expr}}}$ is the gene-expression feature map (PCA projection) and $\phi_{\text{PFM}} : \mathbb{R}^{224\times224\times3} \to \mathbb{R}^{d_{\text{PFM}}}$ is a frozen PFM vision encoder. In the DLPFC spot-level tasks, $d_{\text{expr}} = 50$; in sample-level disease prediction, $d_{\text{expr}} = 30$ following the PFM Sample Aggregation pipeline. The augmented representation x_i^{aug} replaces the expression-only representation in the downstream input while preserving the observation indices, coordinates, and labels. Following the same evaluation framework as Section 3.4, the downstream output $\hat{o}_{\text{PFM}}$ is compared with the matched real-only baseline $\hat{o}^{\text{real}}$ on the same evaluation objects $\mathcal{R}^{\text{ev}}$:

$$\Delta M_{\text{PFM}} = M(\hat{o}_{\text{PFM}}, \mathcal{R}^{\text{ev}}) - M(\hat{o}^{\text{real}}, \mathcal{R}^{\text{ev}}),$$

where positive values indicate improvement over the matched expression-only baseline.

Four PFMs are evaluated as zero-shot feature extractors, covering a range of architectures, pretraining data, and output dimensions:

- **GPFM** (Ma et al., 2024): a ViT-based pathology foundation model trained via teacher-student distillation from multiple expert pathology foundation models combined with self-

Table 3: Feature-level integration strategies in spAug. Each strategy specifies how the PFM embedding is combined with gene-expression features for a downstream backend.

Strategy	Augmented representation	Compatible tasks
PFM Concatenation	$x_i^{\text{aug}} = [\psi_{\text{expr}}(x_i),\ \phi_{\text{PFM}}(h_i)]$	Spatial clustering (PCA-spatial Leiden); spot-level prediction
PFM Graph Integration	x_i^{aug} as node feature in SpaGCN	Spatial clustering (SpaGCN)
PFM Sample Aggregation	$\text{Agg}(\{(\psi_{\text{expr}}(x_i), \phi_{\text{PFM}}(h_i), p_i) : z_i \in \mathcal{B}_s\})$	Sample-level disease prediction

distillation, pretrained on approximately 95,000 whole-slide images spanning 34 tissue types and 56 datasets, producing a 1024-dimensional embedding.

- **UNI** (Chen et al., 2024): a ViT-Large/16 vision encoder with 304M parameters, pretrained on the Mass-100K dataset (approximately 100,000 whole-slide images and 100 million patches) via DINOv2 self-supervised learning, yielding a 1024-dimensional embedding.
- **UNI2-h**: an enhanced version of UNI with a higher-resolution ViT-H backbone, producing a 1536-dimensional embedding.
- **GigaPath** (Xu et al., 2024): a two-stage vision transformer comprising a ViT-based tile encoder pretrained with DINOv2 and a LongNet slide encoder with dilated attention for whole-slide modeling, pretrained on the Prov-Path dataset (approximately 171,000 whole-slide images and 1.38 billion patches), producing a 1536-dimensional embedding.

For concatenation-based feature augmentation, the image embedding is concatenated with the task-specific gene-expression PCA vector to form a joint $d_{\text{expr}} + d_{\text{PFM}}$ dimensional feature:

$$x_i^{\text{aug}} = [\psi_{\text{expr}}(x_i),\ \phi_{\text{PFM}}(h_i)] \in \mathbb{R}^{d_{\text{expr}}+d_{\text{PFM}}}. \tag{2}$$

These embeddings are incorporated through three integration strategies that define how the PFM representation enters the downstream model (Table 3). PFM Concatenation is the default strategy used across clustering, spot-level prediction, and sample-level aggregation: the PFM embedding is concatenated with the task-specific gene-expression PCA vector to form a joint input. PFM Graph Integration routes the same concatenated vector through a SpaGCN graph attention network, providing a contrast between simple concatenation and task-specific architectural adaptation. PFM Sample Aggregation extends the concatenation strategy to the sample level by aggregating PFM-augmented spot representations into sample-level representations.

Each strategy produces a task-specific augmented input from the common PFM embedding pipeline. PFM Concatenation replaces the expression-only representation with the joint vector for every measured training spot:

$$\mathcal{D}^{\text{tr}}_{\text{PFM,concat}} = \{[\psi_{\text{expr}}(x_i),\ \phi_{\text{PFM}}(h_i)] : z_i \in \mathcal{R}^{\text{tr}}\},$$

which forms the input to PCA-spatial Leiden clustering, logistic regression, XGBoost, and MLP classifiers. PFM Graph Integration uses the same joint vector as the node feature in the SpaGCN graph attention network while preserving the spatial neighbor graph E constructed from the

original coordinates $\{p_i\}$:

$$V = \{[\psi_{\text{expr}}(x_i),\ \phi_{\text{PFM}}(h_i)] : z_i \in \mathcal{R}^{\text{tr}}\}, \qquad \mathcal{I}_{\text{SpaGCN}} = (V, E).$$

In this adapted SpaGCN, the original model's built-in histology image feature extractor is replaced with frozen PFM embeddings: the PFM embedding $\phi_{\text{PFM}}(h_i)$ is concatenated with $\psi_{\text{expr}}(x_i)$ and used as the node feature in the graph attention network, followed by deep embedded clustering. This provides a contrast between simple PCA concatenation and graph-aware architectural integration of the same PFM representations, and serves as a bridge between feature-level augmentation and the graph-based downstream model used in the synthetic-observation clustering pipeline.

PFM Sample Aggregation extends concatenation to the sample level by aggregating PFM-augmented spot features within the spot set $\mathcal{B}_s$ of sample s:

$$b_s^{\text{PFM}} = \text{Agg}\left(\{(\psi_{\text{expr}}(x_i),\ \phi_{\text{PFM}}(h_i),\ p_i) : z_i \in \mathcal{B}_s\}\right),$$

where Agg computes PCA-score moments and quantiles together with coordinate summaries (Section 3.4). The sample-level classifier is then trained on $\{b_s^{\text{PFM}}\}_{s\in\mathcal{T}}$.

This setup follows the feature-level augmentation approach of spEMO (Liu et al., 2025), covering both zero-shot PCA concatenation and graph-aware downstream integration (SpaGCN). To enable a direct comparison with synthetic-observation augmentation under matched protocols, we evaluate four PFM sources (GPFM, UNI, GigaPath, UNI2-h) across four downstream tasks (spatial clustering, within-slice low-label prediction, cross-slice prediction, and sample-level disease prediction).

This comparison asks whether image-derived features from diverse PFMs can substitute for or complement synthetic-observation augmentation. If these zero-shot embeddings provide gains comparable to or larger than those from adding synthetic spots, it would indicate that feature enrichment is a more direct augmentation mechanism for those settings. Conversely, if the PFM embeddings are unhelpful or harmful in the same settings where synthetic observations help, it would reinforce the distinct roles of the two augmentation mechanisms.

3.6 Downstream execution and metrics

All comparisons start from a matched real-only baseline built from gene expression alone. Expression profiles are normalized, projected into a PCA space (50 components for DLPFC spot-level tasks; 30 components for sample-level disease prediction), and processed without using held-out spots, held-out sections, or held-out samples for fitting. Baseline-relative deltas are always computed against the real-only baseline that matches the same dataset, split, downstream backend, and evaluation protocol.

For the feature-level route, PFM embeddings are attached to measured spots before downstream modeling according to the task-specific integration strategy (Table 3). In spatial clustering, the two clustering backends (PCA-spatial Leiden with PFM Concatenation, and SpaGCN with PFM Graph Integration) use different architectures and are therefore reported with appropriate caveats rather than collapsed into a single route-level comparison.

For the synthetic-observation route, generated spots are incorporated according to the task-compatible strategies in Section 3.4. In spatial clustering, real and synthetic spots jointly define the expression–spatial neighbor graph, and Leiden clustering (Traag et al., 2019) is applied with seed 42; the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) are computed only on real spots after synthetic observations are removed. In within-slice and cross-slice prediction, synthetic observations may be appended, used alone as a diagnostic, sampled locally, or weighted in the supervised loss. Cross-slice augmented training sets are assembled from training sections only. In sample-level disease prediction, training samples may use real, synthetic, augmented, or weighted sample inputs, whereas test samples are always represented by real spots only.

Spatial clustering is evaluated with ARI and NMI as primary metrics, with Adjusted Mutual Information (AMI), V-measure, homogeneity, completeness, Fowlkes–Mallows, mapped cluster metrics, and spatial silhouette used as auxiliary diagnostics. Supervised spot prediction uses macro F1 as the primary metric and reports accuracy, balanced accuracy, weighted F1, the Matthews Correlation Coefficient (MCC), per-label metrics, and confusion summaries. Sample-level disease prediction uses macro F1 as the primary metric and reports the Area Under the Receiver Operating Characteristic curve (AUROC), the Area Under the Precision–Recall Curve (AUPRC), balanced accuracy, MCC, sensitivity, specificity, and fold-level confusion summaries. Uncertainty is summarized from the natural evaluation units of each task, including sections, section-seed pairs, split-test-section pairs, or cross-validation folds; bootstrap intervals are reported when computed.

4 Results

4.1 Experimental overview

Feature-level and synthetic-observation augmentation operate on different statistical objects and are therefore compared under matched downstream evaluation settings rather than being summarized by a single route-level average. Figure 3 provides an overview of the overall pattern before the task-specific analyses below. Feature-level augmentation performs best in supervised prediction tasks, especially low-label and sample-level prediction, whereas synthetic-observation augmentation provides selective gains in spatial clustering and sample-level prediction. Neither augmentation route consistently improves cross-slice generalization. Table 1 summarizes the route-specific configurations used for the four downstream tasks. The table is a configuration map rather than a performance ranking. For spatial clustering, PFM concatenation and synthetic-observation augmentation share the PCA-spatial Leiden real-only baseline, whereas SpaGCN-based PFM integration uses a graph attention architecture with a different clustering backend and is therefore reported as an absolute reference rather than a direct route-level result.

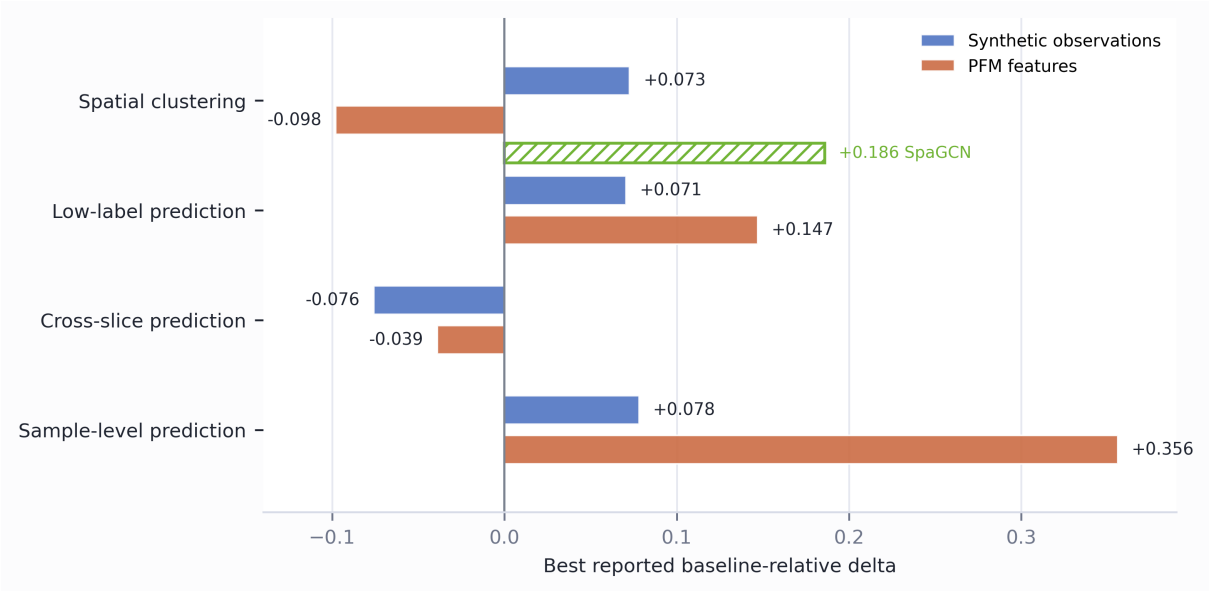


Figure 3: Dual-route comparison across four downstream tasks. Bars show baseline-relative deltas for synthetic-observation augmentation and PFM feature-level augmentation, using ΔARI for spatial clustering and Δmacro F1 for prediction tasks. Synthetic-observation bars summarize task-compatible configurations from the synthetic-observation benchmark; PFM bars summarize PCA+PFM feature-level results from the matched PFM tables. The hatched SpaGCN bar is shown separately as a graph-aware backend reference.

The synthetic-observation results are not interpreted as a generator ranking. Figure 4 provides a classifier-resolved overview of generator–strategy–parameter configurations, while feature-level results are reported within each task using their matched baselines. Figure 5 provides the corresponding overview for the feature-level (PFM) augmentation route. In Figure 4, rows enumerate the real baseline and each generator–strategy combination, and task-specific blocks indicate whether each configuration improves over the matched real-only baseline. The central comparison is therefore configuration-specific: spatial clustering has selected global and local graph-building gains, low-label prediction concentrates gains in weighted supervised settings, cross-slice prediction is mostly non-beneficial, and sample-level disease prediction depends on dataset and classifier headroom. In Figure 5, rows enumerate the real-only baseline and the PFM configurations, and task-specific blocks indicate whether each PFM source improves over the matched baseline. The feature-level pattern mirrors the dual-route overview: PFM embeddings provide consistent gains in low-label prediction and sample-level disease prediction when the baseline has measurable headroom, whereas simple PCA-PFM concatenation has limited benefit for spatial clustering and cross-slice generalization remains negative across all PFM sources.

