**Novel Machine Learning Approach to Differential Cell Flow Cytometry Analysis Based on Projection Pursuit**

Mahan Dastgiri1, Javier Cabrera1\*, Yajie Duan1, Davit Sargsyan2,3,4, Abraham Adokwei4,5, Rebecca Mary Peter2,3, PoChung Chou2,3, Ge Cheng1, Chun-Pang Lin1, Jocelyn Sendecki4, Helena Geys4, Kanaka Tatikola4, Ah-Ng Kong2,3

1Department of Statistics, School of Arts and Sciences, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

2Department of Pharmaceutics, Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

3Graduate Program in Pharmaceutical Science, Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

4Janssen Pharmaceuticals, Johnson and Johnson, Spring House, PA, USA, and Beerse, BE

5Epidemiology and Statistics Department, University of Georgia, GA, USA

**\*Correspondence**

Professor Javier Cabrera

Rutgers, the State University of New Jersey

Dept of Statistics, Hill Center Room 471

110 Frelinghuysen Rd, Piscataway, NJ 08854

Phone: +1-848-445-7665

Email: cabrera@stat.rutgers.edu

**Novel Machine Learning Approach to Differential Cell Flow Cytometry Analysis Based on Projection Pursuit**

# **Abstract**

This paper introduces the novel methodology of differential projection pursuit and its applications to the analysis of large datasets. The method was applied to a cell flow cytometry dataset as an alternative approach to analyze this type of data. Multicolor cell flow cytometry is a well-established laboratory technique to identify cell subpopulations by measuring their physical and biochemical characteristics. Differential projection pursuit helps to find regions with maximal differences between two or more treatments or distributions. Data analysis in flow cytometry relies on gating, the process of manually selecting successive subpopulations of cells using two-dimensional plots. Plotting the variables only two at a time could mask the hidden structure present in the data, and manual selection makes the analysis inconsistent and arbitrary. The new methodology could automate flow cytometry analysis by utilizing the combination of projection pursuit, data nuggets, and factor analysis. When applied to flow cytometry data, differential projection pursuit allows researchers to quickly identify differences in cell populations exposed to different experimental conditions. This methodology could create a platform to explore differences in large datasets and improve the cell flow cytometry analysis clarity and reproducibility by considering the data in its true dimensional space and through automation, respectively.

**Keywords:** machine learning; differential projection pursuit; varimax rotation; flow cytometry; data nuggets; big data.

# **1** **Introduction**

Multicolor flow cytometry (FC) is a laboratory technique used in biological disciplines including cancer biology, molecular biology, and immunology that measures individual cell properties such as size, granularity, and molecular composition. It has wide applications in immunophenotyping, in vivo detection of bacterial or viral infections, and cell cycle analysis (McKinnon 2018; Han et al. 2014; Eastman and Guo 2020). To measure specific proteins on the surface or inside a cell, cells are suspended in a liquid and fluorescent chemical compounds called fluorochromes or fluorophores are added to the suspension. As a cell moves through a tube, the fluorochromes attached to the cells get excited by the laser and the fluorescence of each individual cell is measured along with forward (FSC) or side scattered (SSC) light (Figure 1). The FSC and SSC measurements provide information about the cells’ physical properties and are used to separate single live cells from cell clusters and debris during data preprocessing. The data is typically saved in Flow Cytometry Standards (FCS) format as a matrix, with rows representing individual cells and columns representing the markers (Spidlen et al. 2010; "International Society for Advancement in Cytometry").

The data is processed, traditionally using a technique called gating. Specialized tools such as FlowJo ("FlowJo Software" 2023) and FSC Express ("FCS Express" 2023) import FSC files and plot the data, two dimensions at a time. The investigator draws areas of interest, or gates, to manually identify clusters of cells that they are interested in. This process of gating goes on sequentially as the investigator focuses on specific subpopulations of cells (Figure 2). However, the two-dimensional plots used in the current gating method in flow cytometry data analysis could fail to represent the possible complex structure present in the data. This is because there are usually eight or more dimensions in a cell flow cytometry data, but only two are considered at a time. Additionally, drawing manual gates is an arbitrary approach and it is dependent on the investigator’s defined threshold. As a result, since each individual creates an arbitrary boundary, results are non-reproducible.

