# Looks Aren't Everything: Combining Cell Morphology with Transcriptomics and Machine Learning for Characterizing Cell State

Cell morphology and cell state are intrinsically linked. Cells undergoing mitosis, for example, are identified by their elliptical morphology. Dead cells can be identified by their circular shape. While the connection between cell phenotype and gene expression is well established for certain states, for others it is less obvious. For example, cells entering G1 do not show any observable morphological changes. Pluripotent stem cells beginning to differentiate into the four germ layers also show no major morphological changes. In order to see these changes, one must either incorporate a fluorescent reporter or fix and stain cells and look at protein or RNA abundance. Connecting both morphology and cell state is essential in the study of cell biology if we are to advance our understanding on how these various cell states affect cell function. One possible solution to this problem is using machine learning as a way to identify more niche cellular morphologies and connect them back to cellular state. In this review, I will be looking at various papers incorporating machine learning techniques to connect cellular morphologies to cellular states and present why the field is shifting towards computation. By viewing both transcriptomic and phenotypic representations of cells and using machine learning as the primary method for linking these two concepts, we can advance our study of cellular states.

These advancements have already improved our understanding on connecting cardiomyocyte maturity to cellular state. In a recent paper by Gerbin et al, they studied the connection between cellular transcriptome abundance and various morphological and cytoskeletal components [1]. To do so, the authors developed a metric to characterize cardiomyocyte maturity based on cellular organization. This allowed them to see what structural elements are key indicators of cellular maturity and how they relate to mRNA abundance. I found this paper insightful in its study of how these structural elements connect to transcriptomics since this connection was previously ill-defined. By combining both structural and transcriptomic information, the authors hypothesized that they could better characterize various cellular states, such as cardiomyocyte maturity, and also develop a linear regression model to classify structural maturity. What the authors ended up finding is that for certain genes, such as MHY7 which is a well known marker of cardiomyocyte maturity, their structural model had a near perfect correlation with transcriptome abundance. However, for other genes, structural organization did not seem to correlate with overall transcriptome abundance. This paper serves as a gateway to the main focus of this review, since the authors demonstrate a machine learning approach to characterizing structural maturity and use said information to connect to cell transcriptomics. They demonstrate how combining both morphological and transcriptomic information serves as a new method to characterize cellular state. By using a linear regression model to characterize cellular maturity based on transcriptome and phenotype, they show how machine learning serves as a potential method for assisting in this new field of study.

Combining cell transcriptomes with their morphological features not only provides interesting results about new cell states, but can lead to new methods for defining well established ones. In a paper by Li et al, they trained a machine learning model using flow cytometry measurements to classify apoptotic and healthy cells [2]. Flow cytometry is a fairly common method in cellular biology, where one can sort cells based on fluorescence and morphology without causing major harm to the sample. However, flow cytometry does provide any information on healthy v diseased states of cells. If one needed this information for classifying cells into apoptotic and healthy classes, they would have to fix and stain cells which is both timely, costly and incompatible with flow cytometry. The authors of this paper seek to bridge this gap by training a neural network on both flow cytometry data and straining information (apoptotic marker). They hypothesize that measures of cell morphology from flow cytometry contain information about the apoptotic cell state. Through their study, they show that k-means clustering in Euclidean space is sufficient with their data set to separate cells into both a healthy and apoptotic class with an accuracy comparable to traditional staining methods. This shows how machine learning can be used as a method to combine both straining data (transcriptomics) and morphological information (flow cytometry). While the previous paper sought to find and characterize more niche states of cellular maturity, this paper proposes a new alternative method to traditional staining that will greatly assist in any cell biology experiment. Machine learning combined with both transcriptomics and morphology not only provides a basis for exploring more specialized cell states, but also as a way to take traditional biology methods and computerize them for better accessibility, time spent and cost.

Machine learning as a method for looking at cell state shines through in Matla et al. 's paper, which looks purely at stem cell transcriptomics but produces what I believe to be a high impact result for the field of cancer biology [3]. In this paper, Malta et al focuses on oncogenetic dedifferentiation in cells due to pathway overactivation. Cancerous dedifferentiation touches on what I believe to be one of the most interesting properties of cancer and stem cells. Both seem to involve Yamanaka transcription factors, but while certain activation of these factors leads to stemness, over activation leads to cells becoming cancerous. In this study, OCLR machine learning was used to extract transcriptomic and epigenetic features from pluripotent stem cells and their differentiated counterparts. Through this machine learning study, different mechanisms for the dedifferentiated  oncogenetic state were discovered. As a long standing question in the field of stem cell biology, this new discovery allows for further research into these new states and mechanisms. The discovery of new cancer hallmarks is significant to the field of cancer biology as now it expands the range of potential drug targets when developing therapies. I include this paper because of its use of machine learning and transcriptomics, and to propose my own follow up to this paper to keep in theme with the review thus far. I would propose an additional study which would combine the transcriptomic data with cell morphological data and see if any link can be made between phenotype and transcriptome. Understanding the link between phenotype and cell state would greatly assist with cancer identification and with the extensive genetic profiling done in this paper, I believe this would provide a new method for cancer diagnosis. While we have now seen an example of a cell represented purely by transcriptome with fantastic results, as cell biologists we should continue to bridge the age-old knowledge gap between observed and transcriptomic state.

