PERSIST: PERSISTENCE-GUIDED MULTISCALE DOMAIN IDENTIFICATION USING STABILITY FIELD ANALYSIS IN SPATIAL TRANSCRIPTOMICS

GAURAV KHANAL[1]

Abstract. Spatial transcriptomics data exhibit complex organization across multiple spatial scales, yet most existing approaches identify domains using single-resolution clustering objectives or fixed smoothing assumptions. We introduce PERSIST, a topology-guided framework for multiscale spatial domain identification that decouples scale discovery from downstream inference. Rather than tracking individual topological features, PERSIST uses persistent homology to identify intrinsic spatial scales at which nontrivial organization emerges and leverages this information to guide multiscale regularization on a spatial graph. The method constructs a continuous spatial stability field that quantifies transcriptional coherence across topology-informed scales, enabling the identification of stable domains and transition regions without reliance on learned embeddings, resolution parameters, or biological priors. We apply PERSIST to spatial transcriptomics datasets from mouse brain, human lymph node, and breast cancer tissue, demonstrating consistent behavior across tissue architectures and species. Across all cases, PERSIST reveals interpretable multiscale organization and spatial boundaries that are not captured by clustering-based approaches. This framework provides a general, unsupervised method for topology-informed analysis of spatially resolved molecular data.

Keywords. Spatial transcriptomics; persistent homology; multiscale analysis; topological data analysis; domain segmentation

1. Introduction

Spatial transcriptomics enables the measurement of gene expression while preserving tissue architecture, bridging single-cell genomics and histopathology [27, 28]. This spatial context is essential, as cells with similar molecular profiles can exhibit distinct behaviors depending on their location within a tissue [29, 30]. Despite this promise, extracting principles of spatial organization from spatial transcriptomics data remains challenging. Many computational approaches treat spatial coordinates primarily as a regularization constraint on expression-based clustering, encouraging neighboring locations to share labels rather than modeling spatial structure as an object of interest in its own right [31, 32]. As a result, such methods often recover transcriptionally similar regions while overlooking spatial organization that persists independently of gene expression. For example, large anatomical structures such as white matter are frequently segmented as homogeneous regions despite evidence of internal hubs, transit corridors, and boundary zones.

Existing methods differ in how spatial information is integrated with gene expression. Some approaches, including SpatialDE [37] and spark [36], identify spatially variable genes without performing segmentation. Others, such as BayesSpace [33] and banksy [34], directly incorporate spatial proximity into clustering models to produce spatially smooth expression domains. While effective for identifying transcriptionally coherent regions, these methods can suppress spatial organization that does not align with expression similarity. Methods such as SpaGCN [35] and Giotto [38] occupy an intermediate position, using spatial information to refine expression-driven clusters; however, domain boundaries remain governed by transcriptional similarity, and none of these frameworks explicitly assess whether inferred domains reflect spatial organization that is stable across scales.

A related challenge concerns the choice of spatial scale at which tissue organization is analyzed. Spatial transcriptomics data exhibit structure across multiple characteristic length scales, from local neighborhoods to large anatomical compartments. Most existing methods fix scale implicitly through neighborhood size, smoothing radius, or clustering resolution, often selected heuristically. Consequently, inferred domains can be highly sensitive to parameter choice, making it difficult to distinguish intrinsic tissue organization from scale-specific artifacts. While multiscale approaches exist, they typically aggregate results across resolutions without identifying which scales are supported by the underlying geometry.

PHD-MS represents a recent attempt to address this issue by applying persistent homology to identify spatial domains that remain stable across multiple morphological scales [39]. The method constructs segmentations at a sequence of predefined resolutions and evaluates their persistence across scales. Although this approach highlights the utility of topological persistence as a stability criterion, scale is treated primarily as an external parameter rather than inferred from the spatial embedding itself.

2. PHD-MS: A Commentary

PHD-MS is a framework developed by Beamer and Cang (2025) that applies persistent homology (PH) to identify spatial transcriptomic domains that persist across multiple morphological scales [39]. The authors note that many spatial segmentation methods depend on user-chosen resolution parameters to control domain granularity, which can substantially influence results in the absence of ground truth [39]. To address this sensitivity, PHD-MS constructs a spatial cluster filtration that summarizes how domains evolve across scales and identifies structures that remain stable across multiple resolutions. Applied to Visium mouse brain data, the method recovers known hierarchical organization and nested substructures, while analysis of breast cancer tissue distinguishes stable "core" regions from unstable "frontier" regions at interfaces between malignant and healthy tissue [39].

Most spatial segmentation approaches partition tissue into contiguous regions of similar gene expression but face two key limitations: reliance on a single resolution parameter and the imposition of hard, disjoint boundaries that fail to capture biological heterogeneity, particularly in boundary or "frontier" regions. PHD-MS addresses these issues by explicitly analyzing how clustering assignments change with resolution. The method follows a multi-stage pipeline: (1) generate nested clusterings across multiple resolutions using existing methods such as Leiden, (2) link clusters across adjacent scales by constructing a weighted overlap graph, and (3) apply PH to identify cluster structures that persist across scales and convert them into soft, multiscale domains. Overall, the framework emphasizes multiscale stability as a central organizing principle for spatial transcriptomics analysis.

2.1. **Methodology.**

2.1.1. *Cluster Filtration.* The primary input to PHD-MS is a collection of nested clustering results obtained across multiple resolutions from existing spatial clustering methods. In practice, the authors generate spatially aware embeddings using GraphST and SCAN-IT, followed by spatial domain clustering via the Leiden algorithm over a range of resolution parameters. They construct a sequence of clustering results with increasing spatial scale, using uniformly spaced resolution values between $r = 0.95$ and $r = 0.15$ [39].

These clusterings are organized into a graph in which each vertex represents a cluster at a given scale $k \in \{k_i\}_{i=1}^{n}$, and edges connect clusters at adjacent scales [39]. Edge weights encode transcriptional dissimilarity between clusters. For tissues exhibiting hierarchical nesting, such as the mouse brain, dissimilarity is quantified using the *Containment Index*:

$$f_C([C_{i,j}, C_{i+1,l}]) = 1 - \frac{|C_{i,j} \cap C_{i+1,l}|}{|C_{i,j}|}, \tag{2.1}$$

where $C_{i,j}$ and $C_{i+1,l}$ denote clusters at consecutive finer and coarser scales k_i and k_{i+1}, respectively [39]. Under this definition, perfect containment yields $f_C = 0$. For tissues lacking

strict hierarchical structure, such as breast cancer samples, the authors instead use the *Jaccard Index*:

$$f_J([C_{i,j}, C_{i+1,l}]) = 1 - \frac{|C_{i,j} \cap C_{i+1,l}|}{|C_{i,j} \cup C_{i+1,l}|}, \tag{2.2}$$

where $f_J = 0$ indicates identical clusters and $f_J = 1$ indicates disjoint clusters [39].

