Cytomulate: Accurate and Efficient Simulation of CyTOF data

Yuqiu Yang[1,2], Kaiwen Wang[1], Zeyu Lu[1,2], Tao Wang[2,3,*], Xinlei Wang[1,4,5*]

1 Department of Statistical Science, Southern Methodist University, Dallas, TX, USA, 75275.

2 Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, 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.

4 Department of Mathematics, University of Texas at Arlington, Arlington, USA, 76019

5 Center for Data Science Research and Education, College of Science, University of Texas at Arlington, Arlington, USA, 76019

* Corresponding authors

Short title: Cytomulate: Simulation of CyTOF data

Key words: CyTOF, simulation, proteomics

Abstract:

Recently, many analysis tools have been devised to offer insights into data generated via Cytometry by time-of-flight (CyTOF). However, objective evaluations of these methods remain absent as most evaluations are conducted against real data where the ground truth is generally unknown. In this paper, we develop Cytomulate, a reproducible and accurate simulation algorithm of CyTOF data, which could serve as a foundation for future method development and evaluation. We demonstrate that Cytomulate can capture various characteristics of CyTOF data and is superior in learning overall data distributions than single-cell RNA-seq-oriented methods such as scDesign2, Splatter and generative models like LAMBDA.

Background:

Recent years have seen exponential growth in applications of mass cytometry (CyTOF) which characterizes proteomics profiles of tens of thousands or millions of heterogeneous single cells on 30-40 protein channels with minimal spill-over effects (1–3). As the technique of CyTOF matures, many data analysis methods have been proposed to offer insights into various aspects of single-cell experiments (4). For example, in (5), the authors have observed that the signal intensity of a CyTOF experiment would vary over time due to changes in instrument performance. Their proposed method is currently the dominant tool to alleviate the temporal effect using polystyrene beads embedded with metal lanthanides. On the other hand, methods such as CytofRUV (6), and Cytonorm (7), etc. aim at getting rid of batch effects to facilitate comparisons across different CyTOF experiments. Deep learning algorithms such as residual networks (ResNet) (8) and generative adversarial networks (GAN) (9,10) were also adopted to

resolve the same problem. While some software such as FlowSOM (11), LAMBDA (12), and Bayesian trees (13) focus on automatic cell type identification, others like CytoTree (14) attempt to explore the trajectories along which cells differentiate into other cells. Various data visualization and dimension reduction algorithms have also been applied to CyTOF (15). Typical examples include: t-SNE (16), UMAP (17), MDS (18), SAUCIE (19), scvis (20), etc. As the output of a CyTOF acquisition software is in Flow Cytometry Standard (FCS) format, packages such as flowCore (21) and Seurat (22) have incorporated standardized and efficient pipelines to read, write, and preprocess CyTOF data.

Despite the vast array of methods applicable to CyTOF, most of the CyTOF analysis methods developed by researchers are often evaluated against real data. Without a gold standard and known truth, such as labels and actual distributions, performance assessment becomes somewhat subjective while also blurring the line between method validation and result discovery. The few studies with simulated datasets are highly specialized and workflow-dependent. For example, CytoGLMM (23) focuses on a single cell type for the purpose of differential expression analysis. In (24), the authors further used the simulated datasets in (23) to evaluate performance on various differential expression detection methods. (25) proposed a Bayesian hierarchical model based on Gaussian distributions to study the difference in the protein functional network of a single cell type under two experimental conditions. In (26) the simulation was done by manually splitting a real dataset into two overlapping subsets. These simulation procedures are not specifically designed for general-purpose simulations of CyTOF samples with many cell types, trajectories, batches, etc.

Simulated datasets, albeit being synthetic, contain ground truth against which methods and algorithms can be reliably tested. Instead of *ad hoc* simulation in each study mentioned above, a

formal simulation tool naturally provides a framework for a systematic evaluation for new methodological research in the field. By varying parameters and settings of key factors, researchers can easily generate an unbiased coverage of the system under study in a straightforward manner. On the other hand, real experiments require detailed prior planning, precise executions, and potentially abundant resources. The flexibility along with cost and time benefits make simulation a vital tool for the field. However, the lack of a rigorous, reliable simulation algorithm calls for the development of such a tool for validating other CyTOF data analysis tools and future developments.

By contrast, various methods have been developed for simulating single cell RNA-sequencing (scRNA-seq) data, such as Simple (27), Lun (28), BASICS (29), scDesign2 (30), ZINB-WaVE (31), and Splatter (32), among which Splatter enjoys the highest popularity whilst scDesign2 excels in simulation performance (33,34). Since both scRNA-seq and CyTOF are single-cell sequencing techniques, it might be tempting to tweak scRNA-seq simulation tools such as Splatter and scDesign2 and apply them to CyTOF. However, such a process can be highly nontrivial and the end results may be unsatisfactory if only small adjustments are sought after. First of all, these two techniques are of different biological nature. scRNA-seq examines the expression of thousands of RNAs of several thousand cells, whereas CyTOF captures information on typically 30-40 proteins for tens of thousands or even millions of cells. Therefore, compared with scRNA-seq, CyTOF is more closely related to clinical phenotypes and is more likely to identify rare cell populations. Furthermore, while drop-out is a common issue in scRNA-seq data, it is a much less encountered problem in CyTOF experiments. The expressions in scRNA-seq are integer-valued “counts”, in contrast to real-valued data that even have negative values in CyTOF expression matrices due to several built-in processing steps in a CyTOF system

such as data randomization and transformation. Consequently, applying simulation tools like Splatter and scDesign2 will almost inevitably require a sequence of nonlinear transformations to artificially construct count data, which will yield biased results. Finally, the expression levels of various protein channels in CyTOF are naturally dependent. Extending a complex discrete random variable model such as the Poisson-Gamma mixture model adopted by Splatter to capture the dependency structures present in CyTOF may not be a trivial task. In addition, methods for simulating multivariate discrete data such as the ones devised in (35,36) are not suitable for CyTOF data that are real-valued. The same logic applies to other generic simulation algorithms that are not designed with CyTOF in mind.

To bridge this gap in literature, we developed Cytomulate, the first formal simulation tool that is capable of generating simulated datasets that capture various characteristics of CyTOF data. Our framework combines the flexibility of generating complex, realistic data with efficiency and usability, allowing researchers to acquire data at scale and practitioners to easily simulate datasets for their workflows.

Results:

Characteristics of CyTOF data

Currently, most CyTOF data are available in the FCS format for analysis purposes, containing real-valued protein expression levels rather than counts. This is due to several built-in data processing steps in a CyTOF system (e.g., Helios, latest CyTOF XT): (i) The system applies regression to recover peak counts lost because of overlapping peaks at high intensities, resulting

in non-integer count estimates. (ii) To avoid binning effects caused by integers near zero, the system applies randomization (i.e., adding uniform or Gaussian noise to counts). (iii) Bead normalization is applied to remove temporal effects due to changes in instrument performance over time. (iv) The inverse hyperbolic sine function, referred to as acsinh-transformation, is applied to alleviate the coupling of mean and variance and "gaussianize" the data (37). Although the system allows users to turn on and off each of (ii)-(iv), often one or more such options were implemented for nearly all existing CyTOF data, leading to real-valued expressions.

To motivate the design choices of Cytomulate, we explored prominent characteristics of CyTOF by analyzing various public and in-house datasets (e.g., Finck (5), Levine_13dim (38), Levine_32dim (38), CytoNorm (7), CyAnno (39), Covid (15); Additional file 2). In this section, we mainly use protein expressions from the Levine_13dim dataset as an example to demonstrate such features. On a broader scale, similar phenomena have been observed across numerous CyTOF datasets. The Levine_13dim dataset is publicly available via the R package HDCytoData (40). As the protein expressions have already been pre-processed to correct unwanted artifacts, such as temporal effects, only asinh-transformation was applied.

High volume, high dimension, and heterogeneity: The expression matrix associated with a typical CyTOF experiment usually contains tens of thousands or even millions of cell events (rows), 30-40 protein channels (columns), from a pool of cells of 10-30 cell types. For example, the data of the first patient in Levine_13dim has 167,044 cell events, 13 protein channels, and 14 cell types manually gated by the authors. With a varying number of protein channels, the heterogeneous nature of different cells are captured by CyTOF via cell types and differentiation stages. Any potential simulation method developed for CyTOF needs to not only handle the throughput with adequate efficiency but also the heterogeneity of cell events.

Zero-inflation: The protein expression matrix observed in a CyTOF experiment often contains features with either zeros or values close to zero. These features biologically correspond to surface proteins that are too scarce to be detected by the machine. In addition, because of data randomization, one of the standard options in the CyTOF system, many datasets have near-zero or even negative values rather than zeros. Take, for example the CD4 T cells in the dataset, we see in Fig. 1a and Fig. 1b that the majority of the protein channels (8 out of 13) contain a high proportion of values close to zero.

Correlation in the protein-protein expression: In Fig. 1c, we calculated the correlation matrix among all the protein channels using the data of Naive CD4+ T cells. Clearly, markers such as CD44 and CD45 are correlated, implying that the independence assumption between columns might be too restrictive. This suggests modeling the data jointly instead of marginally.

Near-normality: To explore the distributional characteristics of asinh-transformed protein expressions, we first examined the marginal density estimates of each protein marker for different cell types. In Fig. 1b, we plotted these empirical densities for all markers using Naive CD4+ T cells, where expressed markers (i.e. those with relatively high expression levels) possess "bell"-shaped curves and other markers exhibit two or more modes that can be approximated by Gaussian mixture models (GMMs). To further get a sense of the joint distribution, we performed dimension reduction using Principal Component Analysis (PCA). All cell populations resemble roughly elliptic shapes (Fig. 1d), suggesting that there is no gross violation to the joint normality, according to the assumptions of PCA. It may be tempting to use a multivariate normal (MVN) distribution for each cell type. However, protein markers' distributions depend on user-provided cell types and so MVNs may not be adequate, particularly in cases where the cell typing is coarse or inaccurate. Given that GMMs are known as a universal approximator of densities (41),

combined with the near-normality of CyTOF data described above, it is a sensible choice to use GMMs with Gaussian distributions as basic building blocks of a simulation model. Finally, this key characteristic also differentiates CyTOF from scRNA-seq, which is typically modeled with counts. As explored in the comparison study later in this paper, adapting a model for discrete counts for CyTOF does not work well.

Temporal effect: As illustrated in (5), the signals of a single CyTOF experiment would typically vary over time due to changes in instrument performance. Since Cytomulate aims at modeling asinh-transformed protein expressions, unlike in (5), we chose to demonstrate the signal drift by transforming the Peripheral Blood Mononuclear Cell (PBMC) data from the Finck dataset (Additional file 1: Figure S1a). As a first attempt to model the signal drift, we centered each signal, fitted a smoothing spline against one of them, and subtracted the fitted value from each signal. In Additional file 1: Figure S1b, we see that the adjusted signals overlap and are roughly homoscedastic, suggesting that the overall shape of the signal drift (temporal effect) may be shared by all cell types within a single batch and that we could model the effect using an additive model.

Batch effect: Apart from the temporal effect, many researchers also reported the presence of batch effects across multiple CyTOF experiments due to sample collection, stimulation, staining, etc. (42). In (7) they reported that not only are the batch effects cell-type specific, they also vary depending on protein channels. In Additional file 1: Figure S1c, we illustrated this phenomenon by an interaction plot using the mean expressions of the protein channels of CD4 T cells and CD8 T cells measured on two patients in the Levine_32dim dataset (40). We see that not only do expression levels differ between two batches, the differences also vary by cell types and protein

channels. This implies that in addition to the main batch effect, we should also take into consideration the interaction effects.

Cellular trajectory: A common task is to study cellular trajectory where the expressions of an individual cell are assumed to vary according to some continuous path that connects one cell type to another. In the context of scRNA-seq, lots of methods have been developed to align cells on continuous paths with typical examples Slingshot and Monocle (43–45). To our knowledge, CytoTree (14) is the only available method of trajectory inference designed for CyTOF. It is further questionable whether trajectory inference methods are interchangeable between these two technologies. For the purpose of demonstration, we showed an example of the cellular trajectories following CytoTree's online tutorial with a real Bone Marrow dataset in Additional file 1: Figure S1d.

Cytomulate Framework

In light of the previous exploratory data analysis, we propose Cytomulate, an accurate and efficient simulation algorithm of CyTOF data based on Gaussian Mixture Models (GMM) (46), with add-on features to mimic various patterns/effects existing in real data. More specifically, Cytomulate achieves the task of accurate simulation via two modes selectable by users: emulation mode and creation mode, both offering transparency with known ground truth. Emulation mode is designed for a user to simulate data that aims to capture major characteristics of a specific CyTOF dataset under study. As this mode learns meaningful parameters from real data, it enables interpretation of real data such as locations, variations of protein expressions, protein-protein correlations, and heterogeneity among various cell types, while generating synthetic data to test different hypotheses. Creation mode is purely model-based in which

synthetic data are generated from user-specified settings and parameters. This mode caters to two main user groups. Firstly, it serves users who wish to have access to a wide range of CyTOF datasets, encompassing varying numbers of cells, protein markers, experimental designs, tissue types, diseases, and conditions. These users aim to simulate multiple datasets with diverse characteristics to facilitate comprehensive testing and benchmarking. Secondly, creation mode is also designed to assist new researchers in the field who may not have a real CyTOF dataset at hand. By utilizing this mode, they can generate synthetic data immediately, providing an accessible starting point for their own experiments and studies.

