A cell-level discriminative neural network model for diagnosis of blood cancers

Edgar E. Robles [1,*], Ye Jin [2], Padhraic Smyth [1], Richard H. Scheuermann [3,4,5], Jack D. Bui [4], Huan-You Wang [4], Jean Oak [6], and Yu Qian [3,*]

[1]Department of Computer Science, University of California, Irvine, CA 92697, USA
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[3]Department of Informatics, J. Craig Venter Institute, La Jolla, CA 92037, USA
[4]Department of Pathology, University of California, San Diego, CA 92093, USA
[5]Center for Infectious Disease and Vaccine Research, La Jolla Institute for Immunology, La Jolla, CA 92037, USA
[6]Department of Pathology, Stanford University, Stanford, CA 94305, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Precise identification of cancer cells in patient samples is essential for accurate diagnosis and clinical monitoring but has been a significant challenge in machine learning approaches for cancer precision medicine. In most scenarios, training data are only available with disease annotation at the subject or sample level. Traditional approaches separate the classification process into multiple steps that are optimized independently. Recent methods either focus on predicting sample-level diagnosis without identifying individual pathologic cells or are less effective for identifying heterogeneous cancer cell phenotypes.
Results: We developed a generalized end-to-end differentiable model, the Cell Scoring Neural Network (CSNN), which takes the available sample-level training data and predicts both the diagnosis of the testing samples and the identity of the diagnostic cells in the sample, simultaneously. The cell-level density differences between samples are linked to the sample diagnosis, which allows the probabilities of individual cells being diagnostic to be calculated using backpropagation. We applied CSNN to two independent clinical flow cytometry datasets for leukemia diagnosis. In both qualitative and quantitative assessments, CSNN outperformed preexisting neural network modeling approaches for both cancer diagnosis and cell-level classification. Post hoc decision trees and 2D dot plots were generated for interpretation of the identified cancer cells, showing that the identified cell phenotypes match the cancer endotypes observed clinically in patient cohorts. Independent data clustering analysis confirmed the identified cancer cell populations.
Availability: The source code of CSNN and datasets used in the experiments are publicly available on GitHub and FlowRepository.
Contact: Edgar E. Robles: roblesee@uci.edu and Yu Qian: mqian@jcvi.org
Supplementary information: Supplementary data are available on GitHub and at *Bioinformatics* online.

1 Introduction

Challenges in the diagnosis and prognosis of blood cancers partially lie in the phenotypic heterogeneity of cancer cells. Even within a leukemia subtype, leukemic cells may be derived from slightly different stages of the normal cell developmental trajectory and therefore express different marker proteins, resulting in phenotypic heterogeneity within and between patient samples. To characterize the cellular phenotypic heterogeneity, single-cell assays are essential. In clinical laboratories, complete blood cell count (CBC), cytogenetics for identifying chromosomal

numerical and structural abnormalities, microscopy of bone marrow biopsy, cytology of cerebrospinal fluid, and flow cytometry (FCM) of peripheral blood and bone marrow aspirates are commonly used for leukemia and lymphoma diagnosis. Among these assays, FCM is the most mature single cell analysis technology, supporting identification and quantification of cell surface and intracellular proteins on individual cells. Compared with other single cell assays, FCM is rapid, cheap, and sensitive for detecting and monitoring phenotypic differences in cancer cells. Besides profiles of cellular marker expressions, proportions of the cancer cells within a specimen (i.e., cancer burden, an important measure for optimizing treatments and prognosis) can also be quantified from the analysis of FCM data. As a result, FCM immunophenotyping is routinely used in diagnosis and prognosis of blood and lymphoid cancers [6, 55, 46, 56, 60, 22, 10, 23, 54, 7]. It is also widely used to identify abnormal cell populations in association with non-neoplastic diseases including asthma, allergy, and autoimmunity [52, 19, 8, 4, 61, 25, 59].

Due to significant advances in cytometry instrumentation and reagent technologies since the 2000s [53, 39, 20, 41], applications of cutting-edge machine learning (ML) approaches began to emerge in the recent decade for addressing the increasing volume and complexity in this high-content cytometry data [18, 33, 47]. Automated gating analysis (auto-gating), which identifies cell populations using unsupervised clustering methods or recapitulates the manual identification gating process using supervised learning, in either the original or transformed feature space [29, 43, 36, 9, 63, 40, 44, 38, 34, 12, 48, 2, 26, 28, 57, 31, 1, 45, 51, 27], represents the largest category of these methods. To use these auto-gating methods for diagnosis, a separate disease classification step is required. The accuracy of the classification relies on the auto-gating step to identify cell populations in a complete and accurate way, which can be challenging for blood cancer applications. A second category of methods focuses on identifying cell-based biomarkers from the FCM data by extracting statistical features from cell-level expression patterns and comparing them between samples [30, 15, 5, 13, 58, 62]. The identified biomarkers are selected to be statistically different between cohorts but may not be biologically meaningful cell populations with distinct phenotypes. These methods focus on identifying cohort-level differences and are not designed for or dependent on the accurate identification of (cancer) cells in each individual sample. A third category of methods makes use of representation learning models, such as neural networks, to bypass the feature extraction step. Some of these methods are shown to be effective in predicting the cancer diagnosis [37, 24] but the predicted diagnosis is difficult to interpret and validate without identifying the cancer cells themselves. Another group of methods in this category has been applied to non-neoplastic diseases for predicting the sample diagnosis while identifying the diagnostic cell populations [16, 3, 17]. However, the identified cell populations were not validated due to the lack of cell-level labels and it remains unclear whether these methods can effectively identify cancer cell populations with phenotypic heterogeneity.

Here we define the problem to be solved as follows. Given a set of preexisting clinical FCM samples with diagnostic labels, with cancer burden being optionally available (as identified by expert manual or automated gating analysis), can we predict the diagnosis label of a new FCM sample from the same reagent panel and identify the diagnostic cells simultaneously, while retaining interpretability of the identified cancer cells. Addressing this problem requires the simultaneous optimization of cell population identification (biomarkers) and sample-level classification (diagnosis). In previous work [21], we showed that the simultaneous optimization can be achieved for diagnosis of chronic lymphocytic leukemia (CLL), by adapting gradient descent optimization for identifying a global optimal gate to maximize classification accuracy. In this paper, we hypothesize that a density-based discriminative point set model using backpropagation can address the simultaneous optimization, without requiring initial gating.

Specifically, we developed an end-to-end differentiable representation learning approach - Cell Scoring Neural Network (CSNN) - that learns the density distribution of the cellular expressions on all markers and makes diagnostic predictions based on aggregating cell-level scores into sample-level predictions. In parallel, the sample-level information, including the diagnostic labels and the density patterns, is backpropagated to the cell level for identifying the diagnostic cell population(s). Based on the possible availability of cancer burden information, we developed two versions of CSNN: CSNN-Class (classification), which requires only diagnostic labels in the training data and CSNN-Reg (regression), which makes use of the additional sample-level cancer burden information to improve the prediction at the single cell level. We applied both CSNN modeling methods on two independent datasets for diagnosis of CLL (provided by University of California, San Diego) and B-ALL (B-cell lymphoblastic leukemia, provided by Stanford University). We compared the performance of the resulting models with two relevant representative deep learning modeling approaches, cellCNN [3] and DeepCellCNN [17],

assessing their clinical utilities regarding: a) accuracy of diagnostic prediction, b) interpretability of the identified leukemic cell populations and their phenotypic heterogeneity, and c) accuracy of the identified cancer burden. To confirm that the CSNN-identified diagnostic cell populations are those that differ between the cancer and non-cancer samples, we designed and performed independent data clustering analysis to identify the clusters of cells that can only be found from the cancer samples and compared them with the CSNN-identified cancer cells. To interpret the CSNN-identified phenotypic heterogeneity of the cancer cells, we constructed post hoc decision trees based on the predicted cell-level labels, enumerated all leukemic cell phenotypes along the tree paths, and compared them with the known cancer endotypes observed clinically in patient cohorts. The leukemic cell populations identified in individual samples are then visualized in traditional 2D dot plots for straightforward hematopathology review.

2 Methods

2.1 Notation

We consider N individuals (e.g., patients) where each individual $i, 1 \leq i \leq N$, is represented by an FCM sample X_i. Each sample X_i consists of a set of multi-dimensional vector measurements, where each vector $x_{i,j}$ in the sample corresponds to a single cell, i.e., $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$, where j is an index of cells in sample X_i, n_i is the number of cells in sample X_i, and each dimension corresponds to the expression of an FCM marker.

We assume a target y_i is available for each subject i, provided (for example) by human experts based on manual evaluation of the FCM sample X_i. In this paper, we will consider two different types of targets y_i. The first type of target is a real-valued target y_i, taking values between 0 and 1, indicating the disease burden for sample i, i.e., the proportion of cells that are estimated to be pathogenic for that sample, with $y_i = 0$ for healthy samples and $y_i > 0$ for samples diagnosed with the disease condition. The second type is a binary target y_i, taking the value 0 or 1, indicating a disease diagnosis for sample i, and where $y_i = 1$ indicates disease presence for sample i.

At the cell level, let $z_{i,j}$ be a cell-level binary variable where $z_{i,j} = 1$ indicates that the cell is pathogenic and $z_{i,j} = 0$ indicates that a cell is non-pathogenic (present in healthy conditions). We will assume that cell-level labels are not available in the training data, i.e., that the $z_{i,j}$ value for cell j for patient i is unknown.

An important aspect of our overall approach is to be able to predict, for patient i, both the cell-level binary variables $z_{i,j}$ and the sample-level disease diagnosis y_i. More specifically, we estimate both:

1. **cell-level scores** $s_{i,j}$, in the form of conditional probabilities $s_{i,j} = P(z_{i,j} = 1|x_{i,j})$, i.e., the probability that a particular cell is pathogenic, given marker measurements $x_{i,j}$; and
2. **sample-level** real-valued burdens y_i where these sample-level estimated burden is a functions of the cell-level scores $s_{i,j}$.

