**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

**\*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 Flow Cytometry Analysis base on Projection Pursuit**

# **Abstract**

This paper introduces the novel concept of differential projection pursuit and its applications to the analysis of large datasets. The method was applied to a flow cytometry dataset. Multicolor flow cytometry is a powerful, well-established laboratory technique to identify cell subpopulations by measuring their physical and biochemical characteristics. Projection pursuit is a standard methodology for analyzing multivariate data. It creates low-dimensional projections that display the main structure of high-dimensional data. The objective of differential projection pursuit methodology is to find projections that maximize the difference between two or more treatments or distributions. A new projection pursuit index will be introduced, similar to the Natural Hermite index, that is suitable for measuring differences between two or more distributions. When applied to flow cytometry data, projection pursuit produces a small number of projections that visualize regions of space corresponding to cell subpopulations with density differences across experimental conditions. The final form of these projections involves a novel factor rotation that aligns the axis with a small number of variables in each dimension. The new structure is generally much clearer than that found in ordinary 2-dimensional projections. Since the number of cells in the samples is typically large, the data needs to be reduced by summarizing it into a much smaller number of groups called data nuggets. This produces optimal projections at a highly reduced computational cost.

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

# **1** **Introduction**

## ***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 a mixed population of cells from blood, bone marrow, and solid tissues based on cluster of differentiation (CD) markers, chemokine receptor markers, cytokines, and antigen-specific markers (McKinnon 2018). Fluorescent protein analysis revolutionized in vivo detection of bacterial or viral infections, transplanted cells, and gene knockout in cells (Han et al. 2014). Cell cycle analysis is another critical assay to the measure dynamics of pathophysiological processes by utilizing DNA-binding dyes (Eastman and Guo 2020). 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 for 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 the sample to the laser intercept point, optics that induce visible and fluorescent light signals, and electronics that convert analog light signals to digital (McKinnon 2018). The instruments contain multiple 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, which 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 can measure electrical current impedance, i.e., the opposition to alternative current as the cells travel through the tubes. This allows for the 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 cytometry 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 (Spidlen et al. 2010; 'International Society for Advancement in Cytometry').

Following acquisition, the data is processed, traditionally using a technique called *gating*. Specialized tools such as *FlowJo* ("FlowJo Software" 2023) and *FSC Express* ("FSC 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. The gating strategy follows the 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 also 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 (Brinkman 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 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)-stimulated 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 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 the one used in this study would be computationally expensive beyond most users’ capabilities. Therefore, a compression algorithm called *data nugget* (Beavers 2023) 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 is followed by a linear transformation 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.

## ***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 (Tukey 1974). Friedman index was introduced later as an example of PP indices (Friedman 1987). Cook et al. made substantial progress in this area by introducing several new PP indices (Cook 1993). In particular, they introduced the *natural Hermite index,* which 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 .

(1)

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 (Duan 2023). 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 *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:

(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 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:

(3)

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:

(4)

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. 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.

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.

# **3 Results**

After compressing the raw data into 3,385 refined data nuggets, the top six 2-dimensional projections were produced and rotated using the varimax procedure. The density was estimated for the stimulated ( and unstimulated () cells, and the difference between 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 of 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 simplicity, projections one and three will be discussed here; the rest of the projections could be discussed similarly. 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 the signal levels of the channels in the predicted blue and red regions. The loadings of flow cytometry channels in projections one and three are shown in Figure 6. In Projection one, 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 one and three, 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. These numbers were 5 and 7 for projections one and three, respectively. The boxplots in figures 7(b) and 8(b) 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 one and three, as well as the percentage and counts of the stimulated and unstimulated cells.

The same analysis was performed for the red region. Figure 9 shows clusters that were obtained from the red region. The results of the analysis, including the boxplot of the clusters and tables with the proportion and percentage of cells in each cluster, are in the supplement to the paper in Appendix I.

# **4 Discussion**

The gating approach to flow cytometry data analysis is determined in part by biology but by the limitations of computing power and tools that would allow multidimensional data visualization and analysis. Plotting and clustering such data in two dimensions at a 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 work, the approach was to step back from gating and instead examine 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 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 conformation from other data sets. This methodology has the potential to create a platform for analyzing 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**

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

Brinkman, Josef Spidlen; Karin Breuer; Chad Rosenberg; Nikesh Kotecha; Ryan R. 2012. 'FlowRepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications', *Cytometry Part A, Journal of QUantitative Cell Science*, 81: 727-31.

