bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DeepCyTOF: Automated Cell Classification of Mass Cytometry Data by Deep Learning and Domain Adaptation

Huamin Li [1*], Uri Shaham [2*], Yi Yao [3], Ruth Montgomery [3], and Yuval Kluger [1,4,5†]

[1]Applied Mathematics Program, Yale University, 51 Prospect St., New Haven, CT 06511, USA
[2]Department of Statistics, Yale University, 24 Hillhouse Ave., New Haven, CT 06511, USA
[3]Department of Internal Medicine, Yale School of Medicine, 333 Cedar St., New Haven, CT 06520, USA
[4]Department of Pathology and Yale Cancer Center, Yale University School of Medicine, New Haven, CT, USA
[5]Interdepartmental Program in Computational Biology and Bioinformatics,Yale University, New Haven, CT, USA

Abstract

Mass cytometry or CyTOF is an emerging technology for high-dimensional multiparameter single cell analysis that overcomes many limitations of fluorescence-based flow cytometry. New methods for analyzing CyTOF data attempt to improve automation, scalability, performance, and interpretation of data generated in large studies. However, most current tools are less suitable for routine use where analysis must be standardized, reproducible, interpretable, and comparable. Assigning individual cells into discrete groups of cell types (gating) involves time-consuming sequential manual steps untenable for larger studies. The subjectivity of manual gating introduces variability into the data and impacts reproducibility and comparability of results, particularly in multi-center studies. The FlowCAP consortium was formed to address these issues and it aims to boost user confidence in the viability of automated gating methods. We introduce DeepCyTOF, a standardization approach for gating based on a multi-autoencoder neural network. DeepCyTOF requires labeled cells from only a single sample. It is based on domain adaptation principles and is a generalization of previous work that allows us to calibrate between a source domain distribution (reference sample) and multiple target domain distributions (target samples) in a supervised manner. We apply DeepCyTOF to two CyTOF datasets generated from primary immune blood cells: (i) 14 subjects with a history of infection with West Nile virus (WNV), and (ii) 34 healthy subjects of different ages. Each blood sample was labeled with 42 antibody markers, 12 of which were used in our analysis, at baseline and three different stimuli (PMA/ionomycin, tumor cell line K562, and infection with WNV). In each of these datasets we manually gated a single baseline reference sample to automatically gate the remaining uncalibrated samples. We show that DeepCyTOF cell classification is highly concordant with cell classification obtained by individual manual gating of each sample with over 99% concordance. Additionally, we apply a stacked autoencoder, which is one of the building blocks of DeepCyTOF, to cytometry datasets used in the 4th challenge of the FlowCAP-I competition and demonstrate that it over performs relative to all gating methods introduced in this competition. We conclude that stacked autoencoders combined with a domain adaptation procedure offers a powerful computational approach for semi-automated gating of CyTOF and flow cytometry data such that manual gating of one reference sample is sufficient for accurately gating the remaining samples.

[*]The first two authors contributed equally to this work.
[†]To whom correspondence should be addressed. Tel: 203-737-6262; Email: yuval.kluger@yale .edu

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 Introduction

Flow cytometry (FCM) is routinely used in cellular and clinical immunology. Current fluorescence-based FCM experiments provide up to 15 numeric parameters for each individual cell from blood samples in a high-throughput fashion. This allows efficient multi-parameter characterization of single cell states. Interpretation of such data from hundreds-of-thousands to millions of cells is paramount to understanding the pathogenesis of a broad range of human diseases. Mass cytometry (CyTOF) is an emergent technological development for high-dimensional multi-parameter single cell analysis. By using heavy metal ions as labels and mass spectrometry as readout, many more markers can be simultaneously measured. Current CyTOF instruments allow users to probe over 40 antibody specificities and thus provide a significant improvement in analyzing the underlying cell sub-populations of a system [1, 2]. CyTOF provides unprecedented multidimensional immune cell profiling and has recently been applied to characterizing peripheral blood cells, Natural Killer cells in viral infections, skin cells, cells in celiac disease, responding phenotypes in cancer, and even holds the promise of examination of solid tumors [3, 4, 5, 6, 7, 8, 9, 10]. Cellular characterization by FCM and CyTOF will improve our understanding of disease processes [11].

Gating (assigning individual cells into discrete groups of cell types) is one of the important steps and a bottleneck of analyzing FCM and CyTOF data. It involves time-consuming sequential manual steps untenable for larger studies [12, 13, 14, 15, 16, 17, 18, 19]. The time it takes to manually analyze a cytometry experiment depends on the number of experimental units as well as the number of markers [20]. If the gating strategy is complex, the analysis time might increase dramatically. Technical variation naturally arises due to the variation between individual operators [21]. The subjectivity of manual gating introduces variability into the data and impacts reproducibility and comparability of results, particularly in multi-center studies [22]. Thus the slow processing time and the inherent subjectivity of manual analysis should be considered as primary reasons for using computational assistance methods.

