Novel deep neural network to transcriptionally classify age state of naïve CD4 T cells

Authors: Dhanesh Patel

Abstract: Naïve T cells are critical in mounting an adaptive immune response against a novel foreign antigen. Numbers of naïve T cells decline with age, attributable to convoluting thymus, reduced thymic output, and a large number being converted to the memory pool. Whether naïve T cells are transcriptionally different across different ages remains unknown. Here, using a public scRNA-seq dataset of the spleen made available by the Tabula Muris Consortium, we tackled this question, focusing on naïve CD4 T cells. First, naïve T cells were identified based on provided labels and ImmGen datasets of characteristic gene expression of naïve CD4 T cells relative to other immune cells and T cells. Next, using the most highly variable genes (n=3707), a deep neural network (DNN) model was constructed, structured as an autoencoder (a smaller middle layer surrounded by 2 larger layers). The DNN model had an unseen test set performance of 72.6%. Furthermore, uniform manifold approximation and projection (UMAP) plotting of the embedding of each layer revealed separation characteristics of each layer, that all performed better than plotting of the 3707 highly variable genes. The trained DNN model demonstrates the power of a simple autoencoder DNN structure for curated biological classification objectives, which can potentially be applied to other immune cell subsets and biological classifications, for example levels of stress or urban/rural environment.

Introduction: T cells are critical for mounting a successful immune response against a plethora of pathogens. T cells recognize pathogens using unique T cell receptors (TCR) present at their cell surface. For any given pathogen, only a small fraction of T cells respond and become activated during an infection because any given specificity is rare within the T cell population[1]. Interestingly, naive T cells – those T cells that have not yet recognized any foreign antigen – are critical for defending us against future threats. As both humans and mice age, our naïve T cell numbers decline and become less responsive[2]. The erosion of the naïve T cell repertoire renders both species less capable to respond to pathogens as they age. Understanding transcriptional drivers of naïve T cell aging would allow us to tailor therapies to address naïve T cell deficiencies.

T cell repertoire-wide gene expression should be studied with scRNA-seq because of the underlying transcriptional heterogeneity in naïve T cells, which is due to:

(1) Each TCR being unique and belonging to a population-wide spectrum of reactivity for their ligands - peptides presented by major histocompatibility complex proteins (pMHC).

(2) TCR interactions with self-peptides being required for survival signals[3,4] and modulating responsiveness[5].

Several studies have looked at whether processes implicated in the wider aging context apply specifically to naïve T cells: telomere attrition, DNA damage, dysregulated metabolic pathways[6]. However, none of these studies use an unbiased transcriptomics approach or consider whether these affects occur in a heterogeneous manner. Moreover, whether any genes or transcriptomic programs are driving any age-associated changes in the naïve T cell compartment remains unknown.

The Tabula Muris Consortium has done single cell and bulk RNA-seq on 23 organs from one month to 30 months, providing a valuable resource to investigate the naïve T cell population over time[7]. In particular, the spleen is a lymphoid organ that harbors a large representative population of circulating naïve T cells.

Deep neural networks (DNNs) are a methodology that can learn to classify objects while minimizing errors[8]. They take input variables and output categorical labels and try to optimize a set of weights associated with nodes in a series of hidden layers via backpropagation[8]. In scRNA-seq analysis, DNNs can be used as a classifier to properly assign cell labels (e.g., age) when gene expression or latent variables are inputted. Autoencoders are a category of machine learning techniques that uses a neural network to compress (encode) a series of inputs into a latent space. For example, layers of a neural network begin with a larger layer (high number of nodes), followed by a smaller compressed layer (lower number of nodes) - forcing learning to the most important features of the dataset, thereby reducing the noise in the dataset – followed by a final larger layer to decompress the data. There can be multiple layers in between these 3 layers that provide gradations of compression and decompression.

Here, the naïve CD4 T cells were identified using Tabula Muris provided labels as well as gene expression data from ImmGen. Using the most highly variable genes (n=3707) of the naïve T cells, a simple autoencoder shaped DNN model was constructed. The DNN model had an unseen test set performance of 72.6%. Furthermore, uniform manifold approximation and projection (UMAP) plotting of the embedding of each layer revealed separation characteristics of each layer, that all performed better than plotting of the 3707 highly variable genes. Shapley feature extraction identified a gene set that was distinct by genes identified using the statistical Wilcoxon rank sum test.

