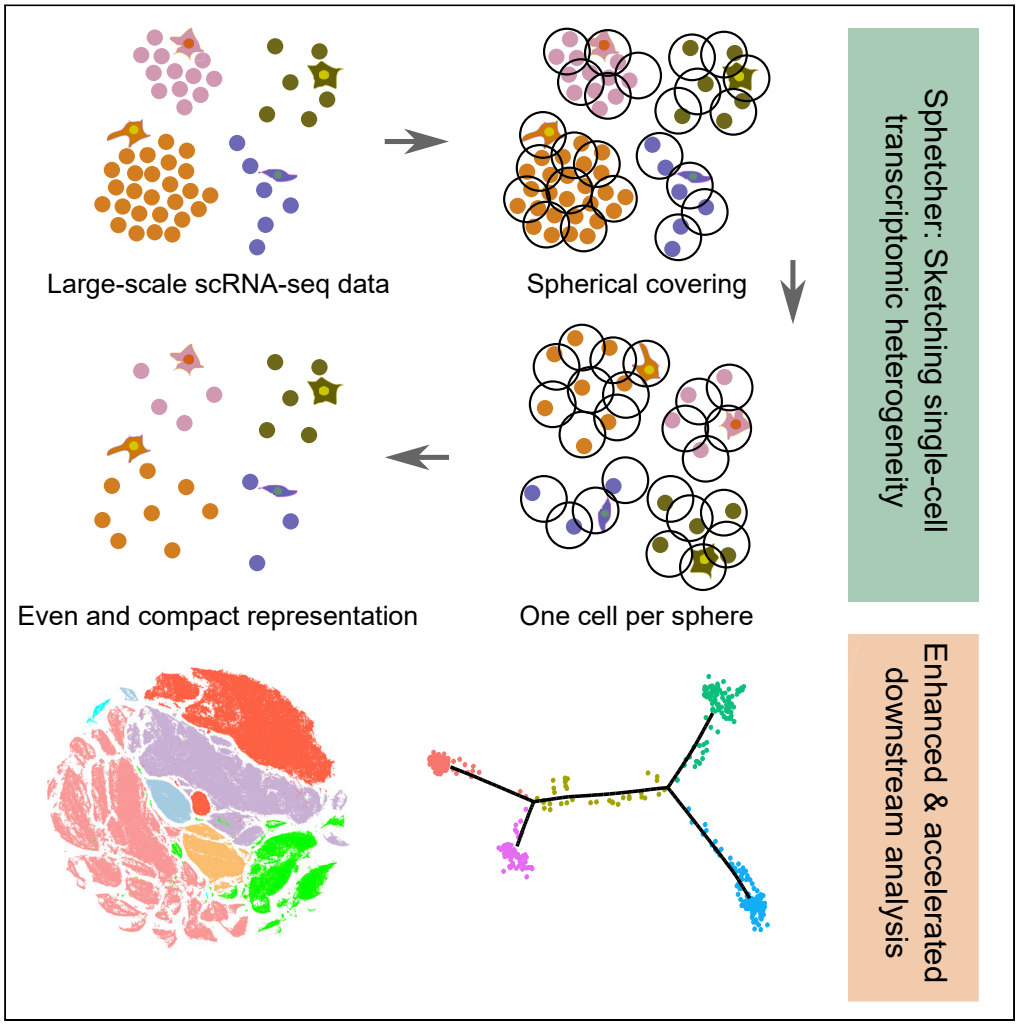
iScience

CellPress
OPEN ACCESS

Article

Sphetcher: Spherical Thresholding Improves Sketching of Single-Cell Transcriptomic Heterogeneity



Van Hoan Do, Khaled Elbassioni, Stefan Canzar

canzar@genzentrum.lmu.de

HIGHLIGHTS

Sphetcher distils large-scale scRNA-seq data down to a small selection of cells

Spheres of small radius around selected cells cover the original transcriptomic space

Selection enhances and accelerates downstream analysis such as trajectory inference

Sphetcher can leverage existing annotation of known cell types

Do et al., iScience 23, 101126
June 26, 2020
https://doi.org/10.1016/j.isci.2020.101126

iScience

CellPress
OPEN ACCESS

Article

Sphetcher: Spherical Thresholding Improves Sketching of Single-Cell Transcriptomic Heterogeneity

Van Hoan Do,[1] Khaled Elbassioni,[2] and Stefan Canzar[1,3,*]

SUMMARY

The massive size of single-cell RNA sequencing datasets often exceeds the capability of current computational analysis methods to solve routine tasks such as detection of cell types. Recently, geometric sketching was introduced as an alternative to uniform subsampling. It selects a subset of cells (the sketch) that evenly cover the transcriptomic space occupied by the original dataset, to accelerate downstream analyses and highlight rare cell types. Here, we propose algorithm Sphetcher that makes use of the thresholding technique to efficiently pick representative cells within spheres (as opposed to the typically used equal-sized boxes) that cover the entire transcriptomic space. We show that the spherical sketch computed by Sphetcher constitutes a more accurate representation of the original transcriptomic landscape. Our optimization scheme allows to include fairness aspects that can encode prior biological or experimental knowledge. We show how a fair sampling can inform the inference of the trajectory of human skeletal muscle myoblast differentiation.

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) has emerged as a revolutionary tool that can shed light on many corners of cell biology that were unaccessible to previous approaches. The technology has improved dramatically over the last few years, especially in terms of throughput. Droplet-based technologies allow to profile the expression of every gene in the genome for hundreds of thousands of cells at once, and even experiments profiling the transcriptome of millions of cells have become increasingly common (Cao et al., 2019). Furthermore, the meaningful interpretation of single-cell datasets requires their integration across different biological contexts, yielding datasets whose enormous size exceeds the capability of current computational analysis methods to solve routine tasks such as clustering, trajectory inference, and visualization in practical time or require excessive amounts of memory.

In practice, methods are often run on a smaller subset of the data to bridge the gap between the scalability of the algorithm and the volume of the data (Hie et al., 2019). The commonly applied uniform subsampling strategy, however, ignores the similarity or dissimilarity between gene expression patterns of single cells and thus risks overlooking rare cell states. Spatial random sampling (SRS) (Rahmani and Atia, 2017) and *k*-means++ (Arthur and Vassilvitskii, 2007), on the other hand, take into account the structure of the data when sampling the data. Experiments performed in Hie et al. (2019), however, demonstrated that these data-dependent methods do not scale efficiently to large datasets and provide unbalanced samples that hamper downstream analyses. Clustering the full data first followed by sampling from clusters, as performed by dropClust (Sinha et al., 2018), has similar issues (Hie et al., 2019). Hie et al. (2019) introduced *geometric sketching* as an alternative approach that efficiently samples cells evenly across gene expression space rather than proportional to the abundance of cells that are in a similar state. For purely computational reasons, however, Hie et al. (2019) approximate the transcriptomic space of single cells by equal-sized boxes rather than spheres, from within which cells are randomly selected as representatives into the *sketch*.

Here, we propose algorithm Sphetcher that makes use of the thresholding technique originally proposed for the design of approximation algorithms for bottleneck problems to efficiently pick representative cells within spheres of a fixed size into a *spherical sketch* of different metric spaces. We provide theoretical

[1]Gene Center, Ludwig-Maximilians-Universität München, 81377 Munich, Germany

[2]Khalifa University of Science and Technology, P.O. Box: 127788, Abu Dhabi, UAE

[3]Lead Contact

*Correspondence: canzar@genzentrum.lmu.de

https://doi.org/10.1016/j.isci.2020.101126

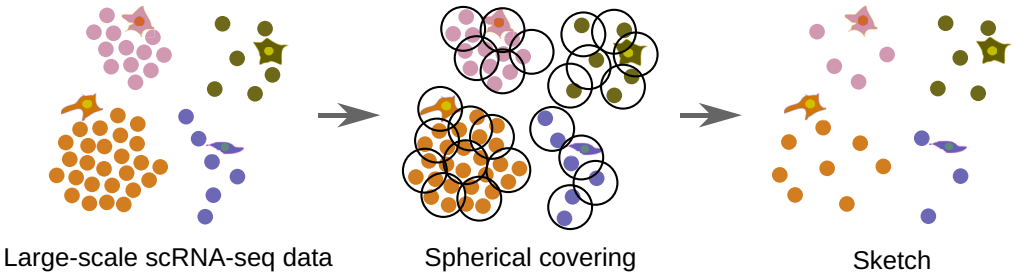


Figure 1. Overview of Sphetcher

For a (large) scRNA-seq dataset (left), Sphetcher uses a disk-friendly greedy algorithm to compute a smallest size set of spheres of a fixed radius that cover all cells (middle). It guesses the smallest possible radius such that a given number of spheres of that radius suffice to cover all cells. One representative cell (the center) from each sphere is selected into the final spherical sketch (right).

guarantees for the spherical sketch computed by Sphetcher and demonstrate through experiments on six single-cell datasets that these theoretical guarantees are indeed reflected in a more accurate representation of the original transcriptomic space, which in turn benefits downstream analyses such as clustering and allows to detect a rare population of inflammatory macrophages. Furthermore, our optimization scheme naturally allows to include fairness aspects that require to include cells of each pre-defined category that can encode prior biological or experimental knowledge such as cell type or collection time point. We demonstrate how our fairness-inspired model can help to incorporate the collection time point of cells in a time series experiment into the reconstruction of their developmental trajectory. Carefully combined with a prior grid sampling strategy that is orders of magnitude faster than geometric sketching, Sphetcher requires only 16 minutes to compute a sketch for a mouse embryonic dataset comprising two million cells.