The FlowCAP consortium aims to boost user confidence in the viability of automated gating methods [23]. Many of the pipelines described therein are tailored for exploratory, discovery-oriented data analysis. New methods for analyzing cytometry data continue to emerge; these methods attempt to improve automation, scalability, performance, and interpretation of data generated in large studies. These computational methods can be categorized as unsupervised or supervised approaches. Both types of approaches use a variety of simple linear transformations, density estimations, hierarchical clustering, and nonlinear projection methods, that together allow extracting features that can be used to study differences between samples. However, most current tools are less suitable for routine use where analysis must be standardized, reproducible, interpretable, and comparable [24]. In general, no automated gating algorithm or approach that would solve all specific computational problems has been accepted as the gold standard for replacing manual gating [23, 25].

In the last few years, deep learning methods have achieved outstanding performance in various computational tasks, such as image analysis, natural language processing, and pattern recognition [26]. These approaches have also been shown to be effective for extracting natural features from data in general settings [27, 28, 29]. Moreover, recent efforts to use deep learning approaches in genomics and biomedical applications show their flexibility for handling complex problems [30, 31, 32, 33, 34]. However, deep learning typically requires very large number of training instances and thus its utility for many genomic, proteomic and other biological applications is questionable. While in most genomics applications, the number of instances (e.g., number of gene expression arrays) is typically smaller than the number of variables (e.g., genes), in each FCM and CyTOF run we typically collect approximately 10^5 to 10^6 cells, so that the number of instances (cells) is several orders of magnitude larger than the number of variables (up to 50 markers). Therefore, developing deep learning approaches to analyze cytometry data is very promising.

In the 4th challenge of the FlowCAP-I competition [23], the goal was to automatically gate 75% of cells in each sample based on manual gating of the remaining 25% of cells. We compare neural nets which were trained as stacked autoencoders to the other supervised competing approaches included in this challenge, and show that the neural nets outperform the competition winner and achieve state-of-the-art accuracy.

In addition, we present DeepCyTOF, an integrated deep learning neural network and domain adaptation framework, which employs one manually gated reference sample and utilizes it for automated gating of the remaining samples in a study. The net we construct is a generalization of DLID, a domain adaptation approach proposed by Chopra et al. [35], for the case of arbitrary number of samples, and is based on combining a collection of autoencoders into a single neural network. We used DeepCyTOF for automatic gating of two CyTOF datasets consisting of 56 and 136 samples respectively, and obtained almost identical results to those obtained by manual gating. We demonstrate that by utilizing manual gating of a single reference sample in each of these datasets, DeepCyTOF can achieve high quality automatic gating for the remaining samples and significantly reduce the time and effort that are currently required for manual gating. Finally, by simulating multiple uncalibrated samples through various transformations

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of one CyTOF sample, we demonstrate that DeepCyTOF is capable of calibrating the simulated samples and that automated gating of these samples is highly accurate.

The structure of this manuscript is as follows: in Section 2 we describe the datasets and algorithms used in this research. Experimental results are given in Section 3. Section 4 concludes this manuscript. Experimental details are given in an appendix.

2 Materials and Methods

Throughout this manuscript, we use the terms *sample* and *subject* as follows: a *sample* is a collection of measurements of cells, belonging to a single *subject*.

2.1 Datasets

2.1.1 FlowCAP-I Datasets

We employed five FCM datasets from FlowCAP-I [23]: (1) Diffuse large B-cell lymphoma (DLBCL), (2) Symptomatic West Nile virus (WNV), (3) Normal donors (ND), (4) Hematopoietic stem cell transplant (HSCT), and (5) Graft-versus-host disease (GvHD). With the results from manual gates produced by expert analysis, the goal of FlowCAP-I challenges was to compare the results of assigning cell events to discrete cell populations using automated gates. In particular, we consider Challenge 4: supervised approaches trained using human-provided gates. We use the manual gating provided from FlowCAP-I to evaluate the predictions of our algorithm in Section 2.2.1.

2.1.2 Mass Cytometry Datasets

We employed two CyTOF datasets generated in the Montgomery Lab. The datasets consist of primary immune cells from blood of (1) $N = 14$ subjects (8 asymptomatic and 6 severe) with a history of infection with West Nile virus (WNV), and (2) $N = 34$ healthy subjects of different ages (20 young and 14 old). Each blood sample was labeled with $d = 42$ antibody markers [5], 12 of which were used in our analysis (see Table 1) as they were the relevant markers for the task of classification described below. Each sample was subjected to four CyTOF experiments including a baseline state and three different stimuli (PMA/ionomycin, tumor cell line K562, and infection with WNV). The goal is to classify each cell to one of 6 cell type categories: (1) unlabeled, (2) B cell, (3) CD4+ T cell, (4) CD8+ T cell, (5) Monocytes, and (6) Natural killer (NK) cells. This task is challenging as it must be accompanied by a procedure that calibrates between samples. Specifically, different samples were measured at different times; fine changes in the state of the CyTOF instrument between those times thus introduce additional variability into the measurements. The specific nature of these changes is neither known nor modeled. To tackle this problem and apply a gating procedure, most practitioners in the field calibrate the samples by applying an experimental-based normalization procedure. This procedure involves mixing samples with polystyrene beads embedded with metal lanthanides, followed by an algorithm which enables correction of both short-term and long-term signal fluctuations [36]. Once the data is normalized, most practitioners in the field apply a manual gating procedure.