Results: To begin, 35718 cells were available from the Tabula Muris spleen dataset, of which 4705 were T cells (**Fig. 1a**) as annotated by expression of Cd3e, Cd3g and absence of Klrk1 (expressed by mature and immature NKT cells) (**Fig. 1b**). Re-clustering the cells revealed 2 T cell subpopulations, the CD4 and CD8 T cells (**Fig. 1c**). As age-associated gene expression changes might be different in each compartment, the larger CD4 T cell population (Leiden clusters 1, 2, and 3) (**Fig. 1d**) will be further analyzed. Leiden clustering revealed 4 clusters (**Fig. 1e**). ImmGen query of microarray gene expression

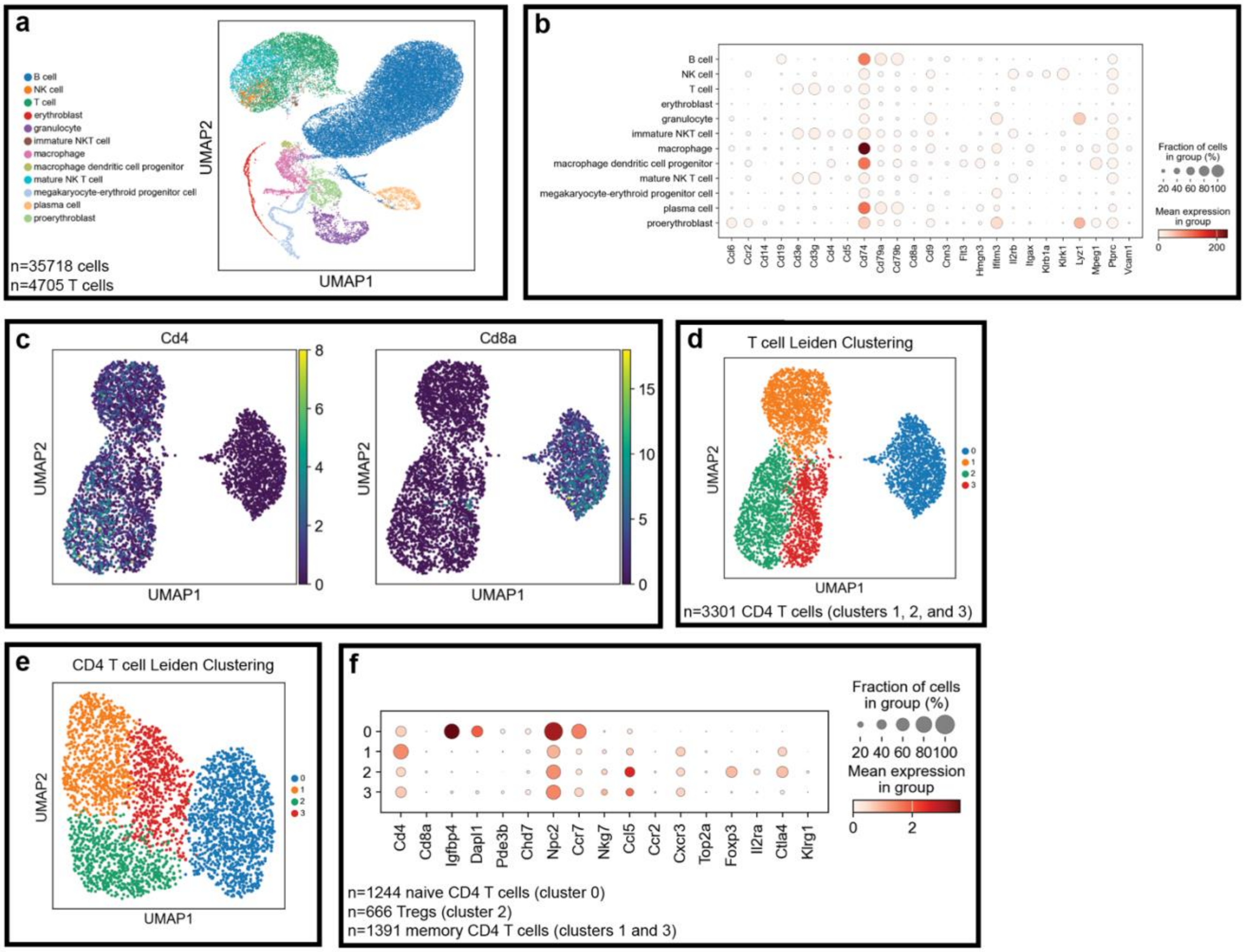
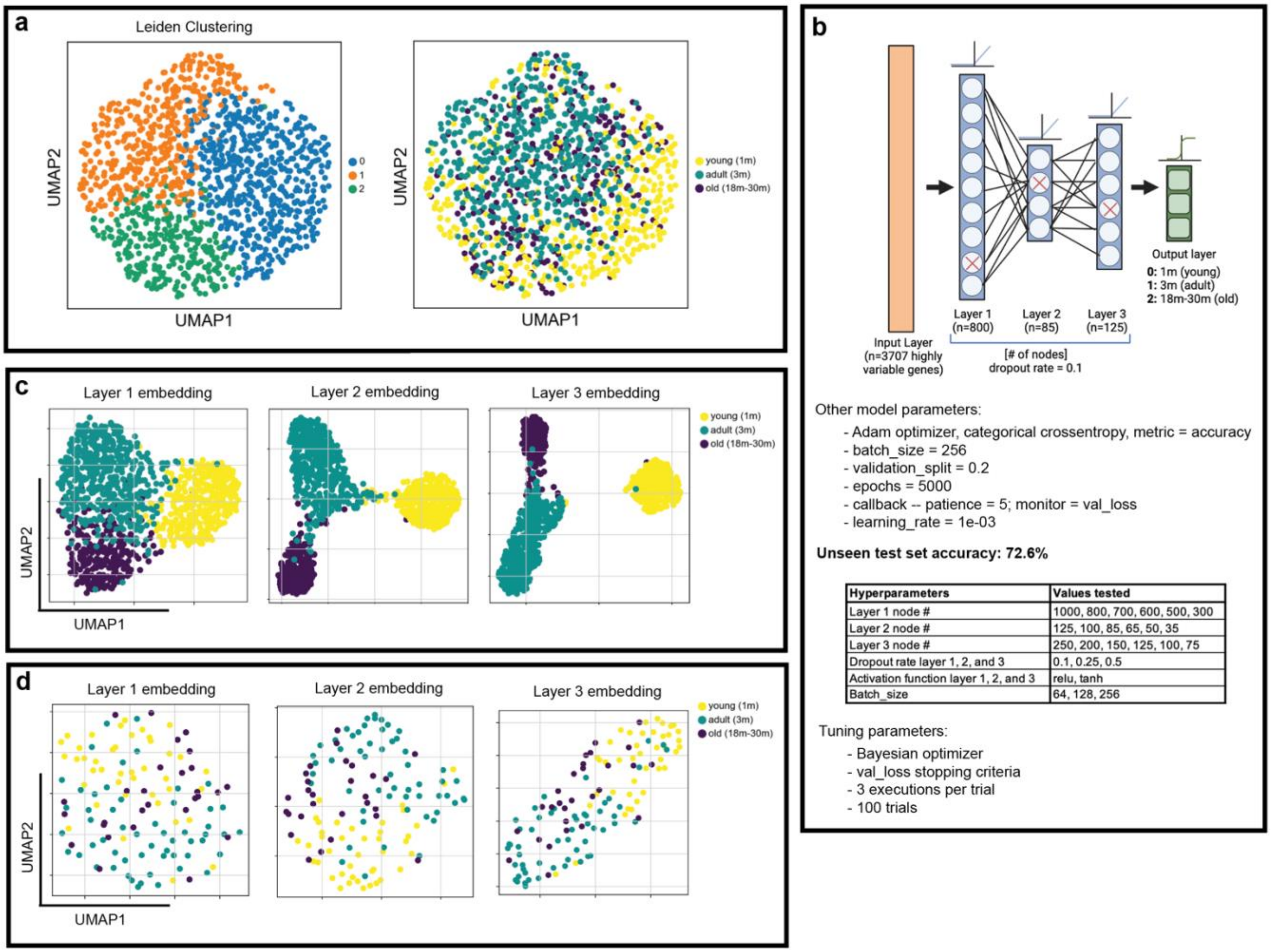


Figure 1. Naive CD4 T cell cluster identification from Tabula Muris spleen scRNA-seq dataset. (**a**) UMAP of all cell populations in the spleen, identified by Tabula Muris Consortium. (**b**) Genes used for cell subset identification as seen in (a), as described by Tabula Muris Consortium. (**c**) UMAP re-clustering of T cells, plotted by Cd4 and Cd8a gene expression. (**d**) Leiden clustering of cells as seen in UMAP of (c) identifies 4 clusters. (**e**) UMAP and Leiden re-clustering of CD4 T cells, identified in (d). (**f**) Characteristic gene expression of naive CD4 T cells, regulatory T cells (Tregs), and memory CD4 T cells using ImmGen microarray expression data.

