PREDICTING PATIENT-LEVEL PHENOTYPES FROM SINGLE-CELL DATA

Mike Fu (xf2209), Chinyere Ihuegbu (coi2002), Amritha Musipatla (sm3773)

Columbia University in the City of New York
Department of Computer Science

ABSTRACT

Deep learning approaches have seen increased use in several functional genomics applications, often meeting or exceeding the performance of state-of-the-art methodologies [1]. Over the course of this project, we intend to employ neural network architecture to predict patient-level phenotypes from single-cell mass cytometry data.

1. INTRODUCTION

High-dimensional single-cell technologies (such as single-cell RNA sequencing, single-cell ATAC-sequencing, mass cytometry, among many others) have become useful tools in understanding the heterogeneity and molecular profile of cells in both health and diseased states [2], as a result these technologies now play an important role in the development of more effective and precise therapies to diseases such as cancer [3, 4]. Single-cell technologies also enable the detection and description of diverse cell population in both healthy and diseased states - including the identification of rare cell populations which traditional biological experiments may fail to detect. This ability to map out the genomic and proteomic profile of cells in different states, such as in development, healthy and diseased states, is crucial in improving our understanding of cellular processes and the human biology .

Unsupervised clustering is a machine learning approach, which utilizes single-cell omics analysis to identify and define cell clusters (and subsequently cell types) [5, 6, 7] . While this learning approach has aided our understanding of the heterogeneity of cell types in development, health and diseased states, unsupervised clustering is still limited in directly associating identified cell clusters with disease status, as this is usually done manually with the aid of literature, wet-lab experiments and expertise knowledge. Furthermore, with unsupervised learning associating cell cluster to a disease phenotype usually requires an additional step of condition-specific differences introduced in wet-lab experiments (or differences introduced as a function of individuals having a disease and thus altered biology relative to healthy individuals),which in turn is represented in unsupervised clustering of single-cell data by the increased or decreased abundance of a given cell cluster(s) compared to healthy controls or healthy individual.

The limitation of unsupervised clusterings' inability to predict phenotype(s) (such as disease condition, expected survival, developmental stage) from single-cell data is one which is expensive and time consuming when relying on the combination of literature, expertise knowledge and wet-lab experiments to associate cell types with phenotypes.

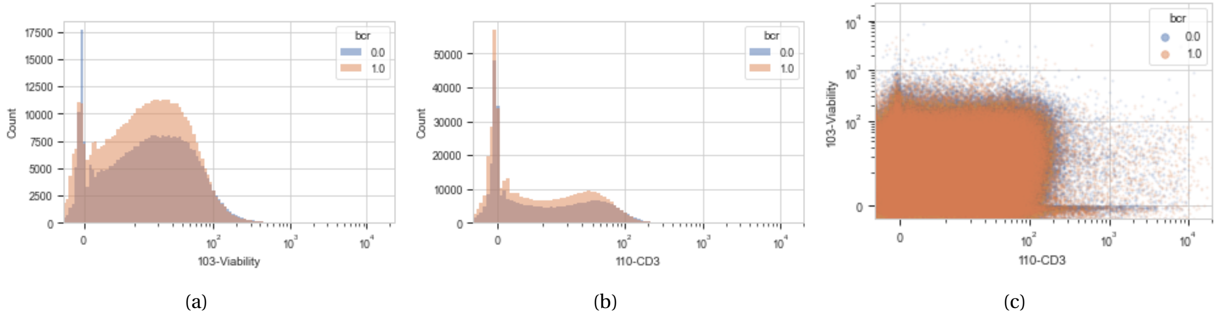
Our project aims to address this problem using supervised learning, specifically deep convolutional neural networks with multiple layers and transfer learning. In a recent study done by Arvaniti and Claassen, a computational framework, called CellCNN, was developed to facilitate the detection of rare- disease associated subsets and their phenotype [8]. CellCNN combined multi instance learning and convolutional neural network using a single layer to identify cell clusters and also the associated disease status of clusters. Our project aims to further extend on this using additional techniques such as dimensionality reduction and saliency maps to extract information which could then be used for predictions.

2. RELATED PAPERS

2.1. Deep Learning for Phenotype Prediction

Recent applications of convolutional neural networks (CNNs) on single-cell inputs have demonstrated CNN utility for both the classification (multi-node output) and regression (single node output) task.

Specifically, Claasen et. al adapted CNN architecture for processing groups of single-cell inputs (collectively, multi-cell inputs), where each of these multi-cell inputs was associated with a phenotype [8]. The group was able to demonstrate efficacy of their CellCNN architecture in a variety of phenotype prediction tasks. Their architecture was used to classify whether peripheral blood mononuclear cells (PBMC) samples had been stimulated or unstimulated by paracrine agents, based on the samples intra-cellular markers. Additionally, the group was able to use CellCNN for the regression task, to identify T-cell subpopulations that were at higher risk of AIDS onset following infection to HIV. CellCNN was shown to predict comparably to state-of-the-art approaches, with lower computational costs.



