Comparative Analysis of Dimension Reduction Methods for Cytometry by Time-of-Flight Data

Kaiwen Wang[1], Yuqiu Yang[1,2], Fangjiang Wu[2], Bing Song, Ph.D.[2], Xinlei Wang, Ph.D.[1,*], Tao Wang, Ph.D.[2,3,*]

1 Department of Statistical Science, Southern Methodist University, Dallas, TX, USA, 75275.
2 Quantitative Biomedical Research Center, Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA, 75390.
3 Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX, USA, 75390.
* Corresponding authors

Short title: Benchmarking Dimension Reduction Methods for CyTOF Data

Key words: CyTOF, scRNA-seq, proteomics, dimension reduction, benchmark

ABSTRACT

While experimental and informatic techniques around single cell sequencing (scRNA-seq) are much more advanced, research around mass cytometry (CyTOF) data analysis has severely lagged behind. However, CyTOF profiles the proteomics makeup of single cells and is much more relevant for investigation of disease phenotypes than scRNA-seq. CyTOF data are also dramatically different from scRNA-seq data in many aspects. This calls for the evaluation and development of statistical and computational methods specific for analyses of CyTOF data. Dimension reduction (DR) is one of the most critical first steps of single cell data analysis. Here, we benchmark 20 DR methods on 110 real and 425 synthetic CyTOF datasets, including 10 Imaging CyTOF datasets, for accuracy, scalability, stability, and usability. In particular, we checked the concordance of DR for CyTOF data against scRNA-seq data that were generated from the same samples. Surprisingly, we found that a less well-known method called SAUCIE is the overall best performer, followed by MDS, UMAP and scvis. Nevertheless, there is a high level of complementarity between these tools, so the choice of method should depend on the underlying data structure and the analytical needs (*e.g.* global *vs* local preservation). Based on these results, we develop a set of freely available web resources to help users select the best DR method for their dataset, and to aid in the development of improved DR algorithms tailored to the increasingly popular CyTOF technique.

BACKGROUND

Recently developed single-cell profiling technologies hold the promise to provide critical insights into the cellular heterogeneity in tissues of various biological conditions, developmental trajectories of single cells, and how cells communicate with each other. Researchers in the field of genomics and proteomics have separately developed single cell profiling technologies, mainly single cell RNA-sequencing (scRNA-seq) and mass cytometry (CyTOF)[1–4], and their variants. Compared with scRNA-seq, CyTOF profiles the proteomics makeup of the single cells and are more relevant for understanding clinical phenotypes than scRNA-seq. CyTOF can additionally capture post-translational modifications[5], which is beyond the reach of scRNA-seq. Also, CyTOF is becoming increasingly popular, at an exponential rate similar to scRNA-seq (**Sup. Fig. 1**).

There are key differences between CyTOF and scRNA-seq data which makes it questionable to directly translate scRNA-seq analytical methods for use on CyTOF data. For example, scRNA-seq usually profiles several thousand genes' RNA expression for several thousand cells, while CyTOF can profile up to 120 proteins' expression for potentially 10 to 100 times more cells. The large number of cells enable CyTOF to capture the rare populations that may be missed by scRNA-seq data. CyTOF data uniquely suffers from time-dependent signal decay during the generation of data by the CyTOF/Helios machine. Furthermore, CyTOF data are often in a continuous scale while scRNA-seq data are inherently integer counts. CyTOF is mostly free

from the drop-out issue that is commonly seen in scRNA-seq data[6–8]. On the other hand, more and more studies are now trying to generate matched CyTOF and scRNA-seq data for the single cells from the same research subjects[9–11]. These two data types are of different nature, and thus are highly complementary with each other. Therefore, analysis processes should verify that findings and conclusions generated from combining CyTOF and scRNA-seq data are consistent/concordant.. Unfortunately, these issues have not been addressed thoroughly by researchers, resulting in difficulties for valid interpretation of CyTOF data.

The first step of single cell data analysis is data exploration, which is usually achieved through Dimension Reduction (DR), and followed by visualization, clustering, cell type assignment, differential expression analyses, *etc*. Proper DR is fundamental to all downstream analyses. For example, if DR incorrectly places some cells in the wrong place in the reduced space, these cells could be identified as a "new" differentiation stage of another irrelevant cell type by mistake, rather than being identified with their correct cell types. Traditional linear DR algorithms such as Principal Component Analyses have existed for decades. With the rise of scRNA-seq, numerous new DR algorithms, such as tSNE[12], UMAP[13], SAUCIE[14], ZIFA[15], PHATE[16], scvis[17], and Diffusion map[18], have been developed to capture the complicated non-linear relationships in the high-dimensional data. However, it's unclear which of these DR methods is the best for CyTOF, and the best DR method for scRNA-seq data may not necessarily extrapolate to CyTOF.

To fill in this void, we reviewed 22 DR methods (**Sup. Table 1**) and systematically compared the performances of 20 DR methods for CyTOF, based on 110 real and 425 simulated datasets. This is the first study that has comprehensively reviewed and benchmarked DR methods for CyTOF data. We also included 10 Imaging CyTOF/mass cytometry (IMC) datasets[19], which is an expansion of mass cytometry that enables the capturing of the spatial information of the cells (**Sup. Fig. 2**). Overall, our results rebut the common thinking in the field that the best methodologies developed for scRNA-seq data are also optimal for CyTOF data. We found significant complementarity between DR tools, and we postulate that the choice of method should depend on the underlying data structure and the analytical needs. Our evaluation highlights some challenges for current methods, and our evaluation strategy can be useful to spearhead the development of new tools that accurately perform DR for CyTOF data.

RESULTS

Overall study design

We collected a total of 22 DR methods published by various researchers (**Sup. Table 1**, and **Fig. 1a**), including those developed specifically for single cell-type of data and also general purpose DR methods. 20 of these methods were practically executable, and we benchmarked them on a total of 110 real CyTOF datasets from various sources (**Fig. 1b**, and **Sup. Table 2**), including both peripheral blood and solid tissues, and also 425 simulated CyTOF datasets of diverse

characteristics, using our Cytomulate algorithm (**Fig. 1c**).

For scoring the accuracy of the DR methods, we chose a total of 16 metrics in 4 main categories, characterizing different aspects of the performances of the DR methods. These 4 main categories are (1) global structure preservation, (2) local structure preservation, (3) downstream analysis performance, and (4) concordance of DR results with matched scRNA data. We provided detailed description of these metrics in **Sup. Table 3** and **Sup. File 1**. Some of the most important metrics were visually demonstrated in **Fig. 1d**. In our study, the DR methods were assessed and ranked mainly by these metrics focusing on the accuracy. But we also assessed and reported their scalability with respect to the number of cells and protein markers, stability of the predictions after subsampling the datasets and also parameter tuning; and the usability of the tools in terms of software and documentation (**Fig. 1a**, and **Sup. Table 3**). More details about each assessment criterion will be given later.

The data compendium consisted of both real datasets (**Sup. Table 2**), which provide the highest biological relevance, and also synthetic datasets, which offer the most comprehensive and controllable coverage of different characteristics of CyTOF data. These real datasets come from a variety of organisms, dynamic processes, and types of trajectory topologies. In particular, while CyTOF has been applied for peripheral blood by immunologists mostly, we included CyTOF data from solid tissue experiments. It is of interest to compare whether solid tissue and peripheral blood CyTOF data possess any differences that would render different DR methods more suitable. In our real datasets, we also included one Imaging CyTOF dataset created on breast cancer tissues (the BC cohort of **Sup. Table 2**). For simulation of CyTOF data, we developed Cytomulate (**method section**), the first formal simulation tool that is able to well mimic the behaviors of CyTOF data. One example simulated CyTOF dataset generated by Cytomulate is shown in **Sup. File 2**. Cytomulate is general-purpose, while prior works, like LAMBDA[20], have created specialized procedures to generate simulated CyTOF data in the context of model-based clustering. These simulation procedures cannot be used to test methods developed for other purposes (*e.g.* DR). For each simulation, we used a real dataset as a reference, to match its dimensions and other statistical properties.

We developed the CyTOF DR Playground webserver to help future users of DR methods visualize our benchmark results of these 535 CyTOF datasets, and choose the optimal DR methods for their own datasets (**Fig. 1e**). This app allows users to query the results of this evaluation study, including filtering of datasets and customized selection of evaluation metrics for the final ranking. We also offer our unified pipeline for homogenized input/output of the DR methods and the benchmark metrics (**Fig. 1e**, CyTOF DR Package), so that future users can easily execute *ad hoc* deployment of DR and the benchmark metrics for their own data. As another important deliverable of our work, we also packaged our CyTOF data simulation algorithm into Cytomulate, and shared it through cytomulate.readthedocs.io (**Fig. 1e**) for

researchers to perform other CyTOF-related research.

Accuracy of DR methods

We found that the accuracy performances of the DR methods were very variable across datasets, indicating that there is no 'one-size-fits-all' method that works well on every dataset (**Fig. 2** and **Fig. 3**; the meanings of the accuracy metric abbreviations provided in **Sup. Table 3**). UMAP is one of the best DR methods for scRNA-seq data, and it is good to see that UMAP also performed well on CyTOF data. But it is somewhat surprising to see that less well-known or less used methods, such as SAUCIE and MDS, out-perform UMAP on many datasets (Wilcoxon Signed Rank Tests on overall accuracy: SAUCIE-UMAP $FDR < 10^{-15}$, and MDS-UMAP $FDR < 10^{-15}$). tSNE and DiffMap, which are also popular scRNA-seq DR methods, perform only moderately on CyTOF data. To rule out the potential bias in the different numbers of datasets from each of the 11 cohorts (**Sup. Table 2**), we aggregated the accuracy scores of the DR methods for the datasets from the same cohorts. We still observed the same variable performances of the DR methods across different CyTOF data cohorts (**Sup. File 2**). For the Breast cancer (BC) Imaging CyTOF datasets, SAUCIE seems to be the unequivocally best performer ($FDR < 0.003$ for all Wilcoxon Signed Rank tests between SAUCIE and other methods). On the Brain CyTOF dataset, where the single cells were extracted from solid tissues, the best performer appears to be scvis with SAUCIE and other methods closely following (Wilcoxon Signed Rank tests: $FDR = 1.0$ between scvis and SAUCIE; $FDR = 0.38$ between scvis and UMAP).

We also examined the different categories of the accuracy metrics, and identified even stronger complementarity between the DR methods (**Fig. 3b**). For the purpose of this analysis, the performances of the metrics on all datasets were averaged. We found that SAUCIE tends to perform generally well on most categories. But MDS seems to be the best in preserving local and global structures. Scvis and UMAP are better at downstream analysis needs involving clustering and cell type assignment, and also achieving concordance with the matched scRNA-seq data. PHATE also has decent performances on many metrics, but not including global structure preservation. tSNE has a better performance than SAUCIE in terms of achieving concordance with the matched scRNA-seq data. There have been some hot debates regarding the global and local structure preserving capabilities of tSNE and UMAP[21,22]. We found, in the context of CyTOF data, that UMAP seems to be better than tSNE at preserving global structures (Wilcoxon Signed Rank Test $p < 10^{-15}$) and that they seem to be un-distinguishable in terms of local structure preserving performance.