Projection pursuit (PP) is a method in multivariate data analysis for finding low-dimensional projections that uncover interesting structures in the data. Differential projection pursuit (DPP) is proposed as a modified version of PP that finds projections that maximize the difference between two or more distributions. To implement DPP for a very large dataset, the data was compressed into a reduced set of nuggets using data nuggets method. This research aims to introduce the novel DPP methodology and propose this method as an alternative approach to traditional gating in flow cytometry data. DPP provides a platform to analyze this type of data in its original multi-dimensional space and identify cell subpopulations with significantly different densities using automated methods. DPP is reproducible, multivariate in nature, directly differential and can be used as an automated analysis.

# **2 Materials and Methods**

## ***2.1 Data Source and Experimental Design***

The dataset used in this analysis was obtained from the FlowRepositorywebsite (Spidlen et al. 2012) and contained samples of HIV-exposed but uninfected (HEU) and unexposed (UE) infants ("FLOW Repository"). Blood samples were taken from the infants six months after birth. There are 308 FCS files from 44 participants in the repository. In this study, we aimed to find subpopulations of cells with large differences between the lipopolysaccharide (LPS)-stimulated and unstimulated conditions. Two UE participants were selected for this study. For each participant, one LPS-stimulated and one unstimulated sample were selected. These four files contained readouts for a total of 1,754,745 cells. Each data file contained measurements of forward light scattering (FSC), side light scattering (SSC), and eight protein markers (Table 1).

## ***2.2 Data Compression with Data Nuggets***

Performing calculations of order on large datasets such as the one used in this study would be computationally expensive beyond most users’ capabilities. Therefore, a compression algorithm called data nuggets (Beavers et al. 2024) was applied to the raw data. The data nuggets method represents a dataset with millions of observations as a weighted set of a few thousand nuggets. According to Beavers et al. 2024, the number of nuggets should be 3000-5000, to allow computation to be performed in real time, while also preserving the data structure. Each nugget is represented by its center, weight, and scale parameters. The center and weight of a nugget are the mean and the number of data points inside the nugget, respectively, and the scale is the approximate radius of the nugget. Following partitioning the dataset into nuggets, a linear transformation is applied to make the data as spherical as possible in the multivariate sense. This method preserves the structure of the data much better than random samples and is therefore more suitable for finding true data structures in low-dimensional projections (Beavers et al. 2024).

## ***2.3 Projection Pursuit and Differential Projection Pursuit***

Projection pursuit is a technique that searches multivariate *p*-dimensional data for lower *d*-dimensional projections, revealing the main structure of the data, i.e., clusters, outliers, and any other low-dimensional nonlinear structure (Friedman and Tukey 1974). A PP index is a function to numerically measure features of low-dimensional projections, for which higher values indicate more interesting structures (outliers, clusters, and any other nonlinear structures). Natural Hermite index is a PP index introduced by Cook, Buja, and Cabrera (1993) to find non-linear structures and to detect departures from normality, and is the keystone tool of this paper (Cook, Buja, and Cabrera 1993; Amaratunga and Cabrera 2003; Amaratunga, Cabrera, and Shkedy 2014).

Let be a *p*-dimensional dataset, be a random orthogonal projection matrix that projects into a *d*-dimensional space , and be the projected data in the *d*-dimensional space. Consider density of the projected data and standard multivariate Gaussian density . The natural Hermite index measures the distance between the *d*-dimensional distribution and the *d*-dimensional normal distribution .

(1)

The computational burden of these indices is satisfactory for small to moderate datasets but not attainable for very large datasets. Duan, Cabrera, and Emir (2023) introduced weighted versions of the PP indices computed over data nuggets. They showed that the “most interesting” projections found by the natural Hermite index on large datasets are identical to those found by the weighted version of the index over data nuggets.