RESULTS

Overview of Our Spherical Sketching Algorithm

Given a large scRNA-seq dataset, we seek to select a subset of cells, a so-called *sketch* (Hie et al., 2019), that evenly represents the geometry of the transcriptional space occupied by the original data. As originally proposed in Hie et al. (2019), we intuitively aim at capturing the transcriptional heterogeneity of single cells by removing predominantly cells that show similar expression patterns to other cells while preserving rare cell states. A sketch of a given size represents the full data well if every original cell is close to a cell in the sketch, according to some measure of distance between two cells. In other words, spheres of a small radius centered at each cell in the sketch must contain, or *cover*, every cell in the full dataset. The smaller the radius, the better the sketch represents the original transcriptional space.

Our algorithm implemented in software tool Sphetcher guesses the smallest possible radius for which a sketch of a given size exists that covers all remaining cells with spheres of this radius (Figure 1). For each guess, it computes the smallest size sketch that covers all cells and tries a smaller or larger radius in the next iteration if the resulting sketch contains too few or too many cells, respectively. It computes the smallest sketch that covers all cells using a greedy set cover approach. In each iteration it adds the cell to the sketch that contains the largest number of yet uncovered cells within the given distance. We employ the disk-friendly greedy (DFG) algorithm developed in Cormode et al. (2010) that scales to very large scRNA-seq datasets. For very large datasets, the spherical sketching approach is combined with a prior grid sampling that we show increases the radius of covering spheres by only a small factor (Transparent Methods).

In addition, our greedy algorithm can incorporate prior categorical information on, e.g., biological cell types or collection time point of cells. In a *fairness*-inspired model it selects at least a given number of representatives from each class into the sketch.

A detailed description of our algorithm and the parameters used in the experiments are provided in Transparent Methods. We also provide a theoretical analysis that shows that if we are willing to include slightly more cells in the sketch, our greedy algorithm is guaranteed to find the covering of cells with spheres with optimal, that is, with smallest possible radius. Furthermore, we give theoretical justification for the practical performance of our greedy set cover approach and its robustness to noise present in scRNA-seq data.

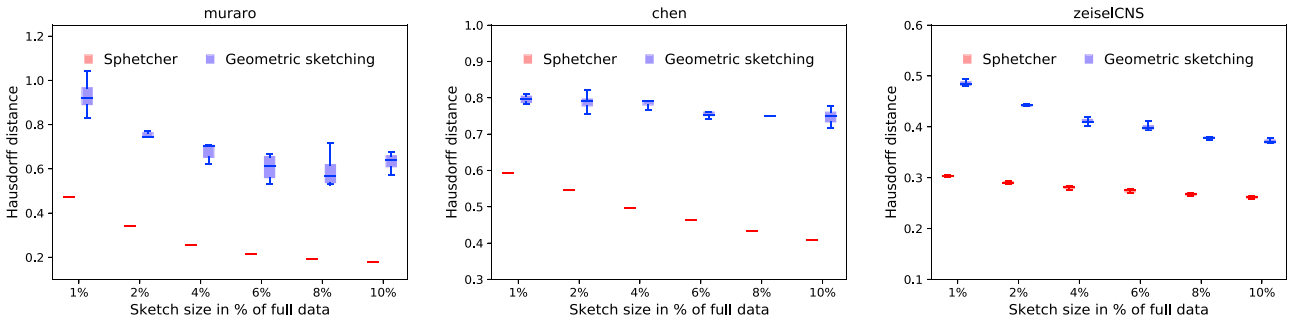


Figure 2. Comparison of Hausdorff Distances

The spherical sketch computed by Sphetcher exhibits consistently smaller Hausdorff distances to the full dataset than geometric sketching, across datasets and sketch sizes. For each sketch size, the results of 10 random trials are shown. Results on datasets zeisel, klein, and saunders are shown in Figure S1. Figure S3 shows Hausdorff distances achieved by our naïve grid sampling strategy on datasets zeiselCNS and saunders.

Sphetcher More Accurately Sketches the Transcriptomic Space

To evaluate how well the *spherical sketch* computed by our method Sphetcher represents the original transcriptomic space, we use the same robust Hausdorff distance measure as Hie et al. (2019) (Transparent Methods). Intuitively, a small Hausdorff distance between a sketch and a full dataset indicates an accurate representation that contains for every cell in the original data a close cell in the sketch. We compare our sketch to the *geometric sketch* computed by Hie et al. (2019), which the authors demonstrated to consistently achieve smaller Hausdorff distances than uniform sampling and data-dependent sampling methods SRS and *k*-means++. The geometric sketch computed in Hie et al. (2019) seeks to minimize the same objective function (Transparent Methods) but simplifies the approximation of the geometric space by equal-sized boxes rather than spheres. We benchmark Sphetcher on six public single-cell datasets from mouse and human that vary in size and number of cell populations: human pancreas (*muraro*) (Muraro et al., 2016) with 2,126 cells, 10 populations; mouse embryonic stem cells (*klein*) (Klein et al., 2015) with 2,717 cells, 4 populations; mouse cortex and hippocampus (*zeisel*) (Zeisel et al., 2015) with 3,005 cells, 9 populations; mouse hypothalamus (*chen*) (Chen et al., 2017) with 14,437 cells, 47 populations; mouse nervous system (*zeiselCNS*) (Zeisel et al., 2018) with 465,281 cells, 7 populations; and adult mouse brain (*saunders*) (Saunders et al., 2018) with 665,858 cells and 11 populations. Figures 2 and S1 show the Hausdorff distances of 10 random trials on sketch sizes ranging from 1% to 10% of the full dataset. Values reported here can deviate slightly from the original publication (Hie et al., 2019) due to different preprocessing (Transparent Methods). Our sampling approach based on spheres results in sketches that consistently lead to smaller Hausdorff distances, across datasets and sketch sizes. As expected, larger sketches yield smaller Hausdorff distances, but across all datasets the geometric sketch based on 10% of the data does not represent the full data as well as our spherical sketch with just 1% of the data. In addition, sketches computed by Sphetcher exhibit a considerably smaller variability over the random trials (Figure S2). Although the geometric sketch randomly picks a cell in each box, Sphetcher's only random decision is in breaking ties between equal-sized sets during the greedy set cover computation (Transparent Methods). Remarkably, our naive grid sampling strategy alone, which is part of our hybrid alternative for very large datasets (Transparent Methods), achieves competitive Hausdorff distances on datasets zeiselCNS and saunders, especially for small sketch sizes (Figure S3).

Clustering of Spherical Sketches Facilitates Cell-Type Identification

A common goal in scRNA-seq data analysis is to discover and characterize cell types, typically through clustering methods. The quality of the clustering therefore plays a critical role in biological discovery. The compact size of a geometric or spherical sketch that accurately summarizes the transcriptional heterogeneity in the full data facilitates such downstream analyses. Furthermore, Hie et al. (2019) observed that a more balanced composition of abundant and rare cell types in a geometric sketch allows to better distinguish between cell types compared with a uniform sampling approach. Here, we apply a similar strategy as in Hie et al. (2019) to evaluate the capability of a standard clustering algorithm to distinguish cell types based on our spherical sketch as compared with the geometric sketch. We first cluster the sketches using the graph-based Louvain algorithm (Blondel et al., 2008) and then propagate the labels to the remaining cells by *k*-nearest neighbor classification. We use the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) to measure the similarity between the inferred clusterings and the ground truth clustering, which is based

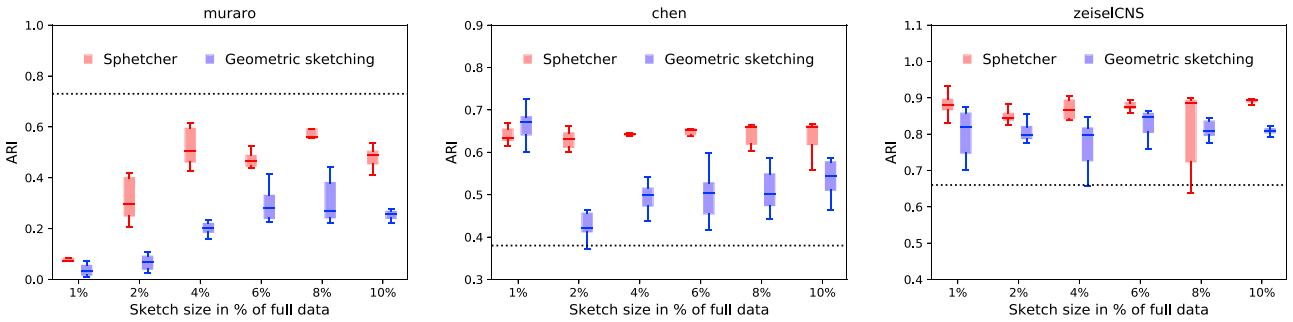


Figure 3. Comparison of Sketch-Based Clustering Accuracy

Louvain clustering of spherical sketches computed by Sphetcher yields more accurate cell clusterings as measured by Adjusted Rand Index (ARI) than geometric sketching based clustering. In both cases, labels assigned to cells in the sketch are propagated to the remaining cells using *k*-nearest neighbor classification. The dotted line indicates the ARI score achieved by clustering the full data using the same Louvain algorithm. Results on datasets zeisel, klein, and saunders are shown in Figure S4.

on the biological cell types taken from the original study. Hie et al. (2019) demonstrated that unsupervised clustering of geometric sketches consistently outperform clusterings of uniformly sampled cells, whereas data-dependent methods *k*-means++ and SRS provide competitive results on only a few instances. In Figures 3 and S4 we show that the more even sampling of the transcriptional landscape by our spherical sketch facilitates the detection of biological cell types. Across datasets and sampling sizes, the clustering of our spherical sketches achieves better or comparable separation of cell types than the clustering of the corresponding geometric sketch. In only three out of thirty-six instances, geometric sketching yielded slightly better median ARI scores. Remarkably, in several cases the clustering of sketches better agrees with the true biological cell types than the clustering based on the full data. This observation is consistent with the assumption of a more balanced composition of cell types in a sketch, but an artifact of the clustering algorithm cannot be excluded, especially in light of the *impossibility theorem for clustering* (Kleinberg, 2003). Note that despite a small variability in Hausdorff distance, the non-deterministic behavior of the Louvain algorithm contributes to the different ARI scores observed in the repeated clustering of spherical sketches.