2.2 A Cell-Scoring Neural Network for Disease Prediction

Our goal is to construct a predictive model that takes a set of cell-level vectors for a sample, $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$, and produces a sample-level prediction of y_i (either a real-valued burden or a probability of a binary label). A challenge in this context is that predictive modeling techniques in statistics and machine learning typically assume a fixed-dimensional vector representation as input to a model, rather than sets of vectors X_i of varying sizes across i.

To handle this issue, we use the following two-step approach. In the first step we map each cell-level vector $x_{i,j}$ to a scalar-valued cell-level conditional probability score $s_{i,j} = P(z_{i,j} = 1|x_{i,j})$ where the mapping $s_{i,j} = s(x_{i,j}; \phi)$ has learnable parameters ϕ, parametrized via a feedforward neural network, which we refer to as a Cell Scoring Neural Network (CSNN). Using this cell-level mapping, each sample $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ can then be represented by a set of cell-level scores $S_i = \{s_{i,1}(\phi), \ldots, s_{i,n_i}(\phi)\}$, where each score indicates how likely it is that a particular cell i, j is pathogenic. Note that the cell-level scores $s_{i,j}(\phi)$ depend implicitly on the cell-level data vectors $x_{i,j}$; we suppress this dependence on $x_{i,j}$ in the notation for simplicity.

In the second step, to predict real-valued burden targets y_i, we aggregate the cell-level scores by averaging, i.e., $\hat{y}_i(\phi) = \bar{s}_i(\phi) = \frac{1}{n_i} \sum_{j=1}^{n_i} s(x_{i,j}; \phi)$, representing an estimate of disease burden for sample X_i. To predict a

binary target we define $P(y_i = 1|X_i)$ to be a logistic function, i.e.,

$$P(y_i = 1|X_i) = P_{\alpha,\beta}(y_i = 1|\bar{s}_i(\phi)) = \frac{1}{1 + exp(\alpha + \beta\bar{s}_i(\phi))}$$

where α and β are learnable parameters of the logistic function.

Note that the two types of predictions have different interpretations. The burden prediction y_i can in practice take values quite close to 0 for patients who have a disease diagnosis (e.g., a patient could have as few as 0.01% pathogenic cells and still have the disease). On the other hand, the conditional probability estimate, $P(y_i = 1|X_i)$, can be interpreted as having a threshold at 0.5, i.e., if $P(y_i = 1|X_i) > 0.5$ then individual i is more likely to have the disease than not (and vice-versa). The logistic parameters allow for accommodation of this difference between predicting burden level and predicting likelihood of disease presence.

A key feature of our approach is that we use information at the sample-level (the y_i's) to learn the mappings for the cell-level scores (the $s_{i,j}$'s). In particular, for real-valued burden targets we pursue a regression approach and minimize a weighted mean-square error loss function:

$$L_{MSE}(\phi) = \sum_{i:y_i>0} \big(\bar{s}_i(\phi) - y_i\big)^2 + \lambda \sum_{i:y_i=0} \bar{s}_i(\phi)^2, \tag{1}$$

where $\lambda > 1$ is a hyperparameter that upweights the second term to encourage the model to push the predictions (burden estimates) for healthy individuals to be close to 0.

For binary targets $y_i \in \{0, 1\}$, we estimate the parameters by minimizing the standard binary cross-entropy objective function used in classification modeling [14]:

$$L_{LL}(\phi, \alpha, \beta) = -\sum_{i:y_i=1} \log P_{\alpha,\beta}(y_i = 1|\bar{s}_i(\phi)) - \sum_{i:y_i=0} \log\big(1 - P_{\alpha,\beta}(y_i = 1|\bar{s}_i(\phi))\big). \tag{2}$$

We learn ϕ for L_{MSE} (and simultaneously, ϕ, α and β for L_{LL}) by using standard gradient descent optimization methods. In what follows, we refer to the first approach above (with real-valued burdens and squared error functions) as CSNN-Reg (for cell scoring neural network regression), and the second approach (with binary labels and log-loss functions) as CSNN-Class (for cell scoring neural network classification).

To represent the cell-level mappings, $x_{i,j} \to s_{i,j}(\phi)$, for both CSNN-Reg and CSNN-Class, we use a flexible function approximator in the form of a multi-layer feedforward neural network. In particular we use a ReLU activation function in the intermediate hidden layers and a sigmoid (softmax) function, $g(z) = 1/(1 + \exp(-z))$ as the activation function at the output layer so that the model's output per cell is constrained to lie between 0 and 1. Additional details on network architectures and hyperparameter settings for optimization are provided in *Appendix 1*.

CellCNN [3] and DeepCellCNN [17] are two existing methods that are comparable to a CSNN since they use neural networks that produce multiple scores per cell, which are then aggregated for sample-level classification. In these methods, the scoring neural network is replaced by a neural network that maps every cell to a vector in a latent space, rather than a single score. These vectors are then averaged together, into a final feature vector which is an abstract representation of the sample. This feature vector serves as a summary of the samples, but is difficult to interpret by humans.

In contrast, a distinct advantage of the interpretable cell-level scores produced by the CSNN models is that for any cell, for a particular sample, we can explore and interpret where pathogenic cells are located in marker space, e.g., by visualizing and highlighting what regions in marker space have scores $s_{i,j} = s(x_{i,j}; \phi)$ above a particular threshold. We discuss how this is related to the concept of manual "gating" in *Appendix 2* and provide illustrations of how this can support biologically-meaningful discovery with real FCM datasets later in the paper.

2.3 Initializing a CSNN Models using Density Estimates

The quality of the learned CSNN models can be improved by using information related to the densities in marker space to initialize the models. We begin by generating initial estimates of cell-level scores $s'_{i,j}$ for each cell j in each sample i in the training data. Consider first the case of real-valued y_i targets (i.e., sample burdens). For

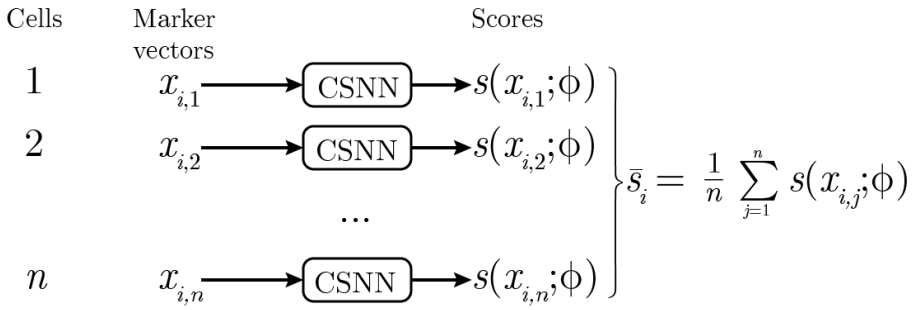


Fig. 1: **Sample-level predictions from cell-level scores** - Each marker vector is evaluated by the CSNN and given a score $s(x_{i,j};\phi)$ that corresponds to $P(z_{i,j}=1|x_{i,j},\phi)$. The average of these scores, $\bar{s}_i$, becomes the prediction for sample i. In CSNN-Reg, these scores serve as the output. For CSNN-Class, the scores are rescaled to determine a threshold for the classification.

samples with $y_i = 0$, all cell-level scores are zero by definition, i.e., $s'_{i,j} = 0$, assuming that all cells in healthy samples are non-pathogenic. For samples with a disease diagnosis ($y_i > 0$) Bayes' rule is used:

$$\begin{aligned} s'_{i,j} &= P(z_{i,j}=1|x_{i,j},y_i>0) && (3)\\ &= 1-P(z_{i,j}=0|x_{i,j},y_i>0) && (4)\\ &= 1-\frac{P(x_{i,j}|z_{i,j}=0,y_i>0)P(z_{i,j}=0|y_i>0)}{P(x_{i,j}|y_i>0)}. && (5)\end{aligned}$$

We can estimate each of the three terms on the right hand side from the training data as follows. The term in the denominator, $P(x_{i,j}|y_i>0)$, is the marginal probability density function (PDF) of marker measurements for sample i, given that sample i has a disease diagnosis: this PDF can be estimated straightforwardly using kernel density estimation (KDE). The first term in the numerator, $P(x_{i,j}|z_{i,j}=0,y_i>0)$ is the probability density for non-pathogenic cells in a sample with a disease diagnosis. We can approximate this by assuming that $P(x_{i,j}|z_{i,j}=0,y_i>0) = P(x_{i,j}|z_{i,j}=0) = P(x_{i,j}|z_{i,j}=0,y_i=0)$, i.e., that the PDF of non-pathogenic cells in marker space is the same in both positive and negative samples. We further assume that the density $P(x_{i,j}|z_{i,j}=0,y_i=0)$ does not vary from sample to sample, allowing us to pool all $x_{i,j}$ measurements from all the negative samples (which have $y_i=0$ and $z_{i,j}=0$ by definition) and again use KDE to estimate this density. The second term in the numerator, $P(z_{i,j}=0|y_i>0)$, is equal to $1-y_i$, under the assumption that y_i (the sample burden) corresponds to the fraction of cells in sample i that are pathogenic. For the situation where the training data only contains binary labels $y_{i,j}\in\{0,1\}$, i.e., the CSNN-Class model, we again estimate scores $s'_{i,j}$ via Bayes rule, but using additional approximations for each of the three required terms in the absence of known burdens y_i.

