OXFORD

Briefings in Bioinformatics, 2025, **26(3)**, bbaf279

https://doi.org/10.1093/bib/bbaf279

Problem Solving Protocol

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

Li Huang [1], Weikang Gong [1,2], Dongsheng Chen [1,*]

[1]State Key Laboratory of Common Mechanism Research for Major Diseases, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, 100 Chongwen Road, Suzhou Industrial Park, Suzhou, Jiangsu Province 215123, China

[2]Center for Artificial Intelligence and Computational Biology, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, 100 Chongwen Road, Suzhou Industrial Park, Suzhou, Jiangsu Province 215123, China

*Corresponding author. State Key Laboratory of Common Mechanism Research for Major Diseases, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, 100 Chongwen Road, Suzhou Industrial Park, Suzhou, Jiangsu Province 215123, China.
E-mail: cds@ism.pumc.edu.cn

Abstract

Large single-cell ribonucleic acid-sequencing (scRNA-seq) datasets offer unprecedented biological insights but present substantial computational challenges for visualization and analysis. While existing subsampling methods can enhance efficiency, they may not ensure optimal performance in downstream machine learning and deep learning (ML/DL) tasks. Here, we introduce scValue, a novel approach that ranks individual cells by 'data value' using out-of-bag estimates from a random forest model. scValue prioritizes high-value cells and allocates greater representation to cell types with higher variability in data value, effectively preserving key biological signals within subsamples. We benchmarked scValue on automatic cell-type annotation tasks across four large datasets, paired with distinct ML/DL models. Our method consistently outperformed existing subsampling methods, closely matching full-data performance across all annotation tasks. In three additional case studies—label transfer learning, cross-study label harmonization, and bulk RNA-seq deconvolution—scValue more effectively preserved T-cell annotations across human gut-colon datasets, more accurately reproduced T-cell subtype relationships in a human spleen dataset, and constructed a more reliable single-cell immune reference for cell-type deconvolution in simulated bulk tissue samples. Finally, using 16 public datasets ranging from tens of thousands to millions of cells, we evaluated subsampling quality based on computational time, Gini coefficient, and Hausdorff distance. scValue demonstrated fast execution, well-balanced cell-type representation, and distributional properties akin to uniform sampling. Overall, scValue provides a robust and scalable solution for subsampling large scRNA-seq data in ML/DL workflows. It is available as an open-source Python package installable via pip, with source code at https://github.com/LHBCB/scvalue.

Keywords: single-cell transcriptomics; data valuation; subsampling; cell type analysis; machine and deep learning

Introduction

Large-scale single-cell transcriptomic atlases, such as those from the Human Cell Atlas [1], CELLxGENE [2], SCAR [3], and SCAN [4], serve as valuable resources for smaller-scale studies [5]. However, visualizing and analysing these extensive datasets can present significant computational and memory challenges. To address this, researchers can create representative subsets, commonly referred to as 'sketches' that enable the extraction of meaningful insights in a more efficient and cost-effective manner [6].

Several sketching methods have been developed to date. The simplest, uniform subsampling is implemented in popular pipelines such as Seurat [7] and Scanpy [8] pipelines, but it may not capture the full transcriptional diversity of a dataset. To overcome this limitation, GeoSketch [6] partitions gene expression space into equal-sized, non-overlapping boxes and selects representative cells from each. Meanwhile, Sphetcher [9] covers cells by small-radius spheres that can better preserve the transcriptomic landscape. Both methods enable more evenly distributed sampling across diverse cell types. Hopper [10] provides theoretical guarantees for maintaining dataset structure, while its partition-based variant, TreeHopper, accelerates the sketching process without considerably compromising quality.

GeoSketch, Sphetcher, and (Tree)Hopper all utilize a minimax distance design [11], minimizing the Hausdorff distance, i.e. the maximum distance between any point in the original dataset and the nearest point in the sketch. In contrast, scSampler [12] adopts a maximin distance design [11], maximizing an inverse distance measure to ensure greater distinctiveness among selected cells. Finally, Kernel Herding (KH) constructs a 'stand-in' sketch [13] that focuses on preserving the original cell type distribution, rather than optimizing distance metrics.

Machine learning and deep learning (ML/DL) techniques are becoming increasingly essential in single-cell analysis [14]. However, relatively few empirical studies have evaluated how sketching methods perform in downstream learning tasks.

Received: February 6, 2025. **Revised:** April 16, 2025. **Accepted:** May 26, 2025

Table 1. Summary of the 16 scRNA-seq datasets used in this study

Dataset	# of cells	# of cell types	Tissue	Species	Literature	Usage
PBMC	31,021	10	Blood	*Homo sapiens*	PMID31500660	CTA; SMC
mBrain	56,399	10	Various cerebral regions	*Mus musculus*	PMID34462589	CTA; SMC
CxG_min	65,536	164	Cross-tissue	*H. sapiens*	PMID39098889	CTA; SMC
mTC	89,429	4	Embryo	*M. musculus*	PMID38355799	Demo; SMC
cMTG	112,929	18	Middle temporal gyrus	*Pan troglodytes*	PMID37824638	SMC
PAC	139,054	18	Primary auditory cortex	*H. sapiens*	PMID37824655	SMC
gMTG	139,945	18	Middle temporal gyrus	*Gorilla*	PMID37824638	SMC
Liver	167,598	14	Liver	*H. sapiens*	PMID35021063	SMC
GSCL	191,230	9	Gonad	*H. sapiens*	PMID35794482	SMC
Spleen	200,664	102	Spleen	*H. sapiens*	PMID35549406	H; SMC
T&ILC	216,611	16	Cross-tissue	*H. sapiens*	PMID35549406	Deconv; SMC
mACA	356,213	197	Cross-tissue	*M. musculus*	PMID32669714	CTA; SMC
Gut	428,469	134	Gut	*H. sapiens*	PMID34497389	LT; SMC
mCNS	2,106,206	15	Embryo	*M. musculus*	PMID38355799	SMC
mEmbryo	3,267,338	5	Embryo	*M. musculus*	PMID38355799	SMC
Fetal	4,062,980	72	Cross-tissue	*H. sapiens*	PMID33184181	SMC

Each dataset was utilized in one or more of the following experiments: demonstration (Demo), ML/DL-based CTA, LT learning, cell-type harmonization (H), deconvolution (Deconv), and/or SMC, assessed by computation time, Gini coefficient, and Hausdorff distance.

Moreover, assessing the ability of different methods to capture rare cell types—quantified by the Gini coefficient—is a critical consideration [12]. In this study, we introduce scValue, a subsampling method that creates sketches of single-cell ribonucleic acid sequencing (scRNA-seq) datasets by assigning a value to each cell. Cells with higher value (i.e. those more informative for cell-type identification) are more likely to be included in the sketch, whereas lower-value cells are less likely to be selected. Experimental results indicate that scValue consistently outperforms existing sketching approaches in ML/DL tasks, yielding more balanced cell-type proportions and exhibiting superior scalability as dataset sizes expand from tens of thousands to millions of cells.

Materials and methods

Large multi-species cross-tissue single-cell ribonucleic acid-sequencing datasets

As summarized in Table 1, this study includes 16 large-scale scRNA-seq datasets containing between 31 thousand and four million cells, spanning four to 197 cell types and covering samples from more than 10 tissues across four species. Each dataset was utilized in one or more of the following experiments: demonstration (Demo), ML/DL-based cell-type annotation (CTA), label transfer (LT) learning, cell-type harmonization (H), deconvolution (Deconv), and/or sketch metric comparison (SMC), which evaluated computation time, Gini coefficient, and Hausdorff distance.

Specifically, the mouse T cell (mTC) dataset was used to illustrate how well each sketching method's subsample reflects the structure of the full dataset (Demo). Four datasets were used for CTA: human peripheral blood mononuclear cells (PBMC), mouse brain (mBrain), CELLxGENE minimal (CxG_min), and the mouse Aging Cell Atlas (mACA). The Gut, Spleen, and T&ILC datasets were used for LT, H, and Deconv, respectively. All 16 datasets were included in the SMC analysis.

Six of the datasets (those used for CTA, LT, and H) were obtained from sources cited in their respective publications, while the remaining 10 were obtained from CELLxGENE (https://cellxgene.cziscience.com/datasets) [2]. PBMC was originally downloaded in Matrix Market format, comprising four separate files for barcodes, genes, expression values, and metadata; CxG_min was provided in Apache Parquet format; and the remaining 14 datasets were in HDF5 AnnData (h5ad) format.

All datasets underwent standard quality control either by the original authors or by CELLxGENE. Subsequently, we stored and processed all datasets in AnnData format using the Scanpy pipeline [8]. In the experiments, each of the 16 scRNA-seq datasets was normalized using a scale factor of 10 000 and log-transformed (unless already log-normalized in the original files). The top 3000 highly variable genes (HVGs) were selected unless HVGs were already provided or a different count was specified in the original publications. Lastly, the top 50 principal components (PCs) were computed from these HVGs and used as input for the subsampling methods.

Overview of scValue

scValue is designed to generate an informative sketch (representative subsample) of large scRNA-seq datasets while preserving critical biological diversity. As illustrated in Fig. 1a, scValue differs from conventional uniform or distance/distribution optimization-based methods. It first trains a random forest classifier to assign a 'data value' to each cell, reflecting its importance in distinguishing cell types. These data values then guide subsampling at the cell-type level: a value-weighted allocation determines the target sketch size for each cell type, ensuring that rare but informative cell types are adequately represented; cells with the highest values are selected using a default binning procedure, producing a smaller dataset that retains the most biologically vital cells.

Data value computation

For a large scRNA-seq dataset containing expression profiles for N cells and M cell types, we denote the input feature matrix $\mathbf{X} \in \mathbb{R}^{N\times d}$ and the input label vector $\mathbf{Y} \in \{1, 2, ..., M\}^N$, respectively. By default—and consistent with other sketching methods—d is set as the top number of PCs calculated from log-normalized gene counts. Figure 1a depicts the schematic workflow of scValue, which uses these inputs to create a value-based sketch consisting of S cells.

