Autoencoders for T Cell Exhaustion Biomarker Discovery in Single-Cell RNA Data

Alex Ying

***Abstract*—The ability to characterize T cells has important clinical implications for the treatment of Chronic Lymphocytic Leukemia (CLL). Specifically, being able to identify exhausted T cells within a patient can determine the suitability of CAR-T cell therapy. In this report, we demonstrate the application of deep-learning techniques to standard bioinformatics analyses of single-cell RNA data (scRNA). Transcriptome data from CD4+ and CD8+ T cells from CLL patients was clustered using both the raw gene expression data and the encoded representations generated from a standard autoencoder network and a pass-through autoencoder architecure. Interestingly, this pass-through autoencoder displayed the ability to cluster cells by GZMK and GZMH expression levels, two known markers for T cell function. This marks an improvement over standard clustering methods, and demonstrates the networks ability to independently learn to cluster cells by the combination of important marker genes. Seven clusters were identified in the T cell populations using this novel clustering method, including likely clusters of exhausted CD4+ T cells and exhausted CD8+ T cells. These clusters indicate that up-regulation LTB and IL7R and down-regulation of TYROBP and CST3 may indicate CD4+ T cell exhaustion, and highlight LAG3 and GZMK as possible markers for CD8+ T cell exhaustion.**

I. Introduction

Chronic Lymphocytic Leukemia (CLL) is a disease characterized by T cells frequently exhibiting a state known as exhaustion [1]. Exhausted T cells respond less strongly to antigens and are known to have lower proliferative capabilities. Identifying exhausted T cell is particularly relevant in the case of CAR-T immunotherapy, where previous research has indicated that expression of key genetic markers for exhaustion are correlated with lower yields of T cells and lower effectiveness [2].

T cell exhaustion is known to be marked by an up-regulation in several inhibitory genes, such as PDCD1, LAG3, TIGIT, and CD160 [2][3]. However, PDCD1 expression alone has been shown to be insufficient to predict the proliferative capabilities of T cells, and the exact interactions between genes is still unknown.

Single-cell RNA Sequencing (scRNA-seq) allows for expression of multiple genes to be counted on a per-cell basis. The expression values are then typically clustered together using manifold learning algorithms, such as UMAP [4], to infer similarity. These clusters are then identified by the percent of cells expressing known markers, as well as the median expression levels of those markers.

The sometimes low signal-to-noise ratio of scRNA sequencing, as well as the inherently high dimensionality and redundancy of the generated data has motivated recent development of autoencoder driven workflows for reducing the dimensionality of the scRNA data for deriving the clusters [6][7][8][9]. The authors of these studies have noted improved performance over the standard scRNA clustering workflow, making the prospect for deep learning techniques in this field rather optimistic.

In this study, we generated two autoencoder networks and cluster each cell based on their compressed representation using UMAP. These clusters are compared to clusters generated from the whole-gene representation. Finally, we analysed of the clusters generated from the autoencoders for novel biomarkers that may be used to identify T cell exhaustion.

II. Methodology

PBMC scRNA data from the NCBI GEO database, series number GSE111015, was used for this study [10]. This scRNA data was then filtered to only include CD4+ and CD8+ T cells by excluding cells that did not have positive expression levels of the CD4 and CD8A genes.

A. Initial Clustering

Initial clustering was performed using the Scanpy library using the standard UMAP utility across the whole gene space [5]. Clusters were identified using the Louvain algorithm, and these cluster labels were used for downstream identification.

