

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Seminars in Diagnostic Pathology

journal homepage: www.elsevier.com/locate/sem_dp

Review article

Artificial intelligence in clinical multiparameter flow cytometry and mass cytometry—key tools and progress

Franklin Fuda^a, Mingyi Chen^a, Weina Chen^a, Andrew Cox^{b,c,*}^a Department of Pathology and Laboratory Medicine, University of Texas, Southwestern Medical Center, Dallas, Texas, USA^b Lyda Hill Department of Bioinformatics, University of Texas, Southwestern Medical Center, Dallas, Texas, USA^c Department of Cell and Molecular Biology, University of Texas, Southwestern Medical Center, Dallas, Texas, USA

ARTICLE INFO

Keywords:

Flow cytometry
 Mass cytometry
 Artificial intelligence
 Machine learning
 The Hematology laboratory

ABSTRACT

There are many research studies and emerging tools using artificial intelligence (AI) and machine learning to augment flow and mass cytometry workflows. Emerging AI tools can quickly identify common cell populations with continuous improvement of accuracy, uncover patterns in high-dimensional cytometric data that are undetectable by human analysis, facilitate the discovery of cell subpopulations, perform semi-automated immune cell profiling, and demonstrate potential to automate aspects of clinical multiparameter flow cytometric (MFC) diagnostic workflow. Utilizing AI in the analysis of cytometry samples can reduce subjective variability and assist in breakthroughs in understanding diseases.

Here we review the diverse types of AI that are being applied to clinical cytometry data and how AI is driving advances in data analysis to improve diagnostic sensitivity and accuracy. We review supervised and unsupervised clustering algorithms for cell population identification, various dimensionality reduction techniques, and their utilities in visualization and machine learning pipelines, and supervised learning approaches for classifying entire cytometry samples. Understanding the AI landscape will enable pathologists to better utilize open source and commercially available tools, plan exploratory research projects to characterize diseases, and work with machine learning and data scientists to implement clinical data analysis pipelines.

List of Abbreviations

AI	Artificial Intelligence
ML	Machine Learning
MFC	Multi-color Flow Cytometry

Background

The World Health Organization's classification of hematopoietic and lymphoid tissue recognizes lineage assignment by immunophenotyping as essential in establishing and subcategorizing hematolymphoid malignancies. Multiparameter flow cytometry (MFC) is a major contributor to establishing immunophenotypes of cell populations. Traditionally, processing and staining of samples for flow cytometry and analysis of flow cytometric data have been labor-intensive processes that require highly trained technicians and pathologists. Recent advancements in automated processing and staining have shown promise in increasing workflow efficiency and reducing human error. Recent advancements in

data analysis via artificial intelligence (AI) demonstrate potential to improve intra- and inter-laboratory standardization and the accuracy and sensitivity of diagnosis. As AI continually improves, it holds the potential to augment and automate aspects of the gating workflow, uncover insights into diseases, decrease expenses, and increase laboratory efficiency.

Current flow cytometric analysis involves manual gating of bivariate plots in a sequential or non-sequential manner. Hierarchical gating systematically evaluates bivariate plots using rigid threshold values to create “child” populations from “parent” populations with the gating criteria of the parent gate passed onto the child gate. Non-hierarchical gating (e.g., Boolean gating and manual cluster analysis) employs an ‘and,’ ‘or’ and ‘not’ logic and relies on the expertise of the operator to evaluate any number of bivariate plots, often applying non-rigid, hand drawn gates to arrive at the final labeled cell populations of interest. With either approach, the results of analyzing flow cytometry samples via manual gating are operator dependent, and differences in analytic approaches and the knowledge and experience of operators lead to

* Corresponding author at: UT Southwestern Medical Center, Department of Bioinformatics E4; 5323 Harry Hines Blvd., Dallas, Texas 75390-9365, USA.

E-mail address: andrew.cox@utsouthwestern.edu (A. Cox).

https://doi.org/10.1053/j.sem_dp.2023.02.004

Available online 5 March 2023
 0740-2570/© 2023 Elsevier Inc. All rights reserved.

variability of results^{1,2}. While non-hierarchical gating may overcome some limitations of hierarchical gating and achieve highly accurate results, non-hierarchical gating will encounter greater challenges in timeliness as more in-depth characterizations of cell populations are made possible by an increasing number of markers present in a single panel. Furthermore, manual gating in general is sometimes limited by the fact that two cell sub-populations may not be fully resolved with any combination of bivariate plots, even though they can be fully resolved if all the markers are considered simultaneously in high dimensional space. AI is well-suited for simultaneously analyzing hundreds of unique combinations of markers to identify cell populations and presenting them in an interpretable format (Fig. 1). With modern flow cytometers showing potential to resolve over 40 fluorescent markers per cell, there is a rich opportunity to leverage AI approaches to both automate and augment analysis workflows in clinical flow cytometry.

AI research in clinical MFC

Introduction

Emerging AI tools show promise to augment routine MFC and mass cytometry analyses by automatically identifying cell populations, reducing analysis variability, and predicting when follow-on procedures are necessary (e.g., add-on tubes). AI algorithms paired with large clinical databases, such as Cytognos's EuroFlow, can automatically identify all major cell types and even flag minor alterations in immunophenotype that can be missed with manual analysis, provided the pipeline from sample processing to data acquisition is highly standardized. State-of-the-art AI algorithms have been shown to produce unbiased results and can reduce variability by 94% compared to manual approaches, as demonstrated by the FlowCAP (Critical Assessment of Population Identification Methods) competition³. Deciphering the immunophenotype of many hematolymphoid neoplasms requires sequential rounds of MFC marker panels, and AI has shown potential to flag when an add-on tube is required⁴. In addition to automating routine analyses, data-driven studies leveraging AI with FCS/mass cytometry have uncovered previously unrecognized cell subpopulations of diseased patients. As major innovations in AI continue to make their way into clinical FCS laboratories, an understanding of the types of AI algorithms is key for recognizing their potential and limitations for guiding routine analyses and clinical discovery. Table 1 highlights many available AI software and packages for use with flow and mass cytometry data.

AI is a broad field loosely defined as any technique that enables computers to mimic human intelligence. Machine learning (ML) is a subfield of AI that involves statistical algorithms that enable machines to improve at performing tasks⁵. Classical ML algorithms for classification include logistic regression, k-nearest neighbors, naive Bayes, decision trees, ensemble algorithms including random forest, support vector machines, and neural networks⁶. Deep learning (DL) is a type of neural network architecture, and thus a subtype of ML, that involves stacking

multiple layers of neural networks in sequence. This allows for the learning of more complex representations of data, and deep learning networks are often trained using large amounts of data. The algorithms made to automate gating, classify samples with a certain disease, find and remove outlier events, etc., all fall into the ML or DL categories.

ML and DL algorithms can be supervised or unsupervised, and this distinction has downstream implications regarding utility. In the context of cytometry, unsupervised algorithms are used to cluster data points, which can be individual cells or an entire MFC or mass cytometry sample. Unsupervised algorithms operate using built-in assumptions regarding the structure of the data to produce their results; however, clusters must still be manually inspected by the operator and assigned an identity and accuracy. Over the greater course of a decade, unsupervised algorithms for MFC and mass cytometry have been innovated to identify rare subpopulations with only a few dozen cells, cluster large cytometry files with tens of millions of cells in seconds and facilitate new insights into a variety of biological phenomena.