In both simulation modes, users are allowed to vary cell abundances, noise levels, batch effects, temporal effects, as well as cellular trajectories to capture the complex needs of methodological developments for CyTOF data. The simulated batch effects, temporal effects, cellular trajectories, as well as the noise are then added once we sample from the GMMs. The following sections elaborate on each mode with greater details as the foundation to the comparison study that follows.

Creation mode aims at quickly generating synthetic datasets that capture aforementioned characteristics of CyTOF data such as zero-inflation, correlation in protein-protein expressions, and near-normality. Users specify the volume, dimension, and the degrees of heterogeneity. Temporal effect, cellular trajectory, and batch effect can be added as part of complex simulations. As this mode does not intend to extract the data distribution of some specific dataset as a reference, it is particularly useful for testing analysis methods on a large scale. For example, in (15), Wang et al. used this mode of Cytomulate to evaluate the performance of 20 dimension reduction methods from various aspects such as local and global structure preservation on 425 simulated datasets with different attributes (e.g., different numbers of cell types and numbers of

differentiation paths). This mode allows practitioners to precisely define each simulated dataset as they wish, which is not often or at all possible with real data.

Specifically, the creation mode randomly generates average cell expressions and then utilizes GMMs from which to sample cell expressions. Users also have the options to enable complex simulations by simply specifying the necessary parameters, such as the number of trees for cell differentiation trajectories. The details of complex simulation will be discussed later. With users having full control of their simulated dataset, the creation mode is untethered by existing—yet oftentimes implicit—conditions of real datasets, such as the exact cell types, circumstances of the patient, the physical machine, etc. By eliminating these factors, Cytomulate allows researchers to develop and test methods based on unbiased results that are generalizable to wide ranging real datasets. Further, it can also serve as a pedagogical tool for researchers to explore the characteristics of CyTOF datasets easily.

Emulation mode, on the other hand, is designed to learn the distribution of some real CyTOF data. To ensure compatibility, Cytomulate assumes that the reference dataset provided by the user has already undergone bead normalization and asinh-transformation (with a cofactor of 5), both of which are standard options in the CyTOF system. Instead of approximating the entire dataset as a whole, which is challenging, Cytomulate leverages the prior knowledge of cell type information. As a result, it requires not only the expression matrix of one particular CyTOF dataset to mimic but also the cell type label associated with each cell as inputs. Although Cytomulate can potentially perform cell type identification or clustering using common algorithms of the likes of K-means, we leave the choice of such algorithm for cell typing to our users as a design choice for the following two reasons: (i) According to (47) and (48), the k-means algorithm and GMM-based model (SWIFT) are often not the optimal choice for clustering

CyTOF data; (ii) There are already plenty of well-established cell identification algorithms and automatic gating techniques such as FlowSOM (11), Bayesian Tree (13), ACDC (49), SCINA (50), etc., all of which have similar goals but slightly different design considerations. For example, using FlowSOM requires minimal prior knowledge on the cell population other than some crude idea on the number of cell types. On the other hand, Bayesian Tree and ACDC use a template as an input to facilitate the automatic gating process. Since we expect the users of the emulation mode to be somewhat experienced in the field, we think it would be better to let the users choose and apply their own cell identification algorithms that suit their situations the best. The resulting expression matrices of this mode are thus expected to be a closer approximation to the real data, as opposed to those generated by the creation mode.

To dive into the details, Cytomulate, at its core, uses a GMM with truncation below 0 to approximate and generate the probability density of a given cell type. The Bayesian Information Criterion (BIC) (51) is adopted for model selection. Although the exploratory data analysis in the previous section shows that a MVN might be a good candidate for modeling a specific cell type in a CyTOF experiment, this should be taken with a grain of salt because the cell-typing procedure can vary across datasets and cohorts. For example, the Levine_32dim dataset provides labels for the following subtypes of B cells: mature B cells, plasma B cells, pre-B cells, and pro-B cells. In practice, however, subsets of major cell types are not available in certain datasets. We thus should allow flexibility in the model to account for cell-type procedures with different resolutions. If the users are interested in subtypes of B cells, a single MVN would be inadequate in describing the data generating mechanism (see the comparison study in the next section for details). Therefore, we chose GMMs as the base for its flexibility and its mathematical tractability.

In Fig. 2a, we outlined the overall framework of Cytomulate with input and output examples, model specification, and possible workflows for downstream analyses. Further details on the two simulation modes and the probabilistic model can be found in the Materials and methods section.

We also modularized all the aforementioned functionalities so that each of them can be easily fine-tuned or turned off with a simple interface, according to a user's specific need. Cytomulate returns the final expression matrices, the cell type labels associated with the expression matrices, as well as cellular trajectory information for each cell event. Optionally, users can also choose PyCytoData as the output option for downstream analyses compatibility (Fig. 2b). In the following sections, we conduct some comparison studies and demonstrate the usage of Cytomulate through examples.

Simulation Comparison Study

Although Cytomulate is the first-of-its-kind simulation tool designed specifically for CyTOF experiments, we will be remiss not to include benchmarks since it's a common practice in the field to adopt similar technologies from other fields to CyTOF. To showcase the superior statistical and computational properties of Cytomulate, we performed a series of empirical experiments using the emulation mode. In this simulation setting, the real benchmark datasets provide a gold standard for assessing the accuracy of each simulation method and the overall validity of the underlying model. More specifically, we focused on whether each model is capable of capturing the features of CyTOF datasets and the overall distributions and relationships between all protein channels.

For the comparison study, we included the following models from within Cytomulate and other fields:

1. Cytomulate with full covariance and model selection via Bayesian Information Criterion (BIC) from GMMs with components varying from 1 to 9 (Cytomulate)
2. Cytomulate constrained to one Gaussian component with full covariance per cell type (Dependent Gaussian or simply DG)
3. Cytomulate constrained to one Gaussian component with diagonal covariance per cell type (Independent Gaussian or simply IG)
4. Splatter (32)
5. Latent Allocation Model with Bayesian Data Analysis (LAMBDA) (12)
6. scDesign2 (30)

As discussed in the **Characteristics of CyTOF Data** section, the correlation of protein-protein expressions is an important feature of CyTOF datasets. We thus included two reduced versions of Cytomulate, namely IG and DG, both featuring only one multivariate normal component per cell type. The IG model focuses on the diagonal covariance matrix to investigate the effect of omitting the covariance structure seen in Fig. 1c. The DG model serves as a direct comparison against IG by considering the linear dependencies among protein channels. Finally, our core model, Cytomulate, aims to showcase the advantages of modeling non-normality through multiple normal components per cell type, in contrast to DG. The inclusion of these progressively complex models will not only demonstrate the flexibility of the Cytomulate

framework, but it will also validate the characteristics of CyTOF, as shown in the exploratory data analysis, by using actual benchmarks.

Beyond Cytomulate and its variants, we also included Splatter, LAMBDA, and scDesign2, which are not designed for CyTOF simulation. Below we give a brief description of each followed by the rationale for their inclusion in our comparison study.

Splatter, or specifically Splat, is originally designed for simulating scRNA-seq data. Its core model uses a Gamma-Poisson mixture to simulate counts observed in scRNA-seq experiments. The authors also take into consideration outliers, library size, mean-variance trend, and technical dropout. As Splatter expects integers as inputs, to simulate CyTOF data which usually have been transformed into a numeric scale using the arcsinh transformation, we can carry out its inverse transformation, round up the results to the nearest integer (37), apply Splatter to simulate counts, and transform the simulated expressions back into the numeric scale. We included Splatter to demonstrate the consequences of adapting scRNA-seq simulation methods to CyTOF, which outweigh the tempting benefit of convenience for any rigorous CyTOF studies.

LAMBDA is a Bayesian hierarchical model designed for automatic cell identification in CyTOF data. The model assumes that given a cell type, the cell expressions follow a zero-inflated multivariate Gaussian distribution. The cell types, on the other hand, are generated from a Categorical distribution whose parameters are controlled by some clinical information. Parameter estimation is accomplished by a stochastic EM algorithm. It is worth pointing out that LAMBDA is not a simulation tool and its simulation functionality is only a byproduct of LAMBDA. More critically, LAMBDA cannot generate new datasets without using reference datasets. Thus, for those who are interested in the Creation Mode or those who do not have

access to any previous samples (*e.g.* CyTOF for some rare diseases), this is not a suitable method at all. We included the method in our study because of its rigorous parameter estimation procedure, which would provide us with a yardstick against which we can evaluate the quality of the simulated datasets in terms of the first and second moments.

scDesign2 is an scRNA-seq simulation algorithm designed to address the shortcomings of other methods in the field, including Splatter. This algorithm first fits one of four discrete distributions (zero-inflated negative binomial, negative binomial, zero-inflated Poisson, or Poisson) to each gene's marginal distribution. Then, it employs a Gaussian copula for each cell type to achieve joint modeling, retaining the gene expressions from the real data while also capturing the correlation structure within each cell type. The key difference between scDesign2 and Cytomulate is that scDesign2 is specifically designed for discrete data, which presents similar challenges to the application of Splatter for CyTOF data.

To demonstrate the versatility of Cytomulate, six publicly available datasets from different species and diverse anatomic sites are collected: Levine_32dim (38), Levine_13dim (38), Samusik (52), CyAnno (39), Covid (15), and LG (53). If multiple samples exist for one dataset, the first sample is used. See Additional file 2 for details. All of these datasets have been bead-normalized and transformed by using the arcsinh transformation. Cell type information is also publicly available for the first four datasets. Cell types in Covid and LG were identified via FlowSOM and manual gating (15).

Cytomulate well approximates real data

In this comparison study, we make use of all six datasets and include the aforementioned five models as candidates. Four metrics are adopted to quantify the quality of the simulated datasets:

1. L2 distance between the simulated mean and the observed mean; 2. Frobenius distance between the simulated covariance and observed covariance; 3. Kullback-Leibler (KL) divergence (54) between the simulated and observed channel expressions; 4. Propensity Mean Squared Error (pMSE) (55) between the simulated and observed data. The L2 distance and the Frobenius distance are defined at the cluster level, whereas the KL divergence and pMSE are computed at the single-cell level for each cell type without using summary statistics. For pMSE, we report the relative efficiency (RE) of each competing method over Cytomulate, defined as the ratio of the pMSE value of Cytomulate vs. that of the other method. Here, for any specific method, a RE value larger than one indicates this method is more efficient than Cytomulate in terms of pMSE and by definition, the RE value for Cytomulate is always one. We also ranked the six methods based on raw values of each metric, where the best methods received an order of 6 with the worst being 1. See the Materials and methods section for further details. For each dataset, the process is repeated 50 times for 50 simulated datasets.

Using the Levine_32dim dataset, detailed comparisons show that Cytomulate outperformed all other methods benchmarked in this study in the categories of KL divergence and pMSE (Fig. 3a). As the KL divergence is a measure of similarity between two distributions, Cytomulate's resounding lead is a testament to its ability to yield a better approximation to the real data compared with other methods. Cytomulate's advantage in distributional properties logically explains its lead in pMSE, which measures whether simulated datasets are distinct from real ones. Overall, for mean estimation, all methods except Splatter and scDesign2 work well (Fig. 3b-c), with LAMBDA achieving the smallest Mean Square Error (MSE) in some datasets, closely followed by Cytomulate. Even though LAMBDA's estimation procedure should theoretically yield decent results for the first two moments, Cytomulate still outperformed

LAMBDA in estimating the covariance structure in some datasets (Fig. 3b). This result also empirically demonstrated that potential improvements can be made if only the first two moments are considered, but this improvement—if having any practical significance at all—may come at the cost of worse KL Divergence and computational efficiency. Splatter, due to the usage of nonlinear transformations and marginal modeling, performs the worst in all benchmarks by an overwhelmingly wide margin, demonstrating that Splatter is suboptimal when simulating CyTOF data. Although scDesign2 performs joint modeling, which contributes to its superior performance against Splatter, the discrete nature of its model makes it a poor choice overall.

Aggregating results from all six datasets using all four metrics, the overall trend observed previously holds for Cytomulate (Fig. 3c). By using GMMs to fit and sample from each cell type, Cytomulate gives the best approximation to the real data as measured by the KL divergence and pMSE. Averaging the orders in Fig. 3b across six datasets further highlights Cytomulate's advantages (Fig. 3d). As expected, by formally deriving a rigorous estimation procedure, LAMBDA is able to achieve the best average MSE in terms of mean while its covariance estimation lags slightly behind. However, as we have seen previously, the sufficiency of only considering the first two moments highly depends on the data and the coarseness of the cell type labels. It is also evident that Splatter always gives the worst performance due to the bias introduced by nonlinear transformations. Even working with the simple IG would yield a performance boost. scDesign2 outperforms Splatter on all metrics, but it still falls behind Cytomulate variants and LAMBDA.

