Wide and Deep Learning for Automatic Cell Type Identification

Christopher M. Wilson[1], Brooke L. Fridley[1], José Conejo-Garcia[2], Xuefeng Wang[1,*], Xiaoqing Yu[1,*],

1 Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612, USA
2 Department of Immunology, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612 Tampa, FL 33612, US

*** xuefeng.wang@moffitt.org, xiaoqing.yu@moffitt.org**

Abstract

Cell type classification is an important problem in cancer research, especially with the advent of single cell technologies. Correctly identifying cells within the tumor microenvironment can provide oncologists with a snapshot of how a patient's immune system is reacting to the tumor. Wide deep learning (WDL) is an approach to construct a cell-classification prediction model that can learn patterns within high-dimensional data (deep) and ensure that biologically relevant features (wide) remain in the final model. In this paper, we demonstrate that the use of regularization can prevent overfitting and adding a wide component to a neural network can result in a model with better predictive performance. In particular, we observed that a combination of dropout and ℓ_2 regularization can lead to a validation loss function that does not depend on the number of training iterations and does not experience a significant decrease in prediction accuracy compared to models with ℓ_1, dropout, or no regularization. Additionally, we show WDL can have superior classification accuracy when the training and testing of a model is completed data on that arise from the same cancer type, but from different platforms. More specifically, WDL compared to traditional deep learning models can substantially increase the overall cell type prediction accuracy (41 to 90%) and T-cell sub-types (CD4: 0 to 76%, and CD8: 61 to 96%) when the models were trained using melanoma data obtained from the 10X platform and tested on basal cell carcinoma data obtained using SMART-seq.

1 Introduction

Immunology is quickly becoming a popular area of study in cancer research and offers an opportunity to expand our understanding and ability to treat patients. Estimating the immune composition of an individual's tumor has been the focus of several studies which have developed deconvolution methods [17, 24] to estimate the cellular composition of the tumor micro-environment with bulk RNA expression data. Recently with the advent of single cell sequencing researchers are now able to measure gene expression in individual cells within the tumor-microenviroment and classify cells using heirarchical clustering and correlation-based methods [1, 2, 7]. Cell type classification

can be conducted by constructing visualizations such as t-Distributed Stochastic Neighbor Embedding (t-SNE) [28] or Uniform Manifold Approximation and Projection (UMAP) [16] plots to define clusters and assign these clusters to different cell types based on enriched canonical markers. However, a major drawback of this canonical process is that it heavily relies on the researchers' knowledge on the cell-type-specific signature genes, and it can become arbitrary when making conclusions based on only a handful of genes. Also, the cell type marker genes are cancer type-specific and may not generalize to other datasets [31]. In addition, discriminating between fine immune cell sub-types, such as exhausted CD8 T cells vs. activated CD8 T cells, effector CD4 T cells vs. naive CD4 T cells is a much more challenging task due to the lack of universal marker genes.

Identification of highly specific cell types is now possible with the development of single cell RNA-sequencing technology. However, a challenge in cell annotation in single cell RNA-sequencing is that transcription profiles are difficult to transfer between different platforms. Multiple platforms have been developed for single cell RNA-sequencing including SMART-seq [18], CEL-seq [11], Fluidigm C1 [13], SMART-seq2 [19], and more advanced droplet-based platforms including Drop-seq [15] and 10X Genomics Chromium system [35], etc. The two most commonly used platforms are SMART-seq/SMART-seq2 and 10X. The 10X platform is a droplet-based approach which generates a unique molecular identifier (UMI) at $5'$ or $3'$ ends to diminish the sequencing reads representation biases due to library amplification. On the other hand, SMART-seq and SMART-seq2 are designed to generate full-length cDNA. Droplet-based $5'$ or $3'$-tag methods like 10X can capture much more cells which in turns can give better overview of the heterogeneity within population; while a full-length proposal like SMART-seq is better suited for studies concerned with isoforms, splicing or gene fusion. Due to the differences in how they amplify the mRNA transcripts, the data generated from these platforms are not directly comparable, which presents great challenge to the integrated cell type identification in cross-platform datasets [4, 20, 29, 34]. Therefore, there is a great need for automatic cell identification method that be used across studies, single cell platforms, and cancer types.

While there are many different single cell RNA-sequencing platforms whose results are on different scales and not directly comparable, the underlying gene to gene relationships should be consistent and navigating these relationships may allow for borrowing of information from different technologies. Deep learning brings us the possibility to explore and summarize complex highly non-linear relationships into high-level features from high throughput data sources. Deep learning is a powerful machine learning technique that is often used in visual recognition [12, 14], natural language processing [21, 32], and starting to infiltrate the realm of cancer research [3, 5, 25]. Deep learning learns patterns in data by using neural networks with many layers of nodes which transform the output model of the nodes from the previous layer with non-linear functions. The coefficients output from each node are augmented using gradient descent in order to optimize the prediction error of the network.