B. Pure Autoencoder Clustering

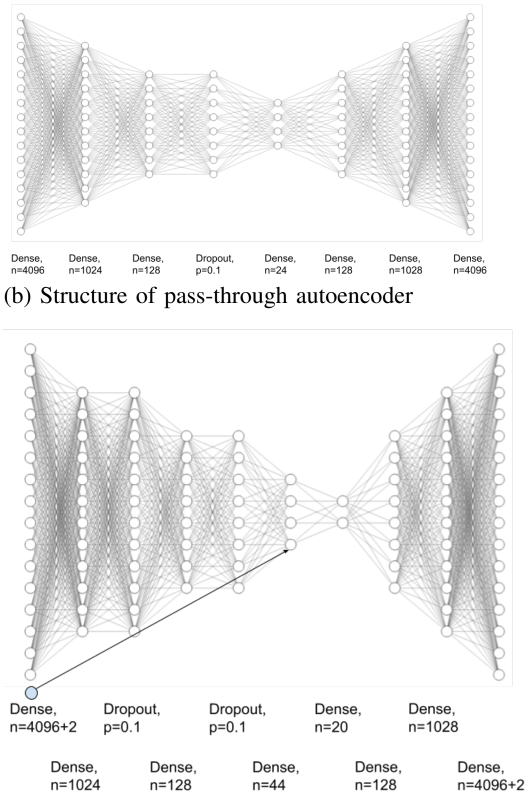
The most variable genes (n = 4096) among the T cells were subset from the scRNA matrix and used as samples for training the autoencoder (Fig 1a). The Louvain algorithm was again used to identify clusters within the latent space, and the UMAP algorithm was also applied to the latent representation for visualization purposes.

C. Feature Pass-through Clustering

One aspect of scRNA clustering is that certain marker genes are known to be entirely associated with the underlying state. In other words, given that a cell expresses CD8A, it can be assumed that its underlying state is some type of CD8+ cell (e.g. a CD8+ memory cell). It is from this fact that we conjecture that passing some of this prior knowledge directly into the latent representation of the gene expression levels may aid with both training and result in clusters that more closely resemble the underlying distribution of cell states in the sample.

Manuscript received December 24, 2021; revised December 24, 2021.

Fig. 1: Diagrams for standard autoencoder and pass-through autoencoder



(a) Structure of standard autoencoder

(b) Structure of pass-through autoencoder

Traditional classification of cells can be understood as an approximation of some ideal classifying function S:

$$S(g) = argmax(p)$$

$$= argmax(w_1 g + w_2 g \otimes g + w_3 g \otimes g \otimes g...)$$

where g is the vector of gene expression levels and p is the probability vector for each possible state. Note that this representation assumes some interactions between genes weighing into the probability, but the exact form of the ideal classifier function is both unknown and not needed for this line of reasoning.

Let the approximation of this classifying function be S':

$$S'(g_m) = cluster(g)$$

$$= argmax(p - \epsilon) = argmax(w_m g) \approx argmax(w_1 g)$$

where g_m is a subset of g containing only the known marker genes, ϵ represents errors in the probability of each state. This approximation of the ideal classifier function is approximately the same as using only $w_1 g$, with the assumption that the non-marker genes have a weights of approximately 0.

Training the autoencoder completely from scratch aims to approximate $S(g)$, where each cluster in theory represents the output of $argmax(p)$. Specifically, we argue that the encoding function $E(g) = h$ and, after sufficient training, $cluster(h) = argmax(w_h h) = argmax(p)$, where h is the vector of latent states and w_h are the associated weights. If instead we pass-through a vector of known marker genes into the encoding, we instead obtain $E(g) = [g_1, g_2, h_1, h_2...]$, and $cluster(g_m, h) = argmax(w_1 g + w_h h) = argmax(p)$. It thus follows that

$$p = w_1 g + w_h h = w_1 g + \epsilon$$

$$\epsilon = w_h h = w_2 g \otimes g + w_3 g \otimes g \otimes g$$

So by passing in the vector of known marker genes, and allowing the network to essentially begin with the approximate classifying function, we have the encoded representation learn only the interaction terms and the non-linear portion of the ideal classifying function. Conceptually, this architecture resembles the residual block commonly used in ResNets and other deep neural networks, and helps with training and reducing the amount of information required to be compressed in the encoded space. The final implementation of the pass-through model does not follow this exact architecture (Fig 1b.). Instead, CD4, CD8A, and LAG3 expression levels were concatenated to the vector of expression levels for the most variable genes as in the standard autoencoder training. The product of the concatentation was used as both the input and output of the network. After the input later, the CD4, CD8A, and LAG3 expression levels are cropped out, and passed through, being concatenated to the dense layer just before the encoding layer. The model was trained as in the pure autoencoder clustering, and again the Louvain algorithm and UMAP algorithms were used for clustering and visualization, respectively.

