Benchmarking sketching methods on spatial transcriptomics data

Ian K. Gingerich, Brittany A. Goods, H. Robert Frost

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 used in scRNA-seq, most sketching methods optimize coverage in expression space alone and ignore physical location, risking spatial bias. We systematically benchmark 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: 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 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 (ARI), PCA loading drift, and local cell-type 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.

1 Introduction

1.1 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 Hausdorff 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 non-redundant 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].

1.2 Sketching for scRNA-seq data

Single cell RNA-seq (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, ie: 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, ie: 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 non-uniform 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].

1.3 Sketching for ST data

These limitations are especially problematic given the scale and resolution of modern ST platforms. 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 half a million cells, across targeted panels of $5,000$ genes. As these technologies mature, multi-section studies and atlases now comprise hundreds-of-thousands to 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], compute 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. 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. 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. 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. 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 address this gap, we benchmark current sketching approaches on 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.

2 Methods

2.1 Evaluated ST datasets

We benchmarked four families of sketching methods (see Table 2) across five 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 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 20 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 [27]; a human breast cancer 10x Xenium dataset with histology-derived annotations [28]; 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 ARI was computed against cell-class labels; and 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. 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 [29] 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 **??**). See Extended Data Section 1.1 for dataset-specific preprocessing details.

Table 1: Dataset summary

Tissue	Species	Technology	# locations	# genes	Study
Sagittal brain	Mouse	Vizgen MERFISH	81,452 cells	1,122	Zhang et al., 2023
Ovary	Mouse	Vizgen MERFISH	43,038 cells	228	Huang et al., 2024
Coronal brain	Mouse	10x Visium HD	393,434 locations*	19,059	10x Genomics (Visium HD)
Lung	Human	10x Xenium	295,883 cells	541	10x Genomics (Xenium Lung)
Breast cancer	Human	10x Xenium	565,916 cells	541	Bhuva et al., 2024
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	—

* 8 μm binned locations.

2.2 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 (20 components), (ii) raw spatial coordinates, and (iii) spatially smoothed embeddings that incorporate local neighborhood structure. 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 $\bar{U} = WU$. Leverage scores are computed from the smoothed basis $\bar{U}$ (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 ten 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.

Table 2: Sketching algorithms evaluated and their configurations.

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

2.3 Evaluation metrics

We evaluated each sketch along complementary transcriptomic and spatial criteria. Robust (partial) Hausdorff distance quantified worst-case nearest-neighbor discrepancies 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 Adjusted Rand Index (ARI) assessed concordance between sketch-derived clusters and reference labels: ground-truth annotations when available (MERFISH ovary, MERFISH brain class, Xenium breast cancer) 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 local compositional fidelity, we computed the mean squared error between k-nearest-neighbor cell-type proportion vectors derived from the full versus sketched datasets at each sketched location, summarizing changes in local neighborhood makeup. Finally, we derived an overall spatial score by rank-summing ARI, coordinate-space Hausdorff distance, and local neighborhood distortion at a fixed sampling fraction, emphasizing joint preservation of group structure, spatial coverage, and microenvironmental context. See Extended data section 1.3 for detailed metric computations.

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

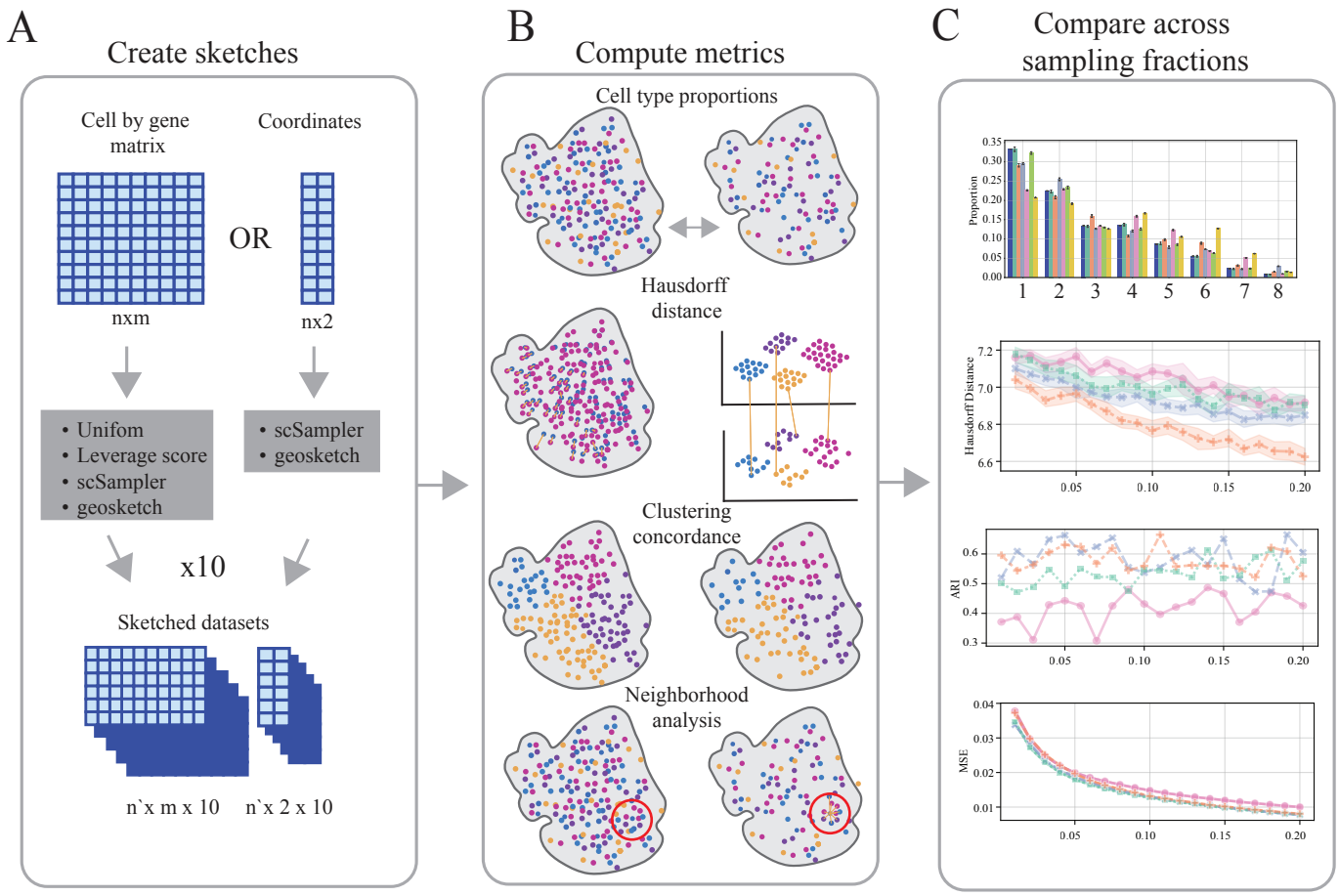

Figure 1: **Analysis pipeline. A:** Sketching protocol **B:** Sketched dataset evaluation metrics **C:** Metric quantification across sampling fractions