Wide and deep learning (WDL) combines a deep neural network with a generalized linear model (GLM) based on a small set of features. Deep learning tends to generalize patterns in the data, while in contrast GLMs may only memorize the patterns in the data. WDL has been shown to be an effective tool in recommender systems [6]. Specifically, we propose utilizing a deep learning model which can leverage large dimensional data (deep), as well as, incorporate a few known biologically relevant genes in the last hidden layer of a neural network to emphasize their biological importance (wide). The wide part of the model allows us make cell type classification more precise and fine tuned to classify more specific immune cell sub-types such as distinguishing activated from exhausted CD8 T cells.

This paper serves two purposes. First it provides some background information about deep learning specifically focusing on regularization methods to avoid overfitting the model. Models are trained, validated, and subsequently used to classify cells from the same dataset. This scenario is realistic since it is possible that some hospitals may not have the resources needed for generating large amounts of data to build their own model. In addition, in many clinical studies the patients' tumor samples are collected over a fairly long period of time (years) in several batches. Waiting till sample collection is finished before single cell RNA-sequencing data analysis is not realistic. It will be extremely helpful to train a deep learning model using samples collected at earlier time points and subsequently apply it to later samples. Second, this paper provides an illustration of how incorporating known biologically relevant biomarkers can be used to transfer knowledge. In this scenario, we explore the possibility to transfer cell type annotations across different single cell RNA-sequencing platforms, which can help make full use of the enormous publicly available single cell RNA-seq data that are generated by different technologies.

In the methods section, we will describe the data single cell RNA-seq datasets used in study, provide background about deep learning, and wide and deep learning. Then in the results section, we will present results from training and testing the models in the two scenarios. Finally, we make some concluding comments and discussion in the discussion section.

2 Methods

2.1 Chang Data

Chang et al. [31] conducted droplet-based 10X 5' single-cell RNA-sequencing on 79,046 cells from primary tumors of 11 patients with advanced basal cell carcinoma before and after anti-PD-1 treatment. In total, RNA profiles from 53,030 malignant, immune and stromal cells, and 33,106 T cells were obtained from single cell RNA-sequencing. The cell types of interest were T cells, B cells, nature killer (NK) cells, macrophages, cancer associated fibroblasts (CAFs), endothelial cells, plasma cells, melanocytes, and tumor cells. The T cells were further classified into regulatory (Tregs), follicular helpers (T_{FH}), T helper 17 (T_h17), naive T cells, activated CD8, exhausted CD8, effector CD8, and memory CD8 T cells (Supplementary Figure 1).

2.2 Tirosh data

Tirosh et al. [26] applied SMART-seq to 4645 single cells isolated from 19 freshly procured human melanoma tumors, profiling T cells, B cells, NK cells, macrophages, endothelial cells, CAFs, and melanoma cells. To further analyze the T cell sub-types, we downloaded the log-transformed TPM (Transcripts per Million reads) gene expression values provided by the study and imported them to Seurat [23]. S and G2/M cell cycle phase scores were assigned to cells based on previously defined gene sets [27]using CellCycleScoring function. Scaled z-scores for each gene were calculated using ScaleData function by regressing the S and G2/M phases scores. Shared nearest neighbor (SNN) based clustering method was used to identify clusters based on the first 30 principle components computed from scaled data with resolution = 1. UMAPs were generated using the same principle components with perplexity = 30 and used for all visualization. Clusters were annotated by identifying differentially expressed marker genes for each cluster and comparing to known cell type markers and markers reported by Tirosh et al. From this analysis, we confirmed the cell annotation provided by Tirosh et al., and were able to further identify CD4+ T cells and CD8+ T cells.

2.3 Background for Classification Problems with Deep Learning

Deep neural networks (DNN) are able to learn and condense highly non-linear features (genes) into a high level summary through the use of composition of functions. These functions are dot products that undergo a non-linear transformation and are then passed into another function in the next layer. There are three types of layers in a DNN which are input, hidden, and output layers. Each node in the input layer corresponds to the expression of a single gene. Information from the input layer is passed to each node in the a hidden layer which optimizes the weights in a dot product to maximize the cell type classification accuracy. The nodes in the output layer produce the probability that a cell is classified as a specific type. The architecture of a generic DNN with two hidden layers is seen in Figure 1. For multi-class classification the objective function to be minimized is so called cross-entropy function, $H(P, Q)$,

$$\frac{1}{n}\sum_{i=1}^{n} H(P,Q) = -\frac{1}{n}\sum_{i=1}^{m}\sum_{i=1}^{m} P(x_i = j)\log(Q(x_i = j)), \tag{1}$$

where m is the number of cell types, n is the sample size, P is the target probability distribution and Q is the predicted cell type probability distribution. This function is minimized by gradient descent which is computed by iterating the chain rule over all layers of the model. The architecture of a DNN is complex and requires careful tuning. Some examples of tuning parameters in a DNN are number of samples used for stochastic gradient, iterations to train the model, and nodes in each layer [9, 10].

