**Supplementary Materials**

scValue: value-based subsampling of large-scale single-cell transcriptomic data for machine and deep learning tasks

Li Huang, Weikang Gong, Dongsheng Chen\*

# Supplementary Note 1: Summary of evaluations and case studies

As summarised in Table 1 and illustrated in Fig. 1C, this study evaluated scValue alongside the six existing sketching methods (Uniform, GeoSketch, Sphetcher, Hopper, KH, and scSampler) using the six large scRNA-seq datasets in the ML/DL tasks of cell-type annotation, label transfer, cell-type harmonisation, and (pseudo) bulk RNA-seq deconvolution. Then these methods were further assessed with the 16 datasets via sketch metric comparison.

In all experiments, each method was applied with its default parameters to generate sketches of each dataset. The tree-partitioning variant of Hopper (TreeHopper) was chosen for its balance between computation time and sketch quality, and scSampler was run with B = 4 partitions to achieve a comparable trade-off.

**Cell-type annotation**

The cell-type annotation tasks were conducted in scenarios where the reference and query data employed consistent cell-type annotation schemes. We replicated four previously studied dataset-model pairs:

1. the human PBMC dataset [1] with the variational autoencoder-based scANVI model [2], originally examined in [3];
2. the mBrain dataset [4] with the variational autoencoder-based scPoli model [5], investigated in the latter’s study;
3. the CxG\_min dataset [6] with the logistic regression-based CellTypist model [7], evaluated in the former’s study;
4. the mACA dataset [8] with the neural networks-based ACTINN model [9], assessed in [10].

For each pair, we used the same dataset partitions (reference and query) described in the corresponding studies. The reference (training) partition was then used to create sketches at 2%, 4%, 6%, 8%, and 10% of the original size via each of the seven subsampling methods, while the query (test) partition remained at its full size. Each resulting sketch was used to train the respective model, which then annotated cell types in the query partition. To ensure robustness, we repeated these steps 10 times with varying random seeds for each sketch, recording the annotation accuracy for each experiment. Moreover, we recorded the accuracy obtained by using the full reference partition for each task (also 10 repeated runs), providing a baseline for assessing the performance drop associated with the different subsampling methods. Further details on partitioning and model training are available in Supplementary Note 2.

**Label transfer**

In this case study, we applied the various sketching methods to cross-dataset cell-type transfer learning and cross-dataset cell-type harmonisation, respectively. In the former task, the reference and query data used different annotation conventions. Specifically, we replicated the “large-scale cross-dataset label transfer” tutorial from CellTypist [7], using 10% sketches of the human Gut dataset [11] to annotate a colon query dataset [12] containing 42k cells in full. T-related cells were transferred from each sketch to the query, and we then assessed how closely these transferred annotations matched those obtained when using the full Gut dataset as the reference.

**Cell-type harmonisation**

The goal of this case study was to standardise the cell-type annotation styles across multiple datasets originating from different studies. Following the harmonisation workflow of CellHint [13], we used the human Spleen dataset [7], which encompasses cells from four independent studies. Each subsampling method was applied to sketch 10% of Spleen. CellHint then used these sketches to learn relationships among T cell-related types across the four studies. The performance of cell-type harmonisation was evaluated by comparing the learned relationships from each sketch against those derived from the full Spleen dataset.

**Deconvolution**

We evaluated scValue for building a single-cell expression reference for MuSiC [14], which is a reference-based deconvolution tool identified as the top performer in a recent systematic benchmark study [15]. The experiment used the T&ILC dataset, comprising 216,611 cells from 12 human donor samples. Specifically, the expression matrices of cells from two donors (A29 and A31) served as the full reference dataset, from which sketches were generated using scValue as well as six baseline methods (Uniform, GeoSketch, Sphetcher, Hopper, KH, and scSampler). The expression profiles from the remaining 10 samples (A36, A35, A37, A52, 582C, 621B, 637C, and 640C) were aggregated to construct pseudobulk data with known ground truth cell type proportions, allowing for direct evaluation of MuSiC’s accuracy. Both the full dataset and the generated sketches were then used by MuSiC to infer cell type proportions from the pseudobulk data.

Following the approach described by Wang et al. [14], we assessed performance using three metrics: the average Pearson correlation, average root mean squared error (RMSE), and average mean absolute error (MAE) between the inferred and true proportions across the 10 samples.

**Sketch metric comparison**

Since most previous studies on existing sketching methods have focused on computation time and Hausdorff distance [16-19], and the most recent work on scSampler [19] highlights Gini coefficient, we evaluated all three metrics in the sketch metric comparison experiment, using sketches generated by each subsampling method across the 16 datasets (Table 1 and Fig. 1C).

Computation time was recorded as the elapsed time (in seconds) to produce each sketch, while Gini coefficient quantified the imbalance in cell-type proportions: lower values indicate a more balanced distribution, and higher values signal greater inequality. This metric is formally defined in [19] as



where  represents the proportion of cell type  in the full dataset. Consistent with scSampler, we utilised the implementation of Gini coefficient computation available at https://github.com/oliviaguest/gini. Hausdorff distance, denoted by , is defined as the maximum Euclidean distance () between any point in the original dataset and the nearest point in 



Smaller  means that all cells in  are well-represented in , implying a higher degree of similarity between the original dataset and the subsample. Conversely, largersuggests less similarity.

# Supplementary Note 2: Experimental setup of the cell-type annotation tasks

scValue was benchmarked against six existing subsampling (sketching) methods, i.e., Uniform, GeoSketch, Sphetcher, Hopper, KH, and scSampler, across four cell-type annotation tasks that paired previously studied single-cell RNA sequencing (scRNA-seq) datasets with various machine or deep learning models. Below, we detail the experimental setup, including data information, reference-query splitting, sketching procedures, and training parameters.

**PBMC with scANVI**

We used the Peripheral Blood Mononuclear Cells (PBMC) dataset [1], which comprises 31,021 cells across 10 cell types with 2,000 highly variable genes (HVGs). This dataset was used to training the variational autoencoder-based scANVI model [2] in the previous study [3]. Following this study, we split PBMC into:

* a reference (training) set of 20,303 cells from SeqWell, DropSeq, and Indrop sequencing protocols.
* a query (test) set of 16,941 cells from the 10X V2 protocol.

Each of the seven sketching methods was applied to generate sketches of the reference set at varying percentages of 2%, 4%, 6%, 8%, and 10%, respectively, while the query set remained unchanged. Each resulting sketch was used to train a scANVI model following the official tutorial (https://docs.scarches.org/en/latest/scanvi\_surgery

\_pipeline.html). Here, raw counts were used for training, as recommended by the official documentation, to ensure the model’s underlying statistical assumptions remain valid. To mitigate overfitting, we set vae.train() to 100 epochs, scanvae.train() to 10 epochs, and model.train() to 10. All other parameters were kept at their default values. The trained model was then used to infer cell types for the full query set. A separate model was trained on the entire reference set to quantify performance loss due to sketching. Each training-and-annotation run was repeated 10 times (random seeds 42 to 51), and the annotation accuracy was recorded for each run.

**mBrain with scPoli**

The second task employed the mBrain dataset [4], which contains 56,399 cells across 10 cell types and 2,000 HVGs. This dataset includes cells from four independent studies and has been investigated in conjection with the variational autoencoder-based scPoli model [5]. We split mBrain into:

* a reference set: Saunders et al. (34,502 cells) and Zeisel et al. (7,394 cells)
* a query set: Schaum et al. (7,856 cells) and Rosenberg et al. (6,647 cells)

All seven sketching methods were applied to the reference set at 2%, 4%, 6%, 8%, and 10% of its size, with the query set left unchanged. For each sketch, we trained scPoli following the official tutorial (https://docs.scarches.org/en/latest/scpoli\_surgery

\_pipeline.html). Again, raw counts were used for training, as recommended by the official documentation, to avoid conflicts with the model’s underlying statistical assumptions. When training the reference scPoli model, to prevent overfitting on smaller sketches, we set n\_epochs to 5 and pretraining\_epochs to 4 for 2-6% sketches; then we set n\_epochs to 10 and pretraining\_epochs to 8 for 8% and 10% sketches. When training the reference mapping scPoli model for the full query set, we set n\_epochs to 10 and pretraining\_epochs to 8. The learning rate was set to 0.01, leaving other parameters at defaults. A model trained on the full reference set (with n\_epochs=10 and pretraining\_epochs=8 for both the reference and mapping models) served as a baseline for evaluating the performance drop due to sketching. All experiments were repeated 10 times (random seeds 42 to 51), with annotation accuracies recorded for each run.

**CxG\_min with CellTypist**

The third task concerned the logistic regression-based CellTypist model [7] and the CxG\_min dataset [6]. Downloaded from the GitHub page of the study “scTab: Scaling cross-tissue single-cell annotation models” [6], this dataset is a minimal subset of the 22.2 million human CELLxGENE census data. It already includes training and test partitions, drawn from 70% and 15% of the human donors, respectively. Each partition comprises 32,768 cells across 164 cell types with 19331 protein-coding genes. We selected 3,000 HVGs to compute the top 50 principal components as input features for the sketching methods.