Finally, given scores $s'_{i,j}$ from the density-based approach above (for all cells for all samples in the training dataset), a feedforward neural network, parametrized by weights ϕ', is trained to create an initial cell-level model that can approximate the density-based scores via the neural network. The trained weights ϕ' of this neural network are then used to initialize the weights in training of an end-to-end CSNN network (CSNN-Reg or CSNN-Class, using L_{MSE} and L_{LL} respectively as described earlier). We found in practice that this density-based initialization significantly improves the quality of the final sample-level disease predictions. Full details on KDE methods, approximations for binary y_i labels, and training of the initial network, are provided in the Supplement.

Note that one could in principle use the density-based approach alone to build a sample-level prediction model, by using the density-based models from the training to generate scores $s'_{i,j}$ per cell for a new sample. A prediction for y_i could then be based, for example, on thresholding the sum of cell-level scores $\sum_j s'_{i,j}$. However, a purely density-based approach may be sensitive to modeling assumptions, whereas the sample-level discriminative training of the CSNN can allow the model to further tune the initial parameters ϕ' to produce scores $s_{i,j}$ that are directly optimized for robust prediction of burden or likelihood of disease.

3 Experiment Results

3.1 Datasets

Two independent FCM datasets were used in evaluating the performance of the CSNN modeling methods. The first dataset (DS1) was provided by the University of California, San Diego (UCSD) Center for Advanced Laboratory Medicine (CALM) diagnostic lab that was collected and analyzed for the identification of Chronic Lymphocytic Leukemia (CLL) cases using their standard diagnostic protocol. DS1 includes FCM data from 288 subjects - 186 diagnosed as CLL and 102 judged to be non-CLL by the hematopathologist. For each subject, two reagent panels, PB1 and PB2 were used in the clinical FCM assay on peripheral blood (PB) samples. Each panel contained antibodies for the detection of 10 markers (fluorescence parameters): PB1: CD3, CD5, CD10, CD19, CD22, CD38, CD43, CD45, CD79b and CD81; PB2: Anti-Ig-lambda, Anti-Ig-kappa, CD5, CD7, CD19, CD20, CD23, CD38, CD49d and FMC-7. Datasets from each panel also included 6 scatter parameters: forward scatter (FSC)-area(A)/height(H)/width(W) and side scatter (SSC)-A/H/W.

The second FCM dataset (DS2) was provided by the diagnostic lab at Stanford University (Stanford) for the identification of leukemic cells (blasts) of B-cell Acute Lymphoblastic Leukemia (B-ALL) using their standard diagnostic protocol. The samples in DS2 are bone marrows, which consist of both diagnostic samples and samples collected from patients who have received CD19-targeted CAR (chimeric antigen receptors) T-cell therapy. DS2 includes FCM data from 178 subjects - 50 diagnosed as B-ALL and 128 judged to be non-B-ALL. The FCS files of DS2 are from one reagent panel detecting 9 markers: CD66b, CD22, CD19, CD24, CD10, CD34, CD38, CD20, CD45, and 2 scatter parameters (FSC/SSC).

Research on both datasets was approved by Institutional Review Boards of the respective institutions (UCSD and Stanford). Both datasets have been fully de-identified before being transferred and analyzed using the proposed neural networks. Diagnostic labels of DS1 samples were provided by UCSD. Cancer burden of each DS1 sample was obtained by applying the DAFi automated gating analysis [26] on the de-identified FCS files, following the gating strategy used in the diagnostic lab (*Supplementary Figure 1*). Specifically, DAFi examined the property of all markers, focusing on identifying the $CD5^{+}CD19^{+}CD10^{-}CD79b^{dim}$ CLL cells in PB1 using data clustering, according to our previous study [49]. For DS2, both the diagnosis and the cancer burden were provided by Stanford from expert manual gating analysis.

3.2 Performance assessment

3.2.1 Training and testing sets

To assess the performance of the ML modeling approach developed, each dataset was divided into separate subsets for training, validation, and testing. The CLL dataset (DS1) is divided into a training set of size 102, a validation set of size 81, and a testing set of size 105. For the B-ALL dataset (DS2), 118 out of the 178 samples were selected at random and reserved as testing samples; the models were trained with the remaining 60 samples. Using the 60 training samples, hyperparameter values were optimized by running 5-fold cross validation runs.

3.2.2 Quantitative assessment of classification accuracy

The performance of CSNN-Class and CSNN-Reg methods on the CLL and B-ALL datasets were compared with with two machine learning methods recently reported in the literature, CellCNN and DeepCellCNN. For each method hyperparameter optimization was performed by learning parameters on the training subsets and selecting the best model parameter setting based on the area under the receiving operating characteristic curve (AUROC) obtained with the validation subsets. We then retrained the best performing model on a sample containing all the samples in the training set and validation set and determined the final performance on unseen test subsets.

Hyperparameters evaluated for all models included the learning rate and an architecture search, with specific searches for w (CSNN-Class), λ (CSNN-Reg) and the dropout rate (CellCNN). The specific grid values tested for these can be found in the *Appendix 1*. Using the best hyperparameters for each model, we then trained 20 more models of each type with different random initializations to measure their variance with respect to initialization. In order to filter out all the models that initialized incorrectly we ran each training loop with 5 restarts and evaluated them with the testing set, as described in Fig 2. We then picked the best run out of those 20 to report the ROC graphs in Fig 2. Overall, the proposed CSNN-Reg and CSNN-Class produce higher AUROC scores than either CellCNN or DeepCellCNN. CSNN-Reg is superior to all methods on both datasets, indicating that the sample-level burden information (as used by this method) carries additional value beyond sample-level binary

labels (as used by the other three methods). In addition, both of the CSNN models are more robust than the other methods in that they have lower variance (than the other methods) across weight initializations during model training.

In addition to performance on sample-level diagnosis, the CSNN-Reg method was also assessed for its ability to determine leukemic cell proportions on both the B-ALL and CLL datasets. Excellent correlation between the predicted proportions and the proportions reported by the diagnostic lab is observed for the CLL dataset (Fig 3). A similar correlation between predicted and reported leukemic cell proportions was observed for the B-ALL dataset. However, it should be noted that the set of samples that were randomly picked for testing did not contain any samples with proportions in the 0.2-0.8 range and therefore the performance of CSNN-Reg on the B-ALL dataset in this range is determined by interpolation. Ablation tests were conducted to evaluate whether the method performance is improved by the density difference initialization and the post initialization fine-tuning. Results of the ablation tests can be found in *Appendix 3*.

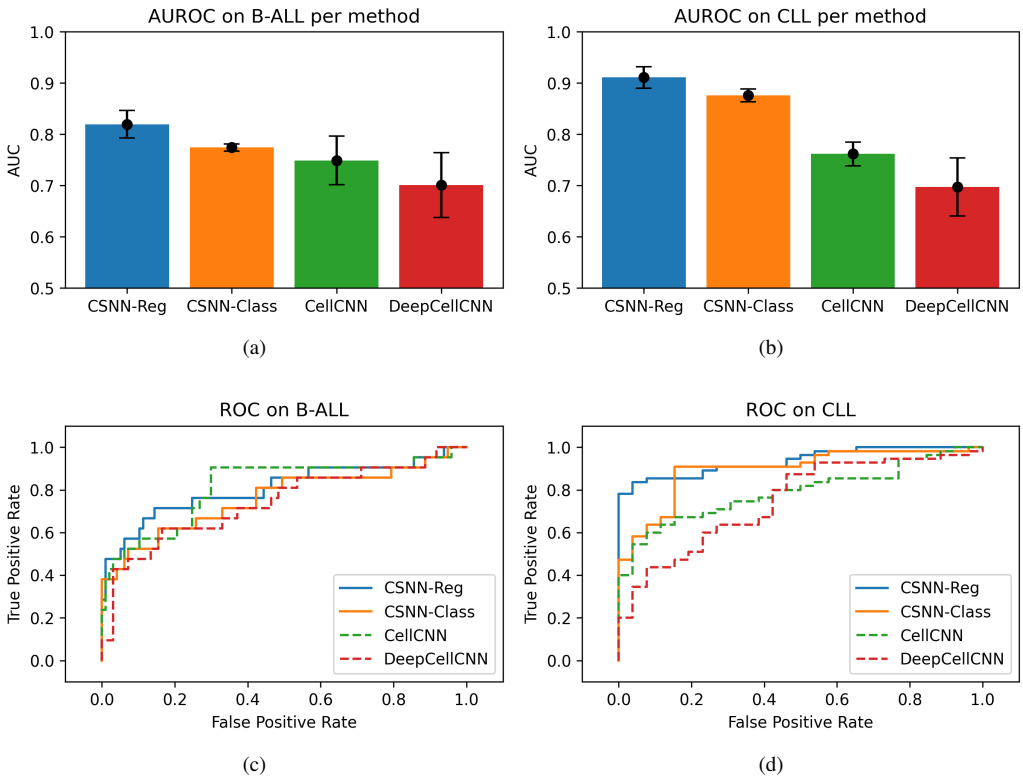


Fig. 2: **Model performance of sample-level leukemia classification** - (a, b) The average of the testing scores for each model class on the B-ALL (a) and CLL (b) datasets, with the error bars representing the standard deviation between the N testing scores. (c,d) The ROC curve for the model with the highest scoring training AUC for each model class out of the N tests on the testing set of the B-ALL (c) and CLL (d) datasets.

3.2.3 Biological interpretation of the identified cancer cells

The CSNN-identified pathologic cells are highlighted on the key 2D dot plots for visual examination and interpretation. We use the term "pathologic" to refer to cell populations that are related to the leukemic state, which can include both the leukemic cells themselves and any reactive “normal” cell population elicited by the presence of a leukemia in the patient that could be equally diagnostic and prognostic. Visual examination of the distributional shapes of antigen expression and locations of the identified pathologic cell populations in 2D dot plots was used to determine: a) if each CSNN-identified cell population has a natural shape on the 2D plot with a unimodal distribution of each marker, b) if the location of each CSNN-identified pathologic cell population matches a known leukemic cell phenotype, and c) if all leukemic cell populations seen on the 2D plots are successfully identified by the CSNN models.