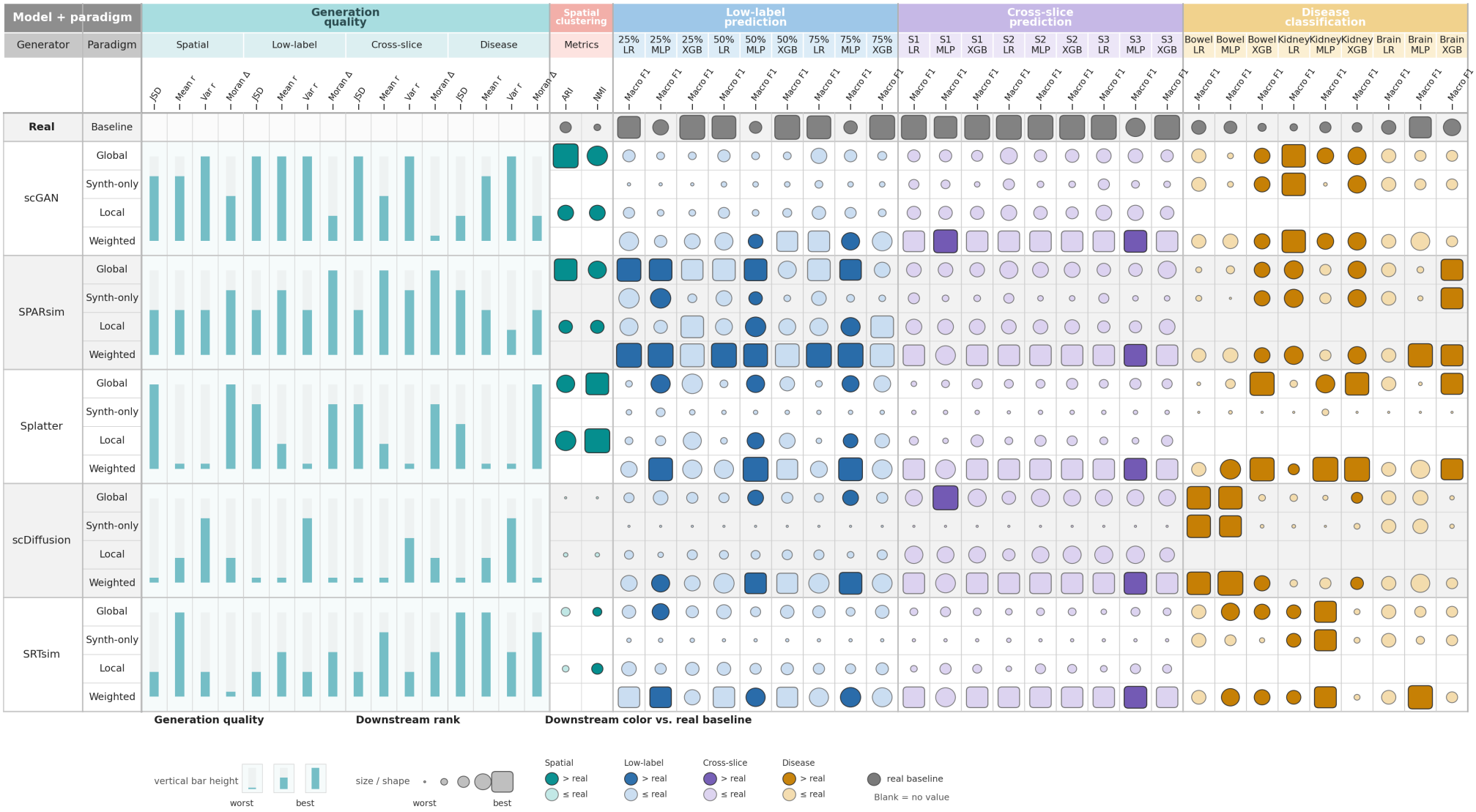


Figure 4: Classifier-resolved overview of synthetic-observation augmentation. Rows correspond to the real-only baseline and generator–strategy combinations for five synthetic generators. Task-specific blocks show downstream performance relative to the matched real-only baseline: spatial clustering by clustering metrics, low-label prediction by labeled fraction and classifier, cross-slice prediction by split seed and classifier, and disease classification by dataset and classifier. Marker color indicates whether the variant exceeds the matched baseline, marker size or shape indicates downstream rank, and blank cells indicate unavailable or inapplicable configurations. Selected spatial-clustering, low-label, and sample-level configurations exceed the matched baseline, whereas cross-slice prediction remains broadly unfavorable.

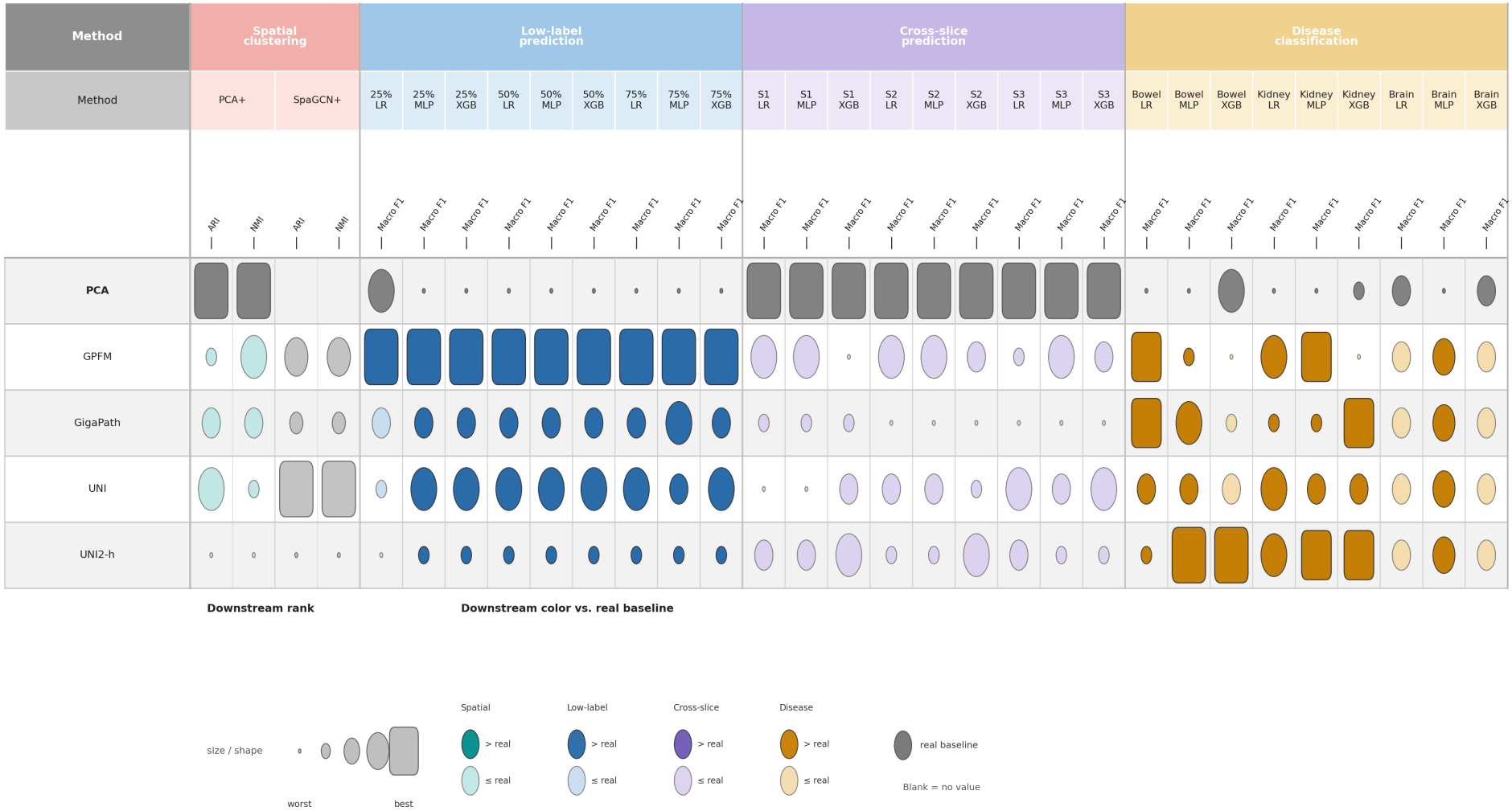


Figure 5: Classifier-resolved overview of feature-level (PFM) augmentation. Rows correspond to the real-only baseline and feature-level configurations built from four PFM embedding sources. Task-specific blocks show downstream performance relative to the matched real-only baseline: spatial clustering by clustering metrics, low-label prediction by labeled fraction and classifier, cross-slice prediction by split seed and classifier, and disease classification by dataset and classifier. Marker color indicates whether the variant exceeds the matched baseline, marker size or shape indicates downstream rank, and blank cells indicate unavailable or inapplicable configurations.

4.2 Evaluation across downstream tasks

4.2.1 Spatial clustering

This section aims to answer the question: Can observation augmentation improve graph-based clustering by providing additional spatial support? DLPFC spatial clustering uses PCA-spatial Leiden as the shared backend for synthetic-observation augmentation and simple PFM concatenation, with SpaGCN-based PFM integration reported separately as a graph-aware architectural reference.

Synthetic-observation augmentation. Augmented real and synthetic spots jointly define the expression–spatial graph, but ARI and NMI are computed only on real spots and are compared with the PCA real-only baseline. Representative effective and failed configurations are shown in Table 4.

Table 4: Representative configurations for DLPFC spatial clustering. Values are mean metrics across 12 sections under PCA-spatial Leiden.

Generator	Strategy	Parameter	Baseline ARI	Mean ARI	ΔARI	Baseline NMI	Mean NMI	ΔNMI
scGAN	Global	ratio 10	0.194	0.267	0.073	0.289	0.376	0.087
SPARsim	Global	ratio 18	0.194	0.248	0.054	0.289	0.352	0.063
Splatter	Local	higher local cap	0.194	0.240	0.046	0.289	0.350	0.062
Splatter	Local	conservative local cap	0.194	0.235	0.041	0.289	0.389	0.100
SRTsim	Global	ratio 1	0.194	0.170	-0.025	0.289	0.296	0.007
scDiffusion	Local	higher local cap	0.194	0.164	-0.030	0.289	0.241	-0.048

The configurations in Table 4 separate several distinct behaviors. scGAN and SPARsim improve both ARI and NMI under Global Real+Synthetic, indicating that additional generated spots can strengthen the expression–spatial graph when the synthetic pool expands layer-consistent neighborhoods. Splatter is more informative under Local Real+Synthetic than under a single global reading: both local-cap settings improve ARI, and the conservative cap yields the largest NMI gain, suggesting that bounded local enrichment can sharpen graph structure without overwhelming real spots. SRTsim provides a mixed case, with slightly improved NMI but reduced ARI, so the added observations may preserve some neighborhood agreement while weakening exact label alignment. scDiffusion is the clearest negative example in this backend, decreasing both ARI and NMI under local sampling. Spatial clustering therefore illustrates the central spAug claim that utility belongs to a complete generator–strategy–parameter–backend configuration, not to a generator alone.

Synthetic spots affect clustering only through the joint expression–spatial graph. Moderate ratios can provide additional support around real tissue layers, whereas poorly aligned synthetic support can create artificial bridges or distort local neighborhoods. Figure 6 shows that synthetic-observation augmentation consistently improves clustering for several generators, suggesting that additional observations can reshape neighborhood graphs beyond what is achievable from the original spots alone. A plausible explanation is that clustering relies heavily on local neighborhood structure. Additional observations can densify the underlying graph, leading to more stable graph construction and improved community detection.

