TopoMetry systematically learns and evaluates the latent geometry of single-cell data

David S. Oliveira[1], Ana I. Domingos[1], and Licio A. Velloso[2]

[1]University of Oxford, UK
[2]University of Campinas, Brazil

Abstract

Reconstructing and investigating the geometry underlying data is a fundamental task in single-cell analysis, yet no unified framework exists to learn, evaluate, and diagnose representations that faithfully preserve it. We present **TopoMetry**, a geometry-aware framework that learns intrinsic coordinate systems directly from the data and refines them into high-fidelity *spectral scaffolds*. These scaffolds capture both local neighborhoods and global structure, supporting downstream analysis such as clustering and visualization. In benchmarks across diverse single-cell datasets, TopoMetry preserved geometry more reliably than standard workflows and revealed biological signals otherwise obscured, including unexpected transcriptional diversity among T cells and links between RNA-defined subpopulations and clonal expansion. The full analysis can be run with a single line of code to generate a comprehensive report, making the framework both powerful and accessible. Beyond individual findings, TopoMetry warrants a shift of focus from static two-dimensional projections to the systematic learning and evaluation of single-cell geometry itself, enabling more faithful exploration of cellular diversity.

Keywords: single-cell genomics; dimensional reduction; manifold learning; clustering; visualization.

1 Introduction

Single-cell genomics has enabled the systematic profiling of thousands to millions of individual cells, providing unprecedented insight into the diversity of cell types and states across tissues and organisms. These technologies generate high-dimensional data in which each cell is represented by measurements across tens of thousands of features, such as gene expression in single-cell RNA-seq (scRNAseq). Interpreting such data depends critically on computational analysis: the algorithms used to represent and compare cells directly shape the clusters, trajectories, and biological hypotheses that emerge from single-cell studies.

The prevailing analytical workflow begins by reducing dimensionality with Principal Component Analysis (PCA)[1], followed by the construction of a neighborhood graph for Leiden clustering and visualization with algorithms such as UMAP (Uniform Manifold Approximation and Projection)[2,3]. The "PCA → neighborhood graph → Leiden clustering and UMAP" pipeline has become the *de facto* standard in single-cell genomics[4] and underpins a wide range of downstream tasks, from cell annotation to trajectory inference and dataset integration. Its popularity reflects both its computational efficiency and its capacity to produce intuitive visual summaries of cellular heterogeneity.

Despite its wide adoption, the PCA→UMAP workflow rests on assumptions that are difficult to verify for most single-cell datasets[1,2]. The use of PCA presumes that biological variation is captured by global variance and that cell states can be represented as linear combinations of genes (Supplementary Figure S1a), while UMAP assumes that cells are uniformly sampled from a manifold with a constant local metric. Violations of these assumptions lead to similarity graphs and embeddings that distort cellular relationships, with direct consequences for clustering outcomes, lineage reconstructions, and biological interpretation. In addition, any two-dimensional projection will inevitably introduce distortions, which has fueled debate about how much trust should be placed in embeddings such as UMAP[5].

The field currently lacks a systematic evaluation of the standard PCA-based pipeline and its impact on the geometric fidelity of single-cell analyses. While benchmarking studies have compared clustering or integration algorithms, none have assessed whether the obtained representations preserve the original manifold structure that encodes cellular identities. In particular, there has been no unified framework for

quantifying how neighborhood graphs, diffusion processes, and embeddings diverge from the ground-truth geometry. Without such evaluation, distortions introduced at early stages of the pipeline may propagate through analyses, affecting biological conclusions in ways that are difficult to detect or correct.

Ideally, the field should adopt a theoretical framework for modeling single-cell data that minimizes assumptions while providing stable and intuitive representations across diverse datasets. Such a framework should (i) treat single-cell data as samples from manifolds with heterogeneous and potentially disconnected supports, (ii) explicitly estimate intrinsic dimensionalities rather than fixing representation size arbitrarily, and (iii) evaluate how well similarity graphs and embeddings preserve the multiscale geometry of the data. By grounding analyses in rigorous geometric principles, such a framework would enable more reliable inference of cell states, lineages, and transitions (Supplementary Figure S1b).

To address this gap, we developed TopoMetry, a framework for single-cell geometric analysis designed to preserve intrinsic data structure independently of distributional or sampling artifacts. TopoMetry considers the existence of multiple manifolds in each single-cell dataset, each representing a distinct macro-population or lineage, and decomposes the joint geometry of the data into a *spectral scaffold*: a set of hundreds of components that, analogous to the harmonics of a Fourier transform, collectively capture the global and local structure of the dataset. A refined cell–cell similarity graph is then constructed from this scaffold and used for downstream analyses. TopoMetry also introduces new geometry preservation metrics and distortion visualization tools, making it possible to quantitatively and qualitatively evaluate representations and visualizations.

We evaluate TopoMetry against the prevailing PCA-based standard (Supplementary Figure S1c) in an extensive collection of single-cell datasets. Across all datasets, the proposed workflow yielded representations that better capture the original data geometry and clustering results better suited for the detection of rare cell populations. We found that the poor performance of PCA on preserving geometry was associated with failure to explain variance, highlighting its inappropriateness for single-cell analyses. The difference in performance was particularly noticeable in datasets containing T cells, for which the TopoMetry analysis identifies a much greater number of T cell clusters than the standard workflow. Using paired TCR and RNA data, we found that these additional clusters are associated to TCR clonality, and explore

the geometrical properties of clonal expansion in single-cell data. Collectively, our results demonstrate the superiority of a geometric framework for single-cell analysis over the prevailing standards.

2 Results

2.1 A framework to systematically learn and evaluate manifolds in single-cell data

To study cellular diversity, single-cell data are often described as points scattered across a high-dimensional space of measured features (such as gene expression in scRNAseq). The challenge is to uncover the hidden structure—the manifold of cell identities—that explains how these points are organized into cell types, lineages, and states. Our goal with TopoMetry is to provide a framework that learns this manifold directly and builds on its geometrical properties, without relying on restrictive assumptions on its topology, data distribution or intrinsic dimensionality.

TopoMetry takes as input a scaled (standardized or Z-score normalized) matrix of features by cells (for example, genes per cell). It then builds a neighborhood graph that connects each cell to its k most similar counterparts, which is transformed into a similarity matrix using decay-adaptive, manifold-aware kernels that account for local intrinsic dimensionality and sampling density (Figure 1a). These kernels reduce density-driven bias and support the construction of Laplacian-type and diffusion operators that approximate manifold geometry (Figure 1a, Supplementary Figure S1d)[6–9].

TopoMetry proceeds by eigendecomposing these operators into hundreds of orthogonal components that together define a *spectral scaffold* (Figure 1a). Each component acts as a latent coordinate capturing a distinct mode of variation. A *multiscale spectral scaffold* aggregates coordinates across diffusion times to reconcile local neighborhoods and long-range global organization. The number of scaffold components is determined automatically by estimating intrinsic dimensionality (I.D.), rather than fixed a priori[10,11] (Figure 1b). The scaffold thus constitutes a latent space that encodes the geometrical properties of the data, is robust to the choice of k and to batch effects (Suppl. Figure S1e), and supports the construction of refined similarity graphs and diffusion operators from the scaffold itself, capturing the "geometry of

the geometry". These refined graphs and operators serve as high-fidelity inputs for downstream tasks such as clustering and visualization (Figure 1a), and can also be exploited to evaluate how information propagates across the manifold, enabling graph-based filtering of categorical or continuous signals (Figure 1c) and data-smoothing operations such as imputation and denoising (Figure 1d). Finally, TopoMetry leverages the Riemannian metric to provide manifold diagnostics that reveal contraction, expansion, and local distortions in two-dimensional maps, allowing geometry-aware interpretation (Figure 1e; cf.[12]).

The proposed approach has several conceptual advantages over current standards. First, it assumes only that the data lie approximately on a manifold—a biologically meaningful premise consistent with the Waddington epigenetic landscape[13] and already implicit in many downstream analyses such as lineage inference and pseudotime estimation. Second, it eliminates the need to pre-select the number of components by estimating I.D. directly and identifying gaps in the scaffold eigenspectrum (Figure 1a). Third, it evaluates fidelity with quantitative metrics rather than relying solely on 2-D plots. Finally, by coupling manifold-based analysis with modern graph-layout techniques (e.g., UMAP[2], PaCMAP[14]), TopoMetry moves beyond fixed pipelines toward systematic, data-driven discovery. The full analysis can be executed with a single line of code, producing a PDF report while remaining fully customizable and compatible with `AnnData`/Scanpy[4]. In summary, TopoMetry brings manifold learning and geometric analysis into single-cell genomics in a way that is rigorous, computationally efficient (Supplementary Figure S1f), and accessible to researchers from all backgrounds.

2.2 Evaluating geometry-preservation across representations

A central question in single-cell analysis is whether a low-dimensional representation preserves the original geometry of cellular relationships, such as multiple cellular lineages cohabiting the same gene expression space. Such geometry is encoded by the diffusion operator built from the initial neighborhood graph[7,15]. To evaluate geometry preservation, we introduce a family of *operator-native* metrics that compare representations at the level of their learned diffusion kernels/transition operators rather than at the level of raw Euclidean coordinates. This choice is deliberate: many downstream tasks (visualization, clustering, trajectories, pseudotime) implicitly act

on a graph or Markov operator, so fidelity should be judged in that native space.

First, we quantify local neighborhood agreement in two complementary ways. The **Sparse Neighborhood F1** (`P-F1@k`) measures set overlap between the top-k transition supports of two operators (insensitive to weights, sensitive to who the neighbors are). The **row-wise Jensen–Shannon similarity** (`P-JS`) compares each row as a probability distribution over neighbors (sensitive to weights, i.e. how strongly neighbors are connected). Together, `P-F1@k` and `P-JS` capture whether local neighborhoods are preserved *and* whether their transition probabilities remain consistent. Second, we assess meso- to global-scale geometry using diffusion coordinates. The **Spectral Procrustes** score (`SP`) aligns truncated diffusion maps (at multiple diffusion times) via orthogonal Procrustes and reports an R^2 goodness-of-fit; high values indicate that two operators induce essentially the same geometry up to a rotation. Collectively, these measures probe complementary facets of fidelity: neighborhood composition, edge weights, global alignment, and connectivity structure. Importantly, all metrics compare the diffusion operator of the projection built by each method to the native diffusion operator of normalized gene expression, ensuring a fair, task-relevant comparison independent of a particular 2-D layout.

Benchmark datasets. To apply the defined geometry preservation metrics to representations learned with the current standard workflow or TopoMetry, we curated a collection of 68 scRNAseq datasets spanning multiple organs, tissues, and species (humans and mice). The corpus comprises atlas-scale datasets with a complex hierarchical structure, as well as focused studies on rare or transitional populations, thereby representing single-cell genomics data in general (Figure 2a, Suppl. Table 1).

Evaluating workflows. We applied the metrics to representations produced by (i) the standard workflow (graph built on the PCA space[1,16]), (ii) a "pure" standalone UMAP approach (graph built directly from the high-dimensional space[2]), (iii) scVI (a widely used variational framework[17,18]), and (iv) TopoMetry (Figure 2a). We then computed the geometry preservation scores. For all analyzed datasets, the spectral scaffolds learned by TopoMetry scored consistently higher than the latent spaces learned by PCA or scVI (Figure 2b). In agreement with these findings, 2-D visualizations obtained with TopoMetry also performed better than those obtained with PCA or scVI (Figure 2c). Noticeably, "pure" UMAP visualizations had intermediary scores be-

tween TopoMetry and PCA→UMAPs, highlighting that PCA usage is not universally appropriate in single-cell genomics. In sum, these results indicate that TopoMetry's manifold-centered approach yields representations that better preserve the original geometrical properties of single-cell data, thus being better suited for downstream analyses and biological interpretation.

Next, we further investigated the reasons underlying the poor performance of PCA-based graphs and visualizations. We found that PCA performance correlates with its ability to explain most of the original variance (Suppl. Fig. S2a). Surprisingly, the total variance explained by PCA was remarkably low across datasets (as little as 20%). Total explained variance decreased as the number of highly-variable genes (i.e. dimensionality) increased, and averaged only ≈36% across all datasets with default gene selection (Suppl. Fig. S2b). Failure to explain variance could not be attributed to insufficient components, as demonstrated by eigenspectra (Suppl. Fig. S2c) and cumulative explained variance curves (Suppl. Fig. S2d), where an *ad hoc* "elbow point" was consistently found around 30–50 components. Total explained variance was weakly associated with cell number (Suppl. Fig. S2e). Such low values of explained variance are widely known as a hallmark of highly non-linear systems in the broader machine-learning community[19,20], albeit apparently unnoticed within the single-cell niche. Collectively, these observations suggest that PCA fails on single-cell data due to intrinsic properties of the data, such as non-linearity, which can be highly dataset-specific. These limitations highlight the need for approaches that move beyond variance-based assumptions and instead preserve manifold geometry directly. We next asked whether such geometry-aware representations, as learned by TopoMetry, could improve the inference of cellular lineages across both small developmental systems and atlas-scale datasets.

2.3 TopoMetry resolves cellular lineages in tiny and atlas-scale systems

A natural application of such geometry-aware representations is the reconstruction of developmental trajectories. To test this, we first applied TopoMetry to scRNAseq data from the developing murine pancreas, a well-established benchmark in RNA velocity studies[21,22]. Standard PCA→UMAP embeddings (Figure 3b, Suppl. Fig. S3a–b) identified broad lineages but failed to represent the cell cycle structure, placing mitotic cells into ambiguous positions along differentiation axes. In contrast, TopoMetry

projections (TopoMAP, Figure 3a,c) reconstructed a closed-loop geometry for the cell cycle and positioned proliferating cells next to other cycling populations, faithfully reflecting the underlying manifold. RNA velocity streamlines confirmed the accuracy of this representation (Figure 3d), while the PCA-based embedding produced vector fields inconsistent with known lineage relationships (e.g. epsilon cells giving rise to beta cells[23]). Overlaying individual scaffold components on TopoMAP projections further showed that each component captures a distinct facet of the manifold, from global developmental trajectories to localized cellular states (Suppl. Fig. S3c).