Table 1: 12 markers used for cell classification

Marker type	Measured marker
Lineage	HLA-DR, CD3-UCHT1, CD16, CD33, CD19, CD14, CD56
Surface	CD4, CD8
Other	DNA1, DNA2, Cisplatin

2.1.3 Simulated Datasets

Starting from a single manually gated CyTOF dataset, we generated 50 simulated samples where each simulated sample corresponds to a (different) change in the calibration of the CyTOF machine. This collection of datasets was used to test whether DeepCyTOF is capable of overcoming substantial calibration differences between samples. A detailed mathematical formulation of the data generation process appears in Appendix C.

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

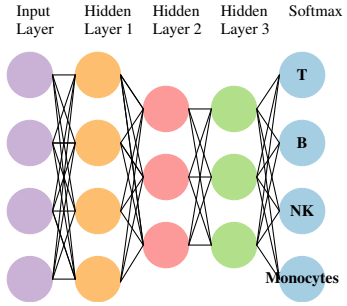

Figure 1: Stacked autoencoder for gating cell populations

2.1.4 Pre-processing

The only pre-processing performed on the FlowCAP-I , the Mass Cytometry datasets and the simulated datasets are a logarithmic transform, followed by rescaling, as described in Appendix B.

2.2 Algorithms

2.2.1 Stacked autoencoders

An autoencoder is a single hidden layer neural network, which is trained to reconstruct its input [37]. The number of units in any hidden layer of the autoencoder is typically smaller than the input dimension, hence a trained autoencoder can be viewed as a tool for nonlinear dimensionality reduction. The map from the input representation to the hidden representation is usually called encoder. The map from the hidden layer to the output layer (which is a reconstruction of the input) is called decoder. It is widely known that the reconstruction of the data from the code space is often "cleaner" than its original representation [38]; hence the autoencoder can also be viewed as a denoising tool.

A stacked autoencoder is a multi-layer neural net, which is trained bottom-up in an iterative fashion, where in each iteration, a single layer is trained as an autoencoder [39]. Once the layer is trained, one uses it to compute a new representation for the data. This representation is then used to train the following layer and so on. Once a stacked autoencoder is trained, one may add a classifier on top of it (a softmax is a popular choice), and then fine-tune the entire neural net using standard backpropagation [40]. Such approach has been widely popular in deep learning applications e.g., [41]. Notably, autoencoders are trained in an unsupervised fashion [42, 43, 44], hence they do not require labeled data. The final fine-tuning step for the specific classification task naturally requires labeled data. However, it was empirically shown that the number of labeled examples can be relatively small, where the data contains regularities which are captured by the autoencoders. Hence, such an approach is particularly useful in applications where the number of unlabeled examples is large while the number of labeled ones is small [28].

The full architecture of our stacked autoencoders consists of three fully connected hidden layers with sigmoid non-linearity, and a softmax regression layer on top. The number of hidden nodes in each layer has been set to 12, 6, and 3, which we found to work well in practice. A scheme of such stacked autoencoder is shown in Figure 1.

2.2.2 DeepCyTOF

Domain Adaptation is a set of techniques that allow to use a learning scheme (or a model) trained on data from a source domain with a given distribution and can also be applied to a target domain with a related but not equivalent distribution. The objective of domain adaptation is to minimize the generalization error of instances from the target domain [45],[46]. In FCM and CyTOF experiments, variation in both biological and technical sources makes automatic gating challenging. Health condition and instrument calibration cause variation across samples; hence, in order to avoid gating each dataset separately (which therefore requires labeled samples from each dataset), a domain adaptation procedure might be helpful.