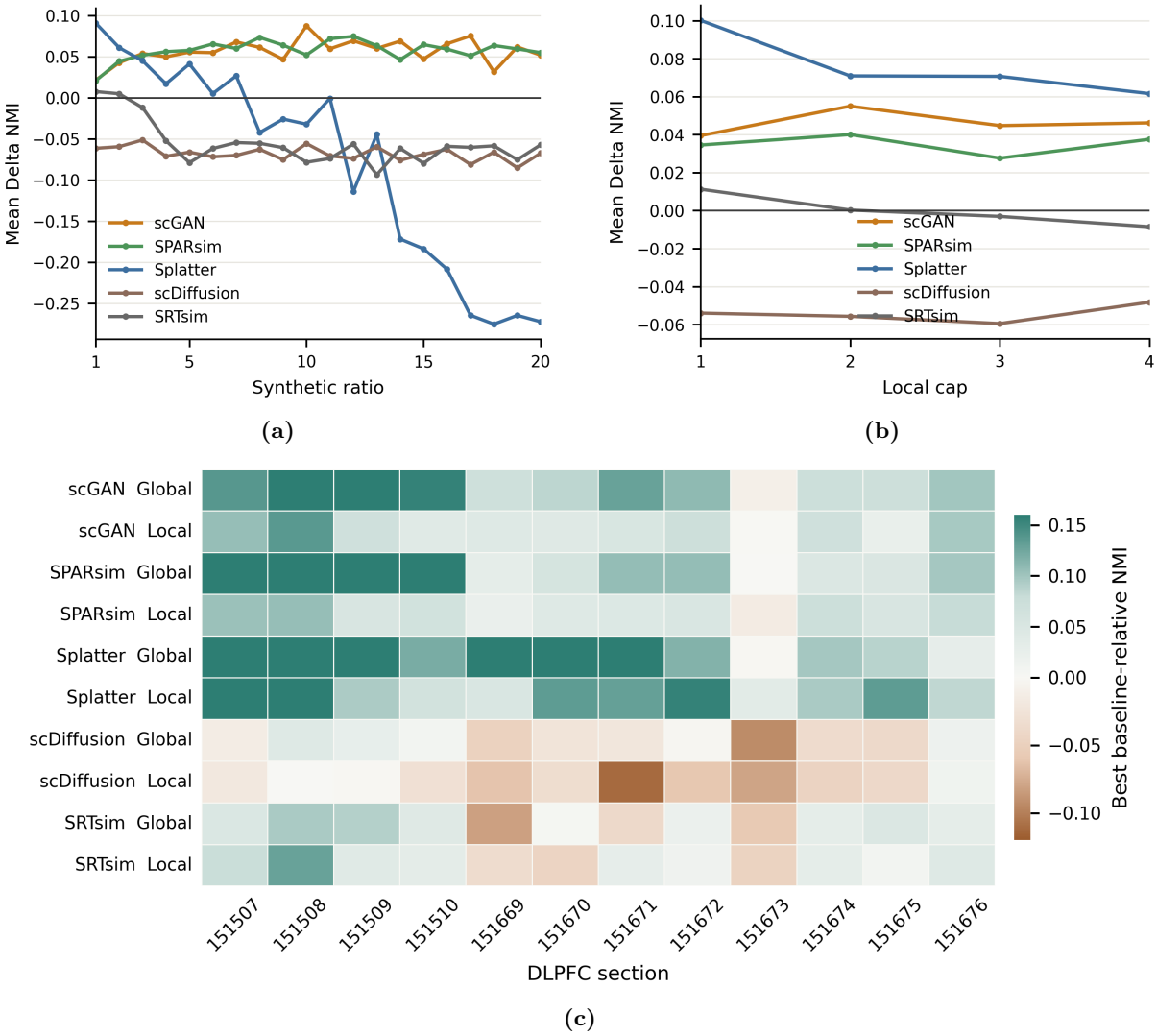
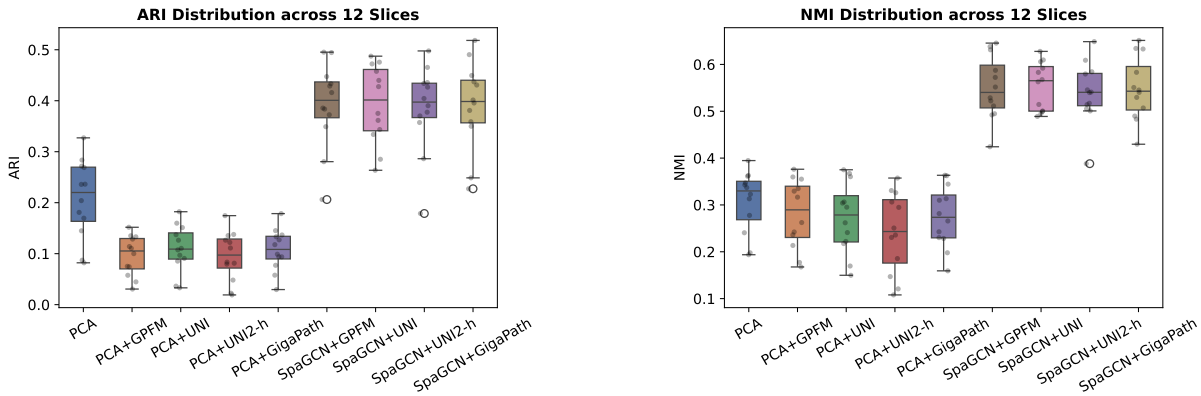


Figure 6: Effect of synthetic-observation augmentation on spatial clustering in DLPFC. **(a)** Mean ΔNMI across 12 sections as the Global Real+Synthetic ratio increases. **(b)** Mean ΔNMI under Local Real+Synthetic as the local cap increases. **(c)** Section-level best ΔNMI for Global Real+Synthetic and Local Real+Synthetic. Rows correspond to generator–strategy pairs and columns to DLPFC sections; color is the best baseline-relative NMI gain observed for that section.

Feature-level augmentation. The same four PFM embeddings were evaluated under two integration strategies (Table 3). Under the shared PCA-spatial Leiden backend, simple PFM concatenation consistently degraded clustering performance relative to the PCA baseline. This pattern mirrors the synthetic-observation results within the same backend: when added information is not compatible with the spatial graph construction, both routes can harm clustering under PCA-spatial Leiden.

As an architectural alternative, the same PFM embeddings integrated through a modified SpaGCN graph attention network (Section 3.5) achieved substantially higher ARI values across all four PFMs, as shown in Figure 7. However, because SpaGCN uses deep embedded clustering rather than PCA-spatial Leiden, its absolute scores are not directly comparable as a route-level delta. The contrast between the two integration strategies demonstrates that PFM utility depends on how the spatial graph incorporates the added features: concatenation alone is in-



(a) **(b)**

Figure 7: PFM feature-level augmentation for DLPFC spatial clustering showing **(a)** ARI and **(b)** NMI across 12 sections for PCA concatenation and SpaGCN integration with each PFM embedding.

sufficient, whereas graph-aware integration can be effective. This task-dependence parallels the synthetic-observation finding that generator utility depends on strategy and parameter choice rather than on generator identity alone.

In spatial clustering, augmentation utility is routed through the spatial graph. Synthetic observations that reinforce coherent neighborhood structure can improve clustering, whereas naive feature concatenation degrades the PCA-spatial Leiden backend. Graph-aware integration of PFM features through SpaGCN can circumvent this limitation within its own architecture, confirming that compatibility with graph construction is the key determinant.

4.2.2 Within-slice low-label prediction

Can augmentation compensate for limited supervision within a tissue section? Within-slice low-label prediction evaluates both augmentation routes under matched classifiers and label fractions.

Synthetic-observation augmentation. Weighted ERM with an MLP improves macro F1 at 25%, 50%, and 75% labeled fractions, whereas Synthetic-only training and XGBoost do not outperform the matched real-only baseline. The best-performing configurations together with their matched baselines are summarized in Table 5.

For MLP, the gain decreases from +0.071 at 25% labels to +0.045 at 75% labels, indicating that synthetic observations provide the greatest marginal benefit when labeled data are scarce. LR exhibits smaller gains, whereas XGBoost remains below the matched baseline even under its best-performing configuration. The failure of Synthetic-only training is similarly informative: generated observations cannot replace real labeled spots but can improve low-label prediction when incorporated through controlled risk weighting.

This pattern is consistent with the role of weighted ERM. Small-to-moderate values of α keep the empirical risk anchored to real labeled observations while allowing synthetic observations to regularize sparse decision regions. As α increases, the objective becomes increasingly influenced by the synthetic distribution, leading to diminished or negative gains. Figure 8 summarizes the responses across labeled fractions, synthetic-to-real ratios, and weighting parameters.

Table 5: Best baseline-improving configurations in within-slice low-label prediction. Values report the best non-baseline mean macro F1 delta for each labeled fraction and classifier over the evaluated generator–strategy–parameter grid, together with the matched real-only baseline.

Fraction	Classifier	Best non-baseline result	Real-only macro F1	Best macro F1	Δ macro F1
25%	LR	SPARsim Weighted ERM, ratio 30, $\alpha = 0.4$	0.611	0.627	0.015
	MLP	SPARsim Weighted ERM, ratio 17, $\alpha = 0.4$	0.583	0.654	0.071
	XGBoost	Best non-baseline remains negative	0.645	0.630	-0.015
50%	LR	SPARsim Weighted ERM, ratio 30, $\alpha = 0.4$	0.654	0.662	0.008
	MLP	Splatter Weighted ERM, ratio 20, $\alpha = 0.4$	0.625	0.682	0.057
	XGBoost	Best non-baseline remains negative	0.696	0.661	-0.035
75%	LR	SPARsim Weighted ERM, ratio 29, $\alpha = 0.2$	0.669	0.679	0.010
	MLP	SPARsim Weighted ERM, ratio 9, $\alpha = 0.4$	0.648	0.693	0.045
	XGBoost	Best non-baseline remains negative	0.729	0.678	-0.051

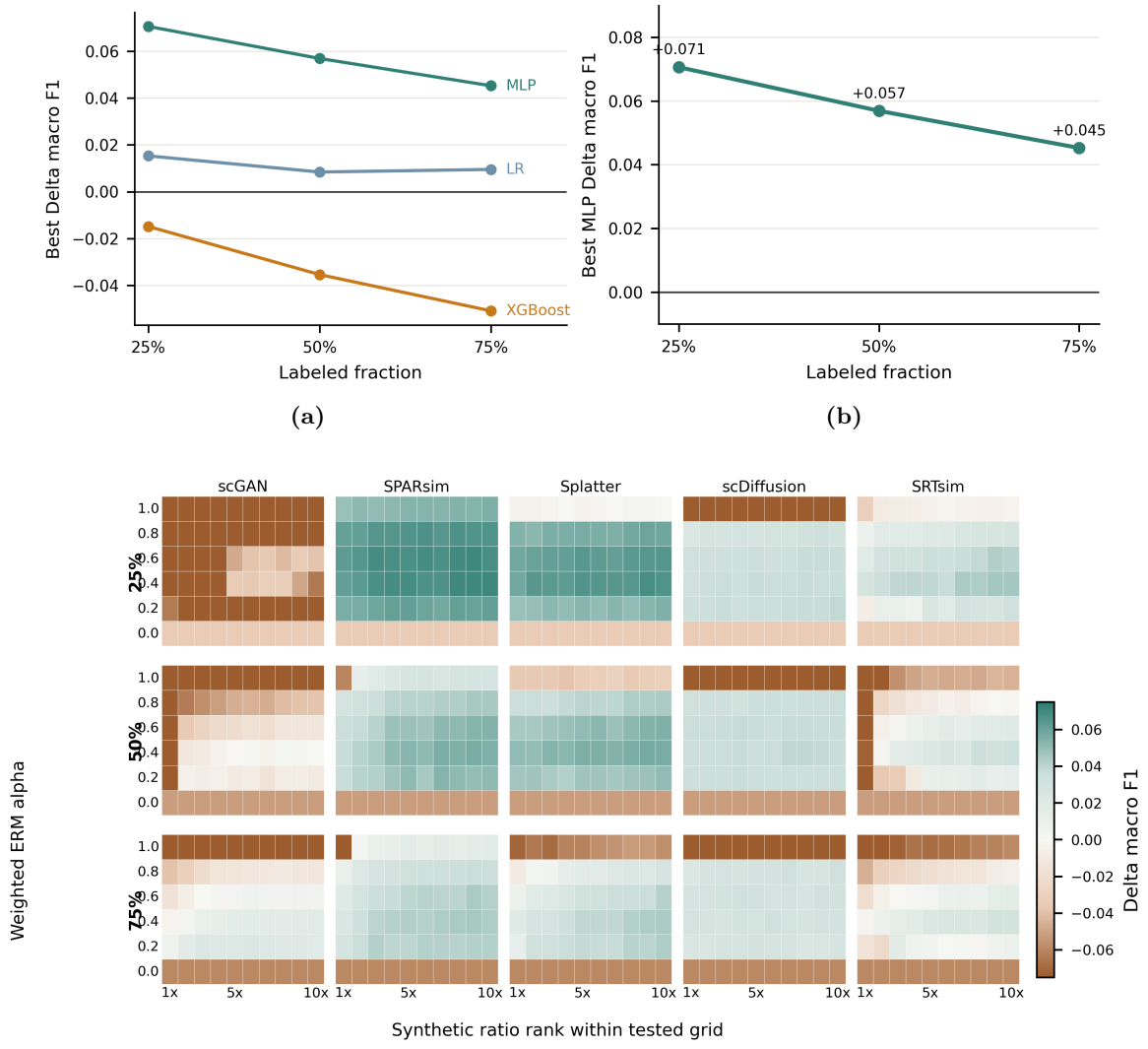
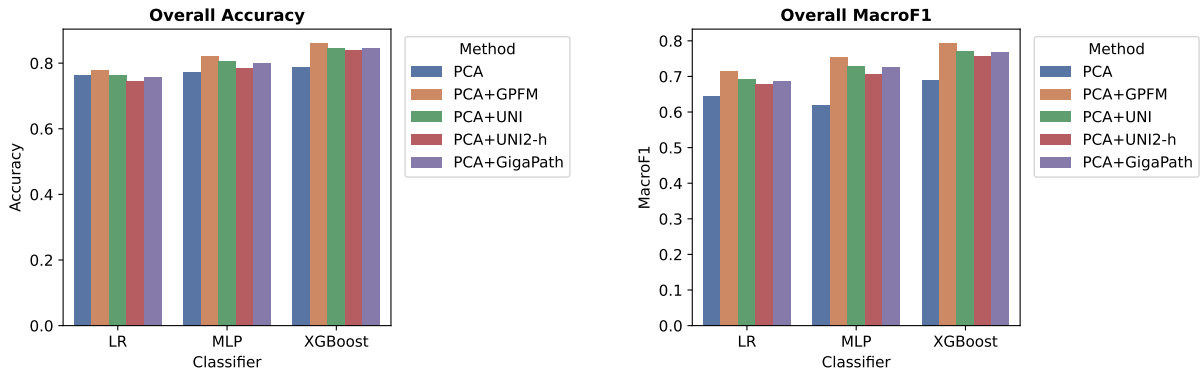


Figure 8: Effect of synthetic-observation augmentation on within-slice low-label prediction. **(a)** Best baseline-relative macro F1 gain for each classifier at 25%, 50%, and 75% labeled fractions. **(b)** Best MLP gain decreases as the real labeled fraction increases. **(c)** MLP weighted-ERM response over α and synthetic ratio rank for each generator and labeled fraction. Each small heatmap uses the ranked ratio settings tested for that generator–fraction combination, ordered from the lowest to highest tested ratio ranks; color encodes Δmacro F1 relative to the matched real-only baseline.

