Novel machine learning approach to differential flow cytometry analysis base on projection pursuit

Mahan Dastgiri 1, 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 and Javier Cabrera1

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

**SUBMIT TO: Special Issue of the Journal of Biopharmaceutical Statistics (JBS) special issue on AI/ML**

**Correspondence**

Professor Javier Cabrera

Rutgers, the State University of New Jersey

School of Arts and Sciences, Room XXX

XXXXXXXX, Piscataway, NJ 08854

Phone: +1-XXX-XXX-XXXX

Email: [xavier.cabrera@gmail.com](mailto:xavier.cabrera@gmail.com)

Table of Contents

[Abstract 3](#_Toc146822857)

[1 Background 3](#_Toc146822858)

[1.1 Key principals of flow cytometry 3](#_Toc146822859)

[2 Materials and Methods 5](#_Toc146822860)

[2.1 Data Source and Experimental Design 5](#_Toc146822861)

[2.2 Data compression with data nuggets 5](#_Toc146822862)

[2.3 Projection pursuit and differential projection pursuit 5](#_Toc146822863)

[2.4 Factor analysis and clustering 7](#_Toc146822864)

[3 Results 8](#_Toc146822865)

[4. Discussion 9](#_Toc146822866)

[6. Figures and Tables 11](#_Toc146822867)

[6. References 18](#_Toc146822868)

# Abstract

Summary here. Write last.

# 1 Background

## 1.1 Key principals of flow cytometry

Multicolor flow cytometry (FC) is a laboratory technique used in biological disciplines including cancer biology, molecular biology, immunology and infectious disease monitoring that measures individual cell properties such as size, granularity and molecular composition. It has wide applications in immunophenotyping by simultaneously characterizing mixed population of cells from blood, bone marrow and solid tissues based of cluster of differentiation (CD) markers, chemokine receptor markers, cytokines and antigen specific markers [1]. Fluorescent protein analysis revolutionized in vivo detection of bacterial or viral infections, transplanted cells, and gene knockout in cells [2]. Cell cycle analysis is another critical assay to measure dynamics of pathophysiological processes by utilizing DNA binding dyes [3]. Antigen specific response works via fluorescent major histocompatibility (MHC) multimers which carry the desired antigen and bound to tested immune cells recognizing the antigen. Following the binding, the response of the antigen is measured and evaluated.

Sample preparation includes three approaches: (a) transfection and expression of fluorescent proteins, (b) staining with fluorescent dyes, or (c) staining with fluorescently conjugated antibodies. To measure specific proteins on the surface or inside a cell, fluorescent chemical compounds called fluorochromes or fluorophores are added to the suspension. The fluorochromes are attached to molecules with affinity to specific proteins, hence labeling these proteins. Cells are first separated and suspended in a liquid, and the suspension is passed through narrow tubes, one cell at a time. A universal flow cytometer is composed of three systems: fluidics that transports sample to the laser intercept point, optics that induce the visible and fluorescent light signals, and electronics that convert analog light signals to digital [1]. The instruments contain a large number of such tubes for parallel processing. As a cell moves through a tube, it is hit by a beam of light from a lamp or a laser. This excites fluorochromes that in turn emit light in a relatively narrow band of wavelengths. The emitted light passes through a series of optical filters and dichroic mirrors deflecting it onto detectors (Figure 1). Besides measuring light emitted by fluorochromes, flow cytometers also detect light scattered by the cells forward or to the side (FSC and SSC, respectively). 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. Additionally, the instruments are able to measure electrical current impedance, i.e., the opposition to alternative current as the cells travel through the tubes. This allows for calculation of the cell size and additional physical properties. As of 2023, flow cytometers may contain as many as 10 lasers and up to 30 fluorochrome detectors. The data collection process in flow cytometer is called *acquisition*. The data is typically saved in Flow Cytometry Standards (FCS) format as a matrix, with rows representing individual cells and columns representing the markers [4, 5].

Following acquisition, the data is processed, traditionally using a technique called *gating*. Specialized tools such as *FlowJo* [6] and *FSC Express* [7] import FSC files and plot the data, two dimensions at the time. The investigator draws areas of interests, 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. The gating strategy follows current understanding of differentiation process, with major differentiating proteins gated first (Figure 2). Once gating is completed, the software will count the number of cells in each gate and output a processed data file. Often, the focus of the analysis is not only the counts, but the ratios of child-parent populations as defined by the gating strategy, i.e., frequencies.