Impact of Distance Metrics

Downstream analysis of scRNA-seq such as clustering and trajectory inference relies on a metric that measures the distance between cells in gene expression space. Distance metrics such as Euclidean distance, correlation-based distance, and cosine similarity (adapted as distance) have been proposed as adequate measures of dissimilarity, and its specific choice might depend on assumptions made by computational analysis methods, properties of datasets, and the specific task at hand (Kim et al., 2018; Jaskowiak et al., 2014). Although the Hausdorff distance is defined based on a given metric, geometric sketching ignores the metric space and considers absolute differences in each dimension independently.

Here, we illustrate the flexibility of Sphetcher in optimizing the Hausdorff distance under different distance metrics (see Transparent Methods) and demonstrate that the choice of metric can impact downstream clustering analysis of scRNA-seq data. To this end, we sample a subset of cells from a medium size dataset with complex population structure (*chen*) using Sphetcher with four different metrics: Euclidean, Manhattan, cosine, and Pearson correlation distance. We cluster the four resulting sketches using the same approach as in the previous section and compare the quality of the clusterings with the one obtained from a geometric sketch. Note that the geometric sketching approach proposed in Hie et al. (2019) cannot distinguish different distance metrics. Figure 4 shows that spherical sketches computed by Sphetcher using Euclidean distance as metric in the objective function yield most accurate clusterings of this dataset. Although cosine and Pearson distances have a slightly negative effect on the quality of the clustering, Manhattan distance and geometric sketching yield substantially less accurate clusterings, especially for small sketch sizes.

On dataset *muraro*, geometric sketching again achieves overall lower ARI scores than Sphetcher using different metrics (Figure 4). In contrast to dataset *chen*, however, Euclidean-distance-based sampling does not show any improvement over alternative metrics, illustrating the benefit of Sphetcher's unique ability to take into account different metrics suitable for different tasks.

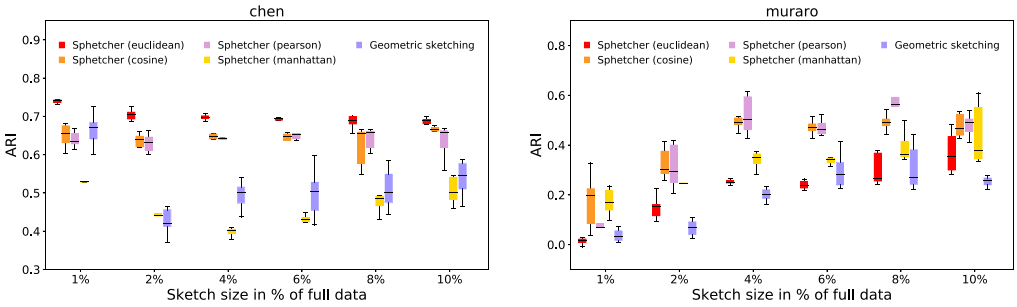


Figure 4. Impact of Distance Metrics on Clustering Performance

Although clustering based on spherical sketches computed by Sphetcher using Euclidean distance yields most accurate results on dataset chen, alternative metrics used by Sphetcher lead to higher ARI scores on dataset muraro, illustrating the importance of Sphetcher's flexible optimization scheme. In contrast, geometric sketching does not distinguish different distance metrics and yields overall less accurate clusterings.

Sphetcher Detects Rare Population of Inflammatory Macrophages

Hie et al. (2019) report and experimentally validate the discovery of a rare population of inflammatory macrophages by clustering a geometric sketch of 20,000 cells sampled from a dataset of 254,941 umbilical cord blood cells. In contrast, clustering the full dataset or a uniform subsample did not reveal this rare population of cells, presumably due to their limited visibility among the more abundant inactive macrophages. We repeated the experiment by clustering our spherical sketch of same size (20,000 cells) obtained after prior grid sampling (Sphetcher-H, Transparent Methods) using the Louvain community detection algorithm. As expected, we were also able to discover a similar cluster of inflammatory macrophages based on the same set of marker genes CD74, HLA-DRA, B2M, and JUNB (AUROC >0.88, Transparent Methods).

Fairness Incorporates Time Points in Trajectory Reconstruction

In time series studies of gene expression, single cells are typically collected at different (known) time points. In this section, we illustrate how fairness aspects can be used to incorporate this additional information into the construction of a spherical sketch. To compare the gene expression dynamics of human skeletal muscle myoblast (HSMM) differentiation with the reprogramming of fibroblasts to myotubes, in Cacchiarelli et al. (2018), single cells were sampled every 24 h post-induction of myoblast differentiation, between 0 and 72 h. Consistent with the original publication, we reconstruct the single-cell trajectory of HSMM differentiation using Monocle 2 (Qiu et al., 2017), ignoring the information on the collection time point of cells. Figure 5 (left) shows the resulting trajectory, in which cells are initially in a cycling state and either fully progress to contractile myotubes or fail to differentiate. Cells are colored by the four different time points. For marked cells (black circle) the inferred pseudotime, i.e. their level of progression through differentiation, and the actual time they were collected, disagree. Even though cells do not always progress through the process of differentiation in a synchronous manner, the presence of fully differentiated cells at time point 0, for example, is most likely an artifact caused by noise in the single-cell measurements.

We sought to automatically detect and remove cells for which the collection time point disagrees with their transcriptomic state through a constrained sketching approach. Instead of imposing a hard constraint that removes "outlier" cells, we let our sketching algorithm decide if cells at different time points are necessary to evenly represent the global transcriptional space. Because our fairness-inspired model imposes covering constraints that require a certain number of cells to be sampled from each time point (Transparent Methods), a fair sampling of cells will implicitly discourage the selection of outlier cells that lie close to cells in a similar state but which have been collected at different time points.

We compare the trajectories computed by Monocle 2 from the geometric sketch, our (unconstrained) spherical sketch, and our fairness-inspired spherical sketch that picks at least four cells from each time point. On all sketches, the overall structure of the inferred trajectory agrees with the trajectory computed from the full data (Figures 5 (right) and S5–S7). However, although outlier cells are included in both the geometric sketch (8 out of 8 trials, Figure S6) and the unconstrained spherical sketch (2 out of 8 trials, Figure S7), Sphetcher under fairness constraints decides to not use outlier cells to represent the transcriptional space. Fairness encourages Sphetcher, for example, to not include fully differentiated cells from time point 0 into the sketch (Figures 5 (right) and S5). Even more, although constrained Sphetcher includes at least one cell

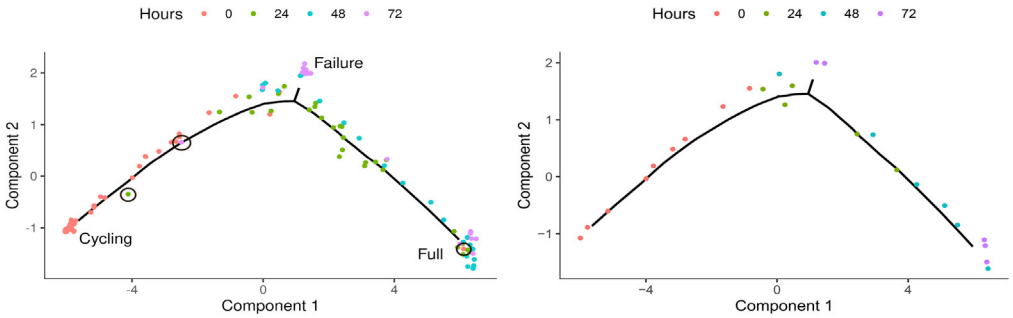


Figure 5. Single-Cell Trajectories of HSMM Differentiation

Single-cell trajectories of HSMM differentiation as reconstructed by Monocle 2 from the full data (left) and from Sphetcher's spherical sketch with fairness constraints (right) consistently describe progression through differentiation. Cells for which inferred pseudotime and collection time point disagree are marked with a black circle and were automatically removed as "outlier" cells by Sphetcher. See also Figures S5–S7.

collected at time point 72 in the final state (Full) in Figure 5 and in all trials in Figure S5, unconstrained sketches do not retain any such cell in any but a single trial (Figures S6 and S7).

In addition, we construct gene expression kinetics plots using Monocle 2 for a set of genes assessed in Cacchiarelli et al. (2018). The expression dynamics inferred from our fair spherical sketch appear smoother than those obtained from the full data, and cells in our sketch better fit the interpolated expression (Figure S8).

Scalability

Here, we demonstrate scalability of our hybrid strategy Sphetcher-H that combines grid sampling with subsequent spherical sketching (Transparent Methods) to large single-cell datasets. In Table 1 we compare the running time of Sphetcher-H with the construction of a geometric sketch (Hie et al., 2019) on the zeiselCNS, saunders, and umbilical cord blood datasets used in previous benchmarks as well as on a dataset (*cao*) comprising two million cells (Cao et al., 2019). On the latter dataset, geometric sketching and Sphetcher-H require in total around 30 min and 16 min of computation, respectively. Remarkably, our naive grid sampling strategy alone is orders of magnitude faster than geometric sketching but achieves competitive Hausdorff distances on the zeiselCNS and saunders datasets (Figure S3).