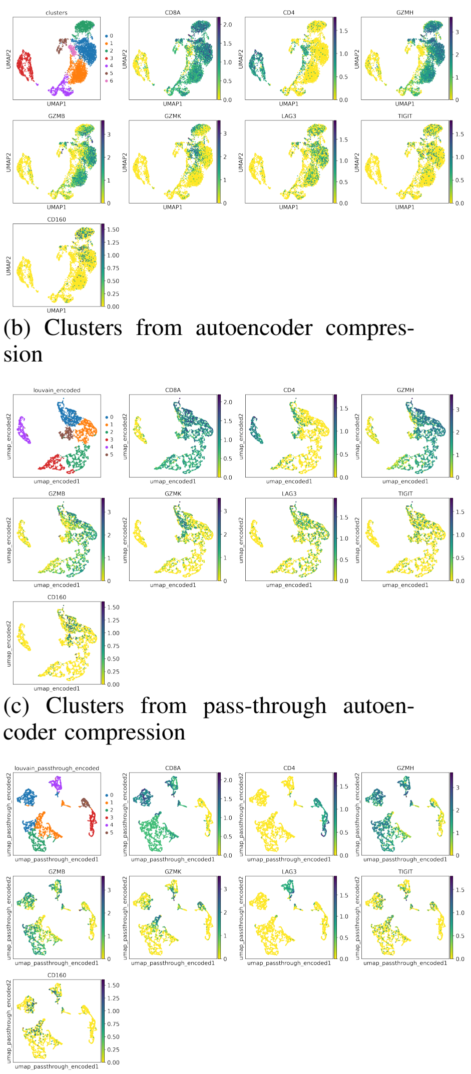
III. Analysis

The mean-squared-area loss while training the models was found to be roughly 0.720 during training for both models. No indication of overfitting was found as the tendency was for both the validation and test loss to approach this value.

The UMAP visualizations for each clustering method (Fig. 2) shows that the autoencoder seems to have under-performed compared to the standard methodology, failing to identify two distinct clusters of CD4+ T cells that were identified by both the standard clustering and pass-through autoencoder. However, as can be seen in the CD4 and CD8A expression levels by cluster, the pass-through architecture clusters CD4+ and CD8+ cells far more tightly than both the standard clustering and standard autoencoder clustering. This is despite the fact that CD4 and CD8A expression levels only appear layer before the encoding layer. Conceptually, this is probably because the transition from the concatenation layer to the encoding layer preserves the CD4 and CD8A expression levels, but also emphasizes interactions between CD4 and CD8A with other genes. This can help explain why the pass-through autoencoder also displayed the surprising ability to cluster based on GZMH and GZMK expression levels, a feature not found in the standard clustering and standard autoencoder clustering.

The GZMK expression levels (Fig. 3) also serve to identify cluster 5 of the pass-through autoencoder clusters to likely be CD4+ exhausted T cells, given its up-regulation of GZMK and down-regulation of GZMB, compared to cluster 3, which may be T helper cells based on their expression of marker genes

Fig. 2: UMAP of known marker genes



(a) Clusters from raw expression values

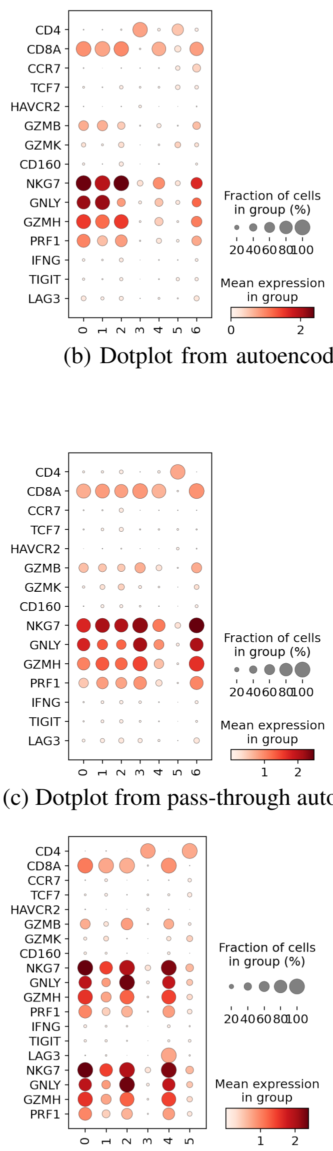
(b) Clusters from autoencoder compression