2.1.2. *Persistent Homology.* The resulting weighted cluster graph serves as the input to persistent homology analysis. The filtration is initialized with all vertices and no edges, after which edges are added in order of increasing dissimilarity. As edges are introduced, clusters across scales merge into connected components. Persistent homology is computed using the GUDHI library, with a union–find algorithm tracking component membership [39].

Each connected component is associated with a persistence pair (b, d), where b denotes the dissimilarity value at which the component appears and d the value at which it merges into an older component. Components with long persistence correspond to stable multiscale domains, whereas short-lived components reflect scale-specific or transient structures.

Stable components are mapped back to tissue locations via a *coreness score*, which quantifies the consistency of domain membership across scales. For a spot x, the coreness score is defined as:

$$\bar{c}_D(x) = 1 - \min_{C_{i,j}, C_{k,l} \in D} \{f_*(C_{i,j}, C_{k,l}) : x \in (C_{i,j} \cup C_{k,l})\}, \tag{2.3}$$

where $c_D : X \to [0, 1]$ [39]. High coreness values indicate stable core membership, while low values identify unstable frontier regions.

To further quantify instability, PHD-MS defines a heterogeneity score h measuring the extent to which a spot participates in multiple persistent domains across scales:

$$h(x) = \sum_{i=1}^{n} p_i \bar{c}_{D_i}(x), \tag{2.4}$$

where p_i denotes the persistence lifetime $(d - b)$ of domain D_i [39]. The heterogeneity score is maximized when a spot is a core member of multiple highly persistent domains and minimized when it consistently belongs to a single domain.

2.1.3. *Evaluation.* The authors use Wasserstein distance and Normalized Mutual Information (NMI) to benchmark the PHDMS method against ground truth annotations and other clustering algorithms. Wasserstein Distance is used to measure the spatial relevance between spatial domains converting each domain into spatial probability distributions. The metric calculates the "cost" in microns to move the predicted distribution to the ground truth distribution [39]. NMI is used to measure the clustering accuracy (or overlap fidelity) while accounting for the soft and multiscale nature of the PHD-MS output. The authors have generalized the NMI formula to accept probabilistic/soft vectors instead of binary labels by constructing a matrix D of normalized coreness scores [39].

2.2. **State of the Art.** The central novelty of PHD-MS is a conceptual shift from single-scale, hard spatial segmentation toward multiscale soft domain identification using topological data analysis (TDA). Most existing spatial transcriptomics methods require the selection of a single resolution parameter, producing a static view of tissue organization that can obscure hierarchical structure and enforce rigid boundaries. In contrast, PHD-MS defines domains as structures that persist across spatial resolutions, aggregating information from nested clusterings to identify biologically stable regions without committing to a single "correct" scale [39].

Algorithmically, PHD-MS departs from standard TDA pipelines by applying persistent homology not to raw data points, but to a graph of clustering labels constructed across resolutions. By tracking connected components in this filtration graph, the method reframes segmentation stability as a topological property: domains are defined by their persistence across scales rather

than by agreement at any single resolution. This makes robustness explicit and measurable, and elevates scale to a first-class object of analysis rather than a fixed hyperparameter.

Beyond domain identification, PHD-MS replaces binary cluster assignments with continuous, overlapping representations. The coreness score (c_D) assigns each spot a value between 0 and 1 based on the consistency of its domain membership across scales, distinguishing stable domain cores from unstable frontier regions [39]. This formulation naturally captures biologically meaningful boundary zones, such as tumor–stroma interfaces, where transcriptional programs overlap rather than forming sharp transitions.

The paper also advances evaluation practice by adapting existing validation metrics to soft, multiscale domains. Coreness scores are normalized to form probability vectors over domains, enabling a generalized formulation of normalized mutual information (NMI) that rewards accurate representation of ambiguous or transitional regions [39]. Similarly, Wasserstein distance is computed by treating domains as spatial mass distributions over physical coordinates, penalizing geometric distortions in proportion to their spatial extent. Together, these adaptations provide spatially interpretable measures of domain accuracy that complement traditional label-based metrics.

2.3. **Biological Discovery and/or Results.** The authors demonstrate that PHD-MS captures tissue organization across multiple spatial scales and reveals substantial instability in single-resolution spatial segmentation. Across datasets, the analysis shows that spatial partitions vary markedly as resolution changes, even when gene expression data are fixed, and that no single resolution recovers all biologically plausible regions. These observations support the authors' central claim that single-scale spatial segmentation is inherently unstable.

In the mouse brain dataset, PHD-MS recovers major anatomical divisions annotated in the Allen Brain Atlas, including the cerebral cortex, fiber tracts, hippocampus, basal nuclei, amygdala, and hypothalamus [39]. At intermediate scales, several of these regions emerge as persistent domains, while finer-scale substructures appear as nested, more transient components. The authors report that PHD-MS identifies certain hypothalamic subregions and compact anatomical formations more clearly than comparison methods such as NesST, and that the hippocampus exhibits stable internal substructure consistent with known biological organization [39].

In breast cancer tissue, PHD-MS does not outperform methods optimized for annotation agreement, such as GraphST, but reveals patterns of spatial stability and heterogeneity that are not captured by single-resolution approaches [39]. Tumor epithelial cores are identified as highly persistent across scales, whereas stromal and immune-associated regions display substantial instability, frequently splitting or merging as resolution varies. These unstable regions are interpreted as intrinsically multiscale and heterogeneous components of the tumor microenvironment. The authors further support this interpretation through differential gene expression and gene ontology enrichment analyses, which associate unstable regions with immune activity and microenvironmental signaling.

2.4. **Strengths.** The primary strength of PHD-MS lies in its use of PH as a robustness criterion for spatial domain identification. Rather than selecting a single "correct" segmentation, the method reframes spatial analysis as a multiscale stability problem, identifying domains that persist across resolution parameters. This shift is biologically well motivated, as tissues are often hierarchical, exhibit fuzzy boundaries, and contain overlapping organizational structures. By design, PHD-MS avoids enforcing sharp boundaries that rarely exist in biological systems.

Methodologically, PHD-MS applies PH not to raw expression data or spatial coordinates, but to a cross-scale overlap graph constructed from standard clustering outputs. This choice renders the approach modular and interpretable, while making robustness explicit and measurable. By tracking domain birth, death, merging, and splitting across scales, the framework exposes the inherent instability of single-scale segmentation methods and elevates scale to a first-class object of analysis.

A further strength is the introduction of a continuous coreness score that replaces hard cluster assignments with soft, overlapping domain memberships. This representation distinguishes

stable domain cores from unstable frontier regions and provides quantitative support for biological heterogeneity beyond visual inspection. Such frontier regions frequently correspond to biologically active interfaces, including tumor–stroma boundaries enriched for immune activity.

Finally, the authors demonstrate the generality of the framework across diverse tissues and spatial technologies, including Visium, MERFISH, and osmFISH. Together, these results support the applicability of PHD-MS as a general multiscale analysis tool for spatial transcriptomics.