Cook, Dianne; Buja, Andreas; Cabrera, Javier. 1993. 'Projection Pursuit Indexes Based on Orthonormal Function Expansions', *Journal of Computational and Graphical Statistics*, 2: 225-50.

Duan, Y.; Cabrera, J. 2023. 'New Projection Pursuit Index for Big Data', *IN REVISION*.

Eastman, A. E., and S. Guo. 2020. 'The palette of techniques for cell cycle analysis', *FEBS Lett*.

'FLOW Repository'. <http://flowrepository.org/id/FR-FCM-ZZZU>.

"FlowJo Software." In. 2023. Becton, Dickinson and Company.

Friedman, Jerome H. 1987. 'Exploratory Projection Pursuit', *Journal of the American Statistical Association*, 82: 249-66.

"FSC Express." In. 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, and Z. Cui. 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/>.

McKinnon, K. M. 2018. 'Flow Cytometry: An Overview', *Curr Protoc Immunol*, 120: 5 1 1-5 1 11.

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

Tukey, J.H. Friedman; J.W. 1974. 'A Projection Pursuit Algorithm for Exploratory Data Analysis', *IEEE Transactions on Computers*, C-23: 881-90.

# **Appendix I: Results of the Analysis of the Red Region of the Projections One and Three**

# The box plots in Figure 10 show the protein expression levels for each cluster of the red region in the first and third projections. Tables 5 and 6 represent the proportion of stimulated and unstimulated cells from the total in each cluster of the predicted red region for projections one and three, as well as the percentage and counts of the stimulated and unstimulated cells.

A diagram of different colored objects

Description automatically generated with medium confidence

Figure 10: Expression level for each protein in the 7 clusters of the first projection (a) and 6 clusters of the third projection (b) in the predicted red region.

Table 5: Proportion of the stimulated and unstimulated cells in each 7 clusters of the predicted red 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.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **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 |

Table 6: Proportion of the stimulated and unstimulated cells in each 6 clusters of the predicted red 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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Condition** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** | **Cluster 5** | **Cluster 6** |
| Stimulated | 0.0029 (26.6%) 1546 | 0.0033 (25.5%) 1777 | 0.0007 (26.9%) 384 | 0.0014 (25.3%) 753 | 0.0005 (19.8%) 252 | 0.0008 (24.1%) 435 |
| Unstimulated | 0.0046 (73.4%) 4269 | 0.0056 (74.5%) 5196 | 0.0011 (73.1%) 1043 | 0.0024 (74.7%) 2225 | 0.0011 (80.2%) 1023 | 0.0015 (75.9%) 1372 |

# **Tables**

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

|  |  |
| --- | --- |
| **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 2: Proportion of the stimulated and unstimulated cells in the predicted blue region. The three numbers in the third column are the proportion of stimulated (unstimulated) cells in the total, percentage of stimulated (unstimulated) cells in the region, and number of stimulated (unstimulated) cells in the region for each of the six projections.

|  |  |  |  |
| --- | --- | --- | --- |
| **Projection** | **Condition** | **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 3: Proportion of the stimulated and unstimulated cells in each 5 clusters of the predicted blue region of the first projection. The three numbers are the proportion of stimulated (unstimulated) cells in the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, 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 |

Table 4: Proportion of the stimulated and unstimulated cells in each 7 clusters of the predicted blue region of the third projection. The three numbers are the proportion of stimulated (unstimulated) cells in the total, percentage of stimulated (unstimulated)for the region, and number of stimulated (unstimulated) cells within the region, for each cluster.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Condition** | **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 |