Supervised algorithms require that for each cell or sample, a corresponding label is given. These labels are often referred to as the ground truth and are used to train the algorithm and assess performance. Supervised algorithms have been used to predict cell type identities and sample-level labels, like a disease label or whether an add-on flow tube is required to further characterize an abnormal population of cells. While the output of supervised models is quick and easy to comprehend, the veracity of predicted labels is dependent on the fidelity of the ground truth labels. Further, the general utility of the trained model is highly dependent on the similarity, statistically speaking, between any new data the model encounters with that of the training set. If performance on the new data is lacking, a deeper analysis can be done to understand differences in the underlying data structure. After careful consideration, new data can be incorporated into the original dataset and the model can be retrained, but it is important to first understand why the model's performance was originally insufficient on the new data.

Preparation of data for machine learning

In MFC clinical practice, standardization is highly emphasized to ensure consistent results from day to day. Samples are processed using similar protocols, cytometers are calibrated daily using reference standards, and a consistent antibody marker panel is often used for extended periods of time. Without standardization, data can vary greatly and become challenging for algorithms and even experts to compare with high accuracy.

Different AI algorithms require varying amounts of pre-processing to achieve optimal performance. Prior to analysis, flow cytometry data must be compensated, transformed, and cleaned (removing debris and doublets) to faithfully gate cell populations of interest. Similarly, many algorithms will perform better if fed MFC data that has been subject to a similar pre-processing workflow. In conventional MFC, spillover of fluorophore emission spectra is corrected by compensation. Applying compensation to raw data equates to solving a system of linear equations

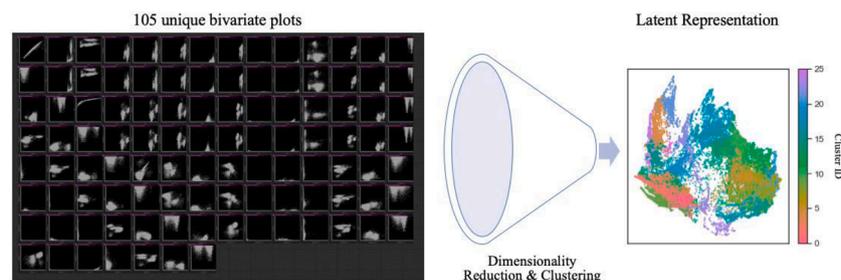


Fig. 1. – AI can identify clusters from high-dimensional MFC data and facilitate human interpretable visualization. A single 10-channel MFC sample contains 105 unique bivariate plots (left). AI identifies clusters, representative of cell populations, and compresses them into a single, unitless bivariate plot (middle), referred to as a latent representation (right).

Table 1
Software AI tools for cytometry data.

Algorithm	Algorithm Brief Description	Environment & Availability	Ref.
<i>Preprocessing & Quality Control</i>			
flowAI	Cleaning based on signal, flow rate, and outliers	R package from Bioconductor	15
flowClean	Detecting anomalies in signal	R package from Bioconductor	16
flowCut	Detecting anomalies in signal	R package from Bioconductor	46
flowVS	Data transformation	R package from Bioconductor	47
OTflow	Optimal transformation selection	Surfdrive repository	11
<i>Unsupervised Clustering</i>			
PhenoGraph	Graph creation via KNN clustering followed by Jaccard similarity. Final partitioning of the graph via Louvain or Leiden algorithm into communities based on modularity optimization.	Python package, Cytobank	39,48
FLOCK	Partitioning of each dimension into bins, followed by merging of dense regions, and density-based clustering	C source code (also available in ImmPort online platform)	40
FlowPeaks	Peak-finding on smoothed density function generated by k-means; using finite mixture model.	R package from Bioconductor	41
X-shift	Weighted k-nearest-neighbor density estimation with identification of local maxima and construction of graph with cluster merging.	Standalone application (VorteX) with graphical interface (command-line version also available)	42
MegaClust	Parallel density-based hierarchical clustering.	Standalone software with CLI (GitHub repository) and paid service platform	36,37
FlowGrid	Time and memory efficient grid-based adaptation of density-clustering for detection of high-density regions and outliers.	Standalone software with CLI (GitHub repository)	43
Gaussian Mean Shift	Density-based clustering with a kernel bandwidth parameter.	Open-source standalone software application through Cytosplore	49
ACCENSE	t-SNE followed by density-based peak-finding and clustering of t-SNE components.	Standalone application with graphical interface	50
ClusterX	Density-based clustering on t-SNE projection map; faster than DensVM.	R package (Cytokit) from Bioconductor	44
immunoclust	Iterative clustering with mixture models and classification likelihood	R package from Bioconductor	25
ASPIRE	Nonparametric Bayesian mixture modeling using Dirichlet process that explicitly models random effects.	C++ standalone software; MATLAB scripts	26
FlowSOM	Self-organizing maps, with merging of clusters via hierarchical consensus.	R package from Bioconductor; Cytosplore; Cytobank	32,33
Rclusterpp	Memory-efficient hierarchical clustering for large-scale use.	R package from GitHub (older version on CRAN)	34
BayesFlow	parametric Bayesian multivariate mixture modeling explicitly		29

Table 1 (continued)

Algorithm	Algorithm Brief Description	Environment & Availability	Ref.
	modeling variation in cell population shape followed by Markov Chain Monte Carlo sampling and merging of model components		
flowGM	Gaussian mixture model with number of clusters determined via Bayesian information criterion	MATLAB from Statistics Toolbox; R from flowCore package	30
SamSPECTRAL	Efficient spectral clustering using density-based downsampling	R package from Bioconductor	38
flowClust/ Merge	multivariate t mixture modeling and entropy-based merging	R package from Bioconductor	22,23
SWIFT	Gaussian mixture model followed by splitting and merging of clusters.	GUI via MATLAB	27
flowMeans	k-means clustering and merging to allow non-spherical clusters	R package from Bioconductor	31
SPADE	Organizes clusters into a branching hierarchy using density-based sampling, k-means clustering and minimum spanning trees	R package from GitHub (older version on Bioconductor); Cytobank	35
<i>Dimensionality Reduction</i>			
viSNE	Non-linear dimensionality reduction	Cytobank; Cytosplore; Cytokit	48,51
A-tSNE	Approximated and user steerable tSNE	Cytosplore	49,52
HSNE	Hierarchical tSNE	Cytosplore	53
Fit-SNE	Non-linear dimensionality reduction	R Bioconductor; Python; MATLAB	54
tSNE-CUDA	GPU-accelerated non-linear dimensionality reduction	Cytobank	55
opt-SNE	Non-linear dimensionality reduction	Cytobank	56
UMAP	Non-linear dimensionality reduction	R; Python; FlowJo	4,57,58
PHATE	Non-linear dimensionality reduction	Python	59,60
<i>Supervised Cell Identity Classification</i>			
CellCNN	Representation learning using a convolutional neural network adapted to process unordered multi-cell inputs	Python	61
DensVM	Density-based clustering on t-SNE projection map; like ACCENSE, with additional support vector machine to classify uncertain points.	R package (cytofit) from Bioconductor	45
flowDensity	Supervised density-based clustering	R package from Bioconductor	62
flowLearn	Semi-supervised clustering using density-based alignments	R package from Bioconductor	63
Pipelines Cytofast	Visual and quantitative analysis for immune profiling after clustering	R package from Bioconductor	64
Cytosplore	Interactive visual analysis system containing A-tSNE, HSNE, and SPADE	Interactive standalone software	49
Cytokit	Preprocessing; clustering (DensVM, FlowSOM, ClusterX, or Phenograph); data visualization (PCA, t-SNE)	R package from Bioconductor, GUI with Shiny application	44
Citrus		R package with GUI; Cytobank	21