quality[9, 10]. These metrics include the robust Hausdorff distance, Adjusted Rand Index (ARI) between the sketch-based clustering solution and ground truth annotations, mean difference between PCA loading vectors, and distortion of the local neighborhood structure (see Fig 1B and 1C). This evaluation framework provides a framework to understand the impact of different down-sampling 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.
- **Adjusted Rand Index (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.
- **Local neighborhood distortion (MSE of cell-type proportions).** Assesses how well the sketch preserves local tissue composition by comparing, for each cell, the kNN-based cell-type proportion vector between the full and sketched datasets. Lower MSE indicates better retention of neighborhood context and microenvironmental structure.
- **Overall spatial score (rank-sum across metrics).** Aggregates ranks across ARI, coordinate-space Hausdorff, and local neighborhood distortion to summarize spatial fidelity at a fixed sampling fraction. Lower (better) rank-sum reflects methods that jointly preserve group structure, spatial coverage, and local composition.

3 Results

Across five 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.

3.1 Sketching result summary

Figs 2 and 3 visualize the overall performance for sketching fractions of 0.1 across all simulated and real ST datasets. For the simulated Visium-HD-like and coronal mouse brain datasets, uniform sampling, leverage score computed on the full cell-by-gene matrix, and leverage score on the spatially smoothed left singular vectors perform well across all evaluation metrics and show a robust overall spatial score (2B (left side of heatmap, columns 1-3) and3D (column 3)). For the other real and simulated datasets, coordinate-based sketching and spatially-smoothed leverage scores result in the best overall score, with the traditional scSampler and Geosketch design preforming the worst.

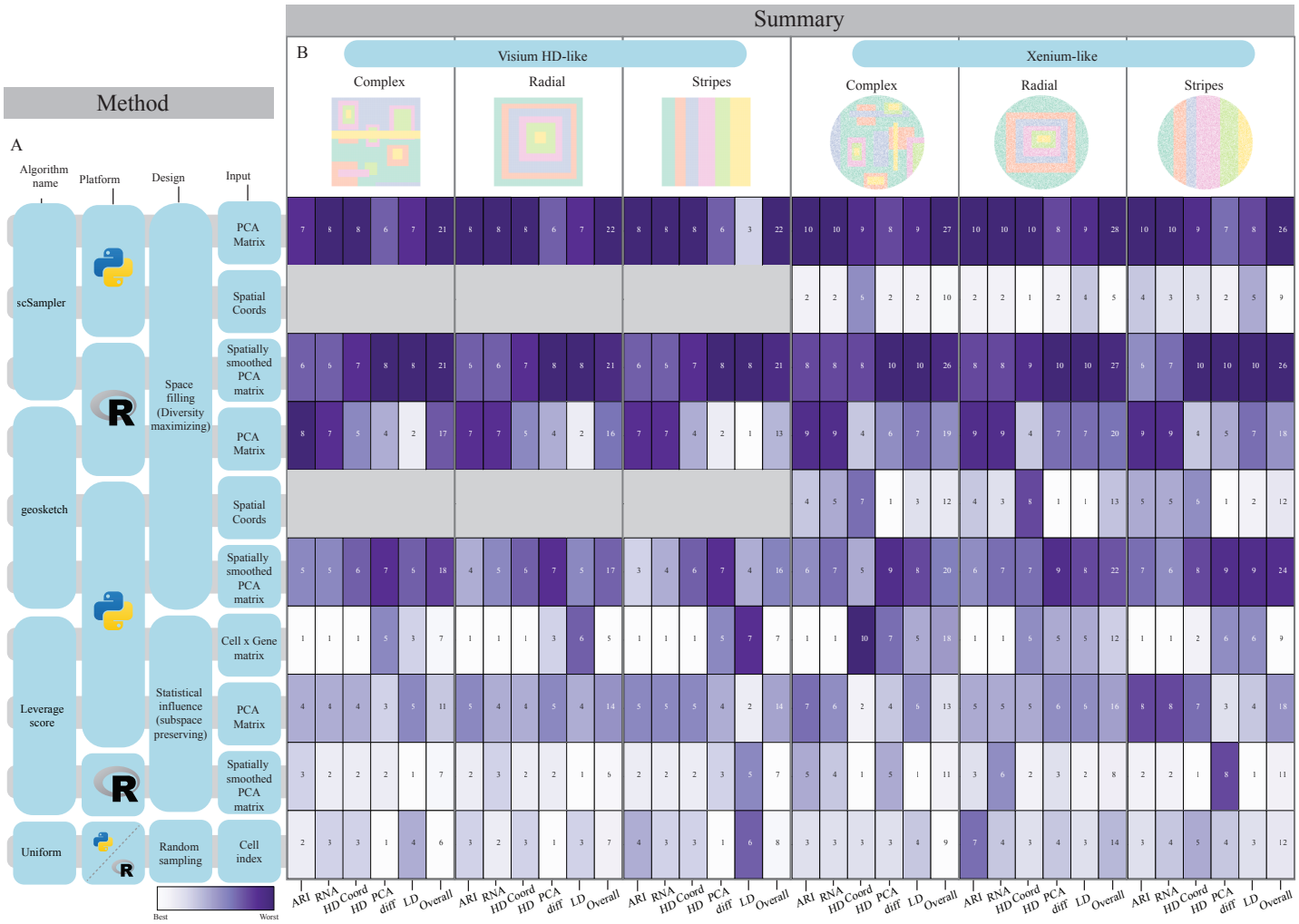

Figure 2: **Overall sketching performance for 0.10 sampling fraction across simulated datasets A:** Tested sketching algorithm details (platform, sketching approach and input data). **B:** Summary statistics for the six evaluated simulation models. Each column represents a evaluation metric and each row a unique algorithm condition. Numeric values represent the rank order for that metric, color is normalized by column. The final column in each dataset represents the overall spatial score for that dataset, which is the summed ranks of all spatial metrics.

