Nucleic Acids Research, 2026, **54**, gkag434
https://doi.org/10.1093/nar/gkag434
Methods

Benchmarking sketching methods on spatial transcriptomics data

Ian K Gingerich [1,2,*], **Brittany A Goods** [2], **Hildreth Robert Frost** [1,*]

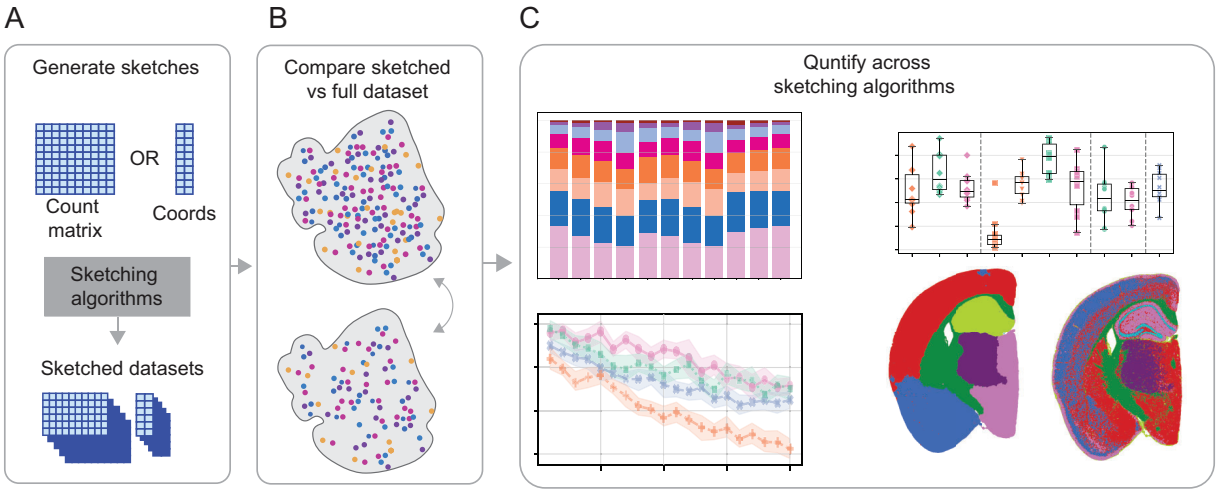
[1]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH 03755, United States
[2]Thayer School of Engineering, Dartmouth College, Hanover, NH 03755, United States

*To whom correspondence should be addressed. Email: ian.gingerich.gr@dartmouth.edu
Correspondence may also be addressed to Hildreth Robert Frost. Email: rob.frost@dartmouth.edu

Abstract

High-throughput spatial transcriptomics (ST) now profiles hundreds of thousands of cells or locations per section, creating computational bottlenecks for routine analysis. Sketching, or intelligent sub-sampling, addresses scale by selecting small, representative subsets. While effective for single-cell RNA sequencing data, existing sketching methods, which optimize coverage in expression space but ignore physical location, can introduce spatial bias when applied to ST data. To explore the impact of sketching on ST analysis, we systematically benchmarked uniform sampling, leverage-score sampling, Geosketch (minimax/Hausdorff), and scSampler (maximin) across multiple real ST datasets (mouse ovary, MERFISH brain, human breast cancer, lung) and simulations, using three input representations: Principle Component Analysis (PCA) embeddings, spatial coordinates, and spatially smoothed embeddings. We show that expression-only designs capture global transcriptomic heterogeneity but distort tissue architecture by over-sampling high-variability regions and under-sampling homogeneous areas. Coordinate-only sampling restores tissue coverage but misses transcriptional extremes. A simple spatially aware extension, computing leverage scores from a randomized singular value decomposition (SVD) basis smoothed by a spatial weights matrix, strikes a favorable balance, recovering rare cell states while maintaining uniform tissue coverage and avoiding edge effects. Across robust Hausdorff distances, clustering stability (Adjusted Rand Index), PCA loading drift, and local cell-type mean squared error (MSE), spatially smoothed leverage scores match or outperform alternatives. These results motivate joint spatial–transcriptomic sketching objectives to enable fast, unbiased analyses of increasingly large ST datasets.

Graphical abstract



Introduction

Sketching

As high-throughput genomic assays have driven datasets into the millions of observations and thousands of features, intelligent down-sampling, or sketching, has become a core strategy for scalable analysis, visualization, and method development. Rather than sampling uniformly at random, principled sketching techniques aim to preserve the geometric and statistical structure of the data to maintain coverage of rare states, boundary regions, and transitional trajectories [1–3]. Design-based (explicit sampling-design) approaches formalize this goal: minimax criteria (often expressed via Haus-

Received: September 8, 2025. **Revised:** April 7, 2026. **Accepted:** April 15, 2026

dorff distance) choose subsets that minimize the worst-case distance from any point to the sketch, controlling coverage gaps, whereas maximin criteria maximize the minimum pairwise distance among selected points, promoting well-spread and nonredundant samples [4]. In addition to influence-based (importance-weighted or leverage/score-driven) schemes such as leverage score sampling, these strategies provide general-purpose recipes for constructing small, representative sketches that support exploration, inference, and benchmarking across large-scale datasets [5].

Sketching for single-cell RNA sequencing data

Single-cell RNA sequencing (scRNA-seq) datasets are increasingly analyzed with "sketching" or intelligent sub-sampling: selecting a representative subset of cells that preserves transcriptomic diversity while dramatically reducing computational load. A naive approach is uniform random sampling, which is simple and unbiased with respect to sampling probability but often fails to capture rare cell types and transitional states, especially when data are highly imbalanced. Uniform sampling also tends to over-represent dense regions in the gene-expression manifold and under-represent sparsely populated regions, leading to loss of important biological variation. Additionally, it has been demonstrated that over-sampled cell types tend to be over clustered i.e. a single biological population is artificially split into multiple small, highly similar clusters driven by local density fluctuations, technical noise, or batch effects, while under-sampled cells are under clustered i.e. distinct rare or transitional populations are merged into a single broad cluster because there are too few points to resolve clear boundaries [6]. Down-sampling can also improve the signal-to-noise ratio in a dataset by reducing random correlations caused by variation in data quality and read-mapping bias [7, 8]. To address these issues, existing nonuniform methods seek to sample evenly across transcriptomic space: they identify or construct a low-dimensional representation (e.g. PCA, latent embeddings), quantify each cell's "coverage" or influence within that space, and then preferentially sample points that increase coverage of underrepresented regions while de-emphasizing redundant points in dense clusters. In scRNA-seq, such approaches, including leverage score-based sampling, minimax/Hausdorff-based geometric sketching, and maximin designs, have become standard tools for creating compact yet faithful summaries of large datasets [5, 9–15].

Sketching for ST data

High-resolution array-based platforms such as 10x Genomics Visium HD produce measurements at millions of subcellular-sized locations per slide; in practice, data are commonly aggregated into roughly 300 000–600 000 spatial bins to balance resolution and computational tractability, with each bin carrying transcriptome-wide counts for 20 000+ genes. *In situ* hybridization–based platforms such as 10x Genomics Xenium routinely profile on the order of 200 000–500 000 cells per tissue section, with larger studies exceeding a million cells, across targeted panels of 5000 genes. As these technologies mature, multi-section studies and atlases now comprise hundreds-of-thousands to tens-of-millions of spatial locations, each with high-dimensional gene expression and precise spatial coordinates [16–26]. While the scale of these datasets promises unprecedented views into tissue architecture, it also creates a big-data bottleneck: standard exploratory workflows, clustering, visualization, neighborhood enrichment, spatial domain detection, become prohibitively slow, and memory-intensive without principled sub-sampling strategies that respect both transcriptomic and spatial structure.

Despite their successful use for scRNA-seq analysis, current sketching algorithms operate solely on the gene expression manifold and ignore physical location. One widely used strategy, implemented in Seurat [5], computes approximate statistical leverage scores for each cell using an efficient randomized algorithm. These leverage scores are then converted into sampling probabilities for sketching. This approach is motivated by the fact that cells with high leverage are those that exert strong influence on the embedding and higher sampling probabilities for these cells enriches for transcriptomic diversity [5, 11]. Geometric sketching methods further formalize the objective of achieving approximately uniform coverage of the data manifold by controlling point-wise approximation error and reducing coverage gaps across regions of varying density. Geosketch [10] and Hopper [12] cast sketch selection as a minimax (Hausdorff) design problem: choose a subset S of size k that minimizes the maximum distance from any point in the dataset to its nearest neighbor in S. Concretely, letting X be all points (cells/locations) in an embedding and $S \in X$ the sketch, the (directed) Hausdorff distance from X to S is the worst-case nearest-neighbor distance, which measures the largest "hole" in coverage. Minimizing this objective ensures that every region in transcriptomic space is close to at least one selected point, limiting coverage gaps that would otherwise exclude rare or transitional cell states. Geosketch uses approximate farthest-point sampling over a grid to efficiently reduce this maximum distance, while Hopper accelerates the search and pruning to further tighten the Hausdorff criterion. A related approach, scSampler [9], adopts a maximin design: it selects points to maximize the minimum pairwise distance within the sketch, spreading selected samples as uniformly as possible across the high-dimensional expression space. Collectively, these methods target the same goal; uniform coverage of transcriptomic variability, via different optimization strategies and implementations.