A challenge to training a deep learning model is to ensure that the results can be generalized to new data sets. One of the simplest ways to prevent overfitting is use to reduce the number of hidden layers or nodes which in turn decreases the number of parameters estimated by the model. Another technique is dropout which randomly deletes a specified proportion of nodes from each layer in the neural network. By deleting different sets nodes in each iteration the model is trained on different sub-networks and becomes less sensitive to the specific weights of nodes. Dropout can speed up the training of a DNN, however, it may require more iterations to train the network. It is recommended that the percentage of nodes to delete from each layer should be between 20-50% [22] Lastly, constraints can be imposed on the weight vector of each node requiring the norm be small. The regularizers work in a similar way to lasso or ridge penalization in a regression setting where there is an additional parameter which changes the influence of the penalization term. The aim is either to keep the value of the weights small or push as many as possible to zero (lasso). Elastic net penalization has also been used which allows for a balance between the ℓ_1 (lasso) and ℓ_2 (ridge) penalty [9, 30, 33].

Understanding which genes were influential to a successful cell type classification model is important for validating the results and can lead to detection of novel genes for future research. Despite many machine learning techniques being seen as 'black boxes', there have been efforts to interpret the results. One simple approach to evaluate the importance of a feature is to calculate dot product of consecutive nodes [8]. This was originally proposed for neural networks with a single hidden layer, but we extend this work to a neural work with two hidden layers

$$w_i = H^{(1)}H^{(2)}O = \sum_{i=i}^{h_1}\sum_{j=i}^{h_2}\sum_{k=1}^{m} H_{ij}^{(1)}H_{jk}^{(2)}O_k \tag{2}$$

where $H_{ij}^{(a)}$ is the coefficient passed from the i^{th} node of the $(a-1)^{st}$ hidden layer to the j^{th} node in the a^{th} hidden layer, and O_k probability that a cell is from the k^{th} cell

	Validation Accuracy	First Layer	Second Layer	Testing Accuracy
No Dropout	94.4	100	50	95.2
Dropout Only	94.2	100	75	94.5
Dropout + ℓ_2	94.3	100	75	93.8
Dropout + ℓ_1	92.9	75	100	92.2

Table 1. Summary of the parameter selection, validation, and testing accuracy for the naive, and wide and deep learning models. The architecture was selected based on the model that highest accuracy when classifying cells in the validation data set. The testing accuracies arise from training a model with the specified architecture with training + validation datasets and testing on previously unused test set.

type. By ranking these weights of each gene we can gain some notion of variable importance in the final classification.

2.4 Emphasizing Important Genes for Improved Cell Type Classifiers

Wide and deep learning (WDL) involves merging a set of features, wide component, with the last hidden layer in a DNN, deep component. Adding these features in the final step will ensure that they are emphasized in the model, since they may be lost due to dropout or assigned with small weights. The wide component is a generalized linear model where the input is a set original features. Wide components tend to memorize the patterns the data, while deep components can generalize non-linear patterns. The architecture of a WDL model is shown in Figure 1. In this study, specific genes that are exclusively expressed by a particular cell type are added to the last hidden layer forcing the model to emphasize them more. This may allow a DNN to produce a more accurate classify cells model than a model constructed with only a deep part, especially in scenarios where the data are obtained from different platforms or cancer types.

3 Results

3.1 Neural Network Tuning

In this section, we want to describe how the hyperparameters (number of nodes, regularization, dropout) were selected. Traditional deep learning models with two hidden layers were constructed with no regularization (No Dropout), 20% dropout for both hidden layers (Dropout Only), 20% dropout and an ℓ_1 regularizer for both hidden layers (Dropout + ℓ_1), and lastly 20% dropout and an ℓ_2 regularizer for both hidden layers (Dropout + ℓ_2). A grid search was employed where the first hidden layer could have 1, 5, 10, 25, 50, 75, 100, 500, 1000 nodes and the second hidden layer could have 1, 5, 10, 25, 50, 75, 100, 500 nodes.