(comparing naïve spleen CD4 T cells, memory spleen CD4 T cells, and Foxp3+(GFP+) CD25+ Tregs) revealed that cluster 0 contains 1244 cells that naïve CD4 T cells (higher Igfbp4, Dapl1, Npc2, and Ccr7) (**Fig. 1f**).

Preliminary Leiden re-clustering of the naïve CD4 T cells (cluster 0 identified in **Fig. 1e-f**), revealed a cloud of cells with poor cluster separation of the 3 clusters identified (**Fig. 2a**, left). In addition, separation of ages binned in the young (1-month-old), adult (3-month-old), and old (18- to 30-month-old) does not reveal any distinct separation in the UMAP space (**Fig. 2a**, right). A deep neural network that draws upon the most highly variable genes (n=3707 genes) as input and age labels as output was considered for three reasons. Firstly, the data would benefit from an autoencoder shaped structure to eliminate noise. Second, a model to predict naïve CD4 T cell ages could be used on unclassified T cells or contexts where aging state of naïve CD4 T cells would want to be queried (autoimmunity, cancer, natural aging, etc.). Third, age-associated gene expression changes of naive CD4 T cell might be revealed in a manner not possible with other methods (e.g., statistical Wilcoxon rank-sum test of genes expressed in one cluster/group vs. all others). Using tensorflow keras package, an autoencoder shaped DNN was constructed, trained on 90% of the dataset, tuned using a Bayesian optimizer, and its performance tested on a 10% test dataset (**Fig. 2b**). An unseen test set accuracy of 72.6% was noted. To determine what each layer was learning, the cell state outputs of each hidden layer was predicted



Other model parameters:

- Adam optimizer, categorical crossentropy, metric = accuracy
- batch_size = 256
- validation_split = 0.2
- epochs = 5000
- callback -- patience = 5; monitor = val_loss
- learning_rate = 1e-03

Unseen test set accuracy: 72.6%

Hyperparameters	**Values tested**
Layer 1 node #	1000, 800, 700, 600, 500, 300
Layer 2 node #	125, 100, 85, 65, 50, 35
Layer 3 node #	250, 200, 150, 125, 100, 75
Dropout rate layer 1, 2, and 3	0.1, 0.25, 0.5
Activation function layer 1, 2, and 3	relu, tanh
Batch_size	64, 128, 256

Tuning parameters:

- Bayesian optimizer
- val_loss stopping criteria
- 3 executions per trial
- 100 trials

Figure 2. Auto-encoder structured DNN can predict and learn age separation of naive CD4 T cells. (**a**) Leiden re-clustering of naive CD4 T cells (left) and coloring of cells by age (right). (**b**) Autoencoder-shaped DNN structure, all parameters of the model, as well as tuning parameters tested. (**c**) UMAP plotting of each layer's embedding by predicting each hidden layer's output for the training set cells. (**d**) Same as (c), but for the test set cells.

from the model and plotted on UMAPs, with cells colored by age for the training set (**Fig. 2c**) and test set (**Fig. 2d**). Layer 1 reveals the start of cell age separation, accentuated strongly in layer 2, and further improved on in layer 3 (**Fig. 2c-d**).