To visually validate the performance of each method, we applied UMAP (17) to illustrate the resulting expression matrices of each method (Additional file 1: Figure S2). Due to the granularity of the cell type labels included in the dataset, at the first glance, all methods except

for Splatter produce reasonable simulations. Splatter produces a concerning embedding with cells splattered around the clusters and cell types indistinguishable. For all other methods, a closer examination reveals that only Cytomulate is able to preserve granular details in the simulation while other methods present only the big clusters. The preservation of detailed characteristics provides an advantage not only in preserving rare cell types and their unique phenotypes, but also in situations where cell-type resolution is low, which motivates our investigation in the following section.

Cytomulate is robust against cell-type resolution

Although cell type labels are required inputs for the emulation mode, Cytomulate does not demand highly granular labels, meaning that it is not necessary to identify all subtypes of a given cell type as a prerequisite for the simulation. In fact, different datasets and analyses call for different cell typing approaches: a large dataset with only B cells may warrant more detailed cell types to reveal their differentiation stages, whereas small whole blood samples may not afford the same resolution with reasonable cluster sizes. To account for these two situations, we conducted two separate benchmarks regarding cell typing. We first focused on a dataset with many cell types and then zeroed in on B cells to investigate the effects of subtypes.

In real data, the presence of unidentified or unassigned cells due to factors like noise, specific algorithms, and instrument panels is common, making it important to address such scenarios for reasonable results. To simulate this situation, we randomly masked some cell types in the Levine_32dim dataset and assessed the KL divergence between the simulated and original datasets. Specifically, we conducted a benchmark with three settings, randomly selecting and treating 2, 5, or 10 cell types as the same type, reflecting cases in which these cells are not

properly identified despite potentially distinct properties. Our results show that Cytomulate is unequivocally the best performer among the six methods tested (Fig. 4). As expected, when cells are unmasked (i.e. cell labels are correctly specified), the KL divergence is smaller, but the procedure of masking cells only yields a small increase in KL divergence for Cytomulate (Fig. 4). As the number of masked cell types increases, Cytomulate's advantage is further accentuated as compared to LAMBDA and scDesign2's declining performance (Fig. 4b-c). In contrast, Splatter consistently performs poorly, and the issue is further amplified when masking cell types (Fig. 4).

We have shown Cytomulate's advantage when potentially dissimilar cell types are masked. Now we proceed to investigate the effect of masking similar subtypes. As we mentioned previously, there are 4 subtypes of B cells in Levine_32dim dataset: mature B cells, plasma B cells, pre-B cells, and pro-B cells. In this benchmark, we first provided every method with these detailed subtype labels, and then amalgamated all the labels into "B cells", masking the subtype details. To quantify the differences observed across the methods, we repeated the process 20 times and computed the KL divergence (54) between the simulated and real datasets under the two settings (Fig. 5). Although the overall performance is consistently improved in the unmasked setting with more detailed cell subtypes, Cytomulate appears to be the most robust among all methods. In fact, both settings of Cytomulate outperformed all other methods regardless of cell-typing details.

To gain further insights into the advantages of Cytomulate, we used UMAP for visual comparison among the different methods under the two settings. Although most methods except for Splatter yielded acceptable approximations to the real data when subtypes are given (Fig. 5), only Cytomulate was able to differentiate the underlying clusters with minimal distortion to their

original shapes, even in the absence of detailed subtype information. Notably, LAMBDA suffered a severe loss of detail, approximating the entire dataset with a single Gaussian distribution centered between the two separated major clusters, where the actual density is low. Although scDesign2's simulation came somewhat close to capturing the two major modes, noticeable distortion within each cluster remains. Given that KL divergence is a measure of distributional differences, our quantitative assessment, combined with the visual inspection, clearly shows Cytomulate's advantage in effectively modeling heterogeneous data.

The robustness of Cytomulate makes it an appealing choice for practitioners, especially those who work with public datasets with existing cell types, since Cytomulate offers the greatest flexibility in terms of input. With reasonable cell types, Cytomulate can accurately capture the prominent characteristics of the data.

Cytomulate is efficient

As a single CyTOF experiment could contain tens of thousands or millions of cell events along with 30 - 40 protein channels, computational efficiency including estimation and simulation is a significant issue. In fact, we initially included ZINB-WaVE (31), one of the top-performing scRNA-seq techniques in our comparison studies. However, it was excluded as it would take hours to simulate one moderately-sized dataset that contains around 50,000 cells. Specifically, we measured the estimation time (time required to construct the model and estimate the parameters) and simulation time (time needed to generate data from the estimated model) as our metrics for algorithm efficiency as all methods benchmarked involve these two steps. We conducted an analysis on the running time of each method by varying the number of cells and the number of markers using the CD8+ Naive T cells in the CyAnno dataset, which contains 8,414

cell events and 39 protein channels. To be more specific, while evaluating the impact of sample size, we used all 39 protein channels and applied bootstrap sampling to generate expression matrices containing various numbers of cell events. On the other hand, we fixed the number of cell events to 8,414 and sampled the columns to be included when investigating the effect of the number of markers on processing time. Besides including the six models previously described and benchmarked for accuracy, we also included Cytomulate with 5-component GMMs (Restricted Cytomulate) to gauge the impact of model selection on efficiency.

Overall, both estimation time and simulation time for all methods under consideration appear to be polynomial with respect to the number of cells and the number of protein channels (Additional file 1: Figure S3). Cytomulate is consistently faster than LAMBDA in all situations. For a moderate-size dataset containing 100,000 cells, even Cytomulate with model selection from 9 models takes around only 2 minutes (Additional file 1: Figure S3a). If restrictions are imposed, the estimation can be done within seconds. On the other hand, LAMBDA, due to the usage of a stochastic EM algorithm for the model estimation, is twice as slow, while scDesign2 is also among the slowest methods tested. When sample size reaches one million cells, both LAMBDA and scDesign2's estimation times exceed one hour, which makes them not viable for large scale simulations. Surprisingly, LAMBDA and scDesign2 both suffer from poor scalability with regard to the number of markers as compared with other methods (Additional file 1: Figure S3b; Additional file 1: Figure S3d)

In terms of simulation time (Additional file 1: Figure S3c-d), all methods are relatively efficient as compared to estimation time. Despite the efficiency of all methods here, Cytomulate is always best with or without model selection, while Splatter is the least efficient. This characteristic allows Cytomulate to quickly generate many replicates using a single dataset as a reference,

which is a common task. As a unique advantage to the Emulation mode, Cytomulate can perform parameter estimation using a small sample and then scale up by simulating a much larger sample. This further improves the scalability of Cytomulate given its efficiency in terms of simulation time. With the increasing capability of CyTOF to produce larger datasets and more channels, Cytomulate is fully capable of handling large datasets.

Complex simulations:

The core model of Cytomulate accounts for the expressions of each cell type as an essential part for simulation. Beyond the core model and unlike LAMBDA, Cytomulate can easily be extended to incorporate more complex effects commonly encountered in CyTOF experiments such as batch effects, temporal effects, as well as cell differentiations. While some of these effects are oftentimes considered nuisances with corrections desired, the ability to approximate them is crucial for methodological developments in the field of normalization or other pre-processing techniques. Also, the rising potential of trajectory inference with CyTOF datasets instead of scRNA-seq calls for a more thorough consideration of this topic. This section details the overall design and implementation of these complex simulation scenarios.

Temporal effects:

As discussed in (5), due to cell debris, etc. the sensitivity of a CyTOF machine may vary over time during data generation, causing signals for cells from one batch to fluctuate. In Cytomulate, we assume that the temporal effect for each batch is generated by a function defined on the unit interval such that the value at time zero is always zero, representing that there is no temporal

effect at the beginning. We provide users with 3 options to simulate a temporal effect, each with a different level of control over the overall shape of the temporal effect:

1. Brownian Bridge (56): The shape of the entire trajectory is randomly generated from a Brownian Bridge where the ending value is randomly sampled from a normal distribution. To evaluate the function on the unit interval, the Akima spline (57) is used for interpolation. This option gives the users minimal control over the evolution of the temporal effects.

2. Polynomial: This option allows the user to specify the rough shape of the trajectory *via* the coefficients of a polynomial. A linear transformation will be used to make sure the resulting temporal effect starts at 0 and ends at a value randomly generated from a normal distribution.

3. Spline: During the preprocessing steps, it is possible that the users can extract the temporal effect component. The spline option permits the users to add the original temporal effect back to the simulated dataset via a smoothing spline. Since the size of the simulated dataset might differ from that of the original dataset, Cytomulate will automatically rescale the time interval so that the entire trajectory could be retained in the simulated dataset (Fig. 6a).

Batch effects:

Previous research in the field (7), has indicated that the batch effects in CyTOF experiments vary not only by batches but also by cell types and protein channels. To simulate batch effects, we decompose the effects into global batch effects which assume different values for different batches and local batch effects which vary depending on the cell types and protein channels in a similar fashion to the way we decompose effects into main effects and interactions in Analysis of Variance (ANOVA). Further details can be found in the Materials and methods section. As a

demonstration, we simulated two batches with 3 cell types using the creation mode of Cytomulate (Fig. 6b).

Cellular trajectories:

To simulate cellular trajectories, Cytomulate adopts a two-step procedure. In the first step, Cytomulate simulates the underlying graph of the cellular trajectories. Given the simulated graph, the actual protein expressions are then sampled in the second step.

Specifically, in the first step, a polytree (a directed acyclic graph whose underlying undirected graph is a tree) or a polyforest (a collection of nonoverlapping polytrees) is generated (58). A tree represents a single trajectory, where edges connect different cell types, allowing cells to differentiate along defined paths. In the Creation Mode, this is done by first randomly grouping the cell types and applying Kruskal's algorithm (59) with random weights to sketch the underlying tree or forest. In the Emulation Mode, this is accomplished by the Clauset-Newman-Moore greedy modularity maximization algorithm (60) (61) followed by the Kruskal's algorithm with the L2 distances among observed average cell means as weights. In both circumstances the root node represents the starting point of differentiation, where its child nodes inherit certain channels that define the entire lineage. To find the root node and convert the graph to a directed acyclic graph (DAG), a depth-first search with a randomly selected node as the root on each graph component is performed.

In the second step, Cytomulate starts the differentiation process. For select channels that change between each pair of nodes, a Brownian Bridge combined with interpolating splines is used to generate the actual protein expressions along the polytree or the polyforest in a fashion similar to how Splatter (32) simulates paths. The time parameter of differentiation is determined by a Beta

distribution. This scheme allows smooth transitions between cell types while also ensuring incremental changes along the tree (*i.e.* each tree contains a family of cell types with gradual changes). Further details can be found in the Materials and methods section. For illustration, we used the creation mode of Cytomulate to generate 5 cell types aligned on one polytree (Fig. 6c).

Applications of Cytomulate within a Unified Framework:

Framework Overview:

To ensure the interoperability of Cytomulate with downstream analyses and related workflows, we designed our software package using open standards in the field. Namely, our framework is expression matrix-driven with the optional support for PyCytoData—an easy-to-use interface for CyTOF and downstream analyses (Fig. 2b). Aside from two simulation modes, we support two main usage modes of Cytomulate. For users who prefer batch processing and working with servers, we recommend our command-line interface (CLI). Using a single command with the appropriate arguments and flags, users can generate their desired dataset and save it to disk in the format of their choice. Further, the CLI ensures scalability and interoperability by allowing users to utilize shell scripts and seamlessly integrate their results with other up- or downstream software. Alternatively, our interactive Python mode is designed for those who primarily work within Python and would benefit from the most flexibility. This interface supports fine-tuning of details while allowing users to explore the datasets interactively: from simulation to dimension reduction (DR) and plotting, a few lines of readable code are all they need.

With the wide adoption of CyTOF and the subsequent proliferation of analysis methods, a unified framework is increasingly important. While few methods, except ubiquitous tools such as

PCA, have become virtually platform- and language-agnostic, many otherwise good software suffers from poor usability and extensibility. For example, in the process of evaluating Cytomulate, we found that LAMBDA falls far short of having a user-friendly interface: in fact, we spent a significant amount of time adapting the algorithm to our workflow for it to be viably assessed. To address these issues, we first developed the two usage modes as previously mentioned. The CLI uses the operating system's shell as a common denominator to work with other parts of the workflow as necessary, whereas the interactive mode supports outputting a PyCytoData object. The latter brings the benefit of a suite of integrated downstream analysis tools and utility functions, including builtin normalization and CytofDR for DR evaluation (15). Beyond the software itself, our continuous integration and development workflows along with detailed documentation (https://cytomulate.readthedocs.io) ensure that Cytomulate is well maintained and easy for any user to adopt with minimal to no learning curve.

To demonstrate the usability and applicability of Cytomulate in CyTOF research, we hereby showcase three benchmark studies conducted using both Creation and Emulation modes. We also would like to emphasize the fact that Cytomulate's potential goes much beyond the particular methods and benchmarks conducted in this section. With the flexibility of our framework and the open source initiative, other researchers can effortlessly extend the functionalities by simply using the PyCytoData object with their own workflows. A crucial advantage of employing Cytomulate instead of relying solely on real datasets is that users can specify characteristics of each sample, a task hardly achievable in real datasets without artificially subsampling and manipulations. Indeed, a proper benchmark requires the detailed controls necessary to distinguish between various effects, such as cell abundance and differentiation paths. Cytomulate not only supports all these configurations with both of our

interfaces but also produces sensible results for practitioners as seen in the following benchmarks.