Figure 2 shows the training and validation accuracy (left) and loss (right) when there was the same architecture for each model, and a numerical summary is provided in Table 1 for the architecture leading to the best validation accuracy. Notice that with No Dropout, Dropout Only, and Dropout + ℓ_1 the validation loss increases as the model is trained. On the other hand, the validation accuracy and validation loss remains consistent with the training loss as epochs increase for Dropout + ℓ_2, suggesting that even if the model is trained with an excessive number of iterations that the model performance will not suffer heavily from overtraining. While the No Dropout and Dropout only models had the lowest validation loss and nearly 100% training accuracy,

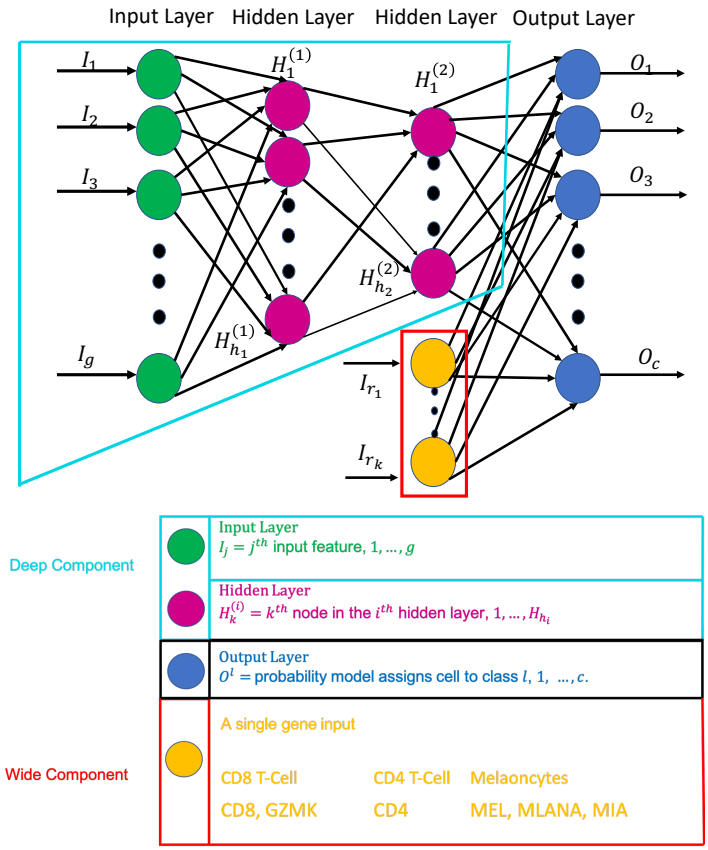

Figure 1. Depiction of a generic wide and deep learning neural network where the wide component is surrounded by the red lines and the deep component is encompassed by the turquoise lines. The deep component equivalent a traditional deep learning model described in sections 2.2 and 3.1. The gene names in yellow correspond to the genes that are used in the wide part of the WDL model in section 3.2.

they are undoubtedly overtrained and will likely not generalize well for future data. Based on these finding, all models discussed in the remainder of this paper will be constructed with dropout and ℓ_2 regularization especially since there is not a significant difference between the testing accuracies of the four methods.

The overall accuracy of the Dropout + ℓ_2 model is 93.8%, with the prediction accuracy of individual cell types ranging from 83 to 100% (Figure 3A). T-cell sub-types are similar in gene expression profiles and are difficult to distinguish. T-cell sub-type classification is commonly done as a second stage of classification where only the T-cells are considered [31]. Figure 3A shows that using a deep learning framework, each T-cell sub-type is classified with at least 83% accuracy, and 5 out of the 7 T-cell sub-types had greater than 91% percent accuracy, and the misclassified cells were classified as another type of T-cell. In single cell RNA-sequencing, the separation between activated CD8 and exhausted CD8 T cells are particularly difficult. The exhausted cells are considered as chronically activated and they also highly express the activation markers such as TNF and IFNG. The subtle difference between activated and exhausted CD8 T cells is the overexpression of exhaustion markers such as TIGIT and HAVCR2. Our deep