We demonstrated that TopoMetry can scale to organism-wide atlases by analyzing the Mouse Organogenesis Cell Atlas (MOCA), a compendium of ∼1.3 million cells spanning murine embryonic development[24]. Both PCA→UMAP and TopoMetry separated the originally annotated differentiation trajectories and cell types into distinct regions of the manifold (Figure 3e–f). However, TopoMetry revealed a much richer hierarchy of sub-trajectories, capturing ∼380 subpopulations compared to the 56 originally described (Figure 3g). These included a particularly fine-grained resolution of neuronal lineages, consistent with the known diversity of the developing nervous system. By encoding long-range and local geometry simultaneously, the TopoMAP embedding reconstructed refined developmental paths that remained unresolved in PCA-based visualizations.

Together, these analyses show that TopoMetry faithfully represents cellular dynamics across scales: from local cycles and branching lineages in small datasets to hundreds of coexisting trajectories in million-cell atlases. This makes TopoMetry an effective framework for studying lineage inference in direct connection with the Waddington epigenetic landscape[13].

2.4 TopoMetry unveils unexpected transcriptional diversity of T cells

When applying TopoMetry across public single-cell datasets, we repeatedly observed a striking pattern in peripheral blood mononuclear cell (PBMC) data: TopoMetry consistently revealed a much richer diversity of T cell subpopulations than those recovered by the standard PCA→UMAP workflow. This pattern first became evident in the widely used *pbmc68k* dataset, which contains ∼68,000 PBMCs from a healthy donor. While TopoMetry uncovered an unexpectedly fine-grained landscape with close to one hundred distinct T cell clusters (Figure 4a), PCA-based UMAP repre-

sented T cells as a handful of broad clusters (Figure 4b). Importantly, major immune cell classes were similarly well separated in both approaches, underscoring that the difference lies specifically in the resolution of T cell diversity.

The discrepancy was further highlighted when examining marker gene expression. T cell clusters obtained with PCA-based graphs presented non-specific markers (Figure 4c), whereas TopoMetry clusters exhibited highly specific expression signatures (Figure 4d). These included canonical T cell markers, interleukin receptors, and T cell receptor (TCR) genes. Such specificity indicates that TopoMetry is capturing biologically meaningful substructure rather than technical noise. Additional TopoMetry projections consistently separated these substructures (Suppl. Fig. S4a).

Independent workflows corroborated these findings. A "pure" UMAP approach, in which clustering and graph construction are performed directly on the high-dimensional gene expression matrix, recovered a subset of the additional T cell substructures but not their full extent (Figure 4e). Similarly, clustering on a standalone kNN graph without PCA partially agreed with TopoMetry's results (Figure 4e, Suppl. Fig. S4b). These comparisons suggest that the additional diversity is present in the original data but systematically missed when PCA is imposed as the first step of the analysis.

We next asked whether technical artifacts could explain these unexpected T cell clusters. Doublet detection using Scrublet[25] showed no doublet enrichment in these clusters (Suppl. Fig. S4c–d), arguing against artifactual origins. Finally, distortion diagnostics revealed that the PCA→UMAP embedding expands large portions of the T cell manifold, while TopoMetry's refined graphs contract the $CD4^+$ T cell region (Figure 4f), consistent with genuine underlying structure rather than projection artifacts. Together, these results establish that TopoMetry reliably detects transcriptional heterogeneity in T cells that is largely obscured by conventional workflows.

The fine-grained T cell structure revealed in pbmc68k can be further dissected through the spectral scaffold itself. When visualizing the first 40 scaffold components, the earliest components captured global immune class separations and long-range relationships, while subsequent components progressively refined local features and delineated specific T cell subsets (Suppl. Fig. S5). This decomposition illustrates how TopoMetry resolves the manifold one region at a time, with each component encoding a distinct aspect of cellular heterogeneity.

We next examined whether this phenomenon generalizes to PBMC datasets from

disease contexts. In case studies of systemic lupus erythematosus, dengue fever, and multiple sclerosis, the standard PCA→UMAP workflow produced only broad T cell clusters with largely non-specific markers (Suppl. Fig. S6a–c). Standalone UMAP approaches recovered a subset of additional populations, but TopoMetry consistently revealed the full breadth of T cell diversity, detecting numerous clusters supported by highly specific marker gene expression. These results suggest that the richer representation of T cells is not dataset-specific but extends across diverse biological contexts.

Closer inspection of marker signatures indicated that the identity of these T cell clusters varied across donors and conditions. Some clusters reflected canonical CD4 or CD8 states, whereas others were distinguished by activation signatures or effector molecules. This variability points to genuine donor-specific biology rather than technical artifacts. In particular, the heterogeneity is consistent with the highly individualized T cell receptor (TCR) repertoire: each individual carries a unique set of clonotypes, which could lead to subtle transcriptional imprints that are easily masked by linear or density-biased workflows. Additionally, recent studies suggest T cells sharing the same TCR clonotype (or recognizing the same epitope) tend to display similar transcriptional phenotypes, and joint representations of both RNA and TCR data more clearly distinguishes antigen-specific subpopulations than those based only on the transcriptome[26–28]. These observations motivated us to test whether the additional T cell resolution provided by TopoMetry is associated with TCR clonotypes.

2.5 TopoMetry resolves clonal dynamics of T cells from RNA expression

To investigate if the additional clusters of T cells exclusively discovered by TopoMetry are associated with TCR clonal dynamics, we analyzed two public datasets where scRNAseq was paired with TCR sequencing: the ECCITE-TCR study of $CD8^+$ T cell responses to SARS-CoV-2 vaccination and infection[27], and the T cell compartment of the Tissue Immune Cell Atlas (TICA,[29]), which combines scRNAseq and VDJ-seq across multiple human tissues.

SARS-CoV-2 vaccination data. In the ECCITE-TCR dataset, TopoMetry once again revealed a rich diversity of $CD8^+$ subpopulations (Fig. 5a) that were absent from

PCA→UMAP representations (Suppl. Fig. S7a) and presented highly specific markers (Fig. 5b). Many of these additional clusters corresponded to effector memory (TEM) and central memory (TCM) populations, as indicated by original cell type annotations (Fig. 5c). These clusters were associated with less frequent clonotypes (Fig. 5d), suggesting that the transcriptional variation uncovered by TopoMetry reflects underlying clonal dynamics. Cell-cycle trajectories of proliferating antigen-specific $CD8^+$ lymphocytes were also faithfully encoded by the TopoMetry, producing coherent loop-like structures in 2-D projections (Fig. 5e), whereas the same geometry was broken or distorted in PCA-based even with the use of joint RNA–TCR representations (Suppl. Fig. S7a–g). Beyond visualizations, the spectral scaffold learned by TopoMetry captured both global lineage structure and local expansions (Suppl. Fig. S8a), while estimates of local intrinsic dimensionality remained stable across the manifold (Suppl. Fig. S8b). Mapping clone-size distributions onto the embedding further confirmed that small and rare clonotypes localized to the additional TCM/TEM clusters, while hyperexpanded clones occupied distinct regions of the manifold (Suppl. Fig. S8c).

Tissue Immune Cell Atlas. We next examined the TICA dataset[29], comprising paired RNA and VDJ profiles of human T cells across tissues and donors. As in previous analyses, TopoMetry identified numerous clusters within the $CD4^+$ compartment that were either merged or absent in the PCA-based workflow and original cell type annotations (Fig. S9a-b). Using the 'scirpy' TCR analysis toolkit ([30]), we queried epitope databases for antigens matching the TCR repertoire of this dataset, and found that the additional clusters uncovered by TopoMetry recognized antigens from specific species (Fig. S9c). Strikingly, we found that one of TopoMetry's clusters corresponded to a single clonotype of tissue-resident memory T $CD8^+$ cells with specificity for SARS-CoV-2 antigens (Fig. S9c). To further investigate clonal expansion dynamics, we visualized the 30 largest clonotypes on TopoMAPs, and found a strong correspondence between these clonotypes and the fine-grained clusters structure identified by TopoMetry (Fig. S9c). Several of these clusters corresponded to hyperexpanded clones, including the population recognizing SARS-CoV-2 antigens (Fig. S9e). Clonal expansion analysis further confirmed that hyperexpanded clones mapped to distinct $CD8^+$ and NK populations, whereas the TopoMetry-specific $CD4^+$ clusters corresponded to clonotypes with modest expansion (Fig. S9f). Finally, to evaluate the

specificity of the cluster-clonotype associations between workflows, we performed TCR repertoire overlap analysis, revealing that the PCA-based workflow and original cell type annotations resulted in strong repertoire overlap (Suppl. Fig. S9f). The "pure" standalone kNN workflow (Suppl. Fig. S9g) presented clearly weaker overlap, followed by TopoMetry (Suppl. Fig. S9h), which presented minimal overlap. These results demonstrate that TopoMetry's clustering results achieve a strong association with clonotypes inferred from TCR sequence similarity, in contrast to the poor performance of the PCA-based approach and the intermediate performance of standalone kNN graphs.

Summary. Collectively, these results show that TopoMetry geometry-aware workflow uncovers transcriptional imprints of TCR clonotypes directly from RNA expression. Across both SARS-CoV-2 vaccination and atlas-scale datasets, the additional T cell clusters revealed by TopoMetry corresponded to clonally distinct populations, many of which were invisible to PCA-based workflows. This finding closes the loop on our incidental observation in PBMC data: the unexpected T cell diversity resolved by TopoMetry reflects the clonal architecture of the immune repertoire. More broadly, these analyses highlight how preserving geometry enables faithful recovery of biological structure, bridging transcriptional diversity with clonotype identity in a way that linear or density-biased methods fail to achieve.

3 Discussion

In this work, we introduced TopoMetry, a framework for geometry-aware analysis of single-cell data that systematically learns manifold structure directly from the data. By constructing similarity graphs with kernels adaptive to local density and intrinsic dimensionality, TopoMetry extracts *spectral scaffolds* that capture the full range of geometric variation present in cellular systems. These scaffolds serve as the basis for refined diffusion graphs that preserve both local neighborhoods and long-range relationships, enabling downstream tasks such as clustering, visualization, imputation, and lineage inference in a way that is faithful to the original geometry. Crucially, TopoMetry also provides operator-native fidelity scores and distortion diagnostics, allowing users to evaluate when and how embeddings diverge from the underlying manifold. Applications across diverse datasets demonstrated that TopoMetry yields

stable, faithful, and interpretable representations, often revealing subtle yet biologically meaningful structures invisible to existing workflows. Most notably, TopoMetry uncovered an unexpectedly rich transcriptional landscape of T cells in peripheral blood, linked to TCR repertoires and clonal expansion, demonstrating the biological insights that become accessible when geometry is preserved.

The results presented here also expose fundamental flaws in the current PCA→UMAP standard. The reliance on PCA as a first analysis step imposes restrictive assumptions: that linear variance captures the relevant biological signal and that the dataset is globally well-approximated by a low-rank linear subspace[1,16]. Our systematic evaluation shows that these assumptions fail dramatically in single-cell genomics: PCA typically explains less than 40% of the variance in standard settings, and the variance explained decreases as the number of highly variable genes grows. These results mean that PCA discards the majority of biological signal before manifold learning even begins. The consequence is clear: PCA-based embeddings systematically collapse diverse cell populations into broad, poorly resolved clusters, as observed with datasets containing T cells, where lineage- and clonotype-associated subpopulations vanish into a single homogeneous cloud. UMAP then compounds the issue by optimizing the layout from the already distorted PCA-based neighborhood graph. The field has relied on this standard for years–not because it is faithful, but because it is convenient. Our results demonstrate that such convenience comes at the cost of accuracy, discarding most of the underlying information and obscuring biologically meaningful structure.

TopoMetry does not simply extend previous diffusion-based methods such as Diffusion Maps[7] or PHATE[31]; it provides a systematic and general framework for geometric analysis and manifold learning in single-cell data. Diffusion Maps introduced the idea of capturing data geometry across scales, while PHATE emphasized visualization of diffusion potentials. TopoMetry unifies and generalizes these insights into a pipeline that is both rigorous and practical: adaptive kernels ensure robustness to uneven sampling, spectral scaffolds encode manifold structure at data-driven resolutions, refined operators capture "the geometry of the geometry," geometry preservation metrics evaluate the learned representations, and Riemannian diagnostics offer principled tools for assessing distortion in visualizations. This combination of theory and practice positions TopoMetry as a distinct step forward. Importantly, its impact is already visible in the field: Hale *et al.* recently used TopoMetry to detect morphologically distinct T cell populations in a high-content imaging study published

in *Science*[32], and Tedeschi *et al.* integrated it into their analysis of multimodal single-cell atlases[33]. That independent groups are already leveraging TopoMetry underscores its utility and relevance.

As with any method, TopoMetry has limitations. Its current implementation is optimized for exploratory analysis rather than for online mapping of new cells, which restricts its use in scenarios requiring direct projection into precomputed manifolds. Computational cost is another consideration: although scalable to millions of cells, geometric analysis is still more computationally intensive than PCA. Finally, while our discovery of clonally linked T cell populations highlights the method's ability to reveal new biology, these findings call for follow-up experimental validation. Encouragingly, studies employing joint representations learned from paired TCR and RNA data have identified populations that resemble those identified by TopoMetry[26,27], indicating that these arise from reproducible biological signals.

Future work will extend TopoMetry's geometric-centered vision in several directions. Integrating geometric autoencoders[34] could enable inverse mapping and online embedding of new cells. Extensions to multimodal data are natural, particularly for integrating RNA, chromatin, protein, and receptor information into unified geometry-aware representations[35]. Biologically, TopoMetry opens new opportunities to study T cell clonality, lineage diversification, and other processes where subtle geometric variation encodes key biological signals. More broadly, the framework could be applied to any high-dimensional single-cell modality, from epigenomics to imaging cytometry, wherever faithful preservation of geometry matters.

In sum, TopoMetry represents a conceptual shift: a framework that places geometry at the center of single-cell analysis and a tool for discovery, capable of revealing structure even in datasets that have already been thoroughly examined by the community. The rich T cell diversity uncovered in PBMC data, long considered a "solved" dataset and ubiquitously used for tutorials and demonstrations, illustrates this power. Far from being exhausted, such datasets are valuable resources of potentially new biological insight if analyzed with appropriate geometric tools. TopoMetry provides these tools, and in doing so, it sets the stage for a new generation of geometry-aware single-cell genomics.