This study compared the *d*-dimensional projections of two groups of *p*-dimensional samples where . For this purpose, we proposed the differential natural Hermite index to compare any two *d*-dimensional distributions and . Let . The differential natural Hermite index for two distributions is given by the following formula:

(2)

This index is maximized over all *d*-dimensional projections of the data, where and are estimated by the *d*-dimensional Kernel density estimators of the projected data and . The steps of the DPP methodology are:

1. Process the raw data into data nuggets and spherize the data nuggets to have identity covariance matrix.
2. Apply a projection matrix to the two datasets: and .
3. Calculate density estimators and from the projected data , and their average .
4. Calculate the index for projection using and .
5. Repeat steps (ii) through (iv) to maximize over all projections .

Let be the scale and the weight vector of data nuggets for , respectively, and be the number of data nuggets. Then, the density estimators and in the differential natural Hermite index can be calculated as the following:

(3)

where with a pre-determined minimal scale level and is the weight of the *ith* nugget in population (Duan, Cabrera, and Emir 2023).

This procedure usually finds a local maximum. Thus, it needs to be repeated multiple times with different initial s to obtain local maxima with high PP index values, which correspond to more interesting projections (Cook, Buja, and Cabrera 1993; Duan, Cabrera, and Emir 2023).

For comparison of populations, this method would require the evaluation of integrals. However, in Weigle, Sargsyan, and Cabrera (2023), it was shown that:

(4)

which requires evaluating only *k* integrals. Therefore, the algorithm to find the optimal projection for the difference between two groups can be extended to *k* groups.

## ***2.4 Factor Analysis and Clustering***

Once the optimal projections are obtained, varimax rotation is conducted, and the orthogonal basis is rotated to express the two main axes of the projections as a function of only a few flow cytometry channels. The varimax rotation performed was applied to the inverse of the linear transformation used to spherize the data. This rotation methodology may be useful for factor analysis in general, as it is not uncommon to standardize multivariate data (Amaratunga and Cabrera 2003; Amaratunga, Cabrera, and Shkedy 2014).

In addition to the standard varimax rotation, the support vector machine (SVM) was then performed to predict the blue and red clusters of cells corresponding to the positive and negative difference between the two densities, respectively. Finally, to find the cluster of cells that differ the most between groups, weighted K-means clustering was performed on the optimal projections. To take into account the weights of the data nuggets, using a weighted version of K-means was necessary (Cabrera and McDougall 2002).

# **3 Results**

We tested this method on a cell flow cytometry data set to demonstrate its applications and to propose it as an alternative approach to the current standard cell flow cytometry analysis. First, the data from 4 FCS files (LPS-stimulated and unstimulated for two UE patients) was combined, forming a dataset comprised of 1,754,745 cells, which was then compressed into 3,385 refined data nuggets. The top six 2-dimensional projections were produced and rotated using the varimax procedure. The ability of data nuggets to represent the entire data set was confirmed by plotting the projected data nuggets, estimated densities using data nuggets with contours, and projected raw data in the whole dataset (stimulated and unstimulated combined), respectively, for the first and third projections (Figure 4). All three plots for both projections closely resemble one another, suggesting the effectiveness of data nuggets in preserving the overall structure of the data.

To find the regions with maximum difference between the two groups, the density and difference between the two densities were estimated for the stimulated ( and unstimulated () cells using equation 3 (Figure 3). A positive difference indicates presence of a higher number of stimulated cells and a negative difference indicates the presence of a higher number of unstimulated cells. The analysis revealed several regions where the stimulated cells were more abundant than the unstimulated (visualized in blue and defined as blue region) and regions where the reverse was true (visualized in red and defined as red region). These two regions are presented in the third column of Figure 3, and they are separated, representing the difference we obtained using the DPP index (Figure 3).