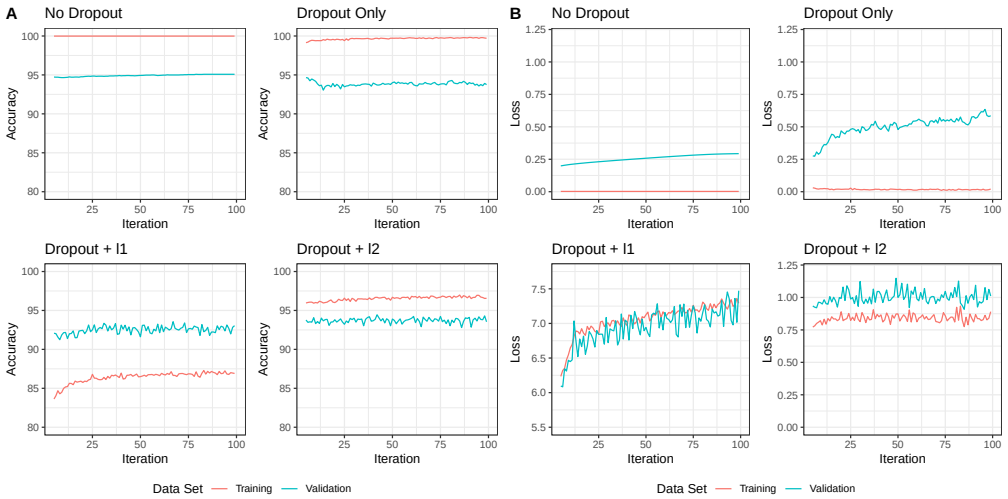

Figure 2. Comparison of the accuracy (A) and categorical cross-entropy loss (B) of four models from section 3.1 with varying methods of regularization and drop out. The plots are for iteration number 5 through 100. The red and turquoise lines correspond to the performance on the training validation set, respectively.

learning model with Dropout+ ℓ_2 setting was able to capture these genes and ranked their importance as 4^{th} and 75^{th} in total of 22,890 genes (Figure 3B, Supplementary Table 1). In addition, the model also put high emphasis on the genes that are typically over-expressed in tissue-resident memory cells, such as LAYN and CXCL13 (Figure 3A), which is consistent with Chang et al. original findings. The cell type classification accuracies for No Dropout, Dropout only, and Dropout + ℓ_1 are included in Supplementary Figure 2.

3.2 Testing on Different Datasets

Both naive and WDL learning models were constructed using the melanoma data generated by Chang for training and basal cell carcinoma data produced by Tirosh for testing the models. Both models were constructed with 100 and 75 nodes for the first and second layers respectively. A comparison of the true cells types and the predicted cell types from the WDL model are shown in Figure 4A and 4B. The naive model had an overall accuracy of 41% which is partly due to the model not classifying any cells as CD4 T cells (Figure 4C). Another discrepancy is that a majority of melanoma cells were classified as CAFs (62%), but the silver lining is that the naive model can distinguish tumoral from stromal cells. A large percentage of NK cells were also classified as CD8-T cells (45%) which is not surprising based on the similarity in cellular function and location in the UMAP in Supplementary Figure 1.

With the addition of the 8 markers listed in Figure 1, the WDL model can better discriminate sub-types of T cells, CD8 T cells and CD4 T cells, with an accuracy of 95.8 and 75.6% respectively (Figure 4B right) and obtained an overall accuracy of 90.1% accuracy (Supplementary Figure 3). The classification accuracy by cell type ranges from 48.6% for NK Cells to 95.8% for CD8 T-cells with 6 of the 8 cell types having greater than 87% accuracy (Figure 4B right). A majority of the misclassified CD8 T-cells were classified classified within the same major cell type. Classification of melanoma cells saw the largest increase in accuracy from 31% to 98%.

In order to understand why the models perform differently, we focus on the differences between specific markers that were highly influential in each model. The