Table 6: PFM feature-level augmentation for within-slice low-label prediction. Values report Δmacro F1 relative to the PCA baseline for the same classifier and label fraction.

Label fraction	Classifier	PCA baseline	ΔGPFM	ΔUNI	ΔUNI2-h	ΔGigaPath
25%	LR	0.611	+0.013	−0.008	−0.024	−0.008
25%	MLP	0.583	+0.116	+0.088	+0.046	+0.074
25%	XGB	0.645	+0.099	+0.076	+0.063	+0.074
50%	LR	0.654	+0.088	+0.064	+0.044	+0.051
50%	MLP	0.625	+0.147	+0.117	+0.101	+0.115
50%	XGB	0.696	+0.107	+0.087	+0.073	+0.081
75%	LR	0.670	+0.112	+0.090	+0.078	+0.086
75%	MLP	0.648	+0.144	+0.128	+0.120	+0.134
75%	XGB	0.729	+0.103	+0.078	+0.069	+0.076

Feature-level augmentation. PFM embeddings provided large gains as feature-level augmentation across most label fractions and classifiers, with GPFM showing the largest improvements and UNI2-h the smallest. These improvements substantially exceed the best synthetic-observation gains (Figure 9, Table 6), confirming that for supervised within-slice tasks, image-derived features provide more direct augmentation than adding generated expression profiles when histology images are available.



(a) **(b)**

Figure 9: Effect of PFM feature-level augmentation on within-slice low-label prediction. Results are shown for all PFM embedding variants in terms of **(a)** Accuracy and **(b)** Macro F1.

Within-slice low-label prediction shows a clear division: PFM embeddings provide large and consistent gains across classifiers, substantially exceeding the improvements available from synthetic observations. When histology images are available, enriching existing spot features is more effective than adding generated expression profiles for supervised within-slice tasks.

4.2.3 Cross-slice generalization

Can augmentation overcome distribution shift across tissue sections? Cross-slice prediction evaluates both augmentation routes. It turns out that neither route addresses the fundamental challenge of held-out-section distribution shifts.

Table 7: Strategy-level failure in DLPFC cross-slice prediction. Values average over generators, classifiers, parameters, and the three split seeds.

Strategy	Mean Δ macro F1	Median Δ macro F1	Positive-unit fraction
Global Real+Synthetic	-0.127	-0.118	0.003
Synthetic-only	-0.247	-0.210	0.000
Local Real+Synthetic	-0.096	-0.094	0.003
Weighted ERM	-0.076	-0.062	0.041

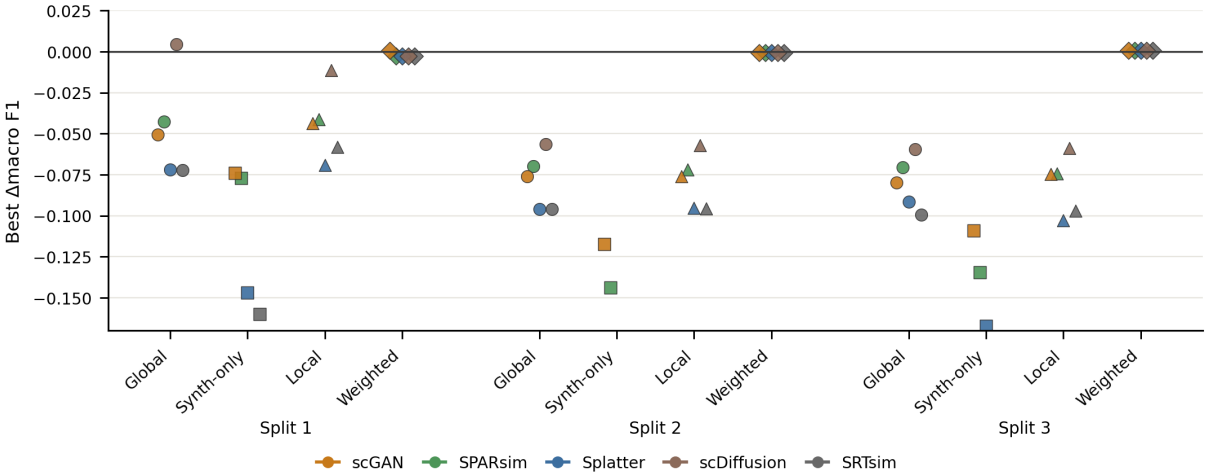


Figure 10: Best synthetic-observation variants in DLPFC cross-slice prediction. Points show the best Δmacro F1 for each generator within each split–strategy combination after optimizing over the available parameter grid and classifiers. Colors denote generators, marker shapes denote strategies, and the horizontal line marks the matched real-only baseline.

Synthetic-observation augmentation. Across three peer 8-train/4-test split seeds, no strategy consistently outperforms the matched baseline under mean delta, median delta, and split–test-section stability criteria. Global Real+Synthetic, Synthetic-only, Local Real+Synthetic, and Weighted ERM all have negative mean and median macro F1 deltas (Table 7).

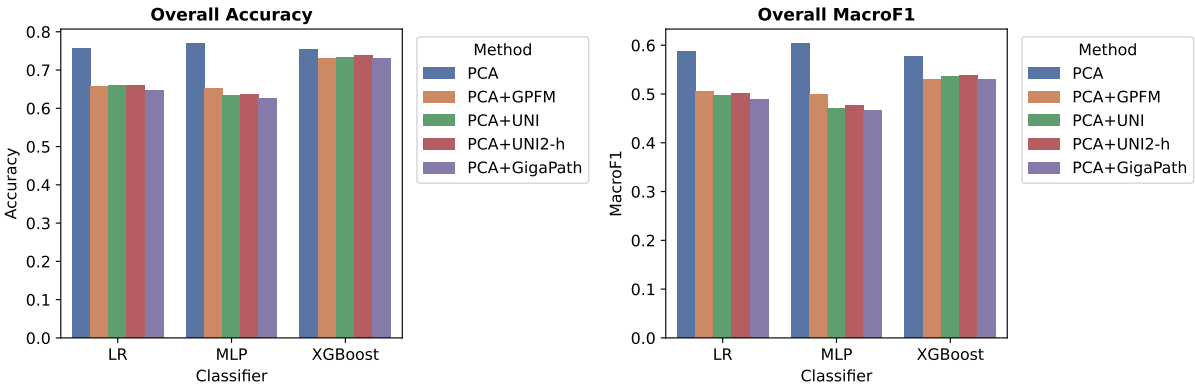
This result defines a boundary for observation augmentation. Synthetic spots expand source-section support but do not generate the unseen biological, technical, or annotation shifts present in held-out sections. The low positive-unit fractions, including 0.041 for Weighted ERM and near-zero values for other strategies, indicate that the negative conclusion is not driven by a few poor parameter choices.

The contrast with low-label prediction is instructive. Within-section training and testing allow a small synthetic-risk contribution to regularize sparse labels; cross-slice prediction instead requires target-section invariance. Figure 10 shows that even optimized configurations remain near or below the real-only baseline.

Feature-level augmentation. PFM image embeddings were consistently harmful for cross-slice classification: aggregate PFM-source × classifier mean Δmacro F1 values across three split seeds were all negative relative to the PCA baseline (Table 8). This failure mirrors the limited gains from synthetic-observation augmentation in the same setting (Figure 11), confirming that neither augmentation route addresses the fundamental challenge of cross-slice generalization in spatial transcriptomics.

Table 8: PFM feature-level augmentation for cross-slice spot prediction. Values report Δmacro F1 averaged over three peer 8-train/4-test split seeds. PCA baseline macro F1: LR=0.586, MLP=0.603, XGBoost=0.577. All values are negative across all 12 PFM–classifier combinations.

Classifier	GPFM	UNI	UNI2-h	GigaPath
LR	−0.082	−0.090	−0.085	−0.098
MLP	−0.104	−0.132	−0.127	−0.137
XGBoost	−0.046	−0.041	−0.039	−0.048



(a) **(b)**

Figure 11: Effect of PFM feature-level augmentation on cross-slice spot prediction. Results are shown for all PFM embedding variants in terms of **(a)** Accuracy and **(b)** Macro F1.

Cross-slice generalization remains unsolved for both routes. Neither PFM embeddings nor synthetic observations generated from training sections can reliably bridge the distribution shift present in held-out sections. This failure is not attributable to poor parameter choice, indicating that cross-slice distribution shift is a fundamental limitation shared by both augmentation routes.

4.2.4 Sample-level disease prediction

Can synthetic observations stabilize sample-level prediction in small cohorts? Sample-level disease prediction evaluates both augmentation routes, separating non-saturated and ceiling-limited regimes.

Synthetic-observation augmentation. Kidney and Bowel retain measurable headroom for improvement for at least some classifiers, whereas Brain is largely ceiling-limited, with a real-only LR baseline of 1.000 macro F1 and an XGBoost baseline of 0.952. Table 9 summarizes the best-performing configurations together with the matched real-only baselines and Synthetic-only diagnostics.

The two non-ceiling datasets exhibit distinct augmentation patterns. In Kidney, all three classifiers achieve positive gains, although through different strategies. LR performs best with a small Global Real+Synthetic scGAN augmentation, and the same gain is attainable under Synthetic-only training. By contrast, MLP and XGBoost benefit most from Weighted Aggregation with Splatter, suggesting that synthetic observations are most effective when their contribution is controlled during sample-level representation construction. Bowel shows a similar overall

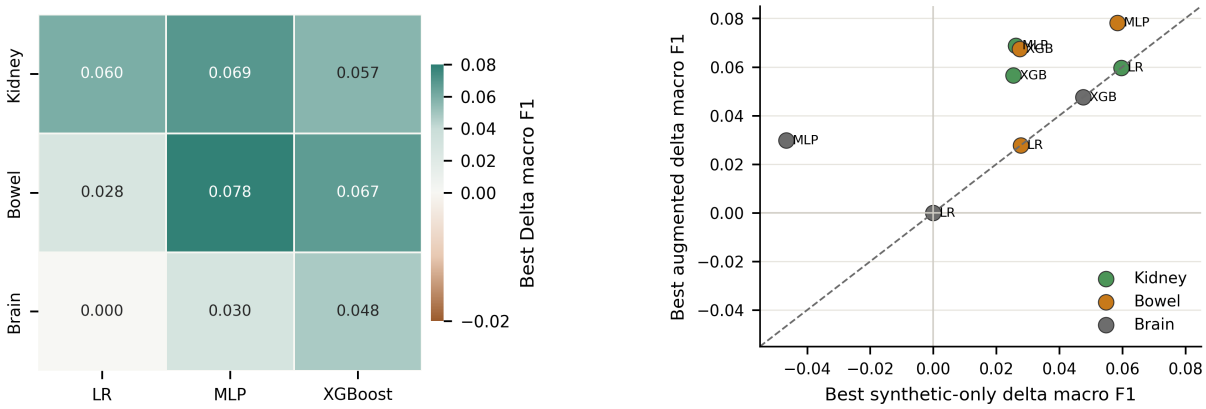
Table 9: Sample-level disease prediction under non-saturated and ceiling-limited regimes. Baselines are real-only macro F1 averaged over five folds.

Dataset	Classifier	Best configuration	Real-only macro F1	Best Δ macro F1	Best synthetic-only Δ
Kidney	LR	scGAN global, ratio 1	0.915	0.060	0.060
	MLP	Splatter weighted, ratio 7, $\alpha = 0.8$	0.644	0.069	0.026
	XGBoost	Splatter weighted, ratio 3, $\alpha = 0.4$	0.912	0.057	0.025
Bowel	LR	scDiffusion global, ratio 16	0.875	0.028	0.028
	MLP	scDiffusion weighted, ratio 15, $\alpha = 0.2$	0.673	0.078	0.058
	XGBoost	Splatter global, ratio 10	0.804	0.068	0.027
Brain	LR	Ceiling-limited	1.000	0.000	0.000
	MLP	SPARsim weighted, ratio 1, $\alpha = 0.2$	0.841	0.030	-0.047
	XGBoost	SPARsim global, ratio 1	0.952	0.048	0.048

trend but with different generators: scDiffusion provides the best LR and MLP configurations, whereas XGBoost performs best with Global Real+Synthetic and Splatter.

Brain should be interpreted differently. Because the LR baseline is already perfect, no further improvement is measurable. Positive gains for MLP and XGBoost simply reflect the remaining classifier headroom rather than a consistent dataset-level augmentation effect. Overall, the utility of synthetic observations depends jointly on baseline headroom, classifier capacity, and the strategy used to incorporate generated observations.

Mechanistically, the weighting parameter α influences sample representation rather than a spot-level loss. Small values of α preserve real observations as the primary contributors to sample-level summaries, whereas larger values increasingly allow generated observations to shape moments, quantiles, and coordinate statistics. Figure 12 illustrates this distinction by separating the baseline-improving cases in Kidney and Bowel from the ceiling-limited behavior observed in Brain.



(a) (b)

Figure 12: Synthetic-observation effects in sample-level disease prediction. **(a)** Best Δmacro F1 by dataset and classifier. **(b)** Best augmented Δmacro F1 versus best Synthetic-only Δmacro F1 for each dataset–classifier pair.