(continued on next page)

Table 1 (continued)

Algorithm	Algorithm Brief Description	Environment & Availability	Ref.
	Unsupervised clustering and regularized regression model		
OpenCyto	Template-based automated gating	R package from Bioconductor/GUI with shinyCyto application	65
TerraFlow	Automated characterization of differences in cell population between labeled sample types. Streamlined combinatorics, linear regression, network analysis, Feature elimination with weighted Lasso regression	Commercial	66

to produce the spillover matrix which is then inverted and multiplied with the raw data to produce compensated data⁷. Usually, compensation is applied prior to algorithm training. However, a recent study by Camp *et al.* shows that deep neural networks, which are universal function approximators^{8,9}, are not impacted in their ability to classify cell types in samples from patients with myelodysplastic syndrome¹⁰. In spectral MFC, all emission spectra for each fluorophore are captured and spillover is corrected by unmixing. While adjusting all fluorescent data with one transformation is preferred over raw values, sometimes for specific samples, fluorophores, antibodies, etc. there are certain transformations that facilitate the separation of distinct cell populations while other transformations show only a single homogeneous population. OTflow (optimal transformation of flow cytometry data), is an algorithmic approach for choosing proper channel-wise transformations by Folcarrelli *et al.* which prevents improper transformations that would obscure fluorescent intensities from separate populations from being applied¹¹. On a related note, many ML algorithms perform best on scaled data. Scaling the data ensure that all features are considered equally during training. A common approach to preserve each feature's distribution is to transform with a minimum/maximum scaler that sets values for all features between 0 and 1. Removing debris and doublets can be essential for many algorithms. Doublets are defined as two cells that pass simultaneously through the light source simultaneously, generating erroneous events with elevated cell size, complexity, and fluorescence intensity. Removing doublets better attunes ML models to cell populations with elevated fluorescent intensities. Debris is often removed prior to downstream analyses as large amounts of it can occlude populations of interest. Debris is relatively easy to identify when plotting side scatter versus forward scatter since cell populations of interest have well-established relative value ranges. In addition to removal of artifacts like doublets and debris, choosing transformations is a key step for optimal performance.

R software packages in Bioconductor offer an entire suite for pre-processing of MFC samples. Automated compensation and transformation can be done with flowCore¹² and flowUtils¹³, and the flowStats¹⁴ package has methods for normalizing sample values to facilitate better downstream automated analysis. FlowAI¹⁵ and flowClean¹⁶ help identify and remove low-quality events. FlowAI is an R package via Bioconductor that automatically detects anomalies that derive from stark changes in flow rate, instability of signal acquisition, and outliers in the dynamic range of intensity values. FlowClean is an algorithm that finds fluctuations in fluorescence intensities related to specific acquisition time points and flags them for quality checking. FlowViz¹⁷, FlowPlots, and ggcyto¹⁸ are R packages that facilitate efficient visualization of gates and transformations across all samples in the dataset. For more details on using these open-source tools and more in R for flow cytometry analysis, we refer readers to the relevant references¹⁹ and²⁰.

Unsupervised clustering

Unsupervised clustering algorithms group cells based on marker expression patterns in the high-dimensional space, considering all markers simultaneously without assumptions (Fig. 2 Left). In this manner, clustering is useful for automatically identifying previously unrecognized cell subpopulations as manual gating evaluates markers two at a time in a sequential fashion. Robert F. Murphy originally suggested a clustering strategy for cytometry data in 1985²¹. Since then, numerous clustering algorithms for cytometry data have been developed, of which many have been shown to closely recapitulate results produced from manual gating. However, performance can vary significantly between datasets and clustering performance should always be critically assessed on new panels or sample types prior to leveraging the AI to help semi-automate gating.

Unsupervised clustering algorithms for MFC

Many unsupervised clustering algorithms can run fully automated and require no input from the user, whereas others may require the user to adjust parameters such as the total number of expected cell populations. Distinct types of clustering algorithms vary in their underlying assumptions, performance, and capabilities. Model-based methods such as flowClust²², flowMerge²³, FLAME²⁴, immunoclust²⁵, ASPIRE²⁶, SWIFT^{27,28}, BayesFlow²⁹, and flowGM³⁰ fit statistical models to the distribution of the data to assign cells to clusters, while other model-based algorithms use a centroid-based approach (e.g., kMeans, flowMeans³¹) or a self-organizing map (FlowSOM^{32,33}) to fit the best representative for each cluster. Some use hierarchical clustering techniques (Rclusterpp³⁴, SPADE³⁵, ADICyt³, MegaClust^{36,37}), while others model the data using an underlying graph structure (e.g., SamsPECTRAL³⁸, PhenoGraph³⁹). Finally, several algorithms use data density, such as FLOCK⁴⁰, flowPeaks⁴¹, X-shift⁴², and FlowGrid⁴³, or the density of a reduced data space, for example, ACCENSE⁴⁴, DensVM⁴⁵, and ClusterX⁴⁴.

Unsupervised algorithm comparisons

Two studies have evaluated several unsupervised algorithms mentioned above and found that some perform more robustly and consistently in faithfully gating cell populations across various samples. In the FlowCAP competition, fully automated clustering was shown to closely approximate expert gating in low-dimensional flow cytometry samples with 3 – 10 fluorescent markers³. Interestingly, representative algorithms from each of the five types mentioned above performed comparably relative to expert manual gating, with mean F1 scores in the 0.85 – 0.89 range (ADICyt, flowMeans, FLOCK, FLAME, SamsPECTRAL). Also, creating an ensemble of the best algorithms enhanced performance beyond any single algorithm. Concluding the positive performance of many algorithms on the FlowCAP challenge, two major problems in the field included accurate detection of rare subpopulations and performance on cytometry samples with more markers and events. This led to new algorithms being developed like PhenoGraph, X-shift, FlowSOM to name just a few. A more recent comparison of clustering methods evaluated established and emerging algorithms across four high-dimensional mass cytometry datasets and two flow cytometry datasets containing rare cell subpopulations⁶⁷. In the mass cytometry samples, the algorithms were measured on their ability to identify 14 – 24 populations of interest ranging in size from several hundred to tens of thousands of cells. In the flow cytometry datasets, the task was accuracy of clustering the single rare subpopulation. Performance on the mass cytometry samples revealed much more discordance than found in the FlowCAP competition; this was attributed to notably inferior performance for all algorithms for one to three of the low-abundance populations. In the mass cytometry challenges, the mean F1 scores for the top six clustering algorithms ranged from 0.624 – 0.671, represented by flowMeans, FlowSOM, and X-shift, with FLOCK, ClusterX, and PhenoGraph close behind. FlowSOM scored the best on three of the four