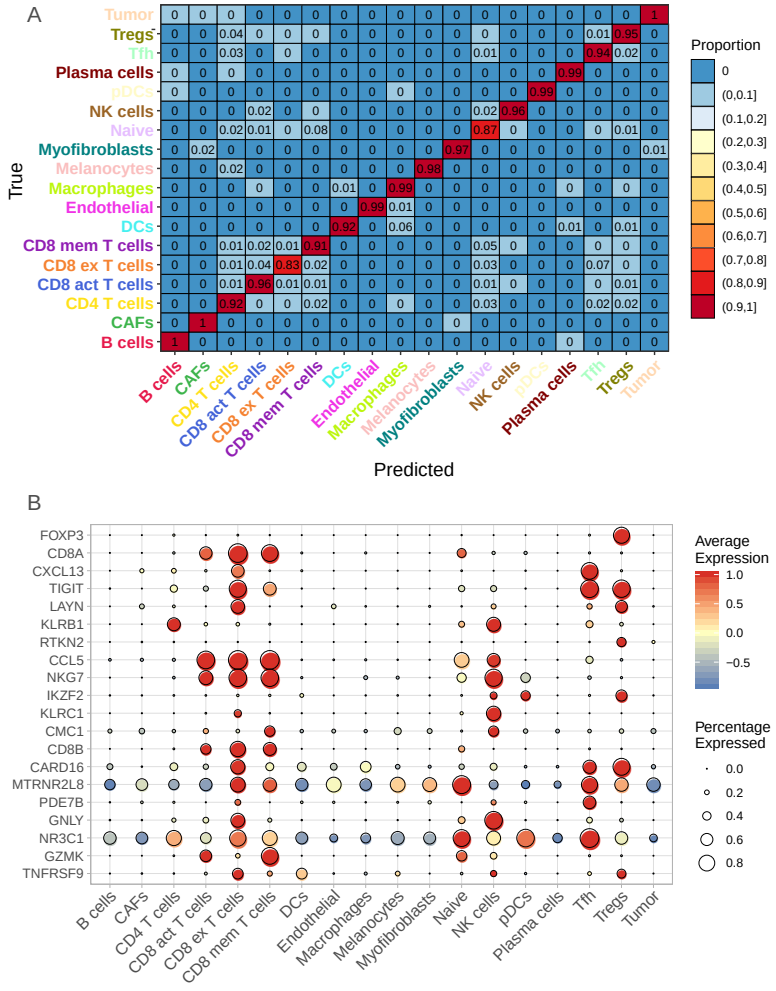

Figure 3. Heatmap of accuracies by cell type for a deep learning model (A), trained and tested on the Chang dataset, with two hidden layers with 100 and 75 nodes respectively, and 20% dropout and ℓ_2 regularization. Average expression of the top 20 most influential genes by cell type (B) where the size of the dot corresponds to the proportion of cells that express this gene and color ranging from blue to red indicating low to high average expression.

importance of the top 20 markers and their importance are displayed in the average expression profiles for the each cell type in Figures 5A and 5B. A total of 8 markers were included in the wide part of the WDL model, and these have by far the largest importance in the model, and four of these (CD8A, CD8B, GZMK, and IL7R) which are all important for identifying CD8 or CD4 T-cell. Thus the importance of these markers explains why the model was able to classify CD8-T cells with an accuracy of 61%. Five out of the top 12 most important markers from the deep component of the WDL model were in the top 20 markers in the naive model with slightly larger weights in the WDL model. Additionally, there were no melanoma markers in the top 20 most influential genes in the naive model leading to many of the melanoma cells being labels as CAFs. Figures 5, Supplementary Figure 4 and Supplementary Table 2 show that are

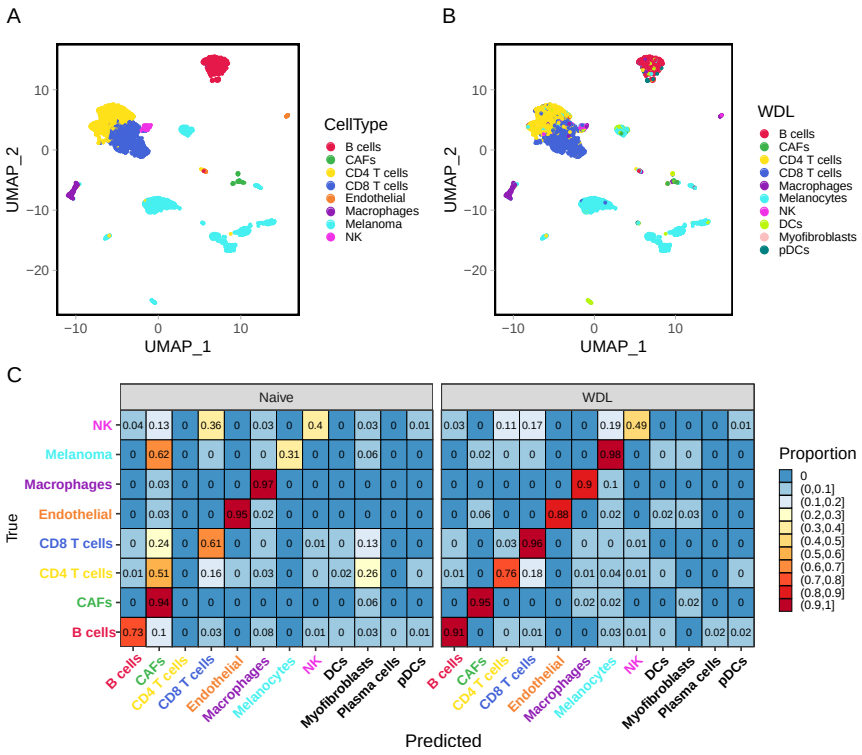

Figure 4. Comparison of the true cell types (A) and predicted cell types from a WDL model (B). Side-by-side comparison of the accuracy by cell type for the naive and WDL models (C).

several genes that are highly influential, yet are not expressed in many cells in the training or test set, such as EOMES, SH2D1B, ENC1, VCAM1. This further illustrates a challenge understanding the importance of markers only a small subset of cells in a particular cell type may express these marker yet the model found them to be influential. Chang identified, after restricting data to only T-cells, EOMES as a marker to distinguish from CD4 memory T-cells other T-cell sub-types, while both deep learning models were able to identify this as in important marker without subsetting the data by major cell type (Supplementary Figure 4).