We note that sketching is not the only paradigm for scaling analyses of large ST datasets. An alternative is to learn probabilistic low-dimensional representations (e.g. variational latent-variable models) [27–29] in which all cells/locations remain represented but computations proceed in a compressed space. Such approaches can preserve the full dataset and enable efficient downstream operations, and increasingly scalable stochastic training procedures continue to improve their applicability. However, in practice, for many ST datasets the computational bottleneck arises before downstream analysis: constructing neighbor graphs, running PCA/latent model training, or even materializing dense intermediate matrices can be challenging at the scale of hundreds of thousands to millions of locations, especially for atlas-scale studies. In these settings, sketching provides a complementary and often simpler strategy that reduces memory and runtime up front, enabling exploratory analysis pipelines that would otherwise be infeasible on typical hardware. Moreover, sketching has begun to enter standard ST workflows, for example, Seurat provides a sketch-based workflow for large spatial datasets (including Visium HD) [5, 11], highlighting the growing practical relevance of understanding how different sketching objectives affect spatial analyses. Importantly, sketching and probabilis-

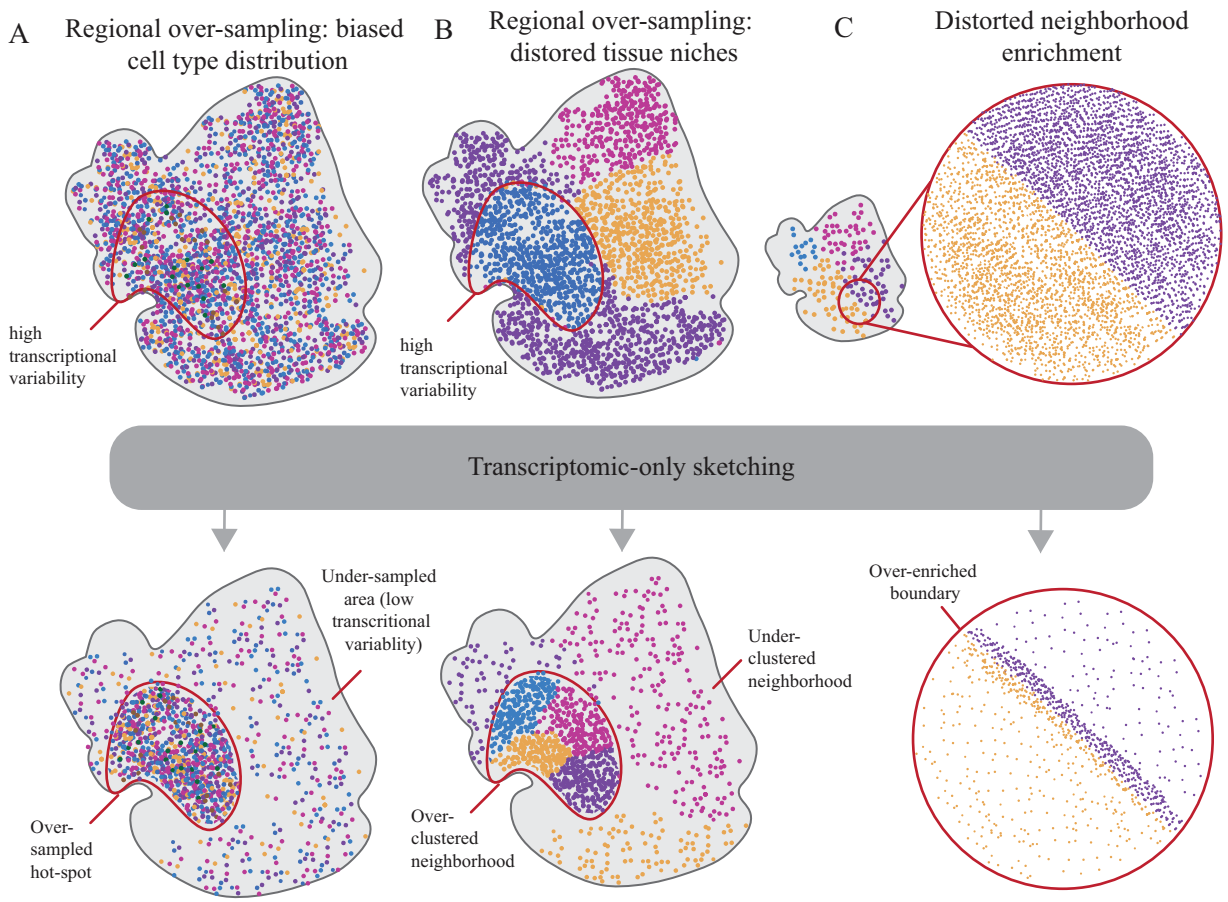


Figure 1. Examples of transcriptomic-based sketching bias. (**A**) Regional over-sampling leads to biased cell type proportions and over-representation in transcriptional hotspot. (**B**) Regional over-sampling leads to over clustered tissue niches in transcriptional hotspot. (**C**) Over-sampling at cell type or neighborhood boarder leads to exaggerated neighborhood enrichment.

tic embeddings are not mutually exclusive: sketches can be used to accelerate model fitting or to prototype analyses, with results subsequently projected or transferred back to the full dataset.

However, transcriptome-only sketching can distort spatial structure in ST data. If high-variability or high-leverage profiles cluster in a particular region of a tissue, a purely expression-based sketch will over-sample that region and under-sample others. In ST data, where neighborhood relationships, cell–cell contact patterns, and spatial gradients of expression are essential for interpretation, such imbalances can systematically bias downstream analysis. Over-representing a spatial hotspot inflates the apparent frequency of certain cell–cell adjacency and co-localization relationships, leading to biased estimates of neighborhood enrichment or depletion and misleading microenvironment niches (Fig. 1A). Methods that integrate ligand–receptor expression with spatial proximity will overestimate interactions prevalent in over-sampled regions and underestimate interactions in under-sampled regions, skewing inferred communication networks and hub cell types. Over-coverage of one region and under-coverage of others distorts estimated spatial trajectories and continuous expression gradients (e.g. along anatomical axes), flattening or exaggerating trends and misplacing boundaries between niches (Fig. 1B). Biased sketches also alter the apparent spatial distribution of gene expression and domain boundaries, potentially merging distinct regions or fragmenting coherent tissue compartments due to uneven sampling density (Fig. 1C). Because most downstream tasks rely on the same sampled subset, either directly or for building reference embeddings early spatial bias propagates through the entire pipeline, compounding errors in subsequent inferences and reducing reproducibility across datasets and laboratories.

To further explore these challenges, we benchmarked current sketching approaches on real and simulated ST datasets to evaluate their ability to preserve both transcriptomic heterogeneity and spatial organization. Our results highlight scenarios in which transcriptome-driven sketches distort tissue architecture and underscore the need for spatially informed sampling strategies. By systematically comparing existing methods on real spatial datasets, we lay the groundwork for adapting sketching algorithms to the unique challenges of ST data, coupling coverage in expression space with balanced coverage in physical space to maintain fidelity for downstream spatial analyses.

Methods

Evaluated ST datasets

We benchmarked four families of sketching methods (see Table 2) across eight public ST datasets spanning multiple platforms and spatial resolutions (MERFISH, Xenium, Visium HD) and six different simulation models. For real data, we performed quality control by excluding cells/locations with fewer

Table 1. Dataset summary

Tissue	Species	Technology	# locations	# genes	Study
Sagittal brain	Mouse	MERFISH	81 452 cells	1122	Zhang *et al.*, 2023 [26]
Ovary	Mouse	MERFISH	43 038 cells	228	Huang *et al.*, 2024 [30]
Coronal brain	Mouse	10x Visium HD	393 434 locations[a]	19 059	10x Genomics
Mouse embryo	Mouse	10x Visium HD	513 290 locations[a]	25 028	10x Genomics
Human ovary	Human	10x Visium HD	439 300 locations[a]	29 936	10 Genomics
Lung	Human	10x Xenium	295 883 cells	541	10x Genomics
Breast cancer	Human	10x Xenium	565 916 cells	541	Bhuva *et al.*, 2024
Mouse pup	Mouse	10x Xenium	1219 863 cells	5101	10x Genomics
Visium-like complex	—	Simulated	100 000 locations	500	—
Visium-like radial	—	Simulated	100 000 locations	500	—
Visium-like stripes	—	Simulated	100 000 locations	500	—
Xenium-like complex	—	Simulated	100 000 cells	500	—
Xenium-like complex	—	Simulated	100 000 cells	500	—
Xenium-like stripes	—	Simulated	100 000 cells	500	—

[a]8 μm binned locations.

than 100 detected counts and removing outliers exceeding five times the mean absolute deviation. After filtering, data were library-size scaled and log-normalized, followed by truncated principal component analysis to 50 components for downstream sketching and evaluation.

Real ST datasets included: a mouse ovary MERFISH dataset with eight ground-truth cell-type labels from the original study [30]; a human breast cancer 10x Xenium dataset with histology-derived annotations [31]; a human lung 10x Xenium dataset without ground-truth labels, for which we generated full-data reference clusters; a sagittal mouse brain MERFISH dataset [26] with region, class, and type annotations, where Adjusted Rand Index (ARI) was computed against cell-class labels; a coronal mouse brain 10x Visium HD dataset spanning cortex, subcortex, and hippocampus, evaluated against full-data graph-based clusters due to the lack of ground truth; a mouse embryo 10x Visium HD dataset with manually annotated cell lineages; a human ovarian cancer 10x Visium HD dataset with manually annotated cell lineages; and a whole mouse pup 10x Xenium dataset with manually annotated cell lineages.

For datasets with multiple levels of annotation available (mouse brain, embryo, and pup), we used higher-level cell type categories rather than fine-grained annotations. This choice was intentional: for several datasets, no reliable ground-truth fine-grained annotation was available, and we did not want to impose overly specific labels without sufficient biological or domain expertise. Using coarser cell type labels ensures a more defensible and consistent benchmark across all datasets. We note that the majority of our metrics, including the clustering comparison framework, are expected to be robust to annotation resolution. For ARI specifically, the number of clusters was set *a priori* before comparison to the full dataset, ensuring the benchmarking was not driven by post hoc tuning of resolution.

Simulated data comprised six datasets (three Visium-like on a grid, three Xenium-like with randomly distributed coordinates) with complex, radial, and striped spatial layout (see Table 1 for details). Simulated ST data was generated with SRTsim [32] using a fixed seed and included six label classes (A–F) with specified high/low-signal and noise gene counts, dispersion, baseline mean expression, zero-inflation, and distinct class-specific log-fold changes to induce separable spatial domains (see Table 1). See Extended data Section 1.1 for dataset-specific preprocessing details.