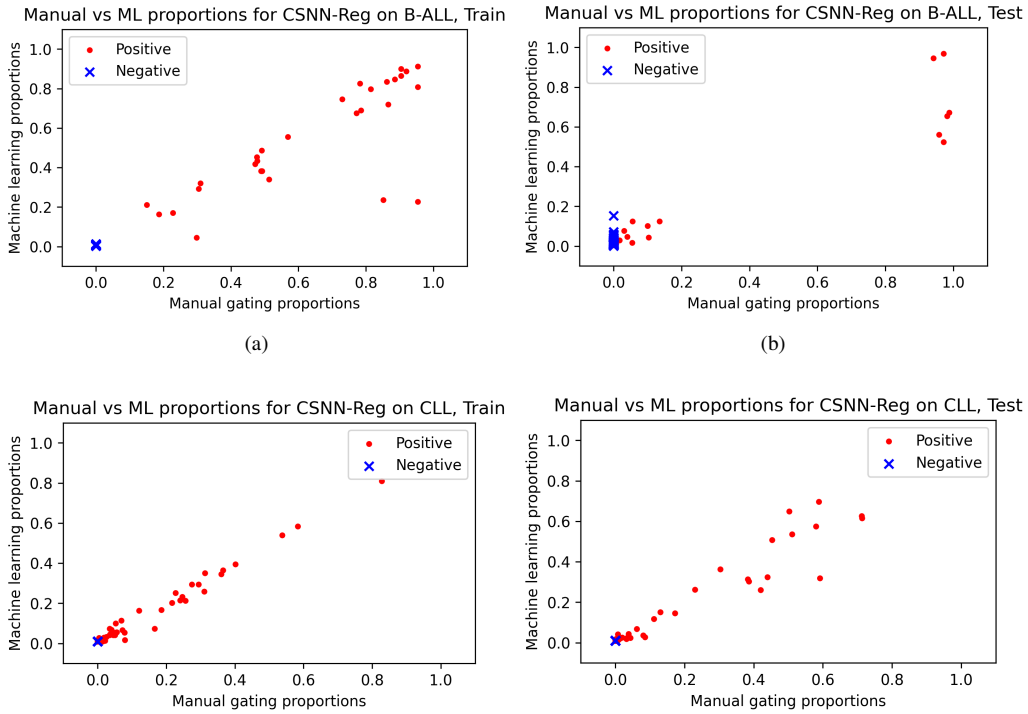
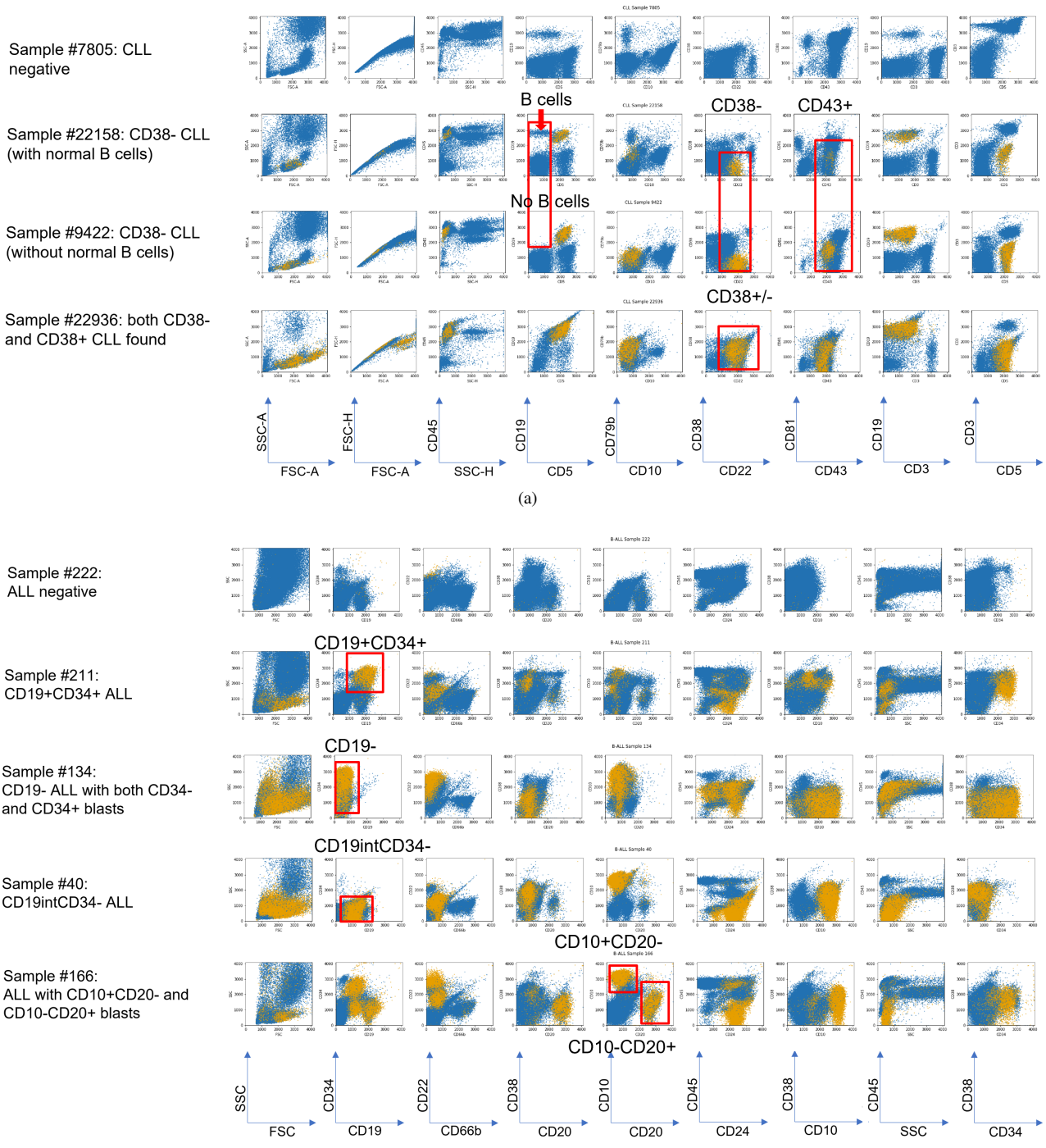


Fig. 3: **Comparison of machine learning- and manual gating-derived leukemic cell proportions** - Using the best finetuned version of CSNN-Reg, the proportion of leukemic cells produced by the best finetuned version of SCNN-Reg and expert manual gating are compared the B-ALL (a, b) and CLL (c, d) training (a, c) and testing (b, d) data subsets.

Fig 4a shows the pathologic cells (in yellow) identified in the CLL dataset by the CSNN-Reg model from a few representative samples (results for all CLL positive samples can be found in *Supplementary Figure 2*). The CLL cell populations are highlighted across nine 2D dot plots that cover all the surface protein markers used in the reagent panel. This representative sample set consists of a negative case (row #1), positive cases with (row #2) and without (row #3) normal B cells, as well as CD38-negative (rows #2-3) and a mixture of CD38-negative and CD38-positive (row #4) CLL cases to illustrate within and between sample phenotypic heterogeneity. Many studies have previously reported the important role of CD38 in CLL prognosis [50, 42, 32, 35]. Fig. 4a clearly shows the capability and accuracy of the proposed CSNN model for identifying these important CLL phenotypic endotypes. Without being informed that the typical CLL phenotype is $CD5^{+}CD19^{+}$, the major pathologic cell populations identified by the CSNN-Reg were found to be $CD5^{+}CD19^{+}$. Without using clustering analysis to define cell populations up front, CSNN still successfully identified the cell populations with natural antigen expression distributional shapes and did not mix the normal $CD19^{+}$ B cells with the $CD5^{+}CD19^{+}$ CLL cells (row #2), which can be difficult for traditional gating methods to cleanly separate. We also observed that the CSNN models do not require a prefiltering step needed by some existing methods to filter out debris/dead cells/doublets, such that any cell subsets in individual samples that differ between the CLL and non-CLL cohorts (e.g., the doublets in row #4, highlighted on FSC-A vs FSC-H) can be identified and used for classification.

The same visual assessment performed on the B-ALL dataset, highlighting the CSNN-identified pathologic cells from 5 representative samples (Fig 4b), including a negative sample (row #1), a typical $CD19^{+}CD34^{+}$ B-ALL sample (row #2), a $CD19^{-}$ B-ALL sample (row #3, probably collected after CD19-targeted CAR-T therapy), an atypical $CD19^{int}CD34^{-}$ B-ALL case (row #4), and a B-ALL sample with at least 3 subtypes of blasts (row #5). Results from all B-ALL positive samples on CD19 vs CD34 can be found in *Supplementary Figure 3*. Expression of CD19, CD34, CD10, and CD20 in the B-ALL samples illustrates the phenotypic heterogeneity observed. These B-ALL phenotypic endotypes can be extremely challenging to identify using manual gating analysis due to this sample-to-sample heterogeneity. The CSNN model identified these phenotypically heterogeneous B-ALL cells at the single cell level by comparing all of the individual samples in the B-ALL positive and B-ALL negative cohorts, without requiring cell-level labels in the training data.