Again, we produced sketches of the reference set at 2%, 4%, 6%, 8%, and 10%, keeping the test set unchanged. Each sketch was used to train a CellTypist model according to the official tutorial (https://colab.research.google.com/github/Teichlab/cel

ltypist/blob/main/docs/notebook/celltypist\_tutorial\_cv.ipynb). Following the tutorial’s recommendation, we enabled feature selection during model training and left all parameters at defaults. A model trained on the full reference set served as the baseline for performance. Each experiment was repeated 10 times (random seeds 42 to 51), and the annotation accuracy was documented for each run.

**mACA with ACTINN**

The final task paired the neural networks-based ACTINN model [9] and the mACA dataset [8] consisting of 356k cells across 197 cell types with 20,116 genes. This model-dataset combination has been explored in the previous study [10]. Following this, we first filtered out cells with missing (“nan”) cell type labels and then split the remaining cells based on timepoints into

* a reference (training) set: 43,196 cells from mouse samples aged 18 months
* a query (test) set: 140,743 cells obtained from mouse samples aged 1, 3, 21, 24, or 30 months

The reference set was subsampled at 2%, 4%, 6%, 8%, and 10% with each of the seven sketching methods, while the query set remained intact. We trained ACTINN on each sketch using the default parameters (50 epochs, learning rate of 0.0001, minibatch size of 128). A model trained on the full reference set was also included for baseline comparison. All experiments were repeated 10 times (random seeds 42 to 51), and the resulting annotation accuracies were recorded.

# Supplementary Note 3: The approach for building core sample sets

We have developed a systematic approach for constructing core sample sets that addresses the propagation of low-confidence cell type labels by leveraging both confidence score computation and subsequent filtering. This method is designed to ensure that only high-quality annotations contribute to downstream analyses, thereby mitigating the risk of inaccuracies. The code is available at: https://github.com/LHBCB/scvalue/tree/main/core\_sample\_sets.

**Step 1: calculation of confidence scores**

The main objective is to assign a confidence score, ranging from 0 to 1, to each cell’s annotation within an scRNA-seq dataset. To accomplish this, we employ the *pred\_by\_celltypist* function, which calculates these confidence scores using one of two approaches: the function can use an established CellTypist model obtained from the official repository (https://www.celltypist.org/models) or a custom CellTypist model that has been trained with high-quality data by the user. The model should be provided in pickle format, which is used by *pred\_by\_celltypist* to output the predicted cell types for individual cells along with their corresponding confidence scores.

**Step 2: validation of predicted cell types**

The predictions made by the CellTypist model are compared directly with the original cell-type annotations. Only the cells whose predicted cell types exactly match the original annotations or belong to the same broader category are retained. Unassigned or unknown cells may also be included if desired. It should be noted that this validation step is optional; if omitted, every cell is considered a candidate for inclusion in the core sample set, regardless of prediction accuracy.

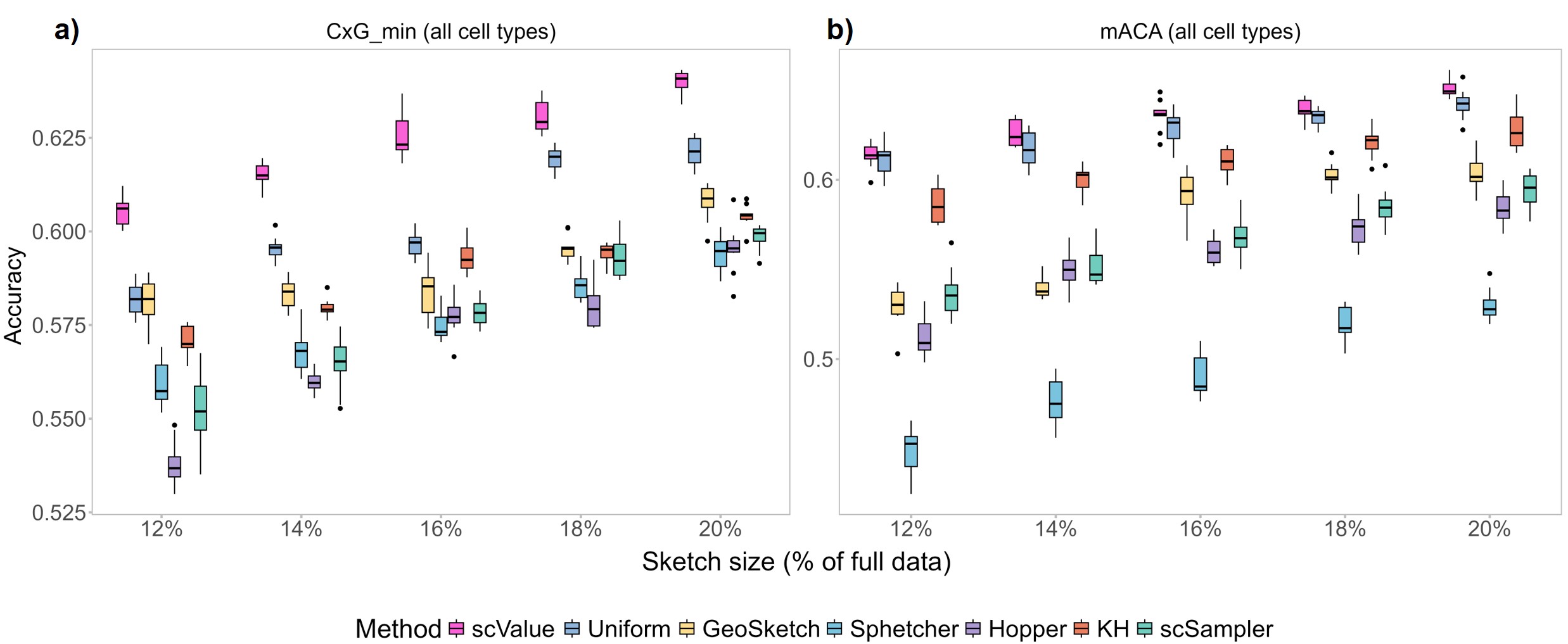
**Step 3: construction of the core sample set**

The core sample set is constructed by filtering cells based on their confidence scores. This is achieved by using the *build\_core\_sample\_set* function, which applies a user-specified threshold (by default 0.5) to the confidence scores. If Step 2 is performed, only cells with confidence scores exceeding the threshold and that meet the validation criteria are included in the core sample set.

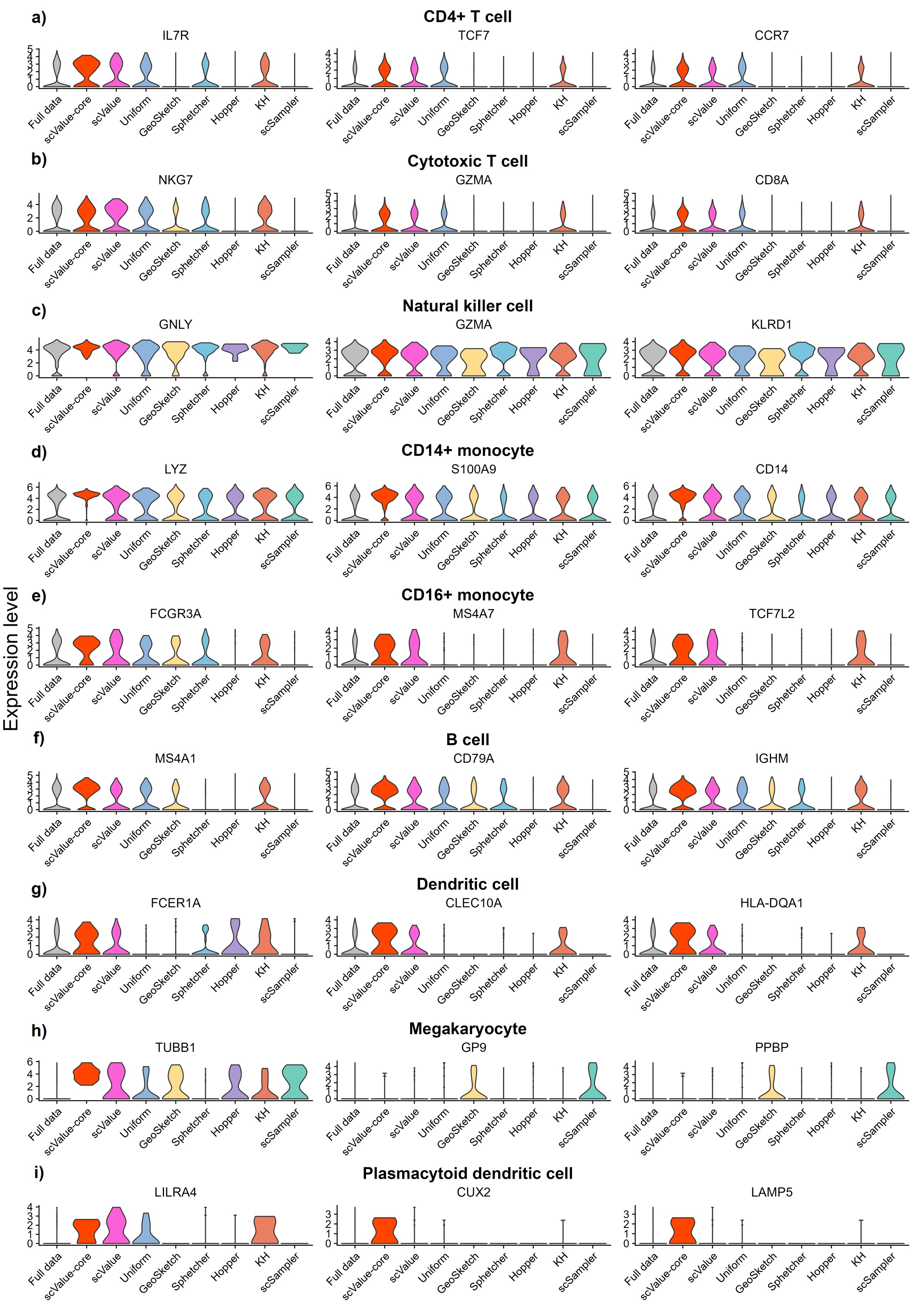
**Step 4: scValue subsampling on the core sample set**