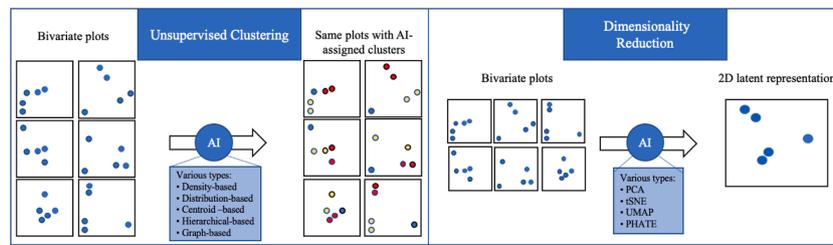


Fig. 2. – Unsupervised approaches for MFC analysis. **Left:** Unsupervised Clustering. An ungated MFC sample is processed by the AI, whereby clusters are assigned to each event (cell). **Right:** Dimensionality reduction. An ungated MFC sample is reduced to a 2-dimensional latent representation allowing for easier visualization. Dimensionality reduction can also be applied to MFC samples with assigned cell identities.

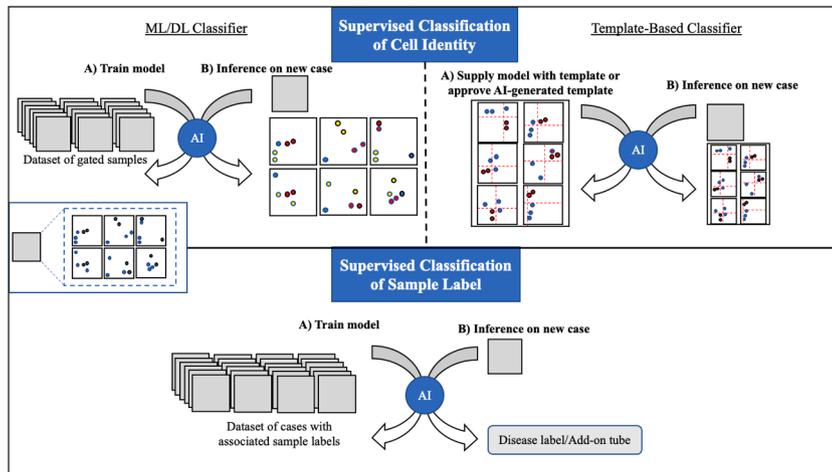


Fig. 3. – Supervised classification approaches for MFC analysis. **Top:** Supervised classifiers for inferring cell identity. ML/DL Classifiers use large datasets of MFC samples to predict cell identities in new cases. Template-based classifiers use a template generated from one or a few samples to replicate the gating scheme on new cases. **Bottom:** Supervised classifiers for sample-level labels. Models have been trained to predict sample-level labels, like the disease associated with an MFC sample or if an add-on tube is recommended. MFC – multi-color flow cytometry, ML - machine learning, DL – deep learning.

datasets and possessed one of the shortest runtimes. The same study also compared the algorithms in their ability to identify a single rare sub-population on two separate MFC datasets, where the population of interest represented less than 400 cells with 0.8% or 0.03% abundance. X-shift performed the best in both cases with an F-1 score of 0.531 and 0.802. As demonstrated in these comparison studies, there is not a single algorithm that performs best across all samples, and performance of the best algorithms can vary widely from sample to sample.

It can be helpful to consider runtime performance when evaluating the usability of clustering algorithms in clinical workflows. Some clustering hundreds of thousands of cells can drive some top-performing algorithms’ runtime into many hours just for a single sample. In a comparison with large samples, model-based approaches like immunoClust, flowClust, and SWIFT and density-based approaches like X-shift, ClusterX, and DensVM had the longest runtimes of one-to-many hours. In the same trial FlowSOM, FLOCK, and PhenoGraph had the fastest runtimes with competitive performance. FlowGrid is a recently developed clustering algorithm that provides the benefits of a density-based algorithm with the scalability of a grid-based method, that can be used to cluster tens of millions of cells in seconds with accuracy comparable to top methods like FlowSOM, FlowPeaks, and FLOCK⁴³. Faster algorithms with similar performance are preferable in a clinical setting since running them ties up machine resources, potentially hindering flow cytometry analysis software which is also computationally demanding. Parallelization is unlikely without involvement of a dedicated compute server so an algorithm that takes several hours or more to produce an output is unlikely to keep pace with the daily demand of cases. Even if an algorithm is run in the background, having technicians and pathologists return to their initial results several hours later is a hindrance to efficiency and unlikely to be adopted in routine clinical workflow.

Dimensionality reduction

Given the high-dimensional nature of MFC, dimensionality reduction techniques that provide summarized, human-interpretable representations have long been explored. Dimensionality reduction algorithms produce a compressed representation of the original data in an unsupervised fashion (Fig. 2 right). They do not assign cluster/community identities but are frequently used to visualize them, and often naturally partition cell populations into visual clusters. In a standard clinical workflow, a collection of 2D scatter plots is used to visualize MFC data. In each plot, two of the available markers are chosen and placed against one another on the two axes of the figure. However, pairwise analysis becomes incredibly cumbersome with the growing number of markers offered by novel acquisition approaches, with modern flow cytometers being able to support more than 40 fluorescent markers in a single sample and mass cytometers often acquiring 32 or more measurements per cell. Dimensionality reduction compresses high-dimensional data by locating a low-dimensional representation, often just two features, that keeps as much of the high-dimensional input’s structure as feasible. In this way, dimensionality reduction algorithms are proving indispensable for visualizing single or multi-sample high-dimensional MFC data, as they confer the underlying structure and interrelationships of the cell populations in a single plot that is readily interpretable by human operators (Fig. 1).

PCA

Principal component analysis (PCA) reduces the dimensions of a flow cytometry sample usually into principal components that reflect the most prominent sources of co-variation of marker expression each cell. The top two principal components are usually plotted to produce a visualization of every cell. PCA components are also useful as inputs for ML algorithms to discriminate different cell and sample types. Methods

such as Automated Population Separator (APC)⁶⁸, The Flow cytometric Orthogonal Orientation for Diagnosis (FLOOD)⁶⁹, and Discriminant Analysis of MultiAspect CYTometry (DAMACY)⁷⁰ all use PCA followed by additional analysis algorithms for sample-level classification. APC (Affinity Propagation Clustering) was originally used to classify gated CD19⁺ neoplastic B-cells in patients with one of three different mature B-cell lymphoproliferative disorders⁶⁸. FLOOD was originally tested on its ability to automatically detect samples that had been exposed to lipopolysaccharide (LPS)⁶⁹. DAMACY was used on several datasets to classify AML, asthma, and exposure to LPS⁷⁰.

t-SNE

Distributed stochastic neighbor embedding has recently emerged as the premier dimensionality reduction strategy for visualizing and subsequently (via a different algorithm) clustering cell populations. While PCA has been most developed for MFC sample-level classification, the rigid linear transformation it imposes typically does not map well to identifying cell populations. t-SNE⁵¹ is a fundamental component of numerous cytometry analysis systems including Cytobank⁷¹, Cytosplore⁴⁹, and cytofit⁴⁴. tSNE is a nonlinear dimensionality reduction technique developed to preserve local neighborhoods, rather than relative distances.