2.5. **Limitations.** Despite its conceptual elegance and strong empirical performance, PHD-MS has several limitations that affect its broader applicability. A fundamental constraint is its reliance on upstream embeddings and clustering methods. Because the framework operates on clustering outputs rather than raw data, systematic biases or failures in upstream methods may be inherited and reinforced by persistence-based stability scores. While PH quantifies robustness, it cannot distinguish biologically meaningful persistence from stability arising due to methodological artifacts.

In addition, PHD-MS is restricted to zero-dimensional persistent homology (H_0), which captures connected components but ignores higher-dimensional topological features. As a result, loop-like or boundary-enclosed structures that may reflect meaningful tissue morphology are not explicitly modeled, and weakly connected regions may be merged early in the filtration.

Although designed to mitigate resolution sensitivity, the method still depends on a user-defined grid of resolution parameters used to generate the clustering sequence. While the authors demonstrate robustness to moderate perturbations, there is no guarantee that all biologically relevant scales are sampled, particularly in tissues with complex or uneven hierarchical organization.

Finally, the coreness score, while intuitive and visually informative, reflects cross-scale stability rather than statistical uncertainty. Without an explicit probabilistic interpretation, small quantitative differences may be over-interpreted. Moreover, although overlapping domains are a conceptual strength, downstream analyses often revert to hard assignments, partially undermining the soft, multiscale formulation.

These limitations motivate the framework introduced in the following section. Rather than operating on clustering outputs or resolution-dependent segmentations, we directly model spatial organization as a geometric and topological property of the tissue embedding itself. In particular, we decouple scale discovery from downstream inference by inferring intrinsic spatial scales via persistent homology and integrating this information into a variational stability field defined on the spatial domain. This formulation avoids reliance on upstream embeddings, fixed resolution grids, or hard partitions, and enables the identification of stable domains and transition regions supported by the underlying tissue geometry.

3. Methodology

We propose PERSIST, a PH-guided multiscale domain identification variational framework for segmenting spatial transcriptomics data into geometrically stable domains without relying on fixed neighborhood parameters or heuristic clustering resolutions. By combining persistent homology with discrete exterior calculus, we construct a scalar field that quantifies transcriptional stability across intrinsic topological scales, which is then partitioned via basin decomposition to recover a segmentation supported by the underlying tissue geometry.

3.1. **Geometric Setting and Function Representation.** We model the tissue as a discrete sampling of an underlying smooth manifold embedded in Euclidean space. Let $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^2$ denote the set of spatial sampling locations, where each x_i is associated with a gene expression vector $g(x_i) \in \mathbb{R}^G$, where G is the number of gene expressions. Our objective is to partition X into spatial domains $\mathcal{D} = \{D_k\}_{k=1}^K$ that remain stable across intrinsic geometric scales, without imposing *a priori* assumptions on neighborhood structure or tissue architecture.

To construct a scalar field that reflects spatial organization rather than generic transcriptional variability, we restricted attention to genes exhibiting significant spatial autocorrelation, quantified using Moran's I [1]. This filtering step removes genes whose expression varies independently

of spatial geometry and would otherwise introduce high-frequency noise into subsequent differential operators. In contrast to highly variable gene (HVG) selection, which prioritizes variance across samples [3], Moran's I explicitly selects for spatially structured expression patterns and is widely used in spatial transcriptomics [38, 37]. We then applied Principal Component Analysis (PCA) to the spatially filtered gene set to obtain a low-dimensional representation capturing dominant spatial trends [2]. The resulting scalar field $f \in C^0(X;\mathbb{R})$, interpreted as a 0-cochain on the vertex set, is defined as the projection onto the first principal component:

$$f(x_i) = \mathrm{PC}_1(g(x_i)) \tag{3.1}$$

We treat f as a noisy discrete sample of a smooth function defined on the underlying tissue manifold. Although more expressive alternatives were considered—including Hodge-theoretic decompositions [8], spectral embeddings derived from the graph Laplacian (e.g., the Fiedler vector) [10], and nonnegative matrix factorization [11]—these approaches exhibited numerical instability under changes in graph connectivity and across filtration scales. The first principal component provided a stable and reproducible scalar field suitable for multiscale analysis and is therefore adopted as a conservative model-reduction step.

3.2. **Simplicial Complex Construction and Topological Filtration.** To encode the geometry of the spatial domain across multiple length scales, we constructed a nested family of simplicial complexes parameterized by a geometric scale. This representation provides both a discrete approximation of the tissue manifold and a mechanism for identifying characteristic spatial scales through topological persistence.

3.2.1. *Adjacency Stabilization.* Prior to simplicial complex construction, we built an initial proximity graph on the spatial point set X using a k-nearest neighbor (KNN) criterion. We retained the largest connected component and removed isolated vertices with degree below a minimal threshold to suppress disconnected or noisy points near tissue boundaries. We used this graph only to define local adjacency and to stabilize subsequent differential operators. It did not determine the multiscale geometry, which we captured independently through the Alpha complex filtration.

3.2.2. *Alpha Complex Filtration.* Subsequently, we constructed the filtration using Alpha complexes. Let $\mathrm{Del}(X)$ denote the Delaunay triangulation of the point set $X \subset \mathbb{R}^2$ [17]. For a scale parameter $\alpha \geq 0$, we define the Alpha complex K_α as the subcomplex of $\mathrm{Del}(X)$ consisting of simplices whose circumspheres have squared radius at most α [18]. Varying α induces a nested sequence of simplicial complexes:

$$\varnothing = K_0 \subseteq \cdots \subseteq K_\alpha \subseteq \cdots \subseteq K_\infty = \mathrm{Del}(X), \tag{3.2}$$

which captures the evolving connectivity of the spatial domain across increasing length scales.

3.2.3. *Persistent Homology and Birth-Based Scale Selection.* We computed persistent homology of the first homology group $H_1(K_\alpha)$ along the Alpha complex filtration to characterize the evolution of one-dimensional topological cycles as the scale parameter α varied [6]. This yielded a persistence diagram

$$\mathcal{D}_1 = \{(b_k, d_k)\}_{k=1}^{M}, \tag{3.3}$$

where each birth time b_k corresponds to the scale at which a nontrivial 1-cycle first appears in the filtration.

Rather than analyzing individual topological features, we treated the empirical distribution of birth times $\{b_k\}$ as an intrinsic descriptor of geometric length scales supported by the spatial embedding. Birth events in H_1 reflect the emergence of loop-like structures associated with enclosed regions and boundary-supported geometry in the underlying point cloud [17, 7]. Consequently, the birth-time distribution provides a data-driven proxy for spatial scales at which compartmentalization becomes geometrically meaningful.

We further used this distribution to adaptively discretize the filtration parameter. By partitioning the birth-time distribution, we selected a finite set of representative scales $\{\alpha_s\}_{s=1}^{S}$, allocating higher resolution to regions with concentrated birth events. This approach yields intrinsic, topology-informed scales for subsequent differential and variational analyses, avoiding uniform or heuristic sampling of α [19, 20].