# 2 Materials and Methods

## 2.1 Data Source and Experimental Design

The dataset used in this analysis was obtained from the *FlowRepository* website [8] and contained samples of HIV-exposed but uninfected (HEU) and unexposed (UE) infants [9]. Blood samples were taken from the infants six months after birth. There were 308 FCS files from 40 participants in the repository. In this study, we aimed to find the subpopulations of the cells with large differences between the [lipopoly](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2196288/)saccharide (LPS)-stimulate and unstimulated conditions. One HEU and one EU participant data 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 the measurements of forward light scattering (FSC), side light scattering (SSC) and eight protein markers (Table 1).

## 2.2 Data compression with data nuggets

Performing calculation of order *n2* on large datasets such as one used in this study would be computationally expensive beyond most users’ capabilities. Therefore, a compression algorithm called *data nugget* [10] 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. This method preserves the structure of the data much better than random samples and therefore is more suitable for finding true data structures in low dimensional projections.

## 2.3 Projection pursuit and differential projection pursuit

*Projection pursuit* (PP) 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. These methods were introduced by Friedman and Tukey for finding structure while exploring a 9-dimensional data from particle physics [11]. Friedman index was introduced later as an example of PP indices [12]. Cook et al made substantial progress in this area by introducing several new PP indices [13]. In particular, they introduced the *natural Hermite index* that became very popular and is the keystone tool of this paper. The natural Hermite index measures the distance between a *d*-dimensional distribution and a *d*-dimensional normal distribution .

Text

Description automatically generated

The computational burden of these indices is satisfactory for small to moderate data sets but not attainable for very large datasets. Duan et al introduced weighted versions of the PP indices computed over data nuggets [14]. 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 *d*-dimensional projections of two groups of *p*-dimensional samples where *d < p*. For this purpose, a differential Hermite index was introduced to compare any two *d*-dimensional distributions and . Let . The differential Hermite index for two distributions is given by the following formula:

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 are:

1. Process the raw data into data nuggets, and spherize data nuggets to have identity covariance matrix;
2. Apply a projection matrix *P* to the two data sets: and ;
3. Calculate density estimators and from the projected data , and their average ;
4. Calculate the index for projection *P* using and ;
5. Repeat step (i) through (iii) to maximize over all projections *P;*

Let be the scale vector of data nuggets for *k* = 1, 2 and *m* be the number of data nuggets. Then, the density estimators and in the differential Hermite index can be calculated as the following:

where with a pre-determined minimal scale level and is the weight of the *ith* nugget in population *k*.

This procedure usually finds a local maximum, so it needs to be repeated multiple times with different initial *P*’s to obtain good local maxima. Usually, obtaining 3 or 4 local maxima projections is sufficient.

For comparison of *k>2* samples, this method would require the evaluation of *k*(*k*-1)/2 integrals. However, in Weigle et al (2023) it was shown that:

=c ,

which requires to evaluate 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. Support vector machine (SVM) is then performed to predict the blue and red clusters of cells corresponding to 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 is performed on the optimal projections.

# 3 Results

After compressing the raw data into 3,385 refined data nuggets, top six 2-dimentional projections were produced and rotated using varimax procedure. The density was estimated for the stimulated ( and unstimulated () cells, and the difference of the two densities (Figure 3). The figure revealed several blue regions where the stimulated cells were more abundant than the unstimulated. Alternatively, the red clusters showed the regions where the unstimulated cells were more abundant than the stimulated.

Table 2 shows the proportion of stimulated and unstimulated cells in the total in the predicted blue region for each projection, as well as the percentage and counts of the stimulated and unstimulated cells for the region. For the rest of the analysis, we focused on projections 1 and 3. Figure 4 shows the projected data nuggets for the first and third projections displayed in Figure 3. The left panels represent the projected data nuggets, while the right panels are the projected raw data using the same projection matrices. The middle panels are the estimated densities using the data nuggets with contours.