A random forest model [15] comprising B decision trees is fitted with $(\mathbf{X}, \mathbf{Y})$. The b-th tree, denoted by f_b, is trained on a bootstrap sample drawn with replacement from $(\mathbf{X}, \mathbf{Y})$. Cell i is considered

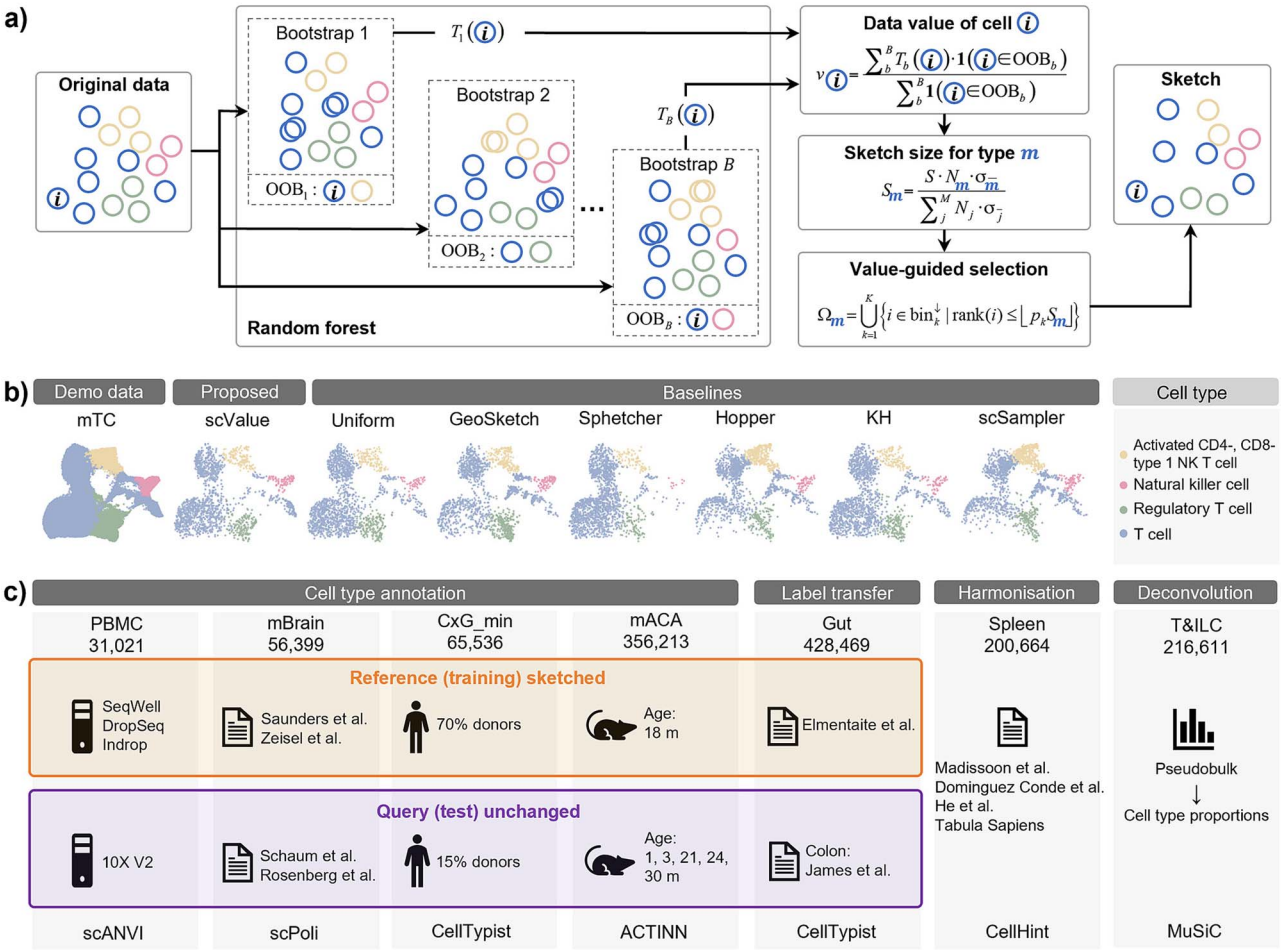

Figure 1. Overview of scValue and experiments. (a) Given a large scRNA-seq dataset, scValue generates an informative sketch (representative subsample) through three steps: first, a random forest classifier is trained on the full dataset using cell-type labels, and out-of-bag (OOB) estimates are computed as data values for individual cells; second, the sketch size for each cell type is determined, allocating greater representation to cell types with higher value variability; third, the value range is divided into equally sized bins, from which top-valued cells are selected to form the final subsample. (b) An mTC dataset illustrates scValue's ability to balance cell-type proportions and enhance cell-type separation, compared to six existing subsampling methods. (c) We systematically evaluated scValue against six counterparts through: four CTA tasks, each involving a previously studied pair of large scRNA-seq dataset and ML/DL model; a case study of LT learning across human gut-colon datasets via CellTypist; a second case study of cross-study label harmonization via CellHint using a human spleen dataset; and a third case study of psuedobulk RNA-seq deconvolution via MuSiC using a human T&ILC dataset.

out-of-bag (OOB) if it is not selected in the b-th bootstrap, which we denote by $i \in \mathbf{OOB}_b$. Using only OOB trees to predict for cell i and averaging the prediction errors yield the cell's OOB estimate. This OOB statistic is traditionally used to evaluate random forest performance [16] and has recently been adapted for data valuation in machine learning tasks [17]. In our study, each cell's data value ($v_i \in [0,1]$) is computed as the OOB accuracy of cell type prediction

$$v_i = \frac{\sum_b^B T_b(i) \cdot 1\left(i \in \mathbf{OOB}_b\right)}{\sum_b^B 1\left(i \in \mathbf{OOB}_b\right)} \quad (1)$$

where $1\left(i \in \mathbf{OOB}_b\right)$ equals 1 if cell i is an OOB sample for the b-th bootstrap and 0 otherwise; the correctness function $T_b(i)$ is given by $1\left(Y_i = f_b\left(\mathbf{X}_i\right)\right)$, which evaluates to 1 if the b-th tree correctly predicts the type of cell i and 0 otherwise. Cells of higher value are expected to bring greater benefits for distinguishing cell types when included in the sketch compared to those of lower value.

scValue defines each cell's data value as its OOB accuracy, computed solely from the trees in which the cell was excluded during training. In essence, a cell's OOB accuracy indicates how accurately the ensemble of trees (trained on all other cells) predicts its cell type. Although this evaluation inherently involves the contributions of other cells, it should be viewed as a strength rather than a drawback. A high OOB accuracy demonstrates that even when omitted from a particular tree's training, the cell's expression profile is well represented by the decision boundaries established by the remaining cells, reflecting its typicality within its cell type. Moreover, because the prediction is derived from the ensemble's collective information, cells with ambiguous or less characteristic features tend to yield lower OOB accuracies and, consequently, lower data values. This approach naturally prioritizes cells that are robustly supported by the overall dataset, thereby enhancing the quality of subsampling for downstream machine and deep learning tasks.

Sketch size determination

Sketching is conducted at the cell-type level in a value-weighted manner to improve cell-type balance in the resulting subset. The number of cells to be subsampled for cell type m is determined by considering both the cell type abundance and the standard error

of its data values

$$S_m = \frac{S \cdot N_m \cdot \sigma_{\overline{m}}}{\sum_j^M N_j \cdot \sigma_{\overline{j}}} \quad (2)$$

Here, N_m represents the number of cells belonging to type m in the original dataset; $\sigma_{\overline{m}}$ is given by $\sigma_m/\sqrt{N_m}$, in which σ_m denotes the standard deviation of data value for cells of type m. The denominator normalizes the value-weighted subsample size across M types. In addition, the largest remainder method [18] is employed to ensure that the total allocation for all cell types sums up exactly to S.

The rationale behind value-weighted subsampling is as follows: rare cell types are often more difficult to learn in a random forest due to their underrepresentation, while heterogeneous cell types that contain diverse or complex expression profiles also present challenges for learning. Both scenarios are typically associated with larger variability in data values. To address this, Equation (2) allocates more cells to rare and heterogenous cell types in the sketch, thereby enhancing the preservation of essential biological information.

Value-guided cell selection

Once the sketch size S_m for cell type m is determined, value-guided cell selection is performed. Depending on the number of cell types present in the original dataset, two cell selection strategies can be applied.

Full binning (FB): If the dataset contains a relatively small number of cell types (e.g. 10 types in the PBMC [19] and mBrain [20] datasets), the FB strategy can be employed to select high-value cells. For each cell type m with N_m cells, the data value range [0, 1] is divided into K equal-sized bins (by default, $K = 10$, corresponding to intervals of 0.1). Let $p_k$denote the proportion of N_m cells that fall into the k-th bin, such that $\sum_{k=1}^{K} p_k = 1$. The number of cells selected from each bin is proportional to p_k, ensuring that the subset preserves the original distribution of cells across bins. Specifically, within the k-th bin, let $\mathbf{bin}_k^{\downarrow}$ denote the set of cells ranked in descending order of their values v_i. Bin-wise selection is then performed by choosing the top $\lfloor p_k S_m \rfloor$ cells from $\mathbf{bin}_k^{\downarrow}$ to form part of the subsample (denoted Ω_m) for cell type m

$$\Omega_m = \bigcup_{k=1}^{K} \left\{ i \in \mathbf{bin}_k^{\downarrow} \mid \mathbf{rank}(i) \leq \lfloor p_k S_m \rfloor \right\} \quad (3)$$

This strategy captures cells across the entire data value ranges, preserving a broad snapshot of variability within each cell type m variation in value. Combining Ωs for each cell type results in the final sketching.

Mean-threshold binning (MTB): When the original dataset includes a large number of cell types (such as 164 types in the CxG_min dataset [21] or 197 types in the mACA dataset [22]), MTB can be employed. Unlike FB that uses uniform, equal-sized bins, MTB focuses on bins above the mean data value for cell type m; all bins below the mean are merged into a single bin starting from 0. This approach prioritizes cells with above-average values

The rationale for using different binning methods for cell selection lies in the structure of the dataset. In simpler datasets with fewer cell types, FB with uniform binning can effectively capture potential subpopulations within each cell type. In contrast, datasets with a large number of cell types tend to exhibit finer subpopulations or distinct cell states rather than broad categories. In such cases, focusing on above-mean bins through MTB helps mitigate the complexity introduced by highly heterogeneous labels.