SHapley (hereafter referred to as Shapley) feature importance was used to determine what the model learnt[9]. The DeepExplainer algorithm was used to extract the feature importance (**Fig. 3a**). Interestingly, the top 20 genes obtained using Shapley (**Fig. 3a**) are quite different from the top 20 genes obtained from the Wilcoxon rank-sum test (**Fig. 3b**). The top 100 Shapley genes of each age group are the genes (features) that are driving the model's age prediction for that age state. While some of the genes are shared between all age states (**Fig. 4a**, top left) or two age states, many of them are uniquely present in the top 100 Shapley list for any given age group (**Fig. 4a**, rest). Fascinatingly, there is very little overlap between the Shapley and Wilcoxon genes. The top 10 Wilcoxon genes are overall expressed at a much higher fraction of all cells (**Fig. 4b**), but do not obviously help in identifying unique age states a lot better than Shapley genes (**Fig. 4a**). Interesting, the fraction of distinct genes between age states (either in fraction of cells expressing the gene or magnitude of gene expression), appears to be a lot higher in the common gene set of Wilcoxon and Shapley genes (**Fig. 4c**) compared to the top Wilcoxon or Shapley genes alone.

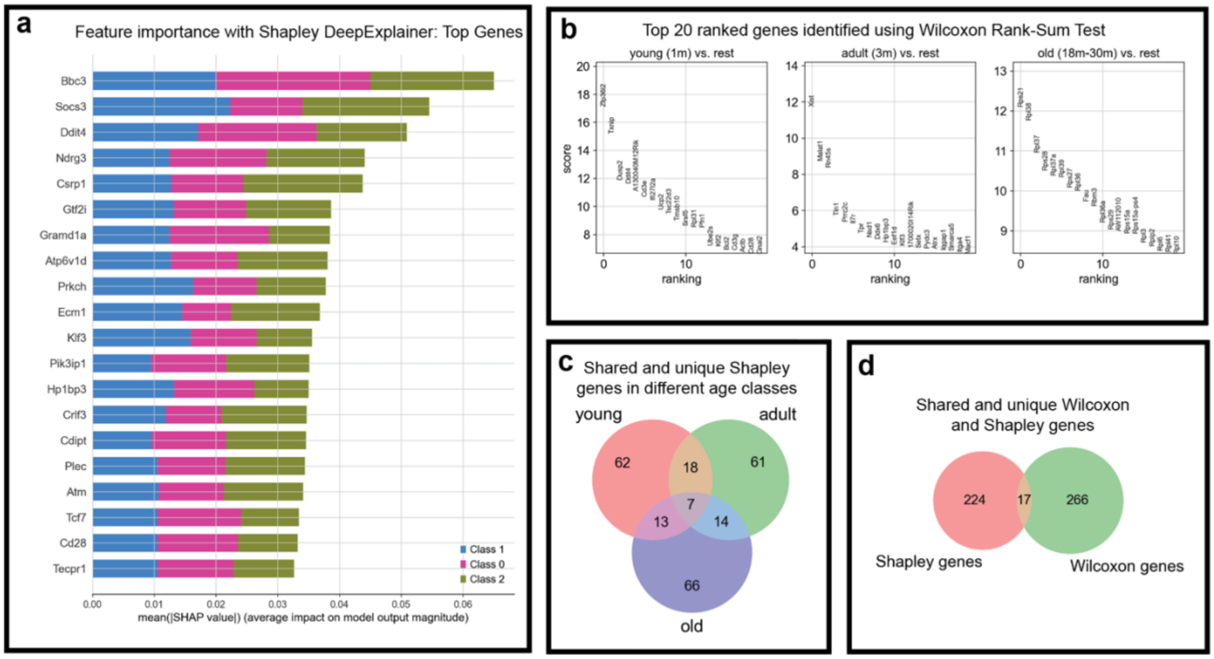


Figure 3. Model's feature importance reveals distinct genes that contributes to age prediction, compared to Wilcoxon identified genes. (**a**) Top 20 genes of the DNN identified with Shapley DeepExplainer feature importance model. (**b**) Top 20 ranked genes expressed in one age group compared to all others identified using Wilcoxon rank-sum Test. (**c**) Venn diagram of the top 100 Shapley genes of each age group. (**d**) Venn diagram of the top Shapley genes (as in (c)) and the top 100 Wilcoxon genes that are most highly expressed in one age group compared to all others.

Discussion: Naïve T cells are a critical component of the adaptive immune system's arsenal. Many physiological changes occurs during aging[10], including to the cellular and stromal components of the immune system[11]. Age-related physiological changes might have effect on the naïve T cell compartment, yet these remain undefined.