(a)

(b)

Fig. 4: **Visualization of the identified pathologic cells by CSNN-Reg** - Pathologic cells identified in representative samples from the CLL (a) and B-ALL (b) datasets in nine 2D dot plots are colored yellow; non-pathologic cells are colored blue. Typical CLL cells are $CD5^{+}CD19^{+}$, while typical B-ALL cells are $CD19^{+}CD34^{+}$. Natural shapes of the identified pathologic cell clusters are produced by the CSNN model, without using clustering analysis. Phenotypic heterogeneity of the B-ALL cells can also be seen within and across the samples

3.2.4 Qualitative assessment and visual comparison of the results identified across competing methods

The results of the four different methods (CellCNN, DeepCellCNN, CSNN-Reg, and K-means) were visualized and compared for their identification of cancer cell phenotypes. The K-means clustering approach is an ad hoc independent, but fully interpretable, way of identifying cell clusters that can only be found in the cancer samples (method design can be found in *Appendix 4*). The complete set of 2D dot plots for visual comparisons of results of CSNN-Reg with the two baseline methods on all CLL and B-ALL samples can be found in *Appendix 5*. Fig. 5

shows the results of CellCNN (row #1), DeepCellCNN (row #2), CSNN-Reg (row #3), and K-means (row #4) on selected CLL (case #22936, Fig. 5a) and B-ALL (case #134, Fig. 5b) cases across nine different 2D dot plots (columns). The 2D plots of the baseline methods were generated using a probability cutoff = 0, because a cell should not be counted as a non-leukemic cell if it increases the likelihood of the sample being positive, and vice versa. Fig. 5 shows that both CellCNN and DeepCellCNN identified few leukemic cells using the probability cutoff = 0, indicating that the probability values output by the baseline methods cannot be directly used to predict whether a cell is leukemia-related or not, which is not too surprising, given that these two baseline methods were not designed for cell-level classification for blood cancer diagnosis.

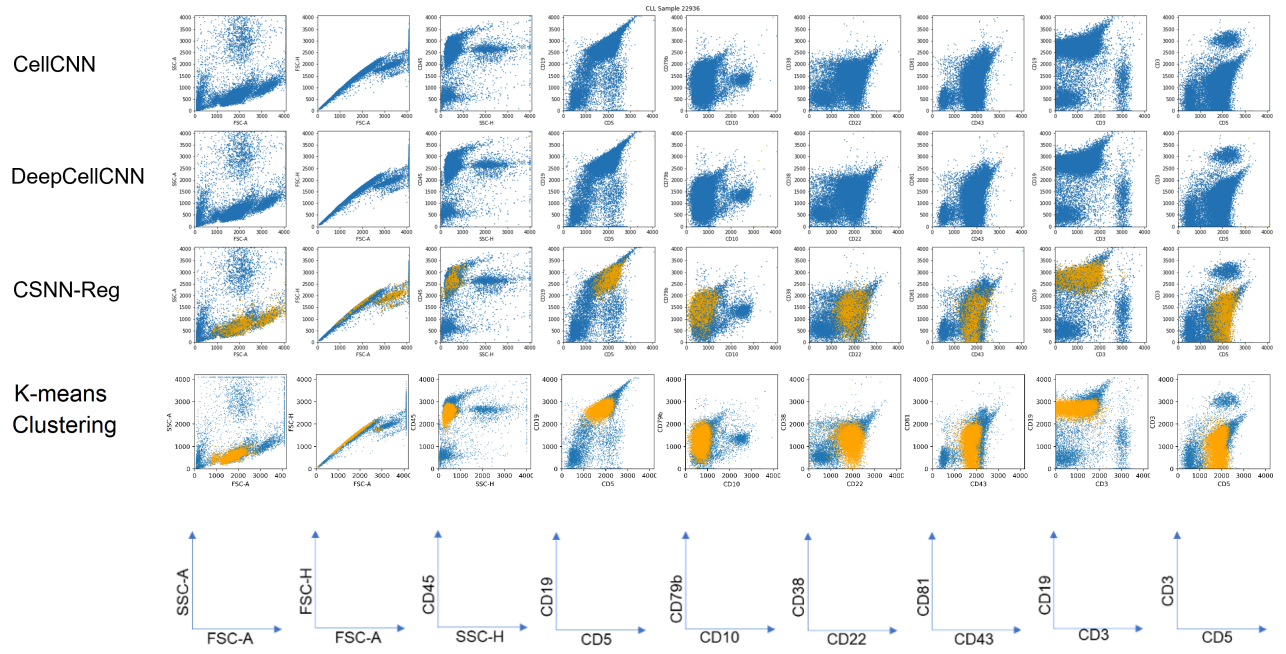
3.2.5 Interpretation of the neural network classification model using decision trees

In order to understand how CSNN was able to identify the heterogeneous leukemia-related cell subsets in individual samples, all B-ALL samples were pooled to construct a global decision tree to illustrate the classification paths of the cells in the pooled sample, based on the CSNN-output labels at the single cell level. Trees were then generated for each individual sample by calculating the cell-level statistics of the sample following the classification structure of the global tree, which preserved the tree layout for result comparison and interpretation across individual samples.

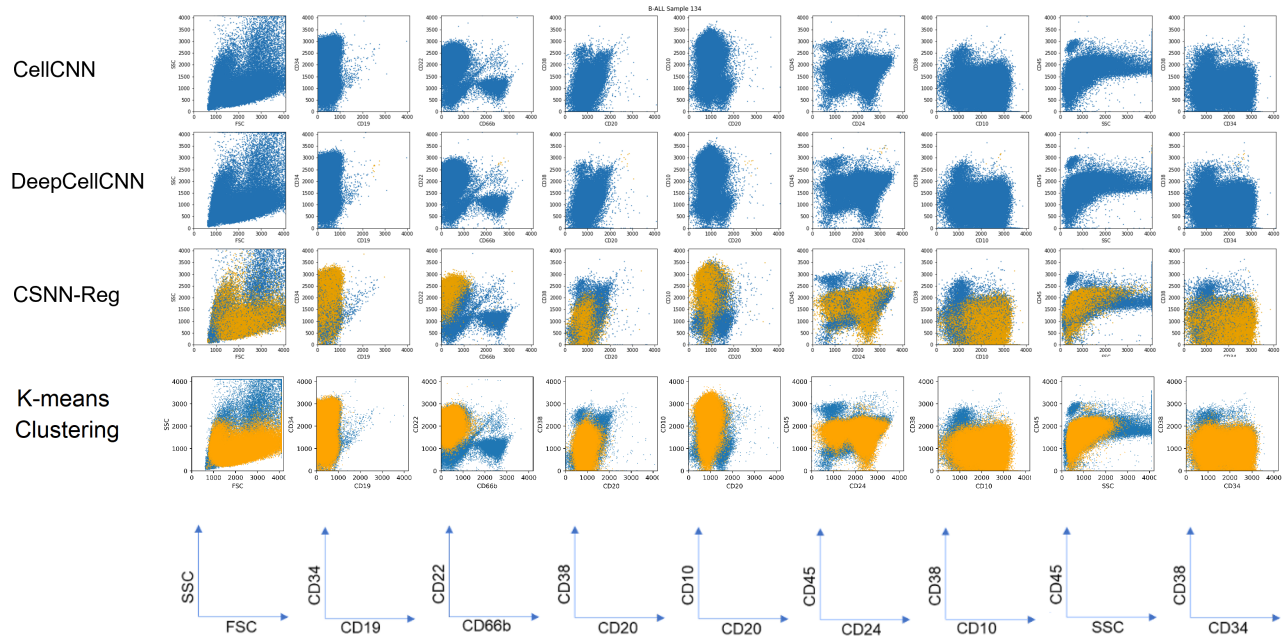
Using decision trees to interpret neural network analysis results is not new and has been discussed previously [11, 17]. However, Fig. 6 shows that a tree-based classifier can be adapted for not only interpreting the sample-level classification but also illustrating the phenotypic heterogeneity of B-ALL cells in individual patient samples. Three representative B-ALL positive samples were selected for visualizing the tree-based classification paths derived from the CSNN-identified cell-level labels side by side with the 2D dot plots that highlight these B-ALL cells (Fig. 6): case #134 (top) is a $CD19^-$ B-ALL example, case #211 (middle) is a $CD19^+$ B-ALL example, and case #166 (bottom) is an example that contains a mixture of CD19 negative and positive leukemia-related cells. Each node in the tree corresponds to a protein marker used in the reagent panel. The root of the tree is automatically selected during the tree construction process as the most informative feature for classification. In the pooled B-ALL sample, CD19 was identified as the root of the tree, which matches with our understanding of the patient cohort, in which some have been treated with the CD19-targeted CAR-T therapy and therefore the leukemic cells that have remained following therapy have lost expression of CD19. Our decision tree model, derived from the CSNN output at the single cell level, successfully identified the two major B-ALL subtypes in the patient cohort: naive $CD19^+$ B-ALL and treatment-related $CD19^-$ B-ALL.

Indeed, each path in the tree-based model leading to a B-ALL positive leaf node corresponds to a distinct B-ALL phenotype, potentially defining B-ALL endotypes (Fig. 6). In case #134, while all of the B-ALL cells are CD19 negative, the $CD19^-$ B-ALL cell phenotypes can be further subdivided based on CD10, CD34, CD20, and CD66b expression. For case #121, while all of the B-ALL cells are CD19 positive, they can be further subdivided based on CD45 and CD38 expression, and SSC characteristics. The B-ALL cells in case #166 consist of three major subtypes: $CD19^-CD10^+CD66b^-$, $CD19^+CD45^-CD38^-$, and $CD19^+CD45^+CD34^-$. Navigating along the decision tree provides an exploratory capability of identifying both known and novel leukemia-related cell phenotypes in individual samples, in a data-driven exhaustive way.