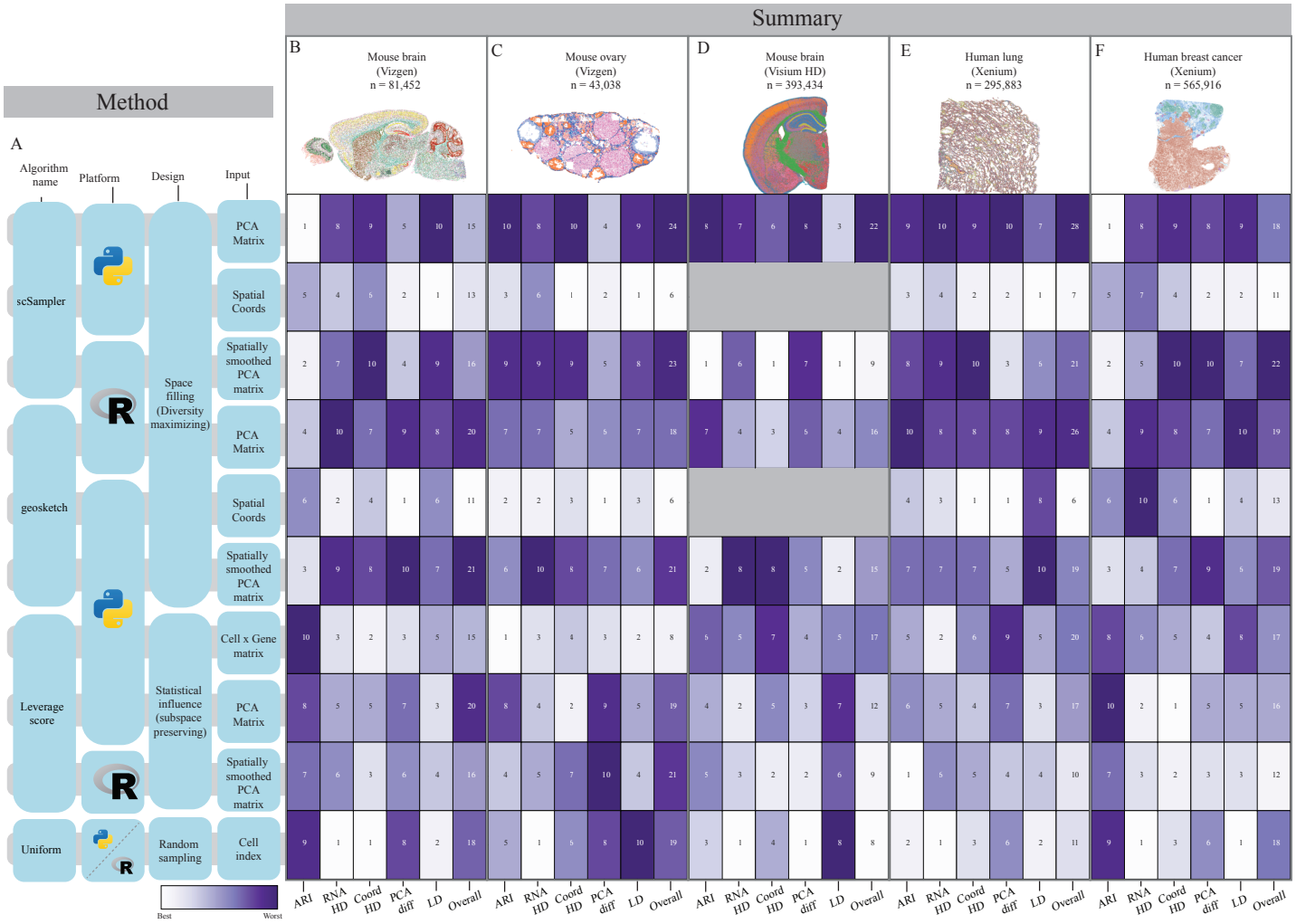

Figure 3: **Overall sketching performance for real ST datasets at 0.10 sketching fraction A:** Tested sketching algorithm details. **B:** Summary statistics for sagital mouse brain dataset (Vizgen MERFISH). **C:** Summary statistics for mouse ovary dataset. **D:** Summary statistics for coronal mouse brain (10x Visium HD). **E:** Summary statistics for human lung dataset (10x Xenium). **F:** Summary statistics for human breast cancer dataset (10x Xenium). Each column represents a evaluation metric and each row a unique algorithm condition. Numeric values represent the rank order for that metric, color normalized by column. The final column in each dataset represents the overall spatial score for that dataset, which is the summed ranks of all spatial metrics.

3.2 Simulated dataset sketching results

This analysis consisted three Visium HD-like and three Xenium-like datasets with different patterns (complex, radial, and striped; see Fig 4A and 4E). 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 (Fig 4B and 4F). 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 (Fig 4C and 4G). 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 (Fig 4D). For the Xenium-like datasets, there is no clear best performing algorithm for the complex and radial datasets, however, scSampler variants still under-perform (Fig 4H). Notably, for both stripe datasets, leverage score and smoothed leverage score significantly outperform all other algorithms (Fig 4D and 4H).

