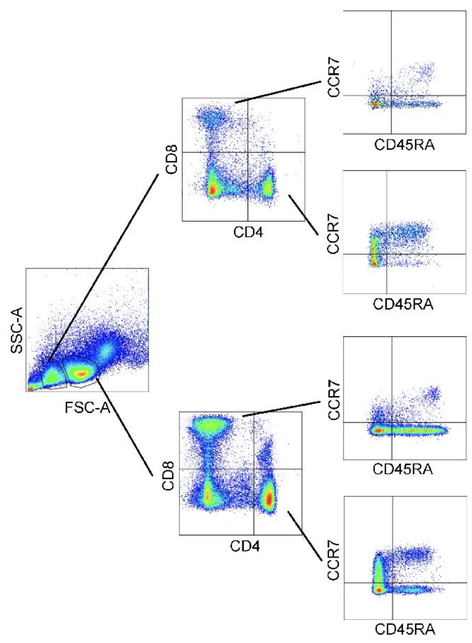
Learning Discriminative Gating Representations for Cytometry Data

Disi Ji[1] **Preston Putzel**[1] **Yu Qian**[2] **Richard H. Scheuermann**[2,3] **Jack D. Bui**[3]
Huan-You Wang[3] **Padhraic Smyth**[1]
[1]University of California, Irvine, [2]J. Craig Venter Institute, [3]University of California San Diego

Introduction

- Traditional analysis of flow cytometry data involves an inefficient manual feature extraction process called gating.
- Prior work applying machine learning techniques to sample diagnosis has focused on learning features separately from classification.
- We optimize features and classifier parameters simultaneously, using a fully differentiable model to learn discriminative interpretable features.



Model

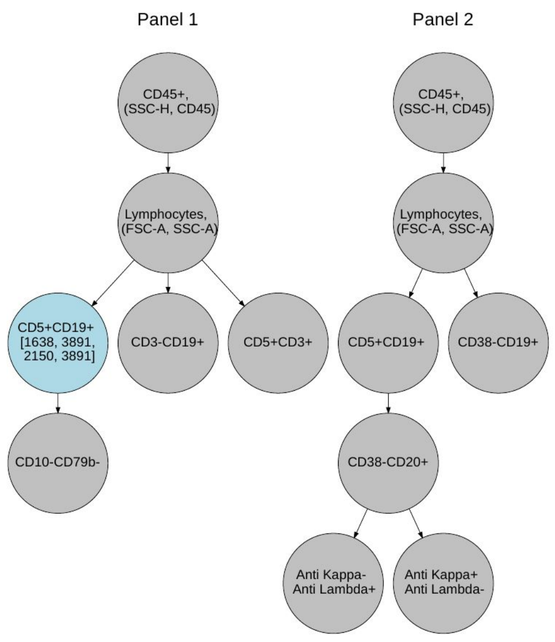
- Data consists of N cell samples, each sample consisting of a matrix of N_i rows (cells) with each column corresponding to a different marker measurement. For multi-panel data there are multiple such matrices per sample.
- Our model takes in a gating tree, with each node k in the tree consisting of a pair of axes d_k. It then learns features by applying sigmoidal gating functions with locations parametrized by θ_k at each node to each cell x_{ij}:

$$g_k(\mathbf{x}_{ij}; \theta_\mathbf{k}, \mathbf{d_k}) = \sigma_s(x_{i,d_1^k} - \theta_{1,1}^k) \times (1 - \sigma_s(x_{i,d_1^k} - \theta_{1,2}^k)) \times \sigma_s(x_{i,d_2^k} - \theta_{2,1}^k) \times (1 - \sigma_s(x_{i,d_2^k} - \theta_{2,2}^k))$$