Feature-level augmentation. PFM embeddings were evaluated as feature-level augmentation for sample-level disease prediction on the same three datasets (Kidney, Bowel, and Brain). Consistent with the synthetic-observation route, the benefits were both dataset- and classifier-dependent.

Kidney exhibited the largest improvements, with all PFM variants raising MLP macro F1

from a moderate baseline to near-perfect performance and providing smaller but consistent gains for LR. Bowel showed more modest improvements, primarily for MLP, whereas LR and XGBoost exhibited positive or near-neutral effects depending on the PFM variant. Brain was largely ceiling-limited, with LR already achieving a macro F1 of 1.000 and XGBoost leaving little remaining headroom; only MLP retained measurable improvement potential. The nearly identical Δ values observed across PFMs for a given classifier (e.g., all MLP variants achieving +0.159 to reach macro F1 = 1.000, and all XGBoost variants yielding $\Delta = 0$) reflect the discrete nature of sample-level evaluation with a limited number of samples and cross-validation folds rather than genuine equivalence among PFMs. Figure 13 summarizes the overall Accuracy and macro F1 across datasets and PFM variants, with detailed classifier-specific results reported in Table 10.

Table 10: PFM feature-level augmentation for sample-level disease prediction. Values report Δmacro F1 over five cross-validation folds. PCA baselines are shown for context.

Dataset	Classifier	PCA baseline	ΔGPFM	ΔUNI	ΔUNI2-h	ΔGigaPath
Kidney	LR	0.915	+0.085	+0.085	+0.085	+0.054
	MLP	0.644	+0.356	+0.325	+0.356	+0.300
	XGB	0.912	−0.042	+0.031	+0.057	+0.057
Bowel	LR	0.875	+0.066	+0.032	+0.028	+0.066
	MLP	0.673	+0.171	+0.172	+0.202	+0.192
	XGB	0.804	−0.154	−0.009	+0.084	−0.071
Brain	LR	1.000	0.000	0.000	0.000	0.000
	MLP	0.841	+0.159	+0.159	+0.159	+0.159
	XGB	0.952	0.000	0.000	0.000	0.000

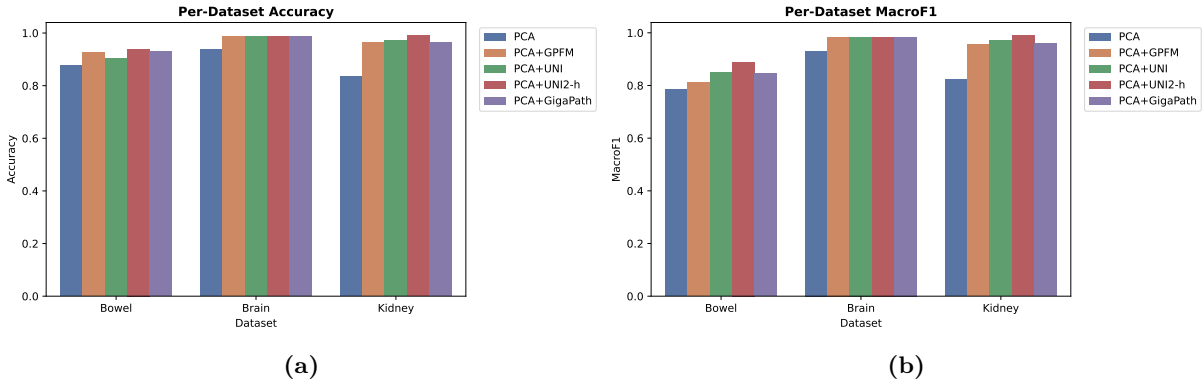


Figure 13: Effect of PFM feature-level augmentation on sample-level disease prediction measured by **(a)** Accuracy and **(b)** Macro F1. Dataset- and classifier-specific results are reported in Table 10.

Sample-level disease prediction further highlights the complementary roles of the two augmentation routes. PFM embeddings provide substantial gains when the baseline leaves sufficient headroom by enriching spot representations before sample-level aggregation, whereas synthetic observations offer more modest but consistent improvements through weighted sample aggregation. The effectiveness of both routes is strongly configuration dependent, varying across

datasets, classifiers, and baseline headroom.

4.3 Understanding augmentation mechanisms

4.3.1 Intrinsic quality and downstream utility

Intrinsic synthetic-data diagnostics are useful for quality control but cannot replace downstream evaluation. A synthetic pool must improve the graph, classifier, or sample representation used by a task, not merely resemble real spots by marginal expression metrics. For clustering contexts, ARI and NMI are both primary graph-recovery metrics; Figure 14 uses ΔARI for consistency with the graph-coherence analyses, whereas the appendix stability summary reports ΔNMI. Figure 14 summarizes this fidelity-to-utility gap.

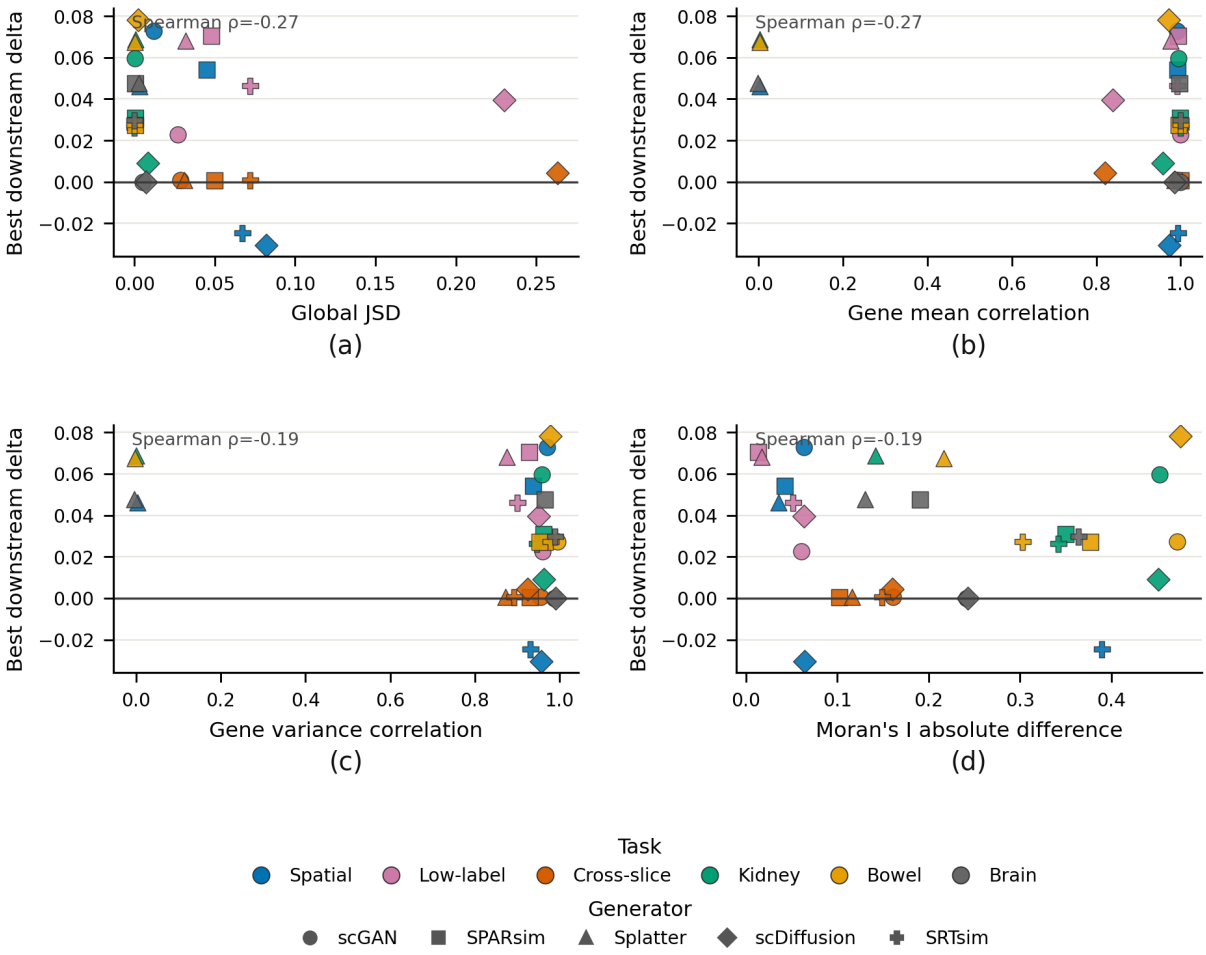


Figure 14: Relationship between intrinsic synthetic-data diagnostics and downstream utility. Each point represents a generator within a task context; color denotes task or dataset and marker shape denotes generator. The y-axis reports best downstream ΔARI for spatial clustering or best Δmacro F1 for prediction tasks. **(a)** Global Jensen-Shannon divergence (JSD) versus best downstream delta. **(b)** Gene mean correlation versus best downstream delta. **(c)** Gene variance correlation versus best downstream delta. **(d)** Moran's I absolute difference versus best downstream delta.

Thus, intrinsic quality and downstream utility are related but non-interchangeable axes. A pool that looks plausible can still harm a specific graph or classifier, while a pool that is not best by intrinsic similarity may become useful under a compatible strategy.

4.3.2 Synthetic-observation augmentation

The synthetic-observation results above show that utility is task-dependent. We therefore performed additional mechanism-oriented analyses to identify which conditions make generated observations helpful or harmful. These analyses are restricted to the synthetic-observation route and use matched baseline-relative deltas from the same evaluation units as the main results.

First, synthetic observations were most useful when the matched real-only baseline left measurable headroom. Across DLPFC spatial sections, low-label configurations, and sample-level dataset–classifier pairs, larger baseline headroom was associated with larger best observed improvement, although headroom alone was not sufficient when the downstream classifier or task was incompatible with the synthetic signal (Figure 15). This pattern explains why the largest low-label gains occur at 25% labels under MLP, why Brain LR has no measurable sample-level gain, and why XGBoost can remain negative despite nonzero headroom.

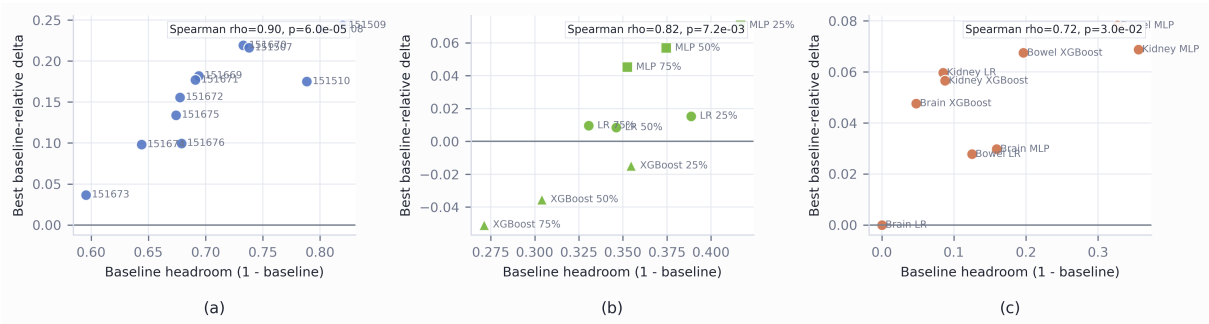


Figure 15: Baseline headroom and synthetic-observation utility. Each point is the best non-baseline synthetic-observation configuration within a matched evaluation unit. **(a)** Spatial clustering uses DLPFC section-level baseline NMI and best ΔNMI. **(b)** Low-label prediction uses labeled-fraction–classifier baseline macro F1 and best Δmacro F1. **(c)** Sample-level prediction uses dataset–classifier baseline macro F1 and best Δmacro F1.

Second, the spatial-clustering gains are consistent with a graph-coherence mechanism rather than a simple synthetic-volume effect. Across generator–strategy–parameter–section configurations, ΔARI increased with improvement in real-spot spatial silhouette, while high observed synthetic ratios and large cluster-count mismatches were associated with worse ARI (Figure 16). Thus, useful synthetic observations appear to improve the expression–spatial graph when they reinforce coherent neighborhoods; excessive or poorly aligned synthetic support can instead fragment the graph or distort the number of recovered domains.

Section-level heterogeneity further supports this interpretation. After selecting the best non-baseline synthetic-observation configuration for each of the 12 DLPFC sections, sections with weaker real-only structure showed larger gains: baseline NMI was strongly negatively correlated with best ΔNMI (Spearman $\rho = -0.90$, $p = 6.0 \times 10^{-5}$), and baseline spatial silhouette showed a weaker but still significant negative association with best ΔNMI ($\rho = -0.58$, $p = 0.048$). Across sections, best ΔNMI ranged from 0.037 to 0.243 with a median of 0.176. Synthetic observations therefore appear to help most when the original graph is under-supported rather than when the real-only graph is already well organized.

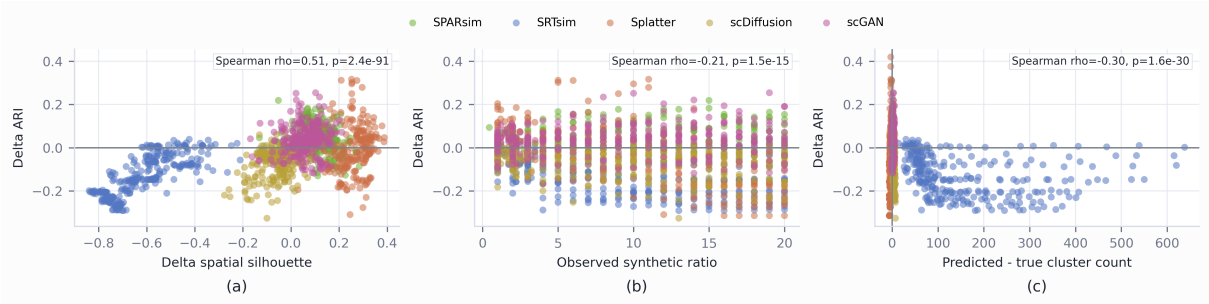


Figure 16: Graph-level constraints on spatial-clustering utility. Each point represents one synthetic-observation configuration for a DLPFC section under PCA-spatial Leiden clustering. **(a)** ΔARI versus the real-spot spatial silhouette. **(b)** ΔARI versus the observed synthetic-to-real ratio. **(c)** ΔARI versus the mismatch between the predicted and true numbers of clusters.

