Generative Modeling Approach for Computational Multiplet Detection

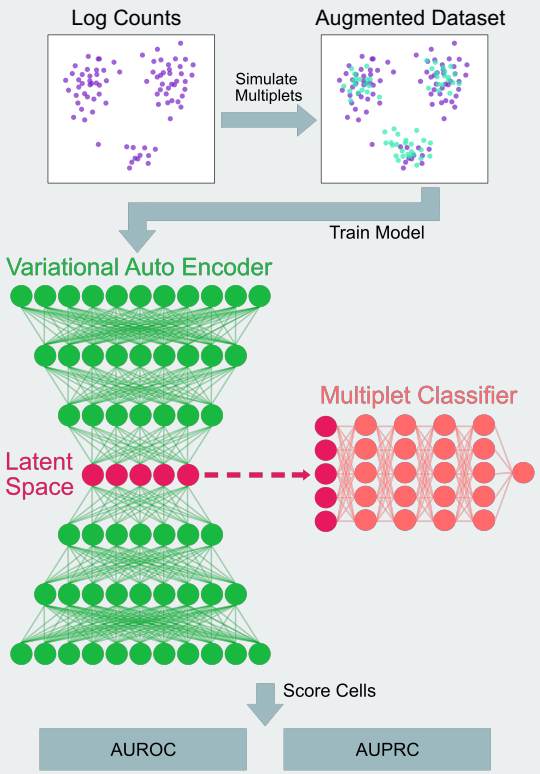
Hannah C. Schriever[1,3], Dennis Kostka[2,3]

1. Carnegie Mellon – University of Pittsburgh Ph.D. Program in Computational Biology, 2. Developmental Biology and Computational & Systems Biology, 3. University of Pittsburgh School of Medicine

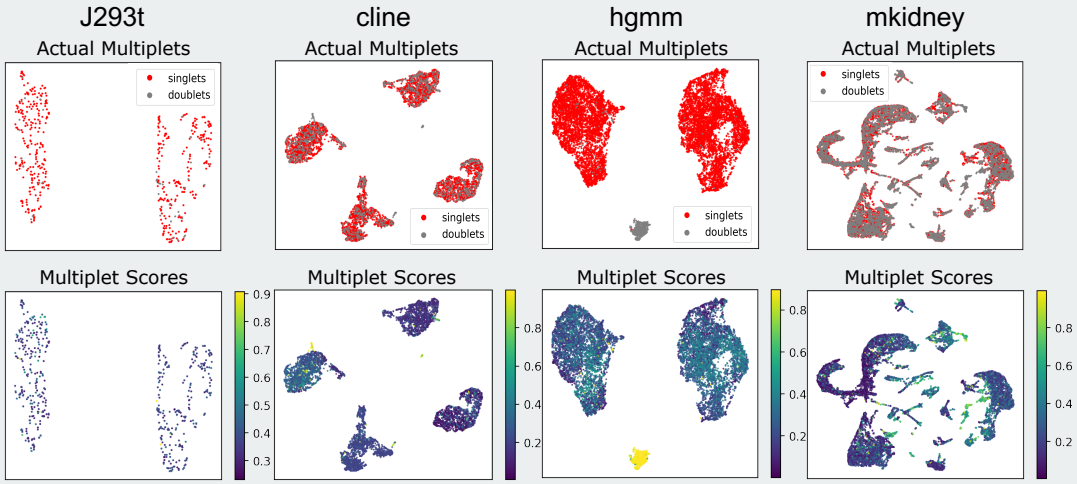
Introduction

Single-cell RNA sequencing (scRNA-seq) greatly increases resolution of expression data, however, this technique is restricted by technical artifacts like multiplets. Multiplets occur when two or more cells receive the same barcode during sequencing, and thus appear as one cell. They introduce nonexistent expression profiles, which leads to incorrect interpretation of the data. I propose vaeda, a computational tool to annotate multiplets.