Finally, each tree path can also be interpreted as a manual gating strategy, with the marker expression cutoff values identified at the tree nodes defining the gating boundaries in the original marker space. In case #134, 1270 on CD19 is the cutoff for dividing the cells into $CD19^+$ and $CD19^-$. Visualization of the 2D dot plots confirms that the CD19 cutoff follows the natural boundary of the CD19 expression distribution. Similarly, the cutoffs of 2322 on CD10 and 1010 on CD66b for case #134 also follow the data distribution for separating positive from negative cells. In case #211, 2246 on CD38 derived from the CSNN output seems a perfect global gating cutoff for dividing the cells into CD38 negative and CD38 positive. Importantly, CSNN successfully calculated the local gate (the B-ALL cell population is highlighted in yellow), without relying on the global cutoff. The interpretation of these results is two-fold. First, the cell-level labels output by CSNN can be used to derive an accurate global cutoff for visual interpretation and validation. Second, the sample-level CSNN results provide for sample-specific cell classification that may deviate from the global cutoff identified in the pooled decision trees. This suggests that it could be very challenging to identify the B-ALL cell populations across all samples using traditional manual gating analysis. The same phenomenon was observed in case #166 on CD19 where the data and plots clearly show that there exists phenotypic heterogeneity of B-ALL within and between individual samples. While a global cutoff can be precisely identified based on the pooled data, the leukemic cells



(a)



(b)

Fig. 5: **Visualization of pathologic cells identified by the different modeling methods** - Leukemic cells identified from positive CLL case #22936 (a) and B-ALL case #134 (b) by CSNN-Reg (row #3) versus two baseline methods CellCNN (row #1) and DeepCellCNN (row #2) and an independent data clustering analysis using K-means (row #4) are highlighted in yellow with the rest of the cells in blue. Neither CellCNN nor DeepCellCNN could identify leukemic cells under their default settings.

in individual samples needed to be identified using a "local gate" as predicted by CSNN. In clinical practice, the cutoffs and 2D dot plots output by CSNN along with the tree-based classification paths can be combined and converted into manual gating strategies for explainable validation by hematopathologists.

3.3 Additional findings

When examining 2D dot plots across individual samples, we noticed that CSNN was able to identify the leukemia-related cells with distinct atypical phenotype, which could be useful for cancer precision medicine. *Supplementary*

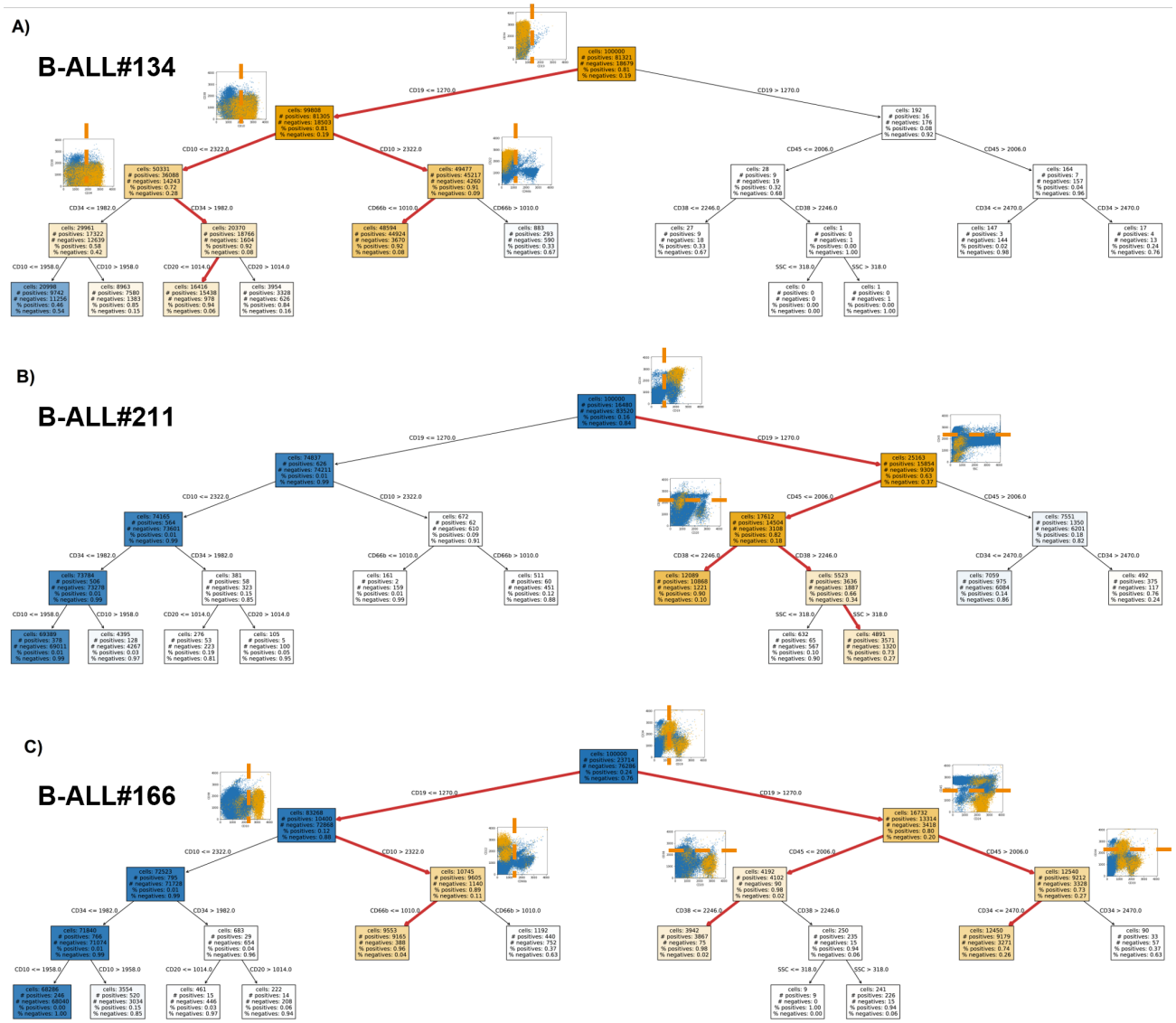


Fig. 6: Neural network model interpretation using decision trees. 2D dot plots show the important tree nodes and highlight the leukemic cells identified and the expression cutoffs of the corresponding markers. Heterogeneity of B-ALL can be clearly seen in A) $CD19^-$ B-ALL found in sample #134, B) $CD19^+$ B-ALL found in sample #211, and C) both $CD19^-$ and $CD19^+$ B-ALL can be found in sample #166. The orange dotted lines in the 2D plots indicate the marker expression cutoffs identified by the decision tree classifier. Conceptually, each path highlighted in red can be thought of as corresponding to a traditional manual gating sequence with marker expression cutoffs determined in a data-driven manner.

Figure 4 shows B-ALL sample #40 side by side with other five other B-ALL samples on plots of CD34 vs CD38, in which few typical $CD34^+$ B-ALL cells are observed in Sample #40. However, the hematopathology report listed that the cancer burden of Sample #40 is 86.19%. *Supplementary Figure 5* compares sample #40 with the typical B-ALL case #211, where CSNN identified an atypical $CD19^{int}CD34^-$ leukemia-related cell population from sample #40 at 83.8%, matching the hematopathology review result. cellCNN and DeepCellCNN could not identify this cell subset in Sample #40.

Another finding from reviewing the 2D dot plots is that CSNN was able to capture cell populations with natural protein expression distributions, similar to what unsupervised clustering analysis can do. In contrast, manual gating analysis, decision-tree classification, and statistical biomarker identification methods often involve abrupt expression cutoffs that do not reflect natural expression gradients. It is important to note that the models produced by CSNN do not generate or rely on any geometric shape of gates but identify the leukemic cell populations as continua. The CSNN models can identify multiple fine-grained (hyper)regions that differ between the cancer and

non-cancer cohorts, which allow the identification of complex classification patterns not easily captured through sequential gating methods.

A third finding is that the experiment results of CSNN show clear improvements in tagging cell-level labels, as it explicitly model whether each cell contributes to a sample-level classification as having leukemia or not. In contrast, other approaches such as DeepCellCNN [17] define the label of a cell as a product of its classification, by amplifying the cell to be 5% of a sample, followed by calculating the difference in the classification likelihood, resulting in a less robust heuristic, as cancer heterogeneity cannot be explained through amplifying the same single cell.

4 Model limitations and future extensions

A challenge that arises when evaluating the performance of these algorithms is the lack of ground truth annotations at the cell level. Although manual gating analysis can generate cell type labels of individual cells, they are only for known cell types and their precise accuracy is questionable due to the subjective manual operation. The only reliable evaluation metric is the classification error of the whole sample, based on the diagnostic labels of each sample. Therefore, to assess cell level classification we rely on visual examination of the cell populations on the original 2D plots in order to confirm that they match the known leukemic cell phenotypes. To improve this situation, we designed an independent data clustering approach (Appendix 4) to identify the cells that can only be found in the cancer samples. This allows us to compare and qualitatively confirm that the pathologic cells identified by CSNN are leukemia related. As CSNN is a probabilistic model, the model assigns a probability to each cell of being leukemia related. For non-leukemic cells, the probability values can be extremely low, usually around 1%; however, they are seldom equal to 0. The aggregated score is an estimator for the expected value of the tumor burden,

$$\bar{s}_i \approx \mathbb{E}_{X \sim p_{\text{data}}} \left[\frac{1}{|X|} \sum_{x_{i,j} \in X} \mathbb{I}(x_{i,j} \in G) \right]. \tag{6}$$

As the probability of a cell being a leukemic cell is usually non-zero, the above equation will return a non-zero value, even when the model has not found any cells that are likely to be related to the leukemia. In this case, a small number of cells may be classified as leukemic using a discrete threshold even for a non-cancer sample, as long as the diagnosis of the sample is correctly predicted by the model.