In order to separate the regions that have a majority of blue and red nuggets, a predictive weighted SVM model fitted into the 2-d projections was used to define the blue and red regions. For this model, the observations are the projected nugget centers, the weights are the nuggets’ weights, and the response is the nuggets’ class (blue, red, or white). For both stimulated and unstimulated cells, the third column of Table 2 gives the proportion of cells in the predicted blue region out of the total number of cells for each. Considering only the blue region, the percentage and counts of the stimulated and unstimulated cells are reported as well. There are 37% of stimulated cells in the entire dataset. The percentage of stimulated cells in the predicted blue region for the six projections ranges from 50.5% to 54.4%, which is about a 40% increase in the number of stimulated cells compared to the 37% in the whole data (Table 2). The fourth column of Table 2 gives the same measurements for the predicted regions using SVM fitted into the nugget centers. Thus, this method identifies regions of cells enriched upon stimulation which could coincide with unique biological characteristics.

The loadings of flow cytometry channels in projections one and three are shown in Figure 5. In Projection one, channels APC-A, APC-Cy7-A, and PE-Cy7-A strongly contribute to the first factor, while channels FITC-A and PerCP-Cy5-5-A contribute significantly to the second factor. For the third projection, channels Alex 700-A, and APC-A strongly contribute to the first factor, and channels APC-Cy7-A and PE-A significantly contribute to the second factor. These factors were constructed by applying a varimax rotation to the original variables, considering the spherization transformation. The loading graphs in Figure 5 are shown in the original scale of the channels after preprocessing.

For simplicity, only projection one will be discussed here. Figure 6 shows the boxplots of the intensity of fluorescence of the channels in the predicted blue and red regions. The intensities in the blue region are higher for Alex 700-A, APC-Cy7-A, and PE-Cy7-A channels. These channels represent proteins TNF-α, IL-6, and CD14, respectively (Table 1). It is expected to see an increase in the production of these proteins after LPS-stimulation (Koch et al. 2014; Liu et al. 2018; Landmann et al. 1996). The blue region in Figure 6 clearly shows this increase due to the presence of more LPS-stimulated cells compared to unstimulated, demonstrating the ability for this methodology to identify biomarkers that differ between cell treatments.

To further divide the predicted blue region into potentially interesting clusters of cells that demonstrate meaningful characteristics, weighted k-means clustering was applied to the predicted blue region data using the nugget weights. Weighted within-cluster sum of squares was used to determine the optimal number of clusters, and the results are shown in Figure 7. The boxplots in Figure 7(b) are weighted using the nugget weights, and they show the intensity of fluorescence for each cluster. The intensities for cluster 2 are lower than the other clusters for all channels except for Pacific Blue-A, and the intensities for cluster 3 are higher than the other clusters for most channels except for PE-A and PE-Cy7-A. This shows patterns that might be of interest to biologists to find meaningful explanations for these clusters.

Table 3, similar to Table 2, shows the proportion of stimulated and unstimulated cells in each cluster of the predicted blue region out of the total number of stimulated and unstimulated cells, as well as the percentage and counts of the stimulated and unstimulated cells in each cluster. The percentage of stimulated cells in each cluster remains about 40% higher than the percentage in the raw data.

The same analysis was performed for the predicted red region (Figure 8, and Table 4). The results of the analysis, including the boxplots of fluorescence intensity level for each channel and a table with the proportion, percentage, and counts of cells in each cluster, are in the supplement to the paper in Appendix I. Together, these findings demonstrate the application of DPP, and confirm its validity through comparison to independently derived biological results.

# **4 Discussion**

Differential projection pursuit is a modified version of PP that finds projections that maximize the difference between two or more distributions. The gating approach to cell flow cytometry data analysis is limited by computing power and lack of tools that would allow multidimensional data visualization and analysis. Plotting and clustering such data in two dimensions at a time bypasses these limitations. However, this presents only a partial view of the structure of multidimensional data and could miss the main underlying structure. Additionally, results from manual gating are highly dependent on the investigator’s perception and experience and are almost certainly non-reproducible. Automated gating can improve reproducibility, but it still does not address the dimensionality issue. In this work, the approach was to step back from gating and instead examine the data in its true dimensional space. Differential projection pursuit is reproducible, multivariate in nature, directly differential and can be used as an automated analysis.