Evaluated sketching algorithms

We compared four sketching algorithms: uniform sampling, Leverage score [5], Geosketch [10], and scSampler [9]. Each was applied to three input representations when applicable: (i) transcriptomic embeddings derived from PCA (50 components), (ii) raw spatial coordinates, and (iii) spatially smoothed embeddings that incorporate local neighborhood composition. Uniform sampling selected indices without replacement using equal probabilities and distinct random seeds per replicate. Leverage score methods prioritize statistically influential observations either from the original expression space (variable feature selection as appropriate) or from a low-rank embedding with sampling probabilities were proportional to normalized leverage values. For the standard leverage-score computation, we follow Seurat's implementation, which uses random projection followed by QR decomposition to approximate leverage scores efficiently on large matrices. For the spatially smoothed leverage-score variant (our simple extension), we first compute a low-rank randomized SVD (rSVD) of the cell-by-gene matrix to obtain U (left singular vectors). We then apply spatial smoothing by multiplying U by a precomputed spatial weights matrix W, yielding $\tilde{U} = WU$. Leverage scores are computed from the smoothed basis $\tilde{U}$ (row-wise squared norms), normalized to probabilities, and sampled without replacement. Geometric sketching implements a minimax (Hausdorff coverage) objective to distribute samples so that all regions of the embedding are close to at least one selected point. Maximin designs aim to maximize the minimum pairwise distance among selected points, yielding a uniformly spread subset in high-dimensional space. Spatially smoothed variants use a precomputed spatial weights matrix to smooth the low-dimensional representation prior to sketch selection, thereby encouraging balanced sampling across physical space while preserving major transcriptomic variation. All sketches were drawn without replacement at pre-specified sampling fractions; procedures with inherent randomness were repeated 10 times with different seeds to assess stability. See Table 2 for all listed algorithms. See Extend data Section 1.2 for details on sketching algorithm design and implementation.

Evaluation metrics

We evaluated each sketch along complementary transcriptomic and spatial criteria. Robust (partial) Hausdorff distance quantified worst-case nearest-neighbor discrepancies

Table 2. Sketching algorithms evaluated and their configurations

Algorithm	Input representation	Design/criterion	Language	Study
scSampler	PCA matrixSpatial coordinates Spatially smoothed PCA	Maximin	Python	Song *et al*., 2022 [9]
Geosketch	PCA matrixSpatial coordinates Spatially smoothed PCA	Minimax	Python	Hie *et al*., 2019 [10]
Leverage score	Cell-by-gene matrixPCA matrix Spatially smoothed U	Statistical influence	R	Clarkson & Woodruff, 2017 [5]
Uniform	Cell index	Random sampling	Python/R	—

between the sketch and full data, computed both in expression space (e.g. PCA) and in physical coordinate space. We used a quantile-based, outlier-robust formulation to reduce sensitivity to extreme points and employed approximate neighbor search to scale to hundreds of thousands of locations. The ARI assessed concordance between sketch-derived clusters and reference labels: ground-truth annotations when available (MERFISH ovary, MERFISH brain class, Xenium breast cancer), manually annotations (Xenium mouse pup, Visium HD ovarian cancer, and mouse embryo), or full-data graph-based clusters otherwise (Xenium lung, Visium HD brain). To evaluate preservation of global variance structure, we computed a PCA projection difference by comparing the sketch's projections onto principal components learned from the full dataset versus components learned from the sketch; lower values indicate more consistent low-dimensional structure. To measure structural compositional fidelity we computed global spatial autocorrelation (Moran's I) on the cell type or cluster labels from the full and sketched datasets and compared the $\log_2$ fold change ($\log_2$FC) between the full and sketched dataset values, summarizing which cells/labels has an increased or decreases spatial association. Finally, to investigate preservation of neighborhood group structure, we identified both small and large spatial domains or niches in the full and sketched datasets and computed the ARI between the sketched and full spatial domain labels. See Extended data Section 1.3 for detailed metric computations. To assess the performance of each algorithm across the datasets evaluated, we aggregated results using a rank-based comparison at a fixed sampling fraction (0.1). For each dataset and each evaluation metric, methods were ranked according to the direction of improvement (higher-is-better for ARI-based metrics; lower-is-better for distance- or error-based metrics such as Hausdorff distance and PCA projection difference). Because many metrics differed only marginally between methods, we allowed ties by rounding metric values prior to ranking so that negligible differences were treated as equivalent. We then computed rank-sums across datasets for each metric to provide an overall comparison of method performance for each criterion, and performed these aggregations separately for real and simulated benchmark datasets. This rank-sum approach emphasizes relative performance while reducing sensitivity to dataset-specific scaling of metrics (Supplementary Information 1.3.6).

For each sketched dataset, we computed a variety of evaluation metrics to quantify the effectiveness of the associated sketching algorithm. Previous methods have used the transcriptomic Hausdorff distance as their primary or sole method for evaluating sketch quality, however, we feel that this metric is insufficient to characterize the spatial characteristics of the sketched data. To this end, we employed additional metrics to better assess sketch quality [9, 10]. These metrics include the robust Hausdorff distance, ARI between the sketch-based clustering solution and ground truth annotations, mean difference between PCA loading vectors, global Moran's I on cell type or cluster labels, and ARI between spatial domain labels (see Fig. 2B and C). This evaluation framework provides a framework to understand the impact of different downsampling strategies on the analysis of ST data, facilitating informed decisions in experimental design and data interpretation. See Section 2.3 for details on evaluation metric calculations. In brief:

- Robust/partial Hausdorff distance (expression and spatial). Quantifies the worst-case nearest-neighbor discrepancy between the sketch and the full dataset, computed separately in transcriptomic space and in physical space. Lower values indicate better coverage (fewer "holes") and greater fidelity of the sketch to the full data, with the partial version reducing sensitivity to outliers.
- ARI. Measures agreement between cluster labels obtained from the sketched data and ground-truth (or full-data-derived) labels, adjusted for chance. Scores range from −1 to 1, with higher values indicating better recovery of biological group structure.
- PCA projection difference. Compares how the sketched data project onto PCs learned from the full dataset versus PCs learned from the sketch itself. Smaller mean distances imply that the sketch preserves the dominant variance structure of the full data and yields consistent low-dimensional representations.
- Spatial autocorrelation (Moran's I) on celltype labels. Quantifies the degree to which identical (or similar) cell-type/cluster labels are spatially clustered versus dispersed across the tissue. We computed Moran's *I* on the full dataset and on each sketch and compared values via $\log_2$FC. Values close to 0 indicate near-random spatial arrangement, positive values indicate spatial clustering (patchiness/compartmentalization), and negative values indicate spatial dispersion. Smaller absolute $\log_2$FC between the sketch and full data indicates better preservation of global spatial organization of labels.
- Spatial domain preservation. Assesses whether sketching preserves higher-order tissue organization into spatial domains or niches. We identified both small and large spatial domains in the full and sketched datasets (using the same domain-calling procedure with two sets of parameters) and measured agreement between sketch-derived and full-data domain assignments using the ARI. Higher ARI values indicate that the sketch retains neighborhood-scale domain structure and boundaries present in the full tissue.

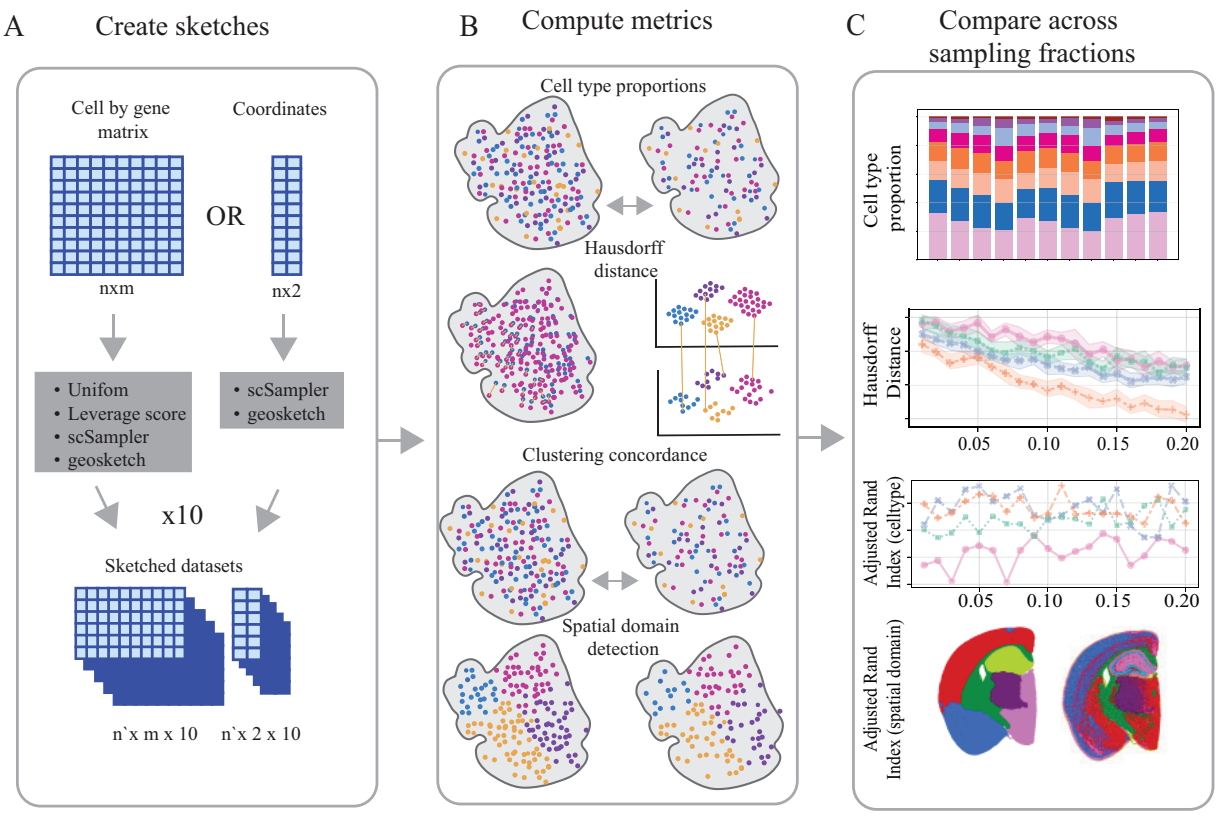


Figure 2. Analysis pipeline. (**A**) Sketching protocol. (**B**) Sketched dataset evaluation metrics. (**C**) Metric quantifica-tion across sampling fractions.