(c) Clusters from pass-through autoencoder compression

[11]. Using Scanpy to identify markers between clusters (Fig. 4) shows that these exhausted CD4+ T cells appear to down-regulate TYROBP and CST3, and up-regulate LTB and IL7R.

LAG3 was used as the primary marker for CD8+ T cell exhaustion, and accordingly, is tightly clustered by the pass-through autoencoder. The clustering generated by both the raw gene expressions and autoencoder show a fairly uniform distribution of LAG3, TIGIT, and CD160. We can maybe still assume that the LAG3+ cluster 4 represents exhausted T cells, as it also contains more GZMK+ cells, but analysis of the between-cluster marker genes still leaves room for investigation for how CD8+ cells were clustered. Cluster 1 most likely contains CD8+ terminal effector cells given up-regulation of GZMK, but clusters 0 and 2 are difficult to distinguish.

Fig. 3: Dotplots of known marker genes



(a) Dotplot from raw expression values

(b) Dotplot from autoencoder

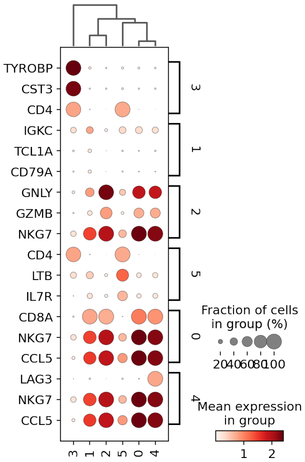
(c) Dotplot from pass-through autoencoder

Another interesting outcome is that despite LAG3 being among many important markers for T cell exhaustion, TIGIT and CD160 do not cluster with LAG3, remaining mostly distributed in all CD8+ cell clusters. This highlights a potential pitfall with the pass-through architecture, where over-biasing the network and adding too many or too few known markers lead to incorrect clustering. However, it's also possible that the lack of cross-correlation between LAG3 and TIGIT within the dataset signifies a heterogeneity in T cell exhaustion that requires further exploration. Notably, there was a lack of cells expressing PDCD1 in the dataset, another common marker for exhaustion, which also contributes to uncertainty with the data itself.

IV. Discussion

Identifying T cell exhaustion is a problem with important implications for the treatment of CLL, and the diversity between exhausted cells contributes to the difficulty in their identification. The results of this study suggests that only looking for the canonical markers is insufficient for

Fig. 4: Dotplots of marker genes between pass-through autoencoder clusters



determining whether a T cell will exhibit exhaustion in its functionality. This makes it somewhat difficult for traditional scRNA identification techniques to group exhausted cells. The application of autoencoders in this study demonstrates that deep learning may be better able to model the underlying T cell state than manifold projection techniques. This is seen in the emergent clustering along GZMK, GZMH, and GZMB expression levels. However, the lack of PDCD1 expression in the dataset, as well as the lack of coexpression between LAG3, TIGIT, and CD160, all key exhaustion markers, suggest that additional data is needed. Especially valuable would be insight into the phenotypic characteristics of the T cells, which can serve to resolve whether this heterogeneity is a result of the underlying complexity of T cell exhaustion or due to problems with the data itself.