The procedure of t-SNE begins by measuring local neighborhoods in the high-dimensional space and connected by a minimum quantity of distance, e.g., Euclidean space. t-SNE aims to preserve local neighborhoods by augmenting these neighborhoods. In the second step, t-SNE optimizes the point placement in the low-dimensional space, such that the resulting mapping groups neighbors of the high-dimensional space into neighborhoods in the low-dimensional space. The resulting plot will group similar cells together into visual clusters (not assigned, just apparent) of similar density while separate clusters indicate different cell types. However, the optimization only preserves relative distances within these clusters, while the distances between islands are meaningless. This effect can be softened⁷², but this hampers the ability to resolve fine-grained structures and comes at a large computational cost. Wattenberg et al. provide a general understanding of the significance of the various parameters⁷³. Belkina et al. have particularly investigated and modified the parameters in FCM for massive data⁷⁴ to gain additional knowledge about the different parameters. The algorithm's performance is constrained by its computational capacity and several methods have aimed to accelerate it^{52,54,75,76}. All these techniques can also be combined with automated optimal parameter estimation⁷⁴. t-SNE embeddings can now be computed for millions of data points thanks to these optimizations. Even with a t-SNE approach that can handle millions of data points, fine-grained structures will often be obscured due to their small visual space. Hierarchical SNE⁷⁷ builds a hierarchy on the data and allows interactive exploration through a divide and conquer procedure⁵³.

UMAP

UMAP⁵⁸ has recently been evaluated for cytometry data analysis⁵⁷ to generate similar visualizations as t-SNE. UMAP seeks to replicate t-SNE's success, but it also measures global distances and provides a significant performance boost by skipping normalization of data on both high- and low-dimensional representations. Unlike t-SNE, 2D plots generated by UMAP are continuous in nature allowing better inference of cell lineages. UMAP has been applied in ML pipelines for classification of B-cell neoplasms⁴.

PHATE

Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE) was developed specifically for biological datasets to overcome limitations of PCA, t-SNE, and UMAP related to sensitivity to noise, scalability to large datasets, and interpretability in 2-dimensional plots. PHATE dimensionality reduction plots provide a denoised visualization that is insensitive to user configurations and preserves and emphasizes

global and local structure including transitions and clusters. In a head-to-head comparison of dimensionality reduction methods using simulated scRNAseq data, PHATE appears to capture the true structure of high-dimensional data best⁵⁹. As tracking cell trajectories is typically not a priority with clinical flow cytometry, the role of PHATE in analysis pipelines is not clear. However, PHATE was recently used in immune cell profiling identifying multimodal signatures of COVID-19 including flow cytometry⁶⁰. Given the scalability and unique topology PHATE provides, it is likely to appear in more open-source MFC tools going forward.

Dimensionality reduction enables human-interpretable insights into high-dimensional data via a 2D or 3D plot and is often combined with clustering algorithms and supplementary information for deeper insights into biological phenomena. viSNE⁵¹ plots show each data point as a color, and multiple plots with different markers overlaid can be used to interpret the biological significance of each cell and manually cluster. Latent representations generated by t-SNE relate to spectral clustering algorithm in that t-SNE embeddings can be produced using automatic clustering like with ACCENSE⁵⁰ or Cytosplore⁴⁹ where the resulting clusters may be inspected through heatmaps.

Supervised classification of cell identity

Supervised algorithms incorporate user input in the form of labels. Often for MFC projects, cell identities are assigned by manual gating. In a simple scenario with MFC data, a supervised classifier is trained on a part of the dataset using the cell identity labels given produced by manual gating (Fig. 2 Top). Once trained, the classifier is then able to infer labels on similar (unlabeled) input data. Thus, supervised approaches are well-suited to replicating gating strategies and identifying target cell populations of interest. Some approaches, like OpenCyto⁶⁵, flowDensity⁶², or flowLearn⁶³, use an explicit, template-based approach to mimic the manual gating process (Fig. 2 Top). Template-based models offer superior customization and interpretability over ML/DL supervised classifiers but at the cost of being constrained to a single, explicit procedure for reaching the desired output. ML models, especially DL models, can account for sources of bias and confounding variability, provided the dataset is carefully crafted, at a scale that would be difficult to replicate with a template-based approach. For example, in a study using a 3-layer deep neural network where labels for every event were provided, the uncompensated MFC data yielded accurate cell population identities and little to no pre-processing was required⁷⁸.

The classic tradeoff has been that DL models are difficult to interpret, but innovative approaches like an ensemble of CNNs⁷⁹ create interpretable architectures while simultaneously leveraging the abstracting strength of DL. In addition, information can be gained from probing trained ML/DL classifiers via representation learning. For example, researchers used the learned representation of CellCNN, a supervised convolutional DL model, to identify rare cell subsets associated with the disease⁶¹. AI is just reaching the tipping point where it can be applied to automate workflows in clinical flow cytometry, provided there is a large enough dataset with highly standardized marker panels and procedures. For example, reports with numeric and phenotypic alterations as well as sample quality and relevant comments and conclusions can be generated automatically for clinical routines using InfinicytTM and EuroFlowTM Databases.

Examples of cell discovery and immune profiling using AI

In the past several years, there have been many exploratory cell subpopulation analyses related to hematological diseases. Baumgaertner et al. used MegaClust to identify CD4+HLA-DR+ and NKT-like subsets from peripheral blood samples of patients with prostate cancer undergoing radiation therapy³⁷. PhenoGraph was used with v-SNE in a pipeline to identify neoplastic T-cell populations from routine clinical

MFC⁴⁸. Lownik et al. used PhenoGraph followed by UMAP to train a random forest model to predict novel sample cell population cluster labels and UMAP-embedded coordinates⁸⁰. Cytofast, an R package via Bioconductor, was used to identify macrophage subsets that significantly decrease upon cancer immunotherapy and distinct prime-boost effects of prophylactic vaccines on the myeloid compartment⁶⁴. The Cytofast package uses clusters from either Cytosplore, using SPADE followed by A-tSNE⁵², or FlowSOM.

In a recent study, TerraFlow, a commercial data analysis tool for immune profiling, was used to characterize changes in the systemic T-cell compartment between healthy and donors with classical Hodgkin lymphoma (cHL) pre-treatment and between pre-treatment cHL and post-treatment cHL⁶⁶. Their results suggest the cHL systemic T-cell compartment shifts toward an exhausted profile and away from less differentiated cells with the potential for self renewal, as well as a shift from T-1 and T-2 helper type toward T-17 helper cells with diminished T-cell effector functions.

Supervised sample classification

An emerging trend is to use a ML supervised approach for classification of MFC data at the sample level (Fig. 2 Bottom). Rather than producing labels for the individual cells, these classifiers are trained on sample-level labels such as the presence of disease or whether an add-on tube is required. Sample-level models have shown robust ability to classify a variety of hematological cancers.