**Figures**

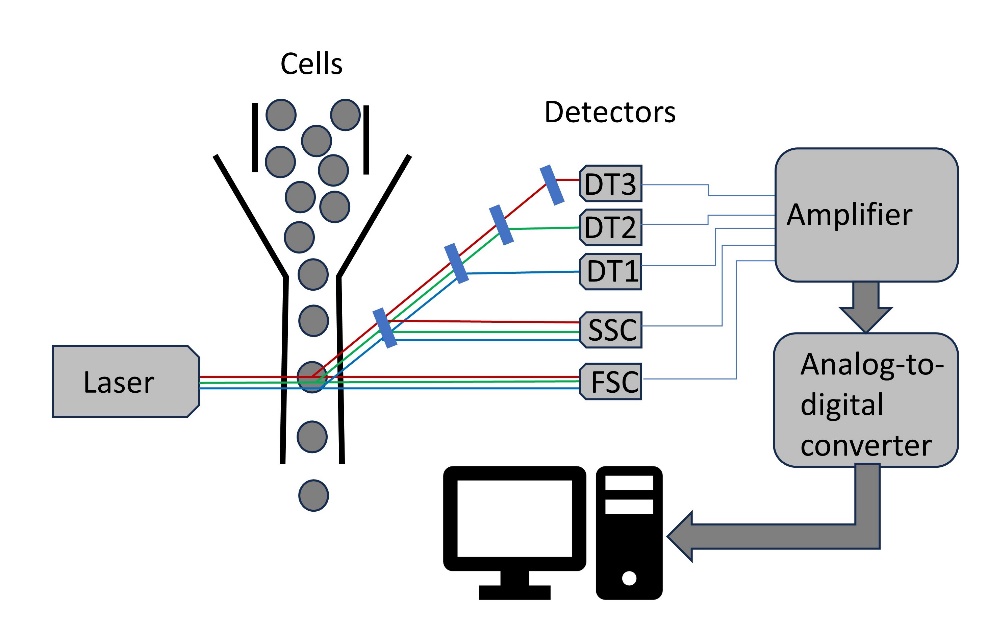


Figure 1: Schematics of a flow cytometer. As the cells pass through a narrow tube, the florochromes attached to the cells get excited by the laser. The light is deflected by a series of optical filters into the detectors. The signal from the detectors is amplified, converted from analog to digital, and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