We focused on persistent homology of H_1 rather than H_0 because the goal of this work is to identify compartmentalized spatial domains rather than connected components. While H_0 primarily encodes connectivity and clustering structure, H_1 captures loop-like features corresponding to enclosed regions and domain boundaries. These features provide a more direct geometric proxy for tissue compartmentalization, aligning with the downstream objective of identifying stable spatial domains.

3.3. **Discrete Exterior Calculus on the Simplicial Complex.** To define regularization and stability operators in a geometrically principled manner, we adopted a Discrete Exterior Calculus (DEC) formulation on the simplicial complexes K_{α_s} generated by the Alpha filtration [8, 9]. Let $C^k(K_\alpha)$ denote the space of discrete k-forms (cochains) on K_α, endowed with the standard L^2 inner product.

3.3.1. *Laplacian Regularization (0-form Heat Flow).* To mitigate measurement noise in the scalar field f, we applied a diffusion process governed by the combinatorial Laplacian acting on 0-forms. The operator $\Delta_0 : C^0(K_\alpha) \to C^0(K_\alpha)$ is defined as the composition of the coboundary operator and its adjoint [8, 9]:

$$\Delta_0 = \delta_0^* \delta_0, \tag{3.4}$$

where $\delta_0 : C^0(K_\alpha) \to C^1(K_\alpha)$ is the discrete exterior derivative (gradient) and δ_0^* is the codifferential (divergence). This operator generated a discrete heat flow on the simplicial complex, analogous to diffusion on a smooth manifold [10]. The regularized field $\tilde{f}$ is obtained via a single explicit Euler step of the heat equation:

$$\tilde{f} = (I - \tau\Delta_0)f, \tag{3.5}$$

where $\tau > 0$ controls the diffusion timescale. This operation acted as a geometric low-pass filter, smoothing the signal relative to the connectivity of the 1-skeleton while preserving global structural trends.

3.3.2. *The Discrete Gradient Operator.* For stability analysis, we require the discrete gradient of the regularized field. In the DEC framework, this is given exactly by the coboundary operator δ_0 [8, 9]. For a 1-simplex (edge) $\sigma = [v_i, v_j] \in K_\alpha$ oriented from v_i to v_j, the action of δ_0 is:

$$\langle \delta_0 \tilde{f}, \sigma \rangle = \tilde{f}(v_j) - \tilde{f}(v_i). \tag{3.6}$$

This formulation yielded edge-wise differences that were intrinsic to the mesh connectivity and independent of any ambient coordinate system, ensuring consistency across scales as the simplicial complex evolved along the filtration.

3.4. **Multiscale Stability Field.** Having established both a topology-informed scale selection and a discrete differential framework, we defined a variational functional that quantifies the local stability of the transcriptional signal across scales. We developed this construction based on the total variation–based regularization on discrete domains and multiscale energy aggregation principles [13, 15]. The central object of the method is a multiscale stability field that integrates local variation over intrinsic geometric scales.

We define a multiscale stability field $\mathcal{S} : X \to \mathbb{R}$, which aggregated the local harmonic energy of the regularized transcriptional signal evaluated across the topology-informed scales $\{\alpha_s\}_{s=1}^{S}$.

3.4.1. *Local Variation as a Normalized 1-Form Norm.* For a fixed scale α_s, we defined the local variation at a vertex v_i as the normalized L^1-norm of the discrete gradient $\delta_0 \tilde{f}$ restricted to the coboundary (star) of v_i. Let $\mathrm{St}(v_i, \alpha_s) = \{\sigma \in K^1_{\alpha_s} \mid v_i \in \partial\sigma\}$ be the set of incident 1-simplices. The variation $V_{\alpha_s}(v_i)$ is given by:

$$V_{\alpha_s}(v_i) = \frac{\sum_{\sigma \in \mathrm{St}(v_i,\alpha_s)} w_\sigma |\langle \delta_0 \tilde{f}, \sigma \rangle|}{\sum_{\sigma \in \mathrm{St}(v_i,\alpha_s)} w_\sigma}, \tag{3.7}$$

where w_σ represents the geometric weight of the edge σ. This formulation follows standard total variation constructions on graphs and discrete meshes, where variation is quantified through local gradient magnitudes [13, 14].

3.4.2. *The Stability Field.* We constructed the global stability field by integrating the complement of the variation over the scale filtration, weighted by the topological persistence density ω_s. This multiscale integration reflects the principle that spatial structures supported across multiple characteristic scales are more likely to represent intrinsic organization rather than noise [15]. The stability field is defined as:

$$\mathcal{S}(v_i) = \sum_{s=1}^{S} \omega_s \left(1 - V_{\alpha_s}(v_i)\right). \tag{3.8}$$

High values of $\mathcal{S}(v_i)$ indicate vertices that maintain signal homogeneity across multiple intrinsic scales, effectively identifying the stable "cores" of stable biological domains. The weights (ω_s) are derived from the empirical persistence density at scale (α_s), emphasizing scales supported by a high concentration of persistent topological features.

3.5. **Robustness Assessment via Bootstrapping.** To evaluate the robustness of the stability field and the resulting domain cores, we performed a bootstrap analysis over spatial subsamples. In each bootstrap iteration, we sampled a fixed fraction of spatial locations (without replacement), and we recomputed the multiscale stability field using the same scale selection and operator definitions. This produces a collection of stability estimates for each vertex.

We designated the vertices that consistently attain high stability across bootstrap samples as robust cores, whereas locations with substantial variability were interpreted as less stable or scale-dependent. This analysis provided a quantitative measure of confidence in the inferred domains while leaving the segmentation procedure unchanged, and helped distinguish persistent spatial structure from sampling-induced variability.

3.6. **Domain Extraction via Basin Decomposition.** The stability field $\mathcal{S}$ induces a scalar landscape over the spatial domain. We extracted discrete spatial domains by partitioning this landscape into basins of attraction associated with its local maxima, using the adjacency structure of the 1-skeleton of the simplicial complex at the finest analysis scale $K_{\alpha_{\min}}$. Interpreting $\mathcal{S}$ as a discrete Morse-like function [12], we assigned each vertex to a basin by following steepest-ascent paths along edges of the complex until reaching a local maximum. This procedure yields a partition of the spatial domain:

$$X = \bigsqcup_{k=1}^{K} D_k. \tag{3.9}$$

Each subdomain D_k corresponds to the unstable manifold of a local maximum of $\mathcal{S}$. This construction provides a discrete approximation to a Morse–Smale decomposition, segmenting the tissue into spatial regions of high multiscale stability separated by boundaries where transcriptional variation concentrates. From a variational perspective, the resulting domains identify regions that minimize transcriptional variation across the topology-informed length scales selected by the filtration.