(a) (b) (c)

Fig. 1. We plot the flow cytometry measurements of protein markers for single-cell measurements out of the Pe'er Lab's B-Cell bone marrow data [9] (a) Cell count for phenotype : 103-Viability (b) Cell count for phenotype : 110-CD3 (c) Scatter plot showing single cells' measurements of Protein Marker 110-CD3 and 103-Viability

Chen et. al. were able to use a deep learning approach to predict gene expression patterns of cells [10]. Their model, dubbed D-GEX, predicts the expression of a "target" gene based on known expressions of a group of "landmark" genes. Comparing their RNN based models to other state-of-the-art methods (linear regression), resulted in D-GEX significantly outperforming other methodologies.

2.2. Datasets

A study done by the Pe'er and Nolan lab groups employed mass cytometry to study the development trajectory of B cells, as well as the behaviour of B cells in response to stimuli [9]. As part of their experiment, Pe'er et. al collected B cell centric mass cytometry data from human bone marrow. Their dataset includes records of multiple B cell cellular features, such as activation of regulatory signaling molecules, and phenotypic protein markers of B cell development in both unstimulated and stimulated conditions. We will be using their publicly available data of mass cytometry of unstimulated and stimulated conditions from the same group to reduce technical variations such as batch effects.

We develop and train our model using mass cytometry data of B cells under unstimulated and stimulated conditions. Since the development of B cells and its response to stimuli results in differential expression of specific B cell proteins and phenotypes (Fig 2), we would test the sensitivity and efficiency of our model in detecting phenotype(s) associated with observed cell types. This intended data set is already publicly available on https://reports.cytobank.org/1/v1.

3. METHODS

Our network will take in single cell data as input and attempt to identify the phenotype-associated cell subpopulations. We will be using supervised learning via variety of different model architectures.

3.1. Preprocessing and Feature Engineering

We will attempt numerous feature engineering techniques. Dimensionality reduction techniques such as linear discriminant analysis and principal component analysis have been used to reduce the raw data size (and by extension the computational cost) of genomics data sets [11]. Outside of functional genomics, dimensionality reduction of input data has been shown to improve performance of image classification models. We will try dimensionality reduction for pre-processing to cell distribution. Another way to extract information is to use saliency maps to see which cells/genes are contributing to the predictions.

3.2. Model Architecture

For our network, we will be using deep convolutional neural networks (CNN). CNNs broadly consist of series of convolutional layers followed by pooling layers. In the case of CellCNN, the architecture consisted of max-pooling and mean-pooling layers following the convolutional layer. One avenue of exploration is to integrate an expectation pooling layer, developed for using CNNs as probabilistic models [12]. We plan on expanding the number of layers, as well as implementing state of the art architectures such as UNet [13]. We will also use transfer learning, applying trained models from other single celled states models, as well as from other fields of study such as computer vision and natural language processing [7] [14].

4. PRELIMINARY RESULTS

4.1. Exploratory Data Analysis

The Pe'er Lab's B cell data was processed into single-cell inputs with corresponding phenotype classification (in this case, B cell response to stimuli). Preliminary exploratory data analysis of the bone marrow B-cell data shows that 38 cellular features were measured via flow cytometry. We use

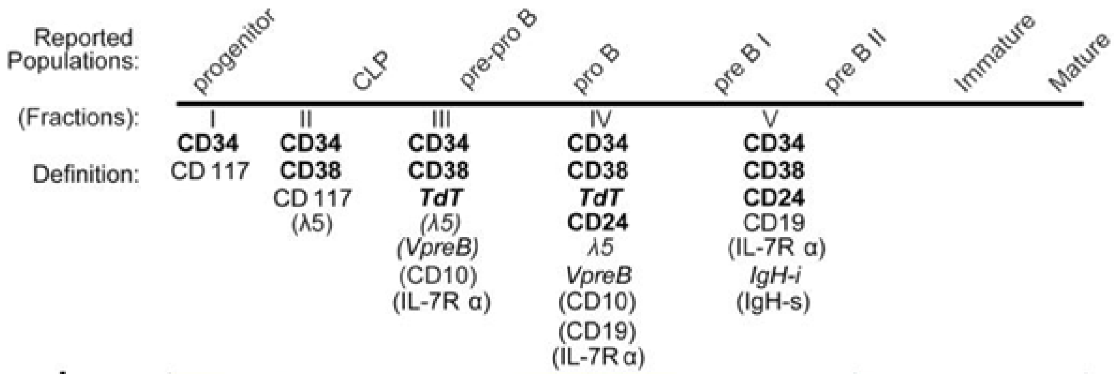


Fig. 2. Significant Protein Markers used to Categorize B-Cell Development Stage [9]

the open source python library Cytoflow [15] to extract and plot the cytometry data. We examine histograms of measurements of specific phenotypes for both the BCR stimulated and unstimulated cells, Figure 1 (a) and 1 (b) plot measurements of two example measurements for both the stimulated and unstimulated populations. We also plot cells' measurements of selected phenotypes against their measurements of other phenotypes in scatter plots, Figure 1 (c) records an example scatter plot featuring cells' 103-Viability measurement against their 110-CD3 measurements.