One area for further improvement is further refinement of the model's hyperparameters. Changing the number of latent dimensions and amount of dropout has a substantial effect on the quality of clusters, and determining what heuristics to use to evaluate the quality of a set of hyperparameters is an area for future study. This brittleness with respect to the encoding layer size may have contributed to the under-performance of the normal autoencoder compared to the standard clustering. Furthermore, investigating the heuristics needed to choose marker genes to pass through to the encoding layer is also an area of future work, as it represents an area where human bias can shape the clustering. While this may be desirable in cases where cell types are known, it can over-value the importance of marker genes in distinguishing between cells, rather than combinations or interactions between cells.

Another area of further refinement is possibly the incorporation of T cell data from other studies [11][12]. T cell exhaustion has been labeled in several different diseases, and the incorporation of this data may help with building a more complete picture of T cell exhaustion and the connection between the transcriptome and functionality. Longitudinal data may also be relevant, as exhaustion is known to develop over time, and comparing the transcriptome between healthy patients and CLL patients can also highlight genetic markers that should be emphasized by the model.

References

[1] S. H. Gohil and C. J. Wu, “Dissecting CLL through high-dimensional single-cell technologies,” Blood, vol. 133, no. 13, pp. 1446–1456, Mar. 2019, doi: 10.1182/blood-2018-09-835389.

[2] J. H. Lee, S. Shao, M. Kim, S. M. Fernandes, J. R. Brown, and L. C. Kam, “Multi-Factor Clustering Incorporating Cell Motility Predicts T Cell Expansion Potential,” Frontiers in Cell and Developmental Biology, vol. 9, p. 847, 2021, doi: 10.3389/fcell.2021.648925.

[3] E. J. Wherry and M. Kurachi, “Molecular and cellular insights into T cell exhaustion,” Nat Rev Immunol, vol. 15, no. 8, pp. 486–499, Aug. 2015, doi: 10.1038/nri3862.

[4] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” arXiv:1802.03426 [cs, stat], Sep. 2020, Accessed: Dec. 24, 2021. [Online]. Available: http://arxiv.org/abs/1802.03426

[5] F. A. Wolf, P. Angerer, and F. J. Theis, “SCANPY: large-scale single-cell gene expression data analysis,” Genome Biology, vol. 19, no. 1, p. 15, Feb. 2018, doi: 10.1186/s13059-017-1382-0.

[6] T. A. Geddes et al., “Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis,” BMC Bioinformatics, vol. 20, no. 19, p. 660, Dec. 2019, doi: 10.1186/s12859-019-3179-5.

[7] D. Tran, H. Nguyen, B. Tran, C. La Vecchia, H. N. Luu, and T. Nguyen, “Fast and precise single-cell data analysis using a hierarchical autoencoder,” Nat Commun, vol. 12, no. 1, p. 1029, Feb. 2021, doi: 10.1038/s41467-021-21312-2.

[8] M. B. Badsha et al., “Imputation of single-cell gene expression with an autoencoder neural network,” Quant Biol, vol. 8, no. 1, pp. 78–94, Mar. 2020, doi: 10.1007/s40484-019-0192-7.

[9] J. Zhao, N. Wang, H. Wang, C. Zheng, and Y. Su, “SCDRHA: A scRNA-Seq Data Dimensionality Reduction Algorithm Based on Hierarchical Autoencoder,” Frontiers in Genetics, vol. 12, p. 1485, 2021, doi: 10.3389/fgene.2021.733906.

[10] A. F. Rendeiro et al., “Chromatin mapping and single-cell immune profiling define the temporal dynamics of ibrutinib response in CLL,” Nat Commun, vol. 11, no. 1, p. 577, Jan. 2020, doi: 10.1038/s41467-019-14081-6.

[11] X. Wang et al., “Single-Cell RNA-Seq of T Cells in B-ALL Patients Reveals an Exhausted Subset with Remarkable Heterogeneity,” Advanced Science, vol. 8, no. 19, p. 2101447, 2021, doi: 10.1002/advs.202101447.

[12] M. Andreatta, J. Corria-Osorio, S. Müller, R. Cubas, G. Coukos, and S. J. Carmona, “Interpretation of T cell states from single-cell transcriptomics data using reference atlases,” Nat Commun, vol. 12, no. 1, p. 2965, May 2021, doi: 10.1038/s41467-021-23324-4.