3.7. **Tissue Schema and Biological Annotation.** To support biological interpretation and post hoc validation of inferred spatial domains, we constructed tissue-specific marker schemas for each dataset. These schemas consisted of curated gene sets associated with known anatomical regions or cell populations relevant to the tissue under study. Importantly, the schemas were not used during scalar field construction, scale selection, stability computation, or domain segmentation, and therefore did not influence the unsupervised discovery process.

Initial marker schemas were assembled using a large language model (LLM) as an assisted curation tool to synthesize marker genes reported across the biological literature for each tissue type. To ensure biological validity and reproducibility, these schemas were subsequently validated and refined using publicly available marker gene databases and reference resources, including CellMarker [21], PanglaoDB [22], and tissue-specific annotations from published single-cell and spatial transcriptomics studies [**?**, 38]. This validation ensured that the final marker sets reflected consensus biological knowledge rather than model-generated artifacts.

Following domain extraction, we performed enrichment analyses to assess correspondence between inferred spatial domains and curated marker schemas. This enabled biological annotation and quantitative evaluation of domain identities while maintaining a strict separation between unsupervised domain discovery and downstream interpretation.

3.8. **Implementation and Software.** We implemented the methodology in Python. We computed Alpha complexes and persistent homology using the GUDHI library [23], which provides efficient implementations of simplicial complexes and persistent homology algorithms. We performed spatial transcriptomics preprocessing and gene filtering using SCANPY and SQUIDPY [24, 25]. Principal Component Analysis and related numerical operations were carried out using SCIKIT-LEARN [26].

We implemented all discrete exterior calculus operators, multiscale stability computations, and basin decomposition procedures explicitly to ensure consistency with the mathematical formulation described above. All experiments were run on standard hardware.

4. RESULTS

We applied PERSIST to spatial transcriptomic data from three biologically distinct tissues—breast cancer, human lymph node, and mouse brain—to evaluate its ability to recover tissue organization across divergent structural regimes. Together, these tissues span gradient-driven tumor microenvironments, compartmentalized immune architecture, and highly ordered neuroanatomy. Across all tissues, PERSIST produced robust, biologically interpretable representations while adapting to tissue-specific modes of organization.

4.1. **Breast cancer tissue organization.** We applied PERSIST to breast cancer spatial transcriptomic data to characterize multiscale tissue organization within a heterogeneous tumor microenvironment (Appendix A). The resulting representation exhibited near-perfect scale-local coherence (SLC = 0.9996), indicating strong robustness across scales, while maintaining substantial structural flexibility (topology orthogonality index, TOI = 0.865). Agreement with baseline clustering was moderate (Adjusted Rand Index = 0.43), suggesting that PERSIST preserves major lineage structure while introducing non-redundant organization beyond standard approaches.

Topological analysis revealed substantial higher-order structure, including 7,199 one dimensional homology (H_1) features, of which 1,313 formed robust stability cores (Appendix A). Mean stability converged to 0.616, corresponding to a broad stability plateau rather than a sharp optimum, consistent with a well-defined multiscale regime. The stability landscape decomposed the tissue into basins separated by high-gradient topological walls, defining coherent domains without imposing hard partitions across the full tissue (Appendix A).

Final topological domains corresponded to major cellular compartments of breast cancer tissue, including myoepithelial, fibroblast/stromal, endothelial, T cell, B cell, and macrophage populations. These domains exhibited strong enrichment for canonical lineage markers, with concordant mean expression and fraction of expressing cells (Appendix A). Overlay of domains

onto matched H&E histology demonstrated clear spatial concordance, with stromal regions forming contiguous domains and immune populations interspersed throughout the tumor microenvironment. Together, these results indicate that PERSIST captures biologically meaningful organization within heterogeneous tumor tissue.

4.2. **Human lymph node tissue organization.** We next applied PERSIST to human lymph node tissue to assess its performance in a highly compartmentalized immune organ (Appendix B). The learned representation exhibited high multiscale robustness (SLC = 0.9975) and strong global organization without collapse to a single axis (TOI = 0.851). Agreement with baseline clustering was moderate (Adjusted Rand Index = 0.33), indicating substantial reorganization beyond simple cell-type partitions. In contrast to breast cancer tissue, lymph node organization showed no significant coupling to a reference gradient (Spearman $\rho = -0.02$, $p = 0.13$), consistent with compartmental rather than trajectory-based tissue structure.

Topological analysis identified 7,572 H_1 features, of which 1,373 formed robust stability cores (Appendix A). Mean stability converged to 0.641, defining a broad stability plateau. Spatial analysis of the stability landscape revealed high-gradient topological walls that partitioned the tissue into discrete basins corresponding to anatomical compartments, without imposing hard global partitions (Appendix A).

Final topological domains corresponded to canonical lymph node compartments, including germinal center dark and light zones, B cell mantle regions, T zone (paracortical) regions, and high endothelial venules. These domains exhibited strong enrichment for expected lineage- and state-specific markers, with concordant mean expression and fractions of expressing cells (Appendix A). Overlay of domains onto matched H&E histology demonstrated clear spatial concordance, confirming recovery of organized immune microanatomy.

4.3. **Mouse brain tissue organization.** Finally, we applied PERSIST to mouse brain spatial transcriptomic data to evaluate its ability to recover highly ordered neuroanatomical structure (Appendix A). The resulting representation exhibited near-perfect multiscale robustness (SLC $\approx$ 1.0), reflecting the strong intrinsic organization of brain tissue. Despite this high coherence, the representation did not collapse to a trivial structure, as indicated by a topology orthogonality index of 0.72. Agreement with baseline clustering was moderate (Adjusted Rand Index = 0.41), consistent with preservation of major neuronal classes alongside reorganization to capture laminar and regional anatomy.

Topological analysis revealed a selectively stabilized structure, with 5,157 H_1 features and 933 robust stability cores (Appendix A). Mean stability converged to 0.672, corresponding to a broad stability plateau. High-gradient topological walls traced anatomical boundaries and partitioned the tissue into coherent basins aligned with cortical layers and major brain regions (Appendix A).

Final topological domains corresponded to canonical mouse brain structures, including cortical layers L2/3, L5, and L6, hippocampal CA1 and dentate gyrus, striatum, and oligodendrocyte-rich regions. These domains showed strong enrichment for expected regional and laminar markers, with concordant mean expression and fractions of expressing cells (Appendix A). Overlay of domains onto matched H&E histology demonstrated clear spatial concordance, with ordered cortical lamination and distinct separation of hippocampal and striatal regions. Stability patterns further reflected known biological differences, with highly stable oligodendrocyte and hippocampal domains and more dynamic upper cortical layers.

4.4. **Comparative analysis of multiscale organization across tissues.** While PERSIST was applied independently to each dataset, comparing the resulting stability landscapes across tissues reveals systematic differences that reflect underlying biological architecture. In particular, the distribution of topological features, stability plateaus, and domain boundary structure varied consistently with known organizational principles of tumor, immune, and neural tissues.