4 Methods

4.1 TopoMetry

4.2 TopoMetry: Topological geoMetry

Manifold hypothesis and sampling model. Let $\mathscr{M} \subset \mathbb{R}^D$ be a compact, d-dimensional C^2 Riemannian submanifold (possibly with boundary) with metric g induced by the ambient Euclidean metric and volume measure $\mathrm{d}V_g$. We observe points $x_1, \ldots, x_N \in \mathbb{R}^D$ drawn i.i.d. from a distribution supported on (or in a tubular neighborhood of) $\mathscr{M}$. In the on-manifold case,

$$x_i \sim q\,\mathrm{d}V_g, \qquad q : \mathscr{M} \to \mathbb{R}_{\geq 0} \text{ unknown (not necessarily uniform).}$$

In the near-manifold case, $x_i = \xi_i + \varepsilon_i$ with $\xi_i \sim q\,\mathrm{d}V_g$ on $\mathscr{M}$ and small ambient noise ε_i supported in a tubular neighborhood; when needed we write $\pi(x_i) \in \mathscr{M}$ for the nearest-point projection and assume $\pi_{\#}\mu = q\,\mathrm{d}V_g$.

Overview. We adopt the manifold hypothesis in a form tailored to single-cell data: the observed profiles $x_1, \ldots, x_N \in \mathbb{R}^D$ are sampled on (or very near) a d-dimensional Riemannian submanifold $\mathscr{M} \subset \mathbb{R}^D$ according to an unknown, not-necessarily uniform sampling density q with respect to the manifold's volume measure. Our goal is to recover informative, low-distortion coordinates that reflect the *intrinsic* geometry of $\mathscr{M}$ rather than artifacts of the ambient space.

A principled way to expose this geometry is via the Laplace–Beltrami operator (LBO), Δ_g, the natural generalization of the Laplacian to curved spaces. Its eigenfunctions act like a "Fourier basis" on $\mathscr{M}$: low-frequency eigenfunctions capture broad organization, while higher frequencies resolve finer structure. In practice Δ_g is not observed directly; instead, methods such as Laplacian Eigenmaps and Diffusion Maps[6,7] build a *graph Laplacian* from an affinity (similarity) graph on the samples. This graph operator converges to Δ_g, especially when affinities are computed with *manifold-adaptive kernels* that correct for nonuniform sampling q[9].

TopoMetry builds on this theory by constructing a *spectral scaffold*: an orthonormal eigenbasis of graph-LBO eigenvectors that provides data-driven, intrinsic coordinates. Each eigenvector can be read as a smooth "mode of variation" across cells, and the collection of the leading modes spans a coordinate system that encodes both

local neighborhoods and global relationships. To avoid choosing a single resolution, TopoMetry forms a *multiscale spectral scaffold* by reweighting these modes across diffusion times (powers of the diffusion operator), thereby blending fine-grained and long-range structure in a single representation. The size of the scaffold is selected automatically from intrinsic-dimensionality estimates, so that subtle transcriptional signals and broad phenotypic organization are both retained without ad hoc parameter tuning.

Finally, TopoMetry reconstructs a neighborhood graph *in the scaffold space* and learns the corresponding graph LBO of the spectral scaffold itself—capturing the "geometry of the geometry." The resulting operators (diffusion potentials and refined similarity graphs) serve as high-fidelity inputs for downstream analyses such as clustering, trajectory inference, pseudotime, and layout optimization (e.g. MAP, PaCMAP). In short, TopoMetry progressively denoises and sharpens manifold structure across scales, yielding coordinates that are stable, interpretable, and faithful to the intrinsic geometry of single-cell atlases.

Computational implementation. TopoMetry was designed for ease of use and computational efficiency. It integrates popular approximate nearest-neighbor (ANN) libraries such as NMSlib and hnswlib. Most graph operations are performed on sparse matrices and scale efficiently with multithreading. Each workflow step is implemented as a modular scikit-learn-style transformer, enabling custom pipelines. Core modules include estimators of neighborhood graphs, affinity kernels and operators, eigendecompositions, and layout optimizations. Analyses are orchestrated by the `TopOGraph` class. The modular design allows seamless integration with machine-learning and single-cell analysis pipelines. Additional utilities estimate intrinsic dimensionalities, evaluate embeddings, and visualize results. High-level APIs simplify usage: a single line of code can run a complete analysis and output a comprehensive report, including direct compatibility with `Scanpy` ([4]) and the broader Python ecosystem for single-cell analyses. The source code is available at https://github.com/davisidarta/topometry. Documentation and tutorials are available at https://topometry.readthedocs.io/en/latest/.

Assumed input. TopoMetry expects as input a cell-by-feature matrix that has been scaled or standardized, typically by Z-score normalization of gene expression. This

preprocessing ensures that each feature contributes comparably to similarity calculations and avoids dominance by highly expressed or variable genes. A standardized input is also advantageous because it allows cell–cell similarity to be computed directly from feature correlations. In practice, this is implemented efficiently using cosine distances, which are equivalent to correlation distances on standardized data. »>In single-cell genomics, the input is usually subset to include only highly variable features, and we employed such standard preprocessing with default parameters in all demonstrated analysis.

Distance calculation. Distances are computed differently depending on the stage of the workflow. For building the initial spectral scaffold, we employ cosine distance on the normalized data, which is robust to differences in sequencing depth and scale and corresponds to the correlation distance if the input is standardized. The cosine (or correlation) distance between two cells $x_i, x_j \in \mathbb{R}^d$ is defined as:

$$d_{\cos}(x_i, x_j) \;=\; 1 - \frac{\langle x_i, x_j \rangle}{\|x_i\|_2 \, \|x_j\|_2},$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product.

For refined graphs built on the scaffold ($P_{\text{of_Z}}$), we use Euclidean distance in the learned spectral coordinates, which provides a natural metric for diffusion geometry. The Euclidean distance between two spectral embeddings $z_i, z_j \in \mathbb{R}^r$ is defined as:

$$d_{\text{Euc}}(z_i, z_j) \;=\; \sqrt{\sum_{k=1}^{r} \left(z_{ik} - z_{jk}\right)^2}.$$

Kernels and affinity estimation. Let $\mathscr{M} \subset \mathbb{R}^M$ be (near) a Riemannian submanifold with unknown sampling density q. For each sample $x_i \in \mathbb{R}^M$, denote by $\text{nbrs}(x_i)$ the ordered set of its k-nearest neighbors under a chosen dissimilarity $d(\cdot, \cdot)$ (Euclidean or cosine-based; see "Cosine/correlation geometry" below). The affinity graph $W \in \mathbb{R}_{\geq 0}^{N \times N}$ is built from locally rescaled distances and then symmetrized.

Adaptive scaling (per-point bandwidth). For robustness to heterogeneous sampling, a local bandwidth σ_i is estimated from the median neighbor distance of x_i:

$$\sigma_i \;=\; d\Big(x_i, \text{nbrs}_{(m)}(x_i)\Big), \qquad m = \left\lfloor \frac{k}{2} \right\rfloor,$$

i.e. the m-th order statistic of $\{d(x_i,x_j) : x_j \in \mathrm{nbrs}(x_i)\}$. This σ_i encodes local crowding (small in dense regions, large in sparse ones) and stabilizes locality.

Cosine/correlation geometry and angularization. When using the cosine metric, we convert cosine *distance* $d_{ij}^{\cos} = 1 - \cos\angle(x_i,x_j) \in [0,2]$ to an angle $\theta_{ij} \in [0,\pi]$,

$$\theta_{ij} \;=\; \arccos\big(1 - d_{ij}^{\cos}\big) \;=\; \arccos\Big(\tfrac{\langle x_i,x_j\rangle}{\|x_i\|_2\,\|x_j\|_2}\Big),$$

and use $d_{ij} := \theta_{ij}$ instead of the Euclidean distance. On feature-wise standardized (z-scored) inputs, $\cos\angle(x_i,x_j)$ equals the Pearson correlation, so $\theta_{ij} = \arccos(\rho_{ij})$ places similarities on the unit sphere ("correlation geometry"). Distances are numerically clipped to their natural ranges $[0,\pi]$ (angular) or $[0,2]$ (cosine), and optionally squared (default), before kernelization.

Bandwidth-adaptive (variable) kernel. We first form a left-normalized, variable-bandwidth dissimilarity

$$\widetilde{d}_{ij} \;=\; \frac{d_{ij}}{\sigma_i + \varepsilon}, \qquad \varepsilon > 0 \text{ small},$$

which rescales each row by its local scale σ_i. The basic affinity is then

$$W_{ij}^{(\mathrm{vb})} \;=\; \exp\big(-\,\widetilde{d}_{ij}^{\,p}\big), \qquad p = 2 \text{ by default (square distances)}.$$

Finally, we symmetrize $W \leftarrow (W + W^\top)/2$. This row-wise scaling followed by symmetrization closely approximates classical (σ_i,σ_j)-variable bandwidth kernels while being simple and fast to compute.

Neighborhood density and "pseudomedian". From $\{\sigma_i\}_{i=1}^N$ we derive a continuous density proxy ω_i by linearly mapping the interval $[\min_j \sigma_j, \max_j \sigma_j]$ to $[2,k]$:

$$\omega_i \;=\; \mathrm{linmap}\Big(\sigma_i;\, [\min\sigma, \max\sigma] \to [2,k]\Big).$$

Large ω_i (near k) indicates sparsity; small ω_i (near 2) indicates locality within dense regions. This quantity can guide mild neighbor-set expansion in under-sampled areas (optionally replacing k with $k' \geq k$) and parameterizes adaptive decay below.

Adaptively decaying kernel (density-dependent tails). To slow decay in sparse regions and sharpen it in dense ones, we make the decay exponent a function of ω_i. Define

$$\eta_i \;=\; 2^{\frac{k-\omega_i}{\omega_i}} \in [1,\infty),$$

so that $\eta_i \approx 1$ in sparse areas ($\omega_i \approx k$) and η_i grows as neighborhoods become denser. Using the same left-normalized $\widetilde{d}_{ij}$, we set

$$W_{ij}^{(\mathrm{ad})} = \exp\Big(-\widetilde{d}_{ij}^{\,p_i}\Big), \qquad p_i = \begin{cases} 2 & \text{(standard, no adaptive decay)}, \\ 2\,\eta_i & \text{(adaptive decay with squared distances)}. \end{cases}$$

Equivalently, with squared distances disabled one uses $p_i = \eta_i$. The resulting kernel decays gently across sparse regions (preserving informative long-range neighbors) and more steeply inside dense regions (preventing oversmoothing). As above, we symmetrize W at the end.

Fixed-bandwidth (global) kernel. For comparison or when desired, a global scale $\sigma > 0$ yields

$$W_{ij}^{(\mathrm{fb})} = \exp\Big(-(d_{ij}/\sigma)^p\Big), \qquad p = 2 \text{ by default.}$$

Summary. TopoMetry provides: (i) a fixed-bandwidth Gaussian/RBF kernel; (ii) a variable, per-point bandwidth kernel using median-k scaling; (iii) an optional neighborhood-expansion variant driven by the density proxy ω_i; and (iv) an *adaptively decaying* kernel with density-dependent exponents. Distance preprocessing supports Euclidean geometry and correlation/cosine geometry via angularization. These design choices make affinity estimation sensitive to manifold structure while robust to heterogeneous sampling, providing a stable foundation for subsequent Laplace–Beltrami operator approximations and diffusion-based embeddings.

Intrinsic dimensionality estimation. Intrinsic dimensionality (i.d.) can be loosely defined as the minimum number of parameters needed to accurately describe a high-dimensional system. Multiple notions exist (e.g., local vs. global i.d.), and accurate estimation remains an active area of research. Estimating the global i.d. of a dataset is closely related to estimating the dimensionality of its underlying manifold, which is crucial when selecting the number of components for dimensionality reduction. Local i.d. is also useful as an auxiliary variable to characterize data geometry. Notably, the similarity kernels employed in TopoMetry are related to the Farahmand–Szepesvári–Audibert (FSA) estimator, leveraging ratios of distances to the $k/2$- and k-th nearest neighbors as a proxy for local sampling density.

Maximum Likelihood Estimator (MLE). Consider i.i.d. observations $X_i = g(Y_i)$

that embed lower-dimensional samples Y_i drawn from an unknown density f via a continuous, smooth map g. For a point x, assume $f(x)$ is approximately constant within a small sphere $S_x(R)$ of radius R, so observations in $S_x(R)$ form a homogeneous Poisson process. Replacing the radius by the k-nearest neighbors yields the local MLE for intrinsic dimension around x:

$$\hat{m}_k(x) \;=\; \left[\frac{1}{k-1}\sum_{j=1}^{k-1}\log\frac{T_k(x)}{T_j(x)}\right]^{-1},$$

where $T_j(x)$ is the distance from x to its j-th nearest neighbor. A global estimate is obtained by averaging the local estimates over all x.

Manifold-adaptive dimensionality estimation (FSA). The FSA method uses two nested neighborhoods around x (at k and $k/2$ neighbors) to estimate local i.d.:

$$\delta_k(x) \;=\; \frac{\log 2}{\log\big(R_k(x)/R_{k/2}(x)\big)},$$

where $R_j(x)$ denotes the distance from x to its j-th nearest neighbor. The ratio $R_k/R_{k/2}$ coincides with the scaling factor used in TopoMetry's bandwidth-adaptive kernel. The global i.d. is reported as the median of the local $\delta_k(x)$ values.

Laplacian-type and diffusion operators. Let $W \in \mathbb{R}^{N\times N}_{\geq 0}$ be a symmetric affinity (similarity) matrix built on the samples, with zero diagonal. Denote the (weighted) degree of node i by $d_i = \sum_j W_{ij}$ and $D = \mathrm{diag}(d_1,\ldots,d_N)$. Graph Laplacians translate this similarity structure into linear operators whose spectra summarize geometry and connectivity; diffusion operators turn W into a Markov process on cells.