Results

Across eight diverse tissues, sub-sampling purely on coordinates ensures uniform spatial coverage but pays a price in transcriptomic diversity; sketching purely on gene expression values ensures that transcriptomic extremes are included but biases sampling toward biologically variable sub-regions. These findings motivate the development of sketching methods that explicitly weight the tradeoff between spatial uniformity and transcriptomic diversity.

Sketching result summary

We summarize overall sketching performance across metrics using an aggregated rank-based framework (Fig. 3B and D). For each dataset (Fig. 3A) and metric, methods were ranked by performance (with ties allowed), and ranks were then summed across datasets to obtain a metric-specific rank-sum for each algorithm; lower rank-sums indicate

better overall performance for that metric. These rank-sums are visualized as heatmaps for real datasets (Fig. 3B) and simulated datasets (Fig. 3D), with hierarchical clustering applied to rows and columns to reveal recurring patterns in metric performance and method behavior (see Extended data Section 1.3.6 for details).

For the real datasets, we aggregated ranks across transcriptomic and coordinate-space Hausdorff distances, ARI for small and large spatial domains, ARI for transcriptomic clustering, changes in global spatial autocorrelation ($\log_2$FC in Moran's *I* computed on cell-type or cluster labels), and differences in PCA solutions (MSE between PCA representations). Across methods, transcriptomic Hausdorff distance was consistently minimized, resulting in relatively similar ranks for this metric (Fig. 3B, top row). In contrast, clear separations emerged for spatially informed criteria: uniform sampling and leverage-score-based approaches generally outperformed geometric sketching methods for coordinate-space Hausdorff distance, spatial domain preservation (small and large domain ARI), and clustering concordance (cluster ARI). Uniform sampling also performed best for preservation of label-based Moran's I, which is expected given that it samples without introducing additional structure-dependent bias, while the remaining methods showed broadly similar performance for this metric. Coordinate-only sketching was excluded from the aggregated heatmap because it served primarily as a comparison baseline and is not directly applicable to grid-like platforms such as Visium HD, where spatial locations are already uniformly arranged.

For simulated datasets, we focused on the subset of metrics available and most informative under the simulation design: transcriptomic and coordinate Hausdorff distances, cluster ARI, and PCA-solution differences (Fig. 3D). While patterns were less pronounced than in the real datasets, uniform sampling and leverage-based methods again tended to outperform geometric sketching, with greater variability across metrics. We attribute the reduced separation in part to limitations of the simulation framework, which modeled expression over a smaller feature space (500 genes) than the higher-dimensional profiles present in the real datasets, thereby diminishing the degree to which transcriptome-driven manifold geometry differentiates sketching objectives.

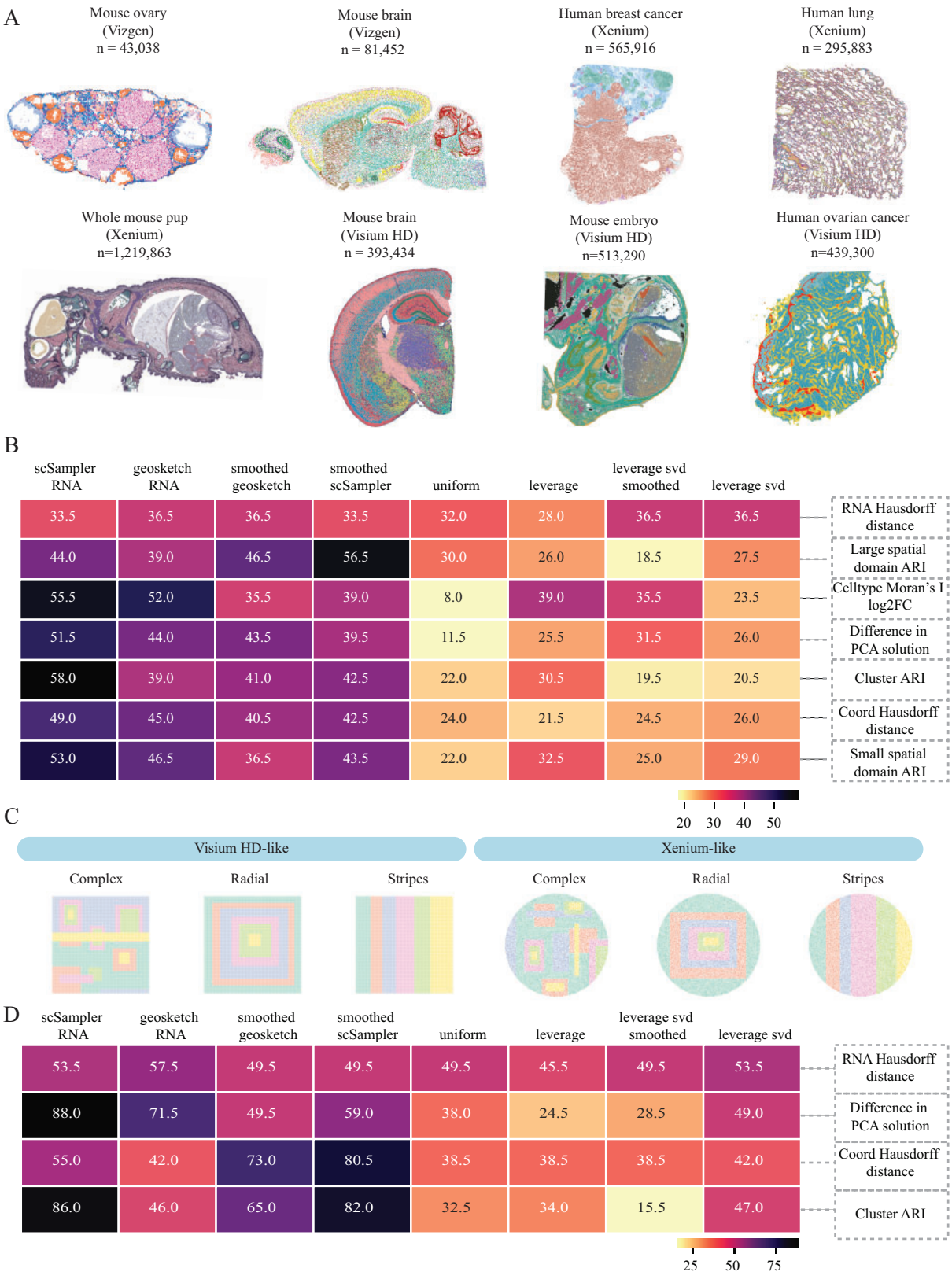


Figure 3. Overall sketching performance for 0.10 sampling fraction across datasets. (**A**) Spatial scatter plots of real datasets colored by cell type or cluster label. (**B**) Heatmap of rank-sums for each method aggregated by metric across all real world datasets. Low rank indicates best performance for that metric. (**C**) Spatial scatter plots of simulated Visium HD-like and Xenium like datasets. (**D**) Heatmap of rank-sums for each method aggregated by metric across all simulated datasets. Low rank indicates best performance for that metric.

Cell type proportion conservation

Across datasets, we assessed how well each sketching strategy preserves the distribution of cell-type (or cluster) labels by comparing label proportions in the full dataset to those observed in the sketch (Fig. 4). Several consistent patterns emerged. First, coordinate-based sketching methods (e.g. scSampler or geoSketch applied to physical coordinates) produced proportion profiles that closely tracked uniform random sampling (Fig. 4A–E). This behavior is expected because these approaches optimize spatial coverage objectives closely related to Hausdorff-type criteria, and therefore tend to sample the tissue geometry in a manner similar to uniform sampling. In some datasets, modest deviations were observed for labels concentrated at tissue boundaries (e.g. epithelial populations in the mouse ovary), where boundary-restricted cell types can become slightly over-represented due to the emphasis on geometric coverage; nevertheless, coordinate-based sketches largely serve as useful baselines for comparison rather than methods intended to prioritize transcriptomic structure.

In contrast, the largest shifts in label proportions arose when comparing leverage-score-based methods to transcriptome-driven geometric sketching. Spatial smoothing of the input data produced only modest changes within the same methodological family (e.g. leverage score versus spatially smoothed leverage score), consistent with smoothing acting primarily as a stabilizer rather than altering the sampling objective. Across datasets, leverage-score approaches generally maintained label proportions more faithfully, whereas geometric sketching on expression features (scSampler/geoSketch on RNA) sometimes substantially distorted proportions, particularly in the Visium HD and whole mouse pup datasets (Fig. 4E–H). For example, in the Visium HD ovarian cancer dataset, the Fibroblast CAF population increased from roughly 20% of the full dataset to over 70% of cells in the scSampler and geoSketch RNA sketches (Fig. 4H). A plausible explanation is that this population exhibits high transcriptional heterogeneity, causing geometric sketching methods to preferentially sample it to achieve coverage of the expression manifold. While such behavior may be desirable when the goal is to capture transcriptomic diversity, it can bias downstream analyses that rely on representative cell-type composition (e.g. estimates of cell-type abundance, comparisons of compositional changes across conditions, and niche frequency quantification).

These results emphasize that sketch selection should be guided by the intended downstream task. If preserving approximate cell-type proportions in the sketched dataset is important—and particularly if labels inferred from the sketch will be interpreted directly without being projected back to the full dataset—then leverage-based approaches provide a more conservative choice (Fig. 4). Conversely, if the primary objective is to capture rare transcriptional states or highly variable populations, transcriptome-driven geometric sketching may still be warranted, with the caveat that the resulting sketches may not be compositionally representative. Given that both leverage-based and geometric sketching methods are relatively fast to run, we recommend applying both in parallel for cell-type discovery and then projecting sketch-derived labels back onto the full dataset to reconcile differences and evaluate robustness, since accurate cell-type annotation is foundational to most ST workflows.