scValue subsampling is performed on the constructed core sample set to generate a high-quality downsized dataset (sketch). This process is expected to enhance downstream analytical performance in both machine learning and deep learning applications, as well as improve cell-type specific marker expression analysis compared to the standard scValue sketch, i.e., produced from the entire dataset without filtering.

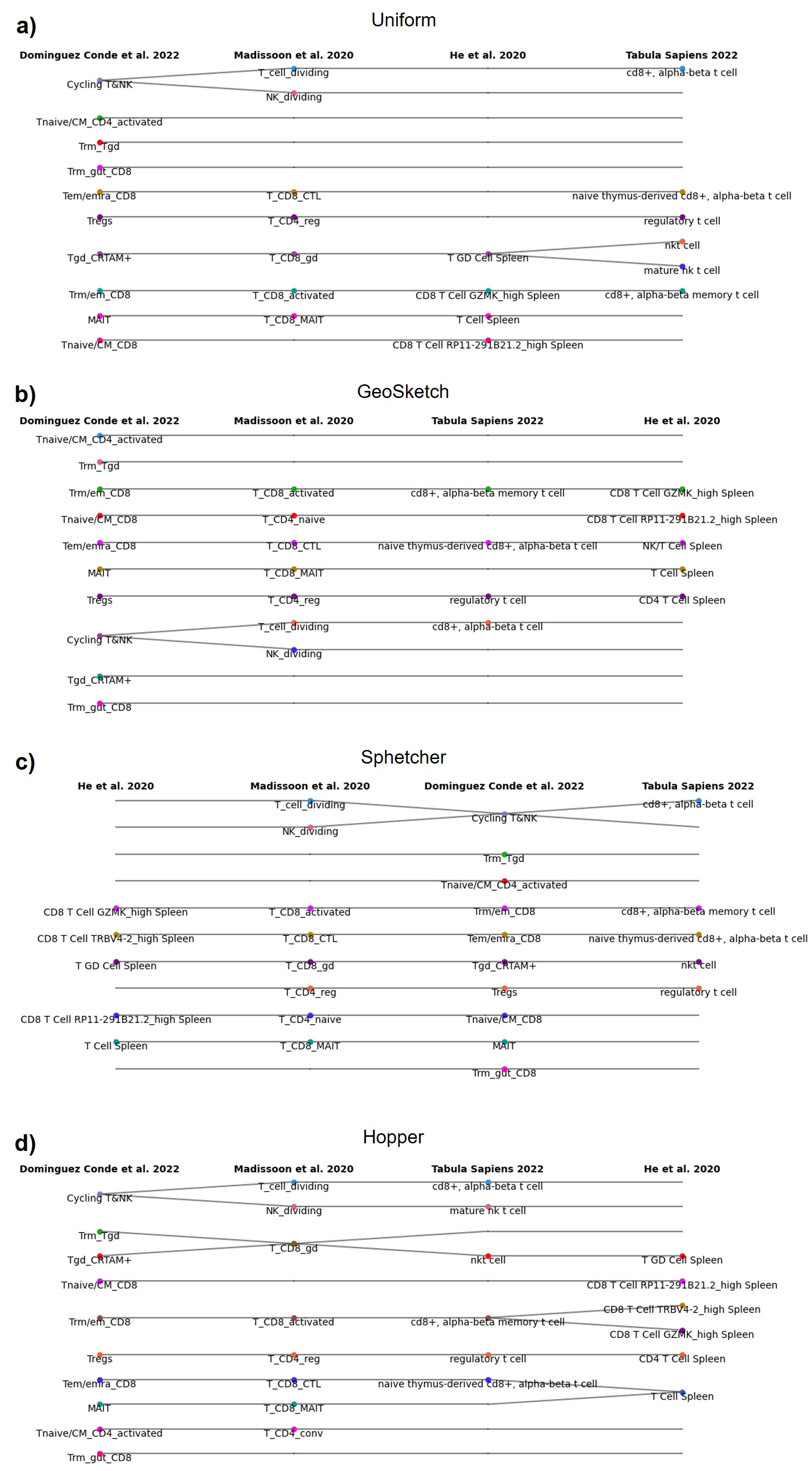
# Supplementary Figures

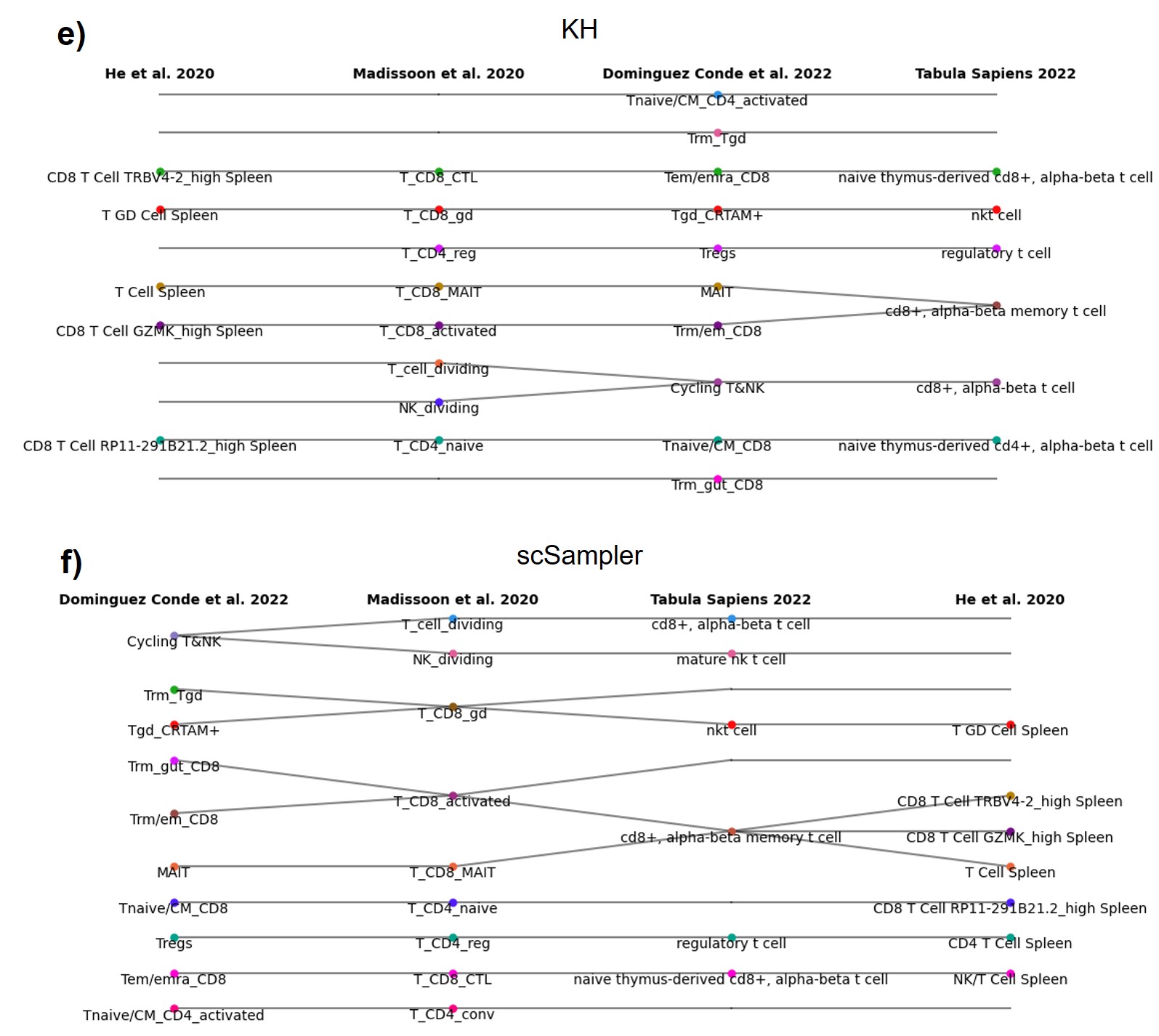


**Figure S1**. Performance comparison of scValue (MTB strategy) and the six baseline methods in cell-type annotation tasks at sketch sizes of 12%-20%.