4.2. Baseline Model Development

From our preliminary data analysis, we note that the spread of measurements cannot easily be correlated to whether the cell was stimulated or not, ie. the measurements for these cells are not well-separated, for that reason we chose to forego using a Gaussian Mixture Model as our baseline. Instead we developed and ran a baseline logistic regression model, which achieved an accuracy of 82% in categorizing cells as stimulated and unstimulated. Our baseline random forest scored an accuracy of 86%. Our baseline SVM model scored a baseline of 75%. Table 1 records our best accuracy for each of these baseline models.

Logistic Regression	**Random Forests**	**Support Vector Machine**
82.5%	86.7%	75.6%

Table 1. Measured classification accuracy of baseline models, trained on the full B-Cell Bone Marrow dataset.

Our baseline models also give us an insight on the relative importance of each protein marker. The top 5 most important protein markers we found were the 114-CD3, 112-CD3, 150-pSTAT5, 154-pSHP2, 110-CD3.

One potential issue we ran into was the sheer amount of the data available. Even training our baseline logistic regression took more than an hour, with other baseline methods, such as random forests and support vector machines taking even longer. One potential fix we came up with is to train only on 50% of the available training data to fine tune our model architecture and hyperparameters. Once we have a model selection that we are satisfied with, then we will use the entirety of the training data to train our model.

5. NEXT STEPS

For the remainder of the semester we intend to develop and improve a neural network based model to accurately classify our cell measurements as stimulated/unstimulated. We will use our logistic regression model as a baseline.

5.1. Model Improvements

In addition to hyperparameter tuning as described in Preliminary Results, we could improve the performances of our NN model by using ablation, where specific features/nodes are removed and the change in accuracy is measured to see how the model performs in the presence vs absence of selected features/nodes. Another method of assessing and increasing performance of our model is by comparing output accuracy and run time for different model architectures on a single data set.

5.2. Integrating Wanderlust Classifications

Another area of interest is including the original findings of the Pe'er Lab group as features in our model. The Pe'er Lab's Wishbone [16] is an unsupervised learning algorithm, a nearest-neighbor (NN) based graph trajectory detection algorithm. Wishbone takes multiparameter single-cell events as input, and maps them onto a cell developmental trajectory [16, 9]. We would like to take each cell's developmental category as classified by the Wishbone algorithm and include this in training the model. As a result we may be able to improve our model's classification rate.