Computational complexity

scValue has $O(BdN \log N)$ computational complexity. Here, we review the theoretical computational complexities of the six existing subsampling methods, i.e. Uniform (as implemented in the Seurat [7] and Scanpy [8] pipelines), GeoSketch [6], Sphetcher [9], Hopper [10], KH [13], and scSampler [12], as documented in the respective publications. A summary of these complexities, alongside that of scValue, is provided in Table 1 for comparison.

Given a large scRNA-seq dataset consisting of N cells and d features (PCs by default), the sketching objective is to extract S cells that effectively represent the original dataset. Among the seven methods evaluated, five include additional parameters or variables that influence their computational complexity. Specifically, scValue has a parameter B, representing the number of trees in the random forest, with a default value of100. TreeHopper, similarly, involves a parameter B, denoting the number of partitions of the original dataset, typically set to 1000. scSampler [12] also utilizes a partitioning parameter B, which was evaluated in its original publication with values of 1 (no partitioning), 4, and 16. In our experiments, we set $B = 4$ to balance computational cost with sketch quality. Lastly, KH incorporates a parameter D, which denotes the number of random features sampled from the original d features.

Implementation

scValue is implemented in Python, with the data value computation step built upon the *RandomForestClassifier* class (version 1.5.2) from the scikit-learn library. Parallelization of tree-fitting is supported to enhance computational efficiency. By default, scValue uses the top 50 PCs ($d = 50$) as the latent representation of cells; however, it also accepts gene expression features or other low-dimensional embeddings as input. The random forest is constructed with $B = 100$ decision trees by default. In the value-weighted subsampling step, FB is set as the default cell selection strategy. Alternatively, proportional subsampling can be carried out instead if the original cell types are highly balanced or when users prefer to maintain the original cell type proportions.

Evaluation and case studies

In this study, we benchmarked scValue against six established subsampling methods (Uniform [8], GeoSketch [6], Sphetcher [9], TreeHopper [10], KH [13], and scSampler [12]) across a comprehensive suite of analyses including: four ML/DL-based CTA tasks, and three case studies on LT, cell-type harmonization, and (pseudo)bulk RNA-seq Deconv, respectively. Each task/case study was conducted with a distinct pair of scRNA-seq dataset (Table 1) and ML/DL model. Moreover, a systematic sketch metric comparison between scValue and its counterparts were performed using all 16 datasets listed in Table 1. Detailed experimental procedures are provided in Supplementary Note 1.

Results

Demonstration: scValue balances cell-type proportions and improves cell-type separation

To illustrate the capacity of scValue to balance cell type proportions while maintaining separation among different cell types, we tested it on the mTC dataset [23], comprising ~90 k cells across four cell types. We used scValue and six existing sketching methods to subsample 10% of the original dataset (~9 k cells)

and then visualized both the full dataset and each subsample using uniform manifold approximation and projection (UMAP), as shown in Fig. 1b.

Overall, scValue, Uniform, and GeoSketch produced sketches that more closely resembled the full distribution compared to Sphetcher, Hopper, KH, and scSampler. Among the former three, both scValue and GeoSketch exhibited better coverage of the relatively rare natural killer cells (pink) by preserving a denser representation of this cell type. Notably, however, scValue achieved clearer separation between T cells (blue) and regulatory T cells (green) than GeoSketch, indicating that scValue not only captures rare populations but also maintains improved resolution between distinct cell types. These observations hence underscore the ability of scValue to effectively balance cell type proportions and enhance separation among cell types, with the potential of benefiting ML/DL tasks.

Evaluation: scValue outperforms existing sketching methods in machine learning and deep learning-based cell-type annotation tasks

To assess how well scValue preserves essential information in sketches for ML/DL-based CTA, we evaluated the method against its six counterparts on four dataset-model pairs: PBMC with scANVI [19], mBrain with scPoli [20], CxG_min with CellTypist [21], mACA with ACTINN [22]. Evaluation involved predicting labels for both all cell types and the subset of rare cell types. In the simpler PBMC and mBrain datasets (each with 10 cell types), rare cell types were defined as those representing less than 10% of the total cells. In contrast, for the more heterogeneous CxG_min (164 cell types) and mACA (107 cell types) datasets, the rarity threshold was set at 0.5% of the total cells.

Each dataset was split into a reference set (for training) and a full-sized query set (for validation). We then generated sketches of the reference set at varying fractions (2%–10% of the original dataset) using the seven subsampling methods. Subsequently, the trained models were used to infer cell types on the full query set; and this procedure was repeated ten times to ensure robust performance estimates. The detailed experimental settings are provided in Supplementary Note 2.

Figure 2 presents the annotation accuracies in boxplots for each annotation task and Tables S1–S8 summarize the average and standard deviation of accuracies for each sketching method at each subsampling percentage in each dataset-model pair's experiment.

For all-CTAs, in PBMC (Fig. 2a), scValue consistently outperformed other methods, especially at smaller sketch sizes (2%–4%). At 10%, scValue reached an average accuracy of 0.8330 compared to 0.8635 for the full dataset, with less variability than its counterparts. In mBrain (Fig. 2b), scValue maintained high and steadily increasing accuracy across all sketch sizes, beginning with a significant lead at 2% and closely approaching the full-reference accuracy of 0.9320. In the more complex CxG_min (Fig. 2c), although all methods improved with larger sketches, scValue outperformed the alternatives in all cases; at 10%, it reached 0.5811 versus 0.6847 for the full dataset. Similarly, in mACA (Fig. 2d) with 197 cell types, scValue led at all sketch sizes, achieving 0.5901 at 10% despite the full-reference accuracy being 0.7176.

For rare-CTAs, a similar trend can be observed in the four datasets (Fig. 2e-h). With only a few exceptions in mBrain (Fig. 2f) at 4%, 6%, and 8% sketch sizes, scValue consistently outperformed the six baseline methods. At a 10% sketch level, scValue achieved rare cell-type accuracies that were closest to or even exceeded the full dataset results, as seen in PBMC (0.6855 versus 0.6224) and mBrain (0.7207 versus 0.6896). These findings confirm that scValue effectively preserved and enhanced the representation of rare cell populations after downsizing, offering a more reliable annotation performance than alternative subsampling approaches.

For datasets with highly heterogeneous labels, such as CxG_min (164 cell types) and mACA (197 cell types), it is challenging to effectively capture the label distributions with small sketch sizes (e.g. 2%–10%). Therefore, increasing the sketch size is advisable for a better representation of the distributions. Following this rationale, we conducted an additional experiment using sketch sizes of 12%–20% on the CxG_min and mACA datasets while keeping all other experimental settings the same as those for the 2%–10% range. As presented in Fig. S1 and Tables S9 and S10, the results indicate that increasing the sketch size improved accuracy across most subsampling methods; notably, scValue continued to lead. At a 20% sketch size, scValue achieved average accuracies of 0.6401 versus 0.6847 for CxG_min and 0.6511 versus 0.7176 for mACA, yielding considerably smaller gaps compared to those observed at a 10% sketch size.

Demonstration: scValue-core enhances sketch quality by mitigating noise and annotation bias

Since noisy data or biased CTAs can impair scValue's ability to accurately prioritize cells and generate reliable subsamples, we have developed a method that first constructs a core sample set from the original dataset before applying scValue. We call this approach scValue-core. The method systematically filters out low-confidence cell type labels using both confidence score computation and subsequent validation. Each cell is assigned a confidence score (ranging from 0 to 1) via either an official or a user-trained CellTypist model. The predicted cell types are then compared with the original annotations, and only cells with matching or broadly similar labels are retained (although this validation step is optional). A user-defined threshold (default 0.5) is applied to filter out low-confidence cells, thereby constructing a high-quality core sample set. Finally, scValue subsampling is performed on this refined dataset to produce a downsized, high-quality sketch, which is expected to enhance downstream ML/DL applications when compared to the standard scValue sketch.

To demonstrate the effectiveness of scValue-core, we applied it to the PBMC dataset using a CellTypist model (*Immune_All_High.pkl*) for annotating high-level immune cells from the official repository. Cells with predictions that did not match their original annotations were removed, and a core sample set was constructed using the default confidence score threshold of 0.5. Subsequent scValue subsampling on this filtered dataset generated sketches representing 2% to 10% of the original data. As shown in Fig. 3a and b, these sketches consistently outperformed both standard scValue and six baseline methods (Uniform, GeoSketch, Sphetcher, Hopper, KH, and scSampler) across various sketch sizes in all-cell-type and rare-CTA tasks. Tables S11 and S12 summarize the mean ± standard deviation accuracies for evaluations on all cell types as well as rare cell types on the PBMC dataset. Both scValue-core and standard scValue outperformed the other subsampling methods, achieving accuracies comparable to or even exceeding those of the full dataset. Notably, scValue-core, which leverages the curated core sample set, delivered superior performance relative to standard scValue; for example, at a 10% sketch size, scValue-core achieved accuracies of 0.8498 for all cell types and 0.7162 for rare cell types, compared to 0.8330 and 0.6855, respectively. These results suggest that employing scValue subsampling on a carefully curated core sample set can further enhance annotation performance.