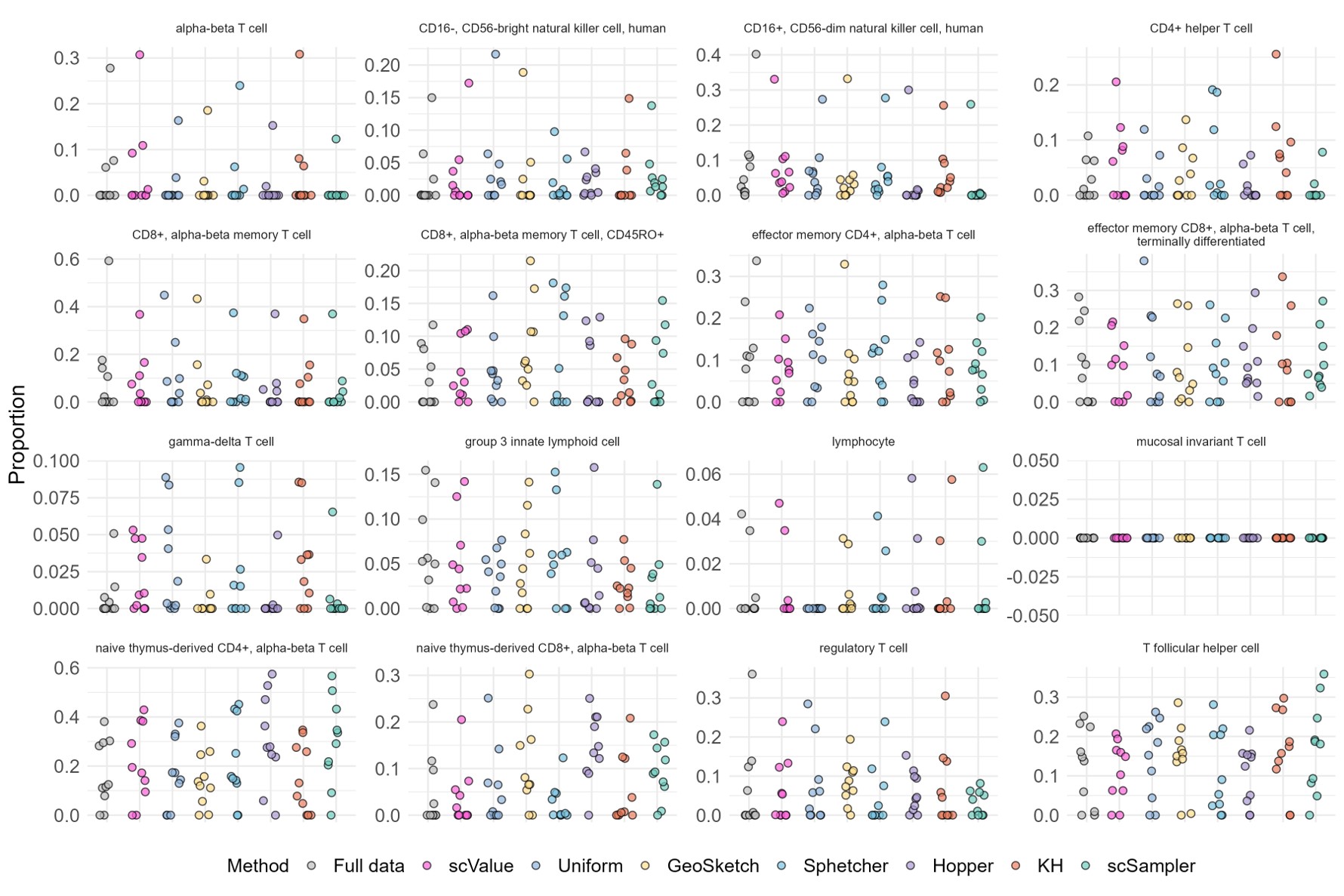


**Figure S2**. Cell-type-specific marker expression levels of full data and sketches by scValue-core, scValue, and the six baseline methods.

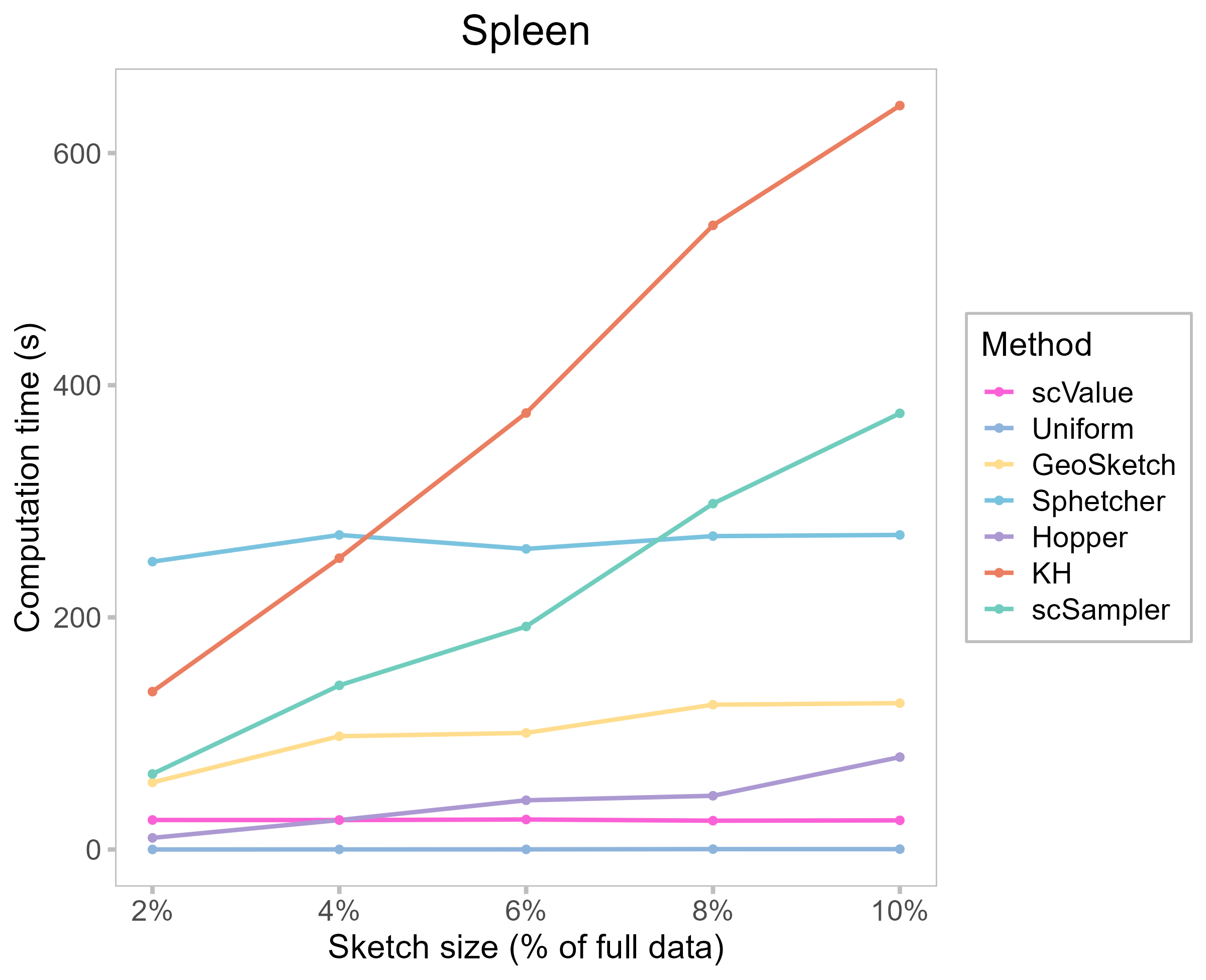




**Figure S3**. Cell type harmonisation results of sketches by the six baseline methods.



**Figure S4.** Sample-level MuSic deconvolution results of full data and the sketches by scValue and the six baseline methods.