Hausdorff distances

At a fixed 10% sampling fraction, we compared sketch quality using robust (partial) Hausdorff distance computed in both transcriptomic space and physical coordinate space (Fig. 5). In transcriptomic space, differences between sketching strategies were modest across datasets: for any given dataset, the total spread in Hausdorff distances across methods was typically small (on the order of 1 unit or less), indicating broadly comparable coverage of the expression manifold at this sampling rate. Consequently, although Fig. 5A visualizes method-to-method differences, the absolute magnitude of separation is limited at 10% sampling. Consistent with this observation, larger divergences between algorithms became more apparent only at higher sampling fractions, as shown in the extended data figures reporting performance across sampling fractions (Extended data Figs 5–12, top panel). Because many practical applications of sketching in ST involve stronger downsampling than 50%, we emphasize that at commonly used low sampling fractions, transcriptomic Hausdorff distance alone is unlikely to provide a decisive criterion for choosing between leverage-based and geometric sketching approaches.

In contrast, coordinate-space Hausdorff distance revealed substantially larger performance differences and more clearly separated the methods (Fig. 5B). As expected, coordinate-based sketching produced the smallest coordinate Hausdorff distances due to explicitly optimizing spatial coverage; however, this objective does not necessarily align with preserving transcriptomic structure. Among transcriptome-informed methods, leverage-based approaches consistently achieved lower coordinate Hausdorff distances than geometric sketching methods across all Visium HD datasets, indicating improved spatial coverage of the tissue while still selecting points based on molecular structure. No dataset showed a clear coordinate-space advantage for geometric sketching, although the sagittal mouse brain and human lung datasets exhibited broadly similar performance across methods (Fig. 5B; Extended data Figs 5–12, second panel). Together, these results suggest that when spatial coverage is a priority, particularly in highly structured tissues, in samples with large variability in local cell density, or in settings where spatial organization and transcriptional heterogeneity are confounded, leverage-based sketching provides a robust default for minimizing spatial "holes" while maintaining comparable transcriptomic coverage.

Cluster concordance and spatial domain conservation

To investigate how sketching impacts downstream spatial analyses, we evaluated preservation of spatial domain structure by identifying tissue domains in the full and sketched datasets and quantifying concordance between them (Fig. 6). Spatial domains were inferred using the Randimized spatial PCA (RASP) workflow [33] (PCA-based smoothing followed by clustering) under two parameterizations chosen to capture either smaller, finer-grained niches (Fig. 6A) or larger, coarser tissue compartments (Fig. 6B; see Extended data Section 1.3.4 for details). For each dataset, we computed an ARI between the domain labels obtained from the sketch and those obtained from the corresponding full dataset. In parallel, we performed canonical graph-based clustering directly on each sketch (i.e. neighbor graph construction in transcriptomic

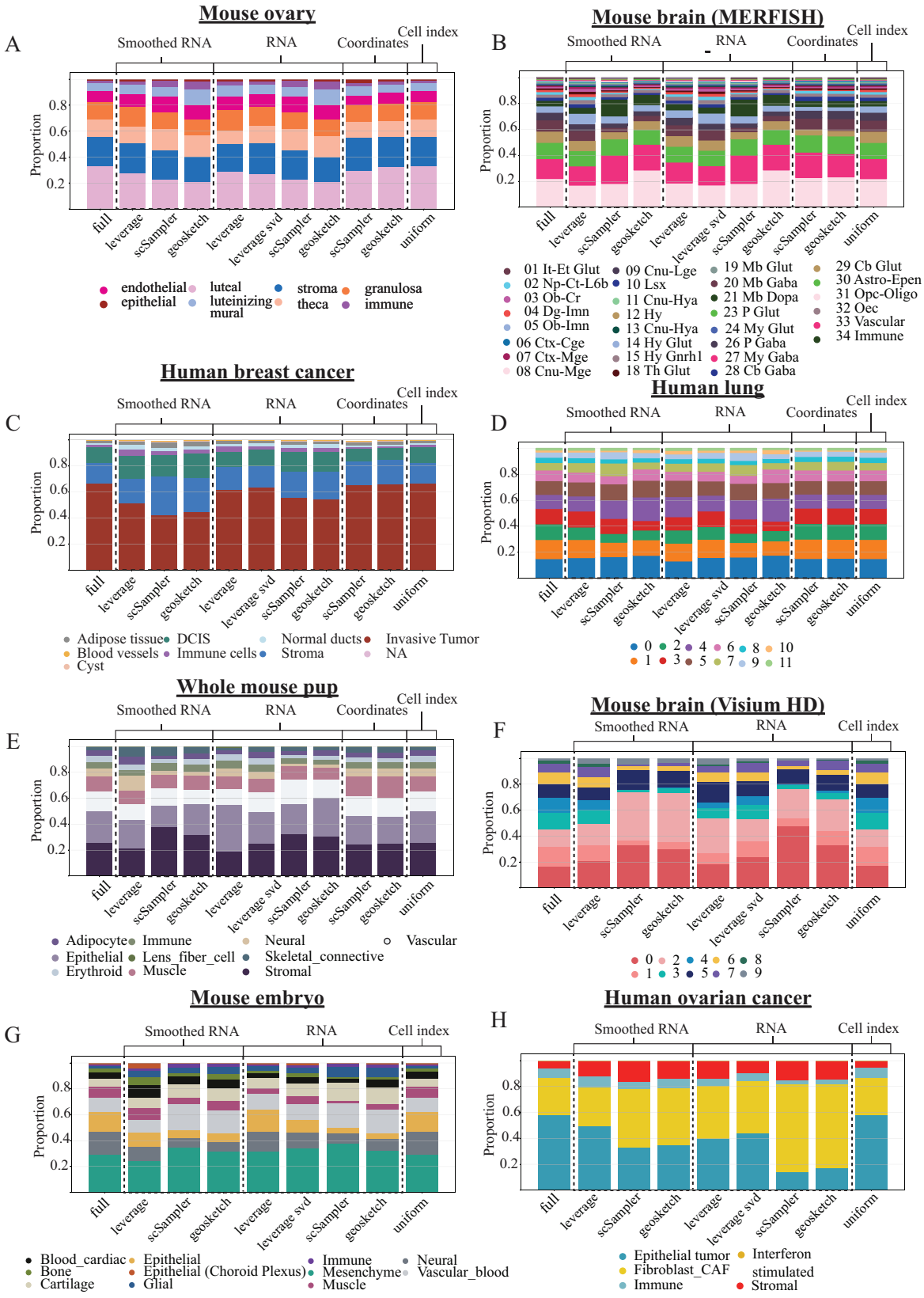


Figure 4. Retained cell type/cluster label proportions at 0.10 sketching fraction for (**A**) Merfish mouse ovary; (**B**) Merfish sagittal mouse brain; (**C**) Xenium human breast cancer; (**D**) Xenium human lung; (**E**) Xenium whole mouse pup; (**F**) Visium HD coronal mouse brain; (**G**) Visium HD mouse embryo; (**H**) Visium HD ovarian cancer.

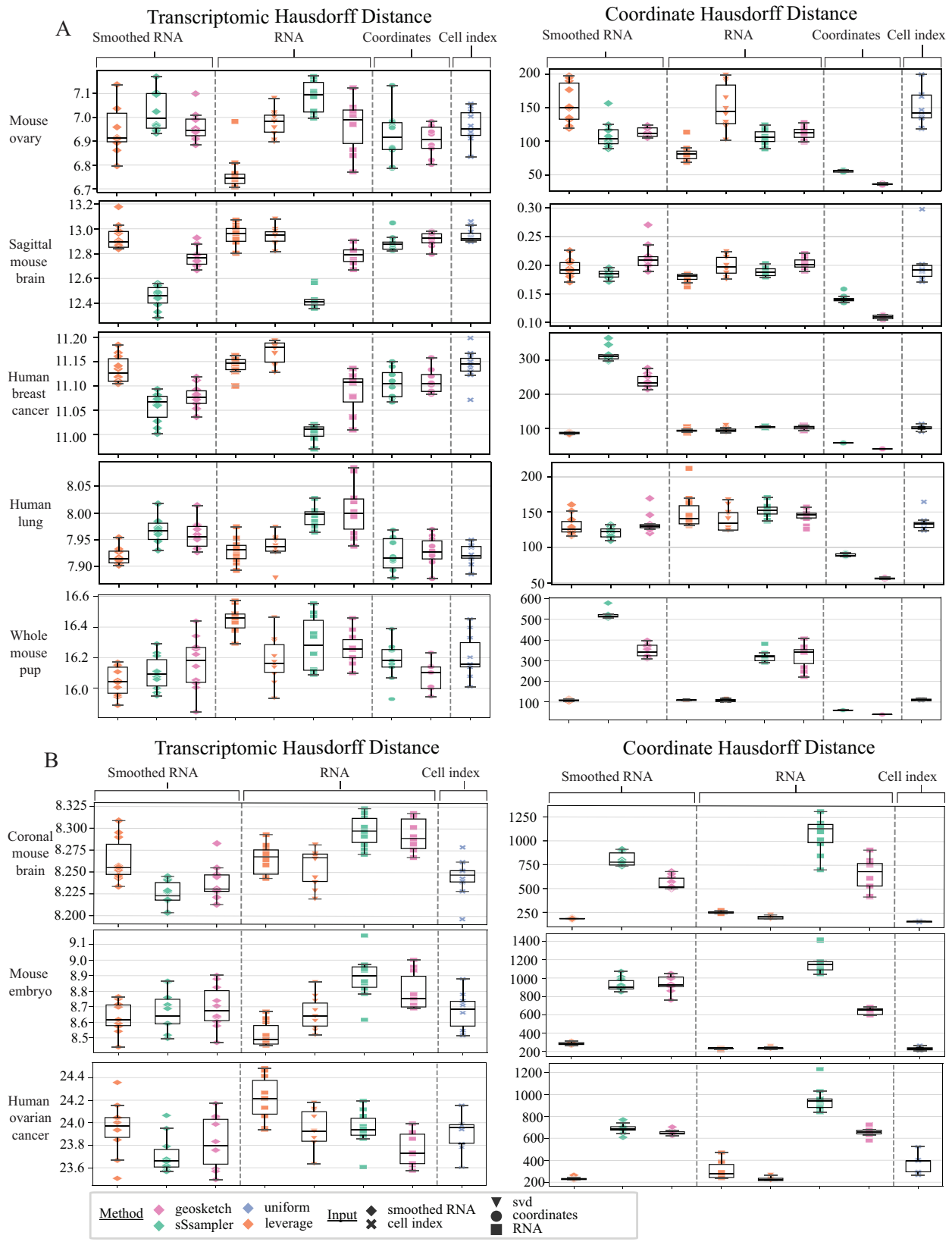


Figure 5. Quantification of transcriptomic and coordinate Hausdorff distance at 0.10 sampling fraction for real datasets. (**A**) Quantification of imaging based (Merfish, Xenium) dataset's Hausdorff distances. (**B**) Quntification of sequencing/spot based (Visium HD) dataset's Hausdorff distances. Each boxplot represents one sketching method, with individual points corresponding to results from 10 independent runs with different random seeds.