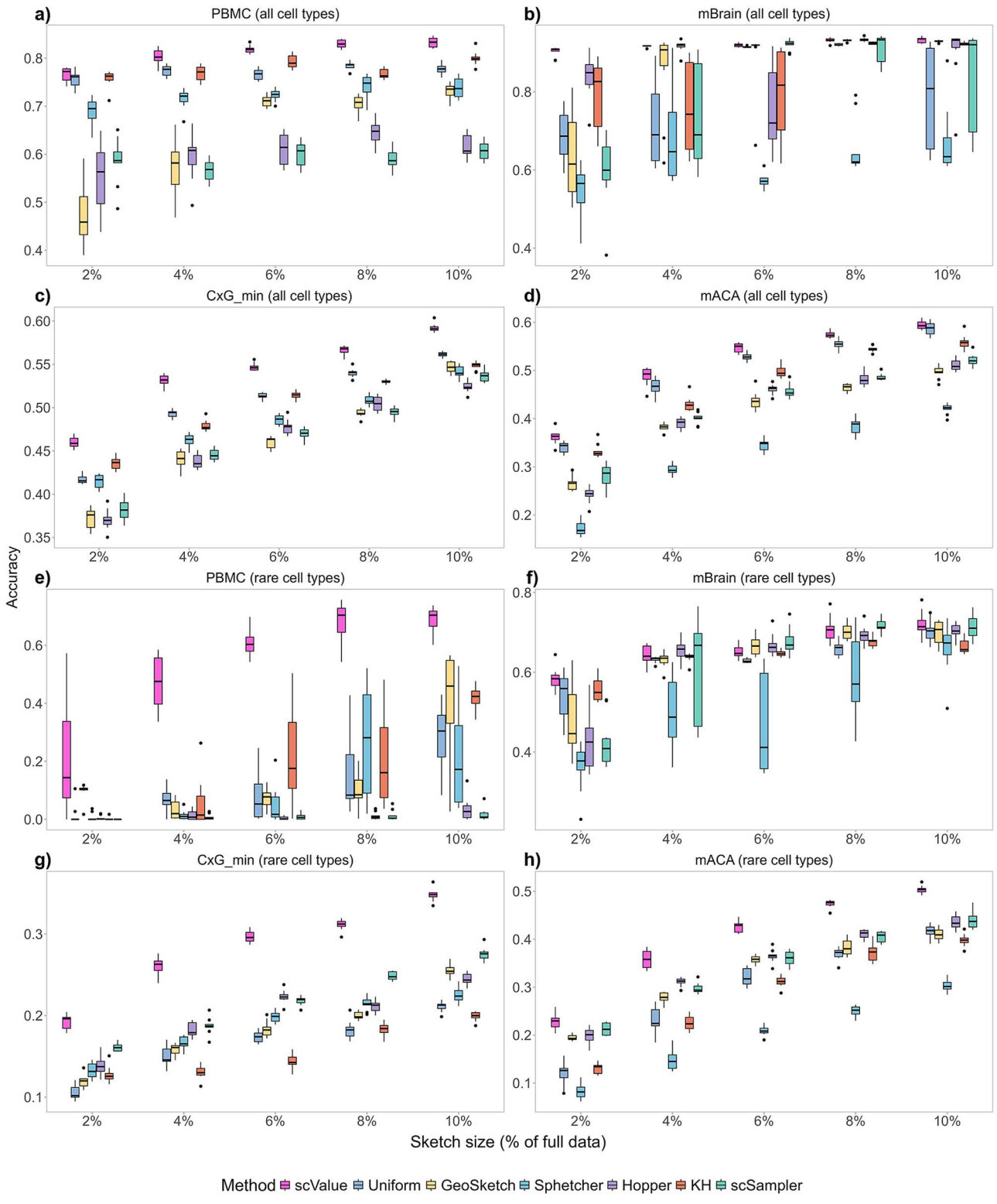

Figure 2. scValue outperforms existing sketching methods in ML/DL-based CTA tasks. We evaluated scValue against Uniform, GeoSketch, Sphetcher, Hopper, KH, and scSampler in (a-d) all-cell-type and (e-h) rare-CTA tasks using four previously studied dataset-model pairs: PBMC with variational autoencoder-based scANVI, mBrain with variational autoencoder-based scPoli, CxG_min with logistic regression-based CellTypist, mACA with neural network-based ACTINN. For each dataset, sketches of the reference partition were created at varying fractions (2%–10%) and used to train the corresponding model to annotate the full query partition.

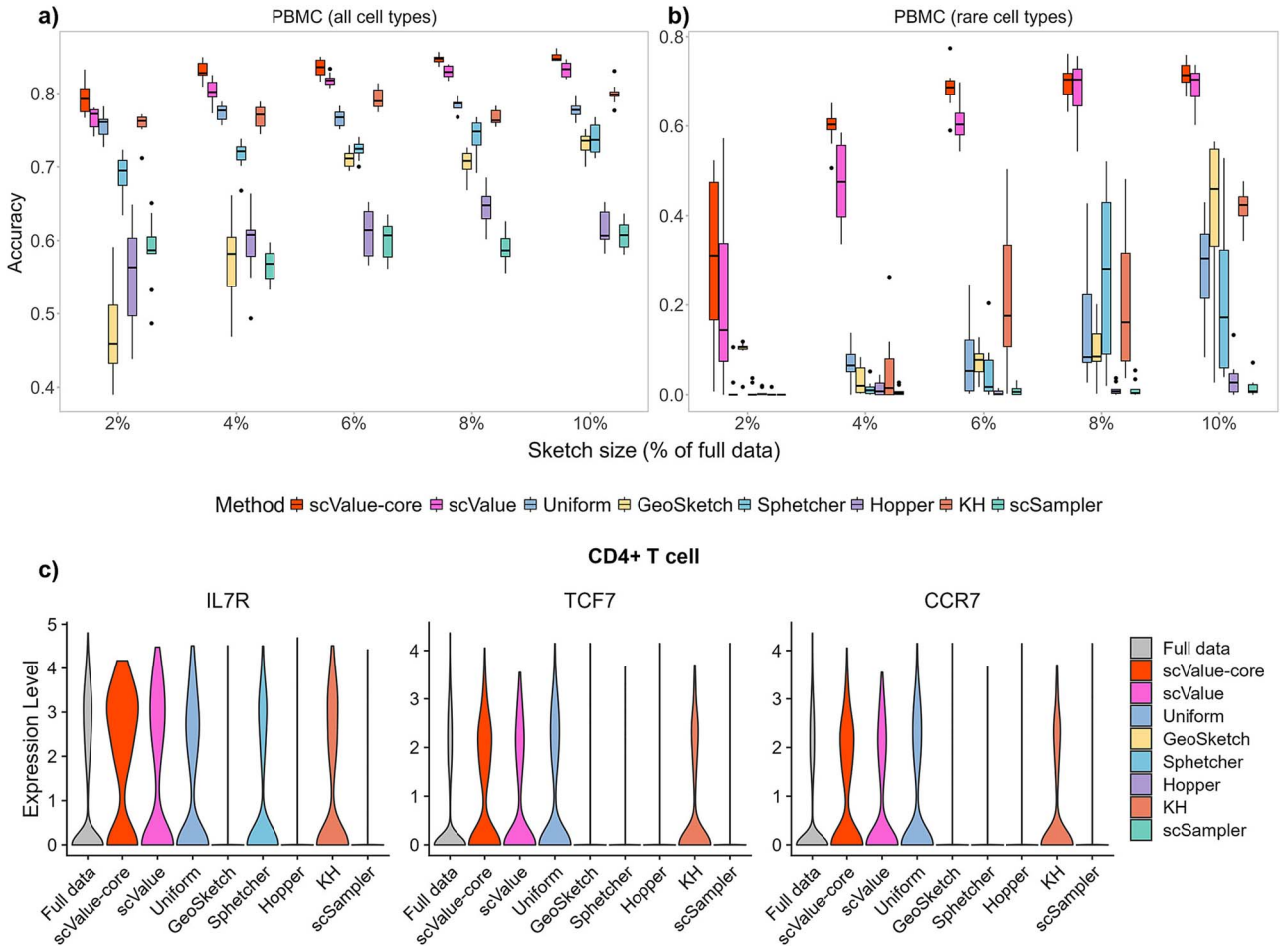

Figure 3. scValue-core, that leverages the core sample set for subsampling, can further improve the performance for both (a) all-cell-type and (b) rare-CTA tasks with the PBMC dataset. (c) Cell-type-specific marker gene expressions are plotted to further analyse the quality of the sketches. Here, the top three markers for CD4+ T cell in the PBMC dataset are visualized as an example.

Cell-type-specific marker gene expressions were analysed to further evaluate the quality of sketches by various methods. Specifically, marker genes for the nine cell types (CD4+ T cells, cytotoxic T cells, natural killer cells, CD14+ monocytes, CD16+ monocytes, B cells, dendritic cells, megakaryocytes, and plasmacytoid dendritic cells, with the 'Unassigned' label excluded) in PBMC, were retrieved from the CellMarker 2.0 database [24].

For each cell type, we used Seurat's FindMarkers function to identify the top three marker genes. We then generated violin plots (see Fig. 3c for CD4+ T cell as an example and Fig. S2 for all cell types) to display the expression levels of these marker genes across datasets produced by the various subsampling methods, including our proposed approaches (scValue-core and standard scValue) as well as six baseline methods. Overall, the expression levels of marker genes in datasets processed with scValue-core and scValue were similar to those in the full dataset. This suggests that our method could preserve key biological signals during downsizing and that the core sample set system could potentially correct for biases introduced by low-confidence labels.

Case study: scValue enables improved CellTypist label transfer learning

To examine how scValue performs when reference and query data have divergent annotation styles, we conducted a LT learning experiment using the Gut (428 k cells) [25] and immune (42 k cells) [26] datasets from the CellTypist tutorial [27]. Following the tutorial's workflow, we first reduced the 428 k Gut dataset to 55 k balanced cells (ensuring each cell type was equally represented). Because the dataset was already well-balanced, we used proportional subsampling rather than value-weighted subsampling and then applied scValue to produce a 10% sketch, while the immune dataset (42 k cells) remained unchanged. We trained a CellTypist model [24] on each 10% sketch (generated by scValue and the six baseline subsampling methods) and then transferred its annotations to selected T-cell populations ('Activated CD4 T,' 'Th1,' 'Tfh,' 'CD8 T,' 'cycling gd T,' 'Tcm,' 'gd T,' 'Th17,' and 'Treg') in the immune query dataset. To evaluate performance, the transferred annotations were compared against the query dataset's original annotations. Figure 4 shows dot plots contrasting these predictions with the original alignment obtained using the full reference data in the tutorial.

Several trends can be observed from the plots. Among all methods, scValue and Uniform yielded LTs that most resembled the original ones. In addition, scValue not only preserved but also improved the correct classification of 'Tfh' cells, indicating its strength in retaining subtle yet important features of cell types. The other five sketching methods showed higher levels of different assignments:

- GeoSketch had a larger proportion of 'CD8 T' cells and 'Treg' cells classified as 'Activated CD4 T'.