Figure 5 shows signal levels of the channels in the predicted blue and red regions. The loadings of flow cytometry channels in projections 1 and 3 are shown in Figure 6. In Projection 1, channels FITC-A and PerCP-Cy5-5-A contributed significantly to the first direction, while channels APC-A, APC-Cy7-A, and PE-Cy7-A strongly correlate with the second direction. For the third projection, channels APC-Cy7-A and PE-A made significant contributions to the first direction, and channels Alex 700-A, APC-A, and PerCP-Cy5-5-A were strongly associated with the second direction. Weighted K-means clustering was applied to the predicted blue regions of projections 1 and 2, and the clustering results are shown in figures 7 and 8, respectively. Within-cluster sum of squares was used as weights to choose the optimal number of clusters. This numbers were 5 and 7 for projections 1 and 3, respectively. The box plots in figures 7 and 8 show the channel signal levels for each cluster. Tables 3 and 4 represent the proportion of stimulated and unstimulated cells in the total number of cells in each cluster of the predicted blue region for projections 1 and 3, as well as the percentage and counts of the stimulated and unstimulated cells.

# 4. Discussion

Gating approach to flow data cytometry is determined in part by biology but also by limitation of computing power and tools that would allow multidimensional data visualization and analysis. Plotting and clustering such data two dimensions at the time overcomes these limitations. However, such projections can present severely distorted images of a multidimensional object, masking important patterns. 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 worked, we stepped back from gating and instead examined the data in its true dimensional space. Applying data nuggets reduced the amount of data by grouping individual cells into typical groups. Projection pursuit found optimal projections that revealed the most information about the data. Finally, by comparing projections of samples across experimental conditions, we identified cell subpopulations with significantly different densities in certain regions of space. Some of these subgroups were identifiable using current classification of immune cells based on surface markers and physical characteristics of the cells while several of the subgroups could represent new subtypes or reveal previously unknown mechanism. The latter would require more careful examination and interpretation by biologists, as well as conformation from other data sets. This methodology has the potential to create a platform for analyzing large and complex data revealing multiple aspects of 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 enormous heterogeneity of cells.

@Davit:

Based on discussion with Maggie, discuss experimental design - Stain Index, color assignment, compensation, Data from different instruments is not comparable as each instrument have its own setting, and lasers need to be calibrated whole the time…

@Javier:

1. The use of data nuggets is to make this work with very large data tables
2. Cons of using 2D for N-D data
3. WH use PP? PP finds best projections and is not attached to individual markers although we try to find projections with axis using minimal number of markers
4. Discuss data: different treatment groups so we are interested in differential analysis
5. Did we find any markers (in the results) that are interesting? Are any of the clusters interpretable? Can we find an example (or synthetic data) where a cluster is masked if you look at simple projections but revealed in PP? E.g., 4D data with 4 clusters along the 4D diagonal, e.g., a cylinder along the 4D diagonal. The clusters will be masked in simple 2D projections but if w project on the diagonal, the difference is visible even in 1D. More complicated – 4D diagonal and some linear combination. Individual 2D scatter plots will look bad; hence, gating would not work.