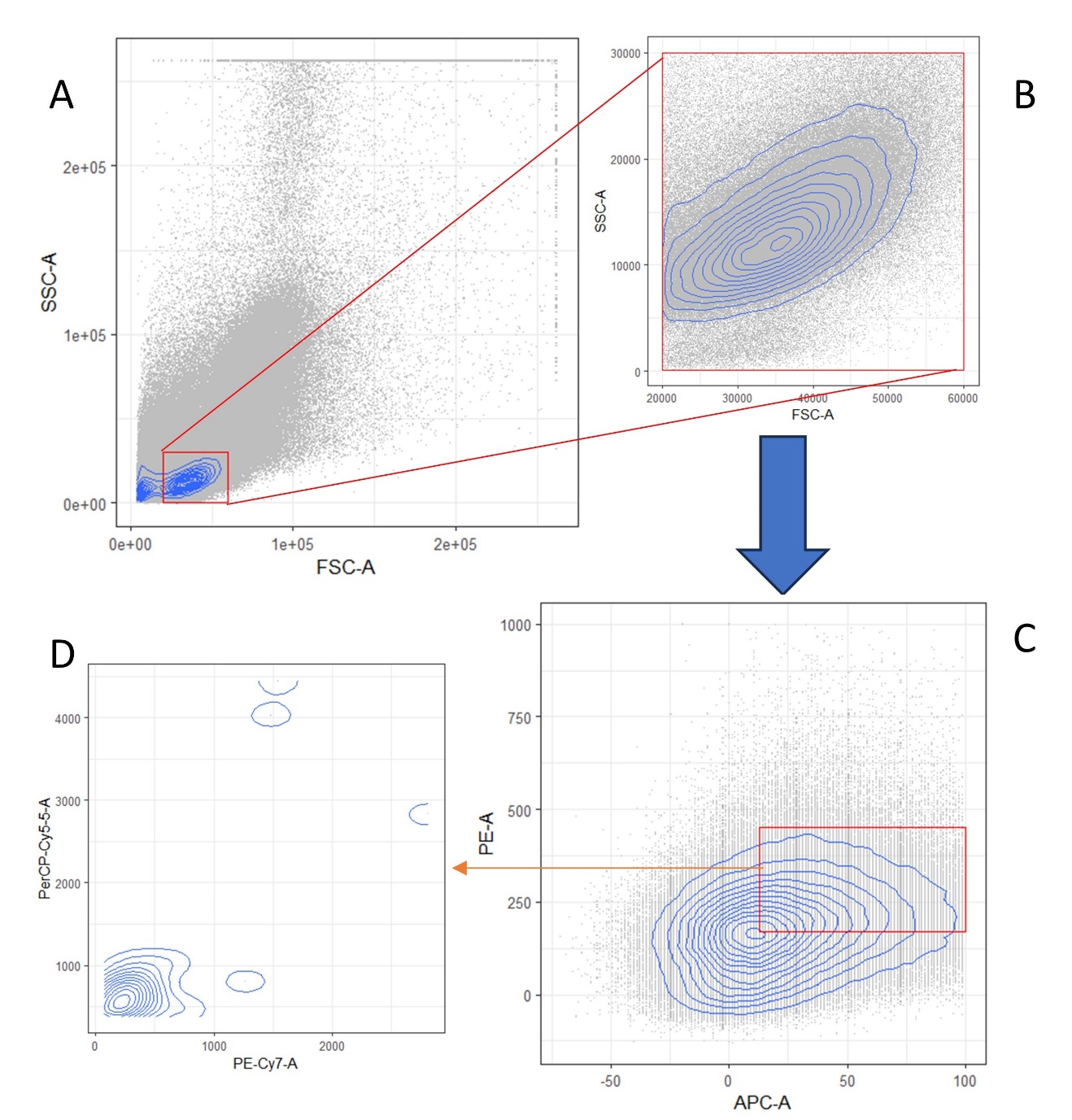


Figure 2: An example of flow cytometry gating. First, FSC-A and SSC-A is plotted, and the area of interest selected by drawing the gate (red). Next, this subset of cells is plotted in 2 other flow cytometry channels, PE-A and APC-A. A new gate is drawn and the process repeats using the next 2 channels. The 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), 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 data nuggets

Description automatically generated

Figure 4: Projected data nuggets (first column), estimated density plot for the data nuggets (second column), and projected raw data (third column) for the first (a) and third projections (b).

A diagram of different colored boxes

Description automatically generated with medium confidence

Figure 5: Flow cytometry channel signal levels in the predicted blue and red regions of the first (a) and the third (b) projections.

A comparison of a number of different types of numbers

Description automatically generated with medium confidence

Figure 6: Loading plots for the first (a) and third (b) projections.

A diagram of different colored boxes

Description automatically generated

Figure 7: Five clusters of the predicted blue region (a) and protein expression level for each cluster (b) for projection one.

A diagram of different colored boxes

Description automatically generated with medium confidence

Figure 8: Five clusters of the predicted blue region (a) and protein expression level for each cluster (b) for projection three.

A comparison of different colored clusters

Description automatically generated

Figure 9: Clusters of the predicted red region for projection one (a) and projection three (b).

# 

# **Figure Captions**

[Figure 1: Schematics of a flow cytometer. As the cells pass through a narrow tube, the florochromes 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. 19](#_Toc146979747)

[Figure 2: PLACEHOLDER! Davit to make a new one 19](#_Toc146979748)

[Figure 3: 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. 20](#_Toc146979749)

[Figure 4: Projected data nuggets (first column), estimated density plot for the data nuggets (second column), and projected raw data (third column) for the first (a) and third projections (b) 21](#_Toc146979750)

[Figure 5: Flow cytometry channel signal levels in the predicted blue and red regions of the first (a) and the third (b) projections. 21](#_Toc146979751)

[Figure 6: Loading plots for the first (a) and third (b) projections. 22](#_Toc146979752)

[Figure 7: Five clusters of the predicted blue region (a) and protein expression level for each cluster (b) for projection one 22](#_Toc146979753)

[Figure 8: Five clusters of the predicted blue region (a) and protein expression level for each cluster (b) for projection three 23](#_Toc146979754)

[Figure 9: Clusters of the predicted red region for projection one (a) and projection three (b)…………... 23](#_Toc146979754)