Breast cancer tissue exhibited a rich but heterogeneous topological structure, characterized by a large number of H_1 features and intermediate stability values. The resulting stability landscape contained both well-defined cores and extended transition regions, consistent with a

tumor microenvironment shaped by interacting malignant, stromal, and immune components. High-gradient walls frequently delineated stromal compartments, while immune-associated regions appeared as more unstable, spatially interspersed domains.

In contrast, human lymph node tissue displayed a more compartmentalized stability landscape. Although topological complexity remained high, stability cores were more sharply localized and separated by pronounced walls corresponding to anatomical boundaries. The absence of coupling to a reference gradient, together with discrete basin structure, reflects the intrinsic architectural organization of secondary lymphoid organs, in which functional zones are spatially segregated rather than arranged along continuous trajectories.

Mouse brain tissue showed the most constrained organization among the three datasets. Near-perfect multiscale coherence and lower topology orthogonality indicated dominance of a small number of organizing axes, corresponding to laminar cortical structure and regional neuroanatomy. Stability cores were tightly concentrated within hippocampal and oligodendrocyte-rich regions, while cortical layers exhibited graded but orderly variation. Unlike tumor tissue, transition regions were narrow and aligned with known anatomical boundaries.

Together, these comparative results demonstrate that differences in PERSIST outputs are not algorithmic artifacts but reflect genuine differences in tissue architecture. The framework adapts to gradient-associated, compartmental, and rigidly organized tissues without imposing a common segmentation paradigm, supporting its use as a general tool for multiscale spatial analysis.

5. Conclusion

We introduced PERSIST, a topology-guided framework for multiscale domain identification in spatial transcriptomics that decouples scale discovery from downstream segmentation. By using persistent homology to infer intrinsic geometric scales and integrating these scales into a variational stability field, PERSIST identifies spatial domains and transition regions without reliance on fixed neighborhood parameters, clustering resolutions, or biological priors.

Across breast cancer, human lymph node, and mouse brain tissues, PERSIST consistently recovered biologically interpretable organization while adapting to fundamentally different tissue architectures. Tumor samples exhibited mixed and gradient-associated structure, lymph node tissue revealed discrete immune compartments, and brain tissue displayed rigid laminar and regional organization. These results demonstrate that the proposed framework generalizes across diverse spatial regimes without imposing a single organizational model.

Beyond spatial transcriptomics, the formulation presented here provides a general strategy for topology-informed multiscale analysis of spatially embedded molecular data. Future work will focus on both methodological extensions and theoretical analysis, including systematic benchmarking against existing spatial segmentation methods to quantify robustness with respect to scale selection, boundary fidelity, and subsampling. From a topological perspective, the framework naturally extends to higher-dimensional homology groups, enabling analysis of volumetric tissue organization using H_2 in three-dimensional spatial transcriptomics data.

Several methodological components also admit refinement. Alternative scalar field constructions beyond the first principal component may better capture complex transcriptional structure while preserving geometric stability, and multiparameter persistent homology offers a principled extension for jointly analyzing spatial scale, temporal dynamics, or multimodal data. In parallel, future work will investigate theoretical stability guarantees for topology-informed scale selection and explicit characterization of boundary and transition regions that emerge between stable domains.

Appendix A. Supplementary Figures

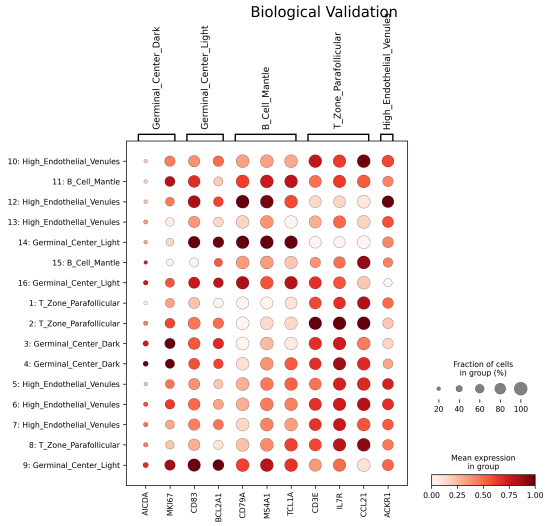

FIGURE 1. Human lymph node gene enrichment.

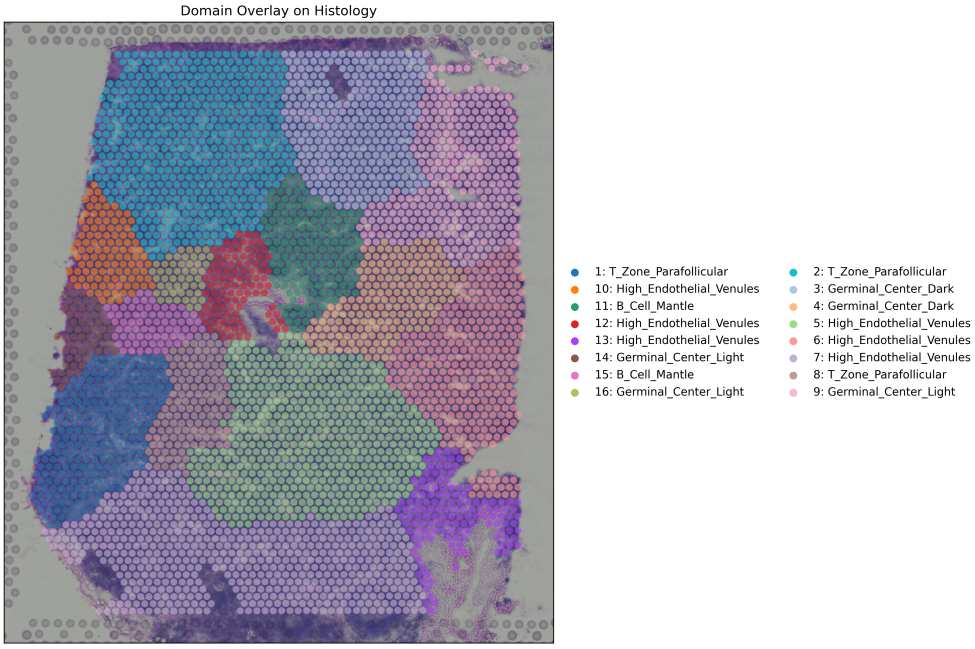

FIGURE 2. Human lymph node segmentation

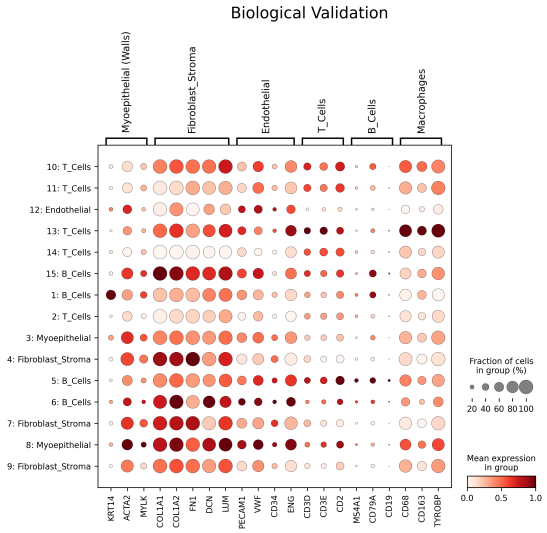

FIGURE 3. Breast cancer gene enrichment.