@ALL: please contribute

# Figures and Tables

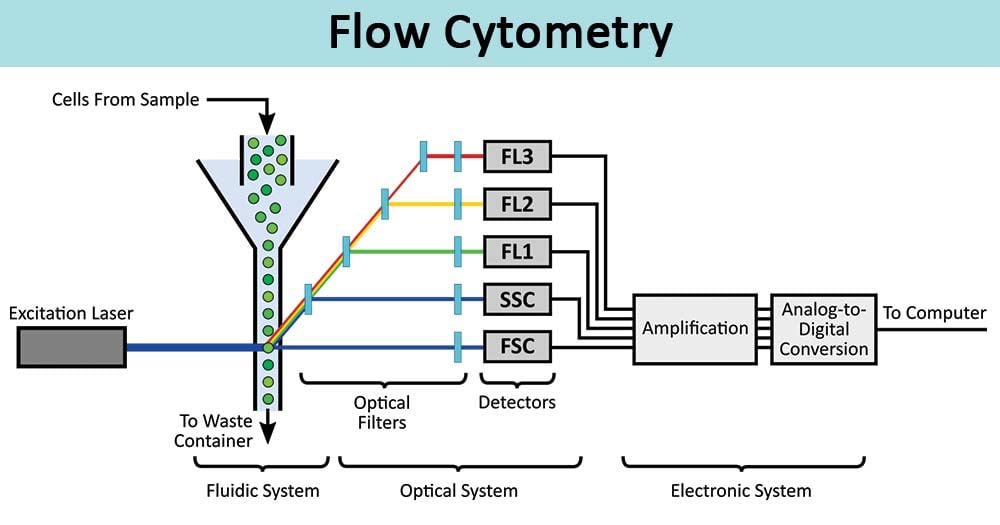


Figure : PLACEHOLDER! David to make a new one

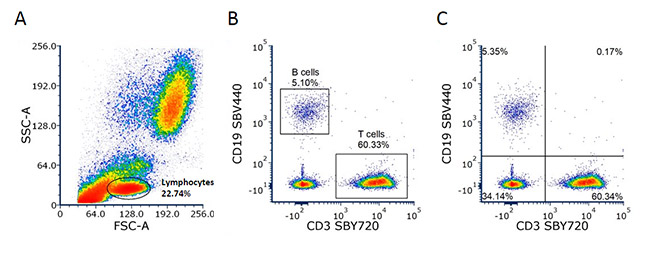


Figure : PLACEHOLDER! Davit to make a new one

A screenshot of a computer generated image

Description automatically generated

Figure : Density plot of Stimulated (first column), Unstimulated (second column) data nuggets, and the difference between them (third column) for 6 projections that were obtained by optimizing the data nuggets projection pursuit Hermite index.

A group of images of data nuggets

Description automatically generated

Figure : Projected data nuggets (first column), estimated density plot for the data nuggets (second column), and projected raw data (third column) for the first and third projections

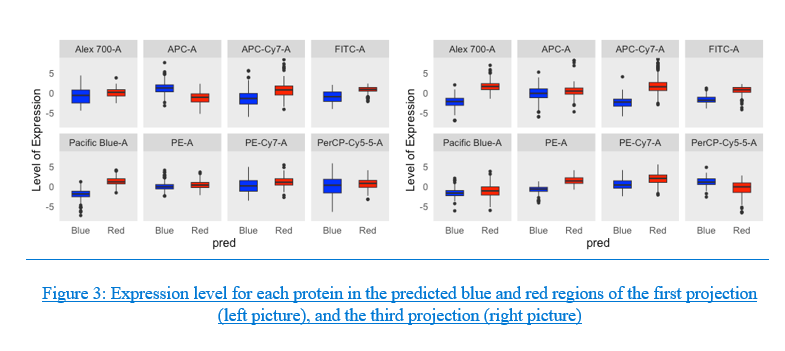


Figure : Flow cytometry channel signal levels in the predicted blue and red regions of the first (A) and the third (B) projections.

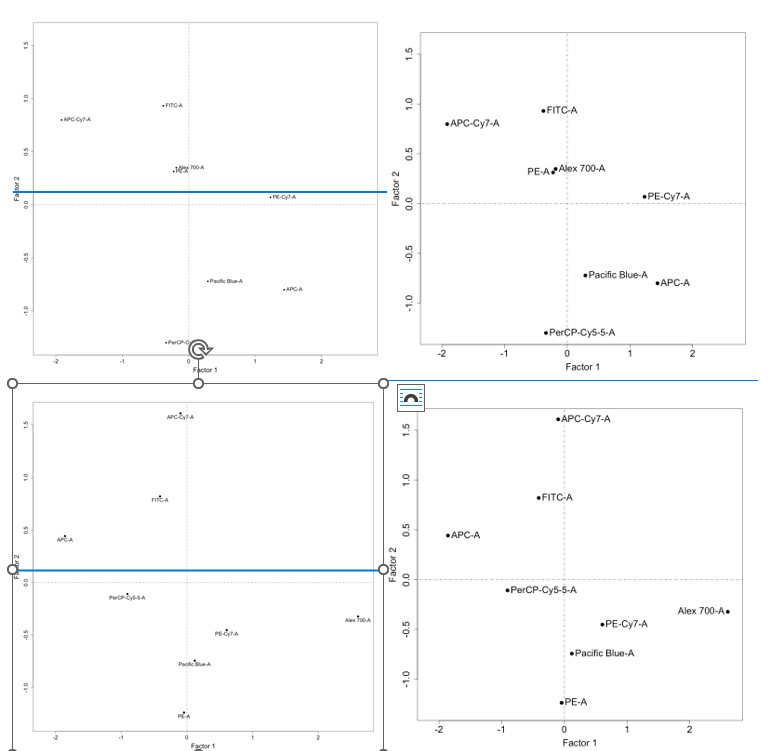


Figure : Loading plots for the first (left) and third (right) projection.