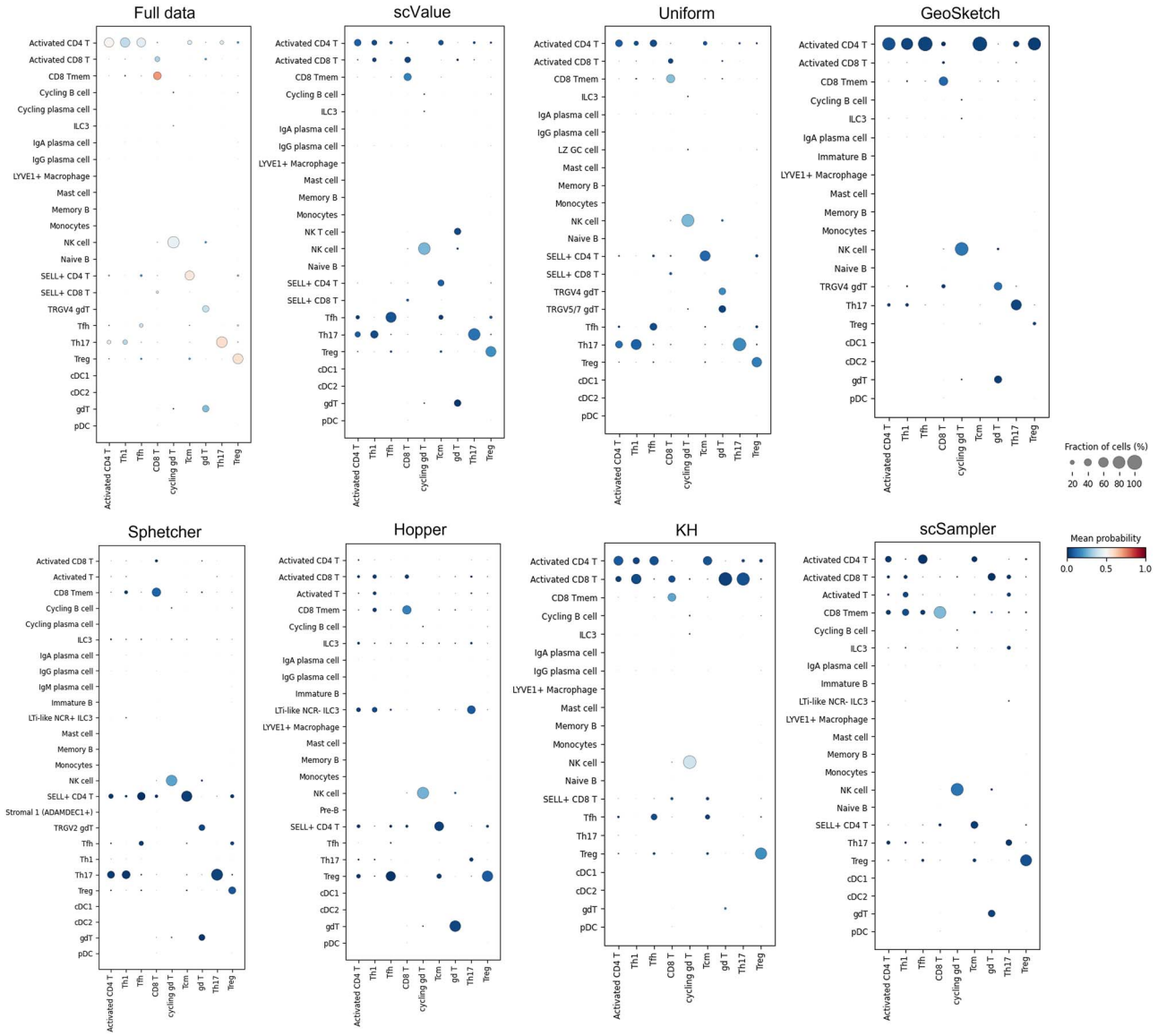

Figure 4. scValue enables improved CellTypist LT learning. In the case study of employing CellTypist to transfer labels from the gut (reference) to the colon (query) dataset, scValue's sketch preserved T-cell labels most similar to the full reference and even surpassed the latter in the 'Tfh' cell type, as reflected in the larger fraction of cells being correctly predicted for this type.

- Sphetcher did not predict any 'Activated CD4 T cells'.
- Hopper inferred the majority of 'Activated CD4 T' cells as other T subtypes.
- KH displayed a noticeable tendency to label 'Activated CD4 T' as 'Activated CD8 T'.
- scSampler classified a large portion of 'Activated CD4 T' as 'CD8 Tmem', and some 'CD8 T' as 'SELL+ CD4 T'.

Subsampling the reference data naturally led to lower LT probabilities, as visually reflected by more blue bubbles (indicating reduced confidence) in each sketch's dot plot. Nevertheless, scValue emerged as a top-performing subsampling strategy, preserving enough critical information to sustain accurate LT.

Case study: scValue facilitates improved label harmonization with CellHint

To explore how effectively scValue captures complex cell-type relationships when integrating scRNA-seq data from multiple sources with divergent annotation styles, we followed the CellHint tutorial workflow [28] to harmonize T-cell annotations across four independent studies, collectively comprising 200 k cells in the Spleen dataset [27]. Each study provides a slightly different nomenclature for T-cell subtypes, making robust cross-study alignment crucial for constructing a standardized cell atlas.

In this experiment, we first generated a 10% sketch of the Spleen dataset (i.e. ~20 k cells) using scValue. Notably, because the original dataset already contained various distinct T-cell labels from multiple sources, our goal was to preserve these fine-grained subpopulations in the sketch. We then applied CellHint [28] to learn hierarchical relationships among the T-cell annotations from the four studies, using the scValue-derived sketch as input. For comparative evaluation, the other six sketching methods also created 10% subsets, which were fed through the same pipeline. Figure 5 depicts the resulting tree plot of scValue representing inter-study label relationships, compared to the original tree constructed from the full 200 k dataset. Plots for the six baseline methods are provided in Fig. S3.

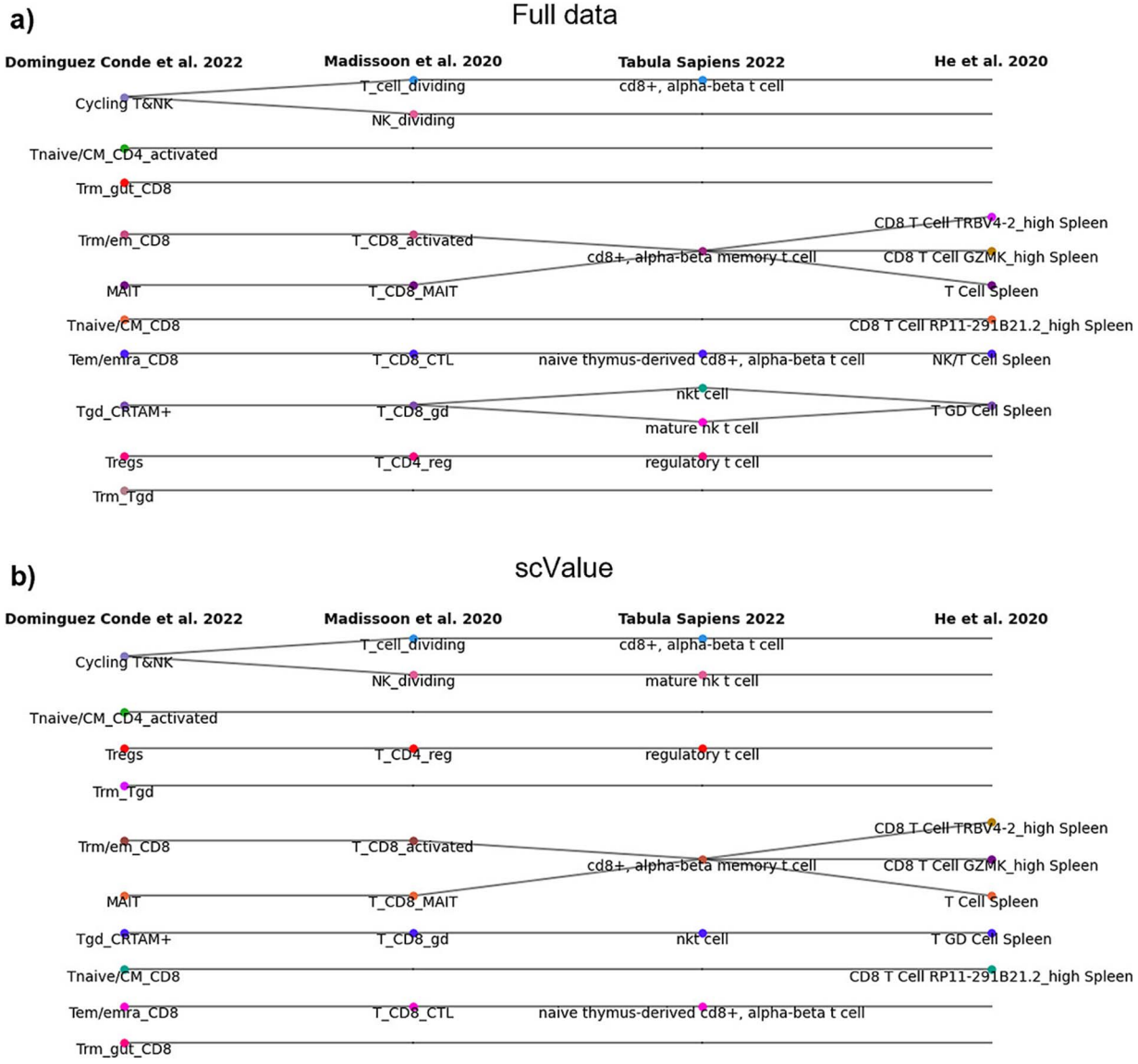

Figure 5. scValue facilitates improved label harmonization with CellHint. In the case study of label harmonization for the multi-study spleen dataset via CellHint, scValue's sketch (a) more accurately reproduced inter-study T-cell subtype relationships compared to those obtained with the full data (b).

Of all methods tested, scValue performed best in reproducing the fine-grained relationships among T-cell subtypes observed in the full dataset's tree. This reflects scValue's ability to maintain critical cellular diversity in reduced subsets, even under conditions where the nomenclature varies across studies. Other methods exhibited diverse degrees of different matching compared to the original tree, including deletion of existing T-cell subtype relationships and/or introductions of new ones:

- Uniform resulted in a reordering of studies and removed relationships involving the 'CD8-positive, alpha-beta memory T cell' and 'T Cell Spleen'.
- GeoSketch related 'Tnaive/CM CD8' to 'T_CD4 naïve'.
- Sphetcher both altered the study order and linked 'Tnaive/CM CD8' to 'T_CD4 naïve'.
- Hopper merged 'Trm_Tgd' and 'Tgd_CRTAM+' into 'T_CD8_gd'.
- KH similarly led to a change of study ordering and related 'Tnaive/CM CD8' to 'T_CD4 naïve'.
- scSampler introduced several additional relationships including linking 'Tnaive/CM_CD4_activated' with 'T_CD4_conv', merging 'Trm_Tgd' and 'Tgd_CRTAM+' into 'T_CD8_gd', and integrating 'Trm_gut_CD8' and 'Trm/em_CD8' into 'T_CD8_activated'.