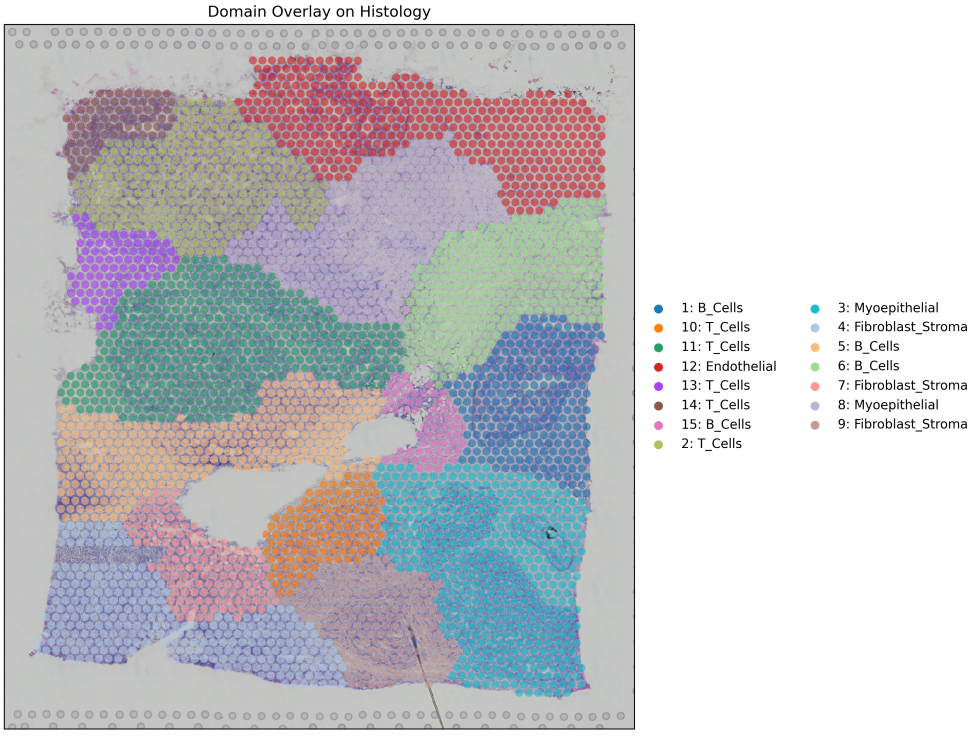

FIGURE 4. Breast cancer segmentation

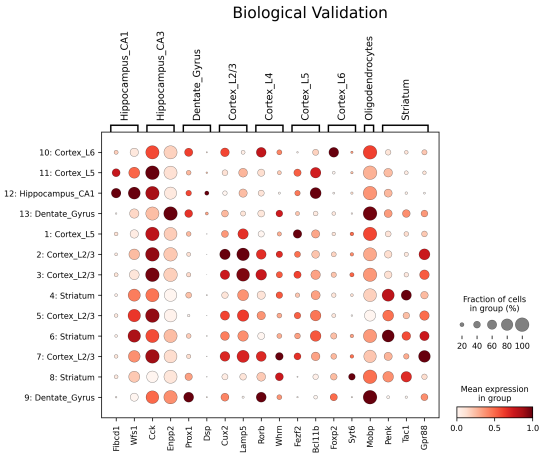

FIGURE 5. Mouse brain gene enrichment.