DISCUSSION

We have introduced Sphetcher, a novel method that computes a small sketch of single-cell datasets that accurately summarizes its transcriptional heterogeneity. Sphetcher utilizes the thresholding technique to efficiently pick representative cells within spheres that better approximate the global geometry than boxes. Furthermore, we provide theoretical justification for its robust performance in practice. Sphetcher is able to accelerate scRNA-seq analyses such as the detection of cell types through clustering or the reconstruction of developmental trajectories. At the same time, it has the ability to shift the focus from a "more data, less algorithm" regime to a "less (but accurate) data, more algorithm" approach. For example, highly accurate yet computationally expensive algorithms such as consensus clustering by SC3 (Kiselev et al., 2017) might become practical again on a spherical sketch computed by Sphetcher from a large-scale dataset. In addition, Sphetcher is sensitive to rare cell types, is flexible in its use of different distance metrics, and allows to use prior categorical information on, e.g., biological cell types or collection time point to guide the selection of cells into a representative sketch.

Limitations of the Study

In most of the experiments in this study, Sphetcher used Pearson correlation as distance metric and was combined with a specific algorithm for downstream analysis. Even though Louvain community detection and Monocle 2 are widely used for scRNA-seq clustering and the inference of single-cell trajectories, respectively, Sphetcher's underlying model might be less compatible with assumptions made by other algorithms. In particular, Sphetcher's aim to minimize the maximum distance of cells to the sketch according to some metric might conflict with internal preprocessing routines applied by computational scRNA-seq analysis software. This interplay of sketching with respect to a given distance metric and subsequent algorithmic analysis was not systematically addressed in this study.

Dataset	# Cells	Sphetcher-H			Geometric Sketching
		Grid	Distances	Set Cover	
Cord blood	254,941	1.0	43.0	88.0	23.0
ZeiselCNS	464,713	3.0	153.0	116.0	120.0
Saunders	665,385	5.0	318.0	200.0	201.0
Cao	2,026,641	10.0	600.0	400.0	1869.0

Table 1. Comparison of CPU Time (in Seconds) of Geometric Sketching and Sphetcher-H

Running times are reported separately for the prior grid sampling, the calculation of pairwise distances, and the computation of a covering of all cells with spheres using a greedy set cover approach (Transparent Methods).

Furthermore, the incorporation of collection time points of cells in trajectory reconstruction demonstrates proof of principle. Additional experiments are required to fully address the benefits of leveraging prior (partial) knowledge on, e.g., cell types in the selection of representative cells into a sketch.

Resource Availability

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Stefan Canzar (canzar@genzentrum.lmu.de).

Materials Availability

This study did not generate new materials.

Data and Code Availability

Sphetcher is available at https://github.com/canzarlab/Sphetcher, where we also make spherical sketches of public, large scRNA-seq dataset available for download.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101126.

ACKNOWLEDGMENTS

V.H.D. was supported by a Deutsche Forschungsgemeinschaft fellowship through the Graduate School of Quantitative Biosciences Munich. We thank the Hemberg Group at the Sanger Institute for providing gene counts for datasets muraro, klein, zeisel, and chen, and the authors of Hie et al. (2019) for providing datasets zeiselCNS, saunders, and the dataset of umbilical cord blood cells.

AUTHOR CONTRIBUTIONS

All authors conceived the algorithm. V.H.D. performed the computational experiments. All authors wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 20, 2020
Revised: April 20, 2020
Accepted: April 28, 2020
Published: June 26, 2020

REFERENCES

Arthur, D. and Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding, In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. *2008*, P10008.

Cacchiarelli, D., Qiu, X., Srivatsan, S., Manfredi, A., Ziller, M., Overbey, E., Grimaldi, A., Grimsby, J., Pokharel, P., Livak, K.J., et al. (2018). Aligning single-cell developmental and reprogramming trajectories identifies molecular determinants of myogenic reprogramming outcome. Cell Syst. *7*, 1–18.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502.

Chen, R., Wu, X., Jiang, L., and Zhang, Y. (2017). Single-cell RNA-seq reveals hypothalamic cell diversity. Cell Rep. *18*, 3227–3241.

Cormode, G., Karloff, H., and Wirth, A. (2010). Set Cover Algorithms for Very Large Datasets (CIKM), pp. 479–488.

Hie, B., Cho, H., DeMeo, B., Bryson, B., and Berger, B. (2019). Geometric sketching compactly summarizes the single-cell transcriptomic landscape. Cell Syst. *8*, 483–493.e7.

Hubert, L., and Arabie, P. (1985). Comparing partitions. J. Classif. *2*, 193–218.

Jaskowiak, P.A., Campello, R.J., and Costa, I.G. (2014). On the selection of appropriate distances for gene expression data clustering. BMC Bioinform. *15*, S2.

Kim, T., Chen, I.R., Lin, Y., Wang, A.Y.-Y., Yang, J.Y.H., and Yang, P. (2018). Impact of similarity metrics on single-cell RNA-seq data clustering. Brief. Bioinform. *20*, 2316–2326.

Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., and Hemberg, M. (2017). Sc3: consensus clustering of single-cell RNA-seq data. Nat. Methods *14*, 483–486.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Kleinberg, J.M. (2003). An impossibility theorem for clustering. In NeurIPS 15, S. Becker, S. Thrun, and K. Obermayer, eds. (MIT Press), pp. 463–470.

Muraro, M.J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M.A., Carlotti, F., de Koning, E.J.P., and van Oudenaarden, A. (2016). A single-cell transcriptome atlas of the human pancreas. Cell Syst. *3*, 385–394.e3.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods *14*, 979–982.

Rahmani, M., and Atia, G.K. (2017). Spatial random sampling: a structure-preserving data sketching tool. IEEE Signal. Process. Lett. *24*, 1398–1402.

Saunders, A., Macosko, E.Z., Wysoker, A., Goldman, M., Krienen, F.M., de Rivera, H., Bien, E., Baum, M., Bortolin, L., Wang, S., et al. (2018). Molecular diversity and specializations among the cells of the adult mouse brain. Cell *174*, 1015–1030.e16.

Sinha, D., Kumar, A., Kumar, H., Bandyopadhyay, S., and Sengupta, D. (2018). dropClust: efficient clustering of ultra-large scRNA-seq data. Nucleic Acids Res. *46*, e36.

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L.E., La Manno, G., et al. (2018). Molecular architecture of the mouse nervous system. Cell *174*, 999–1014.e22.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.

iScience, Volume *23*

Supplemental Information

Sphetcher: Spherical Thresholding Improves Sketching of Single-Cell Transcriptomic Heterogeneity

Van Hoan Do, Khaled Elbassioni, and Stefan Canzar

Supplemental Figures

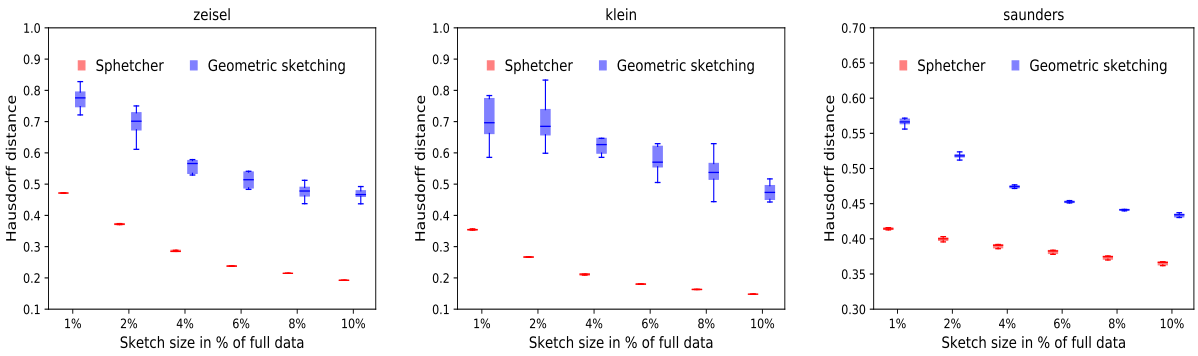


Figure S1. Comparison of Hausdorff distances, related to Figure 2. The spherical sketch computed by Sphetcher exhibits consistently smaller Hausdorff distances to the full dataset than geometric sketching, across datasets and sketch sizes. For each sketch size, the results of 10 random trials are shown.

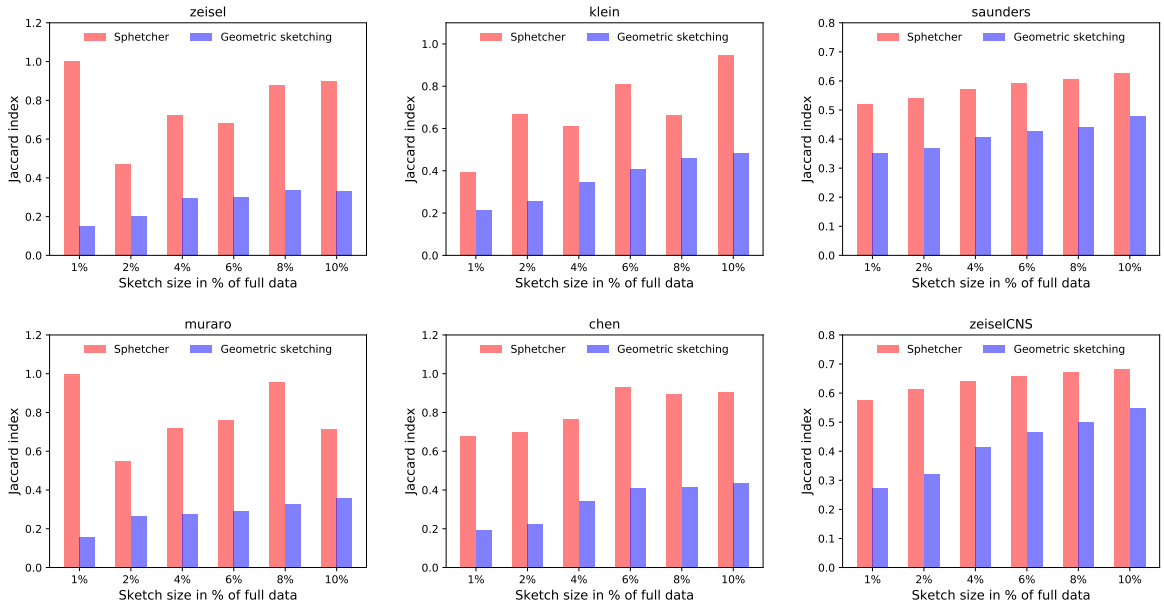


Figure S2. Comparison of Jaccard index, related to Figure 2. We compare the composition of the sketches computed in different random trials. The Jaccard index is computed for all pairs of random trials and the average is taken over all pairs for a given sketch size. The Jaccard index measures the similarity of two sketches by dividing the number of cells that they have in common by the total number of cells contained in either of the sketches. The Jaccard index ranges from 0 to 1, where 0 indicates that the two sketches have no cells in common, while 1 indicates identical sketches. Sphetcher returns highly similar sketches in different random trials, while the set of cells contained in geometric sketches can vary considerably between runs. In addition, these different geometric sketches differ in the quality of representation of the original transcriptomic space (Figure 2 and Figure S1). Note that the similarity of geometric sketches returned in different runs slowly increases with larger sample size, since the algorithm has fewer choices to pick a cell in smaller boxes. In contrast, Sphetcher's random tie breaking between equal-sized sets does not depend on the sample size and thus provides highly stable sketches even for small numbers of cells.