Third, for cross-slice prediction, the broad failure remained even after a permissive post hoc optimization (Figure 17). We selected the best non-baseline synthetic-observation variant separately within each split, held-out section, and classifier, yielding 36 optimized split–section–classifier cases. The resulting Δmacro F1 values remained centered near zero (median −0.0022, mean −0.0006): only 11 of 36 cases were positive, only 2 of 36 exceeded 0.01, and 9 of 36 were at or below −0.005. Positive cases were also not stable across classifiers: LR and XGBoost each improved in only 2 of 12 cases, whereas MLP improved in 7 of 12 cases and accounted for the only two gains above 0.01. This rules out the simple explanation that cross-slice failure is caused only by an unfavorable global parameter choice. Instead, synthetic observations generated from training sections do not supply the missing target-section invariance required for held-out-section transfer.

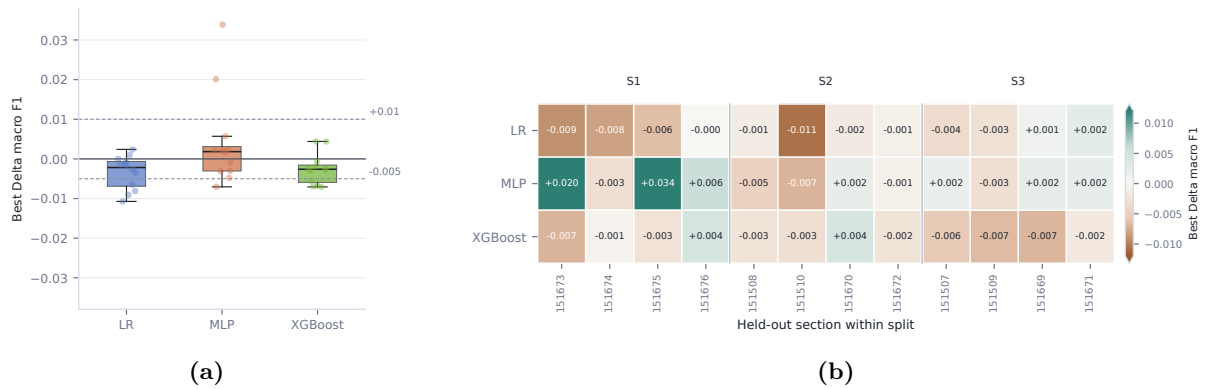


Figure 17: Post-hoc analysis of synthetic-observation augmentation for cross-slice prediction. **(a)** Best Δmacro F1 after selecting the best non-baseline synthetic-observation configuration within each split–held-out-section–classifier case. **(b)** Heatmap of the 36 optimized cases, grouped by split and held-out section.

Together, these analyses refine the interpretation of synthetic-observation augmentation. Effective configurations tend to occur when the baseline has headroom, when generated observations improve rather than distort the graph or decision boundary, and when the downstream task requires additional empirical support rather than target-domain invariance. Ineffective configurations arise when synthetic volume becomes excessive, when cluster structure is distorted, when the classifier cannot use the synthetic signal, or when the missing information is a

held-out-section shift that the training-section synthetic pool cannot cover.

4.3.3 Feature-level augmentation

The PFM results above show that feature-level augmentation utility is also configuration-specific. We therefore performed additional analyses to identify which PFM and task properties determine whether PFM embeddings help or harm downstream performance. These analyses use the same evaluation units and matched baselines as the main feature-level results.

The PFM ranking varied by task (Figure 18). In low-label prediction, GPFM was the best performer, whereas in sample-level prediction UNI2-h showed the largest gains and in cross-slice prediction GPFM showed the least negative values (though all were negative). In spatial clustering concatenation all PFMs degraded similarly. The same evaluation scaffold accommodates both augmentation routes under matched protocols.

First, the magnitude of PFM gains was strongly classifier-dependent (Figure 18). Averaged across the four PFMs, MLP showed the largest improvements in low-label prediction and sample-level tasks, whereas LR and XGBoost showed smaller or negative gains. In cross-slice prediction, all three classifiers exhibited uniformly negative Δ values, with MLP showing the largest degradation. This classifier dependency parallels the synthetic-observation finding that MLP benefits most from additional feature information, while LR and XGBoost are more sensitive to feature-space distortion. The across-task consistency of the MLP advantage — positive in low-label and sample-level settings, negative in cross-slice — indicates that classifier architecture strongly influences whether PFM features can be exploited, but even MLP cannot overcome distribution shifts that neither route addresses.

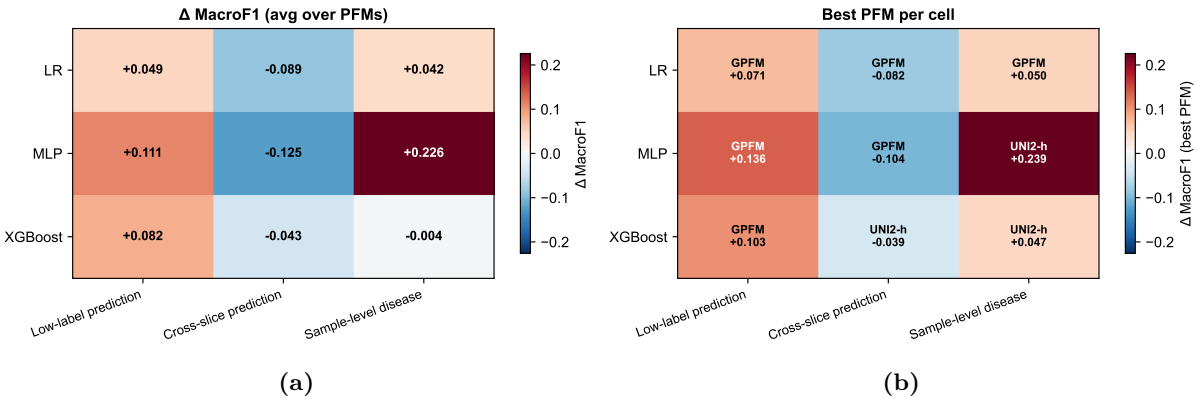


Figure 18: PFM gain by classifier and task: **(a)** Mean Δmacro F1 across four PFMs; **(b)** Best-performing PFM per cell with its Δ value.

Second, to examine whether PFM gains depend on the quality of the PCA baseline, we compared baseline metric values against PFM Δ across all four tasks (Figure 19). The relationship was task-dependent and generally opposite to the synthetic-observation pattern. In spatial clustering under PCA concatenation, PFM ΔARI was negatively associated with baseline ARI (Spearman $\rho = -0.79$), indicating that sections with stronger PCA-based structure suffered greater degradation from naive PFM concatenation — in contrast to synthetic observations, where weaker sections benefited most from added synthetic support. In low-label prediction (MLP), larger baseline headroom (lower baseline macro F1) was associated with larger PFM

gains ($\rho = -0.60$), mirroring but weaker than the synthetic-observation headroom analysis (Figure 15). In sample-level prediction (Kidney, MLP), the negative trend was strong ($\rho = -0.95$), likely reflecting ceiling effects in folds with already-perfect baselines where no room for improvement remained. In cross-slice prediction, no headroom trend was present ($\rho = -0.04$), consistent with the uniformly negative PFM results. Thus, PFM headroom dependence is clearest when the baseline has measurable variance (clustering, sample-level), but the direction depends on whether PFM integration helps (low-label) or harms (clustering concatenation) the task.

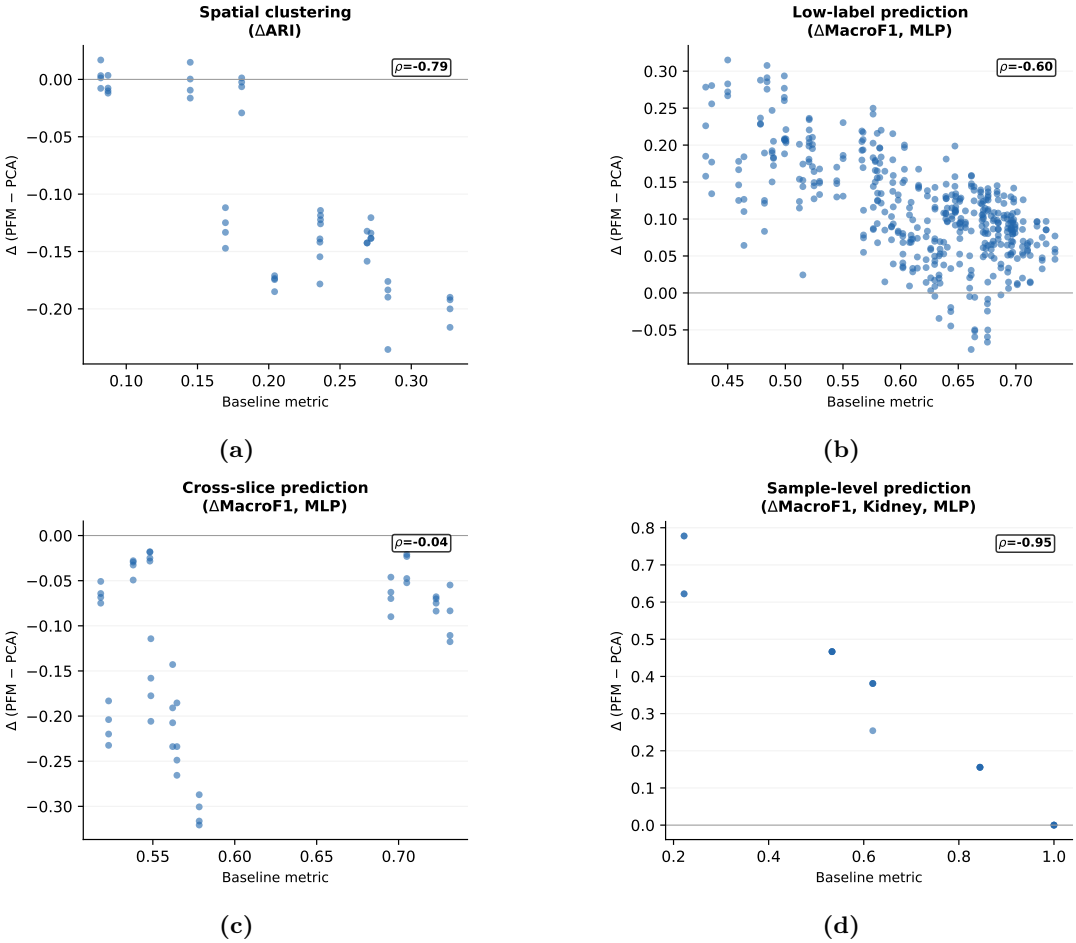
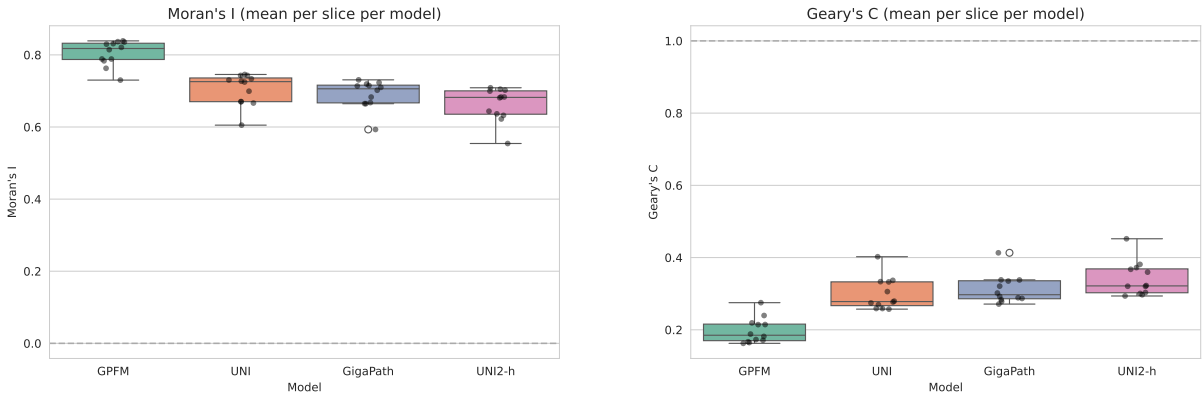


Figure 19: Baseline headroom and PFM utility across tasks: **(a)** Spatial clustering, **(b)** Low-label prediction (MLP), **(c)** Cross-slice prediction (MLP), **(d)** Sample-level prediction (Kidney, MLP).