Three standard graph Laplacians:

$$\text{(i) Unnormalized:}\quad L \;=\; D - W.$$

$$\text{(ii) Symmetric normalized:}\quad L_{\mathrm{sym}} \;=\; D^{-1/2}LD^{-1/2} \;=\; I - D^{-1/2}WD^{-1/2}.$$

$$\text{(iii) Random-walk (row-normalized):}\quad L_{\mathrm{rw}} \;=\; D^{-1}L \;=\; I - D^{-1}W.$$

All three are positive semidefinite on a connected graph, with a trivial eigenpair $(0,\mathbf{1})$. In practice, L_{sym} is convenient for symmetric eigensolvers and is less sensitive to degree heterogeneity; L_{rw} is directly linked to random walks and diffusion. Spectral embeddings such as Laplacian Eigenmaps[6] use the eigenvectors associated with the

smallest nonzero eigenvalues of one of these Laplacians (often dropping the constant eigenvector) as intrinsic coordinates.

Diffusion operator and diffusion time. Row-normalizing W yields the (row-stochastic) diffusion operator

$$P = D^{-1}W,$$

which defines a Markov chain on cells: $(P^t)_{ij}$ is the probability of transitioning from cell i to cell j in t steps. The leading eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and right eigenvectors $\psi_1, \psi_2, \ldots$ of P provide *diffusion coordinates* at time t,

$$\Psi_t(i) = \big(\lambda_1^t \psi_1(i), \lambda_2^t \psi_2(i), \ldots\big),$$

which emphasize multi-step connectivity and progressively suppress noise[7]. When a single t is not desired, one can use alternative scalings (e.g., "diffusion potential" mappings) that highlight long-range structure without fixing t.

Anisotropic (α) normalization to reduce sampling bias. Nonuniform sampling density can bias graph operators. A standard correction (Diffusion Maps) reweights W by powers of the degree before row-normalization:

$$W^{(\alpha)} = D^{-\alpha} W D^{-\alpha}, \qquad D_\alpha = \operatorname{diag}\big(W^{(\alpha)}\mathbf{1}\big), \qquad P^{(\alpha)} = D_\alpha^{-1} W^{(\alpha)}.$$

Here $0 \leq \alpha \leq 1$ tunes the level of debiasing: $\alpha = 0$ leaves W unchanged; $\alpha = 1$ largely removes effects of sampling density. A related "semi-anisotropic" variant applies the degree correction in the normalization step while retaining the original edge weights W; this can stabilize spectra while preserving the original notion of similarity.

Symmetric diffusion operator for stable spectra. While $P^{(\alpha)}$ is generally asymmetric, it is similar to a symmetric matrix with the *same* eigenvalues:

$$P_{\text{sym}}^{(\alpha)} = D_\alpha^{-1/2} W^{(\alpha)} D_\alpha^{-1/2} = D_\alpha^{1/2} P^{(\alpha)} D_\alpha^{-1/2}.$$

This Hermitian form is numerically well behaved; its eigenvectors can be mapped back to those of $P^{(\alpha)}$ by left-multiplication with $D_\alpha^{-1/2}$. Diffusion-map embeddings are then obtained from the leading eigenpairs of $P^{(\alpha)}$ (or equivalently $P_{\text{sym}}^{(\alpha)}$), optionally raised to diffusion time t.

Connections to manifold geometry. Under standard assumptions (smooth manifold,

appropriate kernels and bandwidths), graph Laplacians converge to the Laplace–Beltrami operator Δ_g on the data manifold, and P^t approximates heat diffusion $e^{t\Delta_g}$ [6,7]. This link justifies using Laplacian- and diffusion-based spectra as intrinsic, geometry-aware coordinates for downstream single-cell analysis.

Eigendecomposition, spectral scaffolds, and multiscaling. Let $P \in \mathbb{R}^{N\times N}$ be the diffusion operator derived from the affinity graph. We start from the eigendecomposition

$$P\,\psi_\ell \;=\; \lambda_\ell\,\psi_\ell, \qquad \ell = 1,2,\ldots,$$

with eigenvalues ordered by magnitude $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. For numerical stability it is often convenient to work with a symmetric similarity $P_{\text{sym}} = D^{1/2}PD^{-1/2}$, which is similar to P and therefore has the same spectrum; its eigenvectors are orthonormal in the Euclidean inner product and can be mapped back to right–eigenvectors of P by left multiplication with $D^{-1/2}$.

Spectral scaffold. We define the *spectral scaffold* as

$$\Psi \;=\; [\,\psi_1,\ \psi_2,\ \ldots,\ \psi_r\,] \in \mathbb{R}^{N\times r},$$

whose columns are mutually orthogonal (orthonormal after normalization) and ordered by decreasing $|\lambda_\ell|$. The leading column ψ_1 is the trivial stationary mode (constant on each connected component) and is dropped in embeddings. The scaffold size r is chosen adaptively using intrinsic dimensionality estimates and eigengap heuristics, so it expands or contracts with dataset complexity rather than being fixed a priori.

Diffusion coordinates at timescale t. Geometry at a specific diffusion timescale $t \in \mathbb{N}$ is obtained by reweighting scaffold columns by λ_ℓ^t. For cell i,

$$\Phi_t(i) \;=\; \big(\lambda_1^t\,\psi_1(i),\ \lambda_2^t\,\psi_2(i),\ \ldots,\ \lambda_r^t\,\psi_r(i)\big) \;=\; \Psi(i,\cdot)\,\mathrm{diag}\big(\lambda_1^t,\ldots,\lambda_r^t\big),$$

which progressively suppresses high-frequency (noise-like) components as t increases, emphasizing long-range organization.

Multiscale spectral scaffold. Besides using single t, TopoMetry builds a multiscale representation by analytically aggregating all diffusion times. Using the geometric-series identity $\sum_{t=1}^{\infty}\lambda_\ell^t = \lambda_\ell/(1-\lambda_\ell)$ (valid for $\lambda_\ell \in [0,1)$ and after removing the

trivial mode $\lambda_1 = 1$), we form

$$\Phi_{\mathrm{ms}}(i) \;=\; \left(\frac{\lambda_2}{1-\lambda_2}\,\psi_2(i),\; \frac{\lambda_3}{1-\lambda_3}\,\psi_3(i),\; \ldots,\; \frac{\lambda_r}{1-\lambda_r}\,\psi_r(i) \right),$$

or, in matrix form,

$$\Phi_{\mathrm{ms}} \;=\; \Psi\,\mathrm{diag}\Big(\tfrac{\lambda_2}{1-\lambda_2}, \ldots, \tfrac{\lambda_r}{1-\lambda_r}\Big),$$

optionally restricting to eigenvalues $\lambda_\ell > 0$ to avoid numerical instabilities. This *multiscale spectral scaffold* is equivalent to concatenating diffusion coordinates across all times with exponentially decaying weights, thereby blending fine-grained neighborhoods (small t) and long-range connectivity (large t) in a single, compact set of coordinates.

Normalization and symmetry handling. When a symmetric similarity is used for eigensolving, the obtained eigenvectors are mapped to the right–eigenvectors of P via the similarity transform described above and then ℓ_2-normalized, ensuring that columns of Ψ remain mutually orthogonal. With r selected adaptively, the resulting spectral scaffold (single–time Φ_t or multiscale Φ_{ms}) provides geometry-aware coordinates that adjust to the intrinsic complexity of the single-cell atlas.

Refined graph. Given spectral coordinates $Z \in \mathbb{R}^{N\times r}$ (either a single-time Φ_t or the multiscale spectral scaffold Φ_{ms}), we rebuild an affinity graph $\widetilde{W}$ *in scaffold space* to sharpen geometry and suppress residual noise.

Construction in Z-space. We measure distances with the Euclidean metric $d_{ij} = \|z_i - z_j\|_2$. Using the same adaptive-bandwidth principle as above, each node i receives a local scale σ_i given by the distance to its median k-nearest neighbor in Z-space. Affinities are then computed by

$$\widetilde{W}_{ij} \;=\; \exp\Big(-\big(\tfrac{d_{ij}}{\sigma_i}\big)^{p_i}\Big), \qquad p_i \in \{2,\, 2\,\eta_i\},$$

where $p_i = 2$ yields a bandwidth-adaptive Gaussian and $p_i = 2\,\eta_i$ is an adaptively decaying variant that slows decay in sparse regions and sharpens it in dense ones. The density proxy ω_i and decay factor $\eta_i = 2^{(k-\omega_i)/\omega_i}$ are defined as in the kernel section. Finally, we symmetrize

$$\widetilde{W} \;\leftarrow\; \tfrac{1}{2}\big(\widetilde{W} + \widetilde{W}^\top\big),$$

ensuring a real symmetric affinity suitable for spectral operators (with standard numerical guards).

Optional neighborhood expansion. In severely under-sampled areas, we allow a mild, density-aware increase of the neighbor set to $k' \geq k$ (guided by ω_i), recompute σ_i on $\mathrm{nbrs}_{k'}(i)$, and apply the same formulae.

From $\widetilde{W}$ we compute the degree matrix $\widetilde{D} = \mathrm{diag}(\widetilde{W}\mathbf{1})$ and derive Laplacian-type operators whose normalization yields refined diffusion operators. By default, TopoMetry applies anisotropic reweighting to these operators, producing a geometry-adapted transition matrix that serves as the backbone for downstream analyses. This refined diffusion operator is then used for tasks such as clustering, trajectory inference, pseudotime analysis, and visualization with graph layout algorithms, ensuring that all inferences are grounded in a refined representation of the original manifold of cell identities.

Layout optimization. For visualization, TopoMetry applies graph-layout optimization algorithms to the refined graphs. By default, we provide *TopoMAP* (an efficient UMAP-style layout applied to a precomputed manifold graph) and *TopoPaCMAP*; both are optimized on the refined diffusion operator and its neighborhood structure. Layout trajectories can be inspected to assess convergence and stability.

Uniform Manifold Approximation and Projection (UMAP) / Manifold Approximation and Projection (MAP). UMAP-style objectives interpret a weighted graph as the 1-skeleton of a *fuzzy simplicial set* (FSS): an edge set with graded memberships $\mu_{ij} \in [0,1]$ that encode how strongly i and j belong to one another's neighborhood. Given a low-dimensional embedding $Y = \{y_i\}$, a second fuzzy edge set with memberships $\nu_{ij}(Y) \in [0,1]$ is induced by a fixed, monotonically decreasing function of the low-dimensional distance $\|y_i - y_j\|$. The layout is obtained by minimizing the cross-entropy between the two fuzzy edge sets:

$$\mathscr{L}_{\mathrm{UMAP}}(Y) \;=\; \sum_{i<j} \Big[\mu_{ij} \log \frac{\mu_{ij}}{\nu_{ij}(Y)} \;+\; \big(1-\mu_{ij}\big) \log \frac{1-\mu_{ij}}{1-\nu_{ij}(Y)} \Big].$$

In TopoMetry, *MAP* denotes applying the same objective to an *arbitrary, precomputed* manifold graph (here, the refined graph in scaffold space); we refer to this implementation as *TopoMAP*.

Pairwise Controlled Manifold Approximation and Projection (PaCMAP). PaCMAP

optimizes a robust pairwise objective that balances three types of pair relationships: (i) *near pairs* (local neighbors), (ii) *mid-near pairs* (pairs connecting adjacent neighborhoods), and (iii) *further pairs* (distant non-neighbors). Let local scales be defined from the average distance to the 4th–6th nearest neighbors; write the high-dimensional, scale-adjusted distance as

$$d_{ij}^2 \;=\; \frac{\|x_i - x_j\|_2^2}{\sigma_{ij}},$$

and let $d'_{ab} = \|y_a - y_b\|_2^2 + 1$ denote the corresponding low-dimensional distance surrogate. The PaCMAP loss is

$$\mathscr{L}_{\text{PaCMAP}} = \Big(w_{\text{nb}} \sum_{(i,j)\in\mathscr{N}} \frac{d'_{ij}}{10 + d'_{ij}}\Big) + \Big(w_{\text{mn}} \sum_{(i,k)\in\mathscr{M}} \frac{d'_{ik}}{10^4 + d'_{ik}}\Big) + \Big(w_{\text{fp}} \sum_{(i,\ell)\in\mathscr{F}} \frac{1}{10 + d'_{i\ell}}\Big),$$

where $\mathscr{N}, \mathscr{M}, \mathscr{F}$ are the sets of near, mid-near, and further pairs selected per the original sampling rules, and $(w_{\text{nb}}, w_{\text{mn}}, w_{\text{fp}})$ are stage-specific weights (e.g., Stage 1: $(2, 1000, 1)$, Stage 2: $(3, 3, 1)$, Stage 3: $(1, 0, 1)$). In TopoMetry, PaCMAP is initialized from the spectral scaffold and uses neighbor information from the refined manifold graph, yielding stable, high-fidelity layouts that respect both local neighborhoods and broader organization.

Riemann diagnostics and visualizations. Let $Y \in \mathbb{R}^{N\times m}$ be an embedding (typically $m = 2$ for visualization) with coordinate functions $y^{(a)} : \{1, \ldots, N\} \to \mathbb{R}$ given by the a-th column of Y, and let $L \in \mathbb{R}^{N\times N}$ be a (symmetrized) graph Laplacian built on the refined affinity. We center the embedding, $Y \leftarrow Y - \mathbf{1}\overline{Y}$, and write Ly for the discrete application of L to a vector y.

Discrete dual metric and embedding-space metric. For each sample i and coordinate indices $a, b \in \{1, \ldots, m\}$, define the discrete *dual* metric (a covariant tensor) by

$$H_{ab}(i) \;=\; \tfrac{1}{2}\Big[(L\,(y^{(a)} y^{(b)}))_i \;-\; y_i^{(b)}\,(Ly^{(a)})_i \;-\; y_i^{(a)}\,(Ly^{(b)})_i\Big].$$

Stacking over a, b yields $H(i) \in \mathbb{R}^{m\times m}$. We obtain the embedding-space Riemannian metric $G(i)$ as the (regularized) pseudoinverse of $H(i)$:

$$H(i) \;=\; U_i S_i V_i^\top, \qquad G(i) \;=\; V_i S_i^\dagger U_i^\top, \qquad S_i^\dagger = \operatorname{diag}\big(1/(s_{i\ell} + \varepsilon)\big),$$

with a small $\varepsilon > 0$ to stabilize near-zero singular values. Finally, we project $G(i)$ to

the symmetric positive definite cone by eigenvalue clipping and normalize its overall scale (e.g. unit trace) to make ellipse sizes comparable across points.

