Image-based early predictions of functional properties in cell manufacturing

Hong Seo Lim
Department of Biomedical Engineering
Georgia Institute of Technology
Atlanta, GA, USA
hlim95@gatech.edu

Madeline E. Smerchansky
Department of Biomedical Engineering
Georgia Institute of Technology
Atlanta, GA, USA
ORCID:0000-0001-5206-0193

Jingxuan Zhou
School of Biological Sciences
Georgia Institute of Technology
Atlanta, GA, USA
jxzhou@gatech.edu

Paramita Chatterjee
Marcus Center for Cell Characterization and Manufacturing
Georgia Institute of Technology
Atlanta, GA, USA
ORCID:0000-0001-9761-2289

Angela C. Jimenez
Marcus Center for Cell Characterization and Manufacturing
Georgia Institute of Technology
Atlanta, GA, USA
ORCID:0000-0002-2490-4122

Xingyu Yang
School of Biological Sciences
Georgia Institute of Technology
Atlanta, GA, USA
xysheep@outlook.com

Krishnendu Roy
Marcus Center for Cell Characterization and Manufacturing
Georgia Institute of Technology
Atlanta, GA, USA
krish.roy@gatech.edu

Peng Qiu
Department of Biomedical Engineering
Georgia Institute of Technology and Emory University
Atlanta, GA, USA
peng.qiu@bme.gatech.edu

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313297

***Abstract*—Effective cell manufacturing is essential to realizing the full potential of cell-based therapies but faces a multitude of challenges. One of the major challenges is the identification of critical quality attributes (CQAs), especially ones that enable early predictions of functional properties of the final products. The main goal of this study is to develop machine learning models for early predictions of the functional properties of mesenchymal stromal/stem cells(MSCs) in cell manufacturing. Deep learning models are trained and tested for image-based prediction of functional property—Collagen II expression after chondrogenic differentiation—of MSCs cells. During the MSC expansion, images of culturing wells were collected daily in the first six days, and the Collagen II level was assayed at the end of differentiation, following expansion. For each day, a deep learning model was trained with images from a specific experimental condition, and each model was tested with images from the same condition and also from other conditions. The trained neural network models showed 70-90 percent accuracy. Most of the models across different days and conditions show high consistency, especially models trained with images past day 2 of cell culture. Such consistency suggests that models are picking up similar features in predicting chondrogenesis capability. Our study highlighted the potential of deep neural network models used for early predictions of the functional properties of MSCs in cell manufacturing.**

***Index Terms*—Cell-based therapies, Mesenchymal stromal/stem cells, Early prediction, Cell Manufacturing, Machine Learning, Deep Learning**

This research was supported by funds from The Marcus Foundation, The Georgia Research Alliance, and the Georgia Tech Foundation through their support of the Marcus Center for Therapeutic Cell Characterization and Manufacturing (MC3M) at Georgia Tech.

I. Introduction

Cell-based therapies have the potential to make significant contributions toward regenerative medicine and immunotherapies in cancer, infectious, and autoimmune diseases [1]. The current state-of-the-art in cell therapies involves small-scale hospital-centric processing of cells, operated with minimal characterization and little involvement of process engineering and quality control, leading to wide variability and uncertainty in clinical trials [2], [3]. Effective cell manufacturing is essential to realizing the full potential of cell-based therapies but faces a multitude of challenges, such as cell characterization, process standardization, quality control, cryopreservation, transportation, and logistics [4].

One of the significant challenges is the identification of critical quality attributes (CQAs), especially ones that enable early predictions of functional properties of the final products. The primary reason for such a challenge is that the products

in cell manufacturing are living entities whose properties can be affected by many factors in the manufacturing process [4]. Predictive CQAs and models can significantly contribute to the optimization, standardization, and quality control in cell manufacturing.

Mesenchymal stromal/stem cells (MSCs) are multipotent stromal cells that are capable of differentiating into a variety of mesenchymal lineages such as bones, cartilages, muscles, and adipose tissues. For their self-renewal capacity, many studies have investigated the potential of using MSCs for cell-based therapies [5]. The process by which cartilage is developed through the mesenchymal cell is called chondrogenesis, and research on MSC-based therapy for cartilage regeneration is highly popular due to the proliferative and potent chondrogenic differentiation potential of MSCs [6].