To more accurately evaluate how the performances of the DR methods vary as a function of the characteristics of the CyTOF datasets, we generated 425 simulated CyTOF datasets with the Cytomulate tool. We varied the numbers of cells, the numbers of protein markers, the numbers of

independent cellular trajectories (#trees), and the number of cell types in the simulated CyTOF datasets. We benchmarked the accuracy of the DR methods on these simulated data with the same criteria (**Fig. 3c**). MDS, SAUCIE, and scvis are again the top methods, followed by PHATE and UMAP. MDS and scvis (subsampling) performs more poorly with small numbers of input cells, while SAUCIE, scvis (no subsampling), PHATE and UMAP tend to work better with small numbers of input cells (ANOVA Type III $p < 0.0001$ for interaction effect between number of cells and methods). Scvis (subsampling or no subsampling), PHATE and UMAP tend to work better with datasets of more protein features ($p = 0.0627$). These methods' performances are relatively stable with respect to the number of protein markers in the CyTOF data and the number of independent cellular differentiation trajectories. Curiously, MDS, scvis, PHATE and UMAP seem to perform better when there are more cell types in the CyTOF datasets ($p < 0.0001$).

The high variability of the DR methods between datasets of diverse characteristics could indicate some complementarity between them (**Fig. 2b-d**). To investigate this, we calculated the likelihood of obtaining a top DR result when using more than one DR methods. In the real CyTOF data (**Fig. 2b**), we observed that using a single method (SAUCIE) for DR of CyTOF data can only generate "top" results ~35% of the times. Adding a second method can guarantee a top result ~65% of times. Considering all top 5 methods together can lead to a close to 100% chance of obtaining a top result. Analysis on the simulated CyTOF data (**Fig. 2c**) and the combination of all real and simulated CyTOF data (**Fig. 2d**) also revealed strong complementarity between these different DR methods. Altogether, our results indicate that users should try out a diverse set of methods on their data, and method developers should consider the possibilities for new ensemble methods that utilize this complementarity.

Scalability of DR methods

While many early DR methods were developed at a time when analyzing more than a thousand data points was exceptional, these methods now have to cope with half a million cells from CyTOF data, and maybe soon with even more. To assess their scalability, we ran each method on up- and downscaled versions of the Oetjen dataset[23], where the maximum number of cells is one million (**Fig. 4ab**). Runtime (**Fig. 4a**) and memory usage (**Fig. 4b**) seem to be highly correlated overall (Spearman's $\rho = 0.64$). In terms of both runtime and memory usage, we found that the scalability of most methods was overall very poor. Comparatively, SAUCIE, FA, ICA, PCA, NMF are most favorable in terms of runtime and memory usage in general. UMAP and tSNE are in the next group of methods with slightly less efficient–but reasonable–runtime and memory usage characteristics. The other methods are in the third tier with up to ~100 times more consumption of runtime and memory usage. The runtimes of these methods in our various experiments can be as much as 12 hours and the memory usage can be as large as 700GB. Given the multitudes of implementations and optimizations of tSNE[12,24–26], we tested and observed that

different implementations can have as much as a 10-fold speed up for large samples (**Fig. 4a**). Specifically, our reference implementation with fast fourier transform (FFT) as implemented by openTSNE (named tSNE in this work unless otherwise noted), which is a relatively recent development, vastly outperforms the standard Sklearn version using the Barnes-Hut algorithm as sample sizes grow.

Due to the scalability issue of many of these methods, we recommend subsampling the whole CyTOF datasets for some of these DR methods. In fact the original publications of several of these DR methods also recommend subsampling. For the purpose of observing the overall cell state space of the cells (such as major cell populations and their relationships), subsampling, up to a certain extent, will probably not negatively impact the conclusions to be drawn. But all downstream analyses will have to be limited to the subsampled cells and rare cell populations may be missed in the down-sampled subset, which is the detrimental effect of subsampling. So whether to adopt subsampling or not also depends on the purpose of the analyses. For example, MDS, which has good accuracy performance, scales poorly with the number of cells (**Fig. 4ab**). We were not able to finish processing most of the experiments with large numbers of cells for MDS. Thus, in all our other benchmark studies, we applied MDS on 10% sub-sampled CyTOF data if they abort on original samples due to memory constraints. Scvis also scales poorly with numbers of cells in terms of runtime. But its runtime plateaus with large numbers of cells and never exceeds 12 hours. Curiously, scvis's memory consumption is low, and very comparable with the first tier of methods (SAUCIE, *etc*). Therefore, in our all benchmark studies, we evaluated scvis on both the full CyTOF datasets and also down-sampled datasets.

Due to the many different implementations and variations of tSNE, we also specifically focused on these tSNE methods. We found that the Barnes-Hut (BH) tSNE variants show much poorer performance than the FFT variants from openTSNE and fitSNE, in terms of computation time (**Sup. Fig. 4**). Surprisingly, the BH implementation of openTSNE outperforms all other implementations when sample sizes are small, while falling not far behind the FFT tier for large samples. With the large sample sizes of CyTOF data, it makes sense to always use FFT-based tSNE, but BH still offers some value for small samples, especially given that the former implementation is not as easily accessible in some programming environments. The memory usage is similar among these tSNE implementations, but still with several folds of differences. This indicates that the specific implementations of the DR algorithms can have a dramatic effect on the performances of the methods.

In the above experiments, we have a total of 34 protein markers in the CyTOF data (from the original Oetjen dataset), which is typical for most current CyTOF datasets. In consideration of the possibility that future iterations of the CyTOF technology could incorporate more protein markers, we up-sampled the Oetjen dataset to create up to 60 protein marker genes while fixing the cell number at 100,000. Then we performed the same benchmark study in terms of runtime

(**Fig. 4c**) and memory usage (**Fig. 4d**). Our benchmark analyses show that most methods do not vary too much with respect to the number of protein markers in the CyTOF datasets, alleviating concerns for increased runtime/memory usage for future CyTOF data with more protein markers.

Finally, we also noticed, during our DR benchmark practices above, that some of these DR methods will abort with errors due to increasing sample sizes. To quantify the sample size threshold, we recorded whether the DR methods would abort while performing the scalability benchmark. Specifically, the methods that produced errors in **Fig 4a** and **Fig 4b** were recorded in **Fig. 4e** and **Fig. 4f.** Since sample dimension and the scalability are of interest, we did not allow subsampling for any method here. Our analyses show that Diffmap, Isomap, KPCA Poly/RBF, LLE, MDS, and Spectral start to abort with errors with an increasing number of cells, whereas GrandPrix often fails to converge (**Fig. 4e**). For those methods, we had to subsample real and simulation datasets during other benchmarks of this study when necessary. Our analysis for error occurrences in this particular dataset and its variations show that Spectral and GrandPrix sporadically abort (**Fig. 4f**), and that MDS and Diffmap were unable to complete any of the runs due to their inherent sample size limitations.

Altogether, the scalability analysis indicated that the number of cells, but not number of protein markers, in the CyTOF data is an important factor in the choice of method. But unfortunately, CyTOF data inherently profiles more cells than scRNA-seq data. Therefore, method developers should pay more attention to maintaining reasonable running time and memory usage, when considering CyTOF data.

Stability of DR methods

It is not only important that a method is able to infer an accurate model in a reasonable time frame, but also probably important that it produces a similar model when given very similar input data and input parameters. To test the stability of each method, we executed each method on 100 bootstrap samples of the Oetjen cohort's Sample A, and calculated the stability of the DR methods, across different bootstrap datasets. Stability is calculated by first obtaining the variability of the rankings of the DR methods, in terms of the accuracy metrics, across the bootstrap samples. We then ranked all methods according to variability, with smaller variability suggesting good rank-order stability and thus obtaining better stability scores. We found that the top methods that have the best stability performances are KPCA, NMF, ZIFA, and FA (**Fig. 5a**). The DR methods that have the best accuracy (**Fig. 2**) unfortunately do not possess the best stability performances. Comparatively, UMAP and scvis are more stable. Interestingly, the stability of tSNE largely depends on the specific implementation. MDS, PHATE and SAUCIE have relatively less stability, suggesting future improvements to these methodologies are needed from method developers in this regard. We also calculated the stability of the DR methods, with respect to each accuracy metric category, and found a high level of complementarity between the

DR methods. Scvis and PHATE are more stable for global structure preservation; MDS and tSNE are more stable for local structure preservation; UMAP and scvis are more stable for downstream analysis needs; and UMAP and SAUCIE are more stable for concordance with scRNA-seq data.

Another important aspect of the DR methods related to the stability of computation is how much the results will vary as a function of their tuning parameters, given the same input data. Sometimes, it may be preferable that the DR results do not vary too much even given different parameters. But in other times, the data analytics practitioners may prefer that parameter tuning will lead to more different results given different parameters, so that they can choose a parameter setting that yields the optimal DR results. While we leave the choice of preference to the practitioners, we performed a study of parameter tuning. For several top DR performers, we tuned their parameters, as is shown in **Sup. Table 4**. We included MDS, PHATE, SAUCIE, scvis, tSNE, and UMAP. We found that (**Fig. 5b**) all tuned methods benefit from careful selection of tuning parameters. tSNE and PHATE noticeably benefit the most in terms of percent rank improvement in accuracy, which is defined as the percent increase of ranks of the top configuration over the default. We observed that the default configuration oftentimes is a decent starting point, especially for scvis, UMAP, and some aspects of SAUCIE and MDS. According to our analyses, we present the optimal configuration setting for each major category of accuracy in **Sup. Table 4**, so that future users may choose to use these settings as the starting point for DR of their own CyTOF data. In general, we recommend using more training steps for SAUCIE while not changing any regularization coefficients. tSNE benefits from large perplexity, whereas using small minimum distance between points yields good results for UMAP. For MDS, using more initialization and training steps can offer some incremental improvements. We do not recommend performing extensive parameter sweep for scvis, as the potential gain is small and also because scvis already performs well.

Usability of DR methods

While not directly related to the accuracy of DR, it is also important to assess the quality of the implementation and how user-friendly the DR methods are for a user. We evaluated each method using a transparent checklist of important scientific and software development practices, including documentation, automated testing and integration, regular update, cross platform compatibility, capable of accepting general matrix input, *etc* (**Sup. Table 3** and **Sup. File 3**). As shown in **Sup. Fig. 5**, we found that most methods fulfilled the basic criteria, such as providing example datasets in documentation and being open source. However, installation issues seem to be quite general in bioinformatics and the DR field is no exception. It seems to be difficult to handle dependency issues for a number of packages, especially SAUCIE and scvis, which both depend on out-of-date tensorflow versions. Installation of ZIFA and GrandPrix requires manual intervention. Scvis is command line-only and thus cannot be flexibly invoked in R or Python,

and the official implementation of DiffMap is in R only. The other methods are available in both R and Python. Finally, we note that there are several methods that are capable of mapping new data onto existing DR embeddings (*e.g.* scvis, tSNE and UMAP), which could be very convenient in handling continuously generated data. We observed no clear relationship between usability and method accuracy (**Fig. 2a**). SAUCIE and scvis have very good performances for CyTOF data according to the accuracy criteria. But they may not be very well known and thus their authors have not devoted sufficient resources to polish their software implementations. Our work suggests that the field should consider improving the usability of these two methods and other methods as well to benefit CyTOF researchers.