Benchmarking DR using Cytomulate:

As a practical application of Cytomulate, a previous DR benchmark study (15) included extensive usage of the creation mode to evaluate DR methods. Specifically, Wang et al. simulated 425 CyTOF datasets by specifying and systematically varying key features of CyTOF expressions. To further demonstrate the capabilities of Cytomulate for downstream analyses, we accessed and analyzed a subset of the results published on Database of Actionable Immunology's (DBAI) (62–64) CytofDR Playground (https://dbai.biohpc.swmed.edu/cytof-dr-playground/). In particular, we chose samples with 100,000 cells and the following parameters: the number of markers {30, 35, 40}, the number of cell differentiation paths {2, 3, 5}, and the number of cell types {5, 10, 15, 20, 30}. These criteria yield 225 datasets in total with 25 replicates for each configuration. As for DR methods, we focused on the top-performing and most popular ones in the field according to the benchmarks: MDS (65), tSNE (16), UMAP (17), PCA (66), and SAUCIE (19).

Using two major categories (Global Structure Preservation and Local Structure Preservation) and the two subcategories of the Downstream Category, we validated the previous results that MDS is superior for structure preservation whereas UMAP and SAUCIE are better for clustering-based metrics (Additional file 1: Figure S4). These results are indeed concordant with results obtained from real datasets. Further, Wang et al. (15) demonstrated that Cytomulate produced good datasets for both downstream analyses and DR visualizations. While we defer to the original DR paper for those interested in the details of DR methods and their performances with regard to

simulation settings, we would like to emphasize again the unified framework that enables this type of analysis. Users of Cytomulate can easily generate datasets and replicate the aforementioned DR results by using Cytomulate in conjunction with CytofDR through the PyCytoData framework.

Validating Clustering Performance using Cytomulate:

Clustering is a common and important workflow in the analysis of real CyTOF data because of its role in identifying similar cells. Many methods have been developed for CyTOF data, such as FlowSOM (11) and flowMeans (67). Traditional methods, such as K-Means (46), can in theory be applied to CyTOF data as well, but their performance is questionable. Two previous review papers (47,48) have extensively benchmarked popular clustering methods, and FlowSOM and flowMeans have in general been recommended. However, neither paper included simulation data as part of their workflows. To fill in the gap by incorporating Cytomulate as part of this analysis, we conducted a benchmark of three popular clustering methods to validate conclusions from the previous review papers.

In both Creation and Emulation modes, the ground truth labels are known, thus rendering Cytomulate a gold standard to assess whether clustering algorithms can correctly group the same cells in the same cluster. To demonstrate Cytomulate's advantage, we simulated 20 datasets from the Levine_32dim dataset (40). Under the default setting with fixed probability of cell abundance (Additional file 1: Figure S5a), we found that flowMeans performs the best out of the three methods, achieving the highest Adjusted Rand Index (ARI) (68). FlowSOM follows closely behind flowMeans, whereas K-means lagged severely behind. To account for various cell abundances in real datasets, we also tested a setting by randomly varying the cell abundance

probabilities generated from a Dirichlet distribution with parameters being a 14-dimensional vector of 0.1. While the order of the three methods remained the same, K-means's suboptimal characteristics for CyTOF data are exacerbated. Our results not only agree with the numerical analyses on Levine_32dim by Liu et al. (47), but they also highlight the usefulness of Cytomulate to simulate user-controlled characteristics of CyTOF, such as cell type abundances.

Comparing Batch Normalization Methods using Cytomulate:

As part of Cytomulate's complex simulation, users can easily simulate multiple batches of CyTOF datasets by adding batch effects. This procedure is designed to mimic the real situation in which practitioners have to account for systematic differences across multiple batches. With the development of many batch correction methods in the field, including ComBat (69), CytoNorm (7), and Harmony (70), it is important to assess how well each performs. Cytomulate's complex simulation provides a good basis for such a benchmark. We thus selected the aforementioned three methods and compared their performances using 10 paired simulated samples based on the Samusik (52) dataset. Within each sample, an "anchor" sample that is free of batch effect was first generated. A "replicate" sample was then generated with simulated batch effects.

To assess the extent to which batch effect is corrected, we computed the Earth Mover's Distance (EMD) (71) between corrected batches, and smaller distance corresponds to a more effective batch normalization. Further, we created two scenarios: "High Variance" and "Low Variance" to represent two levels of potential amount of batch variation from sample to sample. In the case of "High Variance", the batch effects were simulated from a normal distribution with zero mean and unit standard deviation. The standard deviation was set to 0.1 in the case of "Low Variance".

Across both high and low variance settings, CytoNorm is the best for CyTOF data, while Harmony is the worst. Also, the low variance setting yields overall smaller EMD as we expect. This benchmark shows again that methods designed specifically for CyTOF tend to outperform others, such as Harmony for scRNA-seq and ComBat for RNA-seq.

Discussion and conclusions

Simulation is a powerful, flexible, and cost-effective tool for performance evaluation, which plays a vital role in methodological development. A common way to test analysis methods is to conduct simulation studies, where generating multiple datasets, with different parameters or assumptions, can be quickly accomplished with minimal cost. The lack of a formal method for accurately synthesizing CyTOF data warrants a dedicated statistical tool with careful considerations of CyTOF's unique characteristics, such as its throughput and the correlation between protein channel expressions.

In this paper, we filled in the aforementioned void by introducing Cytomulate, a tool for simulating CyTOF data accurately and efficiently. It builds on Gaussian Mixture Models (GMM) with a fast and robust estimation procedure and appealing mathematical properties to approximate the real densities of CyTOF channels. Furthermore, Cytomulate features a well-defined probabilistic model to capture prominent aspects of CyTOF data while also incorporating two simulation modes: creation mode and emulation mode to cater to the needs of various researchers. By thoroughly benchmarking Cytomulate's estimation and simulation efficiency, resemblance to real data, and robustness again prior cell-type resolution on six publicly available datasets, we demonstrated that Cytomulate is superior in approximating the overall data

distributions regardless of the granularity of the input. Our algorithm further has the versatility of simulating more complicated phenomena commonly encountered in a CyTOF experiment such as signal drifts (temporal effects), batch effects, and cellular trajectories, which are partially or totally unsupported by other tools. Finally, we showcased Cytomulate's integration with PyCytoData as an open standard for CyTOF data processing in Python while also highlighting Cytomulate's usage in a series of benchmark studies involving several DR methods (15), clustering algorithms, and batch correction tools. By integrating the philosophy of user-centric software and the superior performance of our probabilistic model, Cytomulate offers a practical step towards addressing the usability problem in the field (15,45).

Potential limitations of this study involve the estimation procedure we chose for the emulation mode. Since the parameters are estimated for each cell type separately, in the presence of high-dimensional data, the usage of the EM algorithm could render the performance of a GMM unsatisfactory when the number of cell events is low. A potential remedy for rare cell types would be to borrow "strength" from other cell types to stabilize the parameter estimation via a Bayesian hierarchical setup. What's more, as a trade-off for efficiency, various aspects of CyTOF data such as means, covariances, and cell differentiations are captured in a sequential manner instead of being modeled jointly. Consequently, although the GMMs provide a reasonable approximation to the data distribution of CyTOF due to its universal approximation to densities (41), the estimated parameters are likely biased. Due to the sheer number of cells typically encountered in CyTOF and the complexity of the model, traditional inference methods such as Markov Chain Monte Carlo (MCMC) might not ameliorate this situation. Fortunately, tremendous progress has been made in the field of deep generative models, which has been successfully applied to tasks such as estimating densities and synthesizing realistic-looking

images on large datasets (72,73). Some typical examples include normalizing flow (74), variational auto-encoder (75), and generative adversarial networks (76). Exploiting said methods to gain deeper insights into CyTOF datasets thus could be a possible future direction.

Another future direction constitutes the modeling of clinical phenotypes based on simulated CyTOF datasets. Namely, suppose we denote y as the phenotype and x as CyTOF data, and an interesting direction is to simulate y from $p(y|x)$. Currently, Cytomulate has the ability to generate CyTOF datasets from $p(x)$. Adding the connection between simulated datasets and phenotypes will increase the applicability and practicality of Cytomulate in real settings. We are optimistic that our simulation framework as a first in the field will spark further interests in these exciting directions, and our model will serve as a reference for future works.

Materials and methods

Probabilistic model of Cytomulate:

The probabilistic model of Cytomulate is intended to model the $arcsinh\ (./5)$-transformed CyTOF data. In this section, we will gradually build up towards the final model aided by what we observed in the section on characteristics of CyTOF data.

Basic model

We will start with the simplest scenario where we only have one batch, $n = 1, \cdots, N$ cell events on only one cell type with measurements on $m = 1, \cdots, M$ protein channels.

We start with a continuous latent variable for the protein expressions:

$$X_n \sim f(\mu, \Sigma),$$

where $f(\mu, \Sigma)$ denotes some distribution with an M-by-1 matrix of mean μ, and an M-by-M matrix of variance-covariance matrix Σ. Given the near-Gaussianity nature of CyTOF data, we use a Gaussian Mixture Model to approximate this distribution in Cytomulate.

Since protein expressions should be non-negative, we further truncate the variable below zero. Mathematically, this can be accomplished by

$$I(X_n \succ 0) \circ X_n ,$$

where $\circ$ is the Hadamard product, I is an indicator function, and $\succ$ is a component-wise inequality.

Finally, we incorporate the machine noise into the final protein expressions resulting in the basic model of Cytomulate:

$$Y_n = I(X_n \succ 0) \circ X_n + E_n ,$$

where $E_n \sim L$ and L is some error distribution.

Core model

The core model of Cytomulate is only a slight generalization upon the basic model by considering $p = 1, \cdots, P$ cell types in one batch.

We denote the probability of observing the pth cell type by π_p . Then, for the nth cell event, we assume that the cell type is sampled from

$$\Pi_n \sim Categorical(\pi),$$

where $\pi = [\pi_1, \cdots, \pi_P]$ is a P-by-1 probability vector.

Suppose the cell type for the nth cell event is p.We then update the notation for the latent variables accordingly:

$$X_n \sim f^p(\mu^p, \Sigma^p),$$

where $f^p(\mu^p, \Sigma^p)$ denotes some distribution with mean μ^p, and variance-covariance matrix Σ^p.

The computation of the final protein expressions remain unchanged:

$$Y_n = I(X_n > 0) \circ X_n + E_n,$$

where $E_n \sim L$ and L is some error distribution.

Complex simulations

Batch effects

To add batch effects, suppose we have $b = 1, \cdots, B$ batches, $m = 1, \cdots, M$ protein channels, and $p = 1, \cdots, P$ cell types. For the bth batch, we have $n = 1, \cdots, N^b$ cell events.

Similar to the setup in the core model, we assume that the cell type of the nth cell event in the bth batch is sampled from

$$\Pi_n^b \sim Categorical(\pi^b),$$

where $\pi = [\pi_1^b, \cdots, \pi_P^b]$ is a P-by-1 probability vector. The corresponding latent variable is then denoted by $X_n^b \sim f^p(\mu^p, \Sigma^p)$. Note that while we allow the cell abundances to vary among batches, we only use one distribution for one cell type.

With this setup, we introduce a set of random vectors, each representing the batch effect on the pth cell type in the bth batch:

$$\Psi^{b,p} \sim N(\psi^b 1_{M\times 1}, \sigma^2 I_{M\times M}),$$

where

$$\psi^b \sim N(0, \sigma_\psi^2)$$

represents the center around which the batch effects on the cell types and protein channels in the bth batch vary. We further add two sets of constraints to ensure the identifiability of the parameters:

$$\sum_{b=1}^{B} \psi^b = 0,$$

$$\sum_{p=1}^{P} \Psi_m^{b,p} = \sum_{m=1}^{M} \Psi_m^{b,p} = 0.$$

Temporal effects

Suppose we have $b = 1, \cdots, B$ batches, and $m = 1, \cdots, M$ protein channels. For the bth batch, we have $n = 1, \cdots, N^b$ cell events. To add temporal effects to the nth cell event in the bth batch, we assume that there is a stochastic process:

$$H^b : [0,1]^M \mapsto R^M,$$

such that $H^b(0) = 0$. This reflects our belief that there should be no temporal effects at the beginning of an experiment. Denote the realization of H^b by Φ^b, then the temporal effect on the nth cell event is calculated as:

$$\Phi^b\left(\frac{n}{N^b}\right).$$

Cellular trajectories

The setup of simulating cellular trajectories is a little more involved. Again, we assume that we have $b = 1, \cdots, B$ batches, $m = 1, \cdots, M$ protein channels, and $p = 1, \cdots, P$ cell types. For the bth batch, we have $n = 1, \cdots, N^b$ cell events.

To get started, we first fix cell type p. We suppose that it differentiates into z^p "children" denoted by

$$p^{c_1}, \cdots, p^{c_{z^p}}.$$

In Cytomulate, we ensure that the collection of "parent-children" relationships form a polytree (a directed acyclic graph whose underlying undirected graph is a tree) or a polyforest (a collection of nonoverlapping polytrees) (58).

To accomplish this, a weighted complete undirected graph with cell types as nodes are constructed. In the Creation Mode, the weights on the edges are randomly assigned. In the Emulation Mode, the L2 distances among observed average protein expressions of all cell types are used as weights.

Next, we carve out a spanning tree or a spanning forest via the Clauset-Newman-Moore greedy modularity maximization algorithm (60) (61) followed by the Kruskal's algorithm (59). Finally, to convert the graph to a directed acyclic graph (DAG), a depth-first search with a randomly selected node as the root on each graph component is performed.

Now fix the "child" cell type p^c. We construct the corresponding cellular trajectory from cell type p to cell type p^c in a similar fashion to how we construct temporal effects. We start with a stochastic process

$$G^{p\to p^c}: [0,1]^M \mapsto R^M,$$

such that the mth component satisfies

$$G_m^{p\to p^c}(0) = 0, \text{ and } G_m^{p\to p^c}(1) = \mu_m^{p^c} - \mu_m^p,$$

where μ_m^p and $\mu_m^{p^c}$ are the mean expression levels of the mth protein channel of cell type p and cell type p^c, respectively. Given these two constraints, a Brownian Bridge becomes an intuitive choice for the task of trajectory generation since it is a Brownian Motion with an extra condition on where the position of the process should be at time 1. To simulate such a stochastic process that starts at 0 and ends at $\mu_m^{p^c} - \mu_m^p$, we utilize a standard Brownian Motion B_t as the basic building block and simply let $G_m^{p\to p^c}(t) = t(\mu_m^{p^c} - \mu_m^p) + B_t - tB_1$. This stochastic process is also adopted by Splatter (32) in generating the actual realizations of cellular trajectories denoted here by $g^{p\to p^c}$. When generating the final protein expressions for the nth cell event in the bth batch whose cell type is p, we first sample the "child" cell type p^c from

$$\Omega_n^{b,p} \sim Categorical(\omega^p),$$

where $\omega^p = [\omega_1^p, \cdots, \omega_{z^p}^p]$ is a z^p-by-1 probability vector. Notice that the parameters of the distribution only vary with cell types.

We then sample its mth component of the actual position t_n^b on the cellular trajectory $g^{p\to p^c}$

$$T_{n,m}^b \sim Beta(\alpha^b, \beta^b).$$

Final model

With these setups, the final protein expressions for the nth cell event in the bth batch whose cell type is p and who is differentiating into cell type p^c can be simulated via the following two steps:

$$Z_n^b = X_n^b + g^{p \to p^c}(T_n^p) + \Psi^{b,p} + \Phi^b \,,$$

$$Y_n^b = I(Z_n^b \succ 0) \circ Z_n^b + E_n^b,$$

where $E_n^b \sim L$ and L is some error distribution.

Estimation procedure for the emulation mode

Given the expression matrix along with the cell type label for each cell event, Cytomulate first groups the observations by cell types and associates with each type a GMM whose parameters are then estimated via the EM algorithm adopted by the *scikit-learn* Python package (77). To further improve the approximation of a GMM to CyTOF data, for each cell type, we match the probability of obtaining a near-zero value estimated by the model with the actual frequency observed in the data. This is done by randomly zeroing out simulated values with probabilities inversely proportional to their magnitudes.

Evaluation metrics

In this section, we elaborate on the metrics we used in the comparison studies. Suppose in the simulated batch we have $p = 1, \cdots, P$ cell types. Then, for the pth cell type, we calculate the sample mean m_p and the sample covariance as c_p. We carry out the same procedure on the

observed dataset to the corresponding m_p^r and c_p^r. The L2 distance between the simulated and observed means is found via $dm_p = ||m_p - m_p^r||_2$. The Frobenius distance (78) between the simulated and observed covariances is calculated as $dc_p = ||c_p - c_p^r||_F$. The KL divergence dk_p is estimated using the method proposed in (54); the pMSE $pmse_p$ is calculated using the R package *synthpop* (55). Finally, the discrepancy measures for all the cell types are aggregated using a weighted average where the weights are determined by the cell abundances.

Software implementation details

All analyses, plotting, and benchmarks were performed using the R (4.0.2) and Python (3.8.8) programming environments in MacOS 12.4. All default settings were used for Cytomulate, Splatter, and scDesign2. For LAMBDA, we skipped the multinomial sampling process given that all cell types were known.

Cytomulate is implemented and supported in Python (v3.5 or later), and the CLI is available on all platforms and systems where a compatible python interpreter is present with the proper package installations. The detailed documentation and tutorials are freely available and hosted on the cloud using ReadTheDocs (https://cytomulate.readthedocs.io). All software features have been thoroughly tested with the *pytest* framework.

For the efficiency benchmarks, we carried out the experiment on a MacBook Pro with a 2.9 GHz Quad-Core Intel Core i7 processor running MacOS 13.0.1. Since Cytomulate is implemented in Python while Splatter, LAMBDA, and scDesign2 are R packages, we used Python's time module to measure the performance of Cytomulate-related methods and Sys.time function in R to time Splatter, LAMBDA, and scDesign2. All times are reported in seconds.

Statistical analyses

We employed Pearson correlation coefficients for all correlation matrices and plots. For all order-based metrics, higher orders represent better performance. In cases where ties occurred, we used the average order of tied values. 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 barplots showing discrepancies among all methods under comparison on various evaluation metrics were plotted on the logarithm scale. To increase visual clarity, we added one to all the data points before plotting so that the results would all be positive. Error bars with width being two standard deviations were also added to further facilitate comparison.

Figures and Table Legends

Fig. 1 Characteristics of a typical CyTOF dataset. All four panels are from the Levine_13dim as an example. (a) Heatmap of channel expression for all naive CD4+ T cells. (b) Density plot of all channel expressions from naive CD4+ T cells. (c) Correlation matrix of all channels as computed with Pearson correlation coefficients. (d) Scatter plot of all cells using the first two principal components and colored with provided cell types.

Fig. 2 Overview of Cytomulate model and software pipeline. (a) The model structure and input and output pipelines of Cytomulate. The input stage demonstrates the parameters and reference data needed for each mode, and Cytomulate uses the provided information accordingly for estimation and simulation. The outputs are in the forms of expression matrices and their associated cell types for additional downstream analyses. (b) The software pipeline of

Cytomulate. The flowchart showcases Cytomulate's role and integration with industry standard tools in every step of the research and development process.

Fig. 3 Detailed results on simulation performance benchmarks. (a) The performance of all methods in the Levine_32dim dataset. The four panels correspond to MSE in mean (smaller is better), MSE in covariance (smaller is better), KL Divergence (smaller is better), and relative efficiency in pMSE (larger is better). The relative efficiency is measured using Cytomulate's pMSE: values smaller than 1 indicate worse performance than Cytomulate. (b) Orders of each method across all datasets and all metrics. Orders are based on values from the metrics, and a higher order indicates superior performance in the given metric. (c) The average performance of methods averaged across all datasets. Orange colors indicate superior performance, whereas blue colors represent worse performance. The interpretations are the same as (a). (d) Orders of all methods averaged across all datasets. Higher bars represent a higher average order for a given metric and method. Colors represent different simulation methods.

Fig. 4 Performance of simulation methods against random masking of cell subtypes. For all three panels, randomly cell types of the Levin_32dim were masked and performance of each method was measured using log KL divergence plus 1. Blue bars represent the masked datasets, and yellow bars represent matched, unmasked controls (N=20). The whisker on top of the bars indicates the variance. (a) 2 cell types are randomly masked in each replicate. (b) 5 cell types are randomly masked in each replicate. (c) 10 cell types are randomly masked in each replicate.

Fig. 5 Additional results on performance with different cell subtypes. All parts of the figure utilizes the Levine_32dim dataset with B cells only. (a) UMAP embedding of the original B cells present in the Levine_32dim dataset overlaid with B cell subtype labels as well as a density

estimation. (b) Masked and unmasked performance of all methods measured by log KL divergence plus 1. Color blue represents masked datasets while yellow stands for unmasked datasets. The error bars show the variance. (c)-(h) UMAP embeddings of the following simulation methods with and without B cell subtype labels: Cytomulate, DG, LAMBDA, IG, scDesign2, and Splatter.

Fig. 6 Examples of effects addressed in complex simulations. (a) Example temporal effects generated using the three methods: Brownian Bridge, Polynomial, and Spline. (b) Example of a simulated sample batch effect. Color difference represents batches, and point shape indicates cell types. (c) Example of the cell differentiation trajectory generated by Cytomulate. The trajectories are noted with light blue arrows.

Additional file 1: Figure S1 Examples of complex characteristics observed in real datasets. (a) The temporal effect present in the real PBMC Finck dataset, as measured by the signal drift of five bead channels. Colors represent the different bead channels. (b) The corrected time effects after bead normalization. Colors represent the different bead channels. (c) The batch effect of the two patients in Levine_13dim visualized with an interaction plot in which the mean expression of protein channels change from sample to sample. Each color is a protein marker. (d) An example cellular trajectory fitted using the CytoTree package based on a real Bone Marrow dataset from the CytoTree documentation. Each node is a meta-cell colored by the cell type, and each edge represents the path.

Additional file 1: Figure S2 Detailed results on embeddings of the original and simulation datasets. (a) The UMAP embedding of the original Levine_32dim dataset. In (b), (c), (d), (e), (f), and (g), the UMAP embeddings of simulated datasets generated by Cytomulate, LAMBDA,

scDesign2, Splatter, DG, and IG are plotted respectively. Each point represents a cell in the embedding space, and cells are colored by their cell types. Red circles are overlaid to pinpoint the locations of small clusters present in the original dataset.

Additional file 1: Figure S3 Runtime benchmarks of all simulation methods in the study. In (a) and (b), we measured the estimation time of each method with various numbers of cells and markers respectively. The estimation time includes construction of the model object and estimation of parameters used in each model based on the dataset given. In (c) and (d), we benchmarked the simulation time of each method against different numbers of cells and markers. Simulation time is defined as the time necessary to generate data from the estimated parameters.

Additional file 1: Figure S4 Performance of DR methods benchmarked with Cytomulate's creation mode. Five methods are included, and all rankings are accessed from CytofDR Playground. "*" indicates 10% downsampling of the datasets used for DR and evaluation. (a) Rankings of DR methods using global structure preservation metrics. (b) Rankings of DR methods using local structure preservation metrics. (c) Rankings of DR methods using cluster reconstruction metrics. (d) Rankings of DR methods using cluster concordance metrics.

Additional file 1: Figure S5 Clustering and batch correction benchmark results using Cytomulate. (a) The Adjusted Rand Index (ARI) of three clustering methods across 20 replicates. Higher is better. (b) The Earth Mover's Distance (EMD) between batches for three batch correction methods benchmarked across 10 replicates. Smaller is better.

Additional file 2 Datasets used in the study. A table that provides the dataset name, species, the anatomic site, and its source for accession.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The Levine_32dim, Levine_13dim, and Samusik datasets can be accessed from the HDCytoData package in R (https://github.com/lmweber/HDCytoData). The CyAnno (39) and LG (53) can be accessed from their respective publications. The Covid cohort is hosted on DBAI (https://dbai.biohpc.swmed.edu/) with the following accession codes: CyTOF_00000000000001 to CyTOF_00000000000012.

The Cytomulate algorithm is available at cytomulate.readthedocs.io, and also shared through the Database of Actionable Immunology (dbai.biohpc.swmed.edu) (62–64).

Competing interests

Dr. Tao Wang is one of the scientific co-founders of NightStar Biotechnologies, Inc. All remaining authors declare no competing interests.

Funding

This study was supported by the National Institutes of Health (NIH) [NIH 1R01CA258584/TW, XW, U01AI156189/TW], Cancer Prevention Research Institute of Texas [CPRIT RP230363/TW, RP190208/TW, XW], and Southern Methodist University [Dean's Research Council Award 2022/XW, TW].

Authors' contributions

Y.Y. performed all analyses. Y.Y. and K.W. created the *Cytomulate* Package. Z.L provided analysis pipelines for LAMBDA. X.W. conceived the study. T.W. and X.W. designed and supervised the whole study. All authors wrote the manuscript.

Acknowledgements

Not applicable

Bibliography

1. Lou X, Zhang G, Herrera I, Kinach R, Ornatsky O, Baranov V, et al. Polymer-based elemental tags for sensitive bioassays. Angew Chem Int Ed. 2007;46(32):6111–4.
2. Spitzer MH, Nolan GP. Mass cytometry: single cells, many features. Cell. 2016 May 5;165(4):780–91.
3. Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. Anal Chem. 2009 Aug 15;81(16):6813–22.
4. Chester C, Maecker HT. Algorithmic Tools for Mining High-Dimensional Cytometry Data. J Immunol. 2015 Aug 1;195(3):773–9.
5. Finck R, Simonds EF, Jager A, Krishnaswamy S, Sachs K, Fantl W, et al. Normalization

of mass cytometry data with bead standards. Cytometry A. 2013 May;83(5):483–94.

6. Trussart M, Teh CE, Tan T, Leong L, Gray DH, Speed TP. Removing unwanted variation with CytofRUV to integrate multiple CyTOF datasets. eLife. 2020 Sep 7;9.

7. Van Gassen S, Gaudilliere B, Angst MS, Saeys Y, Aghaeepour N. Cytonorm: A normalization algorithm for cytometry data. Cytometry A. 2020 Mar;97(3):268–78.

8. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, et al. Removal of batch effects using distribution-matching residual networks. Bioinformatics. 2017 Aug 15;33(16):2539–46.

9. Qian WW, Xia C, Venugopalan S, Narayanaswamy A, Dimon M, Ashdown GW, et al. Batch equalization with a generative adversarial network. Bioinformatics. 2020 Dec 30;36(Suppl_2):i875–83.

10. Upadhyay U, Jain A. Removal of Batch Effects using Generative Adversarial Networks. arXiv. 2019;

11. Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, et al. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry A. 2015 Jul;87(7):636–45.

12. Abe K, Minoura K, Maeda Y, Nishikawa H, Shimamura T. Model-based clustering for flow and mass cytometry data with clinical information. BMC Bioinformatics. 2020 Sep 17;21(Suppl 13):393.

13. Ji D, Nalisnick E, Qian Y, Scheuermann RH, Smyth P. Bayesian trees for automated cytometry data analysis. BioRxiv. 2018 Sep 19;

14. Dai Y, Xu A, Li J, Wu L, Yu S, Chen J, et al. CytoTree: an R/Bioconductor package for analysis and visualization of flow and mass cytometry data. BMC Bioinformatics. 2021 Mar 22;22(1):138.

15. Wang K, Yang Y, Wu F, Song B, Wang X, Wang T. Comparative analysis of dimension reduction methods for cytometry by time-of-flight data. Nat Commun. 2023 Apr 1;14(1):1836.

16. Laurens van der M, Geoffrey H. Visualizing Data using t-SNE. J Mach Learn Res. 2008;

17. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2018 Dec 3;37:38–44.

18. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer; 2008.

19. Amodio M, Srinivasan K, van Dijk D, Mohsen H, Yim K, Muhle R, et al. Exploring Single-Cell Data with Multitasking Deep Neural Networks. BioRxiv. 2017 Dec 19;

20. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun. 2018 May 21;9(1):2002.

21. Hahne F, LeMeur N, Brinkman RR, Ellis B, Haaland P, Sarkar D, et al. flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinformatics. 2009 Apr 9;10:106.

22. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.

23. Seiler C, Ferreira A-M, Kronstad LM, Simpson LJ, Le Gars M, Vendrame E, et al. CytoGLMM: conditional differential analysis for flow and mass cytometry experiments. BMC Bioinformatics. 2021 Mar 22;22(1):137.

24. Arend L, Bernett J, Manz Q, Klug M, Lazareva O, Baumbach J, et al. A systematic comparison of novel and existing differential analysis methods for CyTOF data. Brief Bioinformatics. 2022 Jan 17;23(1).

25. Mitra R, Müller P, Qiu P, Ji Y. Bayesian hierarchical models for protein networks in single-cell mass cytometry. Cancer Inform. 2014 Dec 10;13(Suppl 4):79–89.

26. Abdelaal T, Höllt T, van Unen V, Lelieveldt BPF, Koning F, Reinders MJT, et al. CyTOFmerge: integrating mass cytometry data across multiple panels. Bioinformatics. 2019 Oct 15;35(20):4063–71.

27. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40.

28. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016 Apr 27;17:75.

29. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. PLoS Comput Biol. 2015 Jun 24;11(6):e1004333.

30. Sun T, Song D, Li WV, Li JJ. scDesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. Genome Biol. 2021 May 25;22(1):163.

31. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018 Jan 18;9(1):284.

32. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017 Sep 12;18(1):174.

33. Crowell HL, Morillo Leonardo SX, Soneson C, Robinson MD. The shaky foundations of simulating single-cell RNA sequencing data. Genome Biol. 2023 Mar 29;24(1):62.

34. Cao Y, Yang P, Yang JYH. A benchmark study of simulation methods for single-cell RNA sequencing data. Nat Commun. 2021 Nov 25;12(1):6911.

35. Schissler AG, Bedrick EJ, Knudson AD, Kozubowski TJ, Nguyen T, Panorska AK, et al. Simulating High-Dimensional Multivariate Data using the bigsimr R Package. arXiv. 2021;

36. Madsen L, Birkes D. Simulating dependent discrete data. J Stat Comput Simul. 2013 Apr;83(4):677–91.

37. Data scientist's primer to analysis of mass cytometry data [Internet]. [cited 2022 May 26]. Available from: https://biosurf.org/cytof_data_scientist.html

38. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015 Jul 2;162(1):184–97.

39. Kaushik A, Dunham D, He Z, Manohar M, Desai M, Nadeau KC, et al. CyAnno: a semi-automated approach for cell type annotation of mass cytometry datasets. Bioinformatics. 2021 Nov 18;37(22):4164–71.

40. Lukas M. CS. HDCytoData. Bioconductor. 2018;

41. Goodfellow I, Bengio Y, Courville A. Deep Learning (Adaptive Computation and Machine Learning series). Illustrated. Cambridge, Massachusetts: The MIT Press; 2016.

42. Leipold MD, Obermoser G, Fenwick C, Kleinstuber K, Rashidi N, McNevin JP, et al. Comparison of CyTOF assays across sites: Results of a six-center pilot study. J Immunol Methods. 2018 Feb;453:37–43.

43. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018 Jun 19;19(1):477.

44. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014 Apr;32(4):381–6.

45. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019 May;37(5):547–54.

46. Bishop CM. Pattern Recognition and Machine Learning (Information Science and Statistics). Softcover reprint of the original 1st ed. 2006. Springer; 2016.

47. Liu X, Song W, Wong BY, Zhang T, Yu S, Lin GN, et al. A comparison framework and guideline of clustering methods for mass cytometry data. Genome Biol. 2019 Dec 23;20(1):297.

48. Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Cytometry A. 2016 Dec 19;89(12):1084–96.

49. Lee H-C, Kosoy R, Becker CE, Dudley JT, Kidd BA. Automated cell type discovery and classification through knowledge transfer. Bioinformatics. 2017 Jun 1;33(11):1689–95.

50. Zhang Z, Luo D, Zhong X, Choi JH, Ma Y, Wang S, et al. SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. Genes (Basel). 2019 Jul 12;10(7).

51. Schwarz G. Estimating the dimension of a model. Ann Statist. 1978 Mar;6(2):461–4.

52. Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype space with single-cell data. Nat Methods. 2016 Jun;13(6):493–6.

53. Hawley D, Ding J, Thotakura S, Haskett S, Aluri H, Kublin C, et al. RNA-Seq and CyTOF immuno-profiling of regenerating lacrimal glands identifies a novel subset of cells expressing muscle-related proteins. PLoS ONE. 2017 Jun 29;12(6):e0179385.

54. Perez-Cruz F. Kullback-Leibler divergence estimation of continuous distributions. 2008 IEEE International Symposium on Information Theory. IEEE; 2008. p. 1666–70.

55. Raab GM, Nowok B, Dibben C. Assessing, visualizing and improving the utility of synthetic data. arXiv. 2021;

56. Grimmett GR, Stirzaker DR. Probability and Random Processes. 3rd ed. Oxford: Oxford University Press; 2001.

57. Akima H. A new method of interpolation and smooth curve fitting based on local procedures. J ACM. 1970 Oct 1;17(4):589–602.

58. Dasgupta S. Learning Polytrees. arXiv. 2013;

59. Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms, 3rd Edition (The MIT Press). 3rd ed. Cambridge, Mass: The MIT Press; 2009.

60. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys Rev E. 2004 Dec 6;70(6).

61. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science conference. SciPy conference proceedings; 2008. p. 11–5.

62. Zhu J, Gouru A, Wu F, Berzofsky JA, Xie Y, Wang T. BepiTBR: T-B reciprocity enhances B cell epitope prediction. iScience. 2022 Feb 18;25(2):103764.

63. 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. 2021 Jan 6;18(1):92–9.

64. Lu T, Zhang Z, Zhu J, Wang Y, Jiang P, Xiao X, et al. Deep learning-based prediction of the T cell receptor-antigen binding specificity. Nat Mach Intell. 2021 Oct;3(10):864–75.

65. Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 1964 Mar;29(1):1–27.

66. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008 Mar;26(3):303–4.

67. Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR. Rapid cell population identification in flow cytometry data. Cytometry A. 2011 Jan;79(1):6–13.

68. Hubert L, Arabie P. Comparing partitions. J of Classification. 1985 Dec;2(1):193–218.

69. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020 Sep 21;2(3):lqaa078.

70. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019 Dec;16(12):1289–96.

71. Levina E, Bickel P. The Earth Mover's distance is the Mallows distance: some insights from statistics. Proceedings Eighth IEEE International Conference on Computer Vision ICCV 2001. IEEE Comput. Soc; 2001. p. 251–6.

72. Ruthotto L, Haber E. An Introduction to Deep Generative Modeling. arXiv. 2021;

73. Bond-Taylor S, Leach A, Long Y, Willcocks CG. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. arXiv. 2021;

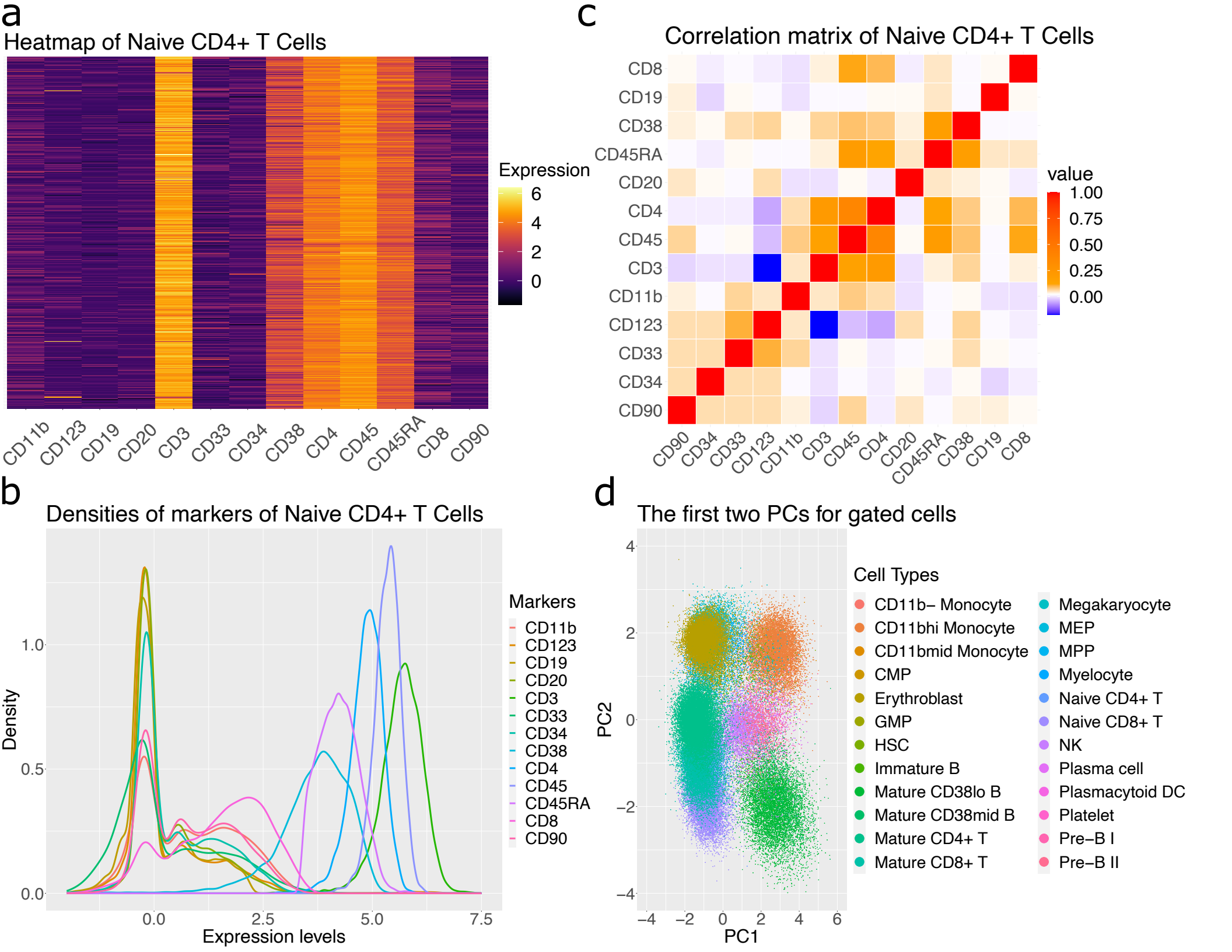
74. Rezende DJ, Mohamed S. Variational Inference with Normalizing Flows. arXiv. 2015;

75. Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv. 2013;

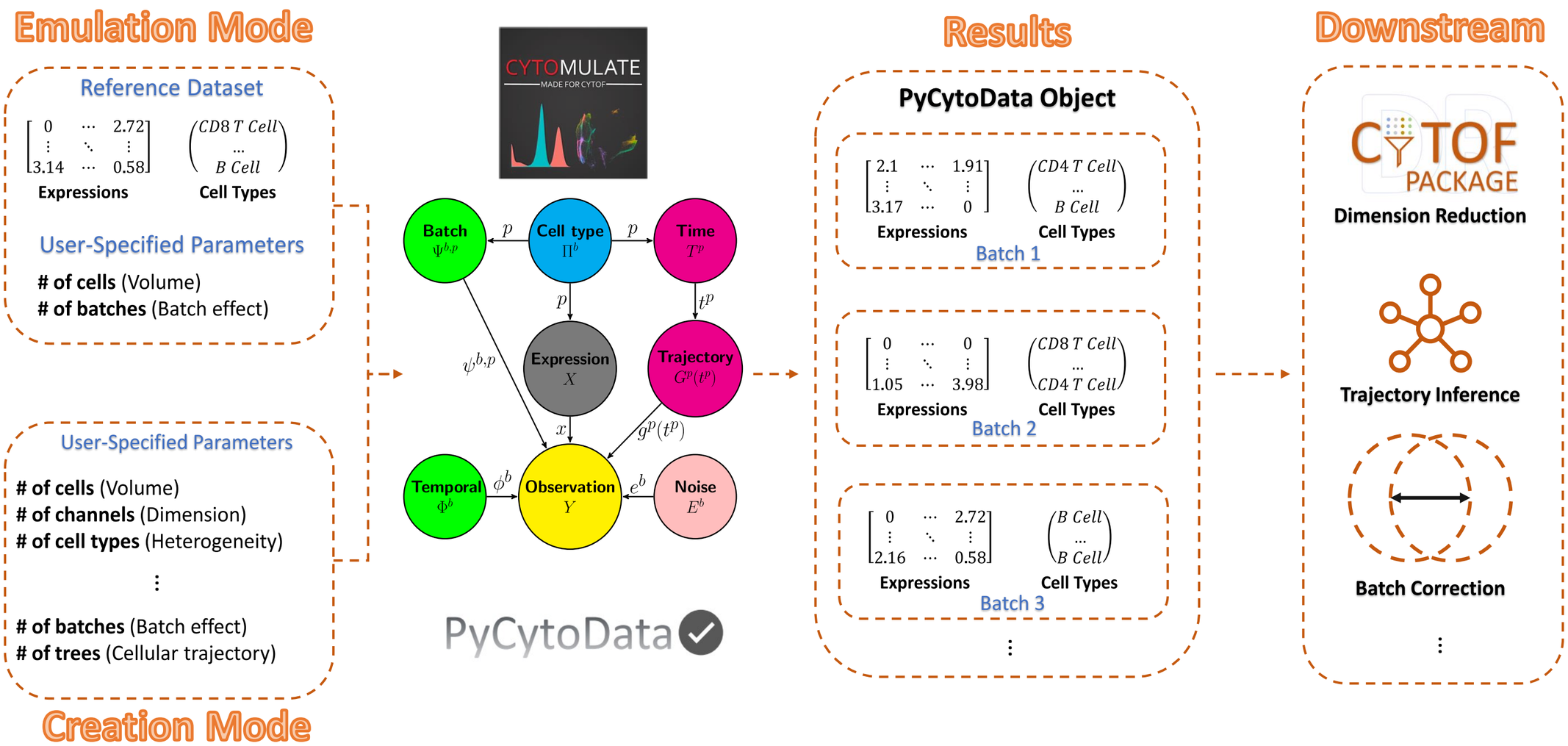
76. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. arXiv. 2014;

77. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. arXiv. 2012;

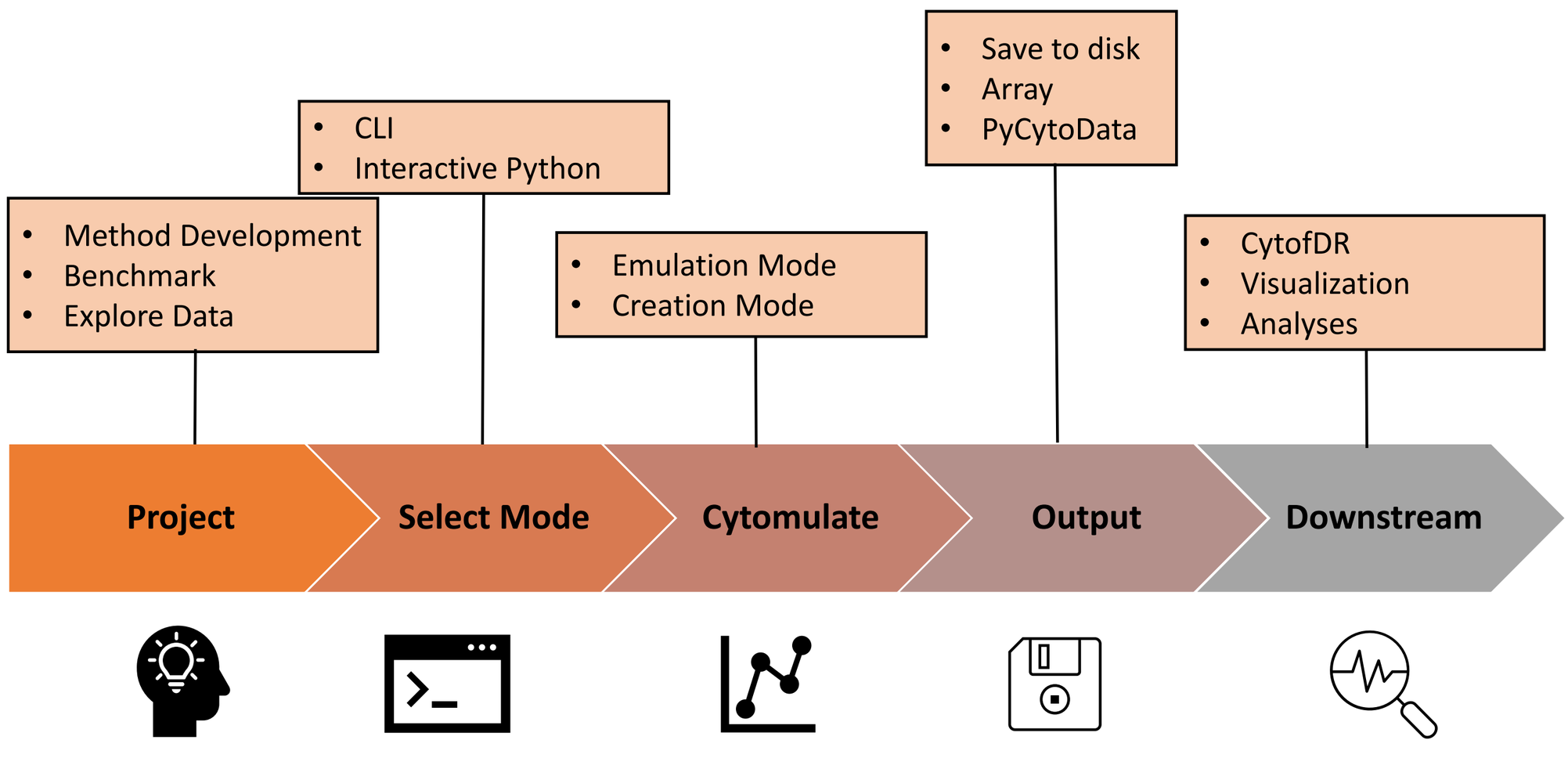
78. Trefethen LN, David Bau III. Numerical Linear Algebra. 1st ed. Philadelphia: SIAM: Society for Industrial and Applied Mathematics; 1997.

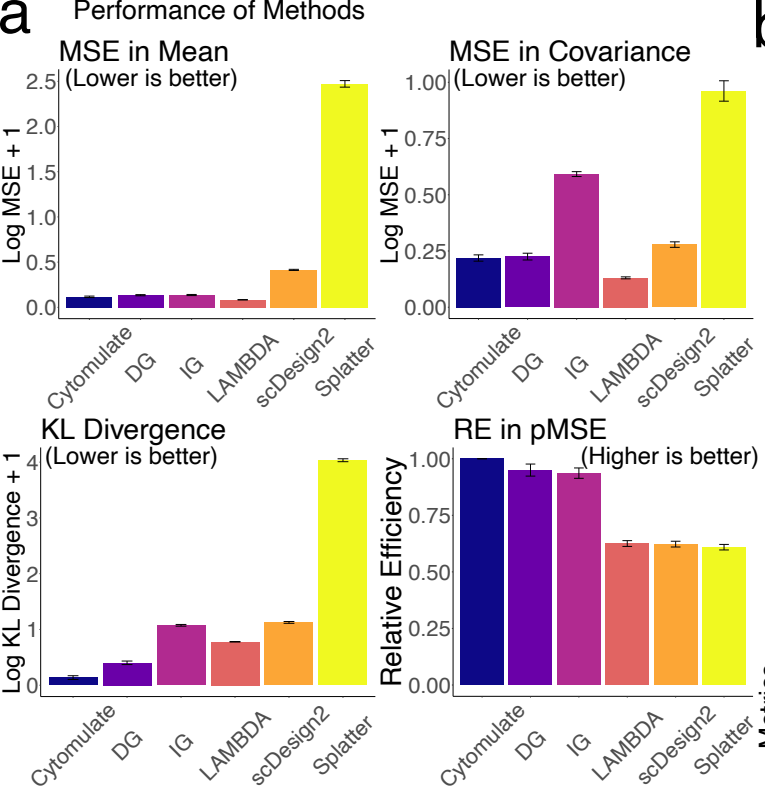

a
Heatmap of Naive CD4+ T Cells
Expression
6
4
2
0
CD11b
CD123
CD19
CD20
CD3
CD33
CD34
CD38
CD4
CD45
CD45RA
CD8
CD90
c
Correlation matrix of Naive CD4+ T Cells
value
1.00
0.75
0.50
0.25
0.00
CD8
CD19
CD38
CD45RA
CD20
CD4
CD45
CD3
CD11b
CD123
CD33
CD34
CD90
b
Densities of markers of Naive CD4+ T Cells
Density
1.0
0.5
0.0
0.0
2.5
5.0
7.5
Expression levels
Markers
CD11b
CD123
CD19
CD20
CD3
CD33
CD34
CD38
CD4
CD45
CD45RA
CD8
CD90
d
The first two PCs for gated cells
PC2
4
2
0
−2
−4
−4
−2
0
2
4
6
PC1
Cell Types
CD11b− Monocyte
CD11bhi Monocyte
CD11bmid Monocyte
CMP
Erythroblast
GMP
HSC
Immature B
Mature CD38lo B
Mature CD38mid B
Mature CD4+ T
Mature CD8+ T
Megakaryocyte
MEP
MPP
Myelocyte
Naive CD4+ T
Naive CD8+ T
NK
Plasma cell
Plasmacytoid DC
Platelet
Pre−B I
Pre−B II

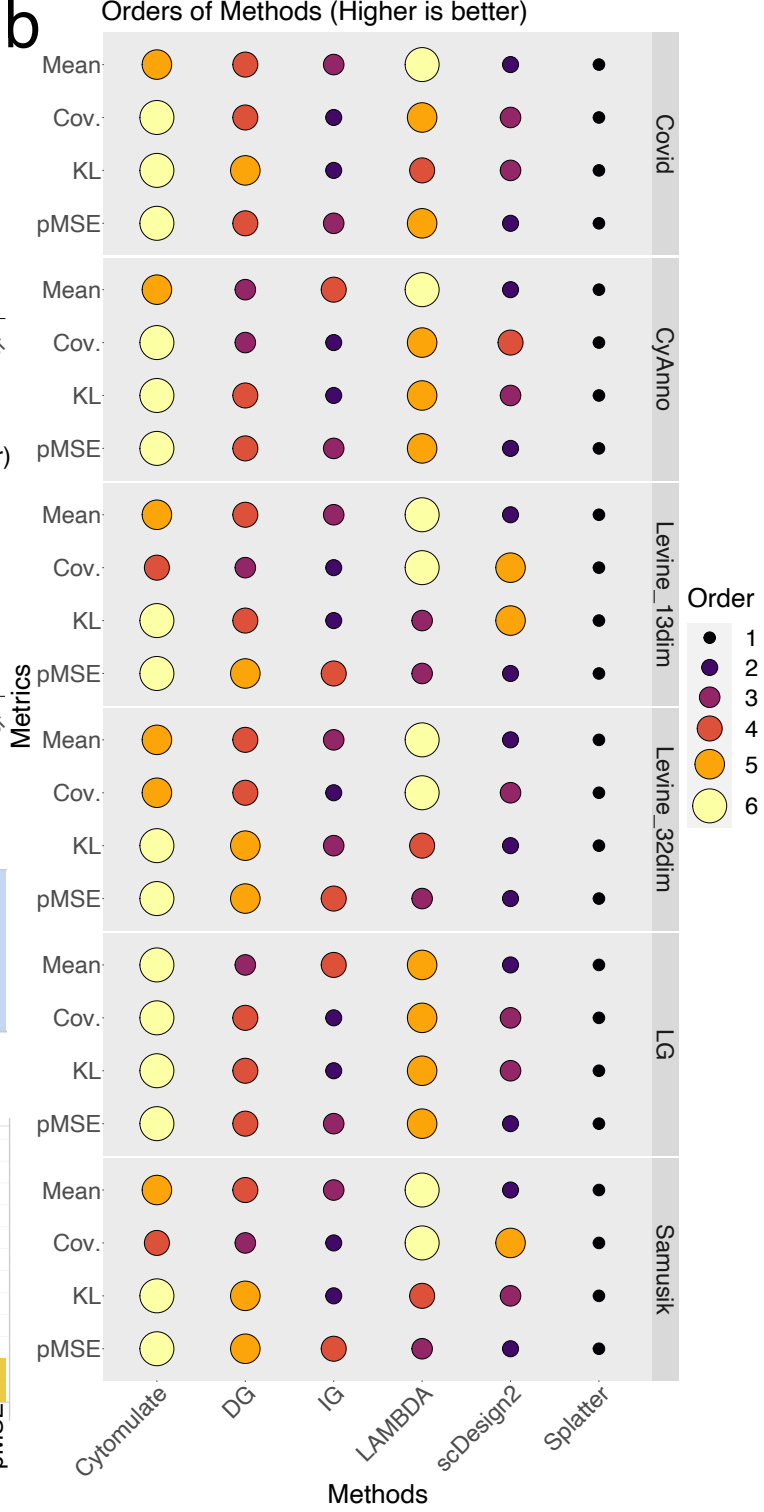
(a)


Emulation Mode
Reference Dataset
Expressions
Cell Types
User-Specified Parameters
# of cells (Volume)
# of batches (Batch effect)
User-Specified Parameters
# of cells (Volume)
# of channels (Dimension)
# of cell types (Heterogeneity)
# of batches (Batch effect)
# of trees (Cellular trajectory)
Creation Mode
CYTOMULATE
MADE FOR CYTOF
Batch
Cell type
Time
Expression
Trajectory
Temporal
Observation
Noise
PyCytoData
Results
PyCytoData Object
Expressions
Cell Types
Batch 1
Batch 2
Batch 3
Downstream
CYTOF PACKAGE
Dimension Reduction
Trajectory Inference
Batch Correction

(b)

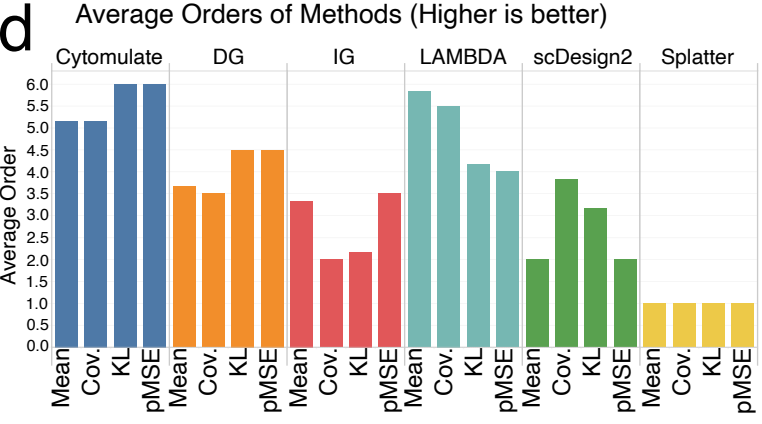

Method Development
Benchmark
Explore Data
CLI
Interactive Python
Emulation Mode
Creation Mode
Save to disk
Array
PyCytoData
CytofDR
Visualization
Analyses
Project
Select Mode
Cytomulate
Output
Downstream

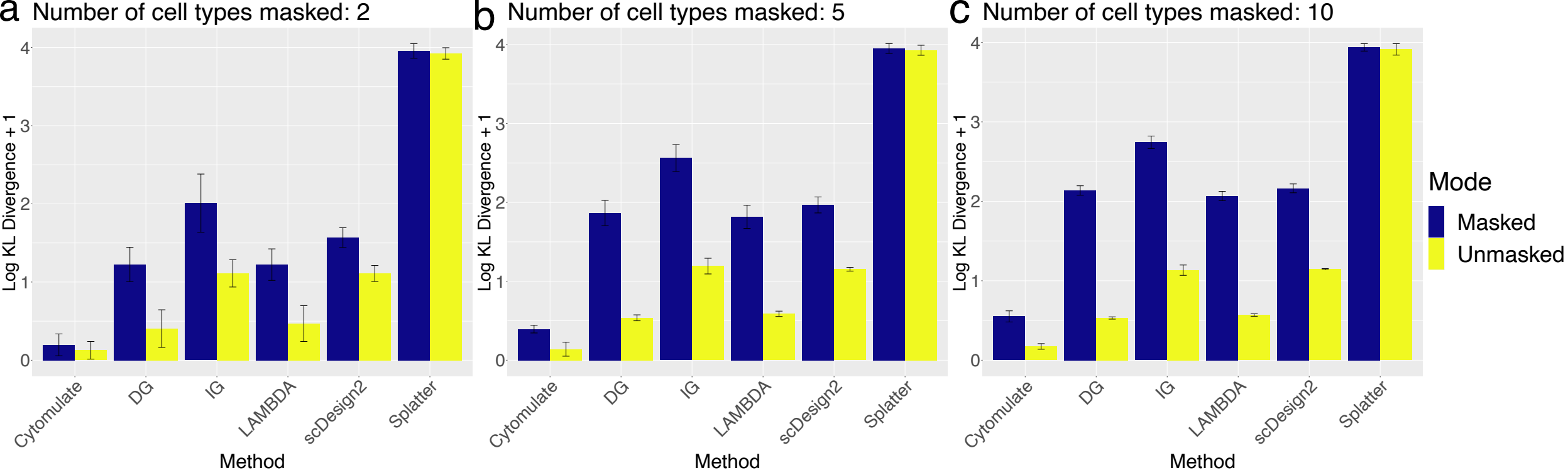


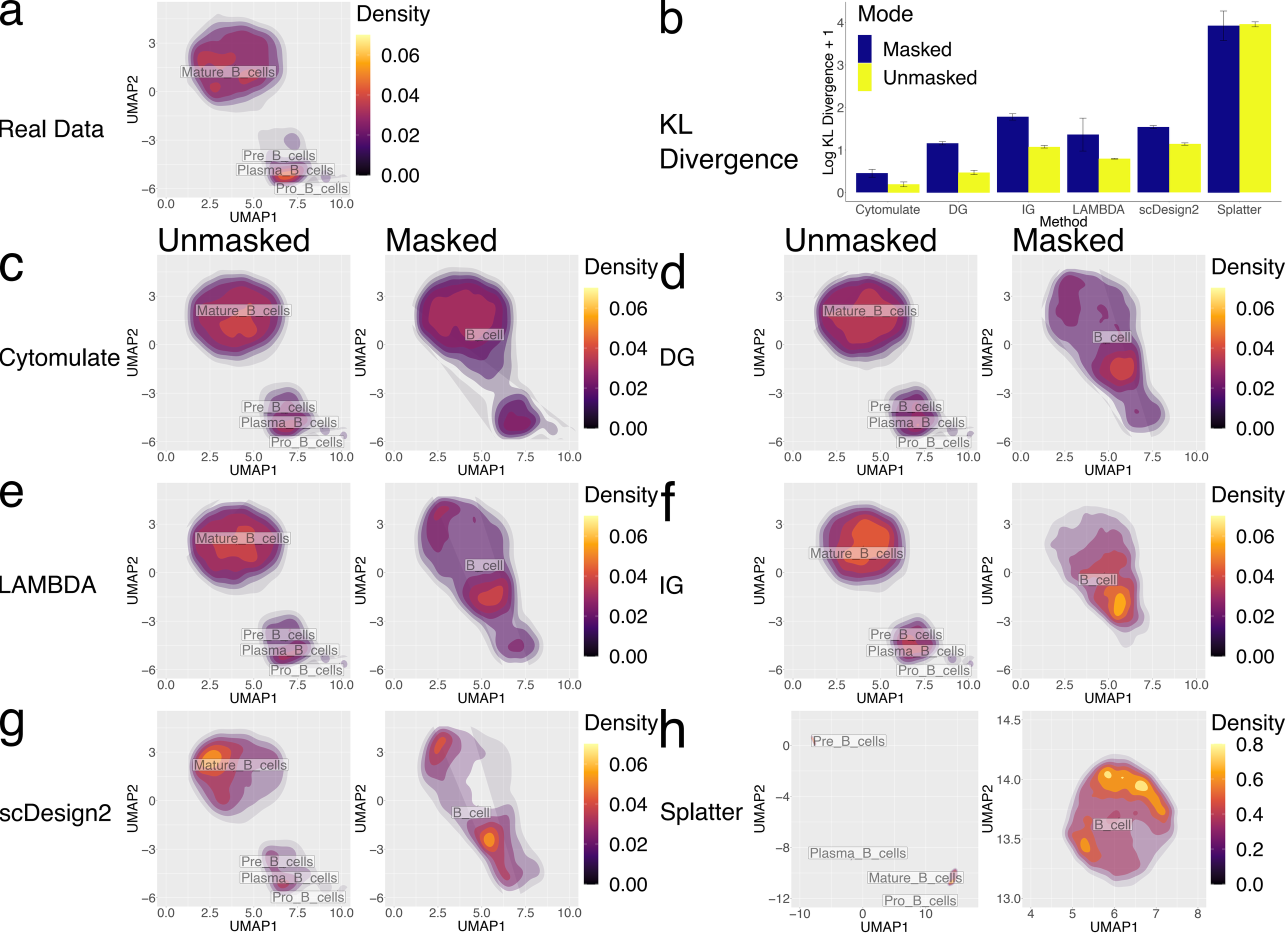


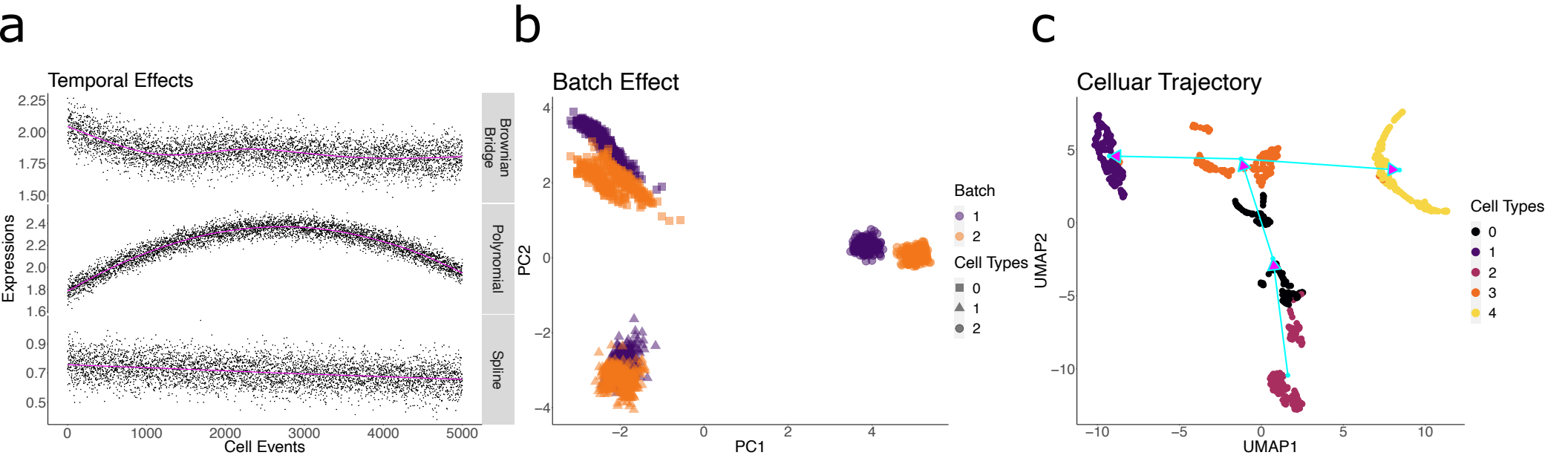
c Average Performance of Methods

	Cytomulate	DG	IG	LAMBDA	scDesign2	Splatter
MSE (Mean)	0.20	0.27	0.27	0.14	0.62	11.25
MSE (Cov.)	0.69	1.14	2.74	0.72	1.12	4.99
KL	1.06	2.47	6.16	2.66	3.77	53.30
RE (pMSE)	1.00	0.93	0.88	0.80	0.70	0.61




a
Number of cell types masked: 2
b
Number of cell types masked: 5
c
Number of cell types masked: 10
Log KL Divergence + 1
0
1
2
3
4
Cytomulate
DG
IG
LAMBDA
scDesign2
Splatter
Method
Mode
Masked
Unmasked


a
Real Data
Density
0.06
0.04
0.02
0.00
Mature_B_cells
Pre_B_cells
Plasma_B_cells
Pro_B_cells
UMAP1
UMAP2
b
KL Divergence
Mode
Masked
Unmasked
Log KL Divergence + 1
Cytomulate
DG
IG
LAMBDA
scDesign2
Splatter
Method
c
Cytomulate
Unmasked
Masked
d
DG
e
LAMBDA
f
IG
g
scDesign2
h
Splatter
B_cell


a
Temporal Effects
Expressions
Cell Events
Brownian Bridge
Polynomial
Spline
b
Batch Effect
PC1
PC2
Batch
1
2
Cell Types
0
1
2
c
Celluar Trajectory
UMAP1
UMAP2
Cell Types
0
1
2
3
4