The main goal of this study is to develop machine learning models for early predictions of the functional properties of MSCs in cell manufacturing. The predictions can be made based on various choices of cell characteristics. One popular choice is the cell surface markers. For example, mesenchymal stromal/stem cells can be identified using surface markers. However, existing data indicate that those surface markers are insufficient for predicting functional quality [7]. More sophisticated characteristics need to be considered. One promising choice is image-based characteristics of cell morphologies, which have been shown to correlate with drug mechanisms of action [8], immunosuppressive capacity [9], and stem cell differentiation [10]. Several known chondrogenesis-related parameters can be measured. Among the parameters, Collagen II gene expression level is known to highly correlate with how functional MSCs are [11]. Several studies suggest that high expression of Collagen II leads to better regeneration efficacy, hence acknowledge Collagen II as a major player in chondrogenesis capability [12], [13]. For these reasons, we assume Collagen II as a functional property of MSCs, where high expression of it indicates strong chondrogenesis capability, and low indicates weak capability in this study.

Developing computational models for image-based prediction requires powerful machine learning models because of the complex nature of the data and the problem needs to be solved. Recent advancements in the field of deep learning open up novel ways to solve tasks that were once considered too complicated. Specifically, the field of computer vision was a major beneficiary of such drastic development. Deep learning, more specifically Convolutional Neural Networks (CNNs), show state-of-the-art accuracy in tasks related to images, such as image classification, segmentation. For this study, we aim to develop CNN models for image-based prediction of functional property—Collagen II—of MSCs.

II. Methods

A. Experiments and Conditions

MSCs purchased from two vendors (Lonza and Rooster Bio) were expanded in each growth medium, creating four experimental conditions. The four conditions corresponded to all possible combinations of the two cell sources and two culturing media. Twenty four 6-well plates (Corning) were seeded with all four conditions in a random order to reduce batch effects. This resulted in 36 replicate wells of each condition. Lonza cells, at passage 4, were seeded at $5000\frac{cells}{cm^2}$ while Rooster Bio cells, at passage 4, were seeded at $4000\frac{cells}{cm^2}$. Lonza media was exchanged every 2-3 days and Rooster Bio media was changed once on day 4.

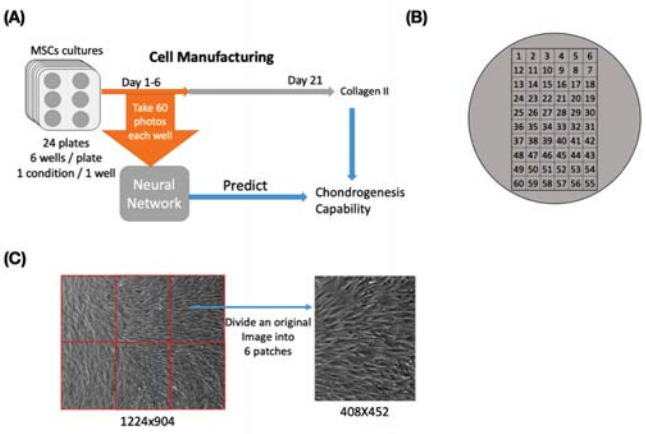

Fig. 1. Experimental Design Schematic. (A) MSCs cultures were put into 24 plates where each plate holds 6 wells. For the first 6 days, 60 images were taken per well. After 21 days, Collagen II level for each well is measured, and each well's chondrogenesis capability is evaluated. Using the images taken at early days along with the chondrogenesis capability measures, nerual network is trained to predict functional property. (B) Location of 60 images taken in a well. (C) Each original image is divided into 6 patch images.

60 phase contrast images per well (Fig. 1) were taken once per day at 4x magnification using the Cytation 5 Cell Imaging Multi-Mode Reader (BioTek), coupled with the BioSpa 8 Automated Incubator (BioTek), starting with day 1, until confluence at day 6. The images were collected for only the first six days to meet the goal of early prediction, and the day-specific models could suggest to how early our models could predict correctly. Fig. 1 summarizes the details of the experiment. There was a total of 2,160 images per condition for each day. Once the cells had grown to be confluent, they were washed with PBS, trypsinized, and formed into pellets [14]. Each cell pellet was 250,000 cells. Cells were plated onto 96 well nonadherent round bottom plates in chondrogenic media containing 1 percent penicillin/streptomyocin, 10nM Dexamethosone, 40 µg/mL l-proline, 5 percent ITS+, 50 µg/mL ascorbic acid –2- phosphate, and 10ng/mL TGF-β3 in alpha minimum essential medium solution and centrifuged at 300xg for 5 minutes. They were allowed to form for 24 hours prior to the first media exchange. Media was exchanged every 2-3 days for 21 days. After 21 days of differentiation, the gene expression of Collagen II was measured using quantitative reverse transcription polymerase chain reaction (qRT-PCR). RNA was isolated from individual pellets using the QuickRNA MicroKit (Zymo). RNA purity was confirmed and cDNA was synthesized from the RNA using between 500-1000 ng of RNA per sample. This was done using the SuperScript