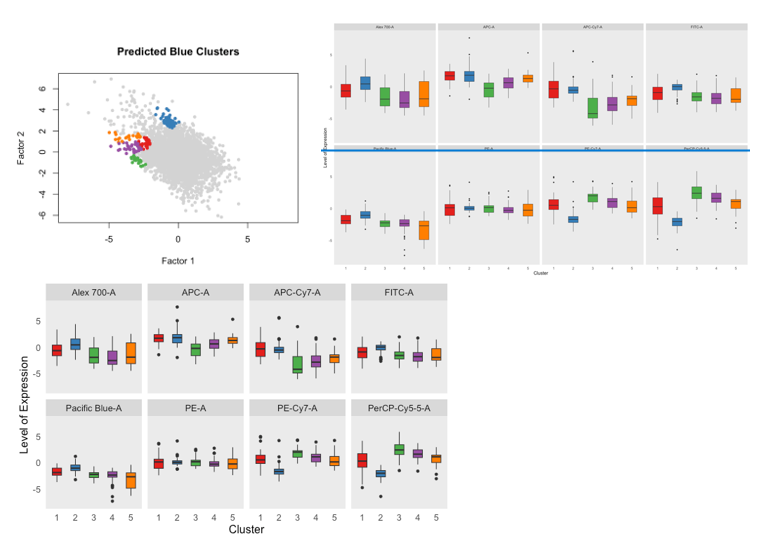


Figure : Five clusters of the predicted blue region for projection one (left picture) and protein expression level for each cluster (right picture)

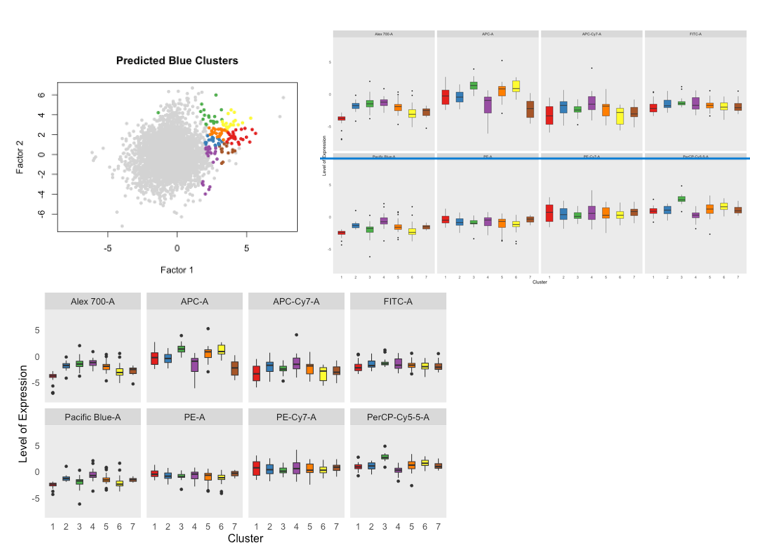


Figure : Five clusters of the predicted blue region for projection three (left picture) and protein expression level for each cluster (right picture)

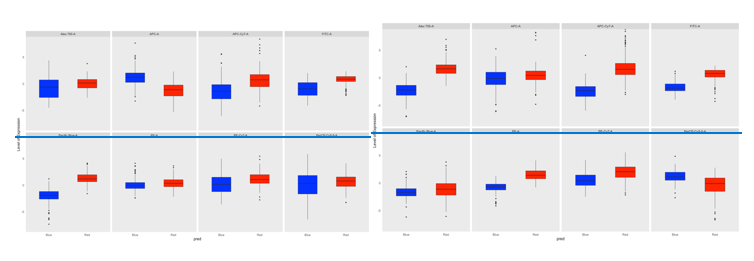


Figure : ???