Applying data nuggets compressed the data by grouping individual cells into small, nearly spherical nuggets of similar radius that reduced the sample size while preserving the structure of the data. This enables us to implement clustering techniques such as K-means, hierarchical, PAM, and model-based clustering for very large datasets. With the current computational facilities available to us, however, this methodology is limited to datasets containing up to 100 million observations in 100 dimensions, approximately.

Differential projection pursuit found optimal projections that revealed the most relevant differential structure in the data over the space of all projections. By comparing projections of samples across experimental conditions, it identified cell subpopulations with significantly different densities in certain regions of space. Some of these subgroups could be identified using the current classification of immune cells based on surface markers and physical characteristics of the cells, while many others could represent new subtypes or reveal previously unknown mechanisms. The latter would require more careful examination and interpretation by biologists, as well as confirmation from other datasets. The main limitation of the differential natural Hermite index is that it would be unable to find some complex structures in data, such as when assessing two groups with a structure of two intertwined spirals in a two-dimensional projection and random values in the orthogonal subspace to .

This methodology has the potential to create a platform for analyzing very large and complex data revealing multiple aspects of the immune system and disease biology, such as cytokine release by stimulated immune cells, different states of signal transduction pathways, and complex cell functions. It can help identify specific subsets of cells in normal and pathological conditions with unique biological functions, considering the enormous heterogeneity of cells.

# **References**

Amaratunga, D., and J. Cabrera. 2009. *Exploration and analysis of DNA microarray and protein array data.* Vol. 605. John Wiley & Sons.

Amaratunga, D., J. Cabrera, and Z. Shkedy. 2014. *Exploration and analysis of DNA microarray and other high-dimensional data*. John Wiley & Sons.

Beavers, T., G. Cheng, Y. Duan, J. Cabrera, M. Lubomirski, D. Amaratunga, and J. E. Tiggler. 2024. Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure. *Journal of Computational and Graphical Statistics. 1-21.*

Cabrera, J., and A. McDougall. 2002. *Statistical consulting*. Springer Science & Business Media.

Cook, D., A. Buja, and J. Cabrera. 1993. Projection pursuit indexes based on orthonormal function expansions. *Journal of Computational and Graphical Statistics* *2*: 225-250.

Duan, Y., J. Cabrera, and B. Emir. 2023. New Projection Pursuit Index for Big Data. *arXiv preprint arXiv:2312.06465*.

Eastman, A. E., and S. Guo. 2020. The palette of techniques for cell cycle analysis. *FEBS letters* *594*: 2084-2098.

FlowRepository. <http://flowrepository.org/id/FR-FCM-ZZZU>.

FlowJo Software. 2023. Becton, Dickinson and Company.

Friedman, J. H., and J. W. Tukey. 1974. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers* *100*: 881-890.

FCS Express. 2023. DeNovo Software by Dotmatics.

Han, Y., S. Wang, Z. Zhang, X. Ma, W. Li, X. Zhang, J. Deng, H. Wei, Z. Li, X. E. Zhang, et al. 2014. In vivo imaging of protein-protein and RNA-protein interactions using novel far-red fluorescence complementation systems. *Nucleic Acids Res* 42: e103.

International Society for Advancement in Cytometry. <https://isac-net.org/>.

Koch, L., D. Frommhold, K. Buschmann, N. Kuss, J. Poeschl, and P. Ruef. 2014. LPS-and LTA-induced expression of IL-6 and TNF-α in neonatal and adult blood: role of MAPKs and NF-κB. *Mediators of inflammation* *2014*.

Landmann, R., H. P. Knopf, S. Link, S. Sansano, R. Schumann, and W. Zimmerli. 1996. Human monocyte CD14 is upregulated by lipopolysaccharide. *Infection and immunity* *64*: 1762-1769.