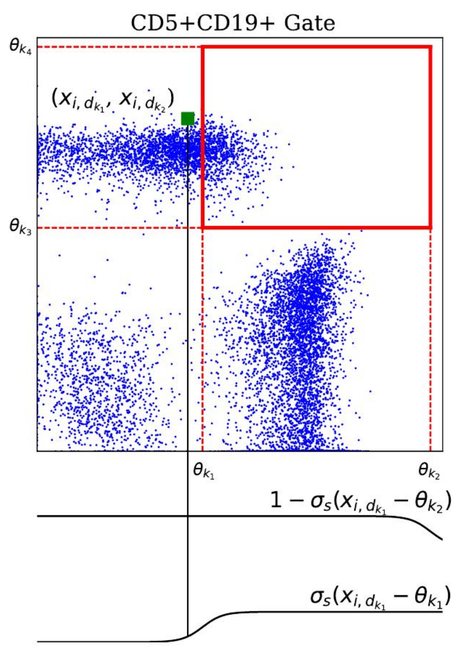
- The model applies these gating functions along the tree producing proportion features for each root to leaf path *p*:

$$f_{ip} = \log\left(\frac{1}{N_i}\sum_{x_{ij}\in x_i}\prod_{k\in \mathrm{Path}(p)} g_k(x_{ij}; \theta_\mathbf{k}, \mathbf{d_k})\right)$$

- A logistic regressor then uses these features to predict the diagnosis probabilities for a sample. We train the model using logistic loss plus regularization to make the gates interpretable. We minimize the loss using SGD with Adam.