A uniform pipeline for implementing CyTOF DR methods and evaluation metrics

To facilitate open and user-friendly deployment and evaluation of DR methods, we established the CyTOF DR Package website (CytofDR.readthedocs.io) to provide implementation of the DR methods and our evaluation metrics (**Fig. 1e**), which is streamlined as much as possible. As we illustrated in **Fig. 6a**, a new DR user of CyTOF data can take a prior knowledge-driven approach of examining our DR benchmark results of the datasets evaluated in this study, shared conveniently through "CyTOF DR Playground", and then run their favorite DR methods using their own installation or the implementations that we provide on "CyTOF DR Package". The user can also take an empirical approach of deploying the evaluation metrics (or a subset of our evaluation metrics according to preference or data availability) on their own CyTOF datasets, using the evaluation and DR code implementation that we offered through "CyTOF DR Package", for picking the best approach. We illustrate the usefulness of this framework here by comparing the DR results of several top-performing methods on the Levine32 dataset (**Fig. 6b**). Using our framework, this figure can be recreated using only a couple of lines of codes (shown in "CyTOF DR Package"). We also envision our system to be quite useful to future developers of DR methods for CyTOF data or other types of data in general. The extensive collection of evaluation metrics will allow future developers to craft their own benchmark recipes for DR in any field. At the same time, developers can easily extend the framework and add functionalities by simply deriving from our Python classes. We hope such a unified and streamlined evaluation-execution paradigm will become more widely adopted for various bioinformatics development scenarios.

DISCUSSION

In this work, we presented a large-scale evaluation of commonly used methods for DR of CyTOF data across many different datasets. Few attempts at tackling this issue have been made and the field seems to think in general that methodologies developed for scRNA-seq data can be directly applied on CyTOF data. This is problematic, as we showed that this assumption is not true in many cases, and there is a high degree of complementarity of DR methods according to

different characteristics of the CyTOF datasets. SAUCIE and scvis are all-rounders that perform admirably across all categories without major flaws, whereas UMAP excels in downstream analyses with clustering workflows. MDS, on the other hand, suffers from poor scalability downstream performance, but its technical advantage in both global and local structure preservation offers a compelling choice as a validation tool. In addition, we found that the performances of the DR methods on the real and the simulated CyTOF data are very comparable, indicating the validity of our simulation approach as well as evaluation results.

Furthermore, we find that higher standards for code assurance, documentation, and ease of use could help in adopting the DR tools, especially the top performers, across the cytometry and single-cell omics field. The combination of decent performance, great usability, and good scalability has given rise to a few popular methods, such as UMAP and tSNE. However, their success is also a testament to two common but critical issues in the field: many good methods are marred by either their usability or scalability (e.g. MDS, SAUCIE and scvis), especially for CyTOF data. Such deficiencies can be deal breakers for those who need reliable analysis pipelines. In light of these observations, we urge researchers to take end-user experiences and the increasing sample sizes of CyTOF into consideration and to develop competitive methods that shine on both theoretical performance and practical user experience.

New users to the field are confronted with an overwhelming choice of DR methods, without a clear idea of which would optimally solve their problem. Also, users could be biased by their prior experiences of DR methods in scRNA-seq data. To address this issue, we developed CyTOF DR Playground and CyTOF DR Package for users to choose the best DR method depending on the context of the CyTOF datasets being analyzed. On the other hand, some researchers seem to analyze CyTOF in the same way as traditional flow cytometry data, using commercial software developed for flow cytometry. They ignore the fact that CyTOF is a much more advanced technique that offers capturing of a lot more protein channels, but has minimum spill-over effects[27]. They also ignore all the sophisticated algorithms developed in recent years, mainly in the field of scRNA-seq data analysis, which might be applied for analyzing CyTOF data to reveal complicated biological insights that cannot be afforded by standardized commercial software. For example, it might be of interest, in another study, to evaluate whether the wealth of pseudotime inference algorithms developed for scRNA-seq data could be applicable for CyTOF data as well[28].

The fields of genomics and proteomics have been developing mostly independently. Genomics characterize the DNA- and RNA-level perturbations of the biological systems, while proteomics characterize the protein-level perturbations, which are highly complementary. The ability to combine both approaches will be extremely powerful. In recent years, there has been a growing trend to marry genomics with proteomics – proteogenomics. At the bulk tissue level, this is reflected in the generation of more and more patient cohorts that contain both genomics and

proteomics data, such as the CPTAC consortium[29–32]. At the single cell level, this trend is reflected by the growing generation of single cell datasets with both scRNA-seq and CyTOF data[9–11,33]. The creation of such matched datasets calls for the development of dedicated analytical methodologies to integrate these two types of data. First, one should ensure that the conclusions drawn from each of the proteomics and genomics modalities are overall consistent, as we reviewed in this work for concordance between CyTOF and scRNA-seq data during DR. Second, robust and innovative statistical methods should be developed for the integrative analyses of CyTOF and scRNA-seq data, to maximize the potential of each technology.

In this regard, CITE-Seq[34], which allows capture of the RNA transcriptomes and cell surface protein expression for the same cell, is unlikely to replace CyTOF (matched with regular scRNA-seq). CITE-seq can capture much less number of cells (a few thousand usually) than CyTOF. And CITE-seq can only profile the expression of protein surface markers, while CyTOF is capable of capturing the expression of both intracellular and cell surface proteins[35]. It also suffers from drop-outs in the protein expression measurements by the UMI counting approach[36,37], while CyTOF is relatively free from this caveat.

CyTOF possesses a much higher potential of being developed into a clinically applicable diagnostic/prognostic/predictive tool, compared with scRNA-seq, due to its cheaper per-cell cost, shorter workflow, and direct measurement of proteomics changes. Our study of the performances of DR methods on CyTOF data urges the field to rethink analytical strategies for CyTOF data and to develop state-of-the-art tools to address this key gap. This effort should ultimately lead to effective translational application of the CyTOF technology for personalized medicine.

MATERIALS AND METHODS

Benchmark environment

All benchmarking exercises were performed on server nodes with dual Intel Xeon E5-2695 v4 CPUs. Each node has 36 physical cores and was allocated 200 Gigabytes (GB) of RAM, except for the scalability benchmarks, where nodes with the same CPUs but 732 GB of RAM were assigned. All jobs were run in CentOS Linux and managed with Slurm with no GUI to ensure reproducibility. Each computation job was allowed to run for up to 7 days if necessary. Since no method exceeded this time limit, all errors and exceptions reported in the paper resulted from either out-of-memory errors or other unforeseen implementation reasons that we could not control. Statistical analyses, preprocessing, and other workflows for this paper were performed in R with various consumer-grade hardware running Windows.

DR method implementation details

The benchmark pipeline was implemented mostly in Python for its extensive support for DR methods and command-line operability. If available, Sklearn implementations were used, as they are not only well maintained and but oftentimes also the default choice for Python users. For

other DR methods, we utilized the implementations of the original authors if possible, with the exception of tSNE, which will be further discussed. One notable method to mention here is Diffmap since its reference implementation is available only in R. As a convenient wrapper, our pipeline and our CyTOF DR Package use all default settings for tuning parameters to ensure fairness across all methods, but multicore optimizations, if available, were enabled to utilize all cores on our server for efficiency.

"tSNE" in the main figures refers to the openTSNE implementation of FIt-SNE. Works such as Kobak *et al*[38] have called for using FIt-SNE in the genomics field, because its computation efficiency is comparable with UMAP. So we choose this as our reference implementation. tSNE (FIt-SNE Original), which is the original authors' implementation, technically offers faster speed and better memory efficiency, but it comes at a cost of usability with no discernable embedding improvement. Thus, openTSNE is a reasonable compromise that both improves upon efficiency as compared to BH and offers a user-friendly interface with good documentation. We also included "tSNE (sklearn)" in our main figures and benchmark study as it is one of the most popular choices with BH implementation. Other methods included in **Sup. Fig. 4** serve as an scalability-only benchmark for users to understand the differences between the various choices.

For popular as well as top methods from our benchmark, we identified their tuning parameters and performed extensive parameter tuning. We list all the parameters in **Sup. Table 4,** and for each method, we performed a grid search of all combinations. SAUCIE has the most parameters for both regularization (Lambda C and Lambda D) and optimization (Steps, Learning Rate, and Batch Size), yielding a staggering 1125 total configurations. Yet, its documentation and publication have little guidance on whether parameter tuning is necessary. On the other hand, the authors of scvis have already tuned parameters and found that scvis is not sensitive to changes of settings. We thus validated their claims by tuning regularization coefficient, perplexity, and the number of neurons on layer 1 and 2. For other methods, we selected parameters based on either previous recommendations on settings, such as the debates between the usage of UMAP and tSNE for single-cell data, and their unique features.

The Downstream and scRNA-seq concordance categories of our DR evaluation framework require clustering of CyTOF and scRNA-seq data. We used FlowSOM[39] for CyTOF as its efficiency and accuracy have been previously demonstrated[40]. The number of clusters for each dataset was chosen based on either previous knowledge, such as provided cell types, or empirical criteria using cluster variance and the elbow method. In general, we over-clustered so that we do not lose resolution on rare cell populations, while at the same time the meta-clustering step of FlowSOM ensures a reasonable number of clusters. scRNA-seq data, on the other hand, require a different approach. Thus, we utilized Seurat's[41] clustering and preprocessing pipeline since it is the industry standard.

Accuracy

Our main evaluation framework assesses the accuracy of each DR method's embeddings. The framework consists of four major categories: global structure preservation (Global), local structure preservation (Local), downstream analysis performance (Downstream), and scRNA-seq concordance. Downstream and scRNA-seq Concordance further have sub-categories to reflect different, equally-weighted aspects of the major categories. For each method, we utilized the original space expression, the reduced space embedding, clusterings before and after DR, assigned cell types, and, if available, the corresponding scRNA-seq data to execute each individual evaluation metric under major and sub-categories. The details and implementations of these methods are discussed in **Sup. File 1.**

For each dataset, we first ranked all the methods based on each individual metric. When appropriate, methods are reverse ranked. Methods that aborted for whatever reason are ranked last for all metrics because they were unable to produce an embedding at all. We then averaged the ranks of metrics under each sub-category if applicable. The sub-category scores or the ranks of metrics themselves in the case of Global and Local were averaged to form the score of each major category. This weighting scheme ensures that all major categories have the same weight and are fairly represented regardless of how many sub-categories and metrics they have. For accuracy of each method, we averaged the scores of each major category. The exact weight of each individual metric as part of the major category is detailed in **Sup. Table 3.**

Given the rank-based nature of accuracy scores, we compared and averaged scores across datasets from different cohorts. The real and simulation accuracy scores are averaged across all real and simulation datasets respectively (**Fig. 2a**), whereas the overall accuracy is an average of all datasets used in this study. **Sup. File 2 Fig. 1** presents accuracy scores averaged within each cohort. For accuracy-related figures (**Fig. 3** and **Sup. Fig. 3**), we ordered the methods according to the overall accuracy in a descending order.

Based on accuracy, we defined complementarity as the likelihood obtaining a most accurate method if practitioners choose a subset of methods. In practice, we tallied the top-ranking methods for all real and simulation datasets and counted the number of times they were the top performer in terms of accuracy. Then, we calculated the observed proportion of each method being on top by dividing the number of times that each method ranked on top by the total number of datasets. A large likelihood means that the method is more often the best as compared to others, while 0 means that it was never the best in our benchmarks. We selected seven most accurate methods according to overall accuracy (**Fig. 2a**) and showed the increasing likelihood of obtaining a top method for a given dataset by adding each of the seven methods one by one in the order of overall accuracy (**Fig. 2b-d**).