As the focus is on minimizing a global objective, the training algorithm might overlook small (<1%) populations of cells in favor of correctly classifying the majority of the samples. This issue prevents the model from identifying specific cell populations that are found in only a small number of samples. Similarly, the model will have difficulty classifying samples that have very small numbers of cancer cells, e.g., minimum residual diseases (MRD), especially if these MRD samples are not included during training. A localized training loss objective designed specifically for identifying MRD samples could help solve this problem. The proposed model could also benefit from identifying and modeling cell populations as a hierarchy. Then each cell population, instead of individual cells, could be scored. Most machine learning models are discriminative, without modeling the cell populations in a generative way, as such, can be hard to interpret. By separating each cell into a cluster and then classifying these clusters individually, both performance and interpretability of the model may improve.

An immediate next step is discrepancy analysis. For discrepant predictions for sample diagnosis and cancer burden, we will need to plot the identified leukemic cells on the original sequential gating paths for hematopathology review. For each false positive case, we plan to investigate whether the subject eventually develops leukemia at a later time point when further clinical data are available and approved for research use.

5 Conclusion

The most important feature of the CSNN model is its capability of simultaneously predicting the diagnosis of a sample and identifying pathologic cells, even with phenotypic heterogeneity. Existing machine learning methods for FCM data analysis either are not designed for blood cancer diagnosis or do not identify and validate the cancer cells from individual samples for result interpretation. In order to demonstrate this capability and assess the performance of CSNN, we designed a suite of interpretation and validation approaches for comparing the CSNN

results to independent clustering analysis, known patient endotypes, diagnostic labels, and expert-identified cancer burden, in addition to hematopathology review of the CSNN-identified cancer cells on original 2D dot plots. Using two independent experiments on CLL and B-ALL, we showed the superiority of CSNN over the existing representative neural network modeling approaches for blood cancer diagnosis. The proposed neural network model is generally applicable to other types of discriminative single cell data analysis.

Data Availability

The source code of the CSNN can be downloaded at GitHub: https://github.com/erobl/csnn. The de-identified B-ALL dataset is publicly accessible on FlowRepository under accession FR-FCM-Z6YK. The CLL dataset was converted to TXT format during de-identification, which can be downloaded at GitHub: https://github.com/JCVenterInstitute/DAFi-gating/tree/master/CSNN/CLL_TXT. 2D dot plots for comparing CSNN with other methods can also be found at GitHub: https://github.com/JCVenterInstitute/DAFi-gating/tree/master/CSNN/Comparison_with_CellCNN_DeepCellCNN.

Acknowledgements

We thank Dr. Holden Maecker for connecting us with the Stanford Diagnostic Lab for performance assessment of the proposed models. This work has been partially supported by the FlowGate project (NCATS U01TR001801) funded by the National Center for Advancing Translational Sciences at the U.S. National Institutes of Health and an Investigator Sponsored Study Award (Protocol 68754391) from Becton, Dickinson and Company (BD).

References

[1]ad David Amir, E., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, **31**(6), 545–552.

[2]Aghaeepour, N., Nikolic, R., Hoos, H. H., and Brinkman, R. R. (2010). Rapid cell population identification in flow cytometry data. *Cytometry Part A*, **79A**(1), 6–13.

[3]Arvaniti, E. and Claassen, M. (2017). Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature Communications*, **8**(1).

[4]Boumiza, R., Debard, A.-L., and Monneret, G. (2005). The basophil activation test by flow cytometry: recent developments in clinical studies, standardization and emerging perspectives. *Clinical and Molecular Allergy*, **3**(1).

[5]Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J., and Nolan, G. P. (2014). Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, **111**(26).

[6]Chiaretti, S., Zini, G., and Bassan, R. (2014). DIAGNOSIS AND SUBCLASSIFICATION OF ACUTE LYMPHOBLASTIC LEUKEMIA. *Mediterranean Journal of Hematology and Infectious Diseases*, **6**(1), e2014073.

[7]Diamond, L. W., Nathwani, B. N., and Rappaport, H. (1982). Flow cytometry in the diagnosis and classification of malignant lymphoma and leukemia. *Cancer*, **50**(6), 1122–1135.

[8]Ebo, D. G., Hagendorens, M. M., Bridts, C. H., Schuerwegh, A. J., De Clerck, L. S., and Stevens, W. J. (2005). Flow cytometric analysis of in vitro activated basophils, specific IgE and skin tests in the diagnosis of pollen-associated food allergy. *Cytometry B Clin Cytom*, **64**(1), 28–33.

[9]Finak, G., Frelinger, J., Jiang, W., Newell, E. W., Ramey, J., Davis, M. M., Kalams, S. A., Rosa, S. C. D., and Gottardo, R. (2014). OpenCyto: An open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Computational Biology*, **10**(8), e1003806.

[10]Freeman, S. D., Virgo, P., Couzens, S., Grimwade, D., Russell, N., Hills, R. K., and Burnett, A. K. (2013). Prognostic relevance of treatment response measured by flow cytometric residual disease detection in older patients with acute myeloid leukemia. *J Clin Oncol*, **31**(32), 4123–4131.

[11]Frosst, N. and Hinton, G. E. (2017). Distilling a neural network into a soft decision tree. *CoRR*, **abs/1711.09784**.

[12]Ge, Y. and Sealfon, S. C. (2012). flowPeaks: a fast unsupervised clustering for flow cytometry data via k-means and density peak finding. *Bioinformatics*, **28**(15), 2052–2058.

[13]Greene, E., Finak, G., D'Amico, L. A., Bhardwaj, N., Church, C. D., Morishima, C., Ramchurren, N., Taube, J. M., Nghiem, P. T., Cheever, M. A., Fling, S. P., and Gottardo, R. (2021). New interpretable machine-learning method for single-cell data reveals correlates of clinical response to cancer immunotherapy. *Patterns*, **2**(12), 100372.

[14]Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

[15]Hu, Z., Jujjavarapu, C., Hughey, J. J., Andorf, S., Lee, H.-C., Gherardini, P. F., Spitzer, M. H., Thomas, C. G., Campbell, J., Dunn, P., Wiser, J., Kidd, B. A., Dudley, J. T., Nolan, G. P., Bhattacharya, S., and Butte, A. J. (2018a). MetaCyto: A tool for automated meta-analysis of mass and flow cytometry data. *Cell Reports*, **24**(5), 1377–1388.

[16]Hu, Z., Glicksberg, B. S., and Butte, A. J. (2018b). Robust prediction of clinical outcomes using cytometry data. *Bioinformatics*, **35**(7), 1197–1203.

[17]Hu, Z., Tang, A., Singh, J., Bhattacharya, S., and Butte, A. J. (2020). A robust and interpretable end-to-end deep learning model for cytometry data. *Proceedings of the National Academy of Sciences*, **117**(35), 21373–21380.

[18]Hu, Z., Bhattacharya, S., and Butte, A. J. (2022). Application of machine learning for cytometry data. *Frontiers in Immunology*, **12**.

[19]Irvin, C., Zafar, I., Good, J., Rollins, D., Christianson, C., Gorska, M. M., Martin, R. J., and Alam, R. (2014). Increased frequency of dual-positive TH2/TH17 cells in bronchoalveolar lavage fluid characterizes a population of patients with severe asthma. *Journal of Allergy and Clinical Immunology*, **134**(5), 1175–1186.e7.

[20]Jaimes, M. C., Leipold, M., Kraker, G., ad Amir, E., Maecker, H., and Lannigan, J. (2022). Full spectrum flow cytometry and mass cytometry: A 32-marker panel comparison. *Cytometry Part A*, **101**(11), 942–959.

[21]Ji, D., Putzel, P., Qian, Y., Chang, I., Mandava, A., Scheuermann, R., Bui, J., Wang, H.-Y., and Smyth, P. (2019). Machine learning of discriminative gate locations for clinical diagnosis. *Cytometry Part A*, **97**.

[22]Kaleem, Z., Crawford, E., Pathan, M. H., Jasper, L., Covinsky, M. A., Johnson, L. R., and White, G. (2003). Flow cytometric analysis of acute leukemias. Diagnostic utility and critical analysis of data. *Arch Pathol Lab Med*, **127**(1), 42–48.

[23]Kern, W., Danhauser-Riedl, S., Ratei, R., Schnittger, S., Schoch, C., Kolb, H. J., Ludwig, W. D., Hiddemann, W., and Haferlach, T. (2003). Detection of minimal residual disease in unselected patients with acute myeloid leukemia using multiparameter flow cytometry for definition of leukemia-associated immunophenotypes and determination of their frequencies in normal bone marrow. *Haematologica*, **88**(6), 646–653.

[24]Ko, B.-S., Wang, Y.-F., Li, J.-L., Li, C.-C., Weng, P.-F., Hsu, S.-C., Hou, H.-A., Huang, H.-H., Yao, M., Lin, C.-T., Liu, J.-H., Tsai, C.-H., Huang, T.-C., Wu, S.-J., Huang, S.-Y., Chou, W.-C., Tien, H.-F., Lee, C.-C., and Tang, J.-L. (2018). Clinically validated machine learning algorithm for detecting residual diseases with multicolor flow cytometry analysis in acute myeloid leukemia and myelodysplastic syndrome. *EBioMedicine*, **37**, 91–100.

[25]Lazarus, M. N., Turner-Stokes, T., Chavele, K.-M., Isenberg, D. A., and Ehrenstein, M. R. (2012). B-cell numbers and phenotype at clinical relapse following rituximab therapy differ in SLE patients according to anti-dsDNA antibody levels. *Rheumatology*, **51**(7), 1208–1215.