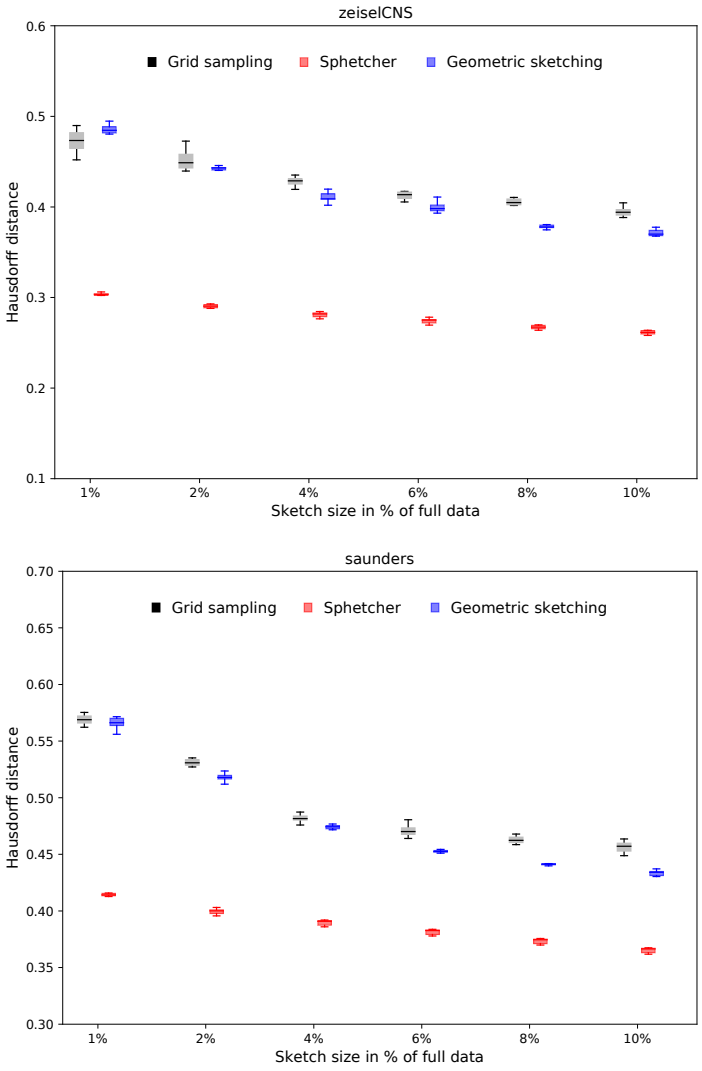


Figure S3. Comparison of Hausdorff distances, related to Figure 2. The naïve grid sampling strategy alone, which is part of our hybrid alternative for very large datasets (Transparent Methods), achieves competitive Hausdorff distances to geometric sketching on datasets zeiselCNS and saunders, especially for small sketch sizes. For each sketch size, the results of 10 random trials are shown.

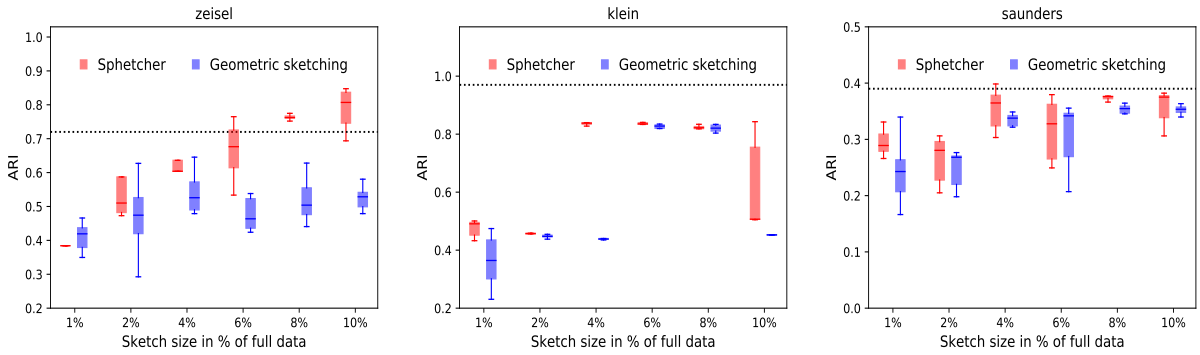


Figure S4. Comparison of sketch based clustering accuracy, related to Figure 3. Louvain clustering of spherical sketches computed by Sphetcher yields more accurate cell clusterings than geometric sketching based clustering. The dotted line indicates the ARI score achieved by clustering the full data using the same Louvain algorithm.

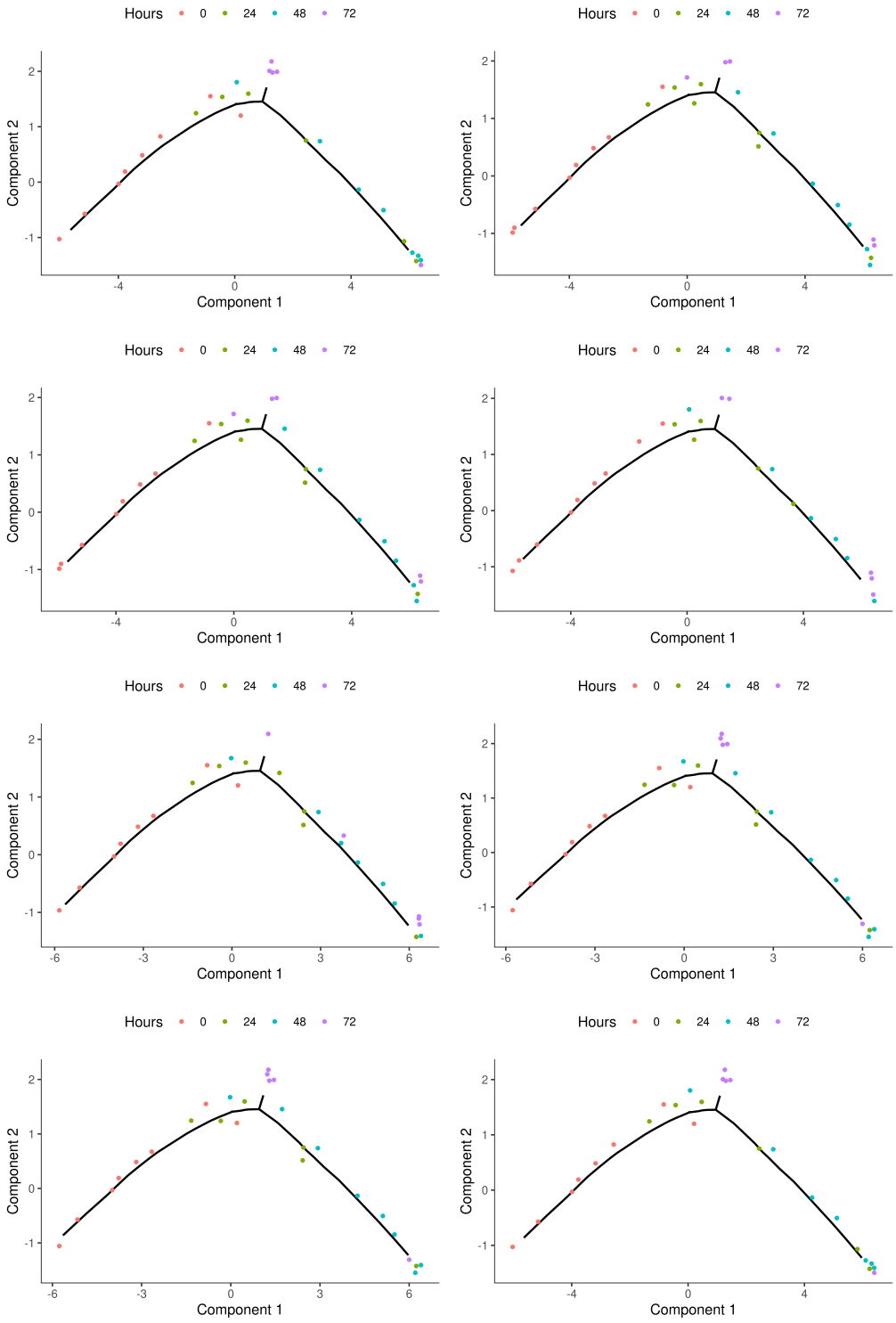


Figure S5. HSMM differentiation trajectories reconstructed by Monocole 2 from Sphetcher's sketch with fairness constraints, related to Figure 5. In 8 trials, Sphetcher did not include 'outlier' cells when its fairness model requires to include at least 4 cells from each time point. For outlier cells inferred pseudotime and actual collection time disagree. At the same time, cells collected at time point 72 in the final state are consistently retained.