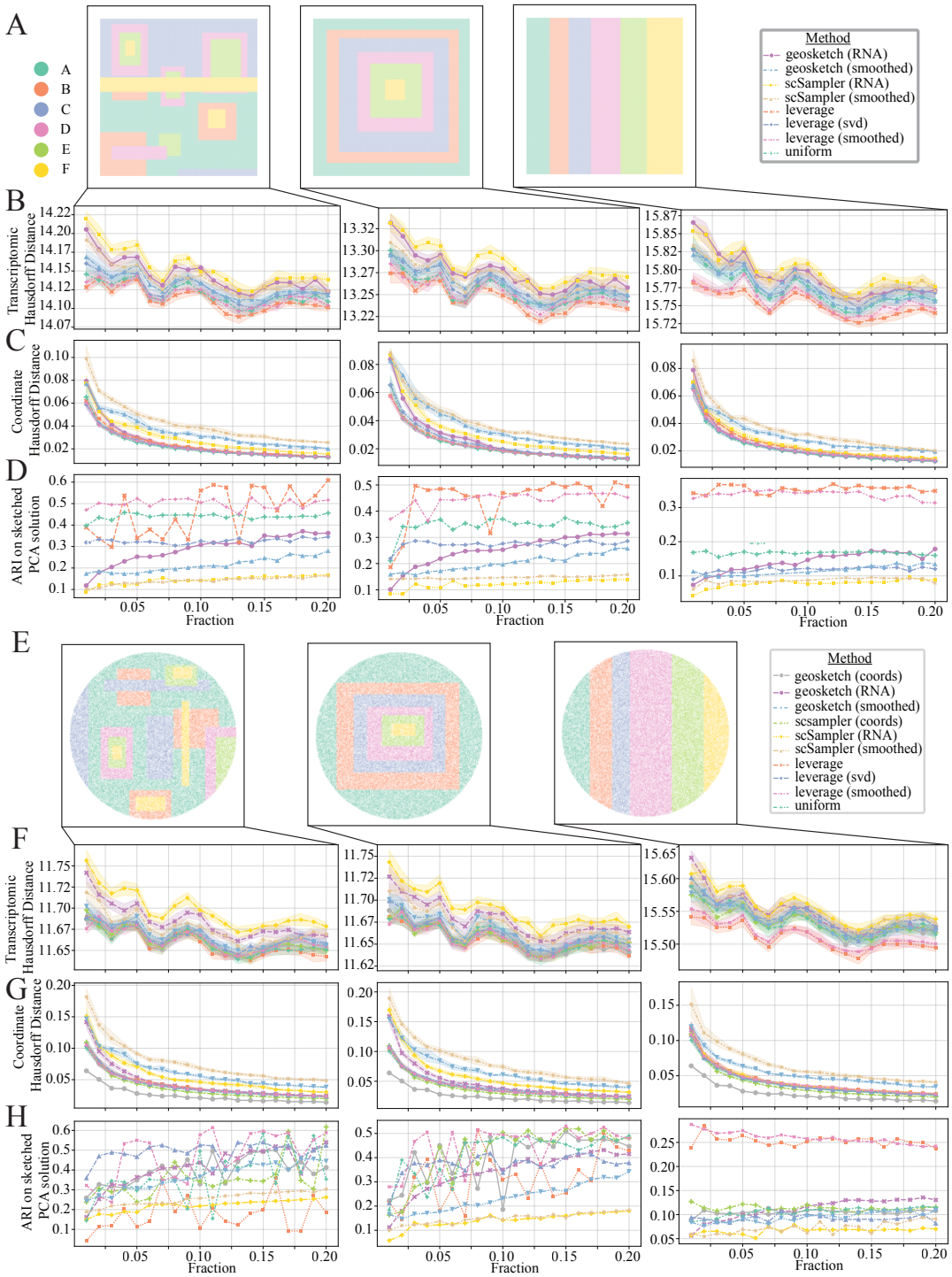

Figure 4: **Simulation results for sketching fractions 0.01-0.2. A:** Ground truth locations plotted spatially for Visium-like complex (left), Visium-like radial (middle) and Visium-like stripes (right) simulated datasets. **B, F:** Transcriptomic Hausdorff distance results for evaluated algorithms for Visium-like datasets (B), and Xenium-like datasets (F). **C, G:** Spatial coordinate Hausdorff distance results for Visium-like datasets (C) and Xenium-like datasets (G). **D, H:** Adjusted Rand Index (ARI) results (sketched clustering results compared to ground truth) for Visium-like datasets (D) and Xenium-like datasets (H).

3.3 Human breast cancer

This analysis utilized the 10x Xenium human breast cancer dataset, downloaded through the R package SubcellularSpatialData[28]. This resource contains ST data derived from FFPE human breast tissue samples featuring infiltrating ductal carcinoma in situ. Annotations are derived from the H&E stained tissue accompanying the original data (Fig 5A). Transcriptomic sketching results in a higher proportion of cells from the DCIS and normal ducts being sampled (Fig 5B top right), while coordinate-based sketching shows a largely uniform sampling across the tissue, with edge effects (Fig 5B bottom right). Geosketch and scSampler effectively minimize the transcriptomic Hausdorff distance much better than the other methods (Fig 5C top left and Extended data Fig 2). As expected, Coordinate-based sketching effectively minimizes the coordinate Hausdorff distance, with leverage score coming in behind the coordinate-based methods (Fig 5C top right and Extended data Fig 2). Uniform sampling and the coordinate-based Geosketch and scSampler achieved the highest ARI scores for sampling fractions between $0.01 - 0.2$, with transcriptomic versions of Geosketch and scSampler under-performing all other methods (Fig 5C middle left). Uniform and coordinate-based sketching methods best preserved local neighborhood structure, with the lowest MSE in location annotation when compared to the full dataset, across all sampled fractions (Fig 5C bottom left and Extended data Fig 2). Areas with the highest MSE for transcriptomic sketching when compared to coordinate-based sketching were localized in areas where transcriptomic sketching tended to over-sample including the DCIS and normal ducts (Fig 5D).