[26]Lee, A. J., Chang, I., Burel, J. G., Arlehamn, C. S. L., Mandava, A., Weiskopf, D., Peters, B., Sette, A., Scheuermann, R. H., and Qian, Y. (2018). DAFi: A directed recursive data filtering and clustering approach for improving and interpreting data clustering identification of cell populations from polychromatic flow cytometry data. *Cytometry Part A*, **93**(6), 597–610.

[27]Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., ad D. Amir, E., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., Finck, R., Gedman, A. L., Radtke, I., Downing, J. R., Pe'er, D., and Nolan, G. P. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, **162**(1), 184–197.

[28]Li, H., Shaham, U., Stanton, K. P., Yao, Y., Montgomery, R. R., and Kluger, Y. (2017). Gating mass cytometry data by deep learning. *Bioinformatics*, **33**(21), 3423–3430.

[29]Lo, K., Brinkman, R. R., and Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A*, **73**(4), 321–332.

[30]Lun, A. T. L., Richard, A. C., and Marioni, J. C. (2017). Testing for differential abundance in mass cytometry data. *Nature Methods*, **14**(7), 707–709.

[31]Lux, M., Brinkman, R. R., Chauve, C., Laing, A., Lorenc, A., Abeler-Dörner, L., and Hammer, B. (2018). flowLearn: fast and precise identification and quality checking of cell populations in flow cytometry. *Bioinformatics*, **34**(13), 2245–2253.

[32]Mainou-Fowler, T., Dignum, H., Proctor, S., and Summerfield, G. (2004). The prognostic value of cd38 expression and its quantification in b cell chronic lymphocytic leukemia (b-cll). *Leukemia lymphoma*, **45**, 455–62.

[33]Mair, F., Hartmann, F. J., Mrdjen, D., Tosevski, V., Krieg, C., and Becher, B. (2016). The end of gating? An introduction to automated analysis of high dimensional cytometry data. *Eur J Immunol*, **46**(1), 34–43.

[34]Malek, M., Taghiyar, M. J., Chong, L., Finak, G., Gottardo, R., and Brinkman, R. R. (2014). flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics*, **31**(4), 606–607.

[35]Matrai, Z. (2005). Cd38 as a prognostic marker in cll. *Hematology*, **10**(1), 39–46. PMID: 16019444.

[36]Meehan, S., Kolyagin, G. A., Parks, D., Youngyunpipatkul, J., Herzenberg, L. A., Walther, G., Ghosn, E. E. B., and Orlova, D. Y. (2019). Automated subset identification and characterization pipeline for multidimensional flow and mass cytometry data clustering and visualization. *Communications Biology*, **2**(1).

[37]Monaghan, S. A., Li, J. L., Liu, Y. C., Ko, M. Y., Boyiadzis, M., Chang, T. Y., Wang, Y. F., Lee, C. C., Swerdlow, S. H., and Ko, B. S. (2022). A Machine Learning Approach to the Classification of Acute Leukemias and Distinction From Nonneoplastic Cytopenias Using Flow Cytometry Data. *Am J Clin Pathol*, **157**(4), 546–553.

[38]Naim, I., Datta, S., Rebhahn, J., Cavenaugh, J. S., Mosmann, T. R., and Sharma, G. (2014). SWIFT—scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 1: Algorithm design. *Cytometry Part A*, **85**(5), 408–421.

[39]Nolan, J. P. and Condello, D. (2013). Spectral flow cytometry. *Current Protocols in Cytometry*, **63**(1).

[40]O'Neill, K., Jalali, A., Aghaeepour, N., Hoos, H., and Brinkman, R. R. (2014). Enhanced flowType/RchyOptimyx: a bioconductor pipeline for discovery in high-dimensional cytometry data. *Bioinformatics*, **30**(9), 1329–1330.

[41]Park, L. M., Lannigan, J., and Jaimes, M. C. (2020). scpOMIP-069/scp : Forty-color full spectrum flow cytometry panel for deep immunophenotyping of major cell subsets in human peripheral blood. *Cytometry Part A*, **97**(10), 1044–1051.

[42]Pittner, B. T., Shanafelt, T. D., Kay, N. E., and Jelinek, D. F. (2005). Cd38 expression levels in chronic lymphocytic leukemia b cells are associated with activation marker expression and differential responses to interferon stimulation. *Leukemia*, **19**, 2264–2272.

[43]Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.-I., Maier, L. M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., De Jager, P. L., and Mesirov, J. P. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, **106**(21), 8519–8524.

[44]Qian, Y., Wei, C., Lee, F. E.-H., Campbell, J., Halliley, J., Lee, J. A., Cai, J., Kong, Y. M., Sadat, E., Thomson, E., Dunn, P., Seegmiller, A. C., Karandikar, N. J., Tipton, C. M., Mosmann, T., Sanz, I., and Scheuermann, R. H. (2010). Elucidation of seventeen human peripheral blood b-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. *Cytometry Part B: Clinical Cytometry*, **78B**(S1), S69–S82.

[45]Qiu, P., Simonds, E. F., Bendall, S. C., Gibbs, K. D., Bruggner, R. V., Linderman, M. D., Sachs, K., Nolan, G. P., and Plevritis, S. K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Biotechnology*, **29**(10), 886–891.

[46]Rawstron, A. C., Kreuzer, K.-A., Soosapilla, A., Spacek, M., Stehlikova, O., Gambell, P., McIver-Brown, N., Villamor, N., Psarra, K., Arroz, M., Milani, R., de la Serna, J., Cedena, M. T., Jaksic, O., Nomdedeu, J., Moreno, C., Rigolin, G. M., Cuneo, A., Johansen, P., Johnsen, H. E., Rosenquist, R., Niemann, C. U., Kern, W., Westerman, D., Trneny, M., Mulligan, S., Doubek, M., Pospisilova, S., Hillmen, P., Oscier, D., Hallek, M., Ghia, P., and Montserrat, E. (2018). Reproducible diagnosis of chronic lymphocytic leukemia by flow cytometry: An european research initiative on CLL (ERIC) &amp european society for clinical cell analysis (ESCCA) harmonisation project. *Cytometry Part B: Clinical Cytometry*, **94**(1), 121–128.

[47]Saeys, Y., Van Gassen, S., and Lambrecht, B. N. (2016). Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat Rev Immunol*, **16**(7), 449–462.

[48]Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L., and Nolan, G. P. (2016). Automated mapping of phenotype space with single-cell data. *Nature Methods*, **13**(6), 493–496.

[49]Scheuermann, R., Bui, J., Wang, H.-Y., and Qian, Y. (2017). Automated analysis of clinical flow cytometry data: A chronic lymphocytic leukemia illustration. *Clinics in Laboratory Medicine*, **37**, 931–944.

[50]Schroers, R., Griesinger, F., Trümper, L., Haase, D., Andreassen, B., Klein-Hitpass, L., Sellmann, L., Dührsen, U., and Dürig, J. (2005). Combined analysis of zap-70 and cd38 expression as a predictor of disease progression in b-cell chronic lymphocytic leukemia. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K*, **19**, 750–8.

[51]Shekhar, K., Brodin, P., Davis, M. M., and Chakraborty, A. K. (2013). Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). *Proceedings of the National Academy of Sciences*, **111**(1), 202–207.

[52]Smyth, L. J., Eustace, A., Kolsum, U., Blaikely, J., and Singh, D. (2010). Increased airway T regulatory cells in asthmatic subjects. *Chest*, **138**(4), 905–912.

[53]Spitzer, M. H. and Nolan, G. P. (2016). Mass cytometry: Single cells, many features. *Cell*, **165**(4), 780–791.

[54]Stetler-Stevenson, M. and Braylan, R. C. (2001). Flow cytometric analysis of lymphomas and lymphoproliferative disorders. *Semin Hematol*, **38**(2), 111–123.

[55]Stetler-Stevenson, M., Arthur, D. C., Jabbour, N., Xie, X. Y., Molldrem, J., Barrett, A. J., Venzon, D., and Rick, M. E. (2001). Diagnostic utility of flow cytometric immunophenotyping in myelodysplastic syndrome. *Blood*, **98**(4), 979–987.

[56]Terstappen, L. W., Safford, M., nemann, S., Loken, M. R., Zurlutter, K., chner, T., Hiddemann, W., and rmann, B. (1992). Flow cytometric characterization of acute myeloid leukemia. Part II. Phenotypic heterogeneity at diagnosis. *Leukemia*, **6**(1), 70–80.

[57]Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., and Saeys, Y. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A*, **87**(7), 636–645.

[58]Weber, L. M., Nowicka, M., Soneson, C., and Robinson, M. D. (2019). diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications Biology*, **2**(1).

[59]Wei, C., Anolik, J., Cappione, A., Zheng, B., Pugh-Bernard, A., Brooks, J., Lee, E. H., Milner, E. C., and Sanz, I. (2007). A new population of cells lacking expression of CD27 represents a notable component of the B cell memory compartment in systemic lupus erythematosus. *J Immunol*, **178**(10), 6624–6633.

[60]Weir, E. G. and Borowitz, M. J. (2001). Flow cytometry in the diagnosis of acute leukemia. *Semin Hematol*, **38**(2), 124–138.

[61]Wolff, A. S., Oftedal, B. E., Kisand, K., Ersvaer, E., Lima, K., and Husebye, E. S. (2010). Flow cytometry study of blood cell subtypes reflects autoimmune and inflammatory processes in autoimmune polyendocrine syndrome type I. *Scand J Immunol*, **71**(6), 459–467.

[62]Yue, A., Chauve, C., Libbrecht, M. W., and Brinkman, R. R. (2021). Automated identification of maximal differential cell populations in flow cytometry data. *Cytometry Part A*, **101**(2), 177–184.

[63]Zare, H., Shooshtari, P., Gupta, A., and Brinkman, R. R. (2010). Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics*, **11**(1).