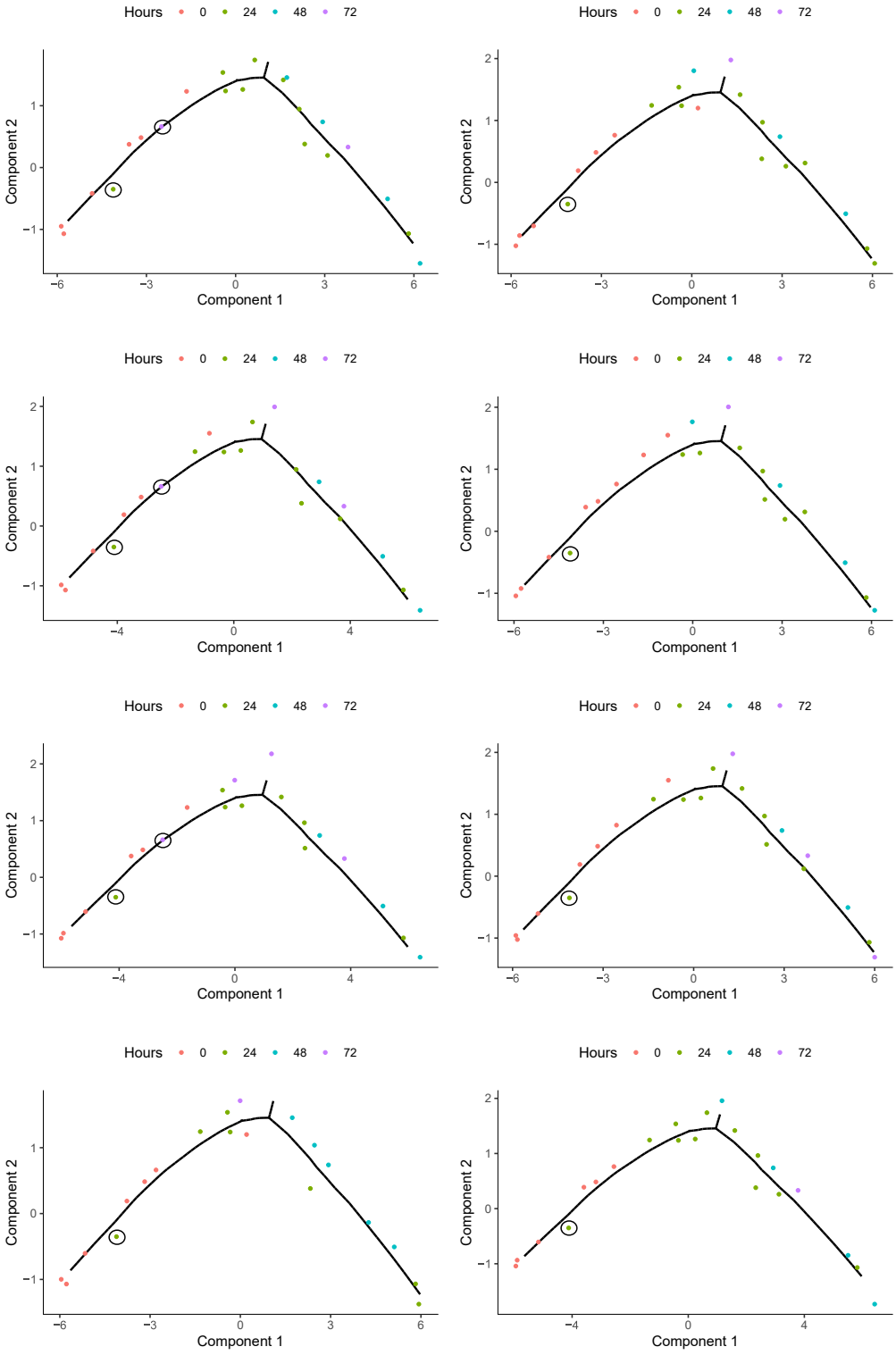


Figure S6. HSMM differentiation trajectories reconstructed by Monocole 2 from geometric sketches, related to Figure 5. In each of 8 trials, geometric sketches included outlier cells (black circles) for which inferred pseudotime and actual collection time disagree. At the same time, in only a single case a cell in final state collected at time point 72 is retained.

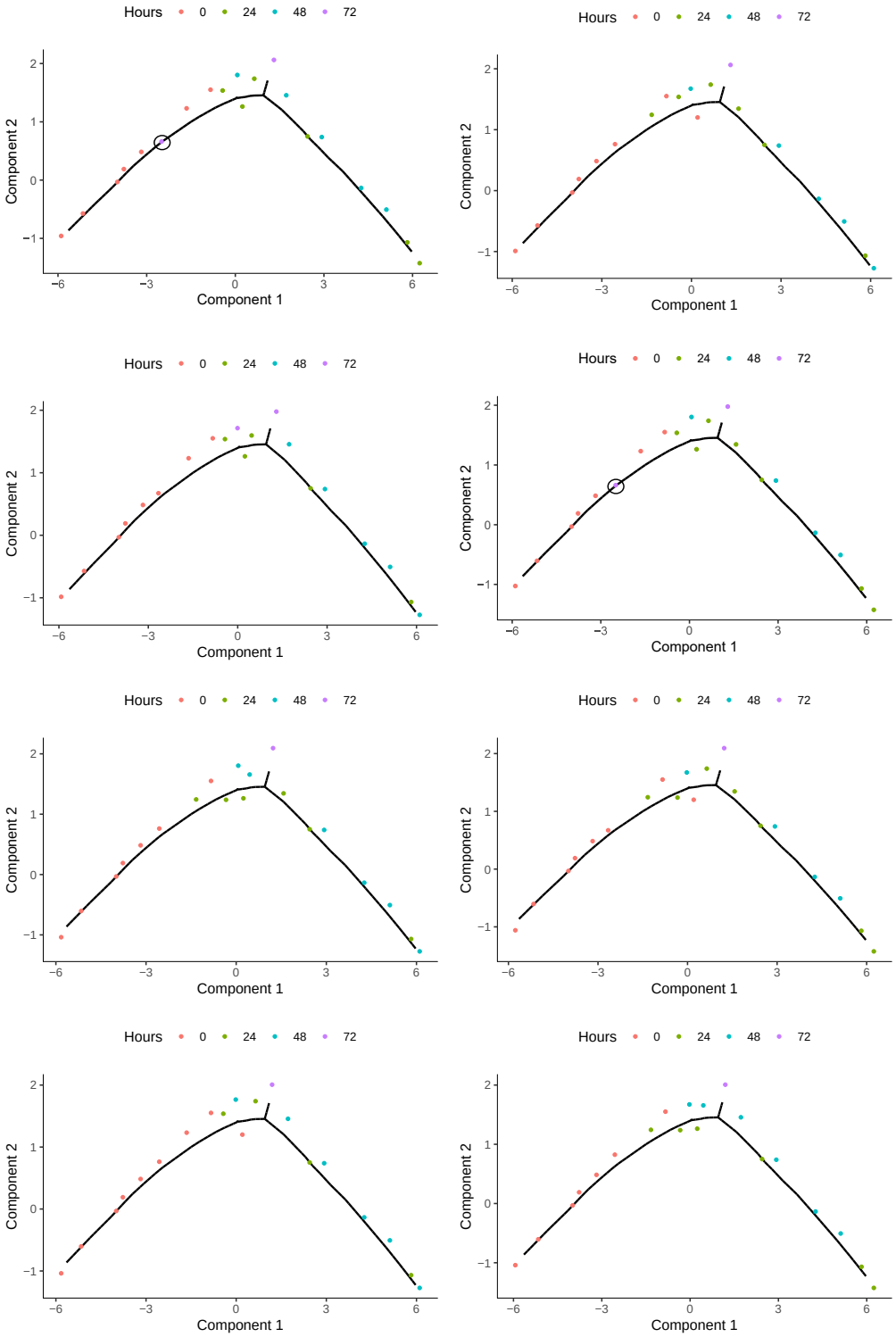


Figure S7. HSMM differentiation trajectories reconstructed by Monocole 2 from Sphetcher's sketch without fairness constraints, related to Figure 5. In 2 out of 8 trials, spherical sketches included outlier cells (black circles) for which inferred pseudotime and actual collection time disagree. At the same time, cells in final state collected at time point 72 are lost in each trial.

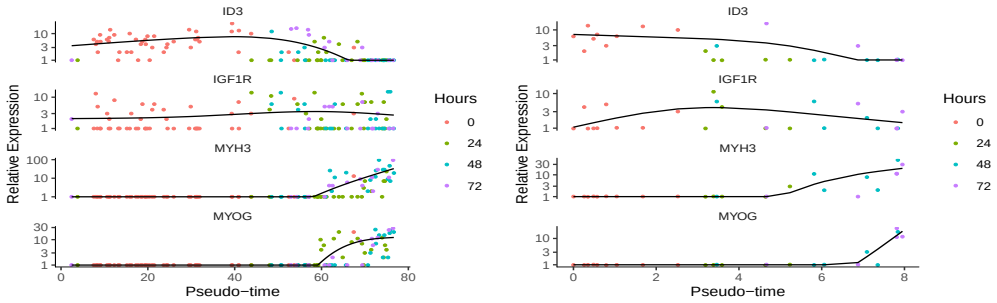


Figure S8. Gene expression dynamics, related to Figure 5. Expression dynamics along pseudotime were computed by Monocle 2 from full data (left) and from the sketch produced by Sphetcher with fairness constraints (right) for genes ID3, IGF1R, MYH3, and MYOG.