Continuing to look at implementations of machine learning for cancer identification, Caidedo uses this as a tool for classifying cancerous mutations in single cells [4]. The goal of the experiment is to classify groups of variants to the same gene using machine learning, regardless of expressed mutation. This is done to gain a deeper understanding of how different mutations are linked to like genes, and why mutations of that gene lead to differing phenotypes. Next, single cells are classified into a mutation category before being aggregated on a population level. This is done to get a sense of how the population as a whole behaves and to compare single cells to their neighbors. Classification of cells into both 10 and subsequent 26 mutant categories via RNN and tSNE hits around 78% accuracy. Sadly, the paper does not go into detail about interpreting the results, which I will state is a major flaw. They do not discuss what these mutations are or what the links between the clusters are. However, the paper itself proposes and shows results of a new method to link genome to cell morphology which follows in line with the topic of this review. I would state that this paper proposes a method implemented in the previously discussed papers, but does not show any meaningful results such as in Gerbin and Li. Regardless, this paper still shows that looking at both cell morphology and transcriptomics and using machine learning as a method to connect the two provides a good base model for studying cellular state.

These targets can potentially be studied using the machine learning outlined in a paper by Dürr [5]. Yeast are commonly used for early drug testing since they are easy to work with, cost effective and are a well established model organism. In testing, scientists add a drug to a yeast culture and then observe single cell morphology differences. These differences can easily be thought of as features for a neural network to train on. Dürr uses this idea in his study of machine learning as a tool to classify unknown phenotypic states for yeast undergoing early drug testing. Yeast cells had different proteins tagged with GFP and were imaged. Images were labeled according to protein being expressed and its localization within the yeast body. A CNN was then trained on said data to classify where in the yeast body the GFP tagged protein localized to. According to the results of this study, they successfully created a neural network capable of classifying yeast based on GFP localization. This tool proves incredibly useful since now researchers can send their network and not only get classification of phenotype, but see what other drugs induce a similar phenotype to check for toxicity and other negative effects. While this paper tackles a more simplistic task, it still demonstrates classification of cell phenotype and how it relates to cell state much like in Gerbin. This new methodology will greatly improve our ability to quickly classify cellular state during pharmacological testing, allowing for speed ups in the drug development timeline.

Looking at both cell phenotype and transcriptome is the future direction of cell biology. As the field of cell biology progresses, we need to study both phenotype and cell transcriptome hand and hand to better define cellular state, allowing us to better understand stemness, cancer and even improve upon pharmacology assays. This review serves as an argument for further research into this topic, as the age-old question of phenotype to cell state remains poorly studied in the field.

## References

[1]: Gerbin KA, Grancharova T, Donovan-Maiye RM, Hendershott MC, Anderson HG, Brown JM, Chen J, Dinh SQ, Gehring JL, Johnson GR, Lee H, Nath A, Nelson AM, Sluzewski MF, Viana MP, Yan C, Zaunbrecher RJ, Cordes Metzler KR, Gaudreault N, Knijnenburg TA, Rafelski SM, Theriot JA, Gunawardane RN. Cell states beyond transcriptomics: Integrating structural organization and gene expression in hiPSC-derived cardiomyocytes. Cell Syst. 2021 Jun 16;12(6):670-687.e10. doi: 10.1016/j.cels.2021.05.001. Epub 2021 May 26. PMID: 34043964.

[2]: Li, Y., Nowak, C.M., Pham, U. *et al.* Cell morphology-based machine learning models for human cell state classification. *npj Syst Biol Appl* 7, 23 (2021). https://doi.org/10.1038/s41540-021-00180-y

[3]: Malta T. M  et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation, Cell, Volume 173, Issue 2, 2018, Pages 338-354.e15, ISSN 0092-8674, https://doi.org/10.1016/j.cell.2018.03.034.

[4]: J. C. Caicedo, C. McQuin, A. Goodman, S. Singh and A. E. Carpenter, "Weakly Supervised Learning of Single-Cell Feature Embeddings," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9309-9318, doi: 10.1109/CVPR.2018.00970.

[5]: Oliver Dürr, Elvis Murina, Daniel Siegismund, Vasily Tolkachev, Stephan Steigele, and Beate Sick, ASSAY and Drug Development Technologies, Aug 2018, 343-349 http://doi.org/10.1089/adt.2018.859