Local distortion indicatrix. Let $G(i) = Q_i \Lambda_i Q_i^\top$ with $\Lambda_i = \mathrm{diag}(\lambda_{i,1} \geq \cdots \geq \lambda_{i,m} > 0)$. The *indicatrix* at i is the ellipse (for $m = 2$)

$$\mathscr{E}(i) \;=\; \{\, u \in \mathbb{R}^2 :\; u^\top G(i)\, u = 1 \,\},$$

whose principal axes have directions given by the columns of Q_i and semi-axis lengths

$$a_i = \sqrt{\lambda_{i,1}}, \qquad b_i = \sqrt{\lambda_{i,2}}.$$

Thus $a_i/b_i = \sqrt{\lambda_{i,1}/\lambda_{i,2}}$ quantifies local anisotropy (directional stretching), while $\det G(i) = \lambda_{i,1}\lambda_{i,2}$ quantifies local area change. In plots we scale (a_i, b_i) by a global factor and ensure each ellipse fits within the axes (conservative circumscribed-circle bound) to avoid clipping.

Contraction/expansion scalar field. We color points (or ellipses) by a centered log-determinant of the metric,

$$\mathrm{Def}(i) \;=\; \log\det G(i) \;-\; \mathrm{center}\big(\{\log\det G(j)\}_{j=1}^{N}\big),$$

with the center chosen as the median (robust) or mean, so that $\mathrm{Def} > 0$ indicates local *expansion* and $\mathrm{Def} < 0$ indicates *contraction*. Optionally, we smooth this scalar over the graph by t steps of diffusion, $\mathrm{Def} \leftarrow P^t \mathrm{Def}$, where $P = D^{-1}W$ is the random-walk operator derived from $L = D - W$; we then re-center and use robust percentile clipping with symmetric limits for color normalization.

Visualization modes. (i) *Localized indicatrices:* draw ellipses $\mathscr{E}(i)$ for a uniform subset of points, optionally modulating their size by a normalized anisotropy score $\log(\lambda_{i,1}/\lambda_{i,2})$. The underlying scatter may be colored by $\mathrm{Def}(i)$ or other annotations. (ii) *Grid-averaged field:* build a regular grid over the embedding, and at each grid site g average nearby metrics, $G_{\mathrm{avg}}(g) = \frac{1}{k}\sum_{i \in \mathscr{N}_k(g)} G(i)$; draw one ellipse from $G_{\mathrm{avg}}(g)$ and color it by the grid-averaged Def. A thinning step enforces a minimum separation between grid sites to limit overlap.

Geometric interpretation. If $\phi : \mathscr{M} \to \mathbb{R}^m$ denotes the embedding map, then $G(i)$ estimates the pullback metric $(\mathrm{d}\phi_{x_i})^\top \mathrm{d}\phi_{x_i}$ at x_i (up to the discrete Laplacian normalization), so that $\log\det G(i) \approx \log\det\big((\mathrm{d}\phi_{x_i})^\top \mathrm{d}\phi_{x_i}\big)$ reports local area change and the eigenstructure of $G(i)$ reports directional distortion. The plotted indicatrices

therefore act as *distortion rulers*, revealing where a 2-D layout expands, contracts, or shears the intrinsic geometry.

Data imputation. We perform diffusion-based imputation on the cell–cell graph. Let $X \in \mathbb{R}^{N\times G}$ denote the cell-by-gene matrix to be imputed, and let $P \in \mathbb{R}^{N\times N}$ be the Markov (row-stochastic) diffusion operator built on the refined manifold graph. For a diffusion time $t \in \mathbb{N}$, the imputed matrix is the graph diffusion

$$X^{(t)} \;=\; P^t X,$$

i.e. each gene is smoothed across the graph while preserving its marginal scale.

Automatic selection of the diffusion time. To avoid oversmoothing, we select t from a candidate grid $\mathscr{T} = \{t_1, \ldots, t_m\}$ by contrasting the gene–gene correlation structure of $X^{(t)}$ against a null in which cross-cell associations have been destroyed but univariate marginals are preserved. Concretely, we first identify a small panel of highly variable genes to score. Let $X_{\text{top}} \in \mathbb{R}^{N\times g}$ be the submatrix of these genes. For each $t \in \mathscr{T}$:

$$X_{\text{top}}^{(t)} \;=\; P^t X_{\text{top}}, \qquad C^{(t)} \;=\; \operatorname{corr}\big(X_{\text{top}}^{(t)}\big) \in \mathbb{R}^{g\times g},$$

and we define a scalar score as the mean absolute off-diagonal correlation,

$$S_{\text{obs}}(t) \;=\; \frac{1}{g(g-1)} \sum_{a\neq b} \big|C_{ab}^{(t)}\big|.$$

Null model and empirical test. For each t, we generate a null distribution by independently permuting the rows of every column of X_{top} (gene-wise shuffling across cells), diffusing with the same t, and recomputing the score:

$$\text{for } b = 1, \ldots, K: \quad \widetilde{X}_{\text{top}}^{(b)} \leftarrow \text{row-permute}(X_{\text{top}}); \quad \widetilde{X}_{\text{top}}^{(b,t)} = P^t\, \widetilde{X}_{\text{top}}^{(b)}; \quad S_{\text{null}}^{(b)}(t) \;=\; \frac{1}{g(g-1)} \sum_{a\neq b}$$

We then compute an empirical p-value and a z-score:

$$p_{\text{emp}}(t) \;=\; \frac{1 + \#\{\, b : S_{\text{null}}^{(b)}(t) \geq S_{\text{obs}}(t) \,\}}{1+K}, \qquad z(t) \;=\; \frac{S_{\text{obs}}(t) - \mu_{\text{null}}(t)}{\sigma_{\text{null}}(t) + 10^{-9}},$$

with $\mu_{\text{null}}, \sigma_{\text{null}}$ the mean and standard deviation of $\{S_{\text{null}}^{(b)}(t)\}_{b=1}^{K}$. The selected time

is

$$t^\star \;=\; \arg\min_{t\in\mathscr{T}} p_{\text{emp}}(t) \quad \text{(tie-break by largest } z(t)\text{)}.$$

After identifying an optimal diffusion time t, the optimal imputation result is:

$$X_{\text{imp}} \;=\; P^{t^\star} X \quad .$$

This procedure selects the weakest amount of smoothing that yields gene–gene correlation structure significantly exceeding what diffusion would induce under a null with broken cross-cell associations. In doing so, it explicitly guards against the common pitfall of oversmoothing, where excessive diffusion can erase genuine biological heterogeneity and introduce spurious correlations that mimic structure but are in fact artifacts. By benchmarking the observed correlation profiles against an empirical null, the method ensures that only structure reproducibly supported by the data is retained. This design minimizes the risk of false positives: correlations that arise purely from graph connectivity rather than underlying transcriptional programs are filtered out, since they also appear in the null-distribution and therefore fail the statistical test. At the same time, genuine patterns of co-expression that reflect biological coordination between genes are amplified, as they persist in the observed data but not under randomized permutations. The approach thus balances denoising with fidelity to the original data, producing imputations that are statistically principled, biologically meaningful, and less likely to bias downstream analyses such as clustering, trajectory inference, or differential expression.

Signal filtering. TopoMetry provides a graph-based low-pass filter to denoise sample-level signals while respecting manifold structure. Let $s \in \mathbb{R}^N$ be a per-cell signal and $P \in \mathbb{R}^{N\times N}$ the row-stochastic diffusion operator on the refined graph (default: built in the multiscale spectral scaffold space). The filtered signal after t diffusion steps is

$$s^{(t)} \;=\; P^t\, s,$$

which corresponds to heat-kernel smoothing on the graph: increasing t attenuates high-frequency (noisy, rapidly varying) components and preserves low-frequency structure aligned with the topology.

Signal construction. Users supply a column in the sample annotations. If the

column is categorical, we form a binary indicator

$$s_i \;=\; \mathbf{1}\{\,\text{cell } i \text{ belongs to the category of interest}\,\},$$

else a numeric column is used directly (non-finite entries set to 0). For stress tests one may optionally add pre-filter Gaussian noise, $s \leftarrow s + \sigma\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$ and $\sigma \geq 0$.

Choice of operator and time. Filtering can use the Markov operator derived in any space ($\mathtt{msZ}, \mathtt{Z}, \mathtt{X}$); the default $\mathtt{msZ}$ leverages the refined multiscale geometry. The parameter $t \in \mathbb{N}$ controls smoothness: $t = 0$ returns the original signal, small t mildly denoises within neighborhoods, and large t approaches the stationary distribution (oversmoothing). In practice, modest values (e.g. $t \approx 4-8$) preserve biological contrasts while reducing noise.

Normalization (optional). After filtering, an optional rescaling can be applied. If the signal is intended as a probability or indicator, one can keep the native range. Otherwise, a min–max normalization maps $s^{(t)}$ to $[0,1]$:

$$\widehat{s}^{(t)} \;=\; \frac{s^{(t)} - \min_i s_i^{(t)}}{\max_i s_i^{(t)} - \min_i s_i^{(t)}},$$

guarded for degenerate ranges. This facilitates downstream visualization and thresholding.

For reproducibility, both the raw vector and its filtered counterpart are recorded in the sample metadata, including the diffusion time t and the operator choice. The procedure is computationally light (matrix–vector products, compatible with sparse P) and can be applied to categorical labels, risk scores, or any continuous per-cell quantity to obtain geometry-aware smoothed estimates.

4.3 Quantification of geometry preservation

To systematically evaluate how well different representations preserve the intrinsic geometry of single-cell data, we developed a set of four complementary metrics. Each metric operates directly on the diffusion operators associated with the original data space and the candidate embedding, thereby enabling a scale-aware and operator-native assessment of geometry preservation.

Row-neighborhood F1 score (PF1). Let P_x and P_y denote two row-stochastic diffusion operators to be compared. For each cell i, let $N_x(i)$ and $N_y(i)$ denote the sets of the top-k neighbors defined by the largest transition probabilities in P_x and P_y, respectively. The per-row F1 score is given by

$$\mathrm{F1}(i) \;=\; \frac{2\,|N_x(i) \cap N_y(i)|}{|N_x(i)| + |N_y(i)|},$$

and the PF1 score is obtained as the average over all cells:

$$\mathrm{PF1} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{F1}(i).$$

This score captures the preservation of local neighborhood identities, independently of edge weights.

Row-wise Jensen–Shannon similarity (PJS). To compare the transition probability distributions themselves, we use the Jensen–Shannon (JS) divergence. For each row i, let p_i and q_i denote the probability vectors corresponding to the nonzero transitions of P_x and P_y. The JS divergence between p_i and q_i is

$$\mathrm{JS}(p_i, q_i) = \tfrac{1}{2}\,\mathrm{KL}\Big(p_i \,\Big\|\, m_i\Big) + \tfrac{1}{2}\,\mathrm{KL}\Big(q_i \,\Big\|\, m_i\Big)\,,$$

where $m_i = \frac{1}{2}(p_i + q_i)$ and KL denotes the Kullback–Leibler divergence. We then define the similarity as

$$\mathrm{PJS} = 1 - \frac{1}{n}\sum_{i=1}^{n} \mathrm{JS}(p_i, q_i).$$

Unlike PF1, this metric is sensitive to the weights of transitions, thus evaluating the fidelity of local transition probabilities.

Spectral Procrustes alignment (SP). Diffusion operators admit an eigendecomposition $P\psi_\ell = \lambda_\ell \psi_\ell$, where $\{\lambda_\ell, \psi_\ell\}$ are eigenpairs ordered by $|\lambda_\ell|$. For a given timescale t, the diffusion map is defined as

$$\Phi_t(i) = \big(\lambda_1^t \psi_1(i), \ldots, \lambda_r^t \psi_r(i)\big)\,.$$

Given the diffusion coordinates Φ_t^x and Φ_t^y from P_x and P_y, we align them via orthogonal Procrustes:

$$R^* = \arg\min_{R^\top R = I} \|\Phi_t^y R - \Phi_t^x\|_F^2.$$

The alignment quality is quantified by the coefficient of determination,

$$R^2 = 1 - \frac{\|\Phi_t^x - \Phi_t^y R^*\|_F^2}{\|\Phi_t^x\|_F^2}.$$

The SP score is averaged across multiple diffusion times t, capturing agreement of the meso- and global-scale geometry of the manifold.

Rationale for metric selection. These three metrics together capture complementary aspects of geometry preservation across scales: PF1 and PJS quantify fidelity of local neighborhoods, and SP evaluates alignment of diffusion eigencoordinates at mesoscopic scales. Compared to previously used metrics, they provide a more faithful evaluation of single-cell representations. Global scores based on PCA depend strongly on graph-layout initialization rather than underlying geometry. Geodesic correlation, though sensitive to structure, overweights large-scale distances and becomes biased in the presence of multiple disjoint submanifolds, as typical in single-cell data. Clustering- or label-based scores are unreliable in real-world settings where ground-truth identities are unknown. Finally, Riemannian diagnostics such as distortion heatmaps and indicatrices are restricted to 2-D embeddings, whereas single-cell analyses require latent spaces with tens or hundreds of dimensions. Our operator-native metrics therefore offer a principled, scalable, and interpretable quantification of geometry preservation across learned representations.

4.4 Standard analysis workflow

Single-cell analysis workflows have converged on a widely adopted pipeline that couples dimensionality reduction by Principal Component Analysis (PCA) with graph construction and visualization using Uniform Manifold Approximation and Projection (UMAP). This approach underpins the majority of published studies and is implemented as the default in most software ecosystems for single-cell genomics.

Preprocessing. Raw count matrices are typically subjected to two standard preprocessing steps. First, counts are normalized for library size (total molecules per cell) and log-transformed to stabilize variance. Second, a subset of highly variable genes (HVGs) is selected, usually between 1,000 and 5,000, under the assumption that these capture the majority of biologically relevant variation while reducing technical noise. The resulting HVG matrix is then standardized (Z-score normalized) so that each gene has mean zero and unit variance, preventing highly expressed genes from dominating subsequent analyses.