Transparent Methods

Sketching scRNA-seq as k-center problem

Given a large scRNA-seq dataset, we seek to select a subset of cells, a so-called *sketch* (Hie et al. 2019), that evenly represents the geometry of the transcriptional space occupied by the original data. As originally proposed in Hie et al. (2019), we use the *Hausdorff distance* to measure how well the sketch captures the transcriptional heterogeneity in the data. Given n data points $X = \{x_1, x_2, \ldots, x_n\}$ representing the m-dimensional gene expression measurements $x_i \in \mathbb{R}^m$ of n individual cells, and a metric d that measures the dissimilarity between pairs of cells, the Hausdorff distances between a sketch $X_S \subseteq X$ and the full dataset is given by:

$$d_H(X_S, X) = \max_{x \in X} \left\{ \min_{y \in X_S} d(x, y) \right\} \tag{1}$$

A sketch achieves a small Hausdorff distance if it includes for every cell in the original dataset a cell that is close to it in gene expression space. Finding a best sketch of size k, i.e. a sketch that minimizes the Hausdorff distance is known as the metric k-center problem in the combinatorial optimization literature. It is known to be NP-hard, but a solution with Hausdorff distance at most 2 times the optimal distance can be found by a simple greedy strategy: In each iteration, pick the point farthest away from the current set of centers and add it as a new center. Although this greedy approach has time complexity $O(nk)$, it does not scale efficiently to large scRNA-seq datasets that require a larger number of cells k to be accurately represented.

A thresholding algorithm

To find a sketch of size k with small Hausdorff distance (1) to a single-cell dataset, we employ the *thresholding* technique that was originally proposed for the design of approximation algorithms for bottleneck problems (Hochbaum & Shmoys 1986). In essence, we are guessing the optimal distance in (1) and for every guess L try to find a feasible solution, that is, a subset of cells of cardinality at most k such that spheres of radius L centered at cells in the subset cover all remaining cells. Then the smallest L^* for which such a feasible sketch exists denotes the optimal solution. We model the problem of finding the smallest set of cells such that the maximal distances from any other cell to the subset is at most a given threshold L as a set cover problem, $\textsc{SetCover}_X(L)$: Given a universe $\mathcal{U} = X$ of n data points, we build a collection $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ of n subsets of $\mathcal{U}$ by including in each set S_i all points that lie within distance of L from x_i, i.e. $S_i = \{x_j \mid d(x_i, x_j) \leq L\}$. Then the minimum number of sets in $\mathcal{S}$ that cover every element of the universe corresponds to a smallest subset of points covering all remaining points with spheres of radius L.

A widely used algorithm for the set cover problem is based on a greedy strategy (Johnson 1974): Starting from an empty set, in each iteration pick the set in $\mathcal{S}$ that covers the largest number of elements yet uncovered and add it to the solution. The greedy algorithm is guaranteed to find a cover which is within a logarithmic factor of the optimal solution (Johnson 1974). Moreover, it has been observed across a wide range of instances that the greedy algorithm produces solutions close to the optimum. A direct implementation of the greedy algorithm, however, scales poorly to large scRNA-seq datasets. We therefore employ the disk-friendly greedy (DFG) algorithm developed in Cormode et al. (2010) for very large datasets. It achieves a dramatic performance improvement over the standard greedy algorithm by applying a geometric scale bucketing approximation. Furthermore, the DFG algorithm runs in linear time with respect to the total size of candidate sets, i.e in $O(\sum_i |S_i|)$, while guaranteeing to output a set cover which is within a logarithmic factor of the optimum. More precisely, the algorithm allows to choose a parameter p that represents a trade-off

between the running time (which is $O((1+\frac{1}{p-1})\sum_i |S_i|)$) and the approximation ratio (which is $1+p\ln n$). The complete algorithm is summarized in Algorithm 1. Let us denote by $\textsc{Greedy}(L)$ the set cover returned by the greedy algorithm when applied to sets $S_i = \{x_j \mid d(x_i, x_j) \leq L\}$, and let $\widetilde{L}(k) := \min\{L \mid \textsc{Greedy}(L) \text{ has size at most } k\}$ which can be found by a logarithmic number of calls to the greedy algorithm via binary search: If $\textsc{Greedy}(L)$ is at most k, we decrease the threshold, otherwise we increase it (halving the length of the search interval in both cases), until the radius L lies in an interval of size at most ε.

Algorithm 1: Sphetcher

Input: Dataset $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^m$, size of the sketch k, and precision ε.
Initialization: $L_{\min} = 0$, $L_{\max} = \max_{i,j} d(x_i, x_j)$.
while $L_{max} - L_{min} > \varepsilon$ **do**
  $L \leftarrow (L_{\min} + L_{\max})/2$
  Solve $\text{SETCOVER}_X(L)$ using the DFG algorithm.
  if $|\textsc{Greedy}(L)| \leq k$ **then**
    $L_{\max} \leftarrow L$
  else
    $L_{\min} \leftarrow L$
  end
end
Output: $X_S = \{x_i | S_i \in \textsc{Greedy}(L)\}$.

If we are willing to increase the size of X_S by a logarithmic factor, Algorithm 1 is guaranteed to return a sketch with optimal Hausdorff distance.

Theorem 1. *Let L^* be the optimal distance in (1) for $|X_S| = k$. If we run the thresholding approach for $|X_S| = k\ln(n)$, then the solution we obtain has Hausdorff distance at most L^*. In other words, $\widetilde{L}(k\ln(n)) \leq L^*$.*

Proof. By definition of L^*, $\textsc{SetCover}_X(L^*)$ has size at most k. Thus, by the known approximation factor of the greedy algorithm, $\textsc{Greedy}(L^*)$ has size at most $k\ln(n)$, which implies by the definition of $\widetilde{L}(k\ln(n))$ that $\widetilde{L}(k\ln(n)) \leq L^*$. □

Grid sampling with guarantees

For datasets much larger than 100,000 cells, we apply a hybrid strategy to reduce the computational cost of determining the neighborhood of each point in Algorithm 1. To this end, we divide the space into equal-sized boxes from which we pick one point at random. In contrast to geometric sketching, we do not attempt to optimally define boxes in each dimension, but leave it to the subsequent thresholding algorithm to properly cover the space by spheres. In fact, we show that if we carefully choose the applied threshold taking into account the size of the grid, our hybrid sampling strategy increases the Hausdorff distance by at most a factor of $(1+\varepsilon)$, where $\varepsilon > 0$ controls the size of the grid.

Let $\textsc{SetCover}_X(L, Z)$ denote an optimal set covering all the points in X with spheres of radius L whose centers are chosen from $Z \subseteq X$. Let $\textsc{Greedy}(L, Z)$ denote the set obtained by the greedy algorithm described above covering all the points in X with spheres of radius L whose centers are chosen from $Z \subseteq X$. We know that $|\textsc{Greedy}(L, Z)| \leq |\textsc{SetCover}_X(L, Z)|\ln(n)$, where $n = |X|$. Let $L_{\min}$ be the minimum distance between two points in X and $L_{\max}$ be the maximum distance

between two points in X. Let I be the smallest integer such that $(1+\varepsilon)^I L_{\min} \geq L_{\max}$. Our hybrid algorithm that carefully combines grid sampling with the thresholding approach is given in Algorithm 2 (Sphetcher-H).

Algorithm 2: Sphetcher-H

1 **Input**: Dataset $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^m$, size of the sketch k, and $\varepsilon > 0$.
2 **Initialization**: $L_{\min} = \min_{i,j} d(x_i, x_j)$, an integer I as defined before.
3 **for** i = *0,...,* I **do**
4 $\quad L \leftarrow (1+\varepsilon)^i L_{\min}$
5 $\quad$ Partition the space into a uniform grid G(L) of size $\varepsilon L/\sqrt{m}$
6 $\quad$ Let $Z(L) \subseteq X$ be the set obtained by choosing one point in each non-empty cell
7 $\quad Y(L) \leftarrow \textsc{Greedy}\big(\big(1+(1+\varepsilon)\varepsilon\big)L, Z(L)\big)$
8 **end**
9 **Output**: $Y(\hat{L}(k))$, where $\hat{L}(k) = \min\{L : |Y(L)| \leq k\}$.

The following theorem limits the increase in Hausdorff distance through Sphetcher-H by at most a factor of $(1+\varepsilon)$.

Theorem 2. *Let L^* be the Hausdorff distance $d_H(X_S, X)$ between X and an optimal set X_S of size k, then $d_H(Y(\hat{L}(k\ln(n))), X) \leq (1+\varepsilon)L^*$.*

Proof. Let L be the distance set in the for loop (Algorithm 2: steps 3 to 7) such that $L^* \leq L < (1+\varepsilon)L^*$. By definition of L^*, we know that $|\textsc{SetCover}_X(L^*, X)| \leq k$. So, let us write $\textsc{SetCover}_X(L^*, X) = X_S := \{x_1, ..., x_k\}$. Let $X'_S = \{x'_1, ..., x'_k\} \subseteq Z(L)$ be chosen such that x'_i lies in the same cell of the grid $G(L)$ as x_i. Hence, $d_H(x_i, x'_i) \leq \varepsilon L$ implies that

$$d_H(X_S, X'_S) \leq \varepsilon L < (1+\varepsilon)\varepsilon L^*.$$

Thus for any point $x \in X$, we have

$$d_H(x, X'_S) \leq d_H(x, X_S) + d_H(X_S, X'_S) \leq (1+(1+\varepsilon)\varepsilon)L^* \leq (1+(1+\varepsilon)\varepsilon)L.$$

It follows that $|\textsc{SetCover}_X((1+(1+\varepsilon)\varepsilon)L, Z(L))| \leq k$ and hence,

$$|\textsc{Greedy}((1+(1+\varepsilon)\varepsilon)L, Z(L))| \leq k\ln(n),$$

that is, $|Y(L)| \leq k\ln(n)$. By definition of $\hat{L}$, we have $\hat{L} \leq L < (1+\varepsilon)L^*$. □

Fair sampling

One of the advantages of our model is its flexibility to incorporate fairness aspects. For example, assume we have prior knowledge of (some) of the cell types present in the sample. Cells might have been pre-sorted, and some cell types such as T cell subtypes are well characterized and can be identified based on known markers, without relying on an unsupervised clustering of the data. Furthermore, when reusing scRNA-seq datasets shared through repositories or data archives, the annotation of cell types, i.e. their labels, are typically provided as part of the original study. Similarly, in time series studies of gene expression, cells are collected at different time points which can supervise the sketching algorithm to preferentially select cells for which collection time point and transcriptomic state agree.