**Figure S5.** The Spleen dataset was used as an example to illustrate how computation time of each sketching methods changes as the sketch size increases from 2% to 10%. scValue’s computational time is nearly independent of the sketch size, which resonates with *O*(*BdN*log*N*) complexity.

# Supplementary Tables

**Table S1.** The scANVI model evaluated on sketches of the PBMC dataset, to make predictions for all cell types; the original query annotation ACC of the model trained on the full reference data was 0.8635 ± 0.0050.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Sketch Percentage** | | | | |
| **2%** | **4%** | **6%** | **8%** | **10%** |
| scValue | **0.7658 ± 0.0148** | **0.8029 ± 0.0160** | **0.8182 ± 0.0085** | **0.8297 ± 0.0086** | **0.8330 ± 0.0103** |
| Uniform | 0.7575 ± 0.0179 | 0.7742 ± 0.0116 | 0.7664 ± 0.0118 | 0.7846 ± 0.0082 | 0.7769 ± 0.0112 |
| GeoSketch | 0.4722 ± 0.0613 | 0.5742 ± 0.0566 | 0.7111 ± 0.0125 | 0.7039 ± 0.0197 | 0.7312 ± 0.0160 |
| Sphetcher | 0.6874 ± 0.0285 | 0.7156 ± 0.0201 | 0.7228 ± 0.0120 | 0.7413 ± 0.0244 | 0.7385 ± 0.0205 |
| Hopper | 0.5494 ± 0.0732 | 0.5939 ± 0.0466 | 0.6112 ± 0.0337 | 0.6456 ± 0.0240 | 0.6170 ± 0.0232 |
| KH | 0.7572 ± 0.0174 | 0.7686 ± 0.0161 | 0.7923 ± 0.0137 | 0.7674 ± 0.0106 | 0.7999 ± 0.0140 |
| scSampler | 0.5855 ± 0.0475 | 0.5665 ± 0.0225 | 0.6006 ± 0.0250 | 0.5902 ± 0.0222 | 0.6065 ± 0.0190 |

**Table S2.** The scPoli model evaluated on sketches of the mBrain dataset, to make predictions for all cell types; the original query annotation ACC of the model trained on the full reference data was 0.9320 ± 0.0047.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Sketch Percentage** | | | | |
| **2%** | **4%** | **6%** | **8%** | **10%** |
| scValue | **0.9033 ± 0.0122** | **0.9169 ± 0.0026** | 0.9199 ± 0.0045 | 0.9318 ± 0.0058 | **0.9328 ± 0.0073** |
| GeoSketch | 0.6325 ± 0.1072 | 0.8558 ± 0.1113 | 0.8940 ± 0.0811 | 0.9316 ± 0.0027 | 0.9296 ± 0.0026 |
| Hopper | 0.8426 ± 0.0573 | 0.9164 ± 0.0147 | 0.7489 ± 0.1099 | **0.9338 ± 0.0046** | 0.9008 ± 0.0768 |
| KH | 0.7930 ± 0.0891 | 0.7588 ± 0.1153 | 0.7908 ± 0.1184 | 0.9247 ± 0.0032 | 0.9230 ± 0.0039 |
| Sphetcher | 0.5456 ± 0.0643 | 0.6848 ± 0.1223 | 0.5724 ± 0.0174 | 0.6542 ± 0.0676 | 0.6722 ± 0.0847 |
| Uniform | 0.6883 ± 0.0654 | 0.7154 ± 0.1026 | 0.9152 ± 0.0018 | 0.9209 ± 0.003 | 0.7840 ± 0.1297 |
| scSampler | 0.5986 ± 0.0908 | 0.7355 ± 0.1322 | **0.9213 ± 0.0135** | 0.9120 ± 0.0375 | 0.8314 ± 0.1303 |

**Table S3.** The CellTypist model evaluated on sketches of the CxG\_min dataset, to make predictions for all cell types; the original query annotation ACC of the model trained on the full reference data was 0.6847 ± 0.0045.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Sketch Percentage** | | | | |
| **2%** | **4%** | **6%** | **8%** | **10%** |
| scValue (MTB) | **0.4597 ± 0.0067** | **0.5316 ± 0.0065** | **0.5470 ± 0.0044** | **0.5663 ± 0.0049** | **0.5921 ± 0.0049** |
| scValue (FB) | 0.4551 ± 0.013 | 0.5120 ± 0.0045 | 0.5394 ± 0.003 | 0.5636 ± 0.0045 | 0.5811 ± 0.0052 |
| GeoSketch | 0.3722 ± 0.0117 | 0.4404 ± 0.0102 | 0.4599 ± 0.0067 | 0.4938 ± 0.0048 | 0.5465 ± 0.0066 |
| Hopper | 0.3703 ± 0.0115 | 0.4380 ± 0.0082 | 0.4775 ± 0.0084 | 0.5045 ± 0.0085 | 0.5245 ± 0.0069 |
| KH | 0.4363 ± 0.0078 | 0.4792 ± 0.0061 | 0.5140 ± 0.0046 | 0.5301 ± 0.0019 | 0.5484 ± 0.0044 |
| Sphetcher | 0.4148 ± 0.0081 | 0.4616 ± 0.0080 | 0.4858 ± 0.0056 | 0.5083 ± 0.0055 | 0.5414 ± 0.0068 |
| Uniform | 0.4171 ± 0.0051 | 0.4928 ± 0.0043 | 0.5132 ± 0.0037 | 0.5401 ± 0.0050 | 0.5615 ± 0.0027 |
| scSampler | 0.3823 ± 0.0124 | 0.4454 ± 0.0066 | 0.4698 ± 0.0066 | 0.4949 ± 0.0061 | 0.5371 ± 0.0071 |

**Table S4.** The ACTINN model evaluated on sketches of the mACA dataset, to make predictions for all cell types; the original query annotation ACC of the model trained on the full reference data was 0.7176±0.0046.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Methods** | **Sketch Percentage** | | | | |
| **2%** | **4%** | **6%** | **8%** | **10%** |
| scValue (MTB) | **0.3617±0.0145** | **0.4874±0.0186** | **0.5469±0.0099** | **0.5742±0.0066** | **0.5946±0.0091** |
| scValue (FB) | 0.3512±0.0205 | 0.4823±0.0102 | 0.5366±0.0091 | 0.5656±0.0108 | 0.5901±0.0098 |
| Uniform | 0.3407±0.0103 | 0.4667±0.0159 | 0.5284±0.0084 | 0.5548±0.0102 | 0.5871±0.0129 |
| GeoSketch | 0.2664±0.0149 | 0.3827±0.0077 | 0.4369±0.0173 | 0.4645±0.0079 | 0.4973±0.0130 |
| Sphetcher | 0.1725±0.0142 | 0.2941±0.0109 | 0.3454±0.0125 | 0.3857±0.0169 | 0.4203±0.0100 |
| Hopper | 0.2416±0.0152 | 0.3904±0.0105 | 0.4608±0.0097 | 0.4813±0.0126 | 0.5107±0.0108 |
| KH | 0.3324±0.0135 | 0.4298±0.0142 | 0.4976±0.0118 | 0.5447±0.0058 | 0.5582±0.0140 |
| scSampler | 0.2821±0.0246 | 0.4011±0.0108 | 0.4572±0.0143 | 0.4859±0.0067 | 0.5215±0.0117 |

**Table S5.** The scANVI model evaluated on sketches of the PBMC dataset, to make predictions for only rare cell types; the original query annotation ACC of the model trained on the full reference data was 0.6224 ± 0.0728.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Sketch Percentage** | | | | |
| **2%** | **4%** | **6%** | **8%** | **10%** |
| scValue | **0.2079 ± 0.1854** | **0.4698 ± 0.0938** | **0.6133 ± 0.0508** | **0.6796 ± 0.0674** | **0.6855 ± 0.0480** |
| Uniform | 0.0133 ± 0.0336 | 0.0693 ± 0.0452 | 0.0850 ± 0.0935 | 0.1489 ± 0.1239 | 0.2853 ± 0.1031 |
| GeoSketch | 0.0978 ± 0.0290 | 0.0322 ± 0.0309 | 0.0747 ± 0.0350 | 0.0993 ± 0.0612 | 0.4066 ± 0.1747 |
| Sphetcher | 0.0064 ± 0.0137 | 0.0130 ± 0.0158 | 0.0501 ± 0.0643 | 0.2676 ± 0.1841 | 0.2066 ± 0.1710 |
| Hopper | 0.0044 ± 0.0075 | 0.0138 ± 0.0159 | 0.0042 ± 0.0058 | 0.0111 ± 0.0124 | 0.0346 ± 0.0396 |
| KH | 0.0017 ± 0.0054 | 0.0548 ± 0.0844 | 0.2130 ± 0.1630 | 0.1995 ± 0.1521 | 0.4167 ± 0.0404 |
| scSampler | 0.0000 ± 0.0000 | 0.0071 ± 0.0096 | 0.0086 ± 0.0100 | 0.0125 ± 0.0177 | 0.0165 ± 0.0213 |