DLID [35] is a deep learning domain adaptation approach that is based on creating "intermediate" datasets, in each of which the source and target distributions are mixed. The original approach is formulated for the case of two distributions. In this approach a classifier is trained using labeled data from the source distribution and operates well

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

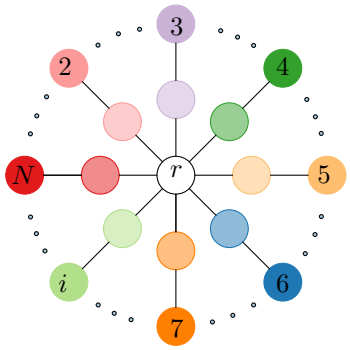

Figure 2: A star graph representing multiple autoencoders. Autoencoders are constructed for the reference sample (r), for each sample in the study as shown in the outer circle ($i = 2, \cdots, N$), as well as for a mixture of each sample in the study with the reference sample as shown in the inner circle (where each mixture consists of 50% randomly selected cells from the sample in the same branch of the star graph and 50% the cells from the reference sample)

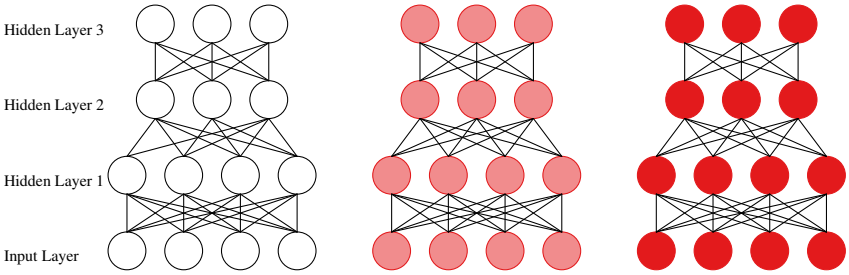

Figure 3: Unsupervised pre-training steps of DeepCyTOF: three stacked autoencoders are applied to reference sample r (open circles), sample i (red circles), and an intermediate dataset (pink circles), in which the cells from samples r and i are mixed with equal proportions.

on instances from both source and target distributions. In this work, we generalize the approach to arbitrary number of datasets corresponding to one source distribution (reference sample) and multiple target distributions (the remaining samples). In this generalization the autoencoders of the source reference sample, target samples and mixtures of the target samples with the reference sample can be represented in a generalized star-like topology (see Figure 2). To determine which of the N samples in a baseline condition is a suitable candidate for being used as a reference sample, we first compute the $d \times d$ covariance matrix for each of these N samples, where d is the dimensionality of each of the datasets associated with these samples. We denote this reference sample by r and associate it with the index $i = 1$. Next, we train autoencoders using the unlabeled data for the reference sample r and for each of the other samples in the baseline condition $i = 2, \cdots, N$. In addition, for each of the target samples $i = 2, \cdots, N$, we also train an autoencoder using 50% of its cells mixed with 50% of the cells of the reference sample r (see Figure 3).

Subsequently, we combine these $1 + 2(N - 1)$ autoencoders to a single large neural net, add a softmax layer on top and fine-tune the net using labeled data obtained by manual gating from subject r only. During the fine-tuning step, the two stacked autoencoders in each branch of the star are connected; each branch is also connected to the stacked autoencoder of the reference sample, as depicted in Figure 4, so that the star graph functions as a single network, with a classifier as its upper layer.

2.3 Comparison of auto and manual gating

To compare the stacked autoencoders approach of Section 2.2.1 with algorithms from the 4th challenge of the FlowCAP-I competition, 25% of the cells of each subject from the FCM datasets in FlowCAP-I are labeled by manual gating and used in the unsupervised pre-training and supervised fine-tuning steps implemented in our algo-

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

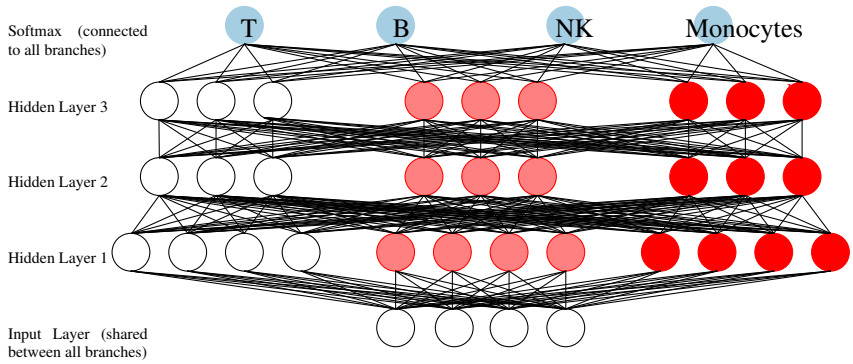

Figure 4: Connectivity between the autoencoders of each branch of the Star graph in the fine tuning step. The two stacked autoencoders on every branch of the Star graph are combined to a single neural net. The reference sample is connected to each branch. A single softmax classifier is connected on top of the last hidden layer of all stacked autoencoders.

rithm, leaving 75% of the remaining cells for testing. To perform semi-automated gating of all samples of each of the two CyTOF datasets based on their respective reference samples, we use all the baseline (unlabeled) samples for the unsupervised pre-training and a single baseline (labeled) reference sample r to fine-tune the net, leaving all samples that undergo three different stimuli (PMA/ionomycin, tumor cell line K562, and infection with WNV) for testing.