Stability

A stable method should offer consistent run-to-run performance. Specifically, given similar datasets, the rank-order stability of methods is of crucial importance for practitioners. It is not

uncommon for a cohort to have numerous samples with the same disease or tissue, and practitioners have to rely on certain chosen DR methods in their pipeline. In this context, we define rank-order stability as whether a certain method's ranking in relation to other methods is stable. To simulate this condition while minimizing random noise from different samples, we bootstrapped Sample A of the Oetjen cohort 100 times by sampling 50% of the original cells with replacement each time. We then benchmarked all DR methods on the 100 bootstrapped samples (N=147570) using the same methodology and configurations as in accuracy benchmarks. Some methods, such as KPCA and Spectral, require further down-samping of the bootstrapped samples due to scalability issues. We ranked all methods using our standard DR evaluation framework for accuracy and calculated the sample variance of each method's ranks across all 100 samples. We then reverse ranked the methods using the sample variance as our metric for stability: methods with smaller variance have higher ranks to indicate good rank-order stability and vice versa. DiffMap reported errors on all runs of this benchmark, likely due to the resampling scheme. Thus, it ranked last in terms of stability.

Scalability

We measured each method's runtime and memory consumption to assess scalability in a two-faceted experiment. To best assess performance on real datasets in a controlled manner, we used Sample B and H of the Oetjen cohort with 34 protein markers as the basis. Then, we employed random up- and down-sampling to comprehensively simulate both the number of cells and the number of different protein markers. By fixing the number of markers at 34, we subsampled without replacement 2000, 5000, 8000, 20000, 30000, 40000, 150000, 10000, 50000, 100000, 200000, 300000, 500000, 750000, and 1000000 cells. For the number of protein markers, we sampled 10, 20, 30, 40, 50, and 60 features while fixing the number of randomly sampled cells at 100000. In the latter case that up-sampling is needed, we utilized the original 34 markers as is and then randomly sampled the remaining markers from these 34 markers with some added random noise. No further method-based subsampling was allowed in this benchmark because doing so would defeat the purpose of assessing scalability fairly across all methods. For methods that would abort at sample size of 100000, which is a reasonable, if not conservative, representation for a typical CyTOF sample, we could only report error in the case of protein marker benchmark (**Fig. 4f).**

We performed all scalability benchmarks on the server with the same hardware configuration as detailed in the "Benchmark Environment" section. Runtime and memory usage were measured on the Operating System level with the "/usr/bin/time -v" command in Linux. We used the "Elapsed (wall clock) time" as the measurement of runtime and "Maximum resident set size" as memory usage. Each measurement consists of the benchmark for one method on one dataset configuration, and all runs were sequential to allow methods to utilize as much system resource as possible without aborting. All other variables, such as file input and output, were kept the same across all runs and their impact was overall negligible. We then calculated and reported

runtime in minutes and memory in GB for convenience.

Usability

We defined usability with 14 criteria in five categories related to code quality and user experience. For each method, we manually inspected the quality of documentation, level of software support, the interface available to users, and implementation details that would benefit practitioners. All information is gathered from their websites, GitHub repositories, in-software docstrings and examples, and our own benchmark processes. The exact criteria are defined in **Sup. File 3**. We scored all methods using a binary scoring system: if a criterion is met, the method receives a score of 1; otherwise, it receives 0. Under circumstances where ambiguity may occur, such as whether the package has been updated as needed, we generally give such edge cases the benefit of the doubt unless clear incompatibility or lack of support are present. The final ranking of usability is based on the sum of scores received by each method (**Sup. Fig. 5**). On the CyTOF DR Playground and **Fig. 2**, we ranked the usability scores for consistency with other metrics, where higher ranks correspond to better usability and the maximum is used in case of ties.

Real CyTOF datasets used in this study

We accessed a total of 11 cohorts of 110 real CyTOF datasets for this benchmark study (**Sup. Table 2**). These include the TuPro cohort[9], the Oetjen cohort[23], the CyAnno cohort[42], the LG cohort[43], the Brain cohort[44], the BC cohort[45], the Levine32 cohort[46], the Levine13 cohort[46], the Samusik cohort[47], the Lung cancer cohort[48], and the Covid cohort (CyTOF_00000000000001 to CyTOF_00000000000012 on DBAI). For all datasets, we either obtained their cell typing results from the original publications, or performed cell typing ourselves based on clustering of the cells followed by manual assignment. We showed the manual cell typing results of the Lung cancer cohort, in **Sup. Fig. 6**, as an example. Matched scRNA-seq data are available for the TuPro and Oetjen datasets. For the BC dataset, there were too many samples. To avoid biasing our benchmarking results, we selected the top 14 datasets with more than 5,000 cells.

For all samples, we deployed our standardized preprocessing pipeline (included as part of CyTOF DR Package) unless they have been preprocessed beforehand by the original authors. Our pipeline includes identifying lineage channels, ArcSinh transformation, gating to remove debris using instrument channels, and bead normalization. In general, we did not perform cross-batch normalization to avoid any bias introduced by such algorithms. For the case that the data have been previously processed, such as the CyAnno cohort, we manually inspected the data and applied specific steps of the pipeline on an as-needed basis: if cell types have already been assigned, we did not gate to remove debris; when not all bead channels were available, we assumed that bead normalization had already been performed on instrument. For all samples, we cross-checked preprocessing results with 2D embeddings and cell types.

CyTOF data simulation (Cytomulate)

We devised a probabilistic simulation model that captured the key features of the $arcsinh(./5)$ -transformed CyTOF data. Suppose that the desired resulting dataset contains $m = 1, \cdots, M$ protein markers, $n = 1, \cdots, N$ cell events, and $p = 1, \cdots, P$ cell types. Let the probability of observing the pth cell type be π_p. Then, for the nth cell event, we can construct a cell type indicator variable $\Pi_n \sim Categorical(\pi)$, where $\pi = [\pi_1, \cdots, \pi_P]$ is a P-by-1 probability vector.

Given the cell type p for the nth cell event, we assume that, without cell differentiation or background noise, the expression level for the mth marker is $X_m \sim N(\mu_m^p, (\sigma_m^p)^2)$. Since a cell type typically expresses a set of marker genes, to further mimic the characteristics of CyTOF data, we associate, with each cell type p, a set $A^{\phi,p} = \{m\colon the\ mth\ marker\ is\ not\ expressed\ in\ cell\ type\ p\}$. Then, given $A^{\phi,p}$, we set $\mu_m^p = \sigma_m^p = 0, \forall m \in A^{\phi,p}$.

To simulate cellular lineage, we assume that the underlying cell lineage is composed of several non-overlapping cell differentiation trees where cell type p differentiates into z^p "children" $p^{c_1}, \cdots, p^{c_{z^p}}$. From the mth marker of each cell type p to the mth marker of each of its "children" p^c, we construct a differentiation path $G_m^{p \to p^c}: [0, 1] \mapsto R$, such that $G_m^{p \to p^c}(0) = 0$, and $G_m^{p \to p^c}(1) = \mu_m^{p^c} - \mu_m^p$, where μ_m^p and $\mu_m^{p^c}$ are the mean expression levels of the mth marker of cell type p and cell type p^c, respectively. We also assume that $G_m^{p \to p^c}$ is a stochastic process, whose realization is denoted as $g_m^{p \to p^c}$.

With these setups, the simulated expression level for the nth cell event was calculated following these steps: (1) Generate a cell type p from $Categorical(\pi)$. (2) If the cell type p has "children", we randomly select from them a cell type p^c. (3) For each marker m, we generate a sample x_m from $N(\mu_m^p, (\sigma_m^p)^2)$, a "pseudotime" t_m^p from $Beta(0.4, 1)$, and a background noise e from $N(0, \sigma^2)$. (4) The final expression level for the mth marker is then calculated as $Y_m^p = x_m + g_m^{p \to p^c}(t_m^p) + e$.

An example simulated CyTOF dataset generated by Cytomulate was showcased in **Sup. File 2**, which demonstrates that Cytomulate-generated CyTOF data closely mimic real CyTOF data.

Development of the CyTOF DR playground webserver

We developed an online web tool for displaying the results of benchmarking DR algorithms on CyTOF data. The technical stack includes HTML, CSS, JavaScript and DataTables. The visualization table to display the benchmarking results of the real and the simulation datasets in this web tool was created with the DataTables jQuery plugin. This table provides searching, sorting and pagination features and progressive bars to visualize the performances of the DR methods according to different metrics. There are two main features in this web tool. The first one is to search for DR results based on multiple input conditions (left column). And the other is to dynamically calculate the average of the performance scores for the user-chosen metrics and to rank the chosen DR methods based on the average scores (top row). This web tool is hosted on the Database for Actionable Immunology (DBAI) website (https://dbai.biohpc.swmed.edu/)[36,37,49].

Statistical analysis

All computations were performed in the R and Python programming languages. For all boxplots appearing in this study, box boundaries represent interquartile ranges, whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range, and the line in the middle of the box represents the median. All statistical tests were two-sided unless otherwise described. We used nonparametric Wilcoxon Signed Rank Test with "wilcox.test" and "pairwise.wilcox.test" procedures in R for testing location differences in samples. All pairwise tests' p-values were adjusted with the Benjamini-Hochberg Procedure to control the false discovery rate (FDR), and the reported values were upper bounds as observed from all tests. For correlation, we used Spearman's rank-based correlation coefficient.

To test interaction effects between different characteristics of the samples and DR methods, we fitted linear models with accuracy as the outcome. We then conducted Type III analyses and reported F test results for interaction terms. This part of the analysis was performed using PROC GLM in SAS. In simulation, we could control the exact characteristics of the dataset: thus our model included DR methods, number of differentiation trajectories, number of cell types, number of features, and cell numbers. All were treated as factors, and only interaction terms with DR methods were included.

Effects of the tSNE implementation and sample size were tested using a linear model with the following specification in general (more one-hot encoded terms present when more than two methods are present):

$$E(log(RunTime)) = \beta_0 + \beta_1 log(sample\ size) + \beta_2 Method + \beta_3 Method \times log(sample\ size).$$

In the case of **Fig. 4** discussions, the model is built with only tSNE and tSNE (sklearn), whereas for **Sup. Fig. 4,** all tSNE implementations were included. We reported the results of t-test on β_3, which is the interaction between log of sample size and tSNE method. For both models, we set

tSNE as the reference. Thus, a significantly positive β_3 has the interpretation that using another implementation increases the slope of the model, which in turn indicates that the expected runtime increases.

For all rank-based metrics and results, higher ranks indicate superior performance. In case of tied rank, we employed the maximum rank to allow tied ranks to all have the highest rank. The highest possible rank is 20 while the lowest is 1, unless otherwise noted such as in the case of parameter tuning. Decimal ranks are results of the weighting scheme of our evaluation framework.

Code availability

The CyTOF simulating model is available at: cytomulate.readthedocs.io. The CyTOF DR Package webserver is available at github.com/kevin931/cytofDR. The CyTOF DR Playground webserver is available at dbai.biohpc.swmed.edu/cytof-dr-playground.

Data availability

The public CyTOF and scRNA-seq data used in this study were from citations[9,11,23,43–48,50–52] and described in **Sup. Table 2**, with data accession methods documented in this table. The in-house COVID-19 vaccinee CyTOF datasets were available from DBAI (https://dbai.biohpc.swmed.edu/), with accession codes from CyTOF_00000000000001 to CyTOF_00000000000012.

FIGURES AND TABLE LEGENDS

Fig. 1 Overview of several key aspects of this benchmark study. (a) The datasets, DR methods, validation metrics, and deliverables of this study. (b) Synopsis of the real CyTOF datasets used for evaluation in this study. (c) The simulation algorithm for CyTOF data. (d) Diagrams explaining several of the most important validation metrics in the accuracy category. (e) Deliverables of this study, including, a complete set of guidelines for choosing DR methods based on data characteristics and user preferences (CyTOF DR Playground), our streamlined pipeline for implementation of the DR methods and evaluation metrics (CyTOF DR Package), and the Cytomulate tool.