B-cell neoplasms

Ng and Zuromski used UMAP embeddings for over 2,500 B-cell neoplasms and normal cases to train a random forest classifier⁴. The model could be adjusted to 100% sensitivity, albeit at 14% specificity, to allow for 11% of the cases to be autoverified without human intervention. Gaidano et al. used a dataset of 1465 cases of B-cell non-Hodgkin lymphomas to train a decision tree that learned insightful rules for aiding in diagnosis⁸¹. Zhao et al. used SOM-transformed data from 18, 274 cases of a variety of mature B-cell neoplasm to train a convolutional neural network (CNN), a specific type of DL model, and tested it on a test set of 2,348 cases achieving an F1 score of 0.94².

Minimum residual disease and acute leukemia

Ko et al. used 4039 MFC cases from patients without diseased cells and patients with acute myeloid leukemia (AML) or myelodysplastic syndrome (MDS) to fit a gaussian mixture model and produce Fisher-vectors that were then used to train a support vector machine (SVM)⁸². To test the model's ability to monitor for MRD, the trained SVM predicted labels for 287 MFC samples from post-induction patients and tasked with classifying them as normal or diseased, on which the accuracy ranged from 84.6% – 92.4%. A similar approach was used in a study by Monaghan et al. to classify between acute leukemias and nonneoplastic cytopenias⁸³, in which 531 patients with either acute myeloid leukemia, acute promyelocytic leukemia, acute lymphoblastic leukemia, or nonneoplastic cytopenias were processed in a similar technical manner. Rajwa et al. used ASPIRE to predict disease progression of AML patients following induction therapy by mapping cell populations pre- and post-induction⁸⁴.

Hodgkin lymphoma

Simonson et al. used 1,222 MFC cases to train and test a model composed of ensemble of CNNs (convolutional neural networks), where each CNN mapped to a single 2D histogram, that fed their intermediate predictions to a random forest classifier model for classifying Hodgkin lymphoma⁷⁹. The model achieved an area under the receiver operator characteristic (AUC) of 0.93 with tight confidence intervals from five-fold cross validation and achieving sensitivity of 0.80 at a specificity of about 0.90. The model was readily interpretable, thanks to each of its CNN components mapping to a single 2D histogram, from which

Shapley additive explanation values could explain the most impactful markers for prediction.

Using their automated computing platform, TerraFlow, Freeman et al. generated phenotypes capturing the major differences between 25 healthy and 44 newly diagnosed cHL donors. They trained a linear regression on intensity values of the phenotype markers to classify whether the MFC sample belonged to a cHL or healthy patient and showed an AUC of 0.93 on the test set (1.0 sensitivity at about 0.67 specificity). Samples from the same cHL patients 3 months post-treatment were processed in a similar manner and the regression model achieved an AUC of 0.79. Thus, TerraFlow engineers the marker combinations (phenotypes) with the most predictive features between classes enabling a simple linear regression model to achieve promising performance .

Companies using machine learning for clinical cytometry

As we continue to delve deeper into the nuances of the immune system, increasing computational power, dataset sizes, and advanced tuned reagents and instruments can help with disease insights, precision medicine, drug discovery, and disease monitoring. Below are four biotechnology startups that are driving forward the capabilities in clinical cytometry research studies and routine immune profiling and monitoring.

Teiko.bio

Teiko.bio offers custom panels and computational methods to perform immune profiling via mass cytometry to provide insights into target identification and patient response across cancer^{85–87}, autoimmune diseases^{88–90}, and infectious diseases^{91–93}.

Cytek

Cytek is the first company to offer flow cytometry for simultaneous measurement of over 40 fluorescent markers⁹⁴ with current clinical applications in immune profiling and multi-site standardization.

Ozette

Ozette is a computational immune profiling company that has “created an interpretable machine-learning method that discovers and annotates cell populations, leveraging cloud computing to massively serialize the analysis. The result is an automated single-cell analysis platform with unprecedented speed, dimensionality, and annotation depth.”

TerraFlow bioinformatics

TerraFlow provides an immunophenotyping platform built expressly for clinicians and researchers to identify novel biomarkers and disease-associated cell types. The TerraFlow platform returns explicit gating strategies for each disease-associated cell subpopulation identified by its core algorithm, enabling operator-independent reproducibility across samples and runs.

Discussion

For MFC and mass cytometry, the number of supervised approaches has historically lagged unsupervised approaches but is beginning to see an uptick especially at the sample level. Cheung et al. provided a detailed account in 2020 of supervised and unsupervised algorithms for MFC data analysis, including if the algorithms were implemented inside of a GUI, and whether they were accessible for free or through a paid platform⁹⁵. While unsupervised approaches can be conceptualized and immediately coded and tested on data, supervised approaches require a significant upfront investment in the form of the dataset. As a rule of

thumb, supervised approaches perform better with more data. The curse of dimensionality states that as the feature space increases in size, the amount of data needed increases exponentially⁵. Given the high-dimensional nature of flow cytometry, typically hundreds or thousands of gated or labeled samples are needed for a reasonable level of performance, as illustrated by the studies detailed above in 2.7. This requires a large upfront investment of time by the pathologist(s) to inspect and gate each sample. However, supervised approaches are gaining traction as they tend to be more accurate once a large enough, high-quality dataset is amassed. While conceptualizing supervised versus unsupervised can be useful for understanding what approach is best, it is the combination of approaches that is the most powerful – unlocking biological insights and creating automated cell classifiers.

Cytometry-inspired AI is just one aspect of how big data, ML, automation, and clinical decision support promise to revolutionize the field of hematopathology⁹⁶. Clinical diagnosis considers not just immunophenotyping but also morphology, cytogenetics, molecular genetics, and a growing number of multi-omics approaches. Advances in AI promise to automate cell identification, uncover deeper insights into meaningful immune cell subpopulations and serve as clinical decision support systems.

Appendix

Especially helpful review articles:

- Cossarizza et al. 2019⁹⁷
- Rybakowska et al. 2020⁹⁸

References

- 1 Angeletti C. A method for the interpretation of flow cytometry data using genetic algorithms. *J Pathol Inform.* 2018;9:16.
- 2 Zhao M, et al. Hematologist-level classification of mature B-Cell neoplasm using deep learning on multiparameter flow cytometry data. *Cytometry A.* 2020;97:1073–1080.
- 3 Aghaeepour N, et al. Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods.* 2013;10:228–238.
- 4 Ng DP, Zurowski LM. Augmented human intelligence and automated diagnosis in flow cytometry for hematologic malignancies. *Am J Clin Pathol.* 2021;155:597–605.
- 5 Hastie, T., Tibshirani, R. & Friedman, J. The elements of statistical learning – data mining, inference, and prediction.
- 6 Goodfellow I, Bengio Y, Courville A. *Deep learning.* MIT Press; 2016.
- 7 O'Neill K, Aghaeepour N, Spidlen J, Brinkman R. Flow cytometry bioinformatics. *PLoS Comput Biol.* 2013;9, e1003365.
- 8 Liang, S. & Srikant, R. Why deep neural networks for function approximation? Preprint at <http://arxiv.org/abs/1610.04161> (2017).
- 9 Kriegeskorte N, Golan T. Neural network models and deep learning. *Curr Biol.* 2019; 29:R231–R236.
- 10 Camp J, et al. *Deep neural network for cell type differentiation in myelodysplastic syndrome diagnosis performs similarly when trained on compensated or uncompensated data.* In: *Medical Imaging 2022: Digital and Computational Pathology.* 12039. SPIE; 2022:205–214.
- 11 Transformation of multicolour flow cytometry data with OTflow prevents misleading multivariate analysis results and incorrect immunological conclusions - Folcarelli - 2022 - Cytometry Part A - Wiley Online Library. <https://onlinelibrary.wiley.com/doi/full/10.1002/cyto.a.24491>.
- 12 Ellis, B. et al. flowCore: flowCore: Basic structures for flow cytometry data. (2022) doi:10.18129/B9.bioc.flowCore.
- 13 Gopalakrishnan N, et al. Package 'flowUtils'. *Cytom Part J Int Soc Anal Cytol.* 2008; 73:1151–1157.
- 14 Hahne F, et al. Per-channel basis normalization methods for flow cytometry data. *Cytometry A.* 2010;77A:121–131.
- 15 Monaco G, et al. flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics.* 2016;32:2473–2480.
- 16 Fletez-Brant K, Spidlen J, Brinkman RR, Roederer M, Chattopadhyay PK. flowClean: Automated identification and removal of fluorescence anomalies in flow cytometry data. *Cytometry A.* 2016;89:461–471.
- 17 Sarkar D, Le Meur N, Gentleman R. Using flowViz to visualize flow cytometry data. *Bioinformatics.* 2008;24:878–879.
- 18 Jiang, M. ggcyto: Visualize Cytometry data with ggplot. (2022) doi:10.18129/B9.bioc.ggcyto.
- 19 Wang S, Brinkman RR. Data-driven flow cytometry analysis. *Methods Mol Biol Clifton NJ.* 2019;1989:245–265.
- 20 Montante S, Brinkman RR. Flow cytometry data analysis: Recent tools and algorithms. *Int J Lab Hematol.* 2019;41:56–62.
- 21 Murphy RF. Automated identification of subpopulations in flow cytometric list mode data using cluster analysis. *Cytometry.* 1985;6:302–309.
- 22 Lo K, Hahne F, Brinkman RR, Gottardo R. flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics.* 2009;10:1–8.
- 23 Finak G, Bashashati A, Brinkman R, Gottardo R. Merging mixture components for cell population identification in flow cytometry. *Adv Bioinforma.* 2009;2009:1–12.
- 24 Pyne S, et al. Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci.* 2009;106:8519–8524.
- 25 Sörensen T, Baumgart S, Durek P, Grützkau A, Häupl T. immunoClust—An automated analysis pipeline for the identification of immunophenotypic signatures in high-dimensional cytometric datasets. *Cytometry A.* 2015;87:603–615.
- 26 Dundar M, Akova F, Yerebakan HZ, Rajwa B. A non-parametric Bayesian model for joint cell clustering and cluster matching: identification of anomalous sample phenotypes with random effects. *BMC Bioinformatics.* 2014;15:1–15.
- 27 Naim I, et al. SWIFT—scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 1: Algorithm design. *Cytometry A.* 2014;85:408–421.
- 28 Mosmann TR, et al. SWIFT—scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 2: Biological evaluation. *Cytometry A.* 2014;85:422–433.
- 29 Johnsson K, Wallin J, Fontes M. BayesFlow: latent modeling of flow cytometry cell populations. *BMC Bioinformatics.* 2016;17:1–16.
- 30 Chen X, et al. Automated flow cytometric analysis across large numbers of samples and cell types. *Clin Immunol.* 2015;157:249–260.
- 31 Aghaeepour, N. flowMeans: Non-parametric flow cytometry data gating. 7.
- 32 Van Gassen S, et al. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data: FlowSOM. *Cytometry A.* 2015;87:636–645.
- 33 Quintelier K, et al. Analyzing high-dimensional cytometry data using FlowSOM. *Nat Protoc.* 2021;16:3775–3801.
- 34 Linderman, M. & Bruggner, R. Rclusterapp: Linkable C++ clustering. (2013).
- 35 Qiu P, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol.* 2011;29:886–891.
- 36 Pagni M, et al. Density-based hierarchical clustering of pyro-sequences on a large scale—the case of fungal ITS1. *Bioinformatics.* 2013;29:1268–1274.
- 37 Baumgaertner P, et al. Unsupervised Analysis of Flow Cytometry Data in a Clinical Setting Captures Cell Diversity and Allows Population Discovery. *Front Immunol.* 2021;12.
- 38 Zare H, Shoostari P, Gupta A, Brinkman RR. Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics.* 2010;11:403.
- 39 Levine JH, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell.* 2015;162:184–197.
- 40 Qian Y, et al. Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. *Cytometry B Clin Cytom.* 2010;78B:S69–S82.
- 41 Ge Y, Sealton SC. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics.* 2012;28:2052–2058.
- 42 Samusik N, Good Z, Spitzer MH, Davis KL, Nolan GP. Automated mapping of phenotype space with single-cell data. *Nat Methods.* 2016;13:493–496.
- 43 Ye X, Ho JWK. Ultrafast clustering of single-cell flow cytometry data using FlowGrid. *BMC Syst Biol.* 2019;13:35.
- 44 Chen H, et al. Cytofit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline. *PLoS Comput Biol.* 2016;12, e1005112.
- 45 Becher B, et al. High-dimensional analysis of the murine myeloid cell system. *Nat Immunol.* 2014;15:1181–1189.
- 46 Meskas J, Wang S. Precise and accurate automated removal of outlier events and flagging of files based on time versus fluorescence analysis. *GitHub Repos* <https://github.com/Jmeskas/FlowCut>. 2019.
- 47 Azad A, Rajwa B, Pothan A. flowVS: channel-specific variance stabilization in flow cytometry. *BMC Bioinformatics.* 2016;17:291.
- 48 DiGiuseppe JA, Cardinali JL, Rezuke WN, Pe'er D. PhenoGraph and viSNE facilitate the identification of abnormal T-cell populations in routine clinical flow cytometric data. *Cytometry B Clin Cytom.* 2018;94:744–757.
- 49 Höllt T, et al. Cytosplore: Interactive Immune Cell Phenotyping for Large Single-Cell Datasets. *Comput Graph Forum.* 2016;35:171–180.
- 50 Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE) | PNAS. <https://www.pnas.org.foyer.swmed.edu/doi/full/10.1073/pnas.1321405111>.
- 51 Amir ED, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol.* 2013;31:545–552.
- 52 Pezzotti N, et al. Approximated and User Steerable tSNE for Progressive Visual Analytics. *IEEE Trans Vis Comput Graph.* 2017;23:1739–1752.
- 53 van Unen V, et al. Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nat Commun.* 2017;8:1740.
- 54 Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Efficient algorithms for t-distributed stochastic neighborhood embedding. *Nat Methods.* 2019;16: 243–245.
- 55 Chan DM, Rao R, Huang F, Canny JF. GPU accelerated t-distributed stochastic neighbor embedding. *J Parallel Distrib Comput.* 2019;131:1–13.
- 56 Belkina AC, et al. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat Commun.* 2019;10:5415.
- 57 McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Prepr. ArXiv180203426.* 2018.
- 58 Becht E, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol.* 2019;37:38–44.
- 59 Moon, K. R. et al. Visualizing structure and transitions for biological data exploration. 92.