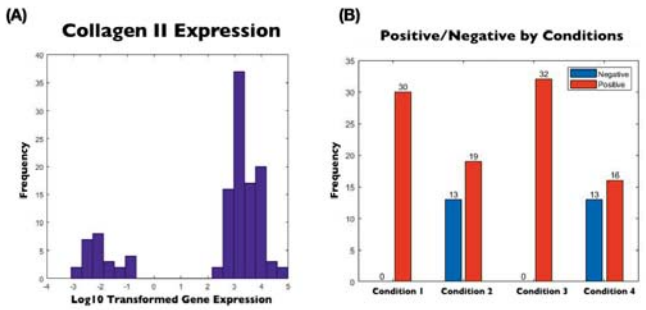

Fig. 2. Histograms of the Collagen II Levels of the Samples(Left) and Frequencies of Condition-specific Positive and Negative Labels(Right). The left histogram shows clear bimodal distribution. Only condition 2 and 4 shows the bimodality.

III First-Strand Synthesis System for qRT-PCR using the oligo(dT)20 primer. PCR was performed using primers for collagen II and beta actin, as a housekeeping gene, expression using the SYBR Green Master Mix (Qiagen) on a StepOnePlus machine.

B. Image Processing

For each day, images from 126 wells were collected, and each of these wells corresponds to one condition. As mentioned above, 60 images were taken per well. Each image was further cropped into smaller image patches. For each original image with a resolution of 1224 by 904, we divided the image into six patches, each with 408 by 452 (Fig. 1). Thus for each well, 360 (60 * 6) images could be used to train/validate/test our models. Excluding images from wells that we do not have relevant Collagen II expression information, there are a total of approximately 10,000 images per day per condition.

C. Binarization of Collagen II Activity

For each well, its Collagen II level was assayed at the end of the expansion process. When the Collagen II expression values were applied with the log-transformation, a histogram of the measured Collagen II level shows clear bi-modality, which indicates that the samples fall into two distinct phenotypes (Fig. 2A). Since the high Collagen II level is associated with strong chondrogenesis capability, we infer that the group with a high Collagen II level are the cultures that have strong chondrogenesis capability. Consequently, samples with low Collagen levels are cultures with weak chondrogenesis capability. The threshold used for the binarization was 1, and according to the threshold, the samples were grouped into two different labels: Collagen II positive samples and Collagen II negative samples. Once we visualize the distribution of positive and negative, we found that positives and negatives are almost equally distributed in Condition 2, and 4, whereas Condition 1 and 3 show all most all positives. The condition 1 and 3 correspond to a culturing media from the same company, and the fact that the vast majority of the cultured cells are positive suggests that our proposed predictive model would not be much needed. However, in the other conditions, i.e., conditions 2 and 4, which corresponds to another culturing media, with a similar distribution of Collagen II positive and negative, our predictive model could have value. Hence, we decided to use images from conditions 2 and 4 only. Such choice not only suits the biological motivation but also the balanced samples with positive and negative facilitate the training of the models with the overarching goal of our study: differentiating positive samples from negative samples.

D. The Architecture of Convolutional Neural Network Models

Our deep learning model's architecture is based on the VGGNet architecture [15], which is considered as one of the-state-of-art architecture in image classification problems. The network consisted of a combination of 7*7 and 3*3 convolutional layers with ReLU activation, and 2*2 max-pooling layers, followed by fully connected layers of 128 nodes with ReLU activation, and a softmax layer for classification. The implementation of the above architecture is done through TensorFlow and Keras, and NVIDIA TITAN XPs were used to train the deep neural network models. Please find the supplement information for the details of the architecture used.

E. Training and Validation of Models

Based on the image patches and the measured Collagen II endpoint, we built deep learning neural networks to identify and correlate image features in the patches with the corresponding Collagen II measurement. Since we built one deep learning model for each condition and each day separately, we were able to evaluate the predictive power of data on different days, and hence address the question of how early we can make robust and credible predictions. We trained a total of 12 models in this study. For each day, a model was trained using images from condition 2, and another model was trained using images from condition 4. For each trained model, it was tested using unseen images from the same condition and also using images from the other condition (i.e., condition 4 images if the model were trained on condition 2, and vice versa). 60 percent of images were used for training, 20 percent for validation, and the rest 20 percent for the testing. For this study, images were randomly split for training/validation/testing.

F. Average-based Prediction

Our models predict between Col II positive and negative at the patch level, because input to the model are patch-level images. Predictions at the patch-level images of the same sample should be integrated for accurate predictions at the sample level. Hence, we devised an average-based prediction. For each well, we have a maximum of 360 patch-level images. Each