**Table S6.** The scPoli model evaluated on sketches of the mBrain dataset, to make predictions for only rare cell types; the original query annotation ACC of the model trained on the full reference data was 0.6896 ± 0.0176.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Sketch Percentage** | | | | |
| **2%** | **4%** | **6%** | **8%** | **10%** |
| scValue | **0.5840 ± 0.0274** | **0.6431 ± 0.0240** | 0.6509 ± 0.0154 | 0.7085 ± 0.0316 | **0.7207 ± 0.0311** |
| Uniform | 0.5409 ± 0.0621 | 0.6314 ± 0.0075 | 0.6290 ± 0.0052 | 0.6605 ± 0.0193 | 0.7039 ± 0.0278 |
| GeoSketch | 0.4796 ± 0.0856 | 0.6314 ± 0.0200 | 0.6668 ± 0.0239 | **0.7026 ± 0.0196** | 0.6998 ± 0.0289 |
| Sphetcher | 0.3625 ± 0.0577 | 0.4977 ± 0.0928 | 0.4657 ± 0.1231 | 0.5851 ± 0.1048 | 0.6601 ± 0.0631 |
| Hopper | 0.4230 ± 0.0702 | 0.6548 ± 0.0251 | 0.6685 ± 0.0265 | 0.6934 ± 0.0220 | 0.7047 ± 0.0182 |
| KH | 0.5569 ± 0.0314 | 0.6370 ± 0.0117 | 0.6471 ± 0.0072 | 0.6757 ± 0.0126 | 0.6643 ± 0.0186 |
| scSampler | 0.4221 ± 0.0619 | 0.6043 ± 0.1296 | **0.6778 ± 0.0340** | 0.7158 ± 0.0165 | 0.7134 ± 0.0294 |

**Table S7.** The CellTypist model evaluated on sketches of the CxG\_min dataset, to make predictions for only rare cell types; the original query annotation ACC of the model trained on the full reference data was 0.4045 ± 0.0059.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Sketch Percentage** | | | | |
| **2%** | **4%** | **6%** | **8%** | **10%** |
| scValue | **0.1926 ± 0.0092** | **0.2603 ± 0.0112** | **0.2967 ± 0.0076** | **0.3114 ± 0.0068** | **0.3478 ± 0.0076** |
| Uniform | 0.1052 ± 0.0089 | 0.1499 ± 0.0129 | 0.1742 ± 0.0070 | 0.1827 ± 0.0110 | 0.2110 ± 0.0060 |
| GeoSketch | 0.1192 ± 0.0080 | 0.1587 ± 0.0068 | 0.1832 ± 0.0096 | 0.2002 ± 0.0048 | 0.2557 ± 0.0084 |
| Sphetcher | 0.1327 ± 0.0096 | 0.1664 ± 0.0078 | 0.1987 ± 0.0071 | 0.2147 ± 0.0077 | 0.2253 ± 0.0103 |
| Hopper | 0.1384 ± 0.0118 | 0.1825 ± 0.0086 | 0.2231 ± 0.0079 | 0.2117 ± 0.0071 | 0.2446 ± 0.0067 |
| KH | 0.1273 ± 0.0099 | 0.1305 ± 0.0083 | 0.1441 ± 0.0085 | 0.1828 ± 0.0079 | 0.1991 ± 0.0056 |
| scSampler | 0.1610 ± 0.0049 | 0.1874 ± 0.0099 | 0.2178 ± 0.0052 | 0.2487 ± 0.0052 | 0.2755 ± 0.0082 |

**Table S8.** The ACTINN model evaluated on sketches of the mACA dataset, to make predictions for only rare cell types; the original query annotation ACC of the model trained on the full reference data was 0.6733 ± 0.0038.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Methods** | **Sketch Percentage** | | | | |
| **2%** | **4%** | **6%** | **8%** | **10%** |
| scValue | **0.2287 ± 0.0180** | **0.3579 ± 0.0192** | **0.4258 ± 0.0116** | **0.4742 ± 0.0080** | **0.5040 ± 0.0078** |
| Uniform | 0.1200 ± 0.0245 | 0.2319 ± 0.0251 | 0.3214 ± 0.0188 | 0.3695 ± 0.0126 | 0.4170 ± 0.0124 |
| GeoSketch | 0.1950 ± 0.0064 | 0.2787 ± 0.0105 | 0.3574 ± 0.0086 | 0.3832 ± 0.0159 | 0.4102 ± 0.0119 |
| Sphetcher | 0.0824 ± 0.0146 | 0.1472 ± 0.0202 | 0.2094 ± 0.0100 | 0.2503 ± 0.0102 | 0.3022 ± 0.0128 |
| Hopper | 0.1997 ± 0.0161 | 0.3111 ± 0.0088 | 0.3648 ± 0.0131 | 0.4108 ± 0.0101 | 0.4360 ± 0.0139 |
| KH | 0.1303 ± 0.0103 | 0.2254 ± 0.0157 | 0.3115 ± 0.0112 | 0.3716 ± 0.0182 | 0.3980 ± 0.0128 |
| scSampler | 0.2125 ± 0.0130 | 0.2973 ± 0.0110 | 0.3604 ± 0.0146 | 0.4054 ± 0.0114 | 0.4399 ± 0.0176 |

**Table S9.** The CellTypist model evaluated on increased sketches (12%-20%) of the CxG\_min dataset, to make predictions for all cell types; the original query annotation ACC of the model trained on the full reference data was 0.6847 ± 0.0045.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Method** | **Sketch Percentage** | | | | |
| **12%** | **14%** | **16%** | **18%** | **20%** |
| scValue | **0.6053 ± 0.0038** | **0.6151 ± 0.0031** | **0.6256 ± 0.006** | **0.6306 ± 0.0042** | **0.6401 ± 0.0028** |
| GeoSketch | 0.5818 ± 0.0043 | 0.5956 ± 0.003 | 0.5967 ± 0.0036 | 0.6193 ± 0.0032 | 0.6215 ± 0.0038 |
| Hopper | 0.5816 ± 0.0060 | 0.5835 ± 0.0039 | 0.5838 ± 0.0064 | 0.5955 ± 0.0033 | 0.6077 ± 0.0048 |
| KH | 0.5595 ± 0.0062 | 0.5680 ± 0.0058 | 0.5750 ± 0.0042 | 0.5856 ± 0.0039 | 0.5941 ± 0.0045 |
| Sphetcher | 0.5377 ± 0.0061 | 0.5598 ± 0.0029 | 0.5771 ± 0.0050 | 0.5801 ± 0.0060 | 0.5954 ± 0.0066 |
| Uniform | 0.5709 ± 0.0038 | 0.5796 ± 0.0024 | 0.5929 ± 0.0042 | 0.5942 ± 0.0026 | 0.6041 ± 0.0030 |
| scSampler | 0.5522 ± 0.0095 | 0.5650 ± 0.0074 | 0.5784 ± 0.0035 | 0.5929 ± 0.0053 | 0.5983 ± 0.0034 |

**Table S10.** The ACTINN model evaluated on increased sketches (12%-20%) of the mACA dataset, to make predictions for all cell types; the original query annotation ACC of the model trained on the full reference data was 0.7176 ± 0.0046.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Methods** | **Sketch Percentage** | | | | |
| **2%** | **4%** | **6%** | **8%** | **10%** |
| scValue (MTB) | **0.6136 ± 0.0069** | **0.6261 ± 0.0075** | **0.6362 ± 0.0083** | **0.6385 ± 0.0067** | **0.6511 ± 0.0047** |
| Uniform | 0.6114 ± 0.0091 | 0.6172 ± 0.0096 | 0.6293 ± 0.0091 | 0.6348 ± 0.0052 | 0.6424 ± 0.0082 |
| GeoSketch | 0.5296 ± 0.0115 | 0.5400 ± 0.0062 | 0.5915 ± 0.0132 | 0.6025 ± 0.0065 | 0.6035 ± 0.0098 |
| Sphetcher | 0.4495 ± 0.0130 | 0.4759 ± 0.0130 | 0.4901 ± 0.0121 | 0.5189 ± 0.0105 | 0.5298 ± 0.0087 |
| Hopper | 0.5128 ± 0.0109 | 0.5489 ± 0.0108 | 0.5602 ± 0.0071 | 0.5727 ± 0.0106 | 0.5849 ± 0.0096 |
| KH | 0.5860 ± 0.0107 | 0.6001 ± 0.0077 | 0.6102 ± 0.0075 | 0.6208 ± 0.0081 | 0.6287 ± 0.0118 |
| scSampler | 0.5370 ± 0.0137 | 0.5518 ± 0.0112 | 0.5682 ± 0.0108 | 0.5850 ± 0.0108 | 0.5944 ± 0.0096 |