These inconsistencies likely stem from each method's differing priorities in selecting representative cells. Approaches that undersample or over-sample particular subpopulations can distort the underlying biological relationships, particularly in datasets with nuanced cell-type definitions that vary by study. In contrast, scValue's built-in balancing and value-based selection strategies appear to safeguard these subtle distinctions.

Case study: scValue effectively builds single-cell ribonucleic acid-sequencing reference for bulk ribonucleic acid-sequencing deconvolution with MuSiC

We evaluated scValue for building a single-cell expression reference for MuSiC [29], which is a reference-based Deconv tool identified as the top performer in a recent systematic benchmark study [30]. 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. [29], 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. As shown in Table 3, scValue achieved the highest correlation (0.6851), the lowest RMSE (0.0592), and the lowest MAE (0.0434). Additionally, Fig. S4 illustrates that the inferred cell type proportions from scValue closely match those obtained from the full dataset, outperforming the six baseline methods. These results demonstrate the applicability of scValue in constructing reliable single-cell references for (pseudo)bulk tissue cell type Deconv.

Comparison: scValue yields efficient computation time, best Gini coefficient, and uniform-like Hausdorff distance

To conclude the evaluations, we benchmarked scValue against the six other methods using the 16 datasets of varying size (Table 1), focusing on computation time, Gini coefficient, and Hausdorff distance. For each dataset, a 10% sketch was generated by each method. The scatter plots in Fig. 6 illustrate computation time (x-axis) versus Gini coefficient (y-axis), with bubble size indicating the sketch's Hausdorff distance from the full dataset. The detailed statistics for producing the plots are provided in Table S13. A major observation is that scValue (shown in pink) typically clustered toward the lower ends of both axes, reflecting its short runtime and low Gini coefficient (balanced cell-type proportions). Meanwhile, the moderate bubble size indicates that scValue's sketch remains reasonably close to the original data distribution. It should be noticed that, for the mEmbryo dataset (containing ~3.2 million cells), Sphetcher and KH did not complete their runs and thus no sketches were included in the relevant plots; similarly, for the Fetal dataset (about four million cells), Hopper, Sphetcher, and KH did not finish running and were therefore excluded from the results.

From Table 4, which ranks each method by its average performance on the three metrics across the datasets, scValue stood out for its computation time rank of 2.4 ± 0.8, second only to the simple Uniform approach, and best Gini coefficient rank of 2.1 ± 1.5. The Hausdorff Distance rank of 5.3 ± 1.1 was slightly lower than Uniform (5.5 ± 1.2) and KH (5.4 ± 1.3), confirming that scValue's sketches did not deviate excessively from the original distribution. These results suggest that scValue strikes a distinct trade-off: it excels at capturing the full diversity of cell populations and is computationally efficient, yet it can exhibit a slightly higher distributional shift compared to alternative methods that are based on distance-optimization (especially, GeoSketch, Hopper, and scSampler). Nevertheless, this shift remains comparable or better than Uniform's performance, reinforcing that scValue offers a robust balance between cell-type representativeness, efficiency, and data fidelity. Finally, Fig. S5 further illustrates how scValue's computational time is nearly independent of the sketch size (tested from 2% to 10%) using the Spleen dataset as an example, which resonates with $O(BdN \log N)$ complexity. This constant-time characteristic emphasizes the practical scalability of scValue, which is an important consideration for large-scale single-cell applications where speed matters.

Table 2. Theoretical computational complexities of the seven sketching methods evaluate in this study

Method	Theoretical complexity
scValue	$O(BdN \log N)$ with Btrees and d features
Uniform	$O(1)$
Geosketch	$O(dN \log N)$ with $d$features
Hopper	$O(NS)$ $O(NS/B)$ for TreeHopper with B partitions
Sphetcher	$O(\sum_i \lvert L_i \rvert)$ for each iteration until L converges, where $\lvert L_i \rvert$ is the size of the i-th spherical subset
KH	$O(dDNS)$ with d features and D random features
scSampler	$O(NS)$ $O(NS/B)$ if split by B subsets

Regarding the computation time observed in the sketch metric comparison, scValue's practical efficiency can be understood as follows. Since the target sketch size S is usually considerably larger than $\log(N)$ for typical sketching applications in single-cell analysis [6], it can be inferred from Table 2 that the theoretical ranking of computational complexity in ascending order is: Uniform < GeoSketch < scValue < (Tree)Hopper ≈ scSampler (with partitioning) < Sphetcher < KH. However, actual computation time may diverge from this order due to several factors [31], such as the constant and lower-order terms ignored in Big-O notation, dataset-specific characteristics, memory management overhead, and other implementation details. Moreover, as the random forest used in scValue can be easily parallelized, this method may achieve faster or comparable runtimes to other sketching methods (including GeoSketch) in practice, especially for large datasets.

Discussion

This study introduces scValue, a novel subsampling method devised for large-scale single-cell transcriptomic data to prioritize each cell based on its utility for cell-type classification. Rather than optimizing distance (e.g. GeoSketch, Sphetcher, Hopper, scSampler) or distribution (e.g. KH), scValue uses a random forest to compute each cell's OOB prediction accuracy as its data value. Subsample sizes per cell type are then determined by both abundance and variability in these values, allocating more cells to rare or intricate cell types. Finally, the FB or MTB strategy can select cells at different value levels, preserving overall data diversity but still favoring higher-value cells.

We systematically compared scValue with six established subsampling methods (Uniform, GeoSketch, Sphetcher, Hopper, KH, and scSampler) across multiple ML/DL tasks. In CTA tasks (PBMC with scANVI, mBrain with scPoli, CxG_min with CellTypist, and mACA with ACTINN), scValue achieved consistently high accuracies, stable runs, and progressive performance gains with larger sketch sizes, approaching results obtained with the full dataset. In three case studies, scValue best preserved T-cell identities when transferring labels from Gut to Colon with CellTypist, most accurately reproduced T-cell subtype relationships in the Spleen

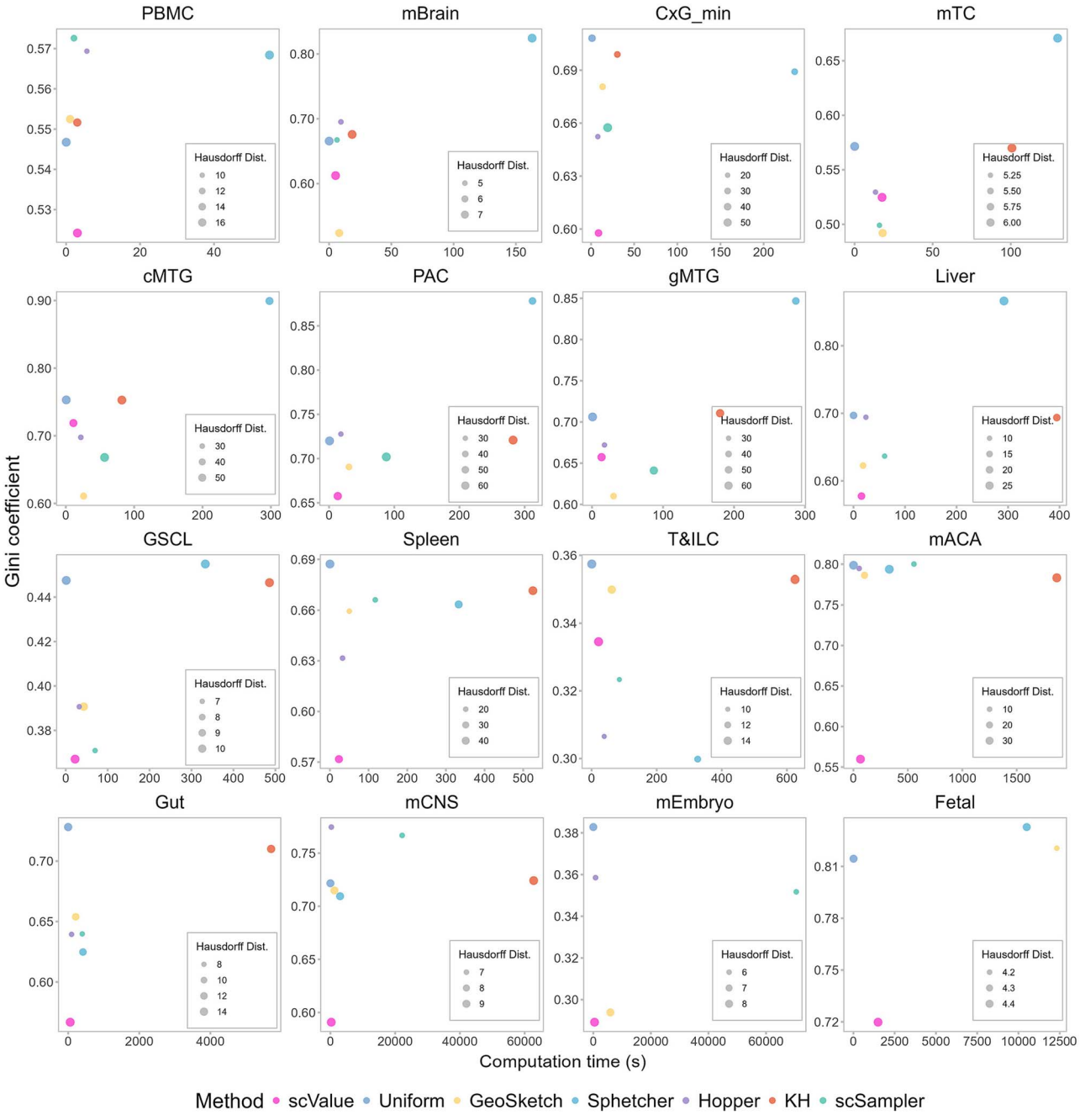

Figure 6. scValue yields efficient computation time, the best Gini coefficient, and uniform-like Hausdorff distance. We evaluated scValue against its counterparts in terms of three sketch quality metrics across 16 datasets containing between 31 000 and 4 million cells, spanning four to 197 cell types and including samples from more than 10 tissues across four species.