- 60 Kuchroo M, et al. Multiscale PHATE identifies multimodal signatures of COVID-19. *Nat Biotechnol.* 2022;40:681–691.
- 61 Arvaniti E, Claassen M. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nat Commun.* 2017;8:14825.
- 62 Malek M, et al. flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinforma Oxf Engl.* 2015;31:606–607.
- 63 flowLearn: fast and precise identification and quality checking of cell populations in flow cytometry | Bioinformatics | Oxford Academic. <https://academic.oup.com/bioinformatics/article/34/13/2245/4860364?login=false>.
- 64 Beyrend G, Stam K, Höllt T, Ossendorp F, Arens R. Cytofast: A workflow for visual and quantitative analysis of flow and mass cytometry data to discover immune signatures and correlations. *Comput Struct Biotechnol J.* 2018;16:435–442.
- 65 Finak G, et al. OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Comput Biol.* 2014;10, e1003806.
- 66 Freeman D, et al. Terraflow, a new high parameter data analysis tool, reveals systemic T-Cell exhaustion and dysfunctional cytokine production in classical hodgkin lymphoma. *Blood.* 2021;138:3516.
- 67 Weber LM, Robinson MD. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry A.* 2016;89:1084–1096.
- 68 Automated pattern-guided principal component analysis vs expert-based immunophenotypic classification of B-cell chronic lymphoproliferative disorders: a step forward in the standardization of clinical immunophenotyping | Leukemia. <https://www.nature-com.foyer.swmed.edu/articles/leu2010160>.
- 69 FLOOD: FLOW cytometric Orthogonal Orientation for Diagnosis - ScienceDirect. <http://www.sciencedirect-com.foyer.swmed.edu/science/article/pii/S0169743915003068>.
- 70 Tinnevelt GH, et al. Novel data analysis method for multicolour flow cytometry links variability of multiple markers on single cells to a clinical phenotype. *Sci Rep.* 2017;7:5471.
- 71 Kotecha N, Krutzik PO, Irish JM. Web-Based Analysis and Publication of Flow Cytometry Experiments. *Curr Protoc Cytom Editor. Board J Paul Robinson Manag Ed. Al.* 2010;010. Unit10.17.
- 72 Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. *Nat Commun.* 2019;10:5416.
- 73 Wattenberg M, Viégas F, Johnson I. How to Use t-SNE Effectively. *Distill.* 2016;1:e2.
- 74 Belkina, A. et al. Automated optimal parameters for t-distributed stochastic neighbor embedding improve visualization and allow analysis of large datasets. [bioRxiv 451690](https://doi.org/10.1101/451690). (2018).
- 75 Van Der Maaten L. Accelerating t-SNE using tree-based algorithms. *J Mach Learn Res.* 2014;15:3221–3245.
- 76 Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods.* 2019;16:243–245.
- 77 Pezzotti N, Höllt T, Lelieveldt B, Eisemann E, Vilanova A. Hierarchical Stochastic Neighbor Embedding. *Comput Graph Forum.* 2016;35:21–30.
- 78 Salama ME, et al. Artificial Intelligence Enhances Diagnostic Flow Cytometry Workflow in the Detection of Minimal Residual Disease of Chronic Lymphocytic Leukemia. *Cancers.* 2022;14:2537.
- 79 Simonson PD, Wu Y, Wu D, Fromm JR, Lee AY. De Novo Identification and Visualization of Important Cell Populations for Classic Hodgkin Lymphoma Using Flow Cytometry and Machine Learning. *Am J Clin Pathol.* 2021;156:1092–1102.
- 80 Lownik JC, Mahov S, Alkan S, Merchant A, Kitahara S. Expanding the use of clustering and dimensionality reduction in high parameter flow cytometry data through machine learning for novel samples. *J Immunol.* 2022;208, 172.03-172.03.
- 81 Gaidano V, et al. A Clinically Applicable Approach to the Classification of B-Cell Non-Hodgkin Lymphomas with Flow Cytometry and Machine Learning. *Cancers.* 2020;12:1684.
- 82 Ko B-S, et al. Clinically validated machine learning algorithm for detecting residual diseases with multicolor flow cytometry analysis in acute myeloid leukemia and myelodysplastic syndrome. *EBioMedicine.* 2018;37:91–100.
- 83 Monaghan SA, et al. A Machine Learning Approach to the Classification of Acute Leukemias and Distinction From Nonneoplastic Cytopenias Using Flow Cytometry Data. *Am J Clin Pathol.* 2021. <https://doi.org/10.1093/ajcp/aqab148>. aqab148.
- 84 Rajwa B, Wallace PK, Griffiths EA, Dundar M. Automated assessment of disease progression in acute myeloid leukemia by probabilistic analysis of flow cytometry data. *IEEE Trans Biomed Eng.* 2017;64:1089–1098.
- 85 Systemic immunity is required for effective cancer immunotherapy. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5312823/>.
- 86 Allen BM, et al. Systemic dysfunction and plasticity of the immune macroenvironment in cancer models. *Nat Med.* 2020;26:1125–1134.
- 87 Hartmann FJ, et al. Comprehensive Immune Monitoring of Clinical Trials to Advance Human Immunotherapy. *Cell Rep.* 2019;28:819–831.e4.
- 88 Rubin SJS, et al. Mass cytometry reveals systemic and local immune signatures that distinguish inflammatory bowel diseases. *Nat Commun.* 2019;10:2686.
- 89 Bader L, et al. Candidate Markers for Stratification and Classification in Rheumatoid Arthritis. *Front Immunol.* 2019;10:1488.
- 90 van der Kroef M, et al. Cytometry by time of flight identifies distinct signatures in patients with systemic sclerosis, systemic lupus erythematosus and Sjögrens syndrome. *Eur J Immunol.* 2020;50:119–129.
- 91 Burnett CE, et al. Mass cytometry reveals a conserved immune trajectory of recovery in hospitalized COVID-19 patients. *Immunity.* 2022;55, 1284-1298.e3.
- 92 Rodriguez L, et al. Systems-Level Immunomonitoring from Acute to Recovery Phase of Severe COVID-19. *Cell Rep Med.* 2020;1, 100078.
- 93 Wang W, et al. High-dimensional immune profiling by mass cytometry revealed immunosuppression and dysfunction of immunity in COVID-19 patients. *Cell Mol Immunol.* 2020;17:650–652.
- 94 Park LM, Lannigan J, Jaimes MC. OMIP-069: Forty-Color Full Spectrum Flow Cytometry Panel for Deep Immunophenotyping of Major Cell Subsets in Human Peripheral Blood. *Cytom Part J Int Soc Anal Cytol.* 2020;97:1044–1051.
- 95 Cheung M, et al. Current trends in flow cytometry automated data analysis software. *Cytometry A.* 2021;99:1007–1021.
- 96 Walter W, et al. How artificial intelligence might disrupt diagnostics in hematology in the near future. *Oncogene.* 2021;40:4271–4280.
- 97 Cossarizza A, et al. Guidelines for the use of flow cytometry and cell sorting in immunological studies (second edition). *Eur J Immunol.* 2019;49:1457–1973.
- 98 Rybakowska P, Alarcón-Riquelme ME, Marañón C. Key steps and methods in the experimental design and data analysis of highly multi-parametric flow and mass cytometry. *Comput Struct Biotechnol J.* 2020;18:874–886.