**Table S11.** The scANVI model evaluated on sketches of the PBMC dataset, with scValue-core included, to make predictions for all cell types; the original query annotation ACC of the model trained on the full reference data was 0.8635 ± 0.0050.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Methods** | **Sketch Percentage** | | | | |
| **2%** | **4%** | **6%** | **8%** | **10%** |
| scValue-core | **0.7927 ± 0.0209** | **0.8309 ± 0.0126** | **0.8352 ± 0.0122** | **0.8468 ± 0.0063** | **0.8498 ± 0.0056** |
| scValue | 0.7658 ± 0.0148 | 0.8029 ± 0.0160 | 0.8182 ± 0.0085 | 0.8297 ± 0.0086 | 0.8330 ± 0.0103 |
| Uniform | 0.7575 ± 0.0179 | 0.7742 ± 0.0116 | 0.7664 ± 0.0118 | 0.7846 ± 0.0082 | 0.7769 ± 0.0112 |
| GeoSketch | 0.4722 ± 0.0613 | 0.5742 ± 0.0566 | 0.7111 ± 0.0125 | 0.7039 ± 0.0197 | 0.7312 ± 0.0160 |
| Sphetcher | 0.6874 ± 0.0285 | 0.7156 ± 0.0201 | 0.7228 ± 0.0120 | 0.7413 ± 0.0244 | 0.7385 ± 0.0205 |
| Hopper | 0.5494 ± 0.0732 | 0.5939 ± 0.0466 | 0.6112 ± 0.0337 | 0.6456 ± 0.0240 | 0.6170 ± 0.0232 |
| KH | 0.7572 ± 0.0174 | 0.7686 ± 0.0161 | 0.7923 ± 0.0137 | 0.7674 ± 0.0106 | 0.7999 ± 0.0140 |
| scSampler | 0.5855 ± 0.0475 | 0.5665 ± 0.0225 | 0.6006 ± 0.0250 | 0.5902 ± 0.0222 | 0.6065 ± 0.0190 |

**Table S12.** the scANVI model evaluated on sketches of the PBMC dataset, with scValue-core included, to make predictions for only rare cell types; the original query annotation ACC of the model trained on the full reference data is 0.8635 ± 0.0050.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Methods** | **Sketch Percentage** | | | | |
| **2%** | **4%** | **6%** | **8%** | **10%** |
| scValue-core | **scValue-core** | **0.2961 ± 0.1918** | **0.5971 ± 0.0403** | **0.6843 ± 0.0466** | **0.6990 ± 0.0385** |
| scValue | scValue | 0.2079 ± 0.1854 | 0.4698 ± 0.0938 | 0.6133 ± 0.0508 | 0.6796 ± 0.0674 |
| Uniform | Uniform | 0.0133 ± 0.0336 | 0.0693 ± 0.0452 | 0.0850 ± 0.0935 | 0.1489 ± 0.1239 |
| GeoSketch | GeoSketch | 0.0978 ± 0.0290 | 0.0322 ± 0.0309 | 0.0747 ± 0.0350 | 0.0993 ± 0.0612 |
| Sphetcher | Sphetcher | 0.0064 ± 0.0137 | 0.0130 ± 0.0158 | 0.0501 ± 0.0643 | 0.2676 ± 0.1841 |
| Hopper | Hopper | 0.0044 ± 0.0075 | 0.0138 ± 0.0159 | 0.0042 ± 0.0058 | 0.0111 ± 0.0124 |
| KH | KH | 0.0017 ± 0.0054 | 0.0548 ± 0.0844 | 0.2130 ± 0.1630 | 0.1995 ± 0.1521 |
| scSampler | scSampler | 0.0000 ± 0.0000 | 0.0071 ± 0.0096 | 0.0086 ± 0.0100 | 0.0125 ± 0.0177 |