Principal Component Analysis. The standardized HVG matrix is projected into a lower-dimensional space using PCA. The number of retained components is typically chosen heuristically (e.g. 30–100) based on an *ad hoc* "elbow point", rather than based on intrinsic dimensionality. The Euclidean distances between cells in this PCA space are then used to build a k-nearest-neighbors (kNN) graph, which serves as the foundation for clustering and visualization.

UMAP graph layout. The kNN graph is converted into an affinity graph using fuzzy simplicial sets, which UMAP optimizes into a two-dimensional embedding. The resulting map aims to preserve local neighborhoods while maintaining a global structure that is visually interpretable. This 2-D representation is widely used for cluster annotation, lineage inference, and exploratory analysis, despite being optimized for visualization rather than strict geometric fidelity.

Clustering and identification of marker genes. Clustering was performed with the Leiden community detection algorithm using default parameters via `scanpy.tl.leiden`. To ensure fair comparisons, we applied the same clustering algorithm with identical `resolution` parameters across all evaluated workflows, including TopoMetry. Marker genes were identified with `scanpy.tl.rank_genes_groups` after constructing dendrograms on the representation associated with the clustering results, using the logistic regression method (`method='logreg'`), which has been shown to achieve optimal results in single-cell data[36].

Summary. Together, these steps—log and size normalization, HVG selection, Z-score standardization, PCA projection, kNN graph construction, and UMAP embed-

ding—constitute the de facto standard in single-cell analysis. While computationally efficient and easy to implement, the workflow is based on strong assumptions: that global variance identifies biologically meaningful structure, that a fixed number of PCs captures the latent geometry, and that UMAP embeddings accurately reflect manifold relationships. These assumptions have rarely been tested systematically, motivating the development of alternative frameworks such as TOPOMETRY.

RNA velocity analysis. RNA velocity analysis was performed using the `scVelo` toolkit[37], with all steps executed using default parameters. The exception was the calculation of neighborhood graphs and moments, which by default rely on the PCA latent space, but were instead computed using the TopoMetry multiscale spectral scaffold. For TopoMetry analysis, spliced and unspliced counts were aggregated, and default preprocessing was applied to provide a standardized matrix of gene expression as input.

T cell receptor analysis The ECCITE-TCR dataset was obtained with pre-computed TCR analysis and metadata. For the TICA dataset, T cell receptor (TCR) clonotype analysis was performed using the `scirpy` toolkit for immune receptor analysis in Python[30]. TCR similarities were computed using the alignment of amino acid sequences. Default parameters were used for all remaining steps.

Variational inference. Variational autoencoder analyses were performed using the `scvi-tools` framework[17,18]. Models were trained on GPU using default parameters, including preprocessing steps handled internally by the package. Latent spaces were computed and used to construct neighborhood graphs for Leiden clustering and UMAP visualizations.

4.5 Public datasets

We assembled a corpus of publicly available single-cell RNA-seq collections from the CellxGene Census (release `2023-07-25`). For each target `collection_id`, we queried the Census metadata table (`census_info/datasets`) to obtain the corresponding `dataset_ids` and then retrieved primary RNA measurements (`is_primary_data = True`) as AnnData objects separately for *Homo sapiens* and *Mus musculus*. Observational annotations included tissue-level fields (e.g., `tissue`, `tissue_general`,

assay, cell_type). Individual datasets within a collection were concatenated (outer join) after loading. To ensure basic quality, we removed cells with fewer than 100 detected genes and genes expressed in fewer than 5 cells prior to downstream analysis. For computational tractability, collections with more than 10^5 cells were excluded from this benchmarking pass, as were a small number of collections exhibiting upstream data issues (e.g., malformed or inconsistent matrices). For each processed collection we stored a compact result bundle containing: collection identifiers and names, cell and gene counts, a set of two-dimensional embeddings, and a results blob with evaluation tables and diagnostics. Additional datasets analysed in this manuscript are also publicly available: The murine pancreas development dataset was obtained through the scVelo toolkit for RNA velocity estimation and visualization. The PBMC68k dataset was freely downloaded from 10X Genomics. Additional PBMC datasets were obtained from the accession codes phs002048.v1.p1 (LES), GSE154386 (Dengue) and GSE138266 (MS). The paired RNA and TCR datasets were made publicly available by their authors and were obtained from Zenodo and from the Human Cell Atlas.

4.6 Performance benchmark

All methods were applied to the same preprocessed matrices (highly variable genes, library-size normalization, $\log(1+x)$, scaling), as described in the preceding section. We evaluated geometry-preservation performance across representations using a unified routine that computes neighborhood- and diffusion-based scores.

Concretely, for each collection:

- We computed a standard PCA model (up to 200 components or the maximal feasible given n_{cells} and n_{genes}) and retained:
  1. the PCA space (X_PCA),
  2. UMAP on PCA neighborhoods (X_PCA_UMAP),
  3. UMAP computed directly on the scaled expression space with cosine distance (X_UMAP_on_X).
- We trained an SCVI model on raw counts (default scvi-tools settings), extracted the latent representation (X_scVI), and computed UMAP on its neighborhood graph (X_UMAP_on_scVI).

- We ran TopoMetry via `tp.sc.fit_adata`, which builds a first kernel on expression, performs an eigendecomposition to obtain a spectral scaffold, and refines the kernel on that scaffold before producing low-dimensional projections (e.g., TopoMAP/TopoPaCMAP). The fitted `TopOGraph` exposes Markov operators P_X, P_Z, and $P_{\mathrm{ms}Z}$ used in evaluation.

We assessed geometry preservation for *all* representations present in `adata.obsm` using the same parameters: Euclidean distance, local neighborhood size k=30, HNSW-based approximate nearest-neighbors search, and diffusion times $t \in \{1,4,8\}$ with spectral rank r=64. The evaluation routine reports a table of metrics, including row-neighborhood k-NN overlap (PF1), row-wise Jensen–Shannon similarity (PJS), and Spectral Procrustes alignment (SP). PCA diagnostics recorded per collection included the full variance-explained spectrum and its cumulative sum. Default hyperparameters were used for all evaluated workflows to enable a fair performance comparison across datasets.

Data and Code Availability

`topometry` is freely accessible as a Python library under the MIT license. The source code is available at https://github.com/davisidarta/topometry, and can be installed through the Python Package Index (PyPI) at https://pypi.org/project/topometry/. The library is extensively documented at https://topometry.readthedocs.io. Any additional information can be requested from the authors and will be made available upon request.

Acknowledgments

We thank Leland McInness, Dmitry Kobak, and Akshay Agrawal for valuable comments on the ideas presented in this manuscript. We also thank Bruno Loyola Barbosa for assistance in testing early implementations of `topometry`.

Funding

DS-O was supported by grant #2020/04074-2 from the São Paulo Research Foundation (FAPESP).

AID was supported by the BBSRC (no. BB/Y006488/1), the ERC Consolidator Award (no. ERC-2017 COG 771431), the Pfizer ASPIRE Obesity Award (#70591281), the National Institutes of Health (NIH) #5UM1DK10555410 – subaward 8795500003311, and the Next Iteration of the Type 2 Diabetes Knowledge Portal (no. 2UM1DK105554). LAV was supported by grant #2013/07607-8 from the São Paulo Research Foundation (FAPESP).

References

[1] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philos. Mag.*, 2:559–572, 1901.

[2] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2018.

[3] Etienne Becht et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*, 37:38–44, 2019.

[4] Fabian A. Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.*, 19:15, 2018.

[5] T. Chari and L. Pachter. The specious art of single-cell genomics, 2022.

[6] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[7] Ronald R. Coifman and Stephane Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21:5–30, 2006.

[8] Boaz Nadler, Stephane Lafon, Ronald R. Coifman, and Ioannis G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Appl. Comput. Harmon. Anal.*, 21:113–127, 2006.

[9] Thomas Berry and John Harlim. Variable bandwidth diffusion kernels. *Appl. Comput. Harmon. Anal.*, 40:68–96, 2016.

[10] Amir Massoud Farahmand, Csaba Szepesvári, and Jean-Yves Audibert. Manifold-adaptive dimension estimation. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 265–272, 2007. doi:10.1145/1273496.1273530.

[11] Zoltán Benkő et al. Manifold-adaptive dimension estimation revisited. *PeerJ Comput. Sci.*, 8, 2022.

[12] Dominique Perraul-Joncas and Marina Meilă. Non-linear dimensionality reduction: Riemannian metric estimation and the problem of geometric discovery, 2013.

[13] Sui Huang. The molecular and mathematical basis of Waddington's epigenetic landscape: A framework for post-Darwinian biology? *BioEssays*, 34:149–157, 2012.

[14] Yuxin Wang, Hong Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization, 2021.

[15] Ronald R. Coifman et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci. USA*, 102:7426–7431, 2005.

[16] Ian T. Jolliffe. *Principal Component Analysis*. Springer, New York, 1986.

[17] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 15:1053–1058, 2018.

[18] Adam Gayoso et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.*, 40:163–166, 2022.

[19] Yanyan Ma and Liang Zhu. A review on dimension reduction. *Int. Stat. Rev.*, 81:134–150, 2013.

[20] Farzana Anowar, Samira Sadaoui, and Bassant Selim. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Comput. Sci. Rev.*, 40, 2021.

[21] Aimée Bastidas-Ponce et al. Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, 146, 2019.

[22] Gioele La Manno et al. RNA velocity of single cells. *Nature*, 560:494–498, 2018.

[23] L. Arnes, Jason T. Hill, S. Gross, M. A. Magnuson, and Lori Sussel. Ghrelin expression in the mouse pancreas defines a unique multipotent progenitor population. *PLOS ONE*, 7, 2012.

[24] Junyue Cao et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566:496–502, 2019.

[25] Samuel L. Wolock, Romain Lopez, and Allon M. Klein. Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.*, 8(4):281–291.e9, 2019.

[26] Felix Drost, Yang An, Irene Bonafonte-Pardàs, Lisa M. Dratva, Rik G. H. Lindeboom, Muzlifah Haniffa, Sarah A. Teichmann, Fabian Theis, Mohammad Lotfollahi, and Benjamin Schubert. Multi-modal generative modeling for joint analysis of single-cell t cell receptor and gene expression data. *Nat. Commun.*, 15(1):5577, 2024.

[27] Bingjie Zhang, Rabi Upadhyay, Yuhan Hao, Marie I. Samanovic, Ramin S. Herati, John D. Blair, Jordan Axelrad, Mark J. Mulligan, Dan R. Littman, and Rahul Satija. Multimodal single-cell datasets characterize antigen-specific cd8+ t cells across sars-cov-2 vaccination and infection. *Nat. Immunol.*, 24(10): 1725–1734, 2023.

[28] Stefan A. Schattgen, Kate Guion, Jeremy Chase Crawford, Aisha Souquette, Alvaro Martinez Barrio, Michael J. T. Stubbington, Paul G. Thomas, and Philip Bradley. Integrating t cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (conga). *Nat. Biotechnol.*, 40(1):54–63, 2022.

[29] César Domínguez Conde et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376, 2022.

[30] Gregor Sturm et al. Scirpy: a Scanpy extension for analyzing single-cell t-cell receptor-sequencing data. *Bioinformatics*, 36:4817–4818, 2020.

[31] Kevin R. Moon et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.*, 37:1482–1492, 2019.

[32] Benjamin D. Hale, Yannik Severin, Fabienne Graebnitz, Dominique Stark, Daniel Guignard, Julien Mena, Yasmin Festl, Sohyon Lee, Jacob Hanimann, Nathan S. Zangger, Michelle Meier, David Goslings, Olga Lamprecht, Beat M. Frey, Annette Oxenius, and Berend Snijder. Cellular architecture shapes the naïve t cell response. *Science*, 384(6700):eadh8697, 2024.

[33] Francesca Tedeschi, Johan Quilbé, Lavinia Ioana Fechete, Sofie Jin Vistisen Christiansen, and Stig Uggerhøj Andersen. Rhd6la regulates root hair responses

to both symbionts and commensals. *bioRxiv*, page 2025.05.22.655471, 2025. doi:10.1101/2025.05.22.655471.

[34] André F. Duque, Sam Morin, Guy Wolf, and Kevin R. Moon. Geometry regularized autoencoders. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45:7381–7394, 2023.

[35] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 2021.

[36] Vasilis Ntranos, Lemon Yi, Páll Melsted, and Lior Pachter. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods*, 16:163–166, 2019.

[37] Volker Bergen, Marius Lange, Shila Peidli, Fabian A. Wolf, and Fabian J. Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.*, 38:1408–1414, 2020.