Liu, X., S. Yin, Y. Chen, Y. Wu, W. Zheng, H. Dong, Y. Bai, Y. Qin, J. Li, S. Feng, et al. 2018. LPS‑induced proinflammatory cytokine expression in human airway epithelial cells and macrophages via NF‑κB, STAT3 or AP‑1 activation. *Molecular medicine reports* *17*: 5484-5491.

McKinnon, K. M. 2018. Flow cytometry: an overview. *Current protocols in immunology* *120*: 5-1.

Spidlen, J., K. Breuer, C. Rosenberg, N. Kotecha, and R. R. Brinkman. 2012. FlowRepository: A resource of annotated flow cytometry datasets associated with peer‐reviewed publications. *Cytometry Part A* *81*: 727-731.

Spidlen, J., W. Moore, D. Parks, M. Goldberg, C. Bray, P. Bierre, P. Gorombey, B. Hyun, M. Hubbard, S. Lange, et al. 2010. Data File Standard for Flow Cytometry, version FCS 3.1. *Cytometry Part A: The Journal of the International Society for Advancement of Cytometry* *77*: 97-100.

Weigle, S., D. Sargsyan, J. Cabrera. 2023. A Hermite index randomization method for pre-clinical studies. (submitted)

# **Appendix I: Results of the Analysis of the Predicted Red Region of Projection One**

Table 4: Proportion of stimulated and unstimulated cells in each cluster of the predicted red region of the first projection. The three numbers are the proportion of cells in each cluster out of the total, as well as percentage and counts of stimulated and unstimulated cells for each cluster.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Condition** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **Cluster 5** | **Cluster 6** | **Cluster 7** |
| Stimulated | 0.0013 (26.8%) 709 | 0.0039 (29.3%) 2100 | 0.0018 (28.9%) 966 | 0.0010 (27.4%) 512 | 0.0014 (20.1%) 774 | 0.0018 (27.6%) 986 | 0.0008 (28.3%) 429 |
| Unstimulated | 0.0021 (73.2%) 1941 | 0.0054 (70.7%) 5056 | 0.0026 (71.1%) 2372 | 0.0015 (72.6%) 1358 | 0.0033 (79.9%) 3086 | 0.0028 (72.4%) 2585 | 0.0012 (71.7%) 1089 |

A diagram of different colored candles

Description automatically generated with medium confidence

Figure 8: (a) plot of projection one with the seven clusters obtained from the predicted red region, and (b) weighted boxplots of the intensity of fluorescence for each cluster across the eight channels.

# **Tables**

Table 1: List of cell flow cytometry channels and corresponding reagents.

|  |  |
| --- | --- |
| **Channel** | **Reagent** |
| FSC-A |  |
| SSC-A |  |
| FITC-A | IFN-α |
| PE-A | CD123 |
| PerCP-Cy5-5-A | MHCII |
| PE-Cy7-A | CD14 |
| APC-A | CD11c |
| APC-Cy7-A | IL-6 |
| Pacific Blue-A | IL-12 |
| Alex 700-A | TNF-α |

Table 2: Proportion of stimulated and unstimulated cells in the predicted blue region. The three numbers in the third and fourth columns are the proportion of cells in the predicted blue region of the total, as well as percentage and counts of stimulated and unstimulated cells in the predicted blue region, for the six projections obtained by projection pursuit.