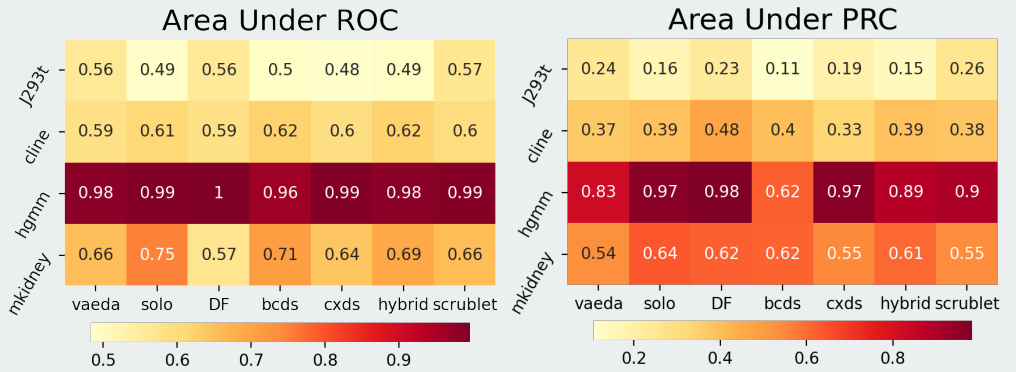
Graphical Abstract



UMAP Visualization of vaeda Multiplet Scores



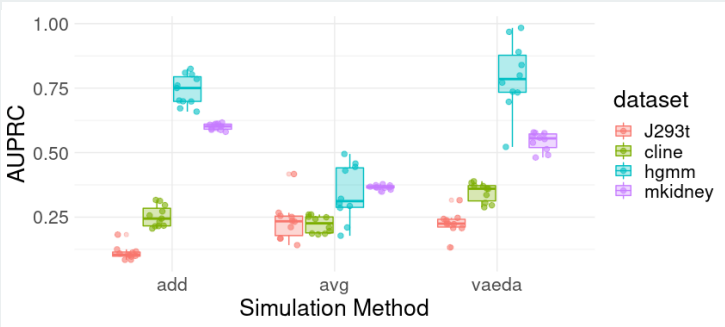
Comparison to other Methods



The methods solo, DoubletFinder, bcds, cxds, scds hybrid, and scrubblet were used to annotate multiplets in four different datasets to compare to our method vaeda. Each method was run five times on each dataset and the average area under the ROC and PRC are reported above. The datasets are ordered based on number of cells.

Multiplet Simulation

In order to train a classification tool, a classifier requires positive (multiplets) and negative (single cells) examples in the dataset. In application, however, the actual multiplets are not known. Therefore, vaeda simulates multiplets to augment the datasets prior to training. Below is a comparison of the AUPRC for different simulation techniques. Sum refers to adding the expression of two cells, avg refers to averaging two cells, and vaeda refers to our method. Vaeda simulates multiplets by adding the expression values of two or three cells. Then the library size of the resulting multiplet is adjusted to match one of its components, which is chosen at random.



Conclusions and Future Directions

Moving forward, running vaeda on more datasets will be essential to compare its performance to other models. Additionally, the current approach of labeling all actual data as negative and all simulated data as positive is naïve, so other approaches may improve performance.

References:

[1] McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst 8, 329–337.e4 (2019). [2] Bais, A. S. & Kostka, D. scds: computational annotation of doublets in single-cell RNA sequencing data. Bioinformatics (2019) doi:10.1093/bioinformatics/btz698 [3] Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst 8, 281–291.e9 (2019) [4] Bernstein, N. et al. Solo: doublet identification via semi-supervised deep learning. bioRxiv 841981 (2019)doi:10.1101/841981. [5] Xi N. M. & Li, J. J. Benchmarking computational doublet-detection methods for single-cell RNA sequencing data. Cell Syst, (2020).

Carnegie Mellon – University of Pittsburgh Ph.D. Program in Computational Biology

University of Pittsburgh