The F-measure statistic (the harmonic mean of precision and recall) is used for the evaluation of our methods as described in [23]. The F-measure for multiple classes is defined as the weighted average of F-measures for each cell type against all other classes. An F-measure of 1.0 indicates perfect agreement with the labels obtained by manual gating. For any given dataset, we create several bootstrap datasets (by sampling with replacement), compute the F-measure on each and report the mean and standard deviation of the F-measure.

3 Results

In this section, we present results from three experiments. First, we use stacked autoencoders to perform cell classification on each of the five FCM datasets from the FlowCAP-I competition. Second, we apply DeepCyTOF onto two CyTOF datasets and demonstrate how combination of deep learning and domain adaptation procedures can effectively eliminate the need for manual gating of all samples in a study. Finally, we use simulated datasets to demonstrate the ability of DeepCyTOF to overcome substantial calibration differences.

3.1 Evaluation of classification performance from FlowCAP-I

The prediction is performed by training a stacked autoencoder (as described in Section 2.2.1) on each dataset. Each stacked autoencoder had three fully connected hidden layers and a softmax layer on top. Table 2 presents the performance of the stacked autoencoder when applied on the five datasets from the 4th challenge of FlowCAP-I competition. The performance of stacked autoencoder is compared to the performance of the respective winner of each of the five datasets in this competition.

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2: Summary of results for the cell identification challenge. The numbers in parentheses represent 95% confidence intervals

$F = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$	Stacked autoencoder	Competition's winner
GvHD[a]	0.98 (0.97,0.99)	0.92 (0.88,0.95)
DLBCL	0.97 (0.95, 0.99)	0.95 (0.93, 0.97)
HSCT	0.98 (0.96, 0.99)	0.98 (0.96, 0.99)
WNV	0.98 (0.97, 0.99)	0.96 (0.94, 0.97)
ND	0.98 (0.96, 0.99)	0.94 (0.92, 0.95)

[a] Datasets: graft-versus-hist disease (GvHD); diffuse large B-cell lymphoma (DLBCL); symptomatic West Nile virus (WNV); normal donors (ND); hematopoietic stem cell transplant (HSCT).

As can be seen in Table 2, the stacked autoencoders performs better in four out of the five datasets and similarly on the HSCT dataset.

3.2 Evaluation of DeepCyTOF by analysis of two mass cytometry datasets

In this experiment we applied DeepCyTOF separately to two different datasets (which contain 14 and 34 baseline samples, respectively) as follows: we first constructed a star graph for each of these two datasets by connecting the stacked autoencoders of the reference sample, target baseline samples and their mixture with the reference to a single network as in Section 2.2.2, and fine-tuned it using labeled data from the reference sample. We then used this classifier to classify cells from the non-reference samples (55 samples from the Asymptomatic vs. Severe WNV dataset and 135 samples from the Old vs. Young dataset) and compared to the performance of a linear classifier (softmax). Table 3 summarizes the results.

Table 3: Summary of results for the two CyTOF datasets. The numbers in parentheses represent 95% confidence intervals.

$F = \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$	AnS-UN[a]	AnS-N	OnY-UN	OnY-N
DeepCyTOF	0.990 (0.984, 0.995)	0.991 (0.987, 0.994)	0.992 (0.986, 0.997)	0.993 (0.987, 0.997)
Softmax regression	0.966 (0.955, 0.975)	0.967 (0.957, 0.976)	0.963 (0.946, 0.977)	0.963 (0.946, 0.977)

[a] Datasets: unnormalized Asymptomatic&Severe (AnS-UN); normalized Asymptomatic&Severe (AnS-N); normalized Old&Young (OnY-N); unnormalized Old&Young (OnY-UN).

Table 3 illustrates some interesting points: first, DeepCyTOF achieves nearly perfect performance on the test data. Second, DeepCyTOF performs significantly better than softmax regression, which may be a result of the depth and the non-linearity of the network. Third, whether or not the data is normalized [36] does not affect the performance of DeepCyTOF.

Figure 5 shows the embedding of the labeled cells in a three dimensional space, obtained from the top hidden layer of a stacked autoencoder (after fine-tuning). As can be seen, most of the labeled cells concentrate in well separated clusters representing specific cell types and only a few cells fall between these clusters.