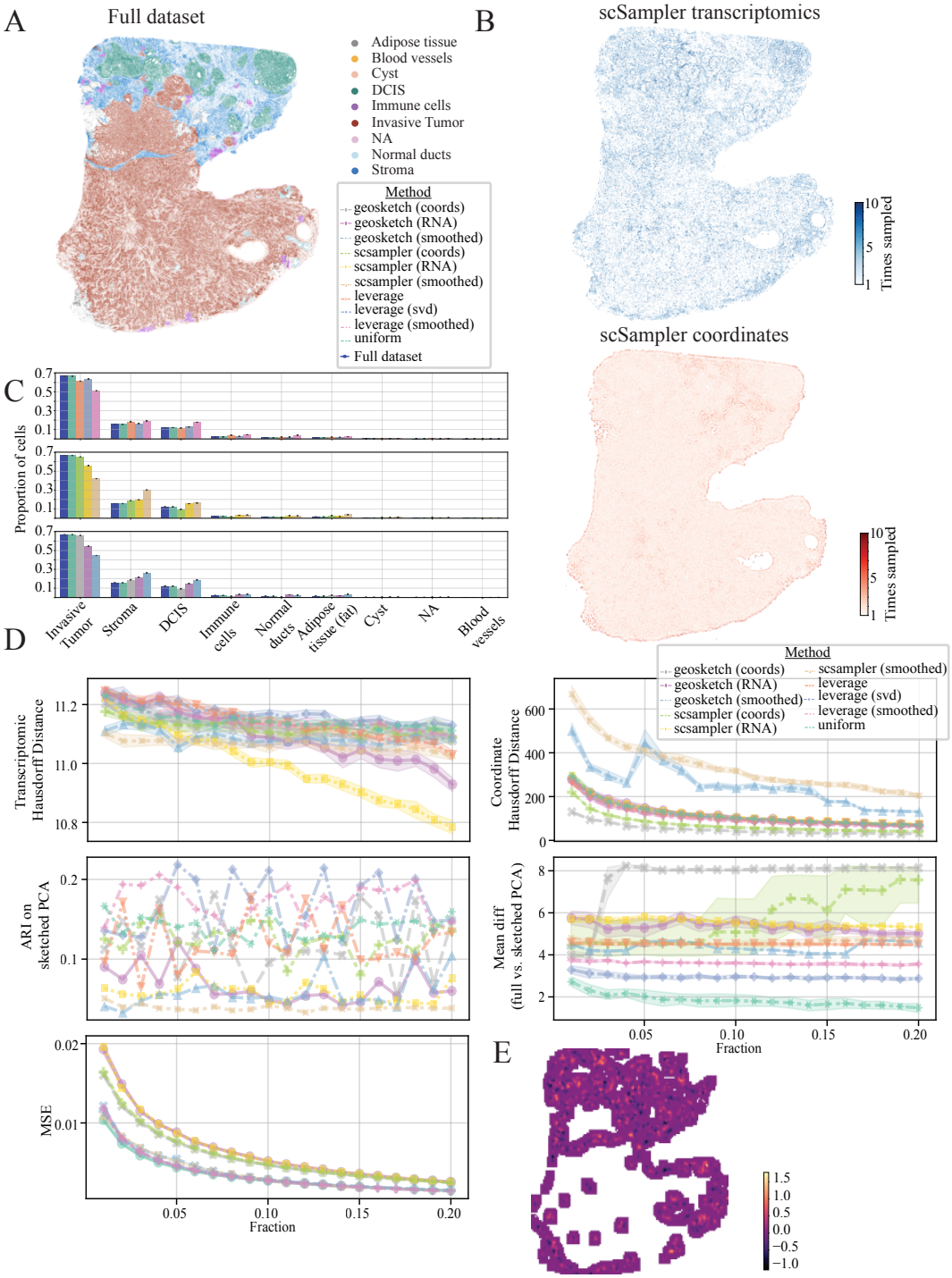

Figure 5: **Human breast cancer results (10x Xenium). A:** Ground truth cell type labels for full dataset **B:** Resulting sketched datasets for transcriptomic (top right), coordinate (bottom right) and composite (bottom left) based scSampler sketching. Color scale indicates how many times a cell was sampled out of ten iterations of the sketching algorithm. **C:** Cell type proportions for full dataset and 0.1 downs-sampled sketches across algorithms evaluated. Rows correspond to Geosketch (top), scSampler (middle) and Leverage score (bottom) algorithms. **D:** Evaluation metrics for sketching algorithms. Transcriptomic Hausdorff distance between sketch and full dataset for sampling fractions from $0.01 - 0.20$ (top left), coordinate Hausdorff distance between sketch and full dataset for sampling fractions from $0.01 - 0.20$ (top right), Adjusted Rand Index (ARI) for sketched cluster solutions compared to ground truth (middle left), mean difference between the full vs sketched data projected onto the principle component loadings (middle right), mean squared error (MSE) between sketched and full dataset's local neighborhood composition (bottom left). Colors indicate sketching method. **E:** Difference in local neighborhood composition MSE between transcriptomic vs coordinate-based scSampler sketching at a 10% sampling fraction, visualized on the tissue. Lighter colors indicate higher transcriptomic MSE compared to coordinate based sketching.

3.4 Coronal mouse brain results

This analysis was performed on a 10x Visium-HD dataset of a mouse coronal brain section downloaded from the 10x website. Cells were labeled via graph-based clustering as no ground truth annotations were available, and generally correspond to known anatomical structures (Fig 6A). RNA-based sketching significantly over-sampled cells in the hippocampus, pia surface and parts of the thalamus when compared to uniform sampling (Fig 6B). This can also be seen in the quantification of cell type proportions, where we see that both scSampler and Geosketch bias towards label 0 as compared to uniform and leverage scor- based sketching (Fig 6C). All algorithms effectivly minimize the transcriptomic Hausdorff distance (Fig 6D top left), while all variants of leverage score, and uniform sketching best minimize the coordinate Hausdorff distance (Fig 6D top right). Sketched labels compared against graph-based clustering on full dataset (ARI) are highly variable, with no algorithm generating superior performance (Fig 6D middle, left). scSampler, and Geosketch showed similar low MSE in local neighborhood structure preservation compared to the full dataset, while leverage score and uniform sampling had larger values (Fig 6D bottom left).