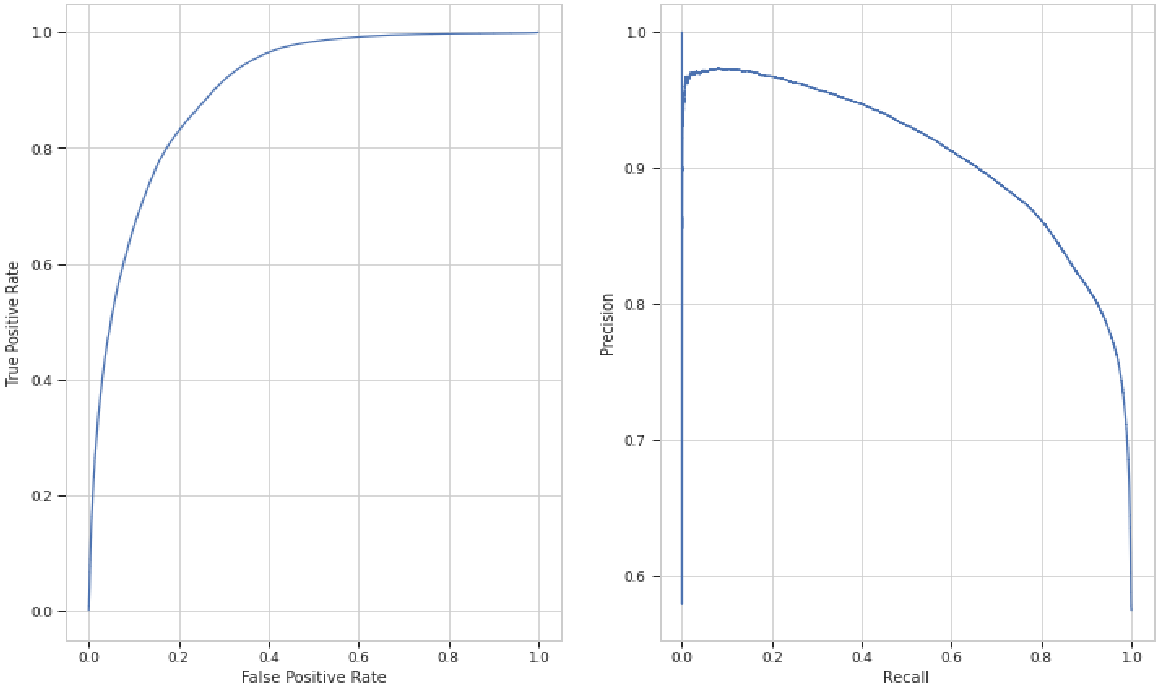


Fig. 3. Precision Recall curve and ROC curves for logistic regression. Our baseline achieved an AUC of 0.78 and 0.83 respectively.

Another option is to see if we could independently develop a NN model to classify cells into their developmental category, using the Pe'er Lab classifications as truth data. As shown in Figure 2, the protein markers '103-Viability', '148-CD34', '167-CD38', '142-CD19', and '147-CD20' show a correlation to a B cell's developmental stage.

5.3. Protein Marker Importance

After developing a suitable model, we would then like to examine the relative effect of certain protein markers. We could use saliency maps, for example, to check which protein markers are more indicative of stimulation.

6. REFERENCES

[1] Lefteris Koumakis, "Deep learning models in genomics; are we there yet?," *Computational and Structural Biotechnology Journal*, vol. 18, 6 2020.

[2] Edoardo Galli, Ekaterina Friebel, Florian Ingelfinger, Susanne Unger, Nicolás Gonzalo Núñez, and Burkhard Becher, "The end of omics? high dimensional single cell analysis in precision medicine," *European Journal of Immunology*, vol. 49, no. 2, pp. 212–220, 2019.

[3] Tian Q. Price N. D. Hood L. Yurkovich, J. T., "A systems approach to clinical oncology uses deep phenotyping to deliver personalized care.," *Nature reviews. Clinical oncology, 17(3), 183–194.*, 2020.

[4] Iorgulescu J. B. Braun D. A. Keskin D. B. Livak K. J. Gohil, S. H., "Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy.," *Nature reviews. Clinical oncology, 10.1038/s41571-020-00449-x. Advance online publication.*, 2020.

[5] Xiaoshu Zhu, Hong-Dong Li, Lilu Guo, Fang-Xiang Wu, and Jianxin Wang, "Clustering of single-cell rna-seq data by unsupervised learning methods," *Current Bioinformatics*, vol. 14, 11 2018.

[6] Andrews T. S. Hemberg M. Kiselev, V. Y., "Challenges in

unsupervised clustering of single-cell rna-seq data.," *Nature reviews. Genetics, 20(5), 273–282.*, 2019.

[7] Hockley J. Görnitz N. Vidovic M. M. Müller K. R. Gutteridge A. Ziemek D. Mieth, B., "Using transfer learning from prior reference knowledge to improve the clustering of single-cell rna-seq data.," *Scientific reports, 9(1), 20353.*, 2019.

[8] Claassen M. Arvaniti, E., "Sensitive detection of rare disease-associated cell subsets via representation learning.," *Nature communications, 8, 14825.*, 2017.

[9] Davis K. L. Amir e. Tadmor M. D. Simonds E. F. Chen T. J. Shenfeld D. K. Nolan G. P. Pe'er D. Bendall, S. C., "Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development.," *Cell, 157(3), 714–725.*, 2014.

[10] Rajiv Narayan Aravind Subramanian Xiaohui Xie Yifei Chen, Yi Li, "Gene expression inference with deep learning," *Bioinformatics*, vol. 32, 2 2016.

[11] Seiya Imoto Satoru Miyano Alok Sharma, Kuldip K. Paliwal, "A feature selection method using improved regularized linear discriminant analysis," *Machine Vision and Application*, vol. 25, 11 2013.

[12] Yang Ding Ge Gao Minghua Deng Xiao Luo, Xinming Tu, "Expectation pooling: an effective and interpretable pooling method for predicting dna–protein binding," *Bioinformatics*, vol. 36, 10 2019.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[14] Eddy S. R. Koo, P. K., "Representation learning of genomic sequence motifs with convolutionalneural networks.," *PLoS computational biology, 15(12),e1007560.*, 2019.

[15] Brian Teague, "Cytoflow," https://github.com/cytoflow/cytoflow/, 2018, [Online; accessed 04-April-2021].

[16] Reich-Zeliger S Angel O Salame TM Kathail P Choi K Bendall SC Friedman N Pe'er D Setty M*, Tadmor MD*, "Wishbone identifies bifurcating developmental trajectories from single-cell data," *Nature Biotechnology*, 2016.