3.3 Evaluation of classification performance from simulated data

To test the capability of DeepCyTOF to overcome substantial calibration differences, we performed an analysis based on simulated data. As described in Appendix C, we used a real baseline sample to generate 50 samples, each of which differs in calibration from the baseline sample, where the calibration difference depends on the sample, marker and cell type. We first constructed a star graph using all 50 samples where the center of the star is the

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

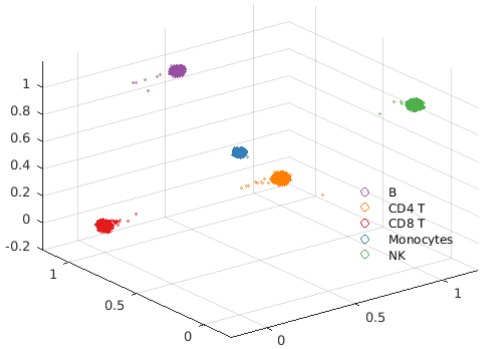

Figure 5: Third hidden layer representation of a blood sample (the unlabeled cells are omitted). Different cell types are concentrated in different regions of the code space.

reference sample. Then, we connected the stacked autoencoders to a single network as in Section 2.2.2, and fine-tuned it using labeled data from the reference sample. We used this classifier to classify cells from all 50 samples and evaluated its performance using the F-measure. The 95% confidence interval of the F-measure is (0.996, 0.999). This demonstrates how the domain adaptation component of DeepCyTOF allows one to perform automatic gating even in scenarios where the multi dimensional marker distribution of the labeled reference training data differs significantly from the corresponding distribution of the unlabeled test data. For comparison, a softmax classifier, trained on the reference sample, achieves a confidence interval of (0.956, 0.958).

4 Discussion

In this work, we showed that deep learning machinery, and stacked autoencoders in particular, can be very effective in classification of cell types; the performance substantially surpasses the predictive accuracy of the methods presented in the 4th challenge of the FlowCAP-I competition. In addition, we introduced DeepCyTOF, an automated framework for gating cell populations in cytometry samples. DeepCyTOF integrates deep learning and domain-adaption concepts. The labels obtained by manual gating of the reference sample were utilized in a domain-adaptation manner. These steps enable DeepCyTOF to inherently calibrate the major cell populations of multiple samples with respect to the corresponding cell populations of the reference sample. We analyzed 192 CyTOF samples and observed nearly identical results to those obtained by manual gating (with F-measure ≥ 0.99).

In practice, run-to-run variations in CyTOF experiments both in the same instrument and between instruments are very common. These variations lead to significant batch effects in the datasets with samples collected at different runs. As a result, noticeable differences between the data distributions of the training data (manually gated reference sample) and the remaining unlabeled test data (the remaining samples) are observed, and an approach such as domain-adaptation is required to remove these biases. Bead-normalization is an approach introduced to mass cytometry as a pre-processing step to diminish the effect of such variations [36]. Interestingly, application of DeepCyTOF to unnormalized and bead-normalized data did not show an advantage of using the latter for the task of automated gating. Our domain-adaptation approach allows us to effectively normalize different distributions for the (supervised learning) task of automated gating via introduction of intermediate representations of cytometry data, each consisting of instances from the reference (gated) distribution mixed with instances from a given un-gated distribution. A combined representation that encompasses the representations of all samples and their mixtures with the reference sample was designed to facilitate the classification of discrete groups of cell populations.

Flow cytometry and mass cytometry experiments provide us with multivariate data with dimensionality ranging between 10-40. Transforming the raw multivariate data to other representations may offer advantages for tasks such as automated gating or calibration. Finding good representations can be done either by manual investigation (hand crafting) or automated approaches. In recent years deep learning approaches have been shown to be suitable for learning useful representations of data in the sense that they provide new sets of features that makes subsequent learning easier. Furthermore, it has been shown that pre-training unsupervised steps such the ones we implemented in DeepCyTOF can improve the learning tasks [47], especially, when labeled training data is scarce. Cytometry

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

experiments provide us with large datasets of unlabeled cells, which makes the unsupervised pre-training steps in the construction of a deep neural network applicable.

As cytometry analyses become widely used in research and clinical settings, automated solutions for analyzing the high dimensional datasets are urgently needed. Current practice in which samples are first subjected to manual gating are slowly substituted by automatic gating methods [48]. Major contributions to between-sample variations in cytometry experiments arise not only due to biological or medical differences but due to machine biases. Here we demonstrated that a novel machine learning approach based on deep neural networks and domain adaptation can substitute manual gating as they both produce indistinguishable results. In future work, we will demonstrate that deep learning approaches could address other challenges in analyzing cytometry data. This include tasks such as unsupervised calibration of samples, and feature extraction for classification or visualization of multiple samples.

Acknowledgement

The authors would like to thank Catherine Blish for CyTOF reagents, Ronald Coifman, and Roy Lederman for helpful discussions and suggestions. This research was partially funded by NIH grant 1R01HG008383-01A1 (Y.K.).

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

[1] Dmitry R Bandura, Vladimir I Baranov, Olga I Ornatsky, Alexei Antonov, Robert Kinach, Xudong Lou, Serguei Pavlov, Sergey Vorobiev, John E Dick, and Scott D Tanner. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical chemistry*, 81(16):6813–6822, 2009.

[2] Olga Ornatsky, Dmitry Bandura, Vladimir Baranov, Mark Nitz, Mitchell A Winnik, and Scott Tanner. Highly multiparametric analysis by mass cytometry. *Journal of immunological methods*, 361(1):1–20, 2010.