Third, to investigate the mechanism underlying the task-dependent PFM effects, we computed the spatial autocorrelation of each PFM embedding across the 12 DLPFC sections using Moran's I and Geary's C (Figure 20). All four PFMs exhibited high Moran's I values, indicating strong spatial structure: GPFM (0.805 ± 0.034) > UNI (0.705 ± 0.043) ≈ GigaPath (0.691 ± 0.039) > UNI2-h (0.663 ± 0.046). The corresponding Geary's C values followed the inverse trend (GPFM 0.198 ± 0.035 < UNI 0.299 ± 0.045 ≈ GigaPath 0.312 ± 0.040 < UNI2-h 0.341 ± 0.047), confirming the Moran's I ranking. This autocorrelation ranking (GPFM > UNI ≈ GigaPath > UNI2-h) partially aligns with the PFM ranking in low-label prediction where GPFM was the best performer, but diverges from sample-level and cross-slice results where the ranking differed (UNI2-h showed the largest gains in sample-level prediction, and GPFM showed

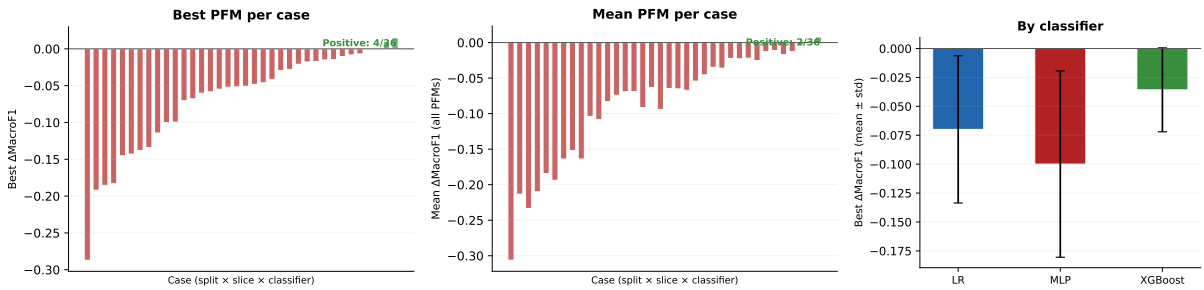
the least negative values in cross-slice prediction), indicating that spatial autocorrelation is a task-relevant — but not sufficient — property for feature-level augmentation utility.



(a) **(b)**

Figure 20: Spatial autocorrelation of PFM embeddings across 12 DLPFC sections: **(a)** Moran's I (higher = stronger spatial correlation); **(b)** Geary's C (lower = stronger spatial correlation). Each boxplot shows the distribution across sections for one PFM.

Fourth, to determine whether the uniformly negative cross-slice PFM results could be rescued by permissive configuration selection, we performed a post hoc optimization mirroring the synthetic-observation analysis (Section 4.3.2). For each of the 36 split–section–classifier cases (3 splits × 4 held-out test slices × 3 classifiers), we selected the best-performing PFM variant within each case, across all four PFMs. The resulting best Δmacro F1 values remained centered at negative values, with only 4 of 36 cases positive (Figure 21a). Even under mean aggregation across all PFMs, only 2 of 36 cases were positive (Figure 21b). No classifier showed systematic positive gains (Figure 21c): MLP, which benefited most from PFM features in other tasks, yielded the lowest best Δ. This analysis confirms that cross-slice prediction remains unsolved for PFM features: even permissive selection within each split–section–classifier case cannot produce systematic improvement, paralleling the synthetic-observation conclusion that cross-slice generalization is fundamentally difficult for either augmentation route.



(a) **(b)** **(c)**

Figure 21: Post-hoc analysis of feature-level augmentation for cross-slice transfer across 36 split–section–classifier cases. **(a)** Best Δmacro F1 under permissive PFM selection. **(b)** Mean Δmacro F1 across all PFMs and split seeds. **(c)** Best Δmacro F1 stratified by classifier.

Together, these analyses refine the interpretation of feature-level augmentation. PFM embeddings improve downstream performance when the added feature dimensions are compatible

with the task — MLP exploits enriched representations in low-label and sample-level prediction, whereas naive PCA concatenation disrupts spatial graph construction in clustering. The spatial autocorrelation of PFM embeddings partially predicts this utility but is task-dependent: GPFM's strong spatial structure aligns with its low-label ranking, yet UNI2-h leads in sample-level prediction. Cross-slice failure is not rescued by permissive PFM selection, confirming that the distribution shift between training and held-out sections is a fundamental limitation shared by both augmentation routes. The same evaluation scaffold thus reveals that PFM utility, like synthetic-observation utility, is jointly determined by the feature source, the downstream task, and the compatibility between the two.

5 Discussion

This study provides a unified perspective on augmentation in spatial transcriptomics by distinguishing two complementary augmentation routes. Feature-level augmentation enriches the representation of existing observations, whereas synthetic-observation augmentation expands the available observation set. Rather than identifying a universally superior augmentation strategy, our results show that these routes address different statistical limitations and should be selected according to the dominant bottleneck of the downstream task.

Across the four downstream tasks considered in this study, augmentation utility was consistently task dependent. Feature-level augmentation was most effective when richer representations provided complementary information for prediction, particularly in supervised settings with limited annotations. Synthetic-observation augmentation was most beneficial when additional observations improved downstream estimation, such as graph construction for spatial clustering or aggregation for sample-level prediction. In contrast, neither augmentation route consistently improved cross-slice prediction, suggesting that substantial distribution shift cannot be resolved simply by enriching feature representations or increasing the number of training observations. These findings indicate that augmentation should be understood through the statistical bottleneck of the downstream task rather than through augmentation route alone.

The comparison between the two augmentation routes also suggests practical guidance for future applications. When complementary modalities such as histology images contain informative signals, feature-level augmentation provides an effective mechanism for enriching individual observations. When the primary limitation is insufficient observations, sparse spatial coverage, or limited labeled data, synthetic-observation augmentation provides a complementary strategy by expanding the empirical support available to downstream models. Importantly, intrinsic measures such as representation quality or synthetic data fidelity alone are insufficient to predict downstream performance. Instead, augmentation should be evaluated according to its impact on downstream analyses under matched experimental settings.

Several directions remain for future work. First, the current study focuses on representative downstream tasks, while other applications, including spatially variable gene detection, deconvolution, trajectory inference, and multimodal integration, provide additional opportunities for evaluating augmentation strategies. Second, the feature-level augmentation considered here relies primarily on simple feature concatenation for general downstream models. More sophisticated multimodal integration strategies, particularly graph-based architectures that jointly

exploit pretrained representations and synthetic observations, may further strengthen the complementarity between the two augmentation routes. Finally, although this work is motivated by spatial transcriptomics, the conceptual distinction between representation enrichment and observation expansion is broadly applicable to other data modalities. As foundation models and generative models continue to evolve, the central question is unlikely to be which augmentation method is universally best, but rather how different forms of additional information should be incorporated into downstream statistical analyses. We hope that spAug provides a principled framework for answering this question in spatial transcriptomics and beyond.

Data availability

All datasets used in this study are publicly available. The DLPFC spatial transcriptomics data (Maynard et al., 2021) are accessed through SpatialLIBD (Pardo et al., 2022) (https://research.libd.org/spatialLIBD/). The HEST-1k collection (Jaume et al., 2024) is available at https://huggingface.co/datasets/MahmoodLab/hest. The three sample-level cohorts (Kidney, Bowel, Brain) are derived from publicly available spatial transcriptomics studies within HEST-1k and SpatialLIBD, processed as described in Section 3.6. Pathology foundation model weights are publicly available: GPFM at https://github.com/birkhoffkiki/GPFM, UNI at https://huggingface.co/MahmoodLab/UNI, UNI2-h at https://huggingface.co/MahmoodLab/UNI2-h, and GigaPath at https://huggingface.co/prov-gigapath/prov-gigapath. The upstream software for the synthetic-observation generators is publicly available from the original project pages: SRTsim at https://github.com/xzhoulab/SRTsim, Splatter at https://github.com/Oshlack/splatter, SPARSim at https://gitlab.com/sysbiobig/sparsim, scGAN at https://github.com/imsb-uke/scGAN, and scDiffusion at https://github.com/EperLuo/scDiffusion.

References

Giovanni Baruzzo, Ilaria Patuzzi, and Barbara Di Camillo. SPARSim single cell: a count data simulator for scRNA-seq data. *Bioinformatics*, 36(5):1468–1475, 2020. doi: 10.1093/bioinformatics/btz752.

Tommaso Biancalani, Giacomo Scalia, Lorenzo Buffoni, Rajat Avasthi, Ziqing Lu, Aman Sanger, Necmettin Tokcan, Charles R. Vanderburg, Asa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Evan Z. Macosko, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature Methods*, 18:1352–1362, 2021. doi: 10.1038/s41592-021-01264-7.

Dylan M. Cable, Evan Murray, Luli S. Zou, Aleksandr Goeva, Evan Z. Macosko, Fei Chen, and Rafael A. Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40:517–526, 2022. doi: 10.1038/s41587-021-00830-w.

Ao Chen, Shudong Liao, Mengnan Cheng, Kai Ma, Liang Wu, Yiwei Lai, Xiaojie Qiu, Jinlong Yang, Jie Xu, Shuai Hao, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis

using DNA nanoball-patterned arrays. *Cell*, 185(10):1777–1792.e21, 2022. doi: 10.1016/j.cell.2022.04.003.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Kangning Dong and Shihua Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications*, 13:1739, 2022. doi: 10.1038/s41467-022-29439-6.

Ruben Dries, Qian Zhu, Ruochi Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianhao Zhao, Abhirup Sarkar, Feng Bao, Rani E. George, Nolan Pierson, Long Cai, and Guo-Cheng Yuan. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22:78, 2021. doi: 10.1186/s13059-021-02286-2.

Miquel Elosua-Bayes, Pau Nieto, Elisabetta Mereu, Ivo Gut, and Holger Heyn. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research*, 49(9):e50, 2021. doi: 10.1093/nar/gkab043.

Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, 4:827–834, 2020. doi: 10.1038/s41551-020-0578-x.

Jian Hu, Xiangjie Li, Kyle Coleman, Anne Schroeder, Nan Ma, David J. Irwin, Edward B. Lee, Russell T. Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18:1342–1351, 2021. doi: 10.1038/s41592-021-01255-8.

Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2402.04376. arXiv:2402.04376.

Guillaume Jaume, Paul Doucet, Andrew H. Song, Ming Y. Lu, Cristina Almagro-Pérez, Sophia J. Wagner, Anurag J. Vaidya, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, and Faisal Mahmood. HEST-1k: A dataset for spatial transcriptomics and histology image analysis. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2406.16192. arXiv:2406.16192.

Vitalii Kleshchevnikov, Artem Shmatko, Emma Dann, Alexander Aivazidis, Hamish W. King, Tong Li, Rasa Elmentaite, Artem Lomakin, Veronika Kedlian, Adam Gayoso, M. Jain, J. E. Park, L. Ramona, E. Tuck, A. Arutyunyan, Roser Vento-Tormo, Moritz Gerstung, Louisa James, Oliver Stegle, and Omer Ali Bayraktar. Cell2location maps fine-grained cell

types in spatial transcriptomics. *Nature Biotechnology*, 40:661–671, 2022. doi: 10.1038/s41587-021-01139-4.

Xiaoqi Liang, Yue Cao, and Jean Yee Hwa Yang. Multi-task benchmarking of spatially resolved gene expression simulation models. *bioRxiv*, 2024. doi: 10.1101/2024.05.29.596418.

Tianyu Liu, Tinglin Huang, Tong Ding, Hao Wu, Peter Humphrey, Sudhir Perincheri, Kurt Schalper, Rex Ying, Hua Xu, James Zou, Faisal Mahmood, and Hongyu Zhao. spEMO: Leveraging multi-modal foundation models for analyzing spatial multi-omic and histopathology data. *bioRxiv*, 2025. doi: 10.1101/2025.01.13.632818.

Yingxin Long, Kok Siong Ang, Meng Li, Kian Long Kelvin Chong, Ravi Sethi, Cheng Zhong, Hang Xu, Chia-Huey Ong, Karishma Sachaphibulkij, Ao Chen, Huan Zeng, Qi Fu, Xin Yi, Ko-Teng Lim, Angela Takano, Patrick Tan, Shyam Prabhakar, and Jinmiao Chen. Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, 14:1155, 2023. doi: 10.1038/s41467-023-36796-3.

Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scDiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9):btae518, 2024. doi: 10.1093/bioinformatics/btae518. URL https://doi.org/10.1093/bioinformatics/btae518.

Jiabo Ma, Zhengrui Guo, Fengtao Zhou, Yihui Wang, Yingxue Xu, Jinbang Li, Fang Yan, Yu Cai, Zhengjie Zhu, Cheng Jin, Yi Lin, Xinrui Jiang, Chenglong Zhao, Danyi Li, Anjia Han, Zhenhui Li, Ronald Cheong Kin Chan, Jiguang Wang, Peng Fei, Kwang-Ting Cheng, Shaoting Zhang, Li Liang, and Hao Chen. Towards a generalizable pathology foundation model via unified knowledge distillation, 2024. URL https://arxiv.org/abs/2407.18449.

Mohamed Marouf, Pierre Machart, Vishal Bansal, Christoph Kilian, Daniel S. Magruder, Christoph F. Krebs, and Stefan Bonn. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nature Communications*, 11:166, 2020. doi: 10.1038/s41467-019-14018-z.

Kristen R. Maynard, Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini, Matthew N. Tran, Zachary Besich, Madhavi Tippani, Jennifer Chew, Yifeng Yin, Joel E. Kleinman, Thomas M. Hyde, N. V. Rao, Stephanie C. Hicks, Keri Martinowich, and Andrew E. Jaffe. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24:425–436, 2021. doi: 10.1038/s41593-020-00787-0.

Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna C. Schaar, Louis B. Kuemmerle, Sergei Rybakov, Ignacio L. Ibarra, Oskar Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, and Fabian J. Theis. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 19:171–178, 2022. doi: 10.1038/s41592-021-01358-2.