Here, age-related transcriptional changes to mouse naïve T cells were queried. To focus our analysis, naïve CD4 T cells were identified from the Tabula Muris Consortium scRNA-seq dataset. Preliminary UMAP visualization did not reveal any age-related separation. To try to reveal age-specific transcriptional signatures, a supervised methodology was employed: an autoencoder-structured DNN was used to try to distinguish and predict the age states. After hyperparameter tuning, a model with unseen test set accuracy performance of 72.6% was obtained.

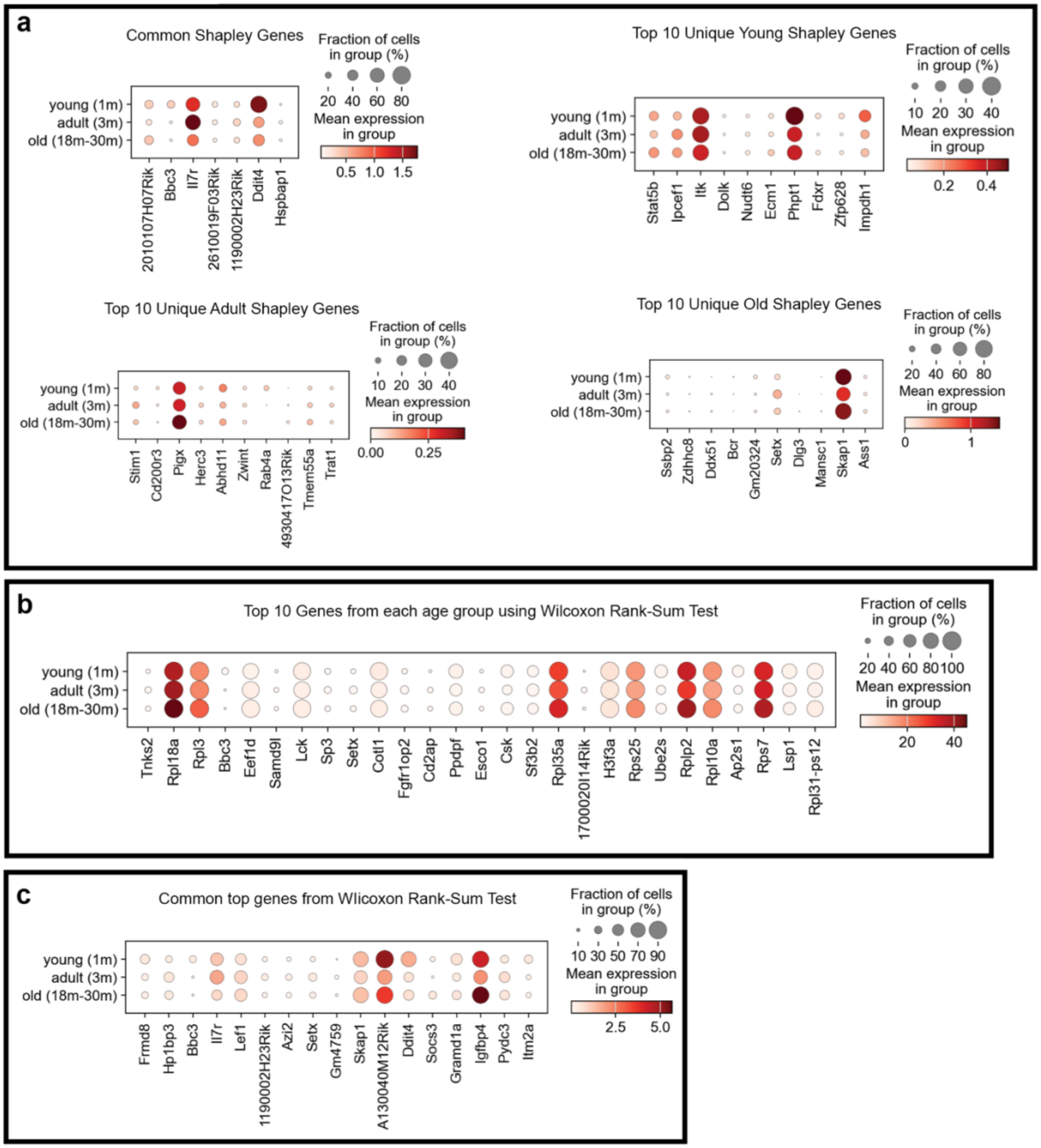


Figure 4. Shapley and Wilcoxon genes' expression patterns in naive CD4 T cells. (**a**) Common Shapley genes' expression (top left) as well as top 10 unique in young (top right), adult (bottom left), and old (bottom right) ages. (**b**) Top 10 ranked genes' expression in each age group compared to all others identified using Wilcoxon rank-sum Test. (**c**) Gene expression of the common (Shapley and Wilcoxon) 17 ranked genes' expression in each age group.