[3] Sean C Bendall, Erin F Simonds, Peng Qiu, D Amir El-ad, Peter O Krutzik, Rachel Finck, Robert V Bruggner, Rachel Melamed, Angelica Trejo, Olga I Ornatsky, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696, 2011.

[4] Amir Horowitz, Dara M Strauss-Albee, Michael Leipold, Jessica Kubo, Neda Nemat-Gorgani, Ozge C Dogan, Cornelia L Dekker, Sally Mackey, Holden Maecker, Gary E Swan, et al. Genetic and environmental determinants of human nk cell diversity revealed by mass cytometry. *Science translational medicine*, 5(208):208ra145–208ra145, 2013.

[5] Dara M Strauss-Albee, Julia Fukuyama, Emily C Liang, Yi Yao, Justin A Jarrell, Alison L Drake, John Kinuthia, Ruth R Montgomery, Grace John-Stewart, Susan Holmes, et al. Human nk cell repertoire diversity reflects immune experience and correlates with viral susceptibility. *Science translational medicine*, 7(297):297ra115–297ra115, 2015.

[6] Yi Yao, Rebecca Liu, Min Sun Shin, Mark Trentalange, Heather Allore, Ala Nassar, Insoo Kang, Jordan S Pober, and Ruth R Montgomery. Cytof supports efficient detection of immune cell subsets from small samples. *Journal of immunological methods*, 415:1–5, 2014.

[7] Arnold Han, Evan W Newell, Jacob Glanville, Nielsen Fernandez-Becker, Chaitan Khosla, Yueh-hsiu Chien, and Mark M Davis. Dietary gluten triggers concomitant activation of cd4+ and cd8+ $\alpha\beta$ t cells and $\gamma\delta$ t cells in celiac disease. *Proceedings of the National Academy of Sciences*, 110(32):13073–13078, 2013.

[8] Jonathan M Irish and Deon B Doxie. High-dimensional single-cell cancer biology. In *High-Dimensional Single Cell Analysis*, pages 1–21. Springer, 2014.

[9] Charlotte Giesen, Hao AO Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J Schüffler, Daniel Grolimund, Joachim M Buhmann, Simone Brandt, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods*, 11(4):417–422, 2014.

[10] Michael Angelo, Sean C Bendall, Rachel Finck, Matthew B Hale, Chuck Hitzman, Alexander D Borowsky, Richard M Levenson, John B Lowe, Scot D Liu, Shuchun Zhao, et al. Multiplexed ion beam imaging of human breast tumors. *Nature medicine*, 20(4):436–442, 2014.

[11] Kondala R Atkuri, Jeffrey C Stevens, and Hendrik Neubert. Mass cytometry: A highly multiplexed single-cell technology for advancing drug development. *Drug Metabolism and Disposition*, 43(2):227–233, 2015.

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

[13] Karthik Shekhar, Petter Brodin, Mark M Davis, and Arup K Chakraborty. Automatic classification of cellular expression by nonlinear stochastic embedding (accense). *Proceedings of the National Academy of Sciences*, 111(1):202–207, 2014.

[14] El-ad David Amir, Kara L Davis, Michelle D Tadmor, Erin F Simonds, Jacob H Levine, Sean C Bendall, Daniel K Shenfeld, Smita Krishnaswamy, Garry P Nolan, and Dana Pe'er. visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology*, 31(6):545–552, 2013.

[15] Peng Qiu, Erin F Simonds, Sean C Bendall, Kenneth D Gibbs Jr, Robert V Bruggner, Michael D Linderman, Karen Sachs, Garry P Nolan, and Sylvia K Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nature biotechnology*, 29(10):886–891, 2011.

[16] Greg Finak, Jacob Frelinger, Wenxin Jiang, Evan W Newell, John Ramey, Mark M Davis, Spyros A Kalams, Stephen C De Rosa, and Raphael Gottardo. Opencyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. 2014.

[17] Greg Finak, Wenxin Jiang, Kevin Krouse, Chungwen Wei, Ignacio Sanz, Deborah Phippard, Adam Asare, Stephen C Rosa, Steve Self, and Raphael Gottardo. High-throughput flow cytometry data normalization for clinical trials. *Cytometry Part A*, 85(3):277–286, 2014.

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[18] Till Sörensen, Sabine Baumgart, Pawel Durek, Andreas Grützkau, and Thomas Häupl. immunoclustan automated analysis pipeline for the identification of immunophenotypic signatures in high-dimensional cytometric datasets. *Cytometry Part A*, 2015.

[19] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

[20] Chris P Verschoor, Alina Lelic, Jonathan L Bramson, and Dawn ME Bowdish. An introduction to automated flow cytometry gating tools and their implementation. *Frontiers in immunology*, 6, 2015.