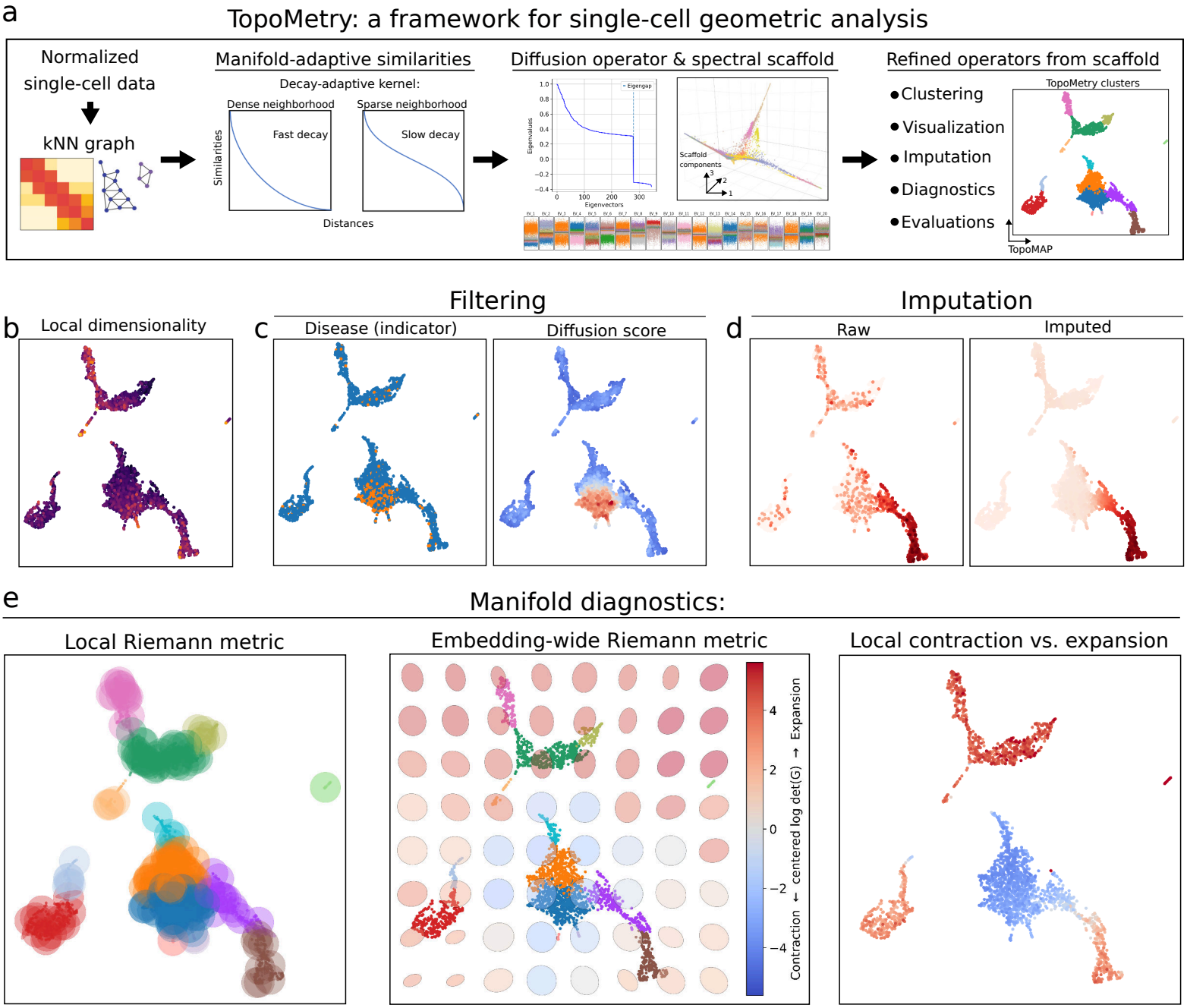

a
TopoMetry: a framework for single-cell geometric analysis
Normalized single-cell data
kNN graph
Manifold-adaptive similarities
Decay-adaptive kernel:
Dense neighborhood
Sparse neighborhood
Fast decay
Slow decay
Similarities
Distances
Diffusion operator & spectral scaffold
Eigenvalues
Eigenvectors
Eigengap
Scaffold components
Refined operators from scaffold
Clustering
Visualization
Imputation
Diagnostics
Evaluations
TopoMetry clusters
TopoMAP
b
Local dimensionality
Filtering
c
Disease (indicator)
Diffusion score
Imputation
d
Raw
Imputed
e
Manifold diagnostics:
Local Riemann metric
Embedding-wide Riemann metric
Local contraction vs. expansion
Contraction ← centered log det(G) → Expansion

Figure 1: A framework for single-cell geometric analysis.
(a) Schematic overview of the TopoMetry algorithm. From an input single-cell dataset (e.g., normalized and scaled scRNAseq data), TopoMetry builds a kNN graph, which is used to learn manifold-adaptive similarities with a decay-adaptive kernel suitable for constructing Laplacian-type and diffusion operators. After estimation of intrinsic dimensionality, these operators are decomposed into a spectral scaffold with up to hundreds of components that jointly explain all of the underlying geometry of the dataset. The spectral scaffolds are used to learn refined Laplacian-type and diffusion operators of the scaffolds themselves, encoding "the geometry of the geometry". The scaffolds and operators constitute key TopoMetry outputs and can be utilized for downstream tasks, such as clustering, visualization, imputation, evaluation, and diagnostics, in a geometry-aware manner. TopoMetry utilities include (b) estimation of local intrinsic dimensionality, (c) filtering of categorical signals, and (d) imputation and denoising. Crucially, TopoMetry introduces the visualization of manifold diagnostics (e) for single-cell data, in which distortions induced by 2-D embeddings can be identified and investigated from a local, global, and contraction/expansion perspective.

a

Corpus:
- 68 scRNAseq datasets
- mouse and human
- standard preprocessing

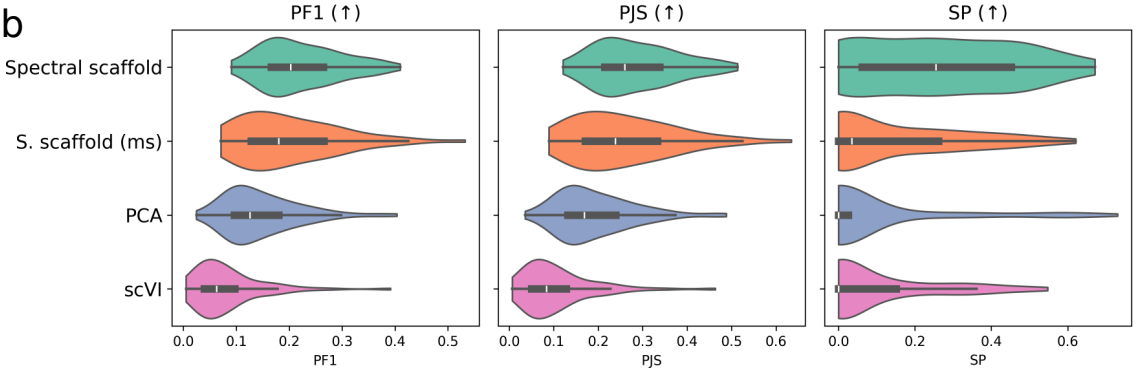
↓

Evaluated methods:
- current PCA-UMAP standard
- standalone UMAP
- scVI
- TopoMetry

↓

Evaluation metrics:
- PF1k: neighborhood consistency
- P-JS: transitions preservation
- SP: global agreement

b



c

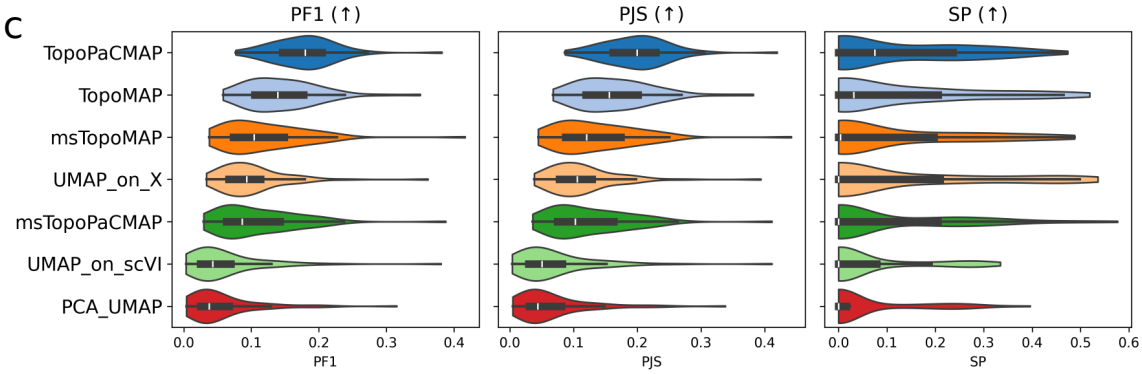


Figure 2: Geometry preservation benchmark.
(a) Schematic representation of the benchmark workflow, in which a corpus composed of 68 scRNAseq datasets was collected, preprocessed, and analyzed with i) the current PCA→UMAP standard, ii) standalone UMAP (graph from high-dimensional gene expression space), iii) scVI (a popular tool for variational inference), and iv) TopoMetry. (b) Violin plots representing geometry-preservation metrics for lower-dimensional latent spaces learned with PCA and scVI, compared to TopoMetry's spectral scaffolds. TopoMetry's scaffolds achieved systematically higher scores across all metrics. (c) Violin plots representing geometry-preservation metrics for 2-D visualizations obtained with the evaluated methods. Except for PaCMAP on TopoMetry's multiscale spectral scaffold, the geometry-aware visualizations achieved systematically higher scores. Visualizations based on scVI and PCA latent space presented the lowest scores.

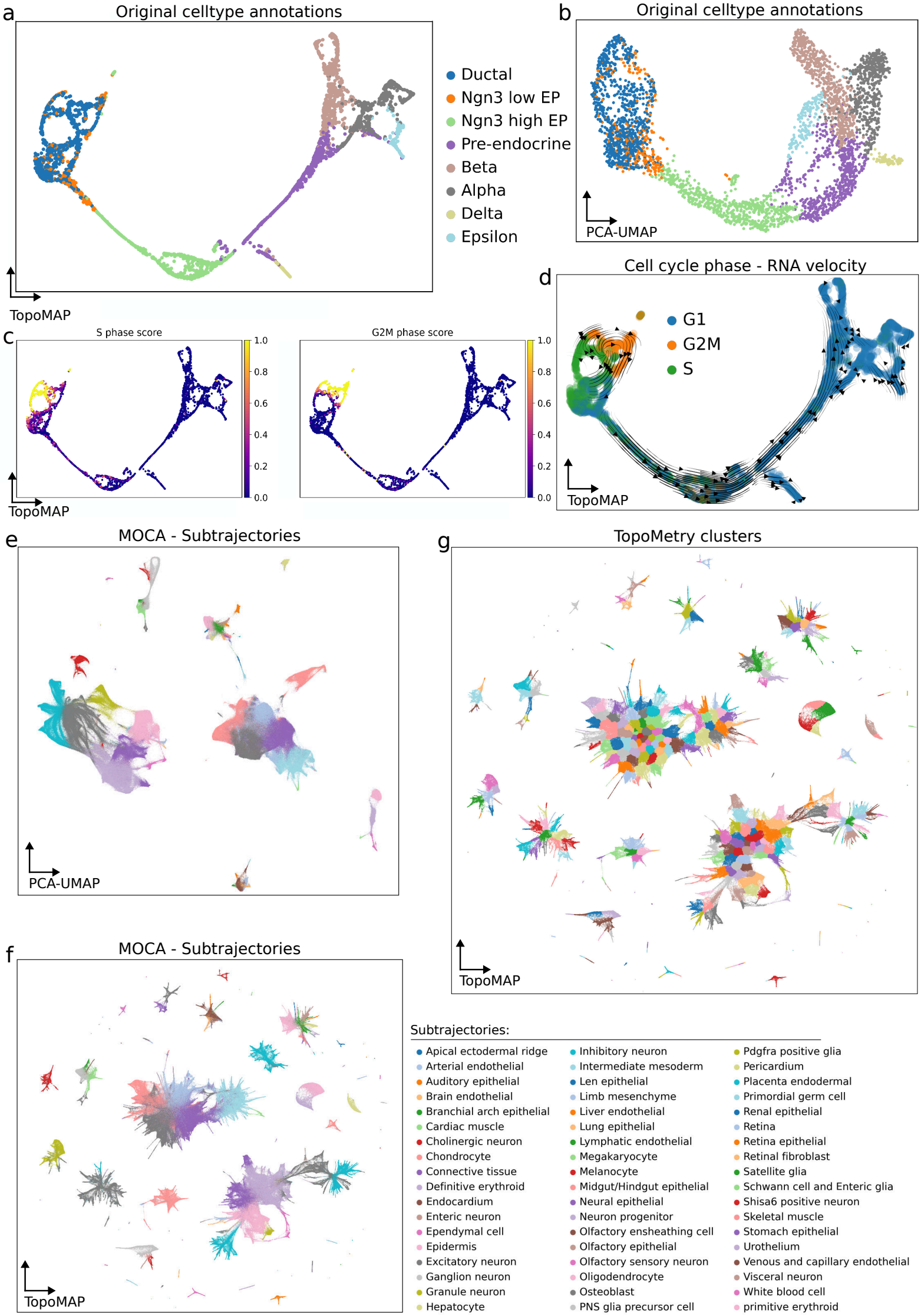

a
Original celltype annotations
Ductal
Ngn3 low EP
Ngn3 high EP
Pre-endocrine
Beta
Alpha
Delta
Epsilon
TopoMAP
b
Original celltype annotations
PCA-UMAP
c
S phase score
G2M phase score
1.0
0.8
0.6
0.4
0.2
0.0
TopoMAP
d
Cell cycle phase - RNA velocity
G1
G2M
S
TopoMAP
e
MOCA - Subtrajectories
PCA-UMAP
g
TopoMetry clusters
TopoMAP
f
MOCA - Subtrajectories
TopoMAP
Subtrajectories:
Apical ectodermal ridge
Arterial endothelial
Auditory epithelial
Brain endothelial
Branchial arch epithelial
Cardiac muscle
Cholinergic neuron
Chondrocyte
Connective tissue
Definitive erythroid
Endocardium
Enteric neuron
Ependymal cell
Epidermis
Excitatory neuron
Ganglion neuron
Granule neuron
Hepatocyte
Inhibitory neuron
Intermediate mesoderm
Len epithelial
Limb mesenchyme
Liver endothelial
Lung epithelial
Lymphatic endothelial
Megakaryocyte
Melanocyte
Midgut/Hindgut epithelial
Neural epithelial
Neuron progenitor
Olfactory ensheathing cell
Olfactory epithelial
Olfactory sensory neuron
Oligodendrocyte
Osteoblast
PNS glia precursor cell
Pdgfra positive glia
Pericardium
Placenta endodermal
Primordial germ cell
Renal epithelial
Retina
Retina epithelial
Retinal fibroblast
Satellite glia
Schwann cell and Enteric glia
Shisa6 positive neuron
Skeletal muscle
Stomach epithelial
Urothelium
Venous and capillary endothelial
Visceral neuron
White blood cell
primitive erythroid

Figure 3: Inferring cellular lineages with TopoMetry.
(a) TopoMAP and (b) PCA→UMAP visualizations of the Pancreas dataset showing cellular developmental trajectories in the murine pancreas, colored by original cell type annotations. (c) TopoMAP visualization, colored by inferred scores of different phases of the cell cycle, and (d) the predicted phase for each cell with RNA velocity overlay. Note how RNA velocity trajectories largely agree with the identified cell cycle structure and the represented geometry. (e) PCA→UMAP visualization of the Mouse Organogenesis Cell Atlas (MOCA), comprising ∼1.3 million cells collected during murine embryo development, colored by refined subtrajectories annotation. (f–g) TopoMAP visualizations of MOCA, colored by original annotations on refined subtrajectories (f), and TopoMetry's clustering results (g). Note how the TopoMAP embedding successfully separates main and refined trajectories and adds enhanced detail and resolution on the diversity of subpopulations arising during development.