|  |  |  |  |
| --- | --- | --- | --- |
| **Projection** | **Condition** | **Blue cluster predicted by SVM model fitted to the 2-d projection** | **Blue cluster predicted by SVM model fitted to nugget centers** |
| 1 | Stimulated | 0.020 (53.5%) 10870 | 0.087 (53%) 46624 |
| Unstimulated | 0.010 (46.5%) 9464 | 0.044 (47%) 41328 |
| 2 | Stimulated | 0.014 (51.5%) 7678 | 0.110 (51.9%) 59130 |
| Unstimulated | 0.008 (48.5%) 7238 | 0.059 (48.1%) 54700 |
| 3 | Stimulated | 0.017 (52.9%) 8879 | 0.253(49.2%) 135400 |
| Unstimulated | 0.009 (47.1%) 7921 | 0.15 (50.8%) 139596 |
| 4 | Stimulated | 0.008 (54.4%) 4539 | 0.075 (54.4%) 40324 |
| Unstimulated | 0.004 (45.6%) 3804 | 0.036 (45.6%) 33823 |
| 5 | Stimulated | 0.017 (50.5%) 8850 | 0.196 (49.6%) 104937 |
| Unstimulated | 0.009 (49.5%) 8682 | 0.115 (50.4%) 106823 |
| 6 | Stimulated | 0.032 (51.4%) 17221 | 0.100 (53%) 53714 |
| Unstimulated | 0.017 (48.6%) 16256 | 0.051 (47%) 47675 |

Table 3: Proportion of stimulated and unstimulated cells in each cluster of the predicted blue region of the first projection. The three numbers are the proportion of cells in each cluster out of the total, as well as percentage and counts of stimulated and unstimulated cells for each cluster.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Condition** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **Cluster 5** |
| Stimulated | 0.0038 (55.2%) 2013 | 0.0098 (53.9%) 5259 | 0.0011 (51.1%) 582 | 0.0039 (51.3%) 2098 | 0.0017 (54.0%) 920 |
| Unstimulated | 0.0018 (44.8%) 1634 | 0.0048 (46.1%) 4498 | 0.0006 (48.9%) 558 | 0.0021 (48.7%) 1990 | 0.0008 (46.0%) 784 |

**Figures**

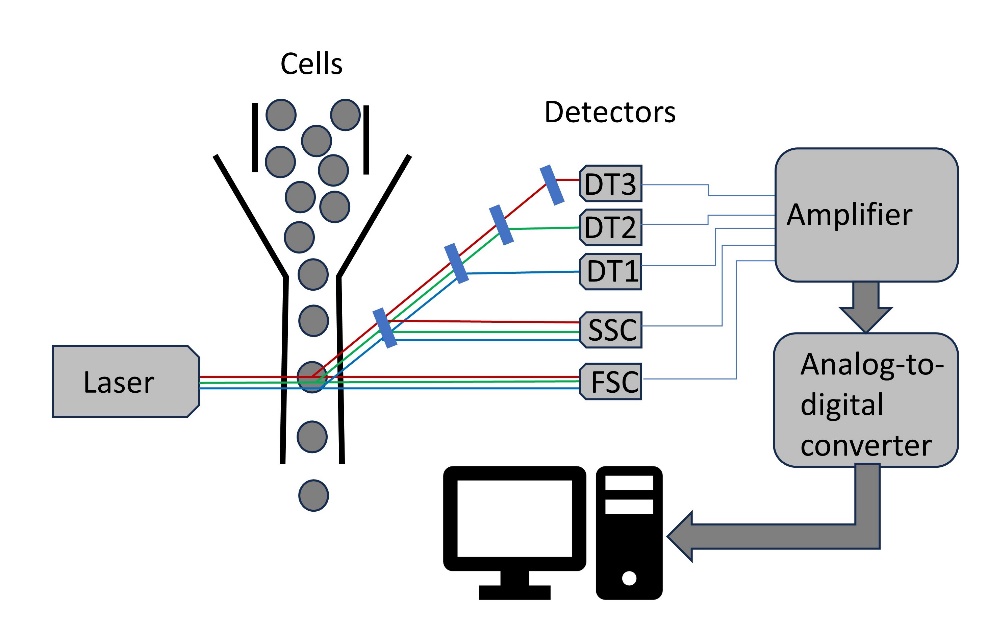


Figure 1: Schematics of a flow cytometer. As the cells pass through a narrow tube, the fluorochromes attached to the cells get excited by the laser. The light is deflected by a series of optical filters into the detectors. The signal from the detectors is amplified, converted from analog to digital, and stored on the computer.