Table : List of reporter and analytes for the datasets

|  |
| --- |
| Channel Reagent |
| FSC-A  SSC-A  FITC-A IFNa  PE-A CD123  PerCP-Cy5-5-A MHCII  PE-Cy7-A CD14  APC-A CD11c  APC-Cy7-A IL6  Pacific Blue-A IL12  Alex 700-A TNFa |

Table : ...

|  |  |  |  |
| --- | --- | --- | --- |
| Projection |  | Predicted blue cluster based on 2-d projection | Predicted blue cluster based on 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 : Proportion of the stimulated and unstimulated cells in each 5 clusters of the predicted blue region of the first projection. The three numbers are: proportion of stimulated (unstimulated) cells from the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, for each cluster.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 |

Table : Proportion of the stimulated and unstimulated cells in each 7 clusters of the predicted blue region of the third projection. The three numbers are: proportion of stimulated (unstimulated) cells from the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, for each cluster.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 |
| Stimulated | 0.0041 (50.5%) 2216 | 0.0036 (59.3%) 1904 | 0.0012 (55.2%) 665 | 0.0027 (46.3%) 1428 | 0.0018 (56.5%) 977 | 0.0013 (58.2%) 703 | 0.0018 (49.9%) 986 |
| Unstimulated | 0.0023 (49.5%) 2170 | 0.0014 (40.7%) 1307 | 0.0006 (44.8%) 539 | 0.0018 (53.7%) 1659 | 0.0008 (43.5%) 752 | 0.0005 (41.8%) 504 | 0.0011 (50.1%) 990 |

# 6. References

1 McKinnon, K.M.: ‘Flow Cytometry: An Overview’, Curr Protoc Immunol, 2018, 120, pp. 5 1 1-5 1 11

2 Han, Y., Wang, S., Zhang, Z., Ma, X., Li, W., Zhang, X., Deng, J., Wei, H., Li, Z., Zhang, X.E., and Cui, Z.: ‘In vivo imaging of protein-protein and RNA-protein interactions using novel far-red fluorescence complementation systems’, Nucleic Acids Res, 2014, 42, (13), pp. e103

3 Eastman, A.E., and Guo, S.: ‘The palette of techniques for cell cycle analysis’, FEBS Lett, 2020

4 Spidlen, J., Moore, W., Parks, D., Goldberg, M., Bray, C., Bierre, P., Gorombey, P., Hyun, B., Hubbard, M., Lange, S., Lefebvre, R., Leif, R., Novo, D., Ostruszka, L., Treister, A., Wood, J., Murphy, R.F., Roederer, M., Sudar, D., Zigon, R., and Brinkman, R.R.: ‘Data File Standard for Flow Cytometry, version FCS 3.1’, Cytometry A, 2010, 77, (1), pp. 97-100

5 <https://isac-net.org/2023>

6 ‘FlowJo Software’, in Editor (Ed.)^(Eds.): ‘Book FlowJo Software’ (Becton, Dickinson and Company, 2023, edn.), pp.

7 ‘FSC Express’, in Editor (Ed.)^(Eds.): ‘Book FSC Express’ (DeNovo Software by Dotmatics, 2023, edn.), pp.

8 Brinkman, J.S.K.B.C.R.N.K.R.R.: ‘FlowRepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications’, Cytometry Part A, Journal of QUantitative Cell Science, 2012, 81, (9), pp. 727-731

9 <http://flowrepository.org/id/FR-FCM-ZZZU2023>

10 Beavers, T.C., J.; Chen G, Duan Y; Lubomirski; Tiggler, S, M.: ‘Data Nuggets: A Method for Reducing Big Data While Preserving Data Structure’, IN REVISION, 2023

11 Tukey, J.H.F.J.W.: ‘A Projection Pursuit Algorithm for Exploratory Data Analysis’, IEEE Transactions on Computers, 1974, C-23, (9), pp. 881-890

12 Friedman, J.H.: ‘Exploratory Projection Pursuit’, Journal of the American Statistical Association, 1987, 82, (397), pp. 249-266

13 Cabrera, D.C.A.B.J.: ‘Projection Pursuit Indexes Based on Orthonormal Function Expansions’, Journal of Computational and Graphical Statistics, 1993, 2, (3), pp. 225-250

14 Duan, Y.C., J.: ‘New Projection Pursuit Index for Big Data’, IN REVISION, 2023