Our goal is to use prior categorical information on, e.g., biological cell types or collection time point to guide the selection of cells into a representative sketch, without fully relying on the correctness of cell type labels nor their synchronous progression through biological processes. We incorporate prior categorical information as *covering constraints* into our model: We seek to select a subset of cells that represent the geometric space of the original data according to (1) but at the same time contain at least a given number of representatives from each class. More formally, let $X_1, X_2, \ldots, X_m \subseteq X$ denote known clusters that do not necessarily partition the whole dataset X, we want to sample k cells that contain at least $l_i \in \mathbb{N}^+$ cells from each X_i, for all $i = 1, 2, \ldots, m$, while minimizing the Hausdorff distance of the sketch to the original dataset. This generalization of the k-center problem is similar to the *colorful k-center* problem, which does not require to include class members into the sketch but instead a certain number of elements from each class need to be covered by spheres around selected centers. For the colorful k-center problem a constant approximation in the Euclidean plane was recently introduced (Bandyapadhyay et al. 2019). In Anegg et al. (2020), the authors study a variant of this problem in which classes are allowed to overlap. Neither of the proposed algorithms is directly applicable to scRNA-seq data, due to low-dimensionality assumptions or the use of the ellipsoid method, respectively.

If $l_i = 1$, for all $i = 1, \ldots, m$, we have hitting set constraints $X_S \cap X_i \neq \emptyset$, $i = 1, \ldots, m$, which can be modeled as m additional elements in the universe of our set cover formulation of the problem. Given a threshold L, the corresponding set cover problem $(\mathcal{U}, \mathcal{S})$ is $\mathcal{U} = \{x_1, \ldots, x_n, X_1, \ldots, X_m\}$ and $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ with $S_i = [x_i] \cup \{X_j \mid x_i \in X_j\}$. Here $[x_i]$ contains x_i and its neighbors within distance L. Picking a set S_i into our set cover solution now does not only cover all cells within distance L of x_i, but x_i also hits all clusters $\{X_j \mid x_i \in X_j\}$. Having cast the constrained sampling problem as an instance of our thresholding framework, we solve it by the same algorithm (Algorithm 1). For general $l_i \in \mathbb{N}^+$, we simply partition X_i into l_i parts and apply the above approach, which however is no longer guaranteed to obtain the optimal Hausdorff distance.

Set cover under perturbation

This section provides the theoretical insight for the practical performance of the greedy set cover approach and its robustness to noise present in, e.g., scRNA-seq data. In step 5 of Algorithm 1 we need to construct the neighborhood for every point x_i that contains all points within a given distance threshold. Due to noise, the true distances will be slightly perturbed and yield imprecise estimates of neighborhoods. Since an instance to our set cover formulation contains a set for the neighborhood of each point, error-prone neighborhoods will affect our (greedy) search for the set with the largest number of uncovered elements. Here, we show that as long as we are able to pick a set with large enough number of uncovered elements, we can essentially preserve the approximation guarantee. More precisely, denote by C_t the set of elements covered *after* t iterations of the greedy search ($C_0 = \emptyset$). Assume that in each iteration t, errors in the distances prevent us from finding the set S_t^* with the maximum value of $|S_i \backslash C_{t-1}|$, but instead we select a set S_t such that $E(|S_t \backslash C_{t-1}|) \geq c \max_i |S_i \backslash C_{t-1}|$ for some constant c, where $E(X)$ denotes the expected value of random variable X. We show that with high probability, we will find a set cover within $2\ln(n)/c$ the size of an optimal solution, which differs only by a constant factor from the approximation guarantee of the (precise) greedy algorithm. Note that inapproximability results (Slavík 1997) show that the greedy algorithm is essentially the best-possible polynomial time approximation algorithm for set cover up to lower order terms. Let $\mathcal{U}$ be the whole set of elements of size n. We have the following theorem.

Theorem 3. *If an iterative algorithm always chooses a set S_t to add to the current solution with*

$$E(|S_t \backslash C_{t-1}| \ \big| C_{t-1}) \geq c \max_i |S_i \backslash C_{t-1}|,$$

for $c \leq 1$, then with (high) probability $1 - \frac{1}{n}$ it returns a set cover that is larger than the optimum set cover by a factor of at most $2\ln(n)/c$.

Proof. Let the number of sets in the optimal solution be σ. We know that at each iteration there is some set that covers at least $|\mathcal{U} \backslash C_t|/\sigma$ new elements. It follows that

$$E(|\mathcal{U} \backslash C_{t+1}| \ \big| C_t) = |\mathcal{U} \backslash C_t| - E(|S_{t+1} \backslash C_t| \ \big| C_t) \leq |\mathcal{U} \backslash C_t| - c \max_i |S_i \backslash C_t| \leq \Big(1 - \frac{c}{\sigma}\Big)|\mathcal{U} \backslash C_t|.$$

Now taking the expectation over all possibilities for C_t we get

$$E(|\mathcal{U} \backslash C_{t+1}|) \leq \Big(1 - \frac{c}{\sigma}\Big) E(|\mathcal{U} \backslash C_t|),$$

and iterating we end up with

$$E(|\mathcal{U} \backslash C_t|) \leq |\mathcal{U}| \Big(1 - \frac{c}{\delta}\Big)^t \leq n e^{-tc/\sigma}.$$

Setting $t = 2\sigma \ln(n)/c$ implies that $E(|\mathcal{U} \backslash C_t|) \leq \frac{1}{n}$, and hence by Markov's Inequality:

$$\Pr(|\mathcal{U} \backslash C_t| \geq 1) \leq E(|\mathcal{U} \backslash C_t|) \leq \frac{1}{n}.$$

Thus, with probability at least $1 - \frac{1}{n}$, the sets we selected form a set cover. □

Benchmarks

Sphetcher

We have implemented Algorithms 1 and 2 along with a fair sampling option in software tool Sphetcher in C++. We applied our hybrid strategy Sphetcher-H (Algorithm 2) on datasets exceeding 200,000 cells, which included datasets zeiselCNS, saunders, cao as well as the umbilical cord blood cells dataset. Unless stated otherwise, Sphetcher uses Pearson correlation as distance metric d, and we set the precision $\varepsilon = 10^{-4}$ in Algorithm 1. Note that throughout this work, the size of our spherical sketch denotes the actual number of cells rather than their logarithmic approximation in Theorem 1.

Data and evaluation

All data were uniformly preprocessed by natural log-transformation of gene counts (after adding a pseudo-count of 1) followed by projection to 100 principle components.

We measure how well a sketch represents the original transcriptomic space by the robust Hausdorff distance. Compared to the classical definition of the Hausdorff distance, the robust variant of the distance between a sketch $X_S \subseteq X$ and the full dataset is less sensitive to outliers (Huttenlocher et al. 1993):

$$d_{HK}(X_S, X) = K^{th}_{x \in X} \Big\{ \min_{y \in X_S} d(x, y) \Big\}, \tag{2}$$

where $K^{th}_{x \in X}$ denotes the Kth largest distance to an element in X. Consistent with Hie et al. (2019), we set $K = \lceil 1\text{e}{-4} \times |X| \rceil$ in our experiments.

Marker genes inflammatory macrophages

We computed AUROC for marker genes reported in Hie et al. (2019) separating inflammatory macrophages from remaining macrophages using the Python package provided with the original publication at `https://github.com/brianhie/geosketch`.

References

Anegg, G., Angelidakis, H., Kurpisz, A. & Zenklusen, R. (2020), A technique for obtaining true approximations for k-center with covering constraints, *in* 'Integer Programming and Combinatorial Optimization', Springer International Publishing.

Bandyapadhyay, S., Inamdar, T., Pai, S. & Varadarajan, K. R. (2019), 'A constant approximation for colorful k-center', *CoRR* **abs/1907.08906**.

Cormode, G., Karloff, H. & Wirth, A. (2010), 'Set cover algorithms for very large datasets', *CIKM* pp. 479 – 488.

Hie, B., Cho, H., DeMeo, B., Bryson, B. & Berger, B. (2019), 'Geometric sketching compactly summarizes the single-cell transcriptomic landscape', *Cell Syst.* **8**(6), 483 – 493.e7.

Hochbaum, D. S. & Shmoys, D. B. (1986), 'A unified approach to approximation algorithms for bottleneck problems', *J. ACM* **33**(3), 533 – 550.

Huttenlocher, D. P., Klanderman, G. A. & Rucklidge, W. J. (1993), 'Comparing images using the hausdorff distance', *IEEE Trans. Pattern Anal. Mach. Intell.* **15**(9), 850–863.

Johnson, D. S. (1974), 'Approximation algorithms for combinatorial problems', *J. Comput. Syst. Sci.* **9**(3), 256 – 278.

Slavík, P. (1997), 'A tight analysis of the greedy algorithm for set cover', *Journal of Algorithms* **25**(2), 237 – 254.