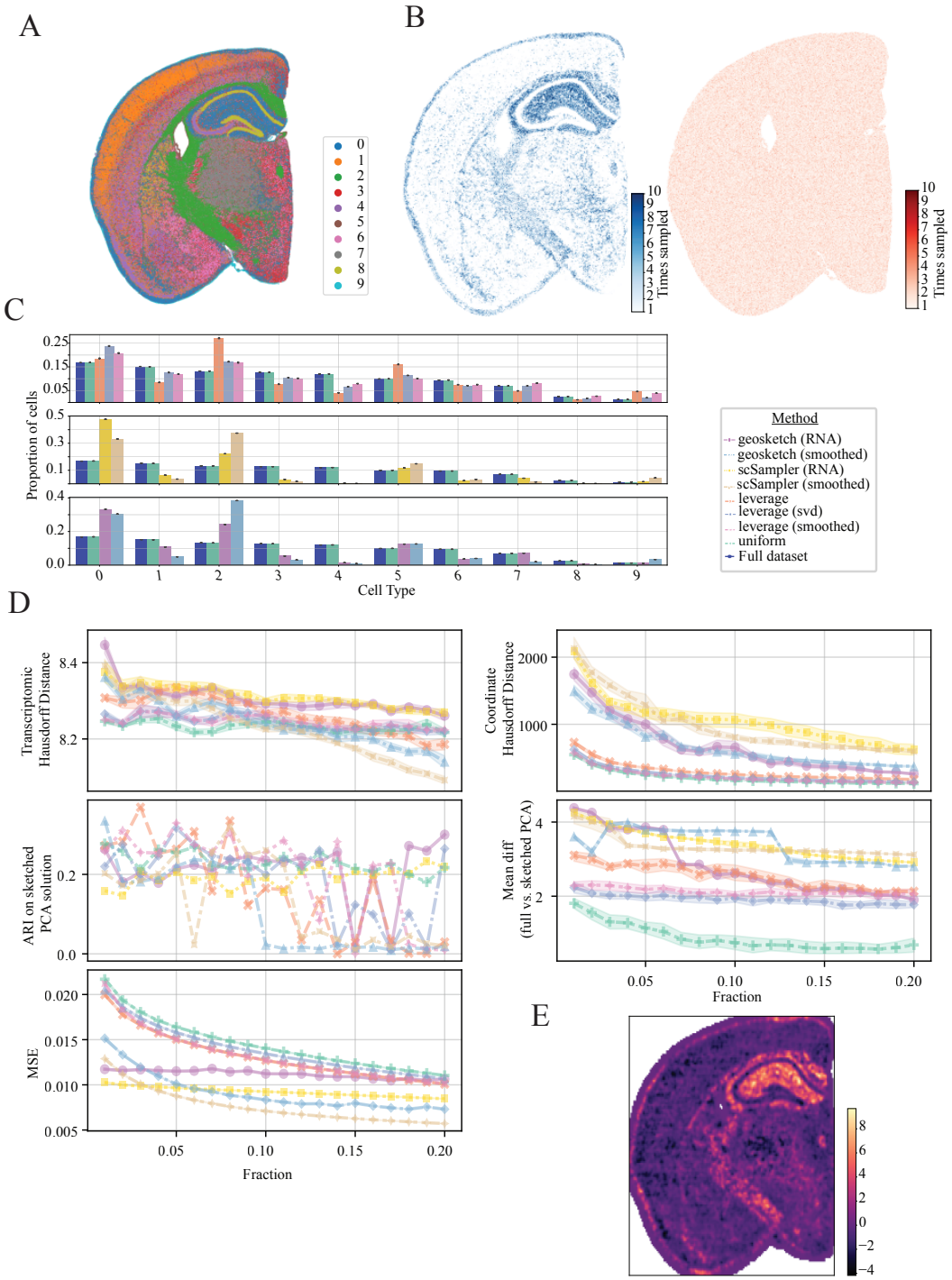

Figure 6: **Mouse brain results (10x Visium HD). A:** Ground truth cell type labels for full dataset. **B:** Sketchedd datasets for scSampler based on transcriptomic data (top right) or tissue coordinates (bottom right). Color scale indicates how many times a cell was sampled out of ten iterations of the sketching algorithm. **C:** Cell type proportions for full dataset and 0.1 down-sampled sketches across algorithms evaluated. Rows correspond to Geosketch (top), scSampler (middle) and Leverage score (bottom) algorithms under different conditions **D:** Evaluation metrics for each sketching algorithm. Transcriptomic Hausdorff distance between sketch and full dataset for sampling fractions from $0.01 - 0.20$ (top left), coordinate Hausdorff distance between sketch and full dataset for sampling fractions from $0.01 - 0.20$ (top right), ARI for sketched cluster solutions compared to ground truth (middle left), mean difference between the full vs sketched data projected onto the principle component loadings (middle right), MSE between sketched and full dataset's local neighborhood composition (bottom left). Colors indicate sketching method. **E:** Difference in local neighborhood composition MSE between transcriptomic scSampler vs uniform sketching at a 10% sampling fraction, visualized on the tissue. Lighter colors indicate higher transcriptomic MSE compared to coordinate-based sketching

3.5 Human lung results

This analysis utilized the 10x Xenium healthy human lung dataset, downloaded from the 10x website. This resource contains ST data derived from FFPE human lung tissue. Annotations are derived from graph-based Louvain clustering (Extended Data Fig 4A). This fibrotic tissue has a unique heterogeneous structure, making sketching a difficult task. Transcriptomic sketching results in a higher proportion of cells in the bottom left section of the tissue(Extended Data Fig 4B top right), while coordinate-based sketching shows a largely uniform sampling across the tissue, with edge effects (Extended Data Fig 4B bottom right). Interestingly, transcriptomic Geosketch and scSampler have larger transcriptomic Hausdorff distances than the other methods for sampling fractions of $0.01 - 0.2$ (Extended Data Fig 4C top left and Extended data Fig 5). Coordinate-based sketching effectively minimizes the coordinate Hausdorff distance, with all other methods comparable (Extended Data Fig 4C top right and Extended data Fig 5). Uniform sampling. the coordinate-based Geosketch and scSampler achieved the highest ARI scores for sampling fractions between $0.01 - 0.2$, with transcriptomic versions of Geosketch and scSampler under-performing all other methods (Extended Data Fig 4C middle left). All methods showed similar MSE in local neighborhood structure preservation compared to the full dataset (Extended Data Fig 4C bottom left). This is likely due to the high heterogeneity of the spatial and transcriptomic features of this dataset. Despite this similarity, when comparing between transcriptomic vs coordinate-based scSampler's MSE it is clear that transcriptomic sketching biases local neighborhoods in areas of higher sampling probability (Extended Data Fig 4D).

3.6 Mouse brain results

The mouse brain dataset features detailed cell type classification provided by the Allen Brain Institute [19]. We choose to use a Sagittal section from the atlas as it spans the full anterior-posterior axis and includes both cortical and sub-cortical brain regions, including the cerebellum (Extended Data Fig 6A). Transcriptomic sketching results in a higher proportion of cells in the olfactory bulb and hippocampus being sampled, likely due to a larger amount of transcriptomic variance in this region (Extended Data Fig 6B top right), while coordinate-based sketching shows a largely uniform sampling across the tissue, with pronounced edge effects (Extended Data Fig 6B bottom right). Geosketch and scSampler effectively minimize the transcriptomic Hausdorff distance much better than the other methods (Extended Data Fig 6C top left and Extended data Fig 7). As expected, coordinate-based sketching effectively minimizes the coordinate Hausdorff distance, with leverage score coming in behind the coordinate-based methods, (Extended Data Fig 6C top right and Extended Data Fig 7). Uniform sampling, leverage score, and the coordinate-based Geosketch achieved the highest ARI scores for sampling fractions between $0.01 - 0.2$, with the coordinate-based scSampler and transcriptomic versions of Geosketch and scSampler under performing all other methods. This finding highlights the importance of incorporating spatial as well as transcriptomic features into the sketching process (Extended Data Fig 6C middle left). Transcriptomic Geosketch and scSampler under-performed all other methods for local neighborhood structure preservation, with the highest MSE in cell-type proportions when compared to the full dataset, across all sampled fractions (Extended Data Fig 6C bottom left amd Extended Data Fig 7). Areas with the highest MSE for transcriptomic sketching when compared to coordinate-based sketching where localized in areas where transcriptomic sketching tended to over-sample including the olfactory bulb and cerebellum (Extended Data Fig 6D).