[21] Christophe Benoist and Nir Hacohen. Flow cytometry, amped up. *Science*, 332(6030):677–678, 2011.

[22] Holden T Maecker, J Philip McCoy, and Robert Nussenblatt. Standardizing immunophenotyping for the human immunology project. *Nature Reviews Immunology*, 12(3):191–200, 2012.

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

[24] Tiffany J Chen and Nikesh Kotecha. Cytobank: providing an analytics platform for community cytometry data analysis and collaboration. In *High-Dimensional Single Cell Analysis*, pages 127–157. Springer, 2014.

[25] Kieran O'Neill, Nima Aghaeepour, Josef Špidlen, and Ryan Brinkman. Flow cytometry bioinformatics. *PLoS Comput Biol*, 9(12):e1003365, 2013.

[26] Li Deng and Dong Yu. Deep learning: methods and applications. *Foundations and Trends in Signal Processing*, 7(3–4):197–387, 2014.

[27] Yoshua Bengio, Aaron Courville, and Pierre Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

[28] Yoshua Bengio, Ian J. Goodfellow, and Aaron Courville. Deep learning. Book in preparation for MIT Press, 2015.

[29] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[30] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, pages 411–418. Springer, 2013.

[31] Angel Cruz-Roa, Ajay Basavanhally, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *SPIE Medical Imaging*, pages 904103–904103. International Society for Optics and Photonics, 2014.

[32] Olgert Denas and James Taylor. Deep modeling of gene expression regulation in an erythropoiesis model.

[33] Rasool Fakoor, Faisal Ladhak, Azade Nazi, and Manfred Huber. Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the International Conference on Machine Learning*, 2013.

[34] Michael KK Leung, Hui Yuan Xiong, Leo J Lee, and Brendan J Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30(12):i121–i129, 2014.

[35] Sumit Chopra, Suhrid Balakrishnan, and Raghuraman Gopalan. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML workshop on challenges in representation learning*, volume 2, 2013.

[36] Rachel Finck, Erin F Simonds, Astraea Jager, Smita Krishnaswamy, Karen Sachs, Wendy Fantl, Dana Pe'er, Garry P Nolan, and Sean C Bendall. Normalization of mass cytometry data with bead standards. *Cytometry Part A*, 83(5):483–494, 2013.

[37] Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length, and helmholtz free energy. *Advances in neural information processing systems*, pages 3–3, 1994.

[38] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408, 2010.

[39] Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle, et al. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19:153, 2007.

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[40] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

[41] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[42] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[43] Yoshua Bengio, Yann LeCun, et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5), 2007.

[44] Christopher Poultney, Sumit Chopra, Yann L Cun, et al. Efficient learning of sparse representations with an energy-based model. In *Advances in neural information processing systems*, pages 1137–1144, 2006.

[45] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[46] Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126, 2006.

[47] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520, 2011.

[48] Cariad Chester and Holden T Maecker. Algorithmic tools for mining high-dimensional cytometry data. *The Journal of Immunology*, 195(3):773–779, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/054411; this version posted May 31, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A Notation

The kth sample is denoted $A^{(k)}$, and is a $n_k \times d$ real valued matrix, where n_k denotes the number of cells and d is the dimension of the input.

B Data Pre-processing

Given N blood samples $A^{(k)}$ for $k = 1, 2, \ldots, N$, we first perform an elementary logarithmic transformation

$$A^{(k)}_{i,j} \leftarrow \log(1 + A^{(k)}_{i,j}) \tag{1}$$

Finally, we rescale each column of $A^{(k)}$ to $[0, 1]$. This is desirable when sigmoid activations are used for the output units of an autoencoder.

C Simulation

Starting from a real blood sample $A^{(1)}$, we calculate the mean of each cell population for each marker. Let μ be a 6 by d matrix such that $\mu_{t,j}$ is the mean intensity of marker j in cell type t.

Let $\triangle$ be defined as

$$\triangle = \frac{1}{C} \cdot \big[0, 1, \ldots, N-1\big] \tag{2}$$

for some positive constant C and let ϵ be vector of d i.i.d. normal random variables. We now generate N matrices $A^{(k)}$ with $k = 1, 2, \ldots, N$ by perturbing $A^{(1)}$ as follows

$$A^{(k)}_{i,j} = A^{(1)}_{i,j} + \triangle_k \cdot \mu_{t(i),j} + \epsilon_j \tag{3}$$

where $A^{(k)}_{i,j}$ is the ith row, jth column of matrix $A^{(k)}$ and $t(i)$ is type of the ith cell in $A^{(1)}$. The difference between subjects can be controlled by changing the value of C.

In summary, the datasets are generated so that each simulated dataset corresponds to a different data transformation of a single real dataset. The additive transformations of each marker in each cell type were implemented with different intensity $\triangle_k$ for each simulated dataset. In our study, we choose $N = 50$ and $C = 25$.