Table 3. Summary of evaluation metrics for MuSiC deconvolution by respectively using full data and the sketches by scValue and the six baseline methods as references.

Method	Correlation	RMSE	MAE
Full data	0.6478	0.0702	0.0510
scValue	**0.6851**	**0.0592**	**0.0434**
Uniform	0.6629	0.0645	0.0466
GeoSketch	0.5763	0.0654	0.0490
Sphetcher	0.6591	0.0651	0.0485
Hopper	0.6348	0.0750	0.0543
KH	0.5957	0.0674	0.0478
scSampler	0.6531	0.0738	0.0512

Note: bold value in the Correlation column denotes the highest (best) values, while bold values in the RMSE and MAE columns denote the lowest (best) values.

Table 4. Rankings of each subsampling method based on its average performance on the three metrics across the 16 datasets

Method	Computation time rank	Gini coefficient rank	Hausdorff distance rank
scValue	2.4 ± 0.8	2.1 ± 1.5	5.3 ± 1.1
Uniform	1.0 ± 0.0	5.4 ± 1.6	5.5 ± 1.2
GeoSketch	3.9 ± 0.5	2.6 ± 1.4	2.7 ± 0.6
Sphetcher	6.1 ± 1.0	5.1 ± 2.3	4.7 ± 1.6
Hopper	3.1 ± 0.9	4.0 ± 1.5	1.3 ± 0.5
KH	6.5 ± 0.8	5.0 ± 1.3	5.4 ± 1.3
scSampler	4.9 ± 0.9	3.7 ± 1.7	2.6 ± 1.9

dataset with CellHint, and achieved best (pseudo)bulk Deconv performance with MuSiC on the T&ILC dataset. Finally, using 16 large scRNA-seq datasets, scValue was second fastest (behind Uniform), produced well-balanced sketches (lowest Gini coefficient), and maintained a Hausdorff distance similar to Uniform, indicating small deviation from the original data.

From a theoretical perspective, scValue's subsampling strategy, which maintains balanced cell-type proportions and emphasizes highly informative cells, can benefit several types of ML/DL algorithms. Deep learning models, such as the neural network-based ACTINN model and variational autoencoder-based approaches like scANVI and scPoli (benchmarked in our study), are particularly sensitive to class imbalance. By providing a representative and balanced subsample, scValue ensures that rare cell types are adequately represented, which leads to improved model convergence and reduced bias. Furthermore, models that utilize mini-batch stochastic gradient descent (SGD), as exemplified by the logistic regression-based CellTypist model, can benefit from training data that accurately reflects the true underlying distribution. With scValue, each mini-batch is more likely to contain a diverse and balanced representation of cell types, thereby enhancing learning stability and generalization. In addition, emerging frameworks such as graph neural networks and transformer-based models for single-cell data analysis can benefit from scValue's ability to capture subtle biological variability. By retaining rare or nuanced subpopulations, scValue preserves the relational structure between cells that is essential for these methods. Moreover, the use of OOB estimates to assess cell-level informativeness contributes to selecting cells that support stable decision boundaries. This advantage is particularly relevant for ensemble methods like random forests and for support vector machines (SVMs), where balanced training data improves classification margins and mitigates overfitting.

Although the evaluation of this study primarily focuses on CTA, scValue is broadly applicable to other ML/DL tasks. For example, in our cell-type harmonization experiment using CellHint on the human Spleen dataset, scValue preserved fine-grained relationships among T-cell subtypes across multiple studies and enabled more robust cross-study comparisons despite variations in annotation styles. Similarly, our bulk RNA-seq Deconv experiment using MuSiC showed that the subsampled single-cell reference generated by scValue closely resembled the full dataset in accurately reconstructing known cell-type proportions in pseudobulk samples. These findings underscore scValue's effectiveness in maintaining biological fidelity and enhancing downstream predictive performance, even when applied to tasks beyond direct CTA. Furthermore, the data valuation framework underlying scValue can be extended to optimize training data for general predictive models, particularly in high-dimensional and imbalanced settings. By prioritizing high-value data points, scValue reduces redundancy and concentrates on the most informative cells, potentially leading to more accurate and efficient model training in applications such as single-cell drug response prediction, disease state classification, and patient stratification. These extended applications can potentially demonstrate the versatility of scValue and reinforce its theoretical basis by linking cell importance (as estimated by OOB accuracy in a random forest) to robust performance across various ML/DL tasks.

scValue can also sketch scRNA-seq datasets from multiple sources. The choice to integrate before or after running the method depends on the magnitude of batch effects and the biological distinctiveness of each dataset. If the datasets are well-integrated with minimal batch effects, merging them into a single dataset before applying scValue is advantageous, as it produces a subsample that accurately reflects the global cellular landscape. This strategy is particularly effective for analyses that require a comprehensive view, such as identifying shared cell populations or constructing integrated cell atlases. Conversely, if the datasets exhibit strong batch effects or contain unique biological characteristics, it is preferable to run scValue on each dataset independently. This approach preserves source-specific features, such as rare cell types or subtle transcriptomic signals, that might otherwise be lost during early integration. After obtaining representative sketches from each dataset, established integration methods can be applied to the combined subsamples, thereby retaining critical distinctions while reducing overall data size to a manageable scale.

Despite the various strengths, scValue's reliance on accurate OOB estimates from the random forest makes it sensitive to noise and redundant cells in the input data, which can lead to inaccurate subsamples. To mitigate this issue, we have implemented scValue-core, an approach that constructs a core sample set from the original data using a systematic filtering process. In scValue-core, each cell is assigned a confidence score using an established CellTypist model, its predicted cell type is validated against the original annotation, and cells scoring below a user-defined threshold are filtered out. By retaining only high-confidence cells for subsequent subsampling, scValue-core effectively reduces the impact of noise and bias, thereby improving the quality of the final sketch. It is also recommended to conduct rigorous quality control and preprocessing before applying scValue. For instance, removing duplicate or near-duplicate cells, applying batch correction to minimize technical variability, and using data imputation methods to recover missing or noisy expression values can all help ensure a cleaner input dataset and, consequently, more reliable OOB estimates. Furthermore, during the data value computation step of scValue, tuning random forest hyperparameters (such as increasing the number of trees and adjusting the maximum depth) may further stabilize OOB estimates in noisy conditions.

We also propose that future versions of scValue could integrate additional selection metrics, such as local density or variability measures, alongside OOB accuracy to provide a more comprehensive evaluation of cell importance and further mitigate the effects of noise and redundancy.

Furthermore, scValue may also introduce bias by favoring cells with stronger, more 'typical' signals. This bias can potentially limit the diversity captured in the subsample in two important ways. First, it may lead to an over-emphasis on cell subpopulations that are inherently easier to classify due to their distinct markers, resulting in a disproportionate representation of these cells while under-sampling rare or ambiguous groups. Second, by focusing on cells that perform well in classification, the approach might compromise the fidelity of the original expression distribution, which is crucial for analyses such as differential expression and gene regulatory inference, where subtle heterogeneity is important. To mitigate these potential biases, we propose several strategies. Rigorous quality control of the input data is essential because noisy or duplicate cells can inflate OOB estimates, thereby distorting the ranking of cells. In our method, the FB strategy divides the entire [0, 1] data value range into equal intervals, which ensures that even cells with lower OOB values are included in the final subsample, preserving a broader spectrum of cellular variability. Additionally, the scValue-core approach constructs a refined core sample set by first assigning a confidence score to each cell based on predictions from an established CellTypist model and then validating these scores against original annotations; only cells that meet or exceed a pre-defined threshold are retained for subsampling. Finally, incorporating additional metrics such as local density and variability alongside OOB accuracy allows for a more nuanced assessment of each cell's annotation quality, ensuring that sparsely populated or outlier regions are adequately represented. Together, these strategies can help preserve both the diversity of cell types and the fidelity of the original data distribution, thereby enhancing the generalizability and robustness of downstream ML/DL tasks.

The subsampling process of scValue can attenuate subtle differences in cell-to-cell variability, which in turn may hinder analyses that demand an accurate reflection of the original data distribution, such as differential expression analysis and gene regulatory inference, where even small shifts in expression levels are crucial. To alleviate this problem, the scValue Python package offers an option for proportional subsampling (by setting *prop_sampling* to true) that more faithfully preserves the original distribution. However, scValue's focus on high-value cells is designed to boost classifier accuracy rather than replicate the complete transcriptomic distribution. As a consequence, statistical methods that require comprehensive data, like DESeq2 [32] for differential expression analysis or weighted gene correlation network analysis (WGCNA) [33] for co-expression network construction, may be adversely affected. These methods depend on precise variance estimates, accurate gene-level distributions, and even subtle expression differences, all of which can be diluted when low-abundance transcripts and fine expression gradients are underrepresented. For analyses that are sensitive to the full range of expression variability, it is advisable either to work with the full dataset or to adopt alternative subsampling strategies, such as geometry-based approaches like GeoSketch [6], which are better suited to maintaining the dataset's fine-grained distributional characteristics. In all cases, validating findings from subsampled data against those from the complete dataset is highly recommended to ensure robust conclusions.

A general caveat in sketching for ML/DL tasks is that, after subsampling, the number of cells may become comparable to or even less than the number of genes, exacerbating dimensionality issues [34]. It remains critical to select an appropriate number of HVGs or to employ suitably designed model architectures to mitigate the curse of dimensionality.

To make a final point, scValue can be envisioned as a general data valuation and subsampling framework for single-cell data. Although the current method focuses on cell-type labels, future extensions could incorporate other meta-information, e.g. developmental timing, disease status, tissue source, and species, when utilizing such information to build predictive models at scale. Besides, subsampling may be reframed as removing low-value cells to identify a minimal dataset yielding near-optimal predictive performance. Such a 'minimal-viable sketch' concept could be particularly valuable for studies constrained by computing resources but aiming to preserve salient biological signals in ML/DL applications.