3.7 Mouse ovary results

This MERFISH dataset features ground-truth cell-level annotations of the mouse ovary and shows clear spatial organization for some cell types (luteal, granulosa) and heterogeneous distribution for others (epithelial, immune, endothelial) (Extended Data Fig 8A)[27]. Transcriptomic sketching results in a higher proportion of cells in the upper left part of the tissue being sampled, likely due to a larger amount of transcriptomic variance in this region (Extended Data Fig 8B top right), while coordinate-based sketching shows a largely uniform sampling across the tissue, with pronounced edge effects**??** (Extended Data Fig 8B bottom right). Transcriptomic Hausdorff distances between the various methods are quite similar, with the leverage score being slightly better across the sampled fractions. Interestingly, coordinate-based sketching achieved a smaller Hausdorff distance compared to the transcriptomic version for both Geosketch and scSampler (Extended Data Fig 8C top left and Extended Data Fig 9). As expected, coordinate-based sketching effectively minimizes the coordinate Hausdorff distance, with leverage score coming in behind the coordinate-based methods (Extended Data Fig 8C top right and Extended Data Fig 9).

Coordinate-based sketching and leverage score also achieved the highest ARI scores for sampling fractions between $0.01 - 0.2$, with the transcriptomic versions of Geosketch and scSampler under-performing all other methods. This finding highlights the importance of incorporating spatial as well as transcriptomic features into the sketching process (Fig Extended Data Fig 8C middle left). As expected, the coordinate-based sketching best captured the local neighborhood structure, with the lowest MSE in cell-type proportions when compared to the full dataset across all sampled fractions (Extended Data Fig 8C bottom left). Areas with the highest MSE for transcriptomic sketching when compared to coordinate-based sketching were localized in areas where transcriptomic sketching tended to over-sample (Extended Data Fig 8D).

4 Discussion

Our benchmarking of sketching algorithms on ST data reveals clear trade-offs between preserving transcriptomic diversity and maintaining tissue coverage. Purely expression-based methods such as Geosketch and scSampler excel at capturing global heterogeneity (i.e., minimizing transcriptional Hausdorff distance and enriching for rare cell states) but do so at the cost of distorting tissue-scale structure. In multiple real tissues (mouse ovary, MERFISH brain, human breast cancer, lung), these methods over-sampled regions of high expression variability (e.g., olfactory bulb, DCIS ducts) and under-sampled more homogeneous areas, leading to biases in downstream neighborhood analyses and local cell-type composition. Conversely, coordinate-only sampling uniformly covers the tissue footprint and achieves minimal coordinate Hausdorff distance, but fails to capture transcriptional extremes and sometimes introduces an edge-effect by over-selecting peripheral cells.

Leverage score methods strike a better balance, particularly when computed on the full cell-by-gene matrix or on a spatially smoothed SVD embedding. Across most datasets and evaluation metrics (i.e., the robust Hausdorff distances, adjusted Rand index of clustering, PCA loading drift, and local cell-type MSE), leverage score-based sketches closely recapitulated both the global transcriptomic landscape and the spatial organization of the tissue. In contrast to pure transcriptomic sketches, spatially-smoothed leverage scores attenuated sampling pattern bias, avoided edge effects, and maintained strong performance on rare cell recovery. Notably, uniform and coordinate sketches performed competitively on datasets with coarsely gridded spots (10x Visium HD and Visium-like simulations), suggesting that when measurement locations are regular and dense, simple schemes suffice.

Our findings underscore the importance of incorporating spatial context into sketch selection for ST studies. While transcriptomic diversity remains critical, especially when rare cell types or subtle expression gradients drive downstream discovery, disregarding spatial structure can lead to misleading inferences about tissue architecture. Spatially-aware sketches, implemented here via a spatial weight matrix and smoothed low-dimensional embeddings, represent a promising compromise: they preserve transcriptomic extremes and spatial uniformity while avoiding unintended edge biases.

Looking forward, these results motivate the development of principled sketching frameworks that explicitly optimize a joint objective over transcriptomic and spatial distances. Possible directions include adaptive weighting of expression versus coordinate-based sampling, multiscale tiling to mitigate grid-edge effects, and integration with biologically informed priors (e.g., known tissue landmarks). Such advances will enable efficient and unbiased down-sampling of ever-larger ST datasets, facilitating rapid exploratory analyses without sacrificing biological fidelity.

5 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 **ExperimentHub**, dataset # EH8567, sample ID 'IDC' (`https://www.bioconductor.org/packages/release/data/experiment/html/SubcellularSpatialData.html`).

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`.

Sagittal mouse brain:

We downloaded the mouse brain dataset from the Allen Brain Atlas via their **abc_atlas_access 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.

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-v4`.

Simulated data:

See the following Github for simulation datasets `https://github.com/Goods-Lab/Goods-Lab`

6 Funding:

NIH grants R35GM146586 (IKG, HRF), P20GM130454 (IKG, BAG, HRF) and P30CA023108 (HRF).