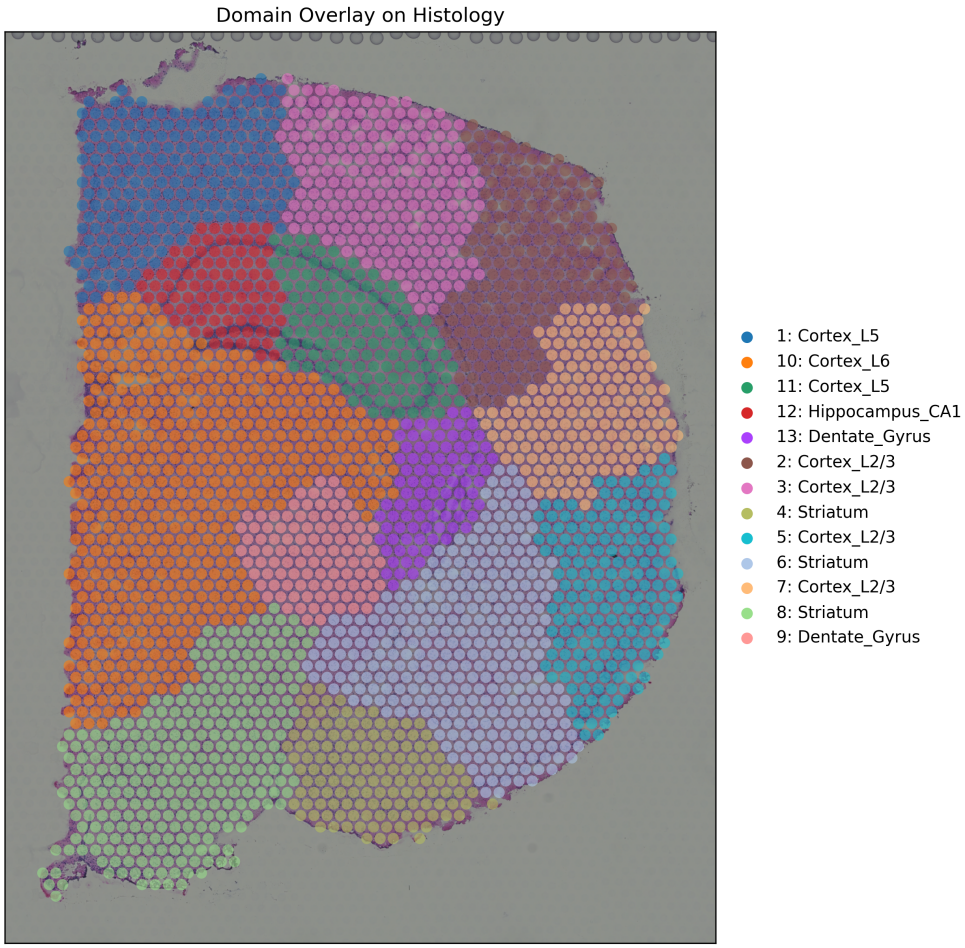

FIGURE 6. Mouse brain segmentation

References

1. Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1–2), 17–23. https://doi.org/10.1093/biomet/37.1-2.17
2. Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer. https://doi.org/10.1007/b98835
3. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33, 495–502. https://pmc.ncbi.nlm.nih.gov/articles/PMC4430369/
4. Edelsbrunner, H., & Mücke, E. P. (1994). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 40(1), 20–32. https://doi.org/10.1109/18.272978
5. Edelsbrunner, H., Letscher, D., & Zomorodian, A. (2002). Topological persistence and simplification. *Discrete & Computational Geometry*, 28, 511–533. https://doi.org/10.1007/s00454-002-2885-2
6. Zomorodian, A., & Carlsson, G. (2005). Computing persistent homology. *Discrete & Computational Geometry*, 33, 249–274. https://doi.org/10.1007/s00454-004-1146-y
7. Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45, 61–75. https://doi.org/10.1090/S0273-0979-07-01191-3
8. Hirani, A. N. (2003). *Discrete Exterior Calculus*. PhD thesis, California Institute of Technology. https://resolver.caltech.edu/CaltechETD:etd-05062003-135650
9. Desbrun, M., Kanso, E., & Tong, Y. (2005). Discrete differential forms for computational modeling. *SIGGRAPH Course Notes*. https://doi.org/10.1145/1198555.1198561
10. Chung, F. R. K. (1997). *Spectral Graph Theory*. American Mathematical Society. https://doi.org/10.1090/cbms/092
11. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788–791. https://doi.org/10.1038/44565
12. Forman, R. (1998). Morse theory for cell complexes. *Advances in Mathematics*, 134, 90–145. https://doi.org/10.1006/aima.1997.1650
13. Chambolle, A., & Pock, T. (2011). A first-order primal–dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40, 120–145. https://doi.org/10.1007/s10851-010-0251-1
14. Hein, M., Audibert, J.-Y., & von Luxburg, U. (2007). Graph Laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8, 1325–1368. https://www.jmlr.org/papers/v8/hein07a.html
15. Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers. https://doi.org/10.1007/978-1-4757-6465-9
16. Meyer, F. (1994). Topographic distance and watershed lines. *Signal Processing*, 38, 113–125. https://doi.org/10.1016/0165-1684(94)90060-4
17. Edelsbrunner, H., & Harer, J. (2010). *Computational Topology: An Introduction*. American Mathematical Society. https://doi.org/10.1090/amsip/001
18. Edelsbrunner, H., Kirkpatrick, D., & Seidel, R. (1983). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4), 551–559. https://doi.org/10.1109/TIT.1983.1056714
19. Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46, 255–308. https://doi.org/10.1090/S0273-0979-09-01249-X
20. Chazal, F., & Michel, B. (2017). An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence and Applications*, 2017, 1–36. IOS Press. https://doi.org/10.3389/frai.2021.667963
21. Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., et al. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Research*, 47(D1), D721–D728. https://doi.org/10.1093/nar/gky900
22. Franzén, O., Gan, L.-M., & Björkegren, J. L. M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, 2019. https://doi.org/10.1093/database/baz046
23. Maria, C., Boissonnat, J.-D., Glisse, M., & Yvinec, M. (2014). The GUDHI library: Simplicial complexes and persistent homology. In *International Congress on Mathematical Software* (pp. 167–174). Springer. https://doi.org/10.1007/978-3-662-44199-2_28
24. Wolf, F. A., Angerer, P., & Theis, F. J. (2018). Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), 15. https://doi.org/10.1186/s13059-017-1382-0
25. Pliškevičius, D., et al. (2021). Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 18(8), 853–855. https://doi.org/10.1038/s41592-021-01264-7
26. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html
27. Moses, L., & Pachter, L. (2022). Museum of spatial transcriptomics. *Nature Methods*, 19(5), 534–546. https://doi.org/10.1038/s41592-022-01409-2
28. Cable, D. M., Murray, E., Zou, L. S., Goeva, A., Macosko, E. Z., Chen, F., Irizarry, R. A., et al. (2022). Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 40(4), 517–526. https://doi.org/10.1038/s41587-021-00830-w

29. Bhowmick, N. A., Neilson, E. G., & Moses, H. L. (2004). Stromal fibroblasts in cancer initiation and progression. *Nature*, 432, 332–337. https://doi.org/10.1038/nature03096
30. Quante, M., Tu, S. P., Tomita, H., Gonda, T., Wang, S. S. W., Takashi, S., Baik, G.-H., Shibata, W., Diprete, B., Betz, K. S., et al. (2011). Bone marrow-derived myofibroblasts contribute to the mesenchymal stem cell niche and promote tumor growth. *Cancer Cell*, 19, 257–272. https://doi.org/10.1016/j.ccr.2011.01.020
31. DeFelipe, J., Alonso-Nanclares, L., & Arellano, J. I. (2002). Microstructure of the neocortex: comparative aspects. *Journal of Neurocytology*, 31(3–5), 299–316. https://doi.org/10.1023/A:1024130211265
32. Costa, A., Kieffer, Y., Scholer-Dahirel, A., Pelon, F., Bourachot, B., Cardon, M., Sirven, P., Magagna, I., Fuhrmann, L., Bernard, C., *et al.* (2018). Fibroblast heterogeneity and immunosuppressive environment in human breast cancer. *Cancer Cell*, 33(3), 463–479.e10. https://doi.org/10.1016/j.ccell.2018.01.011
33. Zhao, E., Stone, M. R., Ren, X., Guenthoer, J., Smythe, K. S., Pulliam, T., Williams, S. R., Uytingco, C. R., Taylor, S. E. B., Nghiem, P., et al. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39, 1375–1384. https://doi.org/10.1038/s41587-021-00935-2
34. Singhal, V., Kainmueller, D., Lickert, H., & Theis, F. J. (2024). BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis. *Nature Genetics*, 56, 431–441. https://doi.org/10.1038/s41588-024-01653-2
35. Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., Lee, E. B., Shinohara, R. T., & Li, M. (2021). 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. https://doi.org/10.1038/s41592-021-01255-8
36. Sun, S., Zhu, J., & Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, 17, 193–200. https://doi.org/10.1038/s41592-019-0701-7
37. Svensson, V., Teichmann, S. A., & Stegle, O. (2018). SpatialDE: identification of spatially variable genes. *Nature Methods*, 15, 343–346. https://doi.org/10.1038/nmeth.4636
38. Dries, R., Zhu, Q., Eng, C.-H. L., Sarkar, A., Bao, F., George, R. E., Pierson, E., Cai, L., & Yuan, G.-C. (2021). Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology*, 22, 78. https://doi.org/10.1186/s13059-021-02286-2
39. Beamer, P., & Cang, Z. (2025). PHD-MS: Multiscale domain identification for spatial transcriptomics via persistent homology. arXiv:2511.08411 [q-bio.QM]. https://doi.org/10.48550/arXiv.2511.08411

[1] MSc - Data Science and Artificial Intelligence - Université Côte d'Azur
Email address: gaurav.khanal@etu.univ-cotedazur.fr