**Table S13.** We compared the quality of 10% sketches created by scValue against those of the six existing subsampling methods using 16 datasets of varying size (31k-4m cells), focusing on computation time, Gini coefficient, and Hausdorff distance. For the mEmbryo dataset (containing ~3.2 million cells), Sphetcher and KH did not complete their runs and thus no metrices were available in the table; similarly, for the Fetal dataset (about four million cells), Hopper, Sphetcher, and KH did not finish running and were therefore excluded from the results. Hence, these methods received the largest ranks for the two datasets, when computing the average and standard deviation of ranks across all 16 datasets.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Method** | **Computation time** | **Computation time rank** | **Gini coefficient** | **Gini coefficient rank** | **Hausdorff distance** | **Hausdorff distance rank** |
| PBMC | scValue | 3.06 | 5 | 0.5242 | 1 | 16.0216 | 5 |
| Uniform | 0.03 | 1 | 0.5467 | 2 | 16.2002 | 6 |
| GeoSketch | 1.09 | 2 | 0.5525 | 4 | 14.3446 | 3 |
| Sphetcher | 55.00 | 7 | 0.5684 | 5 | 16.3879 | 7 |
| Hopper | 5.62 | 6 | 0.5694 | 6 | 9.9145 | 1 |
| KH | 3.02 | 4 | 0.5516 | 3 | 15.6156 | 4 |
| scSampler | 2.12 | 3 | 0.5726 | 7 | 10.2748 | 2 |
| mBrain | scValue | 5.20 | 2 | 0.6124 | 2 | 7.3199 | 5 |
| Uniform | 0.02 | 1 | 0.6657 | 3 | 7.3272 | 6 |
| GeoSketch | 8.10 | 4 | 0.5238 | 1 | 6.3921 | 3 |
| Sphetcher | 163.00 | 7 | 0.8244 | 7 | 7.1648 | 4 |
| Hopper | 9.48 | 5 | 0.6952 | 6 | 4.1282 | 1 |
| KH | 18.49 | 6 | 0.6759 | 5 | 7.3727 | 7 |
| scSampler | 6.32 | 3 | 0.6674 | 4 | 4.3788 | 2 |
| CxG\_min | scValue | 10.12 | 3 | 0.5977 | 1 | 29.6760 | 5 |
| Uniform | 1.14 | 1 | 0.7081 | 7 | 30.7035 | 6 |
| GeoSketch | 13.38 | 4 | 0.6806 | 4 | 21.8332 | 2 |
| Sphetcher | 236.00 | 7 | 0.6892 | 5 | 22.0492 | 3 |
| Hopper | 7.82 | 2 | 0.6524 | 2 | 13.1424 | 1 |
| KH | 30.45 | 6 | 0.6989 | 6 | 27.1067 | 4 |
| scSampler | 19.25 | 5 | 0.6575 | 3 | 56.0944 | 7 |
| mTC | scValue | 17.61 | 4 | 0.5247 | 3 | 6.0054 | 4 |
| Uniform | 0.06 | 1 | 0.5715 | 6 | 6.1491 | 6 |
| GeoSketch | 17.98 | 5 | 0.4920 | 1 | 5.9950 | 3 |
| Sphetcher | 130.00 | 7 | 0.6707 | 7 | 6.0932 | 5 |
| Hopper | 13.38 | 2 | 0.5295 | 4 | 5.0897 | 2 |
| KH | 100.74 | 6 | 0.5700 | 5 | 6.1493 | 7 |
| scSampler | 15.93 | 3 | 0.4991 | 2 | 5.0620 | 1 |
| cMTG | scValue | 11.10 | 2 | 0.7188 | 4 | 46.9527 | 4 |
| Uniform | 0.62 | 1 | 0.7530 | 6 | 51.9670 | 5.5 |
| GeoSketch | 25.99 | 4 | 0.6108 | 1 | 37.6251 | 2 |
| Sphetcher | 298.00 | 7 | 0.8993 | 7 | 44.1226 | 3 |
| Hopper | 21.93 | 3 | 0.6977 | 3 | 21.8112 | 1 |
| KH | 82.00 | 6 | 0.7528 | 5 | 57.1699 | 7 |
| scSampler | 56.74 | 5 | 0.6680 | 2 | 51.9670 | 5.5 |
| PAC | scValue | 13.40 | 2 | 0.6576 | 1 | 57.7370 | 4 |
| Uniform | 0.75 | 1 | 0.7200 | 4 | 62.1916 | 5.5 |
| GeoSketch | 30.52 | 4 | 0.6906 | 2 | 35.5558 | 2 |
| Sphetcher | 312.00 | 7 | 0.8781 | 7 | 44.0899 | 3 |
| Hopper | 18.10 | 3 | 0.7277 | 6 | 22.4604 | 1 |
| KH | 282.25 | 6 | 0.7209 | 5 | 65.0263 | 7 |
| scSampler | 87.84 | 5 | 0.7019 | 3 | 62.1916 | 5.5 |
| gMTG | scValue | 13.41 | 2 | 0.6573 | 3 | 56.2316 | 6 |
| Uniform | 0.74 | 1 | 0.7061 | 5 | 60.8895 | 7 |
| GeoSketch | 30.32 | 4 | 0.6101 | 1 | 35.0978 | 2 |
| Sphetcher | 287.00 | 7 | 0.8468 | 7 | 44.0018 | 3 |
| Hopper | 17.48 | 3 | 0.6721 | 4 | 21.4302 | 1 |
| KH | 180.33 | 6 | 0.7106 | 6 | 55.5025 | 5 |
| scSampler | 87.03 | 5 | 0.6411 | 2 | 54.2774 | 4 |
| Liver | scValue | 15.55 | 2 | 0.5774 | 1 | 16.5304 | 5 |
| Uniform | 0.14 | 1 | 0.6968 | 6 | 18.5627 | 6 |
| GeoSketch | 18.57 | 3 | 0.6225 | 2 | 14.6002 | 3 |
| Sphetcher | 292.00 | 6 | 0.8663 | 7 | 26.8399 | 7 |
| Hopper | 24.12 | 4 | 0.6942 | 5 | 8.4120 | 2 |
| KH | 394.54 | 7 | 0.6937 | 4 | 15.3512 | 4 |
| scSampler | 60.47 | 5 | 0.6366 | 3 | 8.2206 | 1 |
| GSCL | scValue | 22.27 | 2 | 0.3671 | 1 | 10.5512 | 6 |
| Uniform | 1.18 | 1 | 0.4475 | 6 | 10.8609 | 7 |
| GeoSketch | 42.98 | 4 | 0.3907 | 4 | 9.8900 | 3 |
| Sphetcher | 333.00 | 6 | 0.4548 | 7 | 10.0798 | 4 |
| Hopper | 32.36 | 3 | 0.3907 | 3 | 6.6982 | 1 |
| KH | 485.98 | 7 | 0.4465 | 5 | 10.3818 | 5 |
| scSampler | 69.95 | 5 | 0.3709 | 2 | 6.7168 | 2 |
| Spleen | scValue | 23.33 | 2 | 0.5717 | 1 | 31.3364 | 4 |
| Uniform | 0.13 | 1 | 0.6872 | 7 | 41.3595 | 7 |
| GeoSketch | 50.24 | 4 | 0.6594 | 3 | 16.5124 | 3 |
| Sphetcher | 333.00 | 6 | 0.6634 | 4 | 39.7043 | 5 |
| Hopper | 32.77 | 3 | 0.6316 | 2 | 10.1281 | 1 |
| KH | 524.91 | 7 | 0.6716 | 6 | 40.2324 | 6 |
| scSampler | 117.35 | 5 | 0.6661 | 5 | 10.2380 | 2 |
| T&ILC | scValue | 21.39 | 2 | 0.3345 | 4 | 14.0501 | 5 |
| Uniform | 0.51 | 1 | 0.3574 | 7 | 14.3303 | 6 |
| GeoSketch | 61.86 | 4 | 0.3499 | 5 | 13.2311 | 4 |
| Sphetcher | 326.00 | 6 | 0.2998 | 1 | 11.9971 | 3 |
| Hopper | 38.74 | 3 | 0.3065 | 2 | 9.2444 | 1 |
| KH | 625.03 | 7 | 0.3529 | 6 | 14.5102 | 7 |
| scSampler | 85.36 | 5 | 0.3234 | 3 | 9.2764 | 2 |
| mACA | scValue | 59.14 | 3 | 0.5598 | 1 | 43.7135 | 7 |
| Uniform | 1.33 | 1 | 0.7986 | 6 | 36.5662 | 5 |
| GeoSketch | 102.05 | 4 | 0.7864 | 3 | 18.4095 | 3 |
| Sphetcher | 330.00 | 5 | 0.7937 | 4 | 37.0538 | 6 |
| Hopper | 52.81 | 2 | 0.7947 | 5 | 8.1877 | 1 |
| KH | 1868.42 | 7 | 0.7833 | 2 | 36.0512 | 4 |
| scSampler | 556.32 | 6 | 0.8001 | 7 | 8.5561 | 2 |
| Gut | scValue | 57.60 | 2 | 0.5668 | 1 | 14.8138 | 7 |
| Uniform | 0.15 | 1 | 0.7281 | 7 | 12.9361 | 6 |
| GeoSketch | 212.32 | 4 | 0.6539 | 5 | 11.1547 | 3 |
| Sphetcher | 416.00 | 6 | 0.6248 | 2 | 11.6080 | 4 |
| Hopper | 95.71 | 3 | 0.6394 | 3 | 6.2233 | 1 |
| KH | 5707.71 | 7 | 0.7101 | 6 | 12.5748 | 5 |
| scSampler | 396.05 | 5 | 0.6398 | 4 | 6.3493 | 2 |
| mCNS | scValue | 279.89 | 2 | 0.5909 | 1 | 9.1424 | 6 |
| Uniform | 10.53 | 1 | 0.7216 | 4 | 8.7766 | 4 |
| GeoSketch | 1242.71 | 4 | 0.7149 | 3 | 8.2931 | 3 |
| Sphetcher | 2995.00 | 5 | 0.7094 | 2 | 8.9749 | 5 |
| Hopper | 295.62 | 3 | 0.7745 | 7 | 6.9452 | 2 |
| KH | 62784.22 | 7 | 0.7242 | 5 | 9.4050 | 7 |
| scSampler | 22172.88 | 6 | 0.7667 | 6 | 6.2520 | 1 |
| mEmbryo | scValue | 447.94 | 2 | 0.2892 | 1 | 8.0471 | 5 |
| Uniform | 12.56 | 1 | 0.3828 | 5 | 7.8783 | 4 |
| GeoSketch | 5945.76 | 4 | 0.2939 | 2 | 7.2228 | 3 |
| Hopper | 789.66 | 3 | 0.3585 | 4 | 5.9449 | 2 |
| scSampler | 70485.39 | 5 | 0.3517 | 3 | 5.5414 | 1 |
| Fetal | scValue | 1492.95 | 2 | 0.7198 | 1 | 4.4259 | 4 |
| Uniform | 8.26 | 1 | 0.8144 | 2 | 4.3989 | 3 |
| GeoSketch | 12323.61 | 4 | 0.8205 | 3 | 4.1255 | 1 |
| Sphetcher | 10501.00 | 3 | 0.8328 | 4 | 4.3333 | 2 |

# References：

1. Abdelaal T, Michielsen L, Cats D et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data, Genome Biol 2019;20:194.

2. Xu C, Lopez R, Mehlman E et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models, Mol Syst Biol 2021;17:e9620.

3. Du ZH, Hu WL, Li JQ et al. scPML: pathway-based multi-view learning for cell type annotation from single-cell RNA-seq data, Commun Biol 2023;6:1268.

4. Lotfollahi M, Naghipourfar M, Luecken MD et al. Mapping single-cell data to reference atlases by transfer learning, Nat Biotechnol 2022;40:121-130.

5. De Donno C, Hediyeh-Zadeh S, Moinfar AA et al. Population-level integration of single-cell datasets enables multi-scale analysis across samples, Nature Methods 2023;20:1683-1692.

6. Fischer F, Fischer DS, Mukhin R et al. scTab: scaling cross-tissue single-cell annotation models, Nature Communications 2024;15:6611.

7. Dominguez Conde C, Xu C, Jarvis LB et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans, Science 2022;376:eabl5197.

8. Tabula Muris C. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse, Nature 2020;583:590-595.

9. Ma F, Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing, Bioinformatics 2020;36:533-538.

10. Chen J, Xu H, Tao W et al. Transformer for one stop interpretable cell type annotation, Nat Commun 2023;14:223.

11. Elmentaite R, Kumasaka N, Roberts K et al. Cells of the human intestinal tract mapped across space and time, Nature 2021;597:250-255.

12. James KR, Gomes T, Elmentaite R et al. Distinct microbial and immune niches of the human colon, Nat Immunol 2020;21:343-353.

13. Xu C, Prete M, Webb S et al. Automatic cell-type harmonization and integration across Human Cell Atlas datasets, Cell 2023;186:5876-5891 e5820.

14. Wang X, Park J, Susztak K et al. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference, Nature Communications 2019;10:380.

15. Nguyen H, Nguyen H, Tran D et al. Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges, Nucleic Acids Research 2024;52:4761-4783.

16. Hie B, Cho H, DeMeo B et al. Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape, Cell Syst 2019;8:483-493 e487.

17. Do VH, Elbassioni K, Canzar S. Sphetcher: Spherical Thresholding Improves Sketching of Single-Cell Transcriptomic Heterogeneity, iScience 2020;23:101126.

18. DeMeo B, Berger B. Hopper: a mathematically optimal algorithm for sketching biological data, Bioinformatics 2020;36:i236-i241.

19. Song D, Xi NM, Li JJ et al. scSampler: fast diversity-preserving subsampling of large-scale single-cell transcriptomic data, Bioinformatics 2022;38:3126-3127.