Fig. 2 Overview of the benchmark study. (a) The DR methods reviewed in this study, their basic characteristics described in text, and their performances in terms of accuracy, scalability, stability, and utility. The accuracy was evaluated on the real CyTOF data and simulated CyTOF data, respectively. And their average accuracy performances are also reported. (b-d) Complementary of the DR methods, evaluated on the real CyTOF datasets (b), the simulated CyTOF datasets (c), and all datasets combined (d). We define complementarity as the likelihood of obtaining a top-performing method for a given dataset by choosing a specific DR method. For details of calculation, refer to the method section.

Fig. 3 Detailed results on the accuracy performances of the DR methods. (a) The accuracy performances of the DR methods on all real CyTOF datasets, grouped by the 11 cohorts. Orange color indicates better performance, while the blue color indicates worse performance. All sub-metrics of the accuracy criterion were averaged. (b) The accuracy performances of the DR methods for each accuracy category. Results for all real CyTOF datasets were averaged, and the results for each detailed sub-metric of accuracy was shown. Orange color indicates better performance, while the blue color indicates worse performance. "*" Indicates that 10% down-sampling was performed. (c) The accuracy performances on the DR methods on the simulated CyTOF data. The performances were visualized as a function of the characteristics of the CyTOF datasets. Orange color indicates better performance, while the blue color indicates worse performance. All sub-metrics of the accuracy criterion were averaged. "*" Indicates that 10% down-sampling was performed.

Fig. 4 Detailed results on the scalability performances of the DR methods. The runtime (a,c) and memory usage consumption (b,d) of the DR methods. In (a) and (b), the DR methods were evaluated based on increasing numbers of cells. In (c) and (d), the DR methods were evaluated based on increasing numbers of protein features. (e,f) Whether the DR methods abort with error during the runtime/memory benchmarking analyses in each cell number group (e) and protein feature number group (f).

Fig. 5 Detailed results on the stability performances of the DR methods. (a) The stability of the DR methods. We calculated the stability of the DR methods, across different bootstrap datasets generated from the Oetjen cohort's Sample A. Stability is calculated by the variability of the rankings of the DR methods, in terms of the accuracy metrics (combined or separated), across the bootstrap samples. (b) The impact of parameter tuning on the performances of the DR methods. For each DR method, we showed the difference between the default setting and the optimal setting (among all tested settings, **Sup Table 4**), for every major category of accuracy. This improvement in performance is measured by both optimal settings' percent increase in ranks over default and the ranks of default configurations over the highest average ranks achieved by optimal settings.

Fig. 6 Practical guidelines for DR users. (a) A complete set of guidelines for users of DR methods to choose the best method for their CyTOF data. The user can either take a data-driven approach to run their dataset through all chosen DR methods based on their chosen evaluation metrics. Or the user can choose a prior-knowledge driven approach to query the best DR method based on similarity of the characteristics of their CyTOF dataset with those of the datasets we evaluated in this study. Our streamlined pipeline provides implementation of DR methods and the evaluation metrics, for users to benchmark and deploy the DR methods of choice. (b) Demonstration of how a common framework for DR methods facilitates broad applicability using one example CyTOF dataset. The DR plots were generated by our CyTOF DR Package

pipeline.

Sup. Fig. 1 Number of yearly citations returned by Google Scholar, by searching for "CyTOF", "scRNA-seq", and "CyTOF & scRNA-seq".

Sup. Fig. 2 One example breast cancer Imaging CyTOF dataset. The per-cell staining of the Ecadhe (E-cadherin), Ki67, SMA (α-smooth muscle actin), and CD45 channels were shown in the spatial context of the single cells.

Sup. Fig. 3 Performances of the DR methods on the simulation data. (a) Performances of the DR methods on the simulation data for each specific category of the accuracy metrics. (b) Performances of the DR methods on the simulation and real data for each major category of the accuracy metrics (scores of the specific categories averaged for each major category).

Sup. Fig. 4 Focusing on the runtime and memory usage performances of several tSNE algorithms, with increasing numbers of cells. "tSNE" in the main figures refers to "FIt-SNE (openTSNE)" here. "tSNE (sklearn)" in the main figures refers to "BH tSNE (Sklearn)" here.

Sup. Fig. 5 Usability scoring for the DR methods. All detailed usability criteria were displayed for all DR methods, in the format of a heatmap.

Sup. Fig. 6 Manual cell type assignment results of the CyTOF data from the Lung cancer cohort.

Sup. Table 1 All DR methods evaluated in this work

Sup. Table 2 CyTOF and scRNA-seq datasets used in this study

Sup. Table 3 Explanation of the DR validation metrics employed in this study

Sup. Table 4 Tuning parameters tested for the DR methods.

Sup. File 1 Accuracy metrics details and implementations

Sup. File 2 Additional analyses and discussions involving the CyTOF DR methods

Sup. File 3 Detailed explanation of the usability scoring criteria

FUNDING SOURCES

This study was supported by the National Institutes of Health (NIH) [CCSG 5P30CA142543/TW, NIH 1R01CA258584/TW, XW, U01AI156189/TW], Cancer Prevention Research Institute of Texas [CPRIT RP190208/TW, XW].

AUTHOR CONTRIBUTIONS

K.W. performed all DR benchmark analyses and created CyTOF DR Package. F.W. and K.W. created CyTOF DR Playground. Y.Y. and K.W. created Cytomulate. B.S. provided codes for pseudotime inference. X.W. contributed to study conception. T.W. and X.W. designed and

supervised the whole study. All authors wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests related to this work.

Bibliography

1. Cheung, R. K. & Utz, P. J. Screening: CyTOF-the next generation of cell detection. *Nat. Rev. Rheumatol.* **7**, 502–503 (2011).
2. Leipold, M. D. *et al.* Comparison of CyTOF assays across sites: Results of a six-center pilot study. *J. Immunol. Methods* **453**, 37–43 (2018).
3. Behbehani, G. K. Applications of mass cytometry in clinical medicine: the promise and perils of clinical cytof. *Clin. Lab. Med.* **37**, 945–964 (2017).
4. Li, Y. *et al.* Systematic Study of the Immune Components after Ischemic Stroke Using CyTOF Techniques. *J. Immunol. Res.* **2020**, 9132410 (2020).
5. Fernández-Zapata, C., Leman, J. K. H., Priller, J. & Böttcher, C. The use and limitations of single-cell mass cytometry for studying human microglia function. *Brain Pathol.* **30**, 1178–1191 (2020).
6. Lu, T. *et al.* Overcoming Expressional Drop-outs in Lineage Reconstruction from Single-Cell RNA-Sequencing Data. *Cell Rep.* **34**, 108589 (2021).
7. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9**, 997 (2018).
8. Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15**, 539–542 (2018).

9. Stark, S. G. *et al.* SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics* **36**, i919–i927 (2020).

10. Ha, M. K. *et al.* Mass Cytometry and Single-Cell RNA-seq Profiling of the Heterogeneity in Human Peripheral Blood Mononuclear Cells Interacting with Silver Nanoparticles. *Small* **16**, e1907674 (2020).

11. Zhang, F. *et al.* Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nat. Immunol.* **20**, 928–942 (2019).

12. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

13. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* (2018) doi:10.48550/arxiv.1802.03426.

14. Amodio, M. *et al.* Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* **16**, 1139–1145 (2019).

15. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16**, 241 (2015).

16. Moon, K. R. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).

17. Ding, J., Condon, A. & Shah, S. P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **9**, 2002 (2018).

18. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional

single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).

19. Chang, Q. *et al.* Imaging Mass Cytometry. *Cytometry A* **91**, 160–169 (2017).
20. Abe, K., Minoura, K., Maeda, Y., Nishikawa, H. & Shimamura, T. Model-based clustering for flow and mass cytometry data with clinical information. *BMC Bioinformatics* **21**, 393 (2020).
21. Kobak, D. & Linderman, G. C. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol.* **39**, 156–157 (2021).
22. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018).
23. Oetjen, K. A. *et al.* Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry. *JCI Insight* **3**, (2018).
24. van der Maaten, L. Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research* **15**, 3221–3245 (2014).
25. Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S. & Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* **16**, 243–245 (2019).
26. Poličar, P. G., Stražar, M. & Zupan, B. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. *BioRxiv* (2019) doi:10.1101/731877.
27. Chevrier, S. *et al.* Compensation of signal spillover in suspension and imaging mass cytometry. *Cell Syst.* **6**, 612-620.e5 (2018).
28. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell

trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

29. Satpathy, S. *et al.* A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184**, 4348-4371.e40 (2021).
30. Wang, L.-B. *et al.* Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* **39**, 509-528.e20 (2021).
31. Krug, K. *et al.* Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy. *Cell* **183**, 1436-1456.e31 (2020).
32. Satpathy, S. *et al.* Microscaled proteogenomic methods for precision oncology. *Nat. Commun.* **11**, 532 (2020).
33. Maeda, T. *et al.* Role of polyamines derived from arginine in differentiation and proliferation of human blood cells. *Biol. Pharm. Bull.* **29**, 234–239 (2006).
34. Mimitou, E. P. *et al.* Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nat. Methods* **16**, 409–412 (2019).
35. Lin, D., Gupta, S. & Maecker, H. T. Intracellular cytokine staining on PBMCs using CyTOF™ mass cytometry. *Bio Protoc* **5**, (2015).
36. Zhang, Z., Xiong, D., Wang, X., Liu, H. & Wang, T. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nat. Methods* **18**, 92–99 (2021).
37. Lu, T. *et al.* Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nat. Mach. Intell.* (2021) doi:10.1038/s42256-021-00383-2.
38. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 5416 (2019).

39. Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**, 636–645 (2015).

40. Weber, L. M. & Robinson, M. D. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry A* **89**, 1084–1096 (2016).

41. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).

42. Kaushik, A. *et al.* CyAnno: A semi-automated approach for cell type annotation of mass cytometry datasets. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab409.

43. Hawley, D. *et al.* RNA-Seq and CyTOF immuno-profiling of regenerating lacrimal glands identifies a novel subset of cells expressing muscle-related proteins. *PLoS ONE* **12**, e0179385 (2017).

44. Böttcher, C. *et al.* Human microglia regional heterogeneity and phenotypes determined by multiplexed single-cell mass cytometry. *Nat. Neurosci.* **22**, 78–90 (2019).

45. Jackson, H. W. *et al.* The single-cell pathology landscape of breast cancer. *Nature* **578**, 615–620 (2020).

46. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).

47. Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L. & Nolan, G. P. Automated

mapping of phenotype space with single-cell data. *Nat. Methods* **13**, 493–496 (2016).

48. Gonugunta, A. S. *et al.* Humoral and cellular correlates of a novel immune-related adverse event and its treatment. *J. Immunother. Cancer* **9**, (2021).
49. Zhu, J. *et al.* BepiTBR: T-B reciprocity enhances B cell epitope prediction. *iScience* **25**, 103764 (2022).
50. Schulz, D., Severin, Y., Zanotelli, V. R. T. & Bodenmiller, B. In-Depth Characterization of Monocyte-Derived Macrophages using a Mass Cytometry-Based Phagocytosis Assay. *Sci. Rep.* **9**, 1925 (2019).
51. Rundberg Nilsson, A., Bryder, D. & Pronk, C. J. H. Frequency determination of rare populations by flow cytometry: a hematopoietic stem cell perspective. *Cytometry A* **83**, 721–727 (2013).
52. Mosmann, T. R. *et al.* SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 2: biological evaluation. *Cytometry A* **85**, 422–433 (2014).