4 Discussion

WDL presents an opportunity to use a small set commonly known biological markers for cell type classification to allow models to be slightly less data driven. We have illustrated a substantial increase in overall accuracy (41 to 90%) and for T-cell sub-types (CD4 increased from 0 to 76% and CD8 increased from 61 to 96%). We have demonstrated that this can allow for training and testing of models from data obtained from different platforms and types of skin cancer, and even when the target classifications are not the same. Further refinement for classification of fine T-cell sub-types is needed to address questions such as 'how strong is the CD8 T-cell response to a tumor?', i.e. determine the proportion of CD8 T-cells that are exhausted, which are very relevant in cancer research. Additionally, there is a great need to develop systems

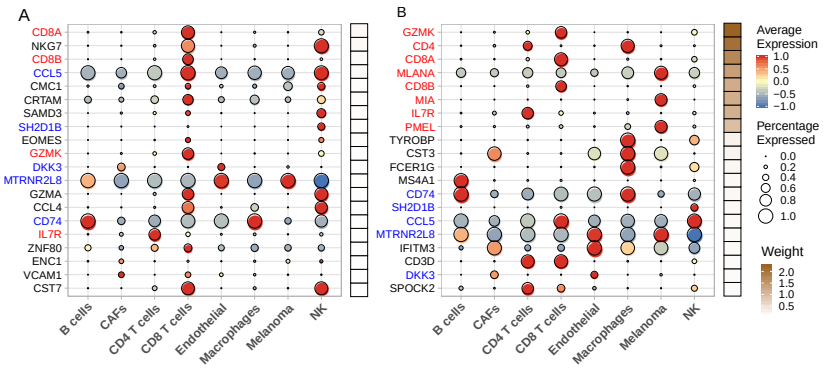

Figure 5. Dot plots for the Tirosh data using both naive (A) and WDL (B) with gene importance weights increasing with brown color scale. The genes names highlighted in red correspond to genes that were included in the wide part of the WDL model, and the blue gene names correspond to the genes that were not in the wide part yet were influential in both the naive and WDL models. Full list of genes and weights is included in Supplemental Table 2.

to transfer knowledge across cancer types. WDL allows the opportunity to address this by including general set of genes for cell type classification and avoiding adding data/cancer specific markers as shown in section 3.2.

In addition to adding a wide component to a DNN there is need for careful consideration for how model is trained to avoid the memorization of data. While dropout is a great tool for making deep learning models more generalizble there are many applications where there is a need for additional steps to avoid overfitting. Regularization is computationally intensive, but makes deep learning models for generalizable to test datasets. Models can very easily be overtrained but a combination of dropout and ℓ_2 regularization can provide a loss function that is stable across the training iterations. Another challenge for deep learning in general is the randomness in the initialization of node weights, dropout, and batches can lead to dramatically different performances for models that are tuned in the same manner and data. Studying an ensemble of deep neural networks could help study the stability of the models and comparing the most important biomarkers in each model can provide further confidence that the markers that are highly influential. Identifying these genes can help clinicians understand commonality between immune cells behavior across cancer types providing better insight and treatment of the cancers themselves.

Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported in part by Institutional Research Grant number 14-189-19 (to XW, XY) from the American Cancer Society, and a Department Pilot Project Award

from Moffitt Cancer Center (to XY). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

All authors read and approved the final manuscript. CW, BLF, JRC, XW, and XY conceived the study. CW, XW and XY designed the algorithm, performed the analyses, interpreted the results and wrote the manuscript.

Acknowledgments

The authors would like to thank Colleagues at Department of Biostatistics and Bioinformatics at Moffitt Cancer Center for providing feedback.

References

1. T. Abdelaal, L. Michielsen, D. Cats, D. Hoogduin, H. Mei, M. J. T. Reinders, and A. Mahfouz. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome Biology*, 20(1):194, 2019.

2. D. Aran, A. P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, S. Chak, R. P. Naikawadi, P. J. Wolters, A. R. Abate, A. J. Butte, and M. Bhattacharya. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2):163–172, 2019.

3. A. Cheerla and O. Gevaert. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14):i446–i454, 07 2019.

4. G. Chen, B. Ning, and T. Shi. Single-cell rna-seq technologies and related computational data analysis. *Frontiers in Genetics*, 10:317, 2019.

5. H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke. The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6):1241 – 1250, 2018.

6. H.-T. Cheng, M. Ispir, R. Anil, Z. Haque, L. Hong, V. Jain, X. Liu, H. Shah, L. Koc, J. Harmsen, and et al. Wide & deep learning for recommender systems. *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems - DLRS 2016*, 2016.

7. J. K. de Kanter, P. Lijnzaad, T. Candelli, T. Margaritis, and F. C. P. Holstege. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Research*, 47(16):e95–e95, 06 2019.