A diagram of a diagram of a fingerprint

Description automatically generated with medium confidence

Figure 2: An example of flow cytometry gating. First, FSC-A and SSC-A are plotted, and the area of interest is manually selected by drawing the gate (red). Next, this subset of cells is plotted in two other flow cytometry channels, PE-A and APC-A. A new gate is drawn and the process repeats using the next two channels. This process continues until all target subpopulations of cells are identified.

A group of squares with different colors

Description automatically generated

Figure 3: Density plot of stimulated (first column) data nuggets, unstimulated (second column) data nuggets, and the difference between them (third column) for six projections that were obtained by optimizing the data nuggets projection pursuit Hermite index. The blue and red regions in the plots of the third column corresponds to the positive and negative difference between the two densities, respectively.

A group of data nuggets

Description automatically generated

Figure 4: Plot of the projected data nuggets for the combined dataset with both stimulated and unstimulated cells (first column), estimated density plot for the data nuggets (second column), and projected raw data plot (third column) for the first (a) and third projections (b).

A comparison of a number of different types of factor plots

Description automatically generated with medium confidence

Figure 5: Loading plots for the first (a) and third (b) projections. In (a), there are three channels (APC-A, APC-Cy7-A, and PE-Cy7-A) that contribute to the first factor, and two channels (FITC-A, and PerCP-Cy5-5-A) that contribute to the second factor. In (b), channels Alex 700-A, and APC-A contribute to factor 1, and APC-Cy7, and PE-A contribute to factor 2.

A diagram of different colored boxes

Description automatically generated

Figure 6: Weighted boxplots of the fluorescence intensity for the eight flow cytometry channels in the predicted blue and red regions of the first projection. The y-axis represents the fluorescence intensity in the normalized scale.

A chart of different colored boxes

Description automatically generated with medium confidence

Figure 7: (a) plot of projection one with the five clusters obtained from the predicted blue region, and (b) weighted boxplots of the intensity of fluorescence for each cluster across the eight channels.

# **Figure Captions**

[Figure 1: Schematics of a flow cytometer. As the cells pass through a narrow tube, the fluorochromes attached to the cells get excited by the laser. The light is deflected by a series of optical filters into the detectors. The signal from the detectors is amplified, converted from analog to digital, and stored on the computer. 20](#_Toc182946940)

[Figure 2: An example of flow cytometry gating. First, FSC-A and SSC-A are plotted, and the area of interest is manually selected by drawing the gate (red). Next, this subset of cells is plotted in two other flow cytometry channels, PE-A and APC-A. A new gate is drawn and the process repeats using the next two channels. This process continues until all target subpopulations of cells are identified. 21](#_Toc182946941)

[Figure 3: Density plot of stimulated (first column) data nuggets, unstimulated (second column) data nuggets, and the difference between them (third column) for six projections that were obtained by optimizing the data nuggets projection pursuit Hermite index. The blue and red regions in the plots of the third column corresponds to the positive and negative difference between the two densities, respectively. 22](#_Toc182946942)

[Figure 4: Plot of the projected data nuggets for the combined dataset with both stimulated and unstimulated cells (first column), estimated density plot for the data nuggets (second column), and projected raw data plot (third column) for the first (a) and third projections (b). 23](#_Toc182946943)

[Figure 5: Loading plots for the first (a) and third (b) projections. In (a), there are three channels (APC-A, APC-Cy7-A, and PE-Cy7-A) that contribute to the first factor, and two channels (FITC-A, and PerCP-Cy5-5-A) that contribute to the second factor. In (b), channels Alex 700-A, and APC-A contribute to factor 1, and APC-Cy7, and PE-A contribute to factor 2. 23](#_Toc182946944)

[Figure 6: Weighted boxplots of the fluorescence intensity for the eight flow cytometry channels in the predicted blue and red regions of the first projection. The y-axis represents the fluorescence intensity in the normalized scale. 24](#_Toc182946945)

[Figure 7: (a) plot of projection one with the five clusters obtained from the predicted blue region, and (b) weighted boxplots of the intensity of fluorescence for each cluster across the eight channels. 24](#_Toc182946946)