Fig. 1

(a)

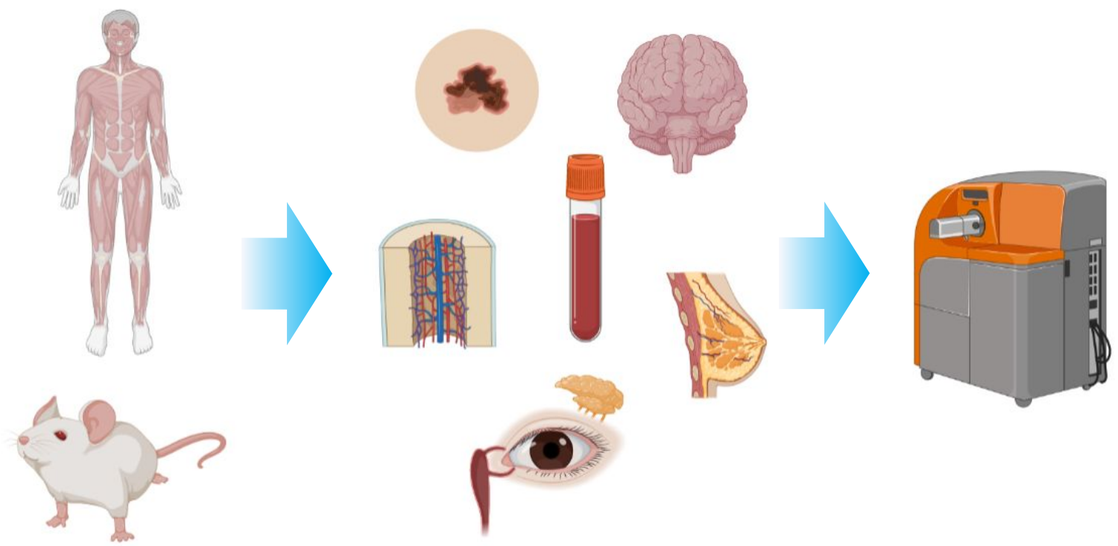
110 real and 425 simulated CyTOF datasets + 20 dimension reduction methods + 30 metrics

Accuracy + Scalability + Stability + Usability

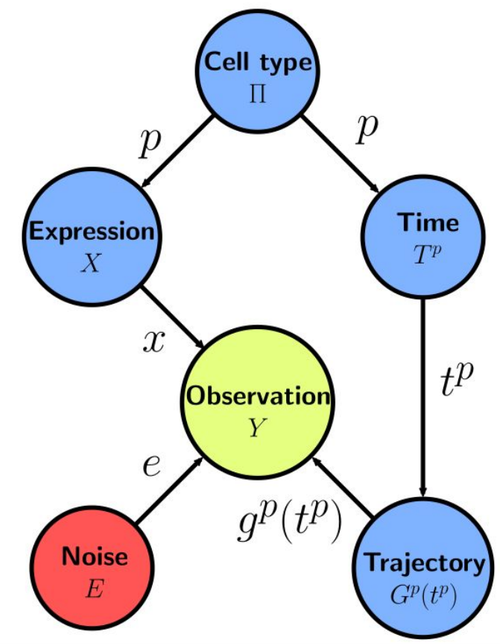
User guidelines + New possibilities for developers

Method/benchmark wrappers:
github.com/kevin931/cytofDR
Guideline app:
dbai.biohpc.swmed.edu/cytof-dr-playground
Simulation tool: cytomulate.readthedocs.io

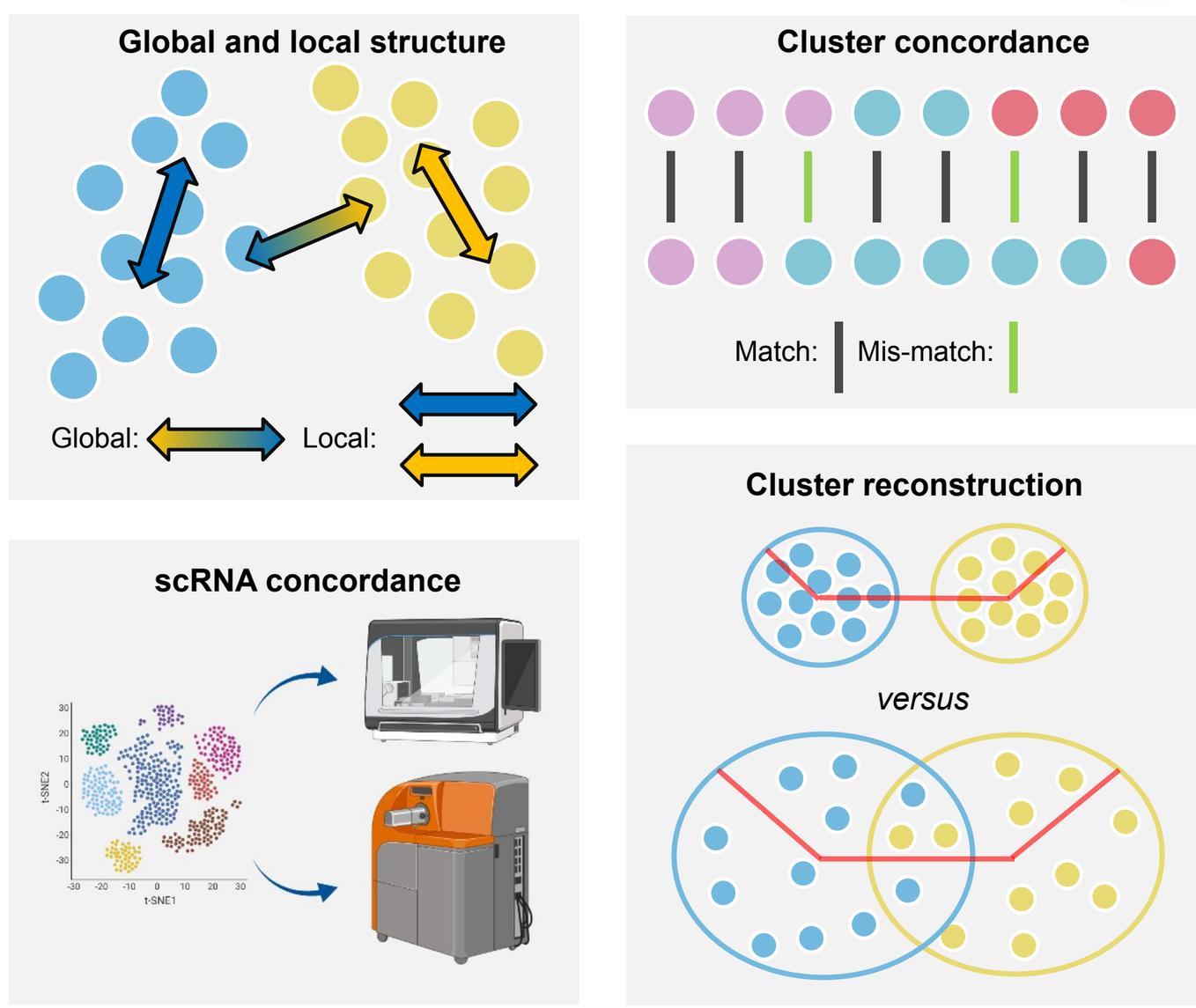
(b)



(c)



(d)



(e)

CYTOF PACKAGE

CYTOF PLAYGROUND

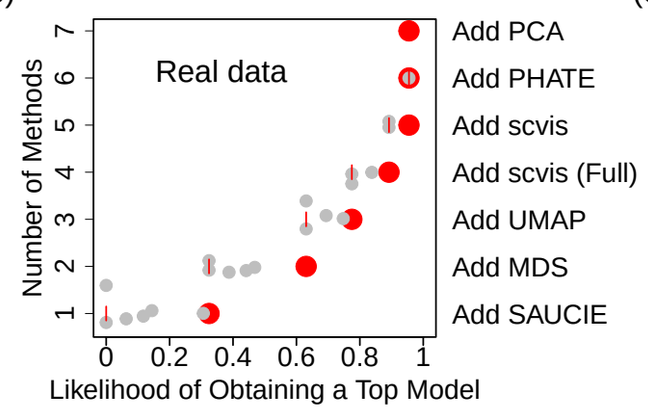
CYTOMULATE
MADE FOR CYTOF

Fig. 2

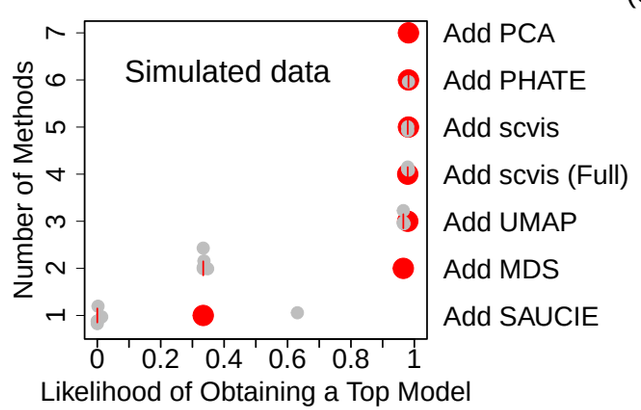
(a)

Method	Platform	Package	Purpose	Mapping	Accuracy: Overall	Accuracy: Real	Accuracy: Simulation	Scalability	Stability	Usability
SAUCIE	Python	SAUCIE	Single Cell	No						
MDS	Python	sklearn	General	No						
UMAP	Python	umap-learn	General/Single Cell	Yes						
scvis (Full)	Python	scvis	scRNA	Yes						
scvis	Python	scvis	scRNA	Yes						
PHATE	Python	phate	Single Cell	No						
PCA	Python	sklearn	General	Yes						
KPCA Poly	Python	sklearn	General	No						
Isomap	Python	sklearn	General	Yes						
tSNE (sklearn)	Python	sklearn	General	No						
GrandPrix	Python	grandprix	Single Cell	No						
tSNE	Python	openTSNE	General/Single Cell	Yes						
KPCA RBF	Python	sklearn	General	No						
DiffMap	R	destiny	General/Single Cell	No						
ICA	Python	sklearn	General	Yes						
FA	Python	sklearn	General	No						
ZIFA	Python	ZIFA	scRNA	No						
NMF	Python	sklearn	General	No						
Spectral	Python	sklearn	General	No						
LLE	Python	sklearn	General	Yes						

(b)



(c)



(d)