Key Points

- Value-based subsampling of large single-cell ribonucleic acid-sequencing (scRNA-seq) data tied to classification utility: scValue introduces a novel 'value-based' strategy that leverages out-of-bag estimates from a random forest to quantify each cell's importance for distinguishing its type. By linking subsampling directly to classification performance, scValue prioritizes the most informative cells for efficient downstream machine learning and deep learning (ML/DL) tasks.
- Balanced yet biologically rich sketches: Through value-guided allocation, scValue ensures that rare or complex cell types are proportionally represented with more top-valued cells, thereby preserving essential biological information.
- Consistently high ML/DL accuracy: In benchmarking tasks spanning cell-type annotation, cross-dataset label transfer, label harmonization, and bulk RNA-seq deconvolution, scValue outperforms (or occasionally matches) existing subsampling methods, maintaining model accuracy close to that achieved with full datasets.
- Efficiency and scalability: scValue operates efficiently on large-scale scRNA-seq datasets (ranging from tens of thousands to millions of cells), producing balanced sketches with near-uniform data distributions.

Acknowledgements

We thank the high-throughput sequencing and high-performance computing platform of the Institute of Systems Medicine for technical support.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work has been supported by National Natural Science Foundation of China (32300560); CAMS Innovation Fund for Medical Sciences (CIFMS) [2021-I2M-1-061, 2022-I2M-2-004, 2023-I2M-2-005]; Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences [2022-RC416–01]; NCTIB Fund for R&D Platform for Cell and Gene Therapy, the Suzhou Municipal Key Laboratory [SZS2022005]; the Special Research Fund for Central Universities, Peking Union Medical College (3332024089). Chinese Academy of Medical Sciences & Peking Union Medical College, Union Medical College Young Scholar Support Program, No. 2023086 and 2023088; China Postdoctoral Science Foundation, the Postdoctoral Fellowship Program (Grade B), Grant No. GZB20230084.

Data availability

PBMC is accessible in the Matrix Market format from https://portals.broadinstitute.org/single_cell/study/SCP424/single-cell-comparisonpbmc-data.

mBrain is acquired by downloading the mouse_brain_normalized.h5ad file from the dataset directory at https://github.com/theislab/scArches-reproducibility.

CxG_min is obtained from https://github.com/theislab/scTab/tree/devel, where the minimal subset of the training and test data were used in our study.

mTC is available under 'Major cell cluster: T cells' at https://cellxgene.cziscience.com/collections/45d5d2c3-bc28-4814-aed6-0bb6f0e11c82.

cMTG is collected under 'Chimpanzee: Great apes study' from https://cellxgene.cziscience.com/collections/4dca242c-d302-4dba-a68f-4c61e7bad553.

PAC is obtained under 'Dissection: Primary auditory cortex(A1)' from https://cellxgene.cziscience.com/collections/d17249d2-0e6e-4500-abb8-e6c93fa1ac6f.

gMTG is downloaded under 'Gorilla: Great apes study' from https://cellxgene.cziscience.com/collections/4dca242c-d302-4dba-a68f-4c61e7bad553.

Liver is available under 'All cells from human liver dataset' at https://cellxgene.cziscience.com/collections/74e10dc4-cbb2-4605-a189-8a1cd8e44d8c.

GSCL is sourced under 'Human Somatic Cell Lineage' from https://cellxgene.cziscience.com/collections/661a402a-2a5a-4c71-9b05-b346c57bc451.

Spleen is downloaded from https://celltypist.cog.sanger.ac.uk/Resources/Organ_atlas/Spleen/Spleen.h5ad.

T&ILC is collected under 'T & innate lymphoid cells' from https://cellxgene.cziscience.com/collections/62ef75e4-cbea-454e-a0ce-998ec40223d3.

mACA is sourced from https://figshare.com/articles/dataset/Processed_files_to_use_with_scanpy_/8273102/3; we used the tabula-muris-senis-bbknn-processed-official-annotations.h5ad file (the Version 3 atlas).

Gut is obtained from https://cellgeni.cog.sanger.ac.uk/gutcellatlas/Full_obj_raw_counts_nosoupx.h5ad.

mCNS is collected under 'Major cell cluster: CNS neurons' from https://cellxgene.cziscience.com/collections/45d5d2c3-bc28-4814-aed6-0bb6f0e11c82.

mEmbryo is available under 'Major cell cluster: Mesoderm' at https://cellxgene.cziscience.com/collections/45d5d2c3-bc28-4814-aed6-0bb6f0e11c82.

Fetal is sourced under 'Survey of human embryonic development' from https://cellxgene.cziscience.com/collections/c114c20f-1ef4-49a5-9c2e-d965787fb90c.

References

1. Regev A, Teichmann SA, Lander ES. *et al.* The human cell atlas. *elife* 2017;**6**:e27041. https://doi.org/10.7554/eLife.27041
2. CZI Single-Cell Biology Program, Abdulla S, Aevermann B. *et al.* CZ CELL× GENE discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res* 2025;**53**:D886–900. https://doi.org/10.1093/nar/gkae1142
3. Deng Y, Chen P, Xiao J. *et al.* SCAR: Single-cell and spatially-resolved cancer resources. *Nucleic Acids Res* 2024;**52**:D1407–17. https://doi.org/10.1093/nar/gkad753
4. Deng Y, Lu Y, Li M. *et al.* SCAN: Spatiotemporal cloud atlas for neural cells. *Nucleic Acids Res* 2024;**52**:D998–1009. https://doi.org/10.1093/nar/gkad895
5. Johnston KG, Grieco SF, Nie Q. *et al.* Small data methods in omics: The power of one. *Nat Methods* 2024;21:1597–1602. https://doi.org/10.1038/s41592-024-02390-8
6. Hie B, Cho H, DeMeo B. *et al.* Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell Syst* 2019;**8**:483–493.e487. https://doi.org/10.1016/j.cels.2019.05.003
7. Satija R, Farrell JA, Gennert D. *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**:495–502. https://doi.org/10.1038/nbt.3192
8. Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15. https://doi.org/10.1186/s13059-017-1382-0
9. Do VH, Elbassioni K, Canzar S. Sphetcher: Spherical Thresholding improves sketching of single-cell transcriptomic heterogeneity. *iScience* 2020;**23**:101126. https://doi.org/10.1016/j.isci.2020.101126
10. DeMeo B, Berger B. Hopper: A mathematically optimal algorithm for sketching biological data. *Bioinformatics* 2020;**36**:i236–41. https://doi.org/10.1093/bioinformatics/btaa408
11. Johnson ME, Moore LM, Ylvisaker D. Minimax and maximin distance designs. *J Stat Plan Inference* 1990;**26**:131–48. https://doi.org/10.1016/0378-3758(90)90122-B
12. Song D, Xi NM, Li JJ. *et al.* scSampler: Fast diversity-preserving subsampling of large-scale single-cell transcriptomic data. *Bioinformatics* 2022;**38**:3126–7. https://doi.org/10.1093/bioinformatics/btac271
13. Baskaran VA, Ranek J, Shan S. *et al.* Distribution-based sketching of single-cell samples. In: *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. New York, NY, USA: Association for Computing Machinery, 1–10, 2022. https://doi.org/10.1145/3535508.3545539
14. Lotfollahi M, Naghipourfar M, Luecken MD. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nat Biotechnol* 2022;**40**:121–30. https://doi.org/10.1038/s41587-021-01001-7
15. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32. https://doi.org/10.1023/A:1010933404324
16. Efron B. Jackknife-after-bootstrap standard errors and influence functions. *J R Stat Soc Series B Stat Methodology* 1992;**54**:83–111. https://doi.org/10.1111/j.2517-6161.1992.tb01866.x
17. Kwon Y, Zou J. Data-oob: Out-of-bag estimate as a simple and efficient data value. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J (eds.), *International Conference on Machine Learning*. Honolulu, Hawaii, USA: JMLR.org, 18135–52, 2023.
18. Balinski ML, Young HP. *Fair Representation: Meeting the Ideal of One Man, One Vote*. 2nd edition. Washington, DC, USA: Brookings Institution Press, 2001.
19. 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. https://doi.org/10.1038/s42003-023-05634-z
20. De Donno C, Hediyeh-Zadeh S, Moinfar AA. *et al.* Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat Methods* 2023;**20**:1683–92. https://doi.org/10.1038/s41592-023-02035-2
21. Fischer F, Fischer DS, Mukhin R. *et al.* scTab: Scaling cross-tissue single-cell annotation models. *Nat Commun* 2024;**15**:6611. https://doi.org/10.1038/s41467-024-51059-5
22. Chen J, Xu H, Tao W. *et al.* Transformer for one stop interpretable cell type annotation. *Nat Commun* 2023;**14**:223. https://doi.org/10.1038/s41467-023-35923-4
23. Qiu C, Martin BK, Welsh IC. *et al.* A single-cell time-lapse of mouse prenatal development from gastrula to birth. *Nature* 2024;**626**:1084–93. https://doi.org/10.1038/s41586-024-07069-w
24. Hu C, Li T, Xu Y. *et al.* CellMarker 2.0: An updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* 2022;**51**:D870–6. https://doi.org/10.1093/nar/gkac947
25. Elmentaite R, Kumasaka N, Roberts K. *et al.* Cells of the human intestinal tract mapped across space and time. *Nature* 2021;**597**:250–5. https://doi.org/10.1038/s41586-021-03852-1
26. James KR, Gomes T, Elmentaite R. *et al.* Distinct microbial and immune niches of the human colon. *Nat Immunol* 2020;**21**:343–53. https://doi.org/10.1038/s41590-020-0602-z
27. Dominguez Conde C, Xu C, Jarvis LB. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* 2022;**376**:eabl5197. https://doi.org/10.1126/science.abl5197
28. 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. https://doi.org/10.1016/j.cell.2023.11.026
29. Wang X, Park J, Susztak K. *et al.* Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* 2019;**10**:380. https://doi.org/10.1038/s41467-018-08023-x
30. Nguyen H, Nguyen H, Tran D. *et al.* Fourteen years of cellular deconvolution: Methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Res* 2024;**52**:4761–83. https://doi.org/10.1093/nar/gkae267
31. Cormen TH, Leiserson CE, Rivest RL. *et al. Introduction to Algorithms*. Cambridge, Massachusetts: MIT press, 2009.
32. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
33. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559. https://doi.org/10.1186/1471-2105-9-559
34. Altman N, Krzywinski M. The curse (s) of dimensionality. *Nat Methods* 2018;**15**:399–400. https://doi.org/10.1038/s41592-018-0019-x

Briefings in Bioinformatics, 2025, **26(3)**, bbaf279
https://doi.org/10.1093/bib/bbaf279
Problem Solving Protocol