Fascinatingly, predicting the hidden layer embeddings using both the training and test sets revealed how each layer learns to separate the data. Forcing the network to condense its number of nodes in layer 2 appears to have helped the model learn a lot of information about the cell age state. Interestingly, in both the training and test hidden layer embedding, most of the errors appears to be the adult (3m) cells being misclassified in the young (1m) or old (18-30m), which might be due to the natural biological progression of some adult cells transitioning from or to the young and old age states, respectively, and therefore might not represent a biological error in prediction ability of the model.

To determine what features (genes) are driving the prediction ability, and therefore transcriptional features that can distinguish the cell states, feature importance was extracted with Shapley and the DeepExplainer algorithm. In contrast, statistical methods (e.g. the Wilcoxon rank sum test) to find genes that are significantly different in one group compared to another(s) can also reveal biologically significant transcriptional differences across different age groups. Interestingly, the genes identified with the two methods overlap very little. Either of the gene sets reveal many genes that are significantly different across the age groups, but there are very few genes that are commonly found in both gene sets. This might present one way to narrow down list of candidate genes for subsequent biological or experimental query.

In summary, a DNN was constructed with a high prediction accuracy for different age states of naïve T cells. Furthermore, using DNN alongside a traditional statistical approach reveals a powerful method to reduce large gene lists to genes that are most likely to be biologically significant for further query.

Data availability: Tabula Muris Consortium scRNAseq dataset can be found in the following link: https://tabula-muris.ds.czbiohub.org/. All code used in the analysis can be found here: https://github.com/DhaneshPatel/Tabula-Muris-Naive-CD4-T-cell-Deep-Neural-network.

References:

1 Moon, J. J. *et al.* Naive CD4<sup>+</sup> T Cell Frequency Varies for Different Epitopes and Predicts Repertoire Diversity and Response Magnitude. *Immunity* **27**, 203-213, doi:10.1016/j.immuni.2007.07.007 (2007).

2 den Braber, I. *et al.* Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity* **36**, 288-297, doi:10.1016/j.immuni.2012.02.006 (2012).

3 Labrecque, N. *et al.* How Much TCR Does a T Cell Need? *Immunity* **15**, 71-82, doi:https://doi.org/10.1016/S1074-7613(01)00170-4 (2001).

4 Polic, B., Kunkel, D., Scheffold, A. & Rajewsky, K. How alpha beta T cells deal with induced TCR alpha ablation. *Proc Natl Acad Sci U S A* **98**, 8744-8749, doi:10.1073/pnas.141218898 (2001).

5 Stefanová, I., Dorfman, J. R. & Germain, R. N. Self-recognition promotes the foreign antigen sensitivity of naive T lymphocytes. *Nature* **420**, 429-434, doi:10.1038/nature01146 (2002).

6 Goronzy, J. J., Fang, F., Cavanagh, M. M., Qi, Q. & Weyand, C. M. Naive T cell maintenance and function in human aging. *J Immunol* **194**, 4073-4080, doi:10.4049/jimmunol.1500046 (2015).

7 A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590-595, doi:10.1038/s41586-020-2496-1 (2020).

8 Kufel, J. *et al.* What Is Machine Learning, Artificial Neural Networks and Deep Learning?-Examples of Practical Applications in Medicine. *Diagnostics (Basel)* **13**, doi:10.3390/diagnostics13152582 (2023).

9 Lundberg, S. M. & Lee, S.-I. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777 (Curran Associates Inc., Long Beach, California, USA, 2017).

10 Boss, G. R. & Seegmiller, J. E. Age-related physiological changes and their clinical significance. *West J Med* **135**, 434-440 (1981).

11 Cakala-Jakimowicz, M., Kolodziej-Wojnar, P. & Puzianowska-Kuznicka, M. Aging-Related Cellular, Structural and Functional Changes in the Lymph Nodes: A Significant Component of Immunosenescence? An Overview. *Cells* **10**, doi:10.3390/cells10113148 (2021).