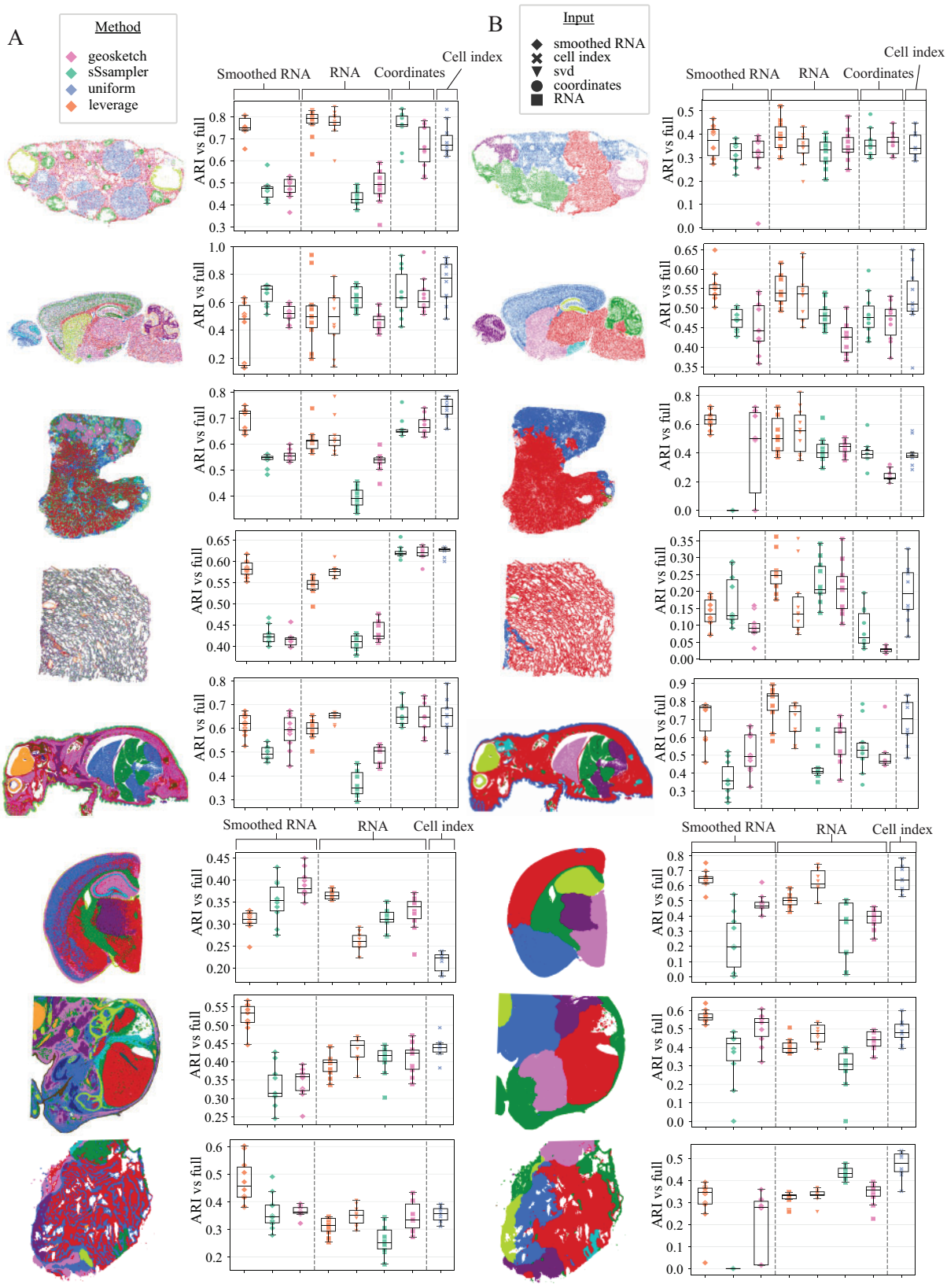


Figure 6. Spatial domain detection on sketched versus full datasets. (**A**) Small spatial domains plotted on full datasets (left), and ARI quantification for sketching algorithms. (**B**) Large spatial domains plotted on full datasets (left), and ARI quantification for sketching algorithms. Note that algorithms are separated by input type, indicated by dotted vertical lines, with spatially smoothed gene expression on the left, gene expression in the middle or middle left, coordinates on the middle right (for imaging based technologies only), and uniform on the far right. Each boxplot represents one sketching method, with individual points corresponding to results from 10 independent runs with different random seeds.

space followed by community detection, as commonly used in standard single-cell and spatial transcriptomics pipelines for defining clusters prior to cell-type annotation) and computed the ARI against the corresponding full-data cell type or cluster labels (Extended data Figs 5–12, second panel from bottom). This complementary analysis serves a different purpose than the RASP-based spatial domain comparison: whereas RASP explicitly integrates spatial context via smoothing and is designed to recover contiguous tissue domains or niches, graph-based clustering measures whether sketching preserves the intrinsic transcriptomic neighborhood structure that underlies routine downstream analyses such as clustering, differential expression, and cell-type annotation. Because these two workflows emphasize different signal components, spatial compartments versus expression-driven community structure, agreement under both provides a more complete picture of sketch fidelity and helps distinguish methods that preserve spatial domains from those that primarily preserve transcriptomic clusters.

At a 10% sampling fraction, leverage-score-based sketching methods generally preserved small spatial domains at least as well as, and often better than, geometric sketching approaches. In six out of eight datasets, leverage-based methods were comparable to or outperformed geometric methods in small-domain ARI, with the spatially smoothed leverage variant achieving the highest ARI among the four leverage-based strategies in the majority of cases (Fig. 6A). In contrast, spatial smoothing applied prior to geometric sketching did not yield a consistent benefit: depending on the dataset, smoothing sometimes improved and sometimes reduced domain concordance, suggesting that its effects are context-dependent rather than uniformly stabilizing.

For larger spatial domains, leverage-based methods again tended to match or exceed geometric sketching performance (seven of eight datasets; Fig. 6B), although we observed increased variability in ARI across repeated runs for many methods and datasets, indicating that at 10% sampling fraction the inferred domain solutions can be unstable. In some instances, spatial smoothing improved ARI (e.g. Visium HD brain and mouse ovary), but this trend was not consistent: in two datasets, smoothing reduced ARI for the leverage-based method, and only the ovarian cancer dataset showed a clear advantage for geometric sketching. Taken together, these results suggest that when spatial domain detection is a key downstream objective, leverage-based sketching provides a conservative default choice, a conclusion that is also reflected in the aggregated rank-sum comparisons (Fig. 3B), which summarize domain-preservation performance across datasets and metrics.

Cell type spatial structure preservation

We evaluated preservation of global cell-type spatial structure by computing Moran's I for each cell type (or cluster label) in both the full and sketched datasets, yielding a label-specific measure of spatial autocorrelation. For each label, we then quantified sketch-induced distortion as the $\log_2$FC between the Moran's I value in the sketch and the corresponding value in the full dataset. Extended data Fig. 1 illustrates this procedure using representative spatial scatter plots for selected labels, alongside method-specific $\log_2$FC values. As expected, uniform sampling produced the smallest deviations from the full dataset for this metric, consistent with its lack of structure-dependent sampling bias.

We also provide complete cell-type-level results across all eight datasets in Supplementary Figs S13 and S14, without applying minimum baseline autocorrelation filters. This comprehensive view reveals that signal in these metrics is primarily driven by cell types with inherently high baseline spatial autocorrelation—namely, those with strong intrinsic spatial organization in the full dataset. Cell types that are sparsely or uniformly distributed across the tissue exhibit low Moran's I values regardless of sketching method, reflecting their dispersed spatial distribution rather than poor sketch quality. Indeed, for datasets containing many uniformly distributed cell types, Moran's I values across sketching methods converge, suggesting limited meaningful variation to preserve for these populations. Consequently, spatial autocorrelation metrics are most informative for cell types with moderately to highly organized spatial structure; sparse populations should not be expected to show stable or differentiated Moran's I signatures across methods.

Dataset-wide patterns across all labels are summarized in Extended data Fig. 2A and B, and Supplementary Figs S13 and S14. To compare sketching methods systematically within each dataset, we ranked methods separately for every label by the absolute Moran's I distortion ($|\log_2 \text{FC}|$) assigning rank 1 to the method with the smallest absolute deviation (best preservation). We then summed ranks across all labels to obtain a dataset-level rank-sum for each method, reflecting overall preservation of label-specific spatial autocorrelation. Because the datasets contain different numbers of labels, raw rank-sums are not directly comparable across datasets; therefore, we standardized rank-sums within each dataset (row-wise) using both min–max scaling (Extended data Fig. 2A) and z-scoring (Extended data Fig. 2B) and visualized the results as heatmaps. For interpretability, we omitted uniform sampling from these heatmaps because it was consistently the top performer and would otherwise compress differences among the remaining methods (see Extended data Section 1.3.5 for details).