8. G. D. Garson. Interpreting neural-network connection weights. 1991.

9. I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, 2016.

10. A. Gulli and S. Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017.

11. T. Hashimshony, F. Wagner, N. Sher, and I. Yanai. Cel-seq: single-cell rna-seq by multiplexed linear amplification. *Cell Reports*, 2(3):666–673, September 2012.

12. M. T. Islam, B. M. N. Karim Siddique, S. Rahman, and T. Jabid. Image recognition with deep learning. In *2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, volume 3, pages 106–110, 2018.

13. S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11(2):163–166, February 2014.

14. X. Jia. Image recognition method based on deep learning. In *2017 29th Chinese Control And Decision Conference (CCDC)*, pages 4730–4735, 2017.

15. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, May 2015.

16. L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. cite arxiv:1802.03426Comment: Reference implementation available at http://github.com/lmcinnes/umap.

17. A. Newman, C. Liu, M. Green, A. Gentles, W. Feng, Y. Xu, C. Hoang, M. Diehn, and A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *PLoS Medicine*, 12(5):453–457, Apr. 2015.

18. S. Picelli, r. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096–1098, November 2013.

19. S. Picelli, O. R. Faridani, A. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171–181, January 2014.

20. P. See, J. Lum, J. Chen, and F. Ginhoux. A single-cell sequencing guide for immunologists. *Frontiers in Immunology*, 9:2425, 2018.

21. A. R. Sharma and P. Kaushik. Literature survey of statistical, deep and reinforcement learning in natural language processing. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 350–354, 2017.

22. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

23. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, June 2019.

24. G. Sturm, F. Finotello, F. Petitprez, J. D. Zhang, J. Baumbach, W. H. Fridman, M. List, and T. Aneichyk. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics*, 35(14):i436–i445, 07 2019.

25. B. Tang, Z. Pan, K. Yin, and A. Khateeb. Recent advances of deep learning in bioinformatics and computational biology. *Frontiers in Genetics*, 10:214, 2019.

26. I. Tirosh, B. Izar, S. M. Prakadan, M. H. Wadsworth, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, M. Fallahi-Sichani, K. Dutton-Regester, J.-R. Lin, O. Cohen, P. Shah, D. Lu, A. S. Genshaft, T. K. Hughes, C. G. K. Ziegler, S. W. Kazer, A. Gaillard, K. E. Kolb, A.-C. Villani,

C. M. Johannessen, A. Y. Andreev, E. M. Van Allen, M. Bertagnolli, P. K. Sorger, R. J. Sullivan, K. T. Flaherty, D. T. Frederick, J. Jané-Valbuena, C. H. Yoon, O. Rozenblatt-Rosen, A. K. Shalek, A. Regev, and L. A. Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.

27. I. Tirosh, B. Izar, S. M. Prakadan, M. H. Wadsworth, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, M. Fallahi-Sichani, K. Dutton-Regester, J. R. Lin, O. Cohen, P. Shah, D. Lu, A. S. Genshaft, T. K. Hughes, C. G. Ziegler, S. W. Kazer, A. Gaillard, K. E. Kolb, A. C. Villani, C. M. Johannessen, A. Y. Andreev, E. M. Van Allen, M. Bertagnolli, P. K. Sorger, R. J. Sullivan, K. T. Flaherty, D. T. Frederick, J. Jan?-Valbuena, C. H. Yoon, O. Rozenblatt-Rosen, A. K. Shalek, A. Regev, and L. A. Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, Apr 2016.

28. L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

29. B. Vieth, S. Parekh, C. Ziegenhain, W. Enard, and I. Hellmann. A systematic evaluation of single cell rna-seq analysis pipelines. *Nature Communications*, 10(1):4667, 2019.

30. C. Wu, M. J. F. Gales, A. Ragni, P. Karanasou, and K. C. Sim. Improving interpretability and regularization in deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2):256–265, 2018.

31. K. E. Yost, A. T. Satpathy, D. K. Wells, Y. Qi, C. Wang, R. Kageyama, K. L. McNamara, J. M. Granja, K. Y. Sarin, R. A. Brown, R. K. Gupta, C. Curtis, S. L. Bucktrout, M. M. Davis, A. L. S. Chang, and H. Y. Chang. Clonal replacement of tumor-specific t cells following pd-1 blockade. *Nature Medicine*, 25(8):1251–1259, 2019.

32. T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

33. K. Yu, W. Xu, and Y. Gong. Deep learning with kernel regularization for visual recognition. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1889–1896. Curran Associates, Inc., 2009.

34. X. Zhang, T. Li, F. Liu, Y. Chen, J. Yao, Z. Li, Y. Huang, and J. Wang. Comparative analysis of droplet-based ultra-high-throughput single-cell rna-seq systems. *Molecular Cell*, 73(1):130 – 142.e5, 2019.

35. G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049, January 2017.