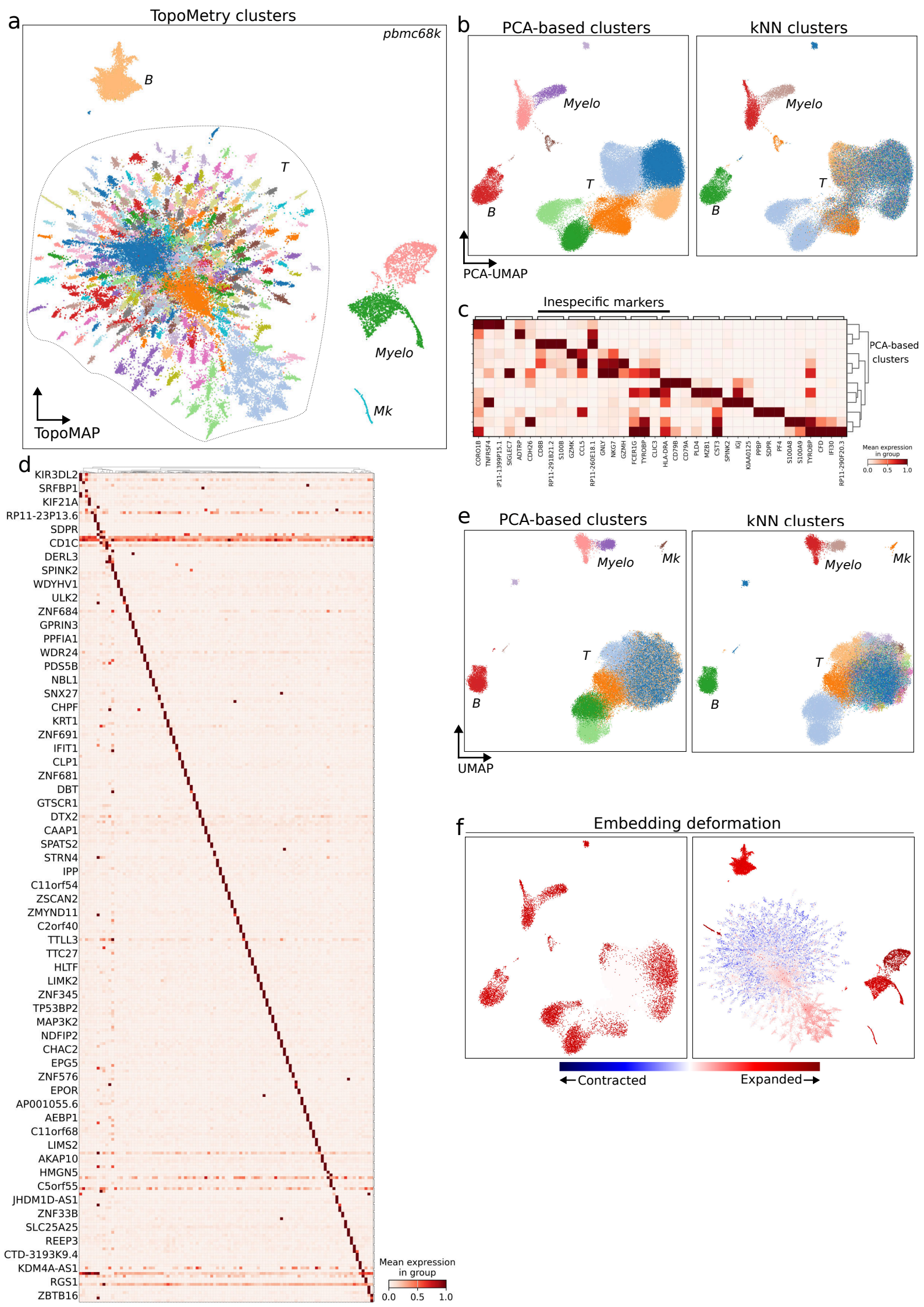

a
TopoMetry clusters
pbmc68k
B
T
Myelo
Mk
TopoMAP
b
PCA-based clusters
kNN clusters
Myelo
T
B
PCA-UMAP
c
Inespecific markers
PCA-based clusters
CORO1B
TNFRSF4
IP11-1399P15.1
SIGLEC7
ADTRP
CDH26
CD8B
RP11-291B21.2
S100B
GZMK
CCL5
RP11-260E18.1
GNLY
NKG7
GZMH
FCER1G
TYROBP
CLIC3
HLA-DRA
CD79B
CD79A
PLD4
MZB1
CST3
SPINK2
IGJ
KIAA0125
PPBP
SDPR
PF4
S100A8
S100A9
TYROBP
CFD
IFI30
RP11-290F20.3
Mean expression in group
0.0
0.5
1.0
d
KIR3DL2
SRFBP1
KIF21A
RP11-23P13.6
SDPR
CD1C
DERL3
SPINK2
WDYHV1
ULK2
ZNF684
GPRIN3
PPFIA1
WDR24
PDS5B
NBL1
SNX27
CHPF
KRT1
ZNF691
IFIT1
CLP1
ZNF681
DBT
GTSCR1
DTX2
CAAP1
SPATS2
STRN4
IPP
C11orf54
ZSCAN2
ZMYND11
C2orf40
TTLL3
TTC27
HLTF
LIMK2
ZNF345
TP53BP2
MAP3K2
NDFIP2
CHAC2
EPG5
ZNF576
EPOR
AP001055.6
AEBP1
C11orf68
LIMS2
AKAP10
HMGN5
C5orf55
JHDM1D-AS1
ZNF33B
SLC25A25
REEP3
CTD-3193K9.4
KDM4A-AS1
RGS1
ZBTB16
Mean expression in group
0.0
0.5
1.0
e
PCA-based clusters
kNN clusters
Myelo
Mk
T
B
UMAP
f
Embedding deformation
Contracted
Expanded

Figure 4: TopoMetry unveils unexpected transcriptional diversity of T cells.
Analysis of the pbmc68k dataset, comprising approximately 68,000 peripheral blood mononuclear cells from a healthy donor (10X Genomics). (a) TopoMAP visualization colored by TopoMetry's clustering results. Main cell types are well separated, and T cells present an unexpected high diversity, with approximately a hundred clusters identifying T cell subpopulations. (b) Standard PCA→UMAP visualizations colored by clustering results obtained with the PCA-derived graph (left) and the kNN graph from the high-dimensional gene expression space (right), presenting the same global separation of main cell types but disagreeing on T cells. (c) Matrixplot of the top 3 marker genes found for PCA-based clusters, highlighting the presence of non-specific markers for T cells. (d) Matrixplot of the top 3 marker genes found for TopoMetry clusters, presenting highly specific marker expression. (e) Standalone UMAP visualizations of the same data, colored by PCA-based (left) and kNN-based clustering results (right). Note how the standalone approach detects part, but not all, of the T cell clusters identified with TopoMetry. (f) Contraction/expansion diagnostics of PCA→UMAP (left) and TopoMAP (right) visualizations. Note how the PCA→UMAP approach expands most regions of the cell identity manifold, while TopoMAP contracts the region inhabited by T CD4 lymphocytes when projecting TopoMetry's refined graphs to a 2-D space.

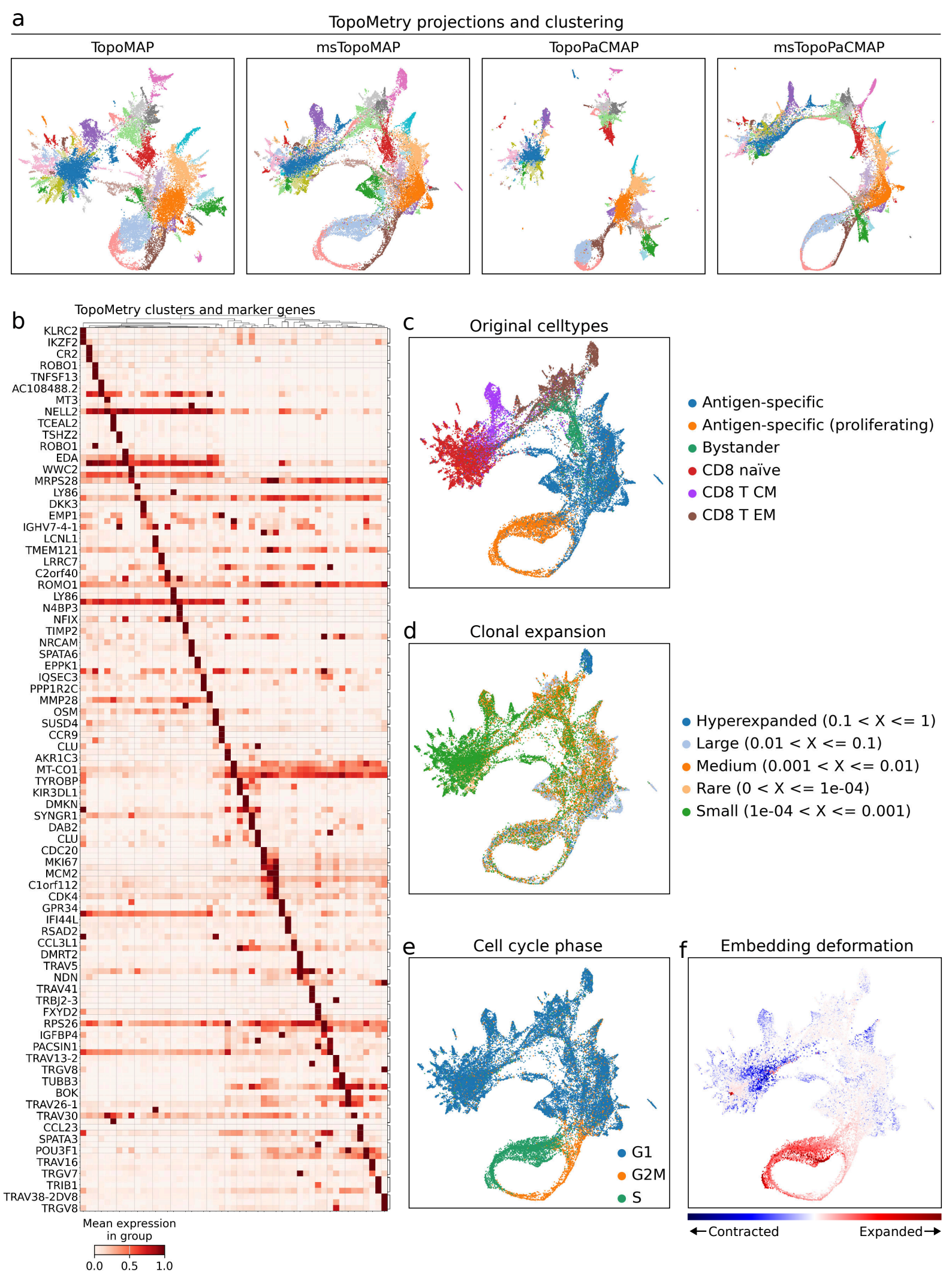

a
TopoMetry projections and clustering
TopoMAP
msTopoMAP
TopoPaCMAP
msTopoPaCMAP
b
TopoMetry clusters and marker genes
KLRC2
IKZF2
CR2
ROBO1
TNFSF13
AC108488.2
MT3
NELL2
TCEAL2
TSHZ2
ROBO1
EDA
WWC2
MRPS28
LY86
DKK3
EMP1
IGHV7-4-1
LCNL1
TMEM121
LRRC7
C2orf40
ROMO1
LY86
N4BP3
NFIX
TIMP2
NRCAM
SPATA6
EPPK1
IQSEC3
PPP1R2C
MMP28
OSM
SUSD4
CCR9
CLU
AKR1C3
MT-CO1
TYROBP
KIR3DL1
DMKN
SYNGR1
DAB2
CLU
CDC20
MKI67
MCM2
C1orf112
CDK4
GPR34
IFI44L
RSAD2
CCL3L1
DMRT2
TRAV5
NDN
TRAV41
TRBJ2-3
FXYD2
RPS26
IGFBP4
PACSIN1
TRAV13-2
TRGV8
TUBB3
BOK
TRAV26-1
TRAV30
CCL23
SPATA3
POU3F1
TRAV16
TRGV7
TRIB1
TRAV38-2DV8
TRGV8
Mean expression in group
0.0
0.5
1.0
c
Original celltypes
Antigen-specific
Antigen-specific (proliferating)
Bystander
CD8 naïve
CD8 T CM
CD8 T EM
d
Clonal expansion
Hyperexpanded (0.1 < X <= 1)
Large (0.01 < X <= 0.1)
Medium (0.001 < X <= 0.01)
Rare (0 < X <= 1e-04)
Small (1e-04 < X <= 0.001)
e
Cell cycle phase
G1
G2M
S
f
Embedding deformation
Contracted
Expanded

Figure 5: TopoMetry detects T cell clonal expansion dynamics from gene expression.
Analysis of the ECCITE-TCR dataset, comprising circulating T $CD8^+$ lymphocytes collected from human donors in baseline conditions and after SARS-CoV-2 vaccination or infection. (a) TopoMetry's default visualizations, colored by TopoMetry's clustering results. Note how projections derived from the fixed-time scaffold better preserve the local structure of the dataset, while projections derived from the multiscale scaffold better preserve long-range relationships and overall global structure. Despite minor differences, all visualizations correctly represent the cell cycle geometry of proliferating lymphocytes and the transcriptional diversity of central (TCM) and effector memory (TEM) lymphocytes. (b) Matrixplot of the top 3 marker genes found for TopoMetry clusters, presenting highly specific marker expression. (c–f) TopoMAP visualizations colored by original cell type annotations (c), clonal expansion information (d), predicted phase of the cell cycle (e), and contraction/expansion diagnostics. Note how the small clusters of TCM and TEM correspond to smaller clonotypes (ranging from small to medium), how the identified cell cycle geometry agrees with cell cycle predictions, and how the former are contracted while the latter are expanded in the 2-D visualization.