Unlike metrics such as spatial coverage or domain preservation, Moran's I distortion did not yield a single dominant "best" method class across all datasets. However, the reduced-rank leverage variant (leverage svd) achieved the best (lowest) standardized rank-sum in four out of eight datasets and performed comparably to the best method in all but two. Notably, this was the only metric for which spatially smoothed geometric sketching (geoSketch) performed well in a substantial subset of datasets (three out of eight). The stronger performance of leverage svd relative to the full leverage score is plausible because the truncated SVD can act as a denoising step that emphasizes dominant, spatially coherent sources of variation while down-weighting high-frequency or sample-specific transcriptional noise; in turn, this may reduce the tendency to over-sample locally heterogeneous regions that can perturb global label autocorrelation. Additionally, operating in a reduced-rank subspace can stabilize the leverage signal in high-dimensional settings, making sampling less sensitive to idiosyncratic gene-level variance. Taken together, these results suggest that when preservation of global spatial autocorrelation of labels is important for downstream analyses, a reduced-rank leverage approach provides a robust default choice, particularly for cell types with intrinsic spatial organization.

Simulated dataset sketching results

This analysis consisted three Visium HD-like and three Xenium-like datasets with different patterns (complex, radial, and striped; see Extended data Fig. 3A and E). For all six datasets, minimization of both transcriptomic Hausdorff distance was similar for all tested algorithms, across sampled fractions, with RNA-based scSampler performing the worst across all datasets, and leverage score performing best (Extended data Figs 3B and F, and 4A and E). For comparison of coordinate-based Hausdorff distances, note that the Visium-like datasets are lacking coordinate-based sketching methods, as they are akin to uniform sampling in this context. For this metric, spatially smoothed Geosketch and scSampler performed the worst, while uniform and leverage score achieved the smallest Hausdorff distance for the Visium-like dataset, with coordinate-based Geosketch performing best on all the Xenium-like datasets (Extended data Figs 3C and G, and 4B and F). However, for both Hausdorff distance comparisons across datasets and between algorithms, all algorithms do a good job minimizing the distance for both coordinates and transcriptomic space, as can been seen with the small variation in values across the sampling fractions. The ARI results show a divergence in algorithm performance across simulated datasets. For the Visium like datasets, there is some variability across sampling fractions, but leverage score and smoothed leverage score, followed by uniform sampling perform best, RNA and smoothed scSampler perform worst (Extended data Figs 3D and 4C). For the Xenium-like datasets, there is no clear best performing algorithm for the complex and radial datasets; however, scSampler variants still under-perform (Extended data Fig. 3H and G). Notably, for both stripe datasets, leverage score and smoothed leverage score significantly outperform all other algorithms (Extended data Figs 3D and H, and 4G).

Comprehensive spatial analysis vignette

This vignette, based on the Allen Institute MERFISH mouse brain dataset, illustrates that the downstream consequences of sketching depend not only on the sampling strategy itself, but also on the spatial structure and biological scale of the feature under consideration. Full analytical details, methods, and extended results are provided in the Supplementary Methods and accompanying vignette file (allen_vignette.html). Briefly, we found that the impact of sketching was heterogeneous across analysis types:

- Globally organized or highly aggregated signals (for example, mean pathway activity scores) were generally well preserved across methods, with only modest differences between approaches. This likely reflects the redundancy inherent in such signals, which are supported by many genes or broad expression programs.
- Spatially organized feature-level signals (including Moran's I of individual genes, ligand–receptor interaction patterns, and spatially structured pathway or transcription factor activities) were substantially more sensitive to sampling strategy. In these settings, leverage-based approaches consistently outperformed density-agnostic geometric methods, whereas spatial smoothing provided only partial recovery of performance.
- Weakly spatially organized signals (such as broadly distributed genes or relatively uniform pathway activities) showed more limited dependence on sketching method, with most approaches yielding similar results.
- High-resolution spatial features (for example, transcription factor activity restricted to specific anatomical regions) were especially vulnerable to undersampling by nonspatially aware methods. In these cases, geometric sketches could substantially under-represent the very regions most relevant for biological interpretation.

Together, these results highlight a central practical point for STs analysis: the importance of sketching strategy increases with the spatial specificity of the biological question. For routine tasks such as general quality control, broad cell-type visualization, or coarse pathway summarization, most sketching approaches may be adequate. By contrast, for analyses aimed at identifying spatially resolved biological programs, leverage-based sampling, particularly when combined with spatial smoothing, more faithfully preserved the signals most relevant for downstream interpretation while still providing substantial reductions in dataset size.

Discussion

Our benchmarking across diverse ST platforms and tissues indicates that sketching method choice should be driven by the specific downstream quantities one aims to preserve, rather than by transcriptomic coverage alone. At commonly used sampling fractions (e.g. 10%), transcriptomic Hausdorff distance was consistently similar across methods, suggesting that most algorithms adequately capture broad expression-space coverage and that this metric is rarely decisive in practice (Fig. 3B, top row). In contrast, metrics that directly reflect structure and spatial coverage in coordinate space, preservation of spatial domains, and recovery of cluster labels—showed clearer separation: uniform sampling and leverage-based approaches generally outperformed transcriptome-driven geometric sketching methods across real datasets, with similar trends (albeit weaker) in simulated data (Fig. 3B and D). The weaker separation observed on simulated benchmarks should not be interpreted as evidence that sketching objectives are inherently less distinct in controlled settings, but rather reflects limitations of current ST simulators in capturing the multiscale complexity of real tissues. In particular, SRTsim-style simulations typically operate over a reduced feature space and generate expression patterns with comparatively smooth, idealized spatial organization, fewer sharp anatomical boundaries, and less within–cell type heterogeneity than platform data. As a result, (i) transcriptome-driven manifold geometry is less informative for differentiating sketch objectives, and (ii) spatially informed heuristics (including smoothing) can align more closely with the simulation's generative assumptions than with real tissues, which contain substantial technical variation (e.g. capture efficiency gradients, segmentation artifacts, ambient RNA) and biological microstructure that partially decouple spatial proximity from expression similarity. These considerations motivate using simulated benchmarks as complementary stress tests rather than as fully predictive surrogates for real-data performance, and further emphasize the importance of evaluating sketching methods on diverse experimental datasets.

These results reinforce a key practical point: when users care about spatially grounded analyses (e.g. domain detection, neighborhood statistics, spatial differential expression,

or niche quantification), method selection should prioritize spatial fidelity metrics rather than minor differences in transcriptomic Hausdorff distance.

Consistent with this, we observed that transcriptome-driven geometric sketching (scSampler/geoSketch on expression features) can substantially distort compositional summaries when highly heterogeneous populations dominate the expression manifold. In several datasets, most strikingly the Visium HD ovarian cancer and whole mouse pup examples—geometric sketches altered cell-type proportions far more than leverage-based sketches (Fig. 4), likely because manifold-coverage objectives preferentially oversample transcriptionally variable regions. Such distortion can bias downstream interpretation when sketches are analyzed in isolation (e.g. apparent shifts in cell-type abundance, altered niche frequencies, or misleading comparisons across conditions). In contrast, leverage-based sampling more consistently preserved label proportions while still providing a transcriptome-informed subset. When the primary objective is to discover rare transcriptional states or characterize high-variance substructure, geometric sketching remains valuable, but we recommend projecting sketch-derived labels back onto the full dataset before making compositional claims. More broadly, running both a leverage-based and a geometric sketch in parallel and comparing projected annotations provides an efficient robustness check for cell typing and state discovery.

Among spatially informed metrics, coordinate-space Hausdorff distance and domain-preservation ARI most strongly favored leverage-based approaches. Coordinate-only sketching (used here primarily as a baseline) unsurpris-ingly optimized spatial coverage but is not broadly applicable across platforms (e.g. Visium HD's regular grid) and can miss transcriptional extremes. Leverage-based sketches, by contrast, frequently achieved substantially better spatial coverage than geometric expression sketches—particularly in highly structured tissues or in samples with strong variability in local sampling density—while maintaining comparable transcriptomic coverage. Similarly, for spatial domain detection using RASP-derived small and large domain parameterizations, leverage-based methods were the conservative best choice in the majority of datasets, although domain solutions at 10% sampling exhibited nontrivial variance, emphasizing the need for repeated runs or stability checks. Finally, preservation of global label spatial autocorrelation (Moran's I) revealed a nuanced pattern: uniform sampling was consistently best, but among nonuniform methods, reduced-rank leverage (leverage svd) emerged as a robust choice across datasets, suggesting that truncated subspace sampling can better maintain global spatial label structure than full-rank leverage in some settings. Taken together, these results support the following practical guidance: use leverage-based sketching as a default when spatial coverage, clustering concordance, or domain detection are key; prefer reduced-rank leverage when preserving global spatial autocorrelation of labels is particularly important; and reserve transcriptome-driven geometric sketching for applications centered on capturing transcriptional heterogeneity and rare states, ideally coupled with label projection back to the full dataset for unbiased spatial and compositional interpretation.

It should be noted that several of the evaluation metrics used here can, in principle, be improved by post hoc heuristics that are not part of the underlying sketch objective, and we therefore interpret them as diagnostics of 'fixed-budget' sketch fidelity rather than as adversarially robust measures. For example, starting from a coordinate-or uniform-based sketch and then augmenting it with spatial k-nearest neighbors of selected points could artificially improve label- and neighborhood-based metrics (e.g. ARI, Moran's I, and domain concordance), and may also reduce coordinate-space Hausdorff distance, largely by increasing local redundancy. Such "completion" strategies, however, effectively change the sampling distribution and/or increase the information budget by trading breadth of coverage for nearby, correlated points, which can reduce the sketch's ability to capture transcriptomic extremes and rare states under a fixed sampling fraction. In this regard, expression-space Hausdorff distance and PCA-structure preservation are comparatively less susceptible to purely spatial neighbor augmentation (and can even worsen as redundancy increases), whereas spatial label metrics are more vulnerable when labels are spatially smooth. Future work could explicitly study budget-matched two-stage baselines (e.g. coordinate sketch + neighbor completion with the same final sketch size), incorporate redundancy penalties or effective-sample-size measures, and adopt holdout-based task evaluations (e.g. generalization of clustering/domain assignments or neighborhood composition to withheld cells) to reduce opportunities for metric-driven "gaming" and to better align evaluation with downstream scientific use cases.