References

1. Joseph, V. R. Space-filling designs for computer experiments: A review. *Quality Engineering* **28.** Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08982112.2015.1100447, 28–35. ISSN: 0898-2112. https://doi.org/10.1080/08982112.2015.1100447 (2025) (Jan. 2016).
2. Chen, Y., Welling, M. & Smola, A. Super-Samples from Kernel Herding. en.
3. Barbian, M. H. & Assunção, R. M. Spatial subsemble estimator for large geostatistical data. *Spatial Statistics* **22,** 68–88. ISSN: 2211-6753. https://www.sciencedirect.com/science/article/pii/S2211675316300999 (2025) (Nov. 2017).
4. Johnson, M. E., Moore, L. M. & Ylvisaker, D. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference* **26,** 131–148. ISSN: 0378-3758. https://www.sciencedirect.com/science/article/pii/037837589090122B (2025) (Oct. 1990).
5. Clarkson, K. L. & Woodruff, D. P. Low-Rank Approximation and Regression in Input Sparsity Time. *J. ACM* **63,** 54:1–54:45. ISSN: 0004-5411. https://dl.acm.org/doi/10.1145/3019134 (2025) (Jan. 2017).
6. Miao, Z. *et al.* Putative cell type discovery from single-cell gene expression data. en. *Nature Methods* **17.** Publisher: Nature Publishing Group, 621–628. ISSN: 1548-7105. https://www.nature.com/articles/s41592-020-0825-9 (2025) (June 2020).
7. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15,** e8746. ISSN: 1744-4292. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6582955/ (2025) (June 2019).
8. Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* **20,** 40. ISSN: 1471-2105. https://doi.org/10.1186/s12859-019-2599-6 (2025) (Jan. 2019).
9. Song, D., Xi, N. M., Li, J. J. & Wang, L. scSampler: fast diversity-preserving subsampling of large-scale single-cell transcriptomic data. *Bioinformatics* **38,** 3126–3127. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btac271 (2025) (May 2022).
10. Hie, B., Cho, H., DeMeo, B., Bryson, B. & Berger, B. Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape. *Cell Systems* **8,** 483–493.e7. ISSN: 2405-4712. https://www.sciencedirect.com/science/article/pii/S2405471219301528 (2025) (June 2019).
11. Hao, Y. *et al. Dictionary learning for integrative, multimodal, and scalable single-cell analysis* en. Pages: 2022.02.24.481684 Section: New Results. Feb. 2022. https://www.biorxiv.org/content/10.1101/2022.02.24.481684v1 (2025).
12. DeMeo, B. & Berger, B. Hopper: a mathematically optimal algorithm for sketching biological data. *Bioinformatics* **36,** i236–i241. ISSN: 1367-4803. https://doi.org/10.1093/bioinformatics/btaa408 (2025) (July 2020).
13. Huang, L., Gong, W. & Chen, D. *scValue: value-based subsampling of large-scale single-cell transcriptomic data for machine and deep learning tasks* en. Jan. 2025. http://biorxiv.org/lookup/doi/10.1101/2025.01.10.632338 (2025).
14. Do, V. H., Elbassioni, K. & Canzar, S. Sphetcher: Spherical Thresholding Improves Sketching of Single-Cell Transcriptomic Heterogeneity. *iScience* **23,** 101126. ISSN: 2589-0042. https://www.sciencedirect.com/science/article/pii/S2589004220303114 (2025) (June 2020).
15. Ren, J. *et al.* A downsampling method enables robust clustering and integration of single-cell transcriptome data. *Journal of Biomedical Informatics* **130,** 104093. ISSN: 1532-0464. https://www.sciencedirect.com/science/article/pii/S1532046422001095 (2025) (June 2022).
16. Polański, K. *et al.* Bin2cell reconstructs cells from high resolution Visium HD data. *Bioinformatics* **40,** btae546. ISSN: 1367-4811. https://doi.org/10.1093/bioinformatics/btae546 (2024) (Sept. 2024).
17. Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. en. *Nature Biotechnology* **39.** Publisher: Nature Publishing Group, 313–319. ISSN: 1546-1696. https://www.nature.com/articles/s41587-020-0739-1 (2024) (Mar. 2021).

18. Janesick, A. *et al. High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue* en. Pages: 2022.10.06.510405 Section: New Results. Oct. 2022. https://www.biorxiv.org/content/10.1101/2022.10.06.510405v1 (2023).

19. Yao, Z. *et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain* en. Pages: 2023.03.06.531121 Section: New Results. Mar. 2023. https://www.biorxiv.org/content/10.1101/2023.03.06.531121v1 (2025).

20. Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. en. *Cell* **185,** 1777–1792.e21. ISSN: 00928674. https://linkinghub.elsevier.com/retrieve/pii/S0092867422003993 (2024) (May 2022).

21. *STOmicsDB: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization — Nucleic Acids Research — Oxford Academic* https://academic.oup.com/nar/article/52/D1/D1053/7416388 (2025).

22. Li, Y. *et al.* Spatiotemporal transcriptome atlas reveals the regional specification of the developing human brain. *Cell* **186,** 5892–5909.e22. ISSN: 0092-8674. https://www.sciencedirect.com/science/article/pii/S0092867423012333 (2025) (Dec. 2023).

23. Pan, J. *et al. Spatiotemporal transcriptome atlas of human embryos after gastrulation* en. Pages: 2023.04.22.537909 Section: New Results. Apr. 2023. https://www.biorxiv.org/content/10.1101/2023.04.22.537909v1 (2025).

24. Shi, H. *et al.* Spatial atlas of the mouse central nervous system at molecular resolution. en. *Nature* **622.** Publisher: Nature Publishing Group, 552–561. ISSN: 1476-4687. https://www.nature.com/articles/s41586-023-06569-5 (2025) (Oct. 2023).

25. *Single-cell spatial transcriptomic atlas of the whole mouse brain: Neuron* https://www.cell.com/neuron/fulltext/S0896-6273(25)00133-3 (2025).

26. Zhang, M. *et al.* Molecularly defined and spatially resolved cell atlas of the whole mouse brain. en. *Nature* **624.** Publisher: Nature Publishing Group, 343–354. ISSN: 1476-4687. https://www.nature.com/articles/s41586-023-06808-9 (2025) (Dec. 2023).

27. Huang, R. *et al. Single-cell and spatiotemporal profile of ovulation in the mouse ovary* en. Pages: 2024.05.20.594719 Section: New Results. June 2024. https://www.biorxiv.org/content/10.1101/2024.05.20.594719v4 (2024).

28. Bhuva, D. D. *et al.* Library size confounds biology in spatial transcriptomics data. *Genome Biology* **25,** 99. ISSN: 1474-760X. https://doi.org/10.1186/s13059-024-03241-7 (2024) (Apr. 2024).

29. Zhu, J., Shang, L. & Zhou, X. SRTsim: spatial pattern preserving simulations for spatially resolved transcriptomics. *Genome Biology* **24,** 39. ISSN: 1474-760X. https://doi.org/10.1186/s13059-023-02879-z (2025) (Mar. 2023).