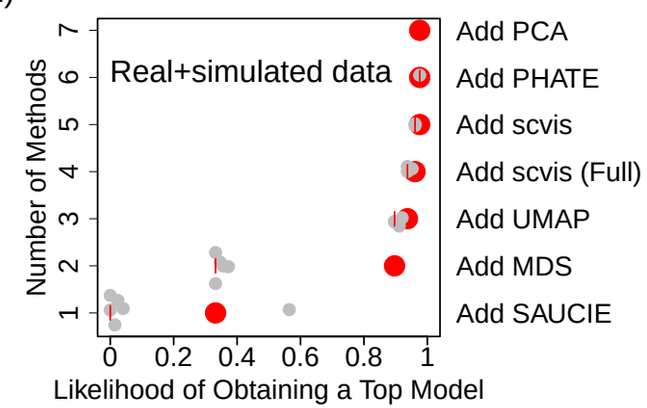


Fig. 3

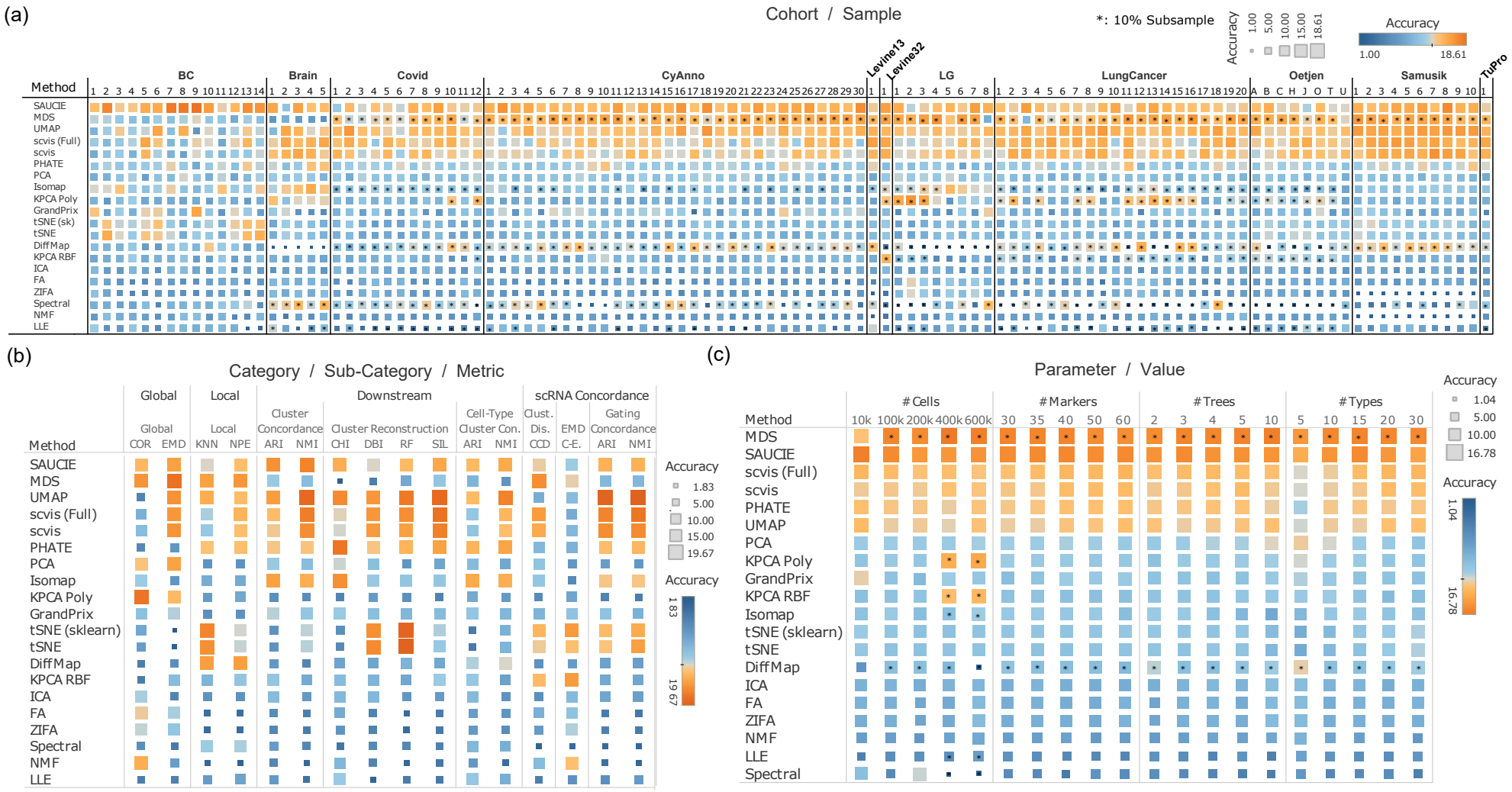
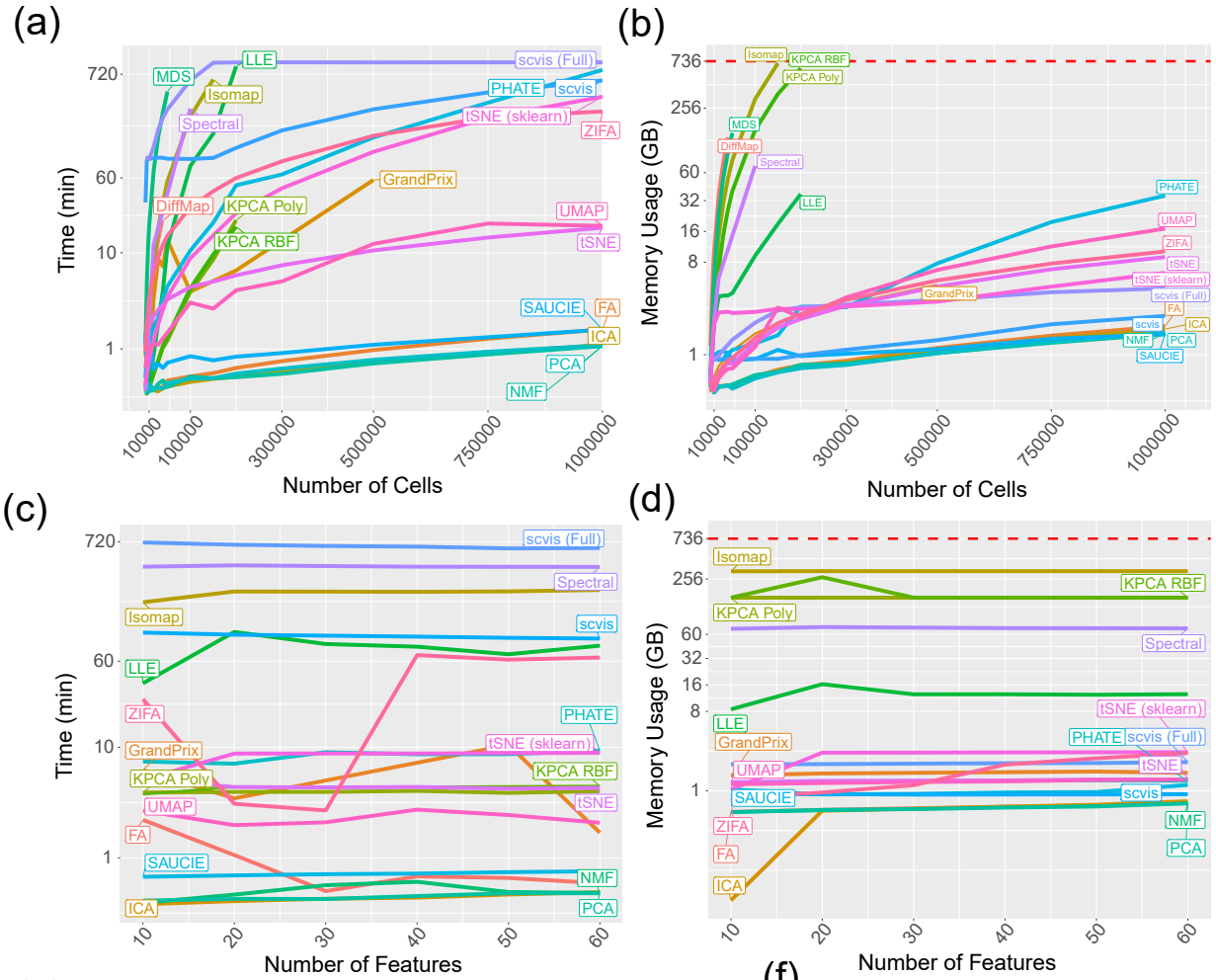

(a)
Cohort / Sample
*: 10% Subsample
Accuracy
1.00
5.00
10.00
15.00
18.61
Accuracy
1.00
18.61
Method
BC
Brain
Covid
CyAnno
Levine13
Levine32
LG
LungCancer
Oetjen
Samusik
TuPro
SAUCIE
MDS
UMAP
scvis (Full)
scvis
PHATE
PCA
Isomap
KPCA Poly
GrandPrix
tSNE (sk)
tSNE
DiffMap
KPCA RBF
ICA
FA
ZIFA
Spectral
NMF
LLE
(b)
Category / Sub-Category / Metric
Global
Local
Downstream
scRNA Concordance
Global
Local
Cluster Concordance
Cluster Reconstruction
Cell-Type Cluster Con.
Clust. Dis.
EMD C-E.
Gating Concordance
COR
EMD
KNN
NPE
ARI
NMI
CHI
DBI
RF
SIL
ARI
NMI
CCD
ARI
NMI
Accuracy
1.83
5.00
10.00
15.00
19.67
Accuracy
1.83
19.67
(c)
Parameter / Value
Cells
Markers
Trees
Types
10k
100k
200k
400k
600k
30
35
40
50
60
2
3
4
5
10
5
10
15
20
30
Accuracy
1.04
5.00
10.00
16.78
Accuracy
1.04
16.78

Fig. 4



(e)

#Cells

Method	2000	5000	8000	10000	20000	30000	40000	50000	100000	150000	200000	300000	500000	750000	1000000
DiffMap	●	●	●	●	●	●	●	○	○	○	○	○	○	○	○
FA	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
GrandPrix	●	●	●	●	●	●	●	●	●	○	●	○	●	○	○
ICA	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Isomap	●	●	●	●	●	●	●	●	●	●	○	○	○	○	○
KPCA Poly	●	●	●	●	●	●	●	●	●	●	●	○	○	○	○
KPCA RBF	●	●	●	●	●	●	●	●	●	●	●	○	○	○	○
LLE	●	●	●	●	●	●	●	●	●	●	●	○	○	○	○
MDS	●	●	●	●	●	●	●	●	○	○	○	○	○	○	○
NMF	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
PCA	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
PHATE	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
SAUCIE	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
scvis	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
scvis (Full)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
Spectral	●	●	●	●	●	○	○	○	●	○	○	○	○	○	○
tSNE	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
tSNE (sklearn)	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
UMAP	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
ZIFA	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

(f)

Error
Success

#protein features

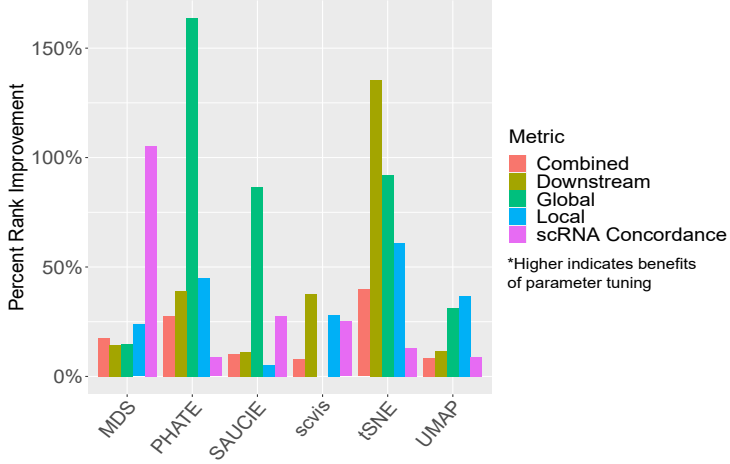
Method	10	20	30	34	40	50	60
DiffMap	○	○	○	○	○	○	○
FA	●	●	●	●	●	●	●
GrandPrix	●	●	○	●	○	●	●
ICA	●	●	●	●	●	●	●
Isomap	●	●	●	●	●	●	●
KPCA Poly	●	●	●	●	●	●	●
KPCA RBF	●	●	●	●	●	●	●
LLE	●	●	●	●	●	●	●
MDS	○	○	○	○	○	○	○
NMF	●	●	●	●	●	●	●
PCA	●	●	●	●	●	●	●
PHATE	●	●	●	●	●	●	●
SAUCIE	●	●	●	●	●	●	●
scvis	●	●	●	●	●	●	●
scvis (Full)	●	●	●	●	●	●	●
Spectral	●	●	○	●	●	○	●
tSNE	●	●	●	●	●	●	●
tSNE (sklearn)	●	●	●	●	●	●	●
UMAP	●	●	●	●	●	●	●
ZIFA	●	●	●	●	●	●	●