(*a*) Gating hierarchy of CLL

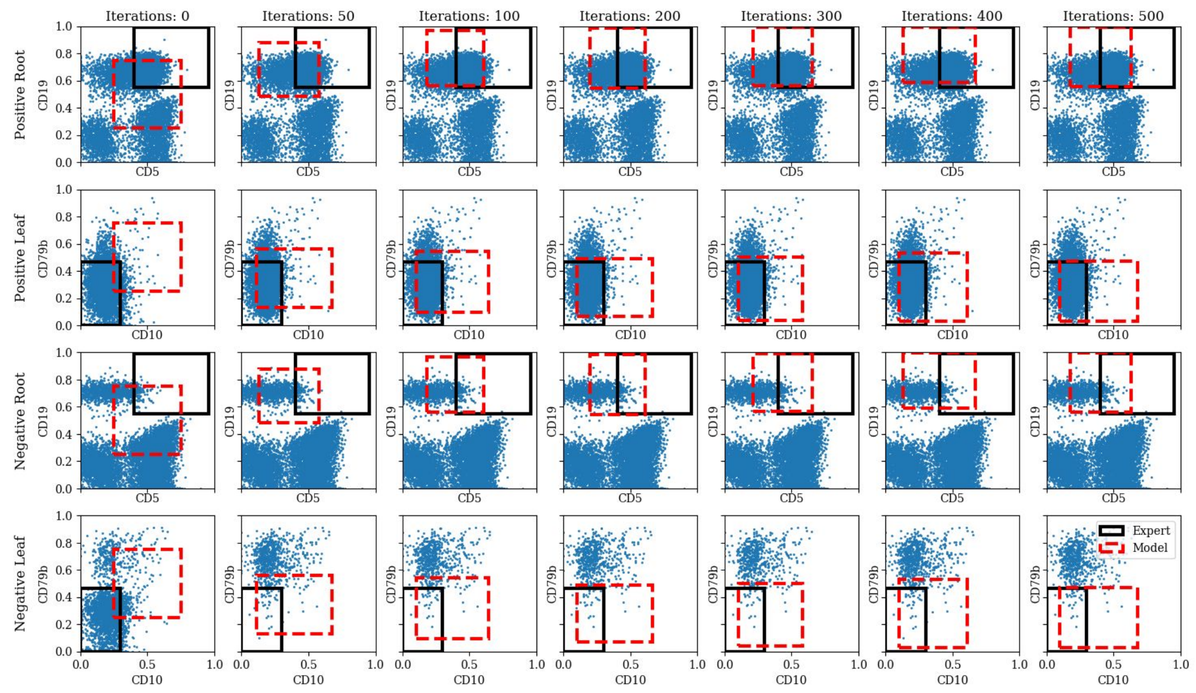


(*b*) One node of the gating tree

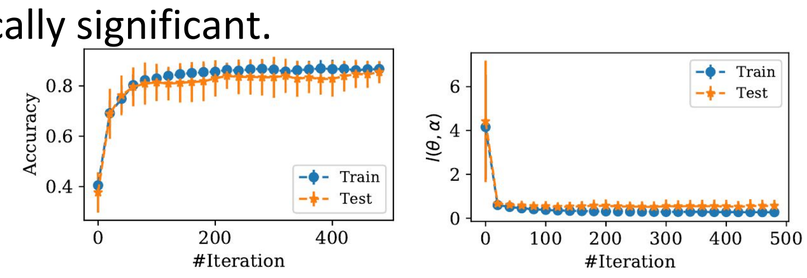
For any questions, email: **pputzel@uci.edu**

Results on CLL Dataset

Chronic lymphocytic leukemia (CLL) clinical data with measurements from 107 individuals, 65 positive, 42 negative across two panels. The model gets similar accuracy to expert analysis, and places gates in the same regions as experts. The panel below shows learning of gates for the one of the panels:

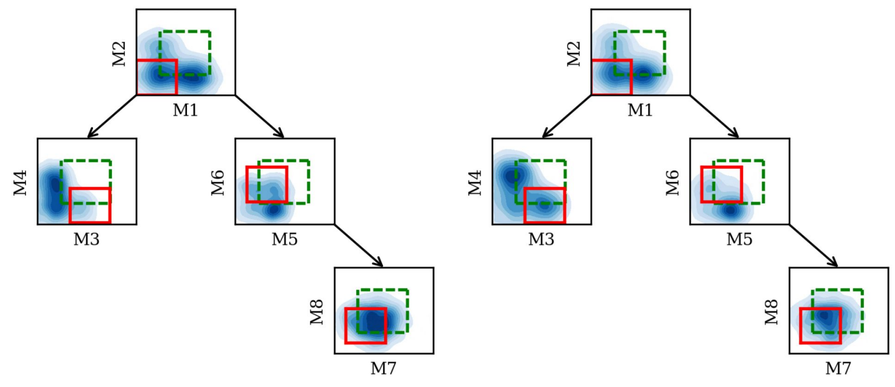


Convergence of the loss, and learned accuracy (below). The difference between model learned accuracy and accuracy using DAFI (a gating algorithm that relies on expert-placed gates) is not statistically significant.

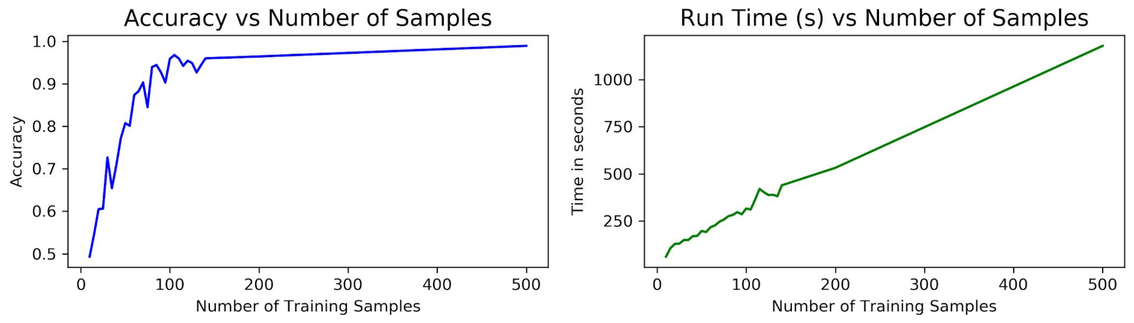


Results on Simulated Data

- Two class synthetic data generated from Gaussian mixture models with high amounts of noise.



- The accuracy as a function of number of training samples used. Runtime scales linearly in number of samples.



Acknowledgements

We would like to thank support from the UCSD Center for Advanced Laboratory Medicine (CALM) and the FlowCAP consortium, as well as Ivan Chang of J. Craig Venter Institute for his work on DAFi-based data preprocessing and analysis. The work in this paper was partially supported by NIH/NCATS U01TR001801 (FlowGate), NSF XSEDE allocation MCB170008, and NIH Commons Credits on Cloud Computing CCREQ-2016-03-00006.