Looking forward, our findings motivate sketching frameworks that explicitly optimize joint objectives over transcriptomic and spatial fidelity while providing predictable behavior for compositional preservation. Promising directions include multi-objective sampling that balances expression-space coverage with coordinate-space coverage, adaptive strategies that account for local density variation, and stability-aware domain-preserving sketches. As ST datasets continue to scale in both cell count and complexity, such methods will be essential for enabling rapid exploratory analysis without inadvertently distorting the spatial and compositional signals that often motivate ST experiments in the first place.

Acknowledgements

Author contributions: I.K.G. and H.R.F. conceptualized the project and developed the methodology. I.K.G. performed data curation, formal analysis, software development, and wrote the original draft of the manuscript. I.K.G., B.A.G., and H.R.F. contributed to writing—review and editing. H.R.F. and B.A.G. acquired funding and supervised the project.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

National Institute of Health grants R35GM146586 (I.K.G. and H.R.F.), P20GM130454 (I.K.G., B.A.G., H.R.F.), and P30CA023108 (H.R.F.).

Data availability

Mouse ovary:

Please contact the authors for access the the Mouse ovary dataset.

Breast Cancer:

We downloaded the Breast cancer 10x Xenium dataset from the SubcellularSpatialData R package via Exper-imentHub, dataset # EH8567, sample ID 'IDC' (https://www.bioconductor.org/packages/release/data/experiment/html/SubcellularSpatialData.html) (last accessed date: 18 September 2024).

Human lung:

We downloaded the healthy human lung 10x Xenium dataset from publicly available datasets on the 10x website https://www.10xgenomics.com/datasets/xenium-human-lung-preview-data-1-standard (last accessed date: 17 March 2025).

Sagittal mouse brain:

We downloaded the mouse brain dataset from the Allen Brain Atlas via their abcatlasaccess Python package. Specifically we followed the MERFISH whole mouse brain STs (Xiaowei Zhuang) tutorial https://alleninstitute.github.io/abc_atlas_access/notebooks/zhuang_merfish_tutorial.html to access section 3.010 from the 'Zhuang-ABCA-3′ dataset (last accessed date: 17 March 2025).

Coronal mouse brain:

We downloaded the coronal mouse brain dataset from the publicly available datasets on the 10x website https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-brain-he (last accessed date: 17 March 2025).

Whole mouse pup:

We downloaded the whole mouse pup Xenium dataset from the publicly available datasets on the 10x website https://www.10xgenomics.com/datasets/xenium-prime-ffpe-neonatal-mouse, (last accessed date: 12 February 2025).

Mouse embryo:

We downloaded the whole mouse embryo Visium HD dataset from the publicly available datasets on the 10x website https://www.10xgenomics.com/datasets/visium-hd-three-prime-mouse-embryo-fresh-frozen (last accessed date: 12 February 2025).

Human ovarian cancer:

We downloaded the human ovarian cancer Visium HD dataset from the publicly avaiable datasets on the 10x website https://www.10xgenomics.com/datasets/visium-hd-three-prime-ovarian-cancer-discovery-fresh-frozen (last accessed date: 12 February 2025).

Simulated data:

Please contact the authors for access to the simulated datasets.

See the following Github: https://github.com/gingerii/Benchmarking-sketching-for-spatial-transcriptomics.

References

1. Joseph VR. Space-filling designs for computer experiments: a review. *Qual Eng* 2016;28:28–35. https://doi.org/10.1080/08982112.2015.1100447.
2. Chen Y, Welling M, Smola A. Super-samples from Kernel Herding. Proceedings of the Twent-Sixth Conference on Uncertainty in Artificial Intelligence. 2012. bioXiv, https://doi.org/10.48550/arXiv.1203.3472, 15 Mar 2012, preprint: not peer reviewed.
3. Barbian MH, Assunção RM. Spatial subsemble estimator for large geostatistical data. *Spat Stat* 2017;22:68–88. https://doi.org/10.1016/j.spasta.2017.08.004
4. Johnson ME, Moore LM, Ylvisaker D. Minimax and maximin distance designs. *J Stat Plan Inference* 1990;26:131–48. https://www.sciencedirect.com/science/article/pii/037837589090122B
5. Clarkson KL, Woodruff DP. Low-rank approximation and regression in input sparsity time. *J ACM* 2017;63:54:1–54:45. https://doi.org/10.1145/3019134
6. Miao Z, Moreno P, Huang N *et al.* Putative cell type discovery from single-cell gene expression data. *Nat Methods* 2020;17:621–8. https://doi.org/10.1038/s41592-020-0825-9
7. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019;15:e8746. https://doi.org/10.15252/msb.20188746
8. Wang T, Li B, Nelson CE *et al.* Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinf* 2019;20:40. https://doi.org/10.1186/s12859-019-2599-6
9. Song D, Xi NM, Li JJ *et al.* scSampler: fast diversity-preserving subsampling of large-scale single-cell transcriptomic data. *Bioinformatics* 2022;38:3126–7. https://doi.org/10.1093/bioinformatics/btac271
10. Hie B, Cho H, DeMeo B *et al.* Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell Syst* 2019;8:483–493. https://doi.org/10.1016/j.cels.2019.05.003
11. Hao Y, Stuart T, Kowalski MH *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;42:293–304. https://doi.org/10.1038/s41587-023-01767-y
12. DeMeo B, Berger B. Hopper: a mathematically optimal algorithm for sketching biological data. *Bioinformatics* 2020;36:i236–41. https://doi.org/10.1093/bioinformatics/btaa408
13. Huang L, Gong W, Chen D. scValue: value-based subsampling of large-scale single-cell transcriptomic data for machine and deep learning tasks. *Briefings in Bioinformatics* 2025;26:3. https://doi.org/10.1093/bib/bbaf279
14. Do VH, Elbassioni K, Canzar SS. Spherical thresholding improves sketching of single-cell transcriptomic heterogeneity. *iScience* 2020;23:101126. https://doi.org/10.1016/j.isci.2020.101126
15. Ren J, Zhang Q, Zhou Y *et al.* A downsampling method enables robust clustering and integration of single-cell transcriptome data. *J Biomed Inform* 2022;130:104093. https://doi.org/10.1016/j.jbi.2022.104093
16. Polan´ski K, Bartolom´e-Casado R, Sarropoulos I *et al.* Bin2cell reconstructs cells from high resolution Visium HD data. *Bioinformatics* 2024;40:btae546. https://doi.org/10.1093/bioinformatics/btae546
17. Stickels RR, Murray E, Kumar P *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat Biotechnol* 2021;39:313–9. https://doi.org/10.1038/s41587-020-0739-1
18. Janesick A, Shelansky R, Gottscho A *et al.* High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. 2022. https://www.biorxiv.org/content/10.1101/2022.10.06.510405v1, accessed on 10/23/2023.
19. Yao Z, Velthoven CTJv, Kunst M *et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. 2023. https://www.biorxiv.org/content/10.1101/2023.03.06.531121v1, accessed on 4/28/2025.
20. Chen A, Liao S, Cheng M *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. en. *Cell* 2022;185:1777–1792. https://doi.org/10.1016/j.cell.2022.04.003

21. Xu Z, Wang W, Yang T *et al.* STOmicsDB: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. *Nucleic Acids Res* 2024;52:D1053–D1061. https://academic.oup.com/nar/article/52/D1/D1053/7416388
22. Li Y, Li Z, Wang C *et al.* Spatiotemporal transcriptome atlas reveals the regional specification of the developing human brain. *Cell* 2023;186:5892–5909. https://doi.org/10.1016/j.cell.2023.11.016
23. Pan J, Li Y, Lin Z *et al.* Spatiotemporal transcriptome atlas of human embryos after gastrulation. 2023. https://www.biorxiv.org/content/10.1101/2023.04.22.537909v1, accessed on 8/18/2025.
24. Shi H, He Y, Zhou Y *et al.* Spatial atlas of the mouse central nervous system at molecular resolution. *Nature* 2023;622:552–61. https://doi.org/10.1038/s41586-023-06569-5
25. Han L, Liu Z, Jing Z *et al.* Single-cell spatial transcriptomic atlas of the whole mouse brain. *Neuron* 2025;113:2141–2160. https://www.cell.com/neuron/fulltext/S0896-6273(25)00133-3
26. Zhang M, Pan X, Jung W *et al.* Molecularly defined and spatially resolved cell atlas of the whole mouse brain. *Nature* 2023;624:343–54. https://doi.org/10.1038/s41586-023-06808-9
27. Gayoso A, Lopez R, Xing G *et al.* A Python library for probabilistic analysis of single-cell omics data. *Nat Biotechnol* 2022;40:163–6. https://doi.org/10.1038/s41587-021-01206-w
28. Tejada-Lapuerta A, Schaar AC, Gutgesell R *et al.* Nicheformer: a foundation model for single-cell and spatial omics. *Nat Methods* 2025;22:2525–38. https://doi.org/10.1038/s41592-025-02814-z
29. Wang C, Cui H, Zhang A *et al.* scGPT-spatial: continual pretraining of single-cell foundation model for spatial transcriptomics. bioRxiv, https://www.biorxiv.org/content/10.1101/2025.02.05.636714v1, 8 February 2025, preprint: not peer reviewed.
30. Huang R, Kratka CE, Pea J *et al.* Single-cell and spatiotemporal profile of ovulation in the mouse ovary. *PLoS Biol* 2025;23:e3003193. https://doi.org/10.1371/journal.pbio.3003193
31. Bhuva DD, Tan CW, Salim A *et al.* Library size confounds biology in spatial transcriptomics data. *Genome Biol* 2024;25:99–760X. https://doi.org/10.1186/s13059-024-03241-7
32. Zhu J, Shang L, Zhou X. SRTsim: spatial pattern preserving simulations for spatially resolved transcrip-tomics. *Genome Biol* 2023;24:39. https://doi.org/10.1186/s13059-023-02879-z
33. Gingerich IK, Goods BA, Frost HR. Randomized Spatial PCA (RASP): a computationally efficient method for dimensionality reduction of high-resolution spatial transcriptomics data. *PLoS Comput Biol* 2025;21:e1013759. https://doi.org/10.1371/journal.pcbi.1013759

Received: September 8, 2025. **Revised:** April 7, 2026. **Accepted:** April 15, 2026