Brenda Pardo, Abby Spangler, Lukas M. Weber, Stephanie C. Page, Stephanie C. Hicks, Andrew E. Jaffe, Keri Martinowich, Kristen R. Maynard, and Leonardo Collado-Torres. spatial-

libd: an r/bioconductor package to visualize spatially-resolved transcriptomics data. *BMC Genomics*, 23:434, 2022. doi: 10.1186/s12864-022-08601-w.

Duy Pham, Xiaomei Tan, Jian Xu, Lisa F. Grice, Pei Y. Lam, Alexandra Raghubar, Jasmina Vukovic, Marc J. Ruitenberg, and Quan Nguyen. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv*, 2020. doi: 10.1101/2020.05.31.125658.

Samuel G. Rodriques, Robert R. Stickels, Aleksandr Goeva, Carly A. Martin, Evan Murray, Charles R. Vanderburg, Joshua Welch, Linlin M. Chen, Fei Chen, and Evan Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019. doi: 10.1126/science.aaw1219.

Xiaotong Shen, Yifei Liu, and Rex Shen. Boosting data analytics with synthetic volume expansion. *arXiv preprint*, 2023. doi: 10.48550/arXiv.2310.17848. URL https://arxiv.org/abs/2310.17848. arXiv:2310.17848.

Dongyuan Song, Qingyang Wang, Guanao Yan, Tianyi Liu, Tianyu Sun, and Jingyi Jessica Li. scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology*, 42:247–252, 2024. doi: 10.1038/s41587-023-01772-1.

Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul I. Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353 (6294):78–82, 2016. doi: 10.1126/science.aaf2403.

Shiquan Sun, Jiaqi Zhu, and Xiang Zhou. SPARK: identifying spatially expressed genes in spatial transcriptomic studies. *Nature Methods*, 17:715–725, 2020. doi: 10.1038/s41592-019-0701-7.

Valentine Svensson, Sarah A. Teichmann, and Oliver Stegle. SpatialDE: identification of spatially variable genes. *Nature Methods*, 15:343–346, 2018. doi: 10.1038/nmeth.4636.

Xinyu Tian and Xiaotong Shen. Conditional data synthesis augmentation. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2504.07426. URL https://arxiv.org/abs/2504.07426. arXiv:2504.07426.

V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019. doi: 10.1038/s41598-019-41695-z.

Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Åijo, Richard Bonneau, Ludvig Bergenstråhle, José Fernández Navarro, John Gould, Gabriel K. Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisén, Joakim Lundeberg, Aviv Regev, and Patrik L. Ståhl. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, 16:987–990, 2019. doi: 10.1038/s41592-019-0548-y.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. doi: 10.1038/s41586-024-07441-w. URL https://doi.org/10.1038/s41586-024-07441-w.

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174, 2017. doi: 10.1186/s13059-017-1305-0.

Edward Zhao, Michael R. Stone, Xiaoyang Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, Stephen R. Williams, Cedric R. Uytingco, Sarah E. B. Taylor, Paul Nghiem, Jason H. Bielas, and Raphael Gottardo. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39:1375–1384, 2021. doi: 10.1038/s41587-021-00935-2.

Jiaqi Zhu, Shiquan Sun, and Xiang Zhou. SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, 22:184, 2021. doi: 10.1186/s13059-021-02404-0.

Jiaqi Zhu, Liang Shang, and Xiang Zhou. SRTsim: spatial pattern preserving simulations for spatially resolved transcriptomics. *Genome Biology*, 24:39, 2023. doi: 10.1186/s13059-023-02879-z.

The appendix contains definitions and compact audit tables needed to interpret the main text. Full configuration-level result atlases, extended parameter-response figures, feature-augmentation detail tables, and machine-readable result schemas are organized in the Supplementary Information.

A Configuration-Level Evaluation and Decision Rules

spAug compares augmentation at the configuration level. A configuration fixes the task, dataset, augmentation route, source, strategy or feature-integration rule, parameter setting, downstream backend/model, and evaluation unit:

$$q = (t, d, r, s, \pi, \eta, b, u).$$

Each augmented configuration is compared with the real-only baseline matched on the same task, dataset, downstream model, split or fold, and evaluation unit:

$$\Delta(q) = M(q) - M_{\text{real}}(t, d, b, u),$$

where M is the task-specific primary metric. Table 11 defines the decision criteria used in the Results. A baseline-improving best configuration has best $\Delta > 0$, whereas a stable improvement requires the improvement to persist beyond a single isolated variant. Broad failure denotes consistently non-positive mean or median deltas, ceiling-limited cases are interpreted cautiously

because the real-only baseline leaves little measurable headroom, Synthetic-only diagnostics test substitutability, and backend-specific gains are interpreted only within the changed backend.

Table 11: Decision rules for interpreting augmentation effects.

Term	Criterion
Baseline-improving best configuration	Best $\Delta > 0$
Stable improvement	Median $\Delta > 0$ or high positive-unit fraction
Broad failure	Mean and median $\Delta \leq 0$ with low positive-unit fraction
Ceiling-limited	Matched real-only score near maximum
Synthetic-only diagnostic	Training excludes real observations
Backend-specific gain	Improvement tied to a changed backend

These rules are applied only within a predefined task-compatible configuration space. A positive best variant is not interpreted as evidence that a generator, PFM source, or augmentation route is uniformly beneficial.

Table 12: Stability summary for synthetic-observation augmentation. Spatial clustering uses ΔNMI; prediction tasks use Δmacro F1.

Task	Dataset	Strategy	n	Unit	Mean Δ	95% CI	Median Δ	Positive fraction	Best Δ
Spatial clustering	DLPFC	Global Real+Synthetic	1200	Configuration–section	-0.018	[-0.025, -0.012]	-0.003	0.487	0.243
	DLPFC	Local Real+Synthetic	240	Configuration–section	0.021	[0.012, 0.029]	0.028	0.625	0.238
Low-label prediction	DLPFC	Global Real+Synthetic	48600	Configuration–section–seed	-0.061	[-0.061, -0.060]	-0.065	0.170	0.217
	DLPFC	Synthetic-only	48600	Configuration–section–seed	-0.229	[-0.231, -0.228]	-0.176	0.062	0.209
	DLPFC	Local Real+Synthetic	6480	Configuration–section–seed	-0.056	[-0.057, -0.055]	-0.055	0.098	0.177
	DLPFC	Weighted ERM	97200	Configuration–section–seed	-0.058	[-0.059, -0.057]	-0.042	0.206	0.223
Cross-slice prediction	DLPFC	Global Real+Synthetic	5400	Configuration–held-out section	-0.127	[-0.129, -0.126]	-0.122	0.003	0.034
	DLPFC	Synthetic-only	5400	Configuration–held-out section	-0.247	[-0.251, -0.243]	-0.189	0.000	-0.035
	DLPFC	Local Real+Synthetic	720	Configuration–held-out section	-0.095	[-0.099, -0.093]	-0.090	0.003	0.004
	DLPFC	Weighted ERM	10800	Configuration–held-out section	-0.076	[-0.077, -0.074]	-0.064	0.041	0.021
Sample-level prediction	Kidney	Global Real+Synthetic	2250	Configuration–fold	-0.027	[-0.033, -0.022]	0.000	0.143	0.349
	Kidney	Synthetic-only	2250	Configuration–fold	-0.061	[-0.070, -0.052]	0.000	0.155	0.349
	Kidney	Weighted Aggregation	4500	Configuration–fold	-0.005	[-0.008, -0.002]	0.000	0.105	0.381
	Bowel	Global Real+Synthetic	2250	Configuration–fold	-0.047	[-0.054, -0.040]	0.000	0.148	0.417
	Bowel	Synthetic-only	2250	Configuration–fold	-0.058	[-0.065, -0.051]	0.000	0.134	0.417
	Bowel	Weighted Aggregation	4500	Configuration–fold	-0.031	[-0.036, -0.027]	0.000	0.129	0.417
	Brain	Global Real+Synthetic	2250	Configuration–fold	-0.148	[-0.159, -0.137]	0.000	0.041	0.238
	Brain	Synthetic-only	2250	Configuration–fold	-0.166	[-0.177, -0.154]	0.000	0.040	0.238
	Brain	Weighted Aggregation	4500	Configuration–fold	-0.077	[-0.083, -0.071]	0.000	0.038	0.238

Table 13: Stability summary for feature-level augmentation. Spatial clustering reports mean ARI across 12 DLPFC sections; prediction tasks report mean macro F1. Deltas are relative to matched PCA baselines except the SpaGCN row, which is reported as a graph-aware architectural reference.

Task	Best or representative feature configuration	Baseline	Feature score	Interpretation
Spatial clustering	PCA+UNI concatenation	PCA ARI 0.208	ARI 0.110; Δ=-0.098	Simple PFM concatenation degrades PCA-spatial Leiden
Spatial clustering	SpaGCN+UNI	PCA ARI 0.208	ARI 0.394; apparent Δ=+0.186	Positive but backend-specific graph-aware result
Low-label prediction	PCA+GPFM	PCA macro F1 0.651	Macro F1 0.754; Δ=+0.103	Consistent feature-level improvement
Cross-slice prediction	PCA+GPFM	PCA macro F1 0.589	Macro F1 0.512; Δ=-0.077	PFM features do not transfer across held-out sections
Sample-level prediction	PCA+UNI2-h	PCA macro F1 0.846	Macro F1 0.954; Δ=+0.108	Strongest feature-level improvement

B Task-Compatible Synthetic-Observation Strategies

Synthetic-observation augmentation first builds a task-matched synthetic pool and then applies a rule for adding or weighting generated spots. Matching strata preserve task structure: section, label, label–region, training section, or sample. For spot-level supervised prediction, Weighted ERM uses

$$\widehat{L}_\alpha(f) = (1-\alpha)\frac{1}{|\mathcal{R}^{\mathrm{tr}}|}\sum_{i\in\mathcal{R}^{\mathrm{tr}}} \ell(f(x_i), y_i) + \alpha\frac{1}{|\mathcal{S}^{\mathrm{tr}}|}\sum_{j\in\mathcal{S}^{\mathrm{tr}}} \ell(f(\tilde{x}_j), \tilde{y}_j).$$

For sample-level prediction, the same symbol α weights generated spots inside the sample representation rather than a spot-level loss. The roles of the five strategies are as follows. Global Real+Synthetic appends matched generated observations to real training strata. Synthetic-only tests whether generated observations can substitute for real training support. Local Real+Synthetic enriches local strata under a local cap. Weighted ERM separates synthetic count from synthetic loss contribution, and Weighted Aggregation performs the analogous weighting inside sample-level construction.

Table 14: Synthetic-observation strategies and task compatibility.

Strategy	Parameter	Compatible tasks
Global Real+Synthetic	ρ	Spatial, low-label, cross-slice, sample-level
Synthetic-only	ρ	Low-label, cross-slice, sample-level
Local Real+Synthetic	$\rho_{\max}$	Spatial, low-label, cross-slice
Weighted ERM	α	Low-label, cross-slice
Weighted Aggregation	α	Sample-level

Not every strategy is used for every task; task compatibility follows the evaluation matrix in Table 1.

C Dataset Splits and Leakage Control

All downstream evaluations keep held-out units real-only. Synthetic observations can modify only the training-side empirical support, and preprocessing parameters are fitted within the corresponding training unit. Feature-level augmentation uses only measured observations: PFM embeddings are frozen image-derived covariates aligned to the same real spots or samples used by the matched PCA baseline, and no generated observations are introduced in this route. For clustering, labels are used only for evaluation and synthetic spots are removed before scoring;

for spot-level prediction, models and transformations are fitted only on labeled or training-side units; for cross-slice prediction, held-out sections are transformed only after training-side fitting; and for sample-level prediction, folds are fixed before either route is applied.

Table 15: Held-out units used for leakage-controlled evaluation.

Task	Held-out unit
Spatial clustering	Real spots in same section
Low-label prediction	Unlabeled spots in same section
Cross-slice prediction	Held-out sections
Sample-level prediction	Held-out samples

D Downstream Models, Metrics, and Reporting Units

The downstream model is part of the evaluated configuration. Primary metrics are ARI/NMI for clustering and macro F1 for prediction. Auxiliary metrics are reported in result files and supplementary material.

Table 16: Downstream backends, reporting units, and key evaluation settings.

Route	Task	Backend	Reporting unit	Key setting
Synthetic observation	Spatial clustering	PCA-spatial Leiden	Section	50 expression PCs, spatial weight 1.0, Leiden resolution 1.0, seed 42
	Low-label prediction	LR / MLP / XGBoost	Section–seed–fraction	50 expression PCs; 25%, 50%, and 75% labeled spots; seeds 42–44
	Cross-slice prediction	LR / MLP / XGBoost	Split–test-section	50 expression PCs; three 8-train/4-test DLPFC section splits
	Sample-level prediction	Weighted Aggregation	Fold	30 expression PCs; five stratified sample-level folds; sample summaries
Feature level	Spatial clustering	PCA-spatial Leiden; SpaGCN	Section	PCA+PFM concatenation or SpaGCN+PFM labels; frozen GPFM/UNI/UNI2-h/GigaPath embeddings
	Low-label prediction	LR / MLP / XGBoost	Section–seed–fraction	Train-fitted 50-PC expression representation concatenated with aligned PFM embeddings
	Cross-slice prediction	LR / MLP / XGBoost	Split–test-section	Combined-train 50-PC expression representation concatenated with frozen PFM embeddings
	Sample-level prediction	PFM Sample Aggregation	Fold	30-PC spot expression summaries augmented by PFM-derived spot/sample summaries