Fig. 5

(a)

Metric	KPCA RBF	NMF	KPCA Poly	ZIFA	FA	ICA	UMAP	PCA	scvis (Full)	scvis	tSNE	MDS	PHATE	SAUCIE	Isomap	tSNE Sk.	GrandP.	Spectral	LLE	diffmap
Overall	20.00	19.00	18.00	17.00	16.00	15.00	14.00	13.00	12.00	11.00	10.00	9.00	7.00	8.00	6.00	5.00	4.00	3.00	2.00	1.00
Global	20.00	17.00	19.00	11.00	16.00	7.00	5.00	12.00	13.00	18.00	10.00	9.00	15.00	2.00	6.00	14.00	4.00	8.00	3.00	1.00
Local	19.00	16.00	20.00	14.00	13.00	10.00	8.00	11.00	5.00	6.00	15.00	18.00	7.00	9.00	12.00	17.00	3.00	4.00	2.00	1.00
Downstream	9.00	16.00	20.00	11.00	8.00	7.00	18.00	4.00	19.00	17.00	14.00	13.00	10.00	15.00	6.00	12.00	5.00	3.00	2.00	1.00
scRNA Con.	19.00	20.00	13.00	14.00	12.00	17.00	18.00	16.00	10.00	5.00	7.00	8.00	4.00	15.00	6.00	3.00	9.00	11.00	2.00	1.00

(b)



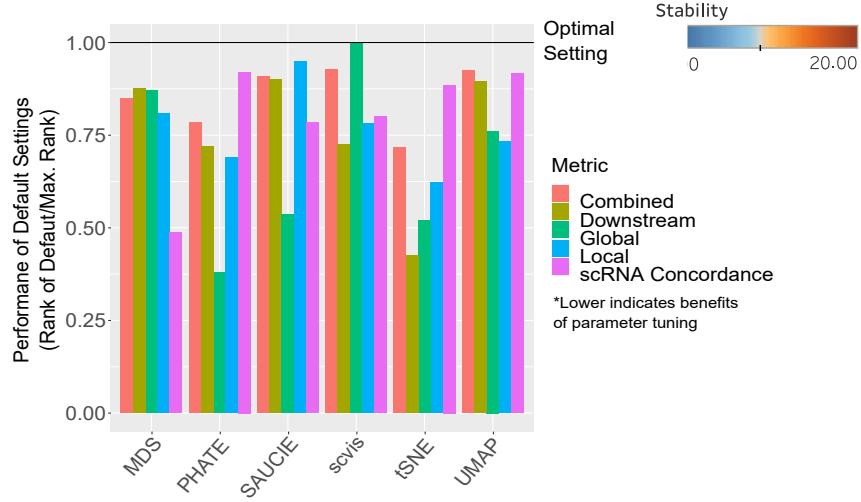
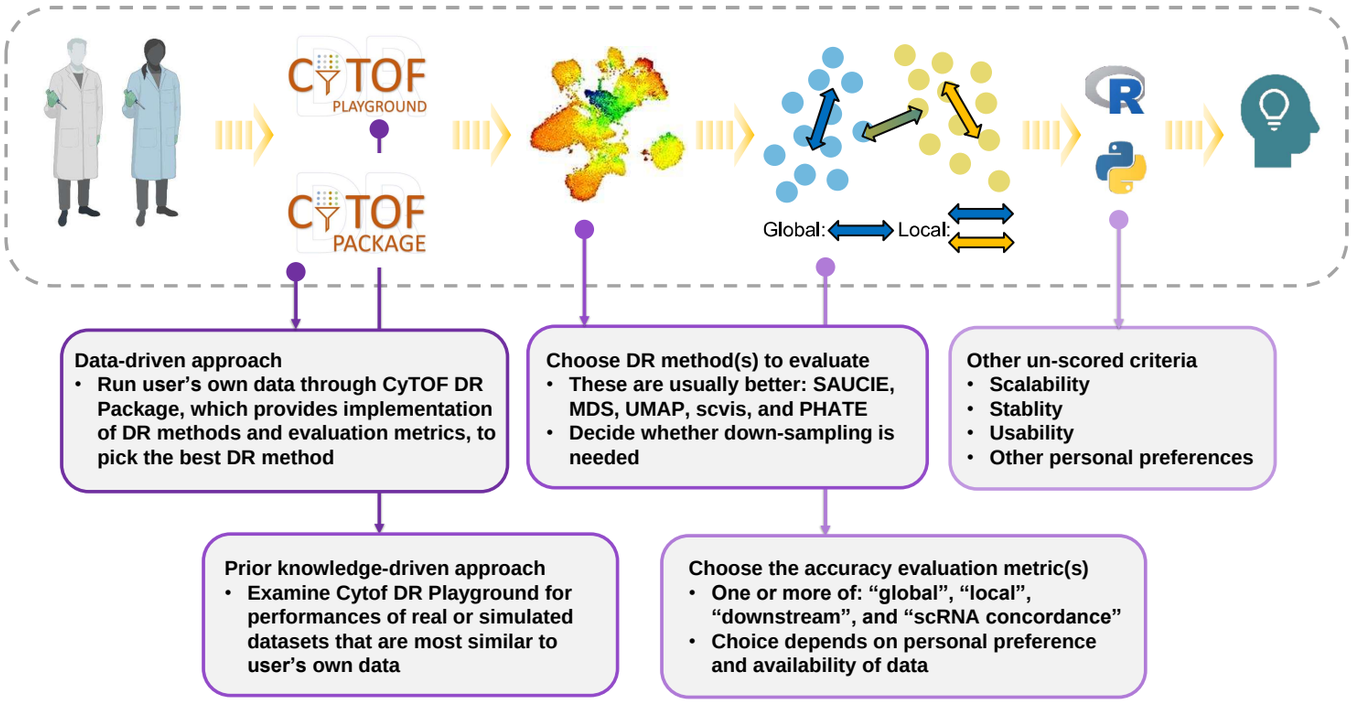
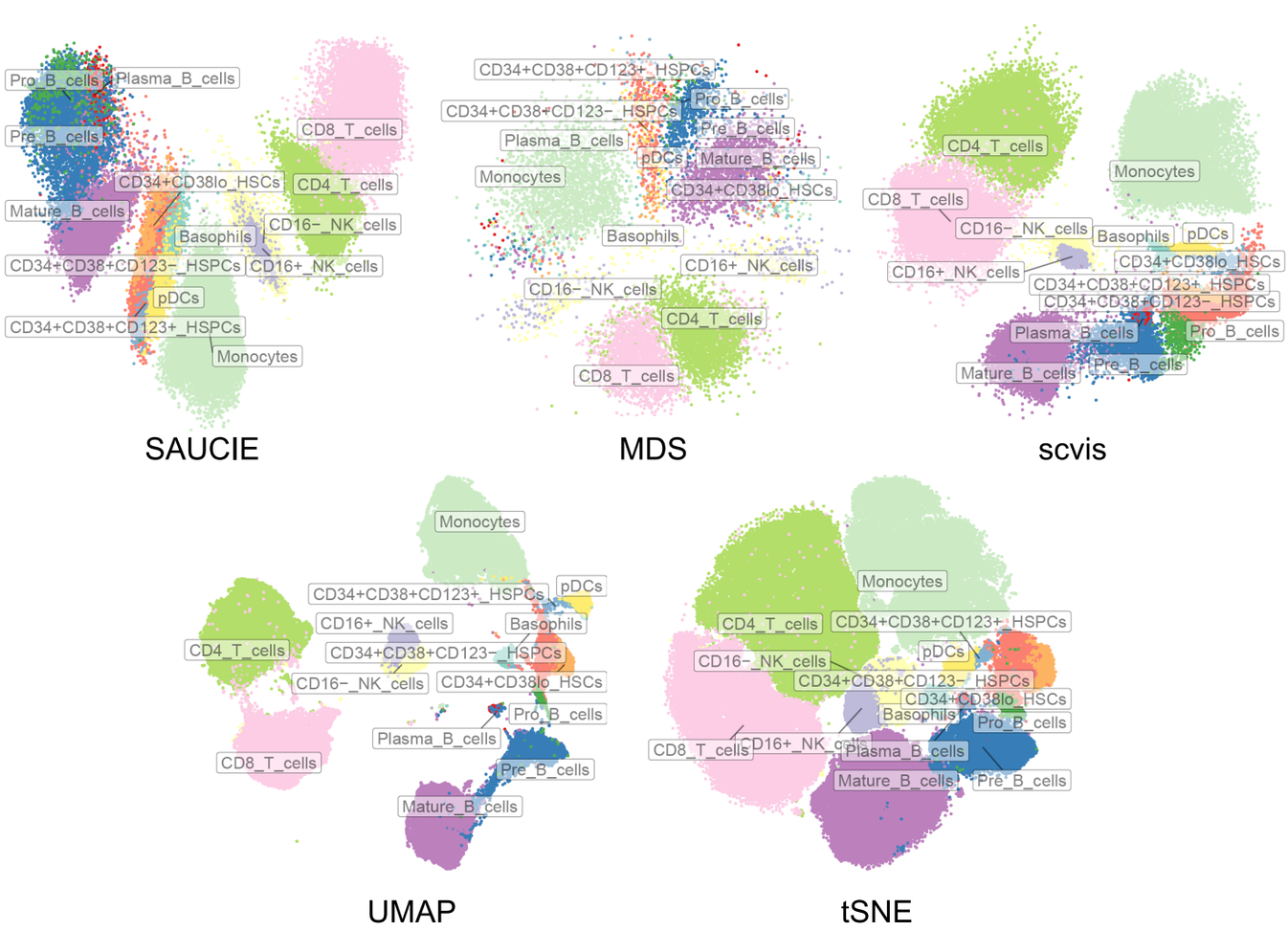


Fig. 6

(a)


CYTOF PLAYGROUND
CYTOF PACKAGE
Global:
Local:
Data-driven approach
• Run user's own data through CyTOF DR Package, which provides implementation of DR methods and evaluation metrics, to pick the best DR method
Choose DR method(s) to evaluate
• These are usually better: SAUCIE, MDS, UMAP, scvis, and PHATE
• Decide whether down-sampling is needed
Other un-scored criteria
• Scalability
• Stablity
• Usability
• Other personal preferences
Prior knowledge-driven approach
• Examine Cytof DR Playground for performances of real or simulated datasets that are most similar to user's own data
Choose the accuracy evaluation metric(s)
• One or more of: "global", "local", "downstream", and "scRNA concordance"
• Choice depends on personal preference and availability of data

(b)


Pro_B_cells
Plasma_B_cells
Pre_B_cells
CD8_T_cells
CD34+CD38lo_HSCs
CD4_T_cells
Mature_B_cells
CD16−_NK_cells
Basophils
CD34+CD38+CD123−_HSPCs
CD16+_NK_cells
pDCs
CD34+CD38+CD123+_HSPCs
Monocytes
SAUCIE
MDS
scvis
UMAP
tSNE