Distribution-based Sketching of Single-Cell Samples

Vishal Athreya Baskaran
athreya@cs.unc.edu
Department of Computer Science, UNC-Chapel Hill

Jolene Ranek
ranekj@live.unc.edu
Computational Medicine Program, Curriculum in Bioinformatics and Computational Biology, UNC-Chapel Hill

Siyuan Shan
siyuanshan@cs.unc.edu
Department of Computer Science, UNC-Chapel Hill

Natalie Stanley
natalies@cs.unc.edu
Computational Medicine Program, Department of Computer Science, UNC-Chapel Hill

Junier B. Oliva
joliva@cs.unc.edu
Department of Computer Science, UNC-Chapel Hill

ABSTRACT

Modern high-throughput single-cell immune profiling technologies, such as flow and mass cytometry and single-cell RNA sequencing can readily measure the expression of a large number of protein or gene features across the millions of cells in a multi-patient cohort. While bioinformatics approaches can be used to link immune cell heterogeneity to external variables of interest, such as, clinical outcome or experimental label, they often struggle to accommodate such a large number of profiled cells. To ease this computational burden, a limited number of cells are typically *sketched* or subsampled from each patient. However, existing sketching approaches fail to adequately subsample rare cells from rare cell-populations, or fail to preserve the true frequencies of particular immune cell-types. Here, we propose a novel sketching approach based on Kernel Herding that selects a limited subsample of all cells while preserving the underlying frequencies of immune cell-types. We tested our approach on three flow and mass cytometry datasets and on one single-cell RNA sequencing dataset and demonstrate that the sketched cells (1) more accurately represent the overall cellular landscape and (2) facilitate increased performance in downstream analysis tasks, such as classifying patients according to their clinical outcome. An implementation of sketching with Kernel Herding is publicly available at https://github.com/vishalathreya/Set-Summarization.

KEYWORDS

Cytometry, Single-Cell Bioinformatics, Clinical Prediction, Compression

ACM Reference Format:
Vishal Athreya Baskaran, Jolene Ranek, Siyuan Shan, Natalie Stanley, and Junier B. Oliva. 2022. Distribution-based Sketching of Single-Cell Samples. In *13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22), August 7–10, 2022, Northbrook, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3535508.3545539

BCB '22, August 7–10, 2022, Northbrook, IL, USA

ACM ISBN 978-1-4503-9386-7/22/08…$15.00
https://doi.org/10.1145/3535508.3545539

1 INTRODUCTION

Advances in high throughput single-cell technologies have transformed our understanding of cellular heterogeneity across a range of biological and clinical applications. For example, single-cell measurements have been used to construct whole organism cell atlases [9, 34] that can serve as prototypical healthy references that assist in the discovery of novel disease cell states [24]. Translationally, single-cell modalities have provided insight into complex immune responses linked to clinical outcomes, such as surgical recovery [15], pregnancy [2], tumor heterogeneity [22], and infectious disease [3].

Single-cell flow and mass cytometry (e.g. CyTOF) technologies have been particularly useful for gaining an in-depth understanding of the diversity of immune cell-types and their relation to disease, as they simultaneously profile 20-45 proteins collectively across all of the cells profiled in a multi-patient cohort [5, 23, 36]. This data can be leveraged in statistical models for identifying differentially abundant cell-populations [12, 26, 40], kernel density estimation methods to characterize treatment resistant cells [8], and classification algorithms to predict a patient's clinical outcome [41]. Similarly, genomic technologies such as single-cell RNA sequencing (scRNA-seq) technologies can profile 20, 000 or more gene expression measurements in thousands of cells [25]. Given the large number of cells measured in such datasets, downstream analysis tasks present a computational challenge and quickly become intractable. Traditional approaches for reducing the size of input data, such as uniform downsampling or clustering, may fail to identify rare but clinically meaningful cell-populations. This limits both the utility of these algorithms and the ability to obtain meaningful results.

Two alternative sketching-based approaches have been developed to intelligently select a subset of cells such that the overall cellular landscape of the full data is preserved. Geometric Sketching first approximates the underlying geometry of the data through a covering of equal volume hypercubes, then evenly samples points from each at random [19]. Alternatively, Hopper constructs a sketch by iteratively minimizing the Hausdorff distance to ensure points in the full data are well-represented in the sketch [13]. Both methods provide an even sampling of cells that focuses on capturing rare

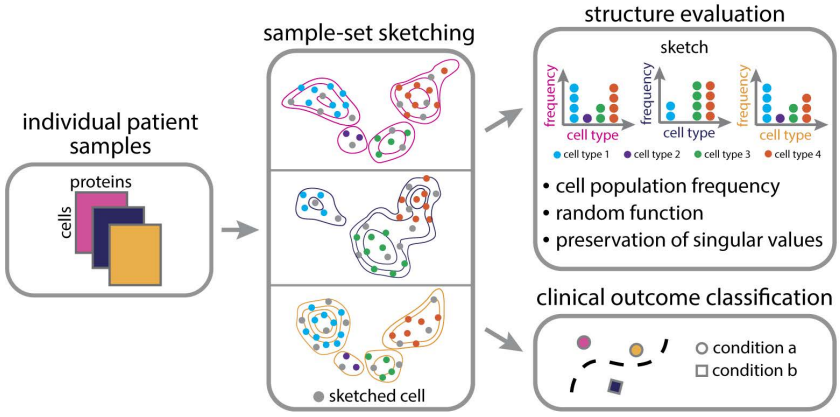

Figure 1: Overview of Kernel Herding for Sketching Single-Cell Data. Multiple *sample-sets* are created after profiling the expression of several genomic or protein features in single cells across multiple individuals. We used Kernel Herding to select a limited representative subset of cells from sample-sets for each individual. Our approach was then evaluated in terms of the quality of cellular landscape preservation and classification performance of sample-sets according to the clinical outcomes of the associated patients.

cell-types. However, as a result, they fail to preserve the overall distributions of cell-type frequencies which can be crucial for clinical outcome prediction and biological interpretation [15, 28]. In a philosophical divergence with this previous sketching work, here we prioritize the sketch's ability to act as a *stand-in* for the original sample-set. That is, here we consider sketches that obtain *similar downstream outcomes* to the original sample-sets when analyzed. In some respects, this goal is a generalization of prior formulations, since rare-cell types must be preserved if the sketch can act as a stand-in. However, our formulation also allows for the preservation of general downstream analysis, which may not be entirely dependent on rare-cell types. Thus, this work presents a novel, more direct approach to sketching single-cell samples.

We propose to preserve downstream analysis of a single-cell sample by explicitly preserving the distribution in the sketched sample. In this work, we show that this can be achieved via Kernel Herding [10], a kernel-based approach to sub-sampling. Furthermore, we show how to efficiently compute Kernel Herding sketches through random Fourier features [33], which allows our sketch approach to scale to large sample-sets. We first evaluate our approach on three diverse flow and mass cytometry datasets by computing several statistics to quantify cellular landscape preservation and performance in downstream tasks, such as patient classification according to clinical outcome. Furthermore, we demonstrate how this approach can be used across single-cell data modalities with a single-cell RNA sequencing application. Our Kernel Herding sketching approach (outlined in Fig. 1) now makes it possible to select a limited subset of cells from each profiled sample-set that adequately represents immune cell-types and their relative frequencies.

2 METHODS

Here we are considering the problem of finding a representative subset to an original sample-set of cells $\mathcal{X} = \{x_i \in \mathbb{R}^d\}_{i=1}^n$. If subsets are to act as a stand-in for original sample-sets, then they should provide similar outcomes when processed as the original sets. Thus, we consider a subset, $\hat{\mathcal{X}} = \{\hat{x}_i\}_{i=1}^m \subset \mathcal{X}$, to be representative if it maintains the distribution of the original set $\mathcal{X}$ and yields similar results to $\mathcal{X}$ when processed.

Problem Formulation At an abstract level, we shall construct a subset $\hat{\mathcal{X}}$ whose statistics match the original sample-set $\mathcal{X}$. Of course, for any finite number of statistics, there may be an ambiguity over the original distribution. For instance, a large variety of samples may share the same mean, thus simply finding a subset of a similar mean as the original set will not suffice to ensure a representative subsample. Instead, we shall use the kernel mean embedding [31] as the target 'statistic' to match in the subset. We briefly explain kernels and kernel mean embedding below.

Kernel Mean Embedding Kernel methods have been successfully applied in a myriad of machine learning problems such as regression [39], classification [11], and dimensionality reduction [29]. Underlying kernel methods is a positive definitive kernel function $K : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ [1], which induces a reproducing kernel Hilbert space (RKHS) (e.g. see [6] for further details). In addition to working over typical vector-values data, recently kernels have also been deployed to operate over distributions (e.g. [31]). The *kernel mean embedding* $\mu_p : \mathbb{R} \mapsto \mathbb{R}$ represents a distribution, p, as:

$$\mu_p(\cdot) = \mathbb{E}_{x \sim p}[k(x, \cdot)] \tag{1}$$

Note that μ_p is itself a function, defined by the expected kernel evaulation to points drawn from a distribution p. For a special class of '*characteristic*' kernels, k, such as the common radial-basis function (RBF) kernel $k(x, x') = \exp(-\frac{1}{2\gamma}||x - x'||^2)$, $||\mu_p - \mu_q|| = 0$ if and only if $p = q$. Thus, for characteristic kernels the kernel mean embedding will be unique to its distribution and matching the kernel embedding exactly guarantees that one matches a distribution.

[1] Kernels may also generalize over non-real domains

In general, the distance[2] $||\mu_p - \mu_q||$ induces a divergence, the maximum mean discrepancy (MMD) [16], between distributions, which has been useful in comparing distributions (e.g. [17]).

In this work we sketch a sample-set with a subset $\hat{\mathcal{X}} \subset \mathcal{X}$ that preserves the empirical distribution found in the original sample-set $\mathcal{X}$ with the mean embedding $\mu_{\mathcal{X}}$ where:

$$\mu_{\mathcal{X}}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot). \tag{2}$$

That is, we look for a subset $\hat{\mathcal{X}}$ whose empirical distribution has a small MMD from that of the original set $\mathcal{X}$, $||\mu_{\hat{\mathcal{X}}} - \mu_{\mathcal{X}}||$. Note that a uniform subsample $\mathcal{U} \subset \mathcal{X}$ provides a reasonable approximation when the cardinality $|\mathcal{U}|$ is large enough. That is $\mu_{\mathcal{U}}(\cdot) = \frac{1}{|\mathcal{U}|}\sum_{x\in\mathcal{U}} k(x, \cdot)$, shall approximate $\mu_{\mathcal{X}}$ when using enough uniformly subsampled points. However, as shown by [10], one may get a more representative subsample through *Kernel Herding* as described below.

Kernel Herding Kernel herding is a greedy algorithm to approximate a mean embedding μ_p with a subset of a set of samples $\mathcal{X} \sim p$, using a subset of points $\hat{\mathcal{X}} = \{\hat{x}_i\}_{i=1}^{m} \subset \mathcal{X}$. Here, we consider the empirical distribution $p = \frac{1}{n}\sum_{i=1}^{n}\delta(x_i)$, where δ is the Dirac-delta function. Hence, we look to approximate $\mu_{\mathcal{X}}$ with $\mathcal{X}$ being our original sample. Kernel herding is especially applicable for synthesizing a large number of cells, due to the following properties. First, the rate of convergence for the error $\|\mu_{\mathcal{X}} - \frac{1}{m}\sum_{j=1}^{m} k(\hat{x}_i, \cdot)\|^2$ is $O(\frac{1}{m})$, which is much faster than the $O(\frac{1}{\sqrt{m}})$ rate for a uniformly random subsample. That is, Kernel Herding can provide a synthesized subset of $\sqrt{m}$ points that approximates the target distribution as well as m uniformly subsampled points. This property implies that the Kernel Herding subset shall maintain the distributional properties found in the original sample-set. Second, it can be shown that the Kernel Herding subset $\hat{\mathcal{X}}$ also approximates the expectation of any function h[3], $\mathbb{E}_{x\sim p}[h(x)]$, with a mean over $\hat{\mathcal{X}}$, $\frac{1}{m}\sum_{j=1}^{m} h(\hat{x}_j)$. This property enables one to obtain similar downstream models when using the summarized subset as the expectation of functions evaluated.

The Kernel Herding subset $\hat{\mathcal{X}}$ is computed as follows:

Algorithm 1 Compute Subsampled Set using Kernel Herding

Require: A set of cells $\mathcal{X}$, from which the number of cells subsampled is m, Radial-Basis Function kernel k.

Initialize $j \leftarrow 1, \hat{\mathcal{X}} \leftarrow \emptyset$
while $j \leq m$ **do**
  $\hat{\mathbf{x}} \leftarrow \underset{x\in\mathcal{X}}{\text{argmax}} \frac{1}{N}\sum_{x'\in\mathcal{X}} k(x, x') - \frac{1}{j}\sum_{x'\in\hat{\mathcal{X}}} k(x, x')$
  $\hat{\mathcal{X}} \leftarrow \hat{\mathcal{X}} \cup \{\hat{\mathbf{x}}\}$
  $j \leftarrow j + 1$
end while
return $\hat{\mathcal{X}}$

Random Fourier Features Performing Kernel Herding with a kernel function k, however, may not scale since an iteration requires an $O(n^2)$ computation (equivalent to computing the Gram matrix of pairwise kernel evaluations on $\mathcal{X}$). To rectify this, we propose to use random Fourier frequency features [33]. For a shift-invariant kernel (such as the RBF kernel), random Fourier features provide a feature map $\varphi(x) \in \mathbb{R}^D$ such that the dot product in feature space approximates the kernel evaluation, $\varphi(x)^T\varphi(x') \approx k(x, x')$ (e.g. see [32, 33, 33] for further details). Using the dot product of $\varphi(x)$, our mean embedding becomes $\mu_{\mathcal{X}} = \frac{1}{n}\sum_{i=1}\varphi(x_i)$, where $\mu_{\mathcal{X}}(x') = \frac{1}{n}\sum_{i=1}\varphi(x_i)^T\varphi(x')$. In practice, the random features are constructed by drawing frequencies, $W \in \mathbb{R}^{d\times\frac{D}{2}}$, at random (*iid* column-wise) once, and holding them fixed to construct the features

$$\varphi_W(x) = [\sin(\mathbf{W}^T x), \cos(\mathbf{W}^T x)] \in \mathbb{R}^D, \tag{3}$$

where $[\cdot,\cdot]$ denotes concatenation. For instance, to approximate the RBF kernel, $k(x, x') = \exp(-\frac{1}{2\gamma}||x - x'||^2)$, we would draw the d-dimensional frequencies $W_i \overset{iid}{\sim} \mathcal{N}(0, \gamma^{-1}I)$ [33]. It is simple to show that when using the approximation $\varphi_W(x)^T\varphi_W(x') \approx k(x, x')$, one may compute Algorithm 1 with $O(n)$ iterates by avoiding pairwise kernel computations [10]; we detail this below.

Algorithm 2 Compute Subsampled Set using RBF Kernel Herding with Random Features

Require: A set of cells $\mathcal{X}$ with dimensionality d, from which the number of cells subsampled is m, dimensionality of the random feature space D and kernel hyperparameter γ.

`# Draw random Fourier frequencies`
Compute $\mathbf{W} \in \mathbb{R}^{d\times\frac{D}{2}}$ by sampling its elements independently $\mathbf{W}_{i,j} \sim \mathcal{N}(0, \frac{1}{\gamma})$
`# Subsampling using Kernel Herding`
Initialize $j \leftarrow 1, \hat{\mathcal{X}} \leftarrow \emptyset, \theta_0 \leftarrow \frac{1}{n}\sum_{i=1}^{n}\varphi_{\mathbf{W}}(x_i)$
while $j \leq m$ **do**
  $\hat{\mathbf{x}} \leftarrow \underset{x\in\mathcal{X}}{\text{argmax}}\ \theta_{j-1}^T\varphi_{\mathbf{W}}(x_i)$
  $\hat{\mathcal{X}} \leftarrow \hat{\mathcal{X}} \cup \{\hat{\mathbf{x}}\}$
  $\theta_j \leftarrow \theta_{j-1} + \theta_0 - \varphi_{\mathbf{W}}(\hat{\mathbf{x}})$
  $j \leftarrow j + 1$
end while
return $\hat{\mathcal{X}}$

2.1 Single-Cell Datasets Used in Experiments

In all experiments, we used publicly available, multi-sample flow, mass cytometry (CyTOF), and single-cell RNA sequencing datasets. For the flow and mass cytometry datasets, each sample-set consists of a collection of protein markers (up to ~45) measured across individual cells. For the single-cell RNA sequencing dataset, each sample-set consists of a collection of gene expression measurements (~20k) measured across individual cells. Here, we briefly introduce the multi-sample publicly available datasets used in our experiments. All preprocessed data are available in the Zenodo repository: https://zenodo.org/record/6546964.

- **Preeclampsia** The preeclampsia CyTOF dataset [18] includes samples collected longitudinally from 12 healthy women and 11 women with preeclampsia throughout their pregnancies. All patient samples were downloaded from Flow Repository under Repository ID FR-FCM-ZYRQ (http://flowrepository.org/id/FR-

[2] Corresponding to the RKHS norm
[3] In the corresponding reproducing kernel Hilbert space (RKHS).

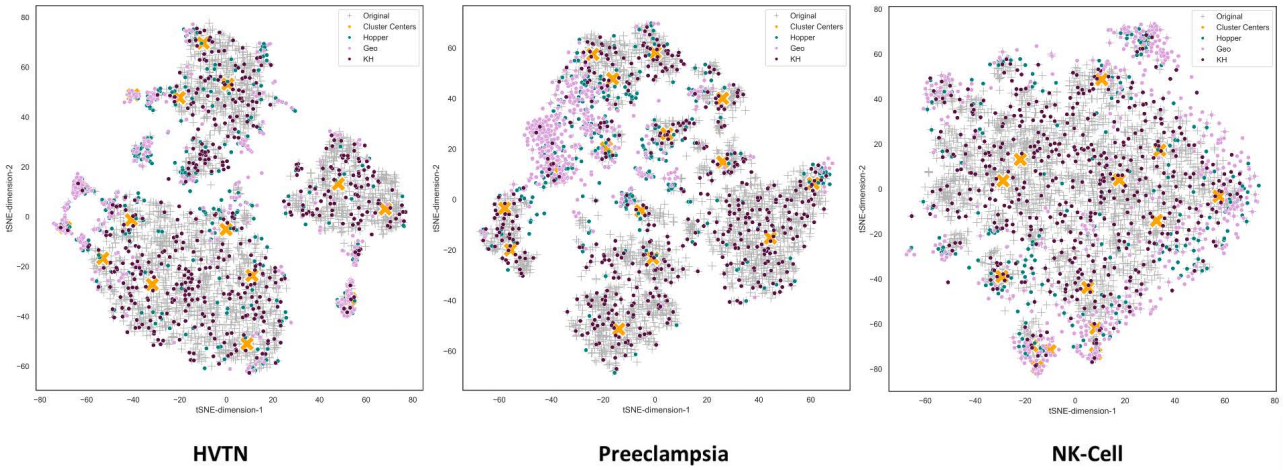

Figure 2: t-SNE visualizations of the sketches produced under each method. The distribution of points sketched with Kernel Herding (KH) (dark purple points) follows that of the original data (gray, background points representing a large number of randomly sampled cells). Sketches with Kernel Herding adequately represent rare populations and retain their relative frequencies, while Geometric Sketching and Hopper struggle with these aspects.

FCM-ZYRQ). Classification tasks distinguished between healthy from preeclamptic women.

- **HVTN** The HIV Vaccine Trials Network (HVTN) is a Flow Cytometry dataset and consists of 96 total sample-sets of T-cells that were each subjected to stimulation with either Gag or Env [1] (downloaded from Flow Repository, ID FR-FCM-ZZZV http://flowrepository.org/id/FR-FCM-ZZZV). Classification tasks distinguished Gag from Env stimulated samples
- **NK-Cell** The NK-Cell CyTOF dataset profiled NK-Cells across 21 individuals who were either positive or negative for Cytomegalovirus (CMV) [4] (downloaded from https://github.com/eiriniar/CellCnn). Classification tasks distinguished CMV positive (CMV+) from CMV negative (CMV-) samples.
- **MS** The multiple sclerosis (MS) single-cell RNA sequencing dataset [35] consists of peripheral blood samples collected from 4 MS patients and 4 healthy controls. Patient samples were accessed from the Gene Expression Omnibus using the accession code GSE138266. We performed standard single-cell RNA sequencing data preprocessing, including filtering cells according to read depth and distribution of molecular counts, removing cells with greater than 20 percent mitochondrial transcripts, and retaining genes that were expressed amongst a minimum of 5 cells. Following quality control filtering, we normalized the data to account for differences in sequencing depth by estimating size factors using Scran pooling normalization v1.20.1 [27] and scaling them across batches using Batchelor v1.8.0. We then performed batch effect correction using ComBat [20]. Lastly, we restricted the feature space by selecting for highly variable genes on `log+1` transformed data using a normalized dispersion measure in Scanpy v1.8.1 (flavor = Seurat, minimum mean = 0.012, minimum dispersion = 0.25, maximum mean = 5). In our subsequent experiments, we performed principal component analysis on the preprocessed dataset and used the top 50 components for downstream tasks.

3 RESULTS

We compared the performance of Kernel Herding to sketches obtained using Geometric Sketching, Hopper, and IID subampling for tasks related to the overall preservation and usefulness of the resulting immunological landscape. That is, we sought to evaluate whether or not the sketches adequately represented all major immune cell-types and their relative frequencies and could be used to produce meaningful immunological features for downstream tasks. Code for reproducing the results of all subsequent experiments is publicly available at https://github.com/vishalathreya/Set-Summarization.

3.1 Description of Related Algorithms

Here, we briefly define the sketching approaches that were compared to in our experiments.

- **Geometric Sketching** Geometric Sketching introduced in Ref. [19] infers a *plaid covering* of cells in the high-dimensional space. Cells are sketched through volume-dependent sampling by selecting the same number of cells from sections of the high-dimensional plaid covering.
- **Hopper** Hopper introduced in Ref. [13] forms a sketch by using fastest first traversal, which is a greedy approximation to the k-center problem. Intuitively, this sketching approach sequentially adds cells to the sketch that are sufficiently different from those that were already included in the sketch.
- **Independent and Identically Distributed Subsampling (IID)** IID sketches were generated by simply selecting a random subsample of cells from each sample-set. Here, each cell had the same probability of being selected in the sketch.

3.2 t-SNE Visualizations of Sketched Regions of the Cellular Landscape

We begin with a qualitative assessment of sketching approaches on cytometry data (Fig. 2). In order to get a visual understanding

of the regions of the cellular landscape included in the sketches produced by each method, we plotted a $2d$ t-SNE projection of respective sketches (colored points) along with a set of overall cells from a large sub-sample of the original set (gray-colored points). I.e. Fig. 2, plots cells from three samples, one for each respective dataset (similar results may be obtained from other samples). Furthermore, we projected cluster centroids (15, from k-means) as large yellow crosses. Taken together, the respective plots give an overview of the cells found in each original sample-set (gray), and which cells were then included in sketches (colored). Note that an IID sketch would stem from a uniform sub-sample of the gray points, which was omitted for visual clarity. One can observe that the Geometric Sketching (*Geo*, pink, Fig. 2) sketches concentrate over a few sub-regions of the cellular-space, leaving large regions of cells underrepresented in the sketch. This is also true of Hopper (*Hopper*, green, Fig. 2), though to slightly lesser extent. Finally, we can observe that our proposed Kernel Herding (*KH*, purple, Fig. 2) sketches yield a more representative coverage of the cell-space. Following our philosophical goal of obtaining sketches that may act as general *stand-ins* for original sample-sets, it is intuitive that the discrepancy in representation of sketches to the original shall result in discrepancies of outputs of downstream analysis. Moreover, following the insights gained from Kernel Herding (Sec. 2), it is intuitive that IID sketches shall be less representative than those from Kernel herding. In subsequent experiments below, we show that these intuitions hold empirically.

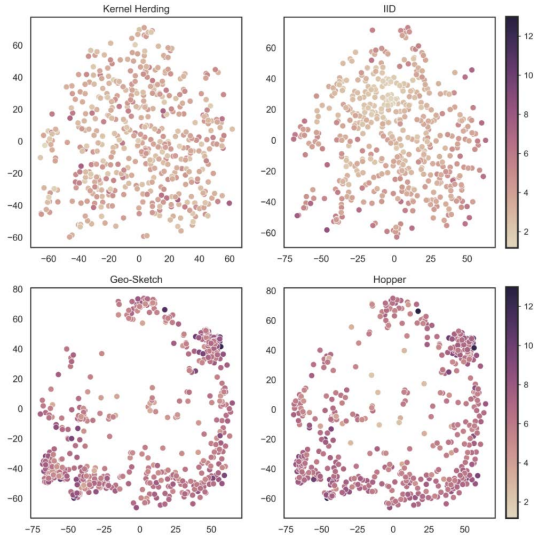

Figure 3: t-SNE visualizations of cells sketched from the NK-Cell dataset colored by their the third nearest neighbor distances in the original protein-marker feature space.

We provide additional context to the aforementioned t-SNE projections (Fig. 2) with a volumetric analysis. In particular, as t-SNE projections may not be volumetrically preserving (areas in the $2d$ space need not scale to areas in the original space), it may be the case that sketched points seem overly warped and are not representative of their coverage in the original high-dimensional protein marker feature space. To obtain a clearer view of the representation of sketched cells in the original space, we visualize the third nearest neighbor distance in the original space when scattering the cells in the $2d$ t-SNE space (Fig. 3 plots this for a sample-set in the NK-Cell dataset). That is, dark-colored points stem from *sparsely* populated regions in the original space, since the nearest neighbor of such points in the original space was *far*. Similarly, light-colored points stem from *densely* populated regions in the original space. As suspected above, Geo-Sketch and Hopper sketches are concentrated in *sparse* regions of the space. While this acceptable for certain applications, the resulting sketches largely ignore denser regions of the cell-space, which prevents sketches from acting as a *stand-in* for downstream analysis. In contrast, we observe that Kernel Herding obtains good coverage of the original space, while still including cells from sparser regions.

3.3 Random Function Fidelity

As previously discussed, our aim is to produce sketches that may act as a stand-in for general analysis of samples. To test how well sketches estimate the output of a *wide-range* of analyses, we begin by evaluating randomly generated functions on sketches and comparing their outputs to that of the original sample (Fig. 4).

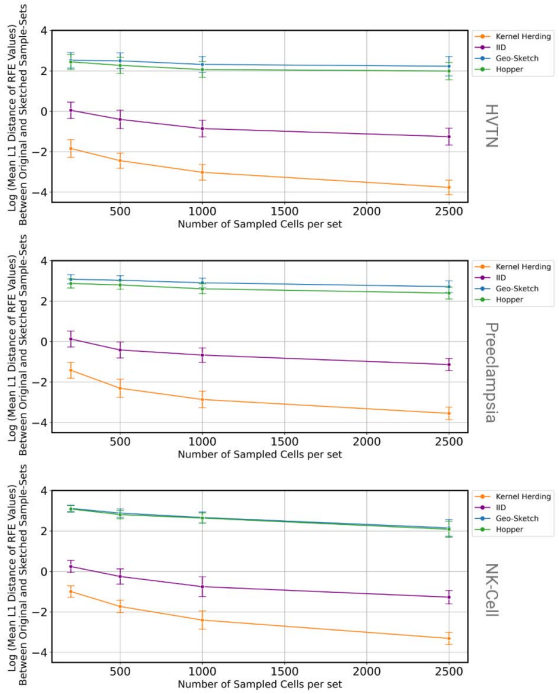

Figure 4: Random Function Evaluation We evaluated a random function using the full and sketched versions for the HVTN, Preeclampsia and NK-Cell datasets, for sketch sizes between 200 and 2500 cells. Sketches with Kernel Herding generally produce random function evaluations (RFEs) that are the most similar to that obtained when using all cells.

We generate random cell-wise functions, $f : \mathbb{R}^d \mapsto \mathbb{R}$, that are evaluated on sets, $\mathcal{X}$, as : $\mathbf{f}(\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} f(x)$. We compare the function evaluation on the original sample, $\mathcal{X}$, to a sketched sub-sample, $\hat{\mathcal{X}}$, using the ℓ_1 distance: $|\mathbf{f}(\mathcal{X}) - \mathbf{f}(\hat{\mathcal{X}})|$. We parameterize functions through random features (3): $f_{W,\beta}(x) = \varphi_w(x)^T \beta$, where

$\varphi_w(x)$ are the random features w.r.t. a drawn set of frequencies, and β are coefficients. Note that $f_{W,\beta}(x)$ is a *highly non-linear* function in x. Thus, testing the discrepancies $|\mathbf{f}_{W,\beta}(X) - \mathbf{f}_{W,\beta}(\hat{X})|$ between multiple W, β shall give a robust measure of the representative power of sketches. For each dataset, we draw 5 random functions (by drawing β, W at random) and report the average ℓ_1 discrepancy for sketches produced with each respective method for various sketch cardinalities (see Fig. 4, note the log-scale). Perhaps not surprisingly in light of our qualitative analysis, we see that Geo-Sketch and Hopper yield sketches that are poor stand-ins for this task. Interestingly, we also see a large advantage in Kernel Herding sketches to IID sketches. This highlights that although IID sketches retain distributional properties, Kernel Herding sketches provide a more efficient and accurate synthesis.

3.4 Singular Value Fidelity

Above we studied sketches' ability to act a stand-in for general non-linear evaluations on original sample-sets. Here, we now consider a specific analysis based on singular values. In particular, we study how well the singular values of the original $n \times d$ (cell $\times$ protein marker matrices) sample compares to the singular values of corresponding $m \times d$ sketches with the ℓ_1 metric: $||\frac{1}{\sqrt{n}}\vec{\sigma}(X) - \frac{1}{\sqrt{m}}\vec{\sigma}(\hat{X})||_1$, where $\vec{\sigma}$ is the vector of corresponding singular values to sample or sketch. Note that this ℓ_1 metric directly relates to differences of eigenvalues of corresponding covariance matrices, and hence is indicative of how well sketches may act as a stand-in for linear subspace analyses such as PCA.

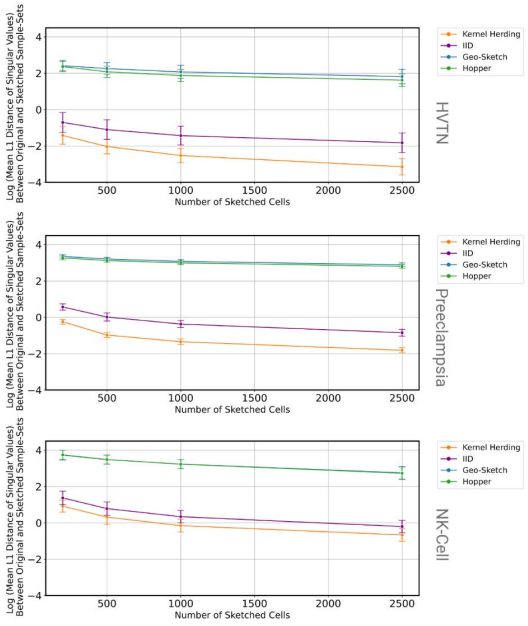

Figure 5: Singular Value Distribution Evaluation Differences in singular value distributions of cell $\times$ protein marker matrices as quantified with an L1-norm were used to summarize the quality of sketches over a range of sketch sizes, for preserving the overall cellular landscape in the HVTN, Preeclampsia and NK-Cell datasets (top, middle, bottom, respectively).

As before, we see that Geo-sketch and Hopper sketches are also unable to act as stand-ins for a singular-value analysis; a problem that does not improve as sketch sizes increase. Given that many single cell analyses rely on capturing the lower dimensional structure in samples (e.g. [30, 38]), this underlying bias in previous sketching approaches is limiting to their ability to act as reliable stand-ins. Moreover, we similarly observe that our proposed Kernel Herding sketches act as more faithful stand-ins to the naive IID sketches.

3.5 Cluster Frequency Fidelity

Next, we study sketches' abilities to retain the overall cellular frequencies that were found in the original sample. Cell-type frequencies can be crucial for clinical outcome prediction and biological interpretation [7, 37]; thus, a sketch's ability to retain cell-type frequencies is pivotal for its ability to act as a stand-in to the original set in many impactful tasks. Here we automatically detect cell-populations through an unsupervised, kmeans cluster analysis. All cells in the original sample-set are used to compute the cluster centroids, which then provide the cluster association for cells from sketches of that set. Once the centroids are computed for the original sample, each sketch is summarized according to the frequencies of the clusters in that sketch. That is, for each sketch, we compute the portion of its cells that were assigned to each cluster, $\rho(\hat{X}) \in [0,1]^K$, $\rho_k(\hat{X}) = \frac{1}{|\hat{X}|}\sum_{x \in \hat{X}} \mathbb{I}\{k = \text{argmin}_c||x - \nu_c||\}$, where $\nu_1, \ldots, \nu_K$ are the cluster centroids. Similarly to above, we may compare the outcomes of analyzing the original sample-set, X, to that of a corresponding sketch, $\hat{X}$, with an ℓ_1 metric: $||\rho(X) - \rho(\hat{X})||_1$ and average this distance over all sets in the dataset.

We plot the ℓ_1 discrepancies in frequencies for various sketching cardinalities and number of clusters in Fig. 6. Following the

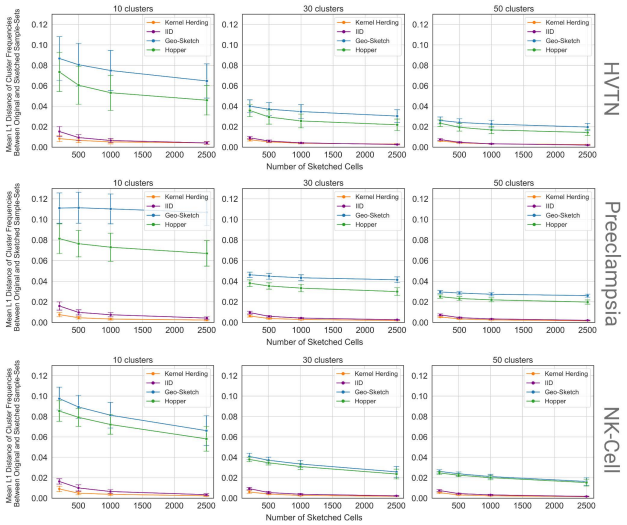

Figure 6: Cell-Population Frequency Evaluation k-means was used to partition sketched sample-sets (between 200 to 2500 cells) into cell-populations (10, 30, 50 clusters). In comparison to IID, Geometric Sketching, and Hopper, Kernel Herding sketches most closely preserve the frequencies of cell-populations observed using all cells in the sample-set.

same major pattern to previous experiments, we see that alternative sketching approaches (Geo-Sketch and Hopper) were unable to properly act as a stand-in to computing the cell-population frequencies in the original set. This trend is also not assuaged by increasing those sketch's sub-sampling size. Kernel Herding avoids these issues, whilst also providing a better representative sketch than IID sub-sampling. We see similar results across various number of clusters, which studies the sketches' fidelity with different granularities of cell-populations.

3.6 Classification Effectiveness

Finally, we explicitly test to see if the cell-population discrepancies found in Fig. 6 are relevant to producing immunological features that are useful for classifying samples according to the clinical outcomes of their associated individuals. That is, here we explicitly test that clinically relevant cell populations (such as *rare-cells*) are maintained in sketches. We hypothesized that the sketches with more accurately represented frequencies (e.g. as in Sec. 3.5) may be more predictive of a clinical outcome of interest.

As in Sec. 3.5, for each of the three cytometry datasets, cells were clustered into one of thirty clusters. Here the clusters are found from an aggregate collection of the multiple sketched sample-sets in each dataset. That is, the clusters are determined via a concatenation of all the sketched cells from all the sample sets that are in respective training sets. The frequency, or proportion of cells assigned to each population, $\rho(\hat{X})$, was used as input for a downstream classification task to predict the clinical outcome for each sketched set. Note that here we trained a classifier based on multi-sample dataset $\left\{(\rho(\hat{X}_i), y_i)\right\}_{i=1}^{N}$, where $\hat{X}_i$ is the sketch for the i-th sample-set, and y_i is the corresponding label. Classification experiments were performed in the HVTN, Preeclampsia and NK-Cell datasets (left, middle, and right of Figure 7, respectively) by splitting individuals according to leave-one-out cross validation and training an RBF support vector machine (SVM). Surprisingly, we found that notwithstanding the explicit focus on maintaining rare-cell types, Geo-sketch and Hopper were unable to produce sketches that lead to better accuracies than IID sub-sampling in this setting. In contrast, we found that our proposed Kernel Herding sketches lead to higher mean accuracies than IID sketching. This held true for other cluster sizes as well (K=15,K=50). Our results show that in general, sketching with Kernel Herding produces sketches that are more clinically predictive than the baseline methods. This is especially apparent in the preeclampsia dataset, where IID sketching produces highly variable classification accuracy.

3.7 Single-cell RNA Sequencing Fidelity

For a cross-modality comparison, we tested the performance of Kernel Herding to the sketches obtained using Geometric Sketching, Hopper, and IID subsampling on single-cell RNA sequencing (scRNA-seq) data. This modality poses an additional challenge, as scRNA-seq data typically contains 20-30 thousand gene measurements across all cells and has a high degree of sparsity and technical noise due to capture inefficiency, amplification noise, and stochasticity [21]. Leveraging a single cell RNA sequencing dataset of patients with multiple sclerosis (MS) and healthy controls, we tested whether sketched cells resulted in similar random function evaluations, similar singular value distributions, and gave similar cell population frequencies. As shown in Figure 8, we found that Kernel Herding produces sketches that are most closely aligned with the results obtained using the full original sample. More specifically, across a range of sketch sizes for each sample-set, Kernel Herding achieves the least ℓ_1 distance to the random function estimates in the original dataset (Fig 8A), has the most similar singular value distributions to the original data (Figure 8B), and best preserves cell population frequencies with the least ℓ_1 norm between true and sketched cell populations (Figure 8C).

Table 1: Runtimes (in seconds) of sketching methods on HVTN, Preeclampsia and NK-Cell datasets (number of cells (n and number of features (d) shown underneath in parentheses) as measured on an Intel(R) Xeon(R) Gold 6226R CPU. Note that our Kernel Herding implementation was not optimized and was coded for readability.

Methods	HVTN (n=200k, d=11)	Preeclampsia (n=215k, d=33)	NK Cell (n=13k, d=43)
IID	0.012 ± 0.001	0.016 ± 0.003	0.004 ± 0.001
Hopper	3.13 ± 0.12	4.43 ± 0.15	1.12 ± 0.09
Geo Sketch	23.76 ± 0.06	77.92 ± 0.15	12.29 ± 0.06
Kernel Herding	26.56 ± 4.38	22.59 ± 1.38	3.64 ± 0.35

4 DISCUSSION AND CONCLUSION

Here, we presented a distribution-based, Kernel Herding approach to select a limited number of representative cells from each sample-set. Of particular note, we recast the focus of sketching sample-sets to providing a smaller sketch that can act as a *stand-in* to the original set. That is, we explicitly and quantitatively assess sketch's ability to faithfully maintain downstream outcomes when used in place of the original set. In contrast to existing sketching approaches, such as, Geometric Sketching [19] and Hopper [13], we found that Kernel Herding strikes a powerful middle-ground between preserving rare cell-populations, while also representing all major cell-types and retaining their relative frequencies. Moreover, maintaining the overall distribution of the original sample is necessary as adequate preservation of cell-population frequencies is important for linking cellular heterogeneity to clinical phenotype or external variables of interest [7, 37], and for developing novel diagnostics or prognostics [14, 15, 18]. Given the modern widespread use of cytometry and scRNA-sequencing technologies in clinical applications with large patient cohorts, the presented Kernel Herding based sketching approach makes such data more manageable for downstream analysis and interpretation. We showed that Kernel Herding was effective across multiple varied single cell modalities including flow and mass cytometry, as well as scRNA-seq data.

We demonstrated the usefulness of Kernel Herding at providing sketches that can serve as a stand-in for the original sample-set by measuring the fidelity of non-linear function evaluations, singular value distributions, and cell-population frequencies between the original and sketched sample-sets. Additionally, we showed the usefulness of sketching through Kernel Herding for downstream tasks, such as clinical outcome prediction; here, Kernel Herding had the most stable performance across datasets, number of clusters, and number of sketched cells. This consistency makes it a reliable method that can be employed to obtain a representative subset,

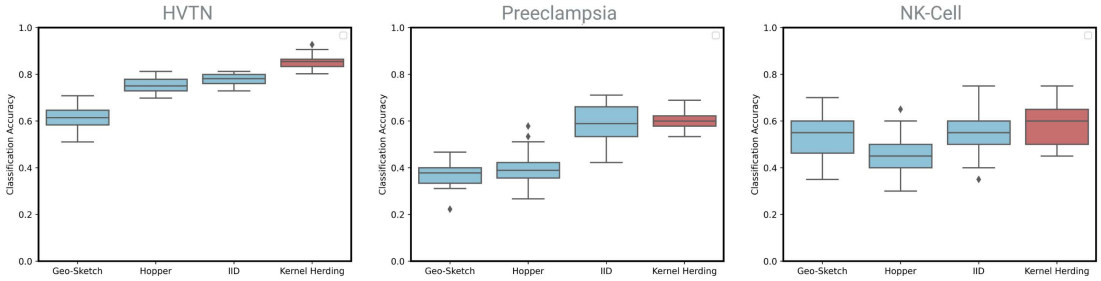

Figure 7: *Clinical Outcome Classification Accuracy.* Sketches of 500 samples per sample set were obtained each with Geometric Sketching, Hopper, IID subsampling, and Kernel Herding and cells were partitioned amongst 30 clusters forming cell-populations. The cell-population frequencies were used as features to predict the clinical outcomes of the associated individuals for each sample-set in the HVTN, Preeclampsia, and NK-Cell datasets (left, middle, and right, respectively). We performed leave-one-out cross validation experiments for this classification task. Kernel Herding and IID produce sketches and associated features that are more predictive than those obtained through Geometric Sketching. Kernel Herding also significantly reduces the variance in classification accuracy in the Preeclampsia dataset.

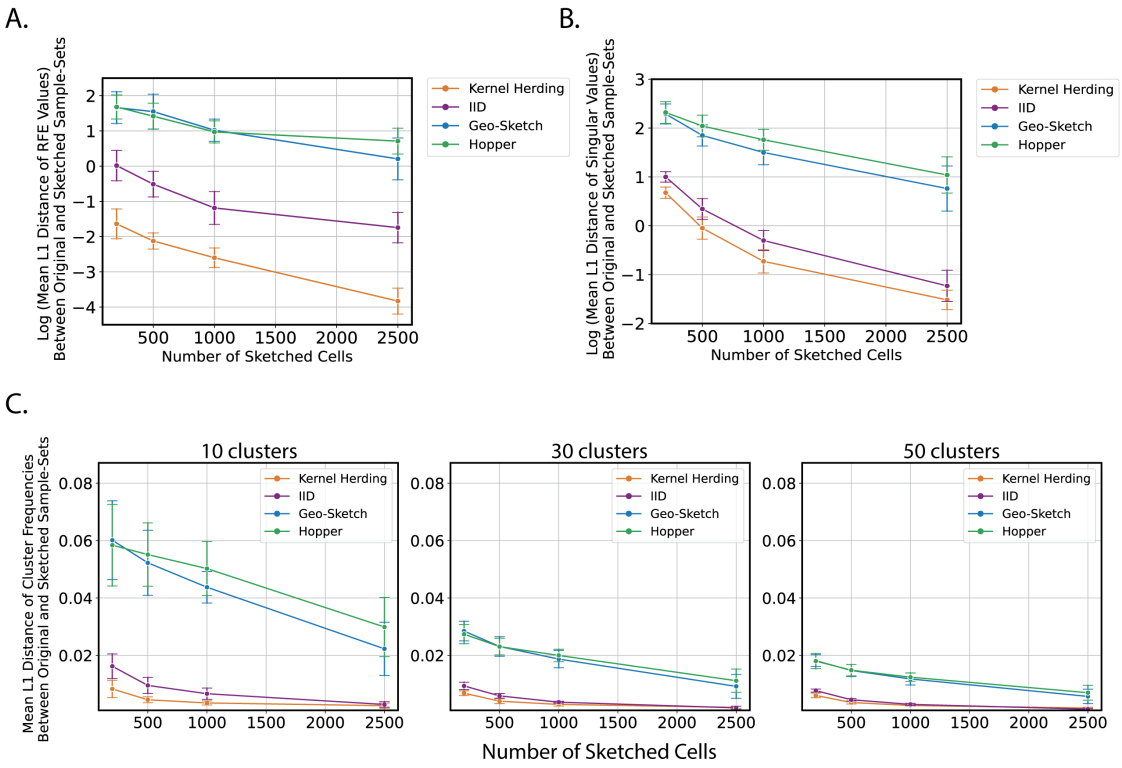

Figure 8: Single-cell RNA Sequencing Evaluation We evaluated Hopper, Geometric Sketching, IID, and Kernel Herding on producing sketches that preserve the overall cellular landscape of single-cell RNA sequencing data. Method performance was quantified using an L1 norm between random function evaluations of original and sketched samples (A), an L1 norm between singular value distributions (B), and an L1 norm between true and sketched cell-population frequencies (C) across a range of sketch sizes (200 - 2500). Sketches with Kernel Herding produce more similar random function evaluations, similar singular value distributions, and best preserve the cell population frequencies that are observed when using all cells.

small or large, of the original distribution of samples. The runtimes of different methods reported in Table 1 add to the practical usefulness of Kernel Herding.

Future work could consider the implications of sketching with Kernel Herding in a broader range of tasks required to understand single-cell datasets, such as, differential abundance analysis [12, 26, 40], or for rapid identification of phenotype-associated cells

[8]. Finally, another area of future work may focus on generating variable sketch sizes across different subsections of the cellular landscape, depending on prior knowledge or scientific question.

ACKNOWLEDGMENTS

This research was partly funded by NSF grant IIS2133595 and by NIH 1R01AA02687901A1.

REFERENCES

[1] Nima Aghaeepour, Greg Finak, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, and Richard H Scheuermann. 2013. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods* 10, 3 (2013), 228–238.

[2] Nima Aghaeepour, Edward A Ganio, David Mcilwain, Amy S Tsai, Martha Tingle, Sofie Van Gassen, Dyani K Gaudilliere, Quentin Baca, Leslie McNeil, Robin Okada, Mohammad S Ghaemi, David Furman, Ronald J Wong, Virginia D Winn, Maurice L Druzin, Yaser Y El-Sayed, Cecele Quaintance, Ronald Gibbs, Gary L Darmstadt, Gary M Shaw, David K Stevenson, Robert Tibshirani, Garry P Nolan, David B Lewis, Martin S Angst, and Brice Gaudilliere. 2017. An immune clock of human pregnancy. *Science immunology* 2, 15 (2017).

[3] Prabhu S Arunachalam, Florian Wimmers, Chris Ka Pun Mok, Ranawaka APM Perera, Madeleine Scott, Thomas Hagan, Natalia Sigal, Yupeng Feng, Laurel Bristow, Owen Tak-Yin Tsang, et al. 2020. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science* 369, 6508 (2020), 1210–1220.

[4] Eirini Arvaniti and Manfred Claassen. 2017. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature communications* 8, 1 (2017), 1–10.

[5] Sean C Bendall, Garry P Nolan, Mario Roederer, and Pratip K Chattopadhyay. 2012. A deep profiler's guide to cytometry. *Trends in immunology* 33, 7 (2012), 323–332.

[6] Alain Berlinet and Christine Thomas-Agnan. 2011. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

[7] Robert V Bruggner, Bernd Bodenmiller, David L Dill, Robert J Tibshirani, and Garry P Nolan. 2014. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences* 111, 26 (2014), E2770–E2777.

[8] Daniel B Burkhardt, Jay S Stanley, Alexander Tong, Ana Luisa Perdigoto, Scott A Gigante, Kevan C Herold, Guy Wolf, Antonio J Giraldez, David van Dijk, and Smita Krishnaswamy. 2021. Quantifying the effect of experimental perturbations at single-cell resolution. *Nature Biotechnology* (2021), 1–11.

[9] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, Cole Trapnell, and Jay Shendure. 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566, 7745 (feb 2019), 496–502.

[10] Yutian Chen, Max Welling, and Alex Smola. 2010. Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. 109–116.

[11] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[12] Emma Dann, Neil C Henderson, Sarah A Teichmann, Michael D Morgan, and John C Marioni. 2021. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology* (2021), 1–9.

[13] Benjamin DeMeo and Bonnie Berger. 2020. Hopper: a mathematically optimal algorithm for sketching biological data. *Bioinformatics* 36, Supplement_1 (2020), i236–i241.

[14] Ramin Fallahzadeh, Franck Verdonk, Ed Ganio, Anthony Culos, Natalie Stanley, Ivana Marić, Alan L Chang, Martin Becker, Thanaphong Phongpreecha, Maria Xenochristou, et al. 2021. Objective Activity Parameters Track Patient-Specific Physical Recovery Trajectories After Surgery and Link With Individual Preoperative Immune States. *Annals of Surgery* (2021).

[15] Edward A Ganio, Natalie Stanley, Viktoria Lindberg-Larsen, Jakob Einhaus, Amy S Tsai, Franck Verdonk, Anthony Culos, Sajjad Ghaemi, Kristen K Rumer, Ina A Stelzer, Dyani Gaudilliere, Eileen Tsai, Ramin Fallahzadeh, Benjamin Choisy, Henrik Kehlet, Nima Aghaeepour, Martin S Angst, and Brice Gaudilliere. 2020. Preferential inhibition of adaptive immune system dynamics by glucocorticoids in patients after acute surgical trauma. *Nature communications* 11, 1 (2020), 1–12.

[16] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2008. A Kernel Method for the Two-Sample Problem. *Journal of Machine Learning Research* 1 (2008), 1–10.

[17] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.

[18] Xiaoyuan Han, Mohammad S Ghaemi, Kazuo Ando, Laura S Peterson, Edward A Ganio, Amy S Tsai, Dyani K Gaudilliere, Ina A Stelzer, Jakob Einhaus, Basile Bertrand, Natalie Stanley, Anthony Culoar, Athena Tanada, Julien Hedou, Eileen S Tsai, Ramin Fallahzadeh, Ronald J Wong, Amy E Judy, Virginia D Winn, Maurice L Druzin, Yair J Blumendeld, Mark A Hlatky, Cecele C Quaintance, Ronald S Gibbs, Brendan Carvalho, Gary M Shaw, David K Stevenson, Martin S Angst, Aghaeepour, and Brice Gaudilliere. 2019. Differential dynamics of the maternal immune system in healthy pregnancy and preeclampsia. *Frontiers in immunology* 10 (2019), 1305.

[19] Brian Hie, Hyunghoon Cho, Benjamin DeMeo, Bryan Bryson, and Bonnie Berger. 2019. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell systems* 8, 6 (2019), 483–493.

[20] W Evan Johnson, Cheng Li, and Ariel Rabinovic. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8, 1 (jan 2007), 118–127.

[21] Peter V Kharchenko, Lev Silberstein, and David T Scadden. 2014. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11, 7 (jul 2014), 740–742.

[22] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, Rachel Finck, Amanda L Gedman, Ina Radtkae, James R Downing, Dana Pe'er, and Garry P Nolan. 2015. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 162, 1 (2015), 184–197.

[23] Thomas Liechti, Lukas M Weber, Thomas M Ashhurst, Natalie Stanley, Martin Prlic, Sofie Van Gassen, and Florian Mair. 2021. An updated guide for the perplexed: cytometry in the high-dimensional era. *Nature Immunology* 22, 10 (2021), 1190–1197.

[24] Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, Sergei Rybakov, Alexander V Misharin, and Fabian J Theis. 2021. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* (aug 2021).

[25] Malte D Luecken and Fabian J Theis. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology* 15, 6 (2019), e8746.

[26] Aaron TL Lun, Arianne C Richard, and John C Marioni. 2017. Testing for differential abundance in mass cytometry data. *Nature methods* 14, 7 (2017), 707–709.

[27] Aaron T L Lun, Karsten Bach, and John C Marioni. 2016. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17 (apr 2016), 75.

[28] Divij Mathew, Josephine R Giles, Amy E Baxter, Derek A Oldridge, Allison R Greenplate, Jennifer E Wu, Cécile Alanio, Leticia Kuri-Cervantes, M Betina Pampena, Kurt D'Andrea, Sasikanth Manne, Zeyu Chen, Yinghui Jane Huang, John P Reilly, Ariel R Weisman, Caroline A G Ittner, Oliva Kuthuru, Jeanette Dougherty, Kito Nzingha, Nicholas Han, Justin Kim, Ajinkya Pattekar, Eileen C Goodwin, Elizabeth M Anderson, Madison E Weirick, Sigrid Gouma, Claudia P Arevalo, Marcus J Bolton, Fang Chen, Simon F Lacey, Holly Ramage, Sara Cherry, Scott E Hensley, Sokratis A Apostolidis, Alexander C Huang, Laura A Vella, UPenn COVID Processing Unit, Michael R Betts, Nuala J Meyer, and E John Wherry. 2020. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications". *Science* 369, 6508 (sep 2020).

[29] Sebastian Mika, Bernhard Schölkopf, Alexander J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. 1998. Kernel PCA and De-noising in feature spaces.. In *NIPS*, Vol. 11. 536–542.

[30] Kevin R Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, Natalia B Ivanova, Guy Wolf, and Smita Krishnaswamy. 2019. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* 37, 12 (dec 2019), 1482–1492.

[31] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. 2017. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning* 10, 1-2 (2017), 1–141.

[32] Junier Oliva, Willie Neiswanger, Barnabás Póczos, Jeff Schneider, and Eric Xing. 2014. Fast distribution to real regression. In *Artificial Intelligence and Statistics*. PMLR, 706–714.

[33] Ali Rahimi, Benjamin Recht, et al. 2007. Random Features for Large-Scale Kernel Machines.. In *NIPS*, Vol. 3. Citeseer, 5.

[34] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen,

Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. 2017. The Human Cell Atlas. *Elife* 6 (dec 2017).

[35] David Schafflick, Chenling A Xu, Maike Hartlehnert, Michael Cole, Andreas Schulte-Mecklenbeck, Tobias Lautwein, Jolien Wolbert, Michael Heming, Sven G Meuth, Tanja Kuhlmann, Catharina C Gross, Heinz Wiendl, Nir Yosef, and Gerd Meyer Zu Horste. 2020. Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nat. Commun.* 11, 1 (jan 2020), 247.

[36] Matthew H Spitzer and Garry P Nolan. 2016. Mass cytometry: single cells, many features. *Cell* 165, 4 (2016), 780–791.

[37] Natalie Stanley, Ina A Stelzer, Amy S Tsai, Ramin Fallahzadeh, Edward Ganio, Martin Becker, Thanaphong Phongpreecha, Huda Nassar, Sajjad Ghaemi, Ivana Maric, Anthony Culos, Alan L Chang, Maria Xenochristou, Xiaoyuan Han, Camilo Espinosa, Kristen Rumer, Laura Peterson, Franck Verdonk, Dyani Gaudilliere, Eileen Tsai, Dorien F Feyaerts, Jakob Einhaus, Kazuo Ando, Ronald J Wong, Gerlinde Obermoser, Gary M Shaw, David K Stevenson, Martin S Angst, Brice Gaudilliere, and Nima Aghaeepour. 2020. VoPo leverages cellular heterogeneity for predictive modeling of single-cell data. *Nature communications* 11, 1 (2020), 1–9.

[38] Britta Velten, Jana M Braunger, Ricard Argelaguet, Damien Arnol, Jakob Wirbel, Danila Bredikhin, Georg Zeller, and Oliver Stegle. 2022. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat. Methods* (jan 2022).

[39] Vladimir Vovk. 2013. Kernel ridge regression. In *Empirical inference*. Springer, 105–116.

[40] Lukas M Weber, Malgorzata Nowicka, Charlotte Soneson, and Mark D Robinson. 2019. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications biology* 2, 1 (2019), 1–11.

[41] Haidong Yi and Natalie Stanley. 2021. CytoSet: predicting clinical outcomes via set-modeling of cytometry data. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* (Gainesville, Florida) *(BCB '21", Article 47)*. Association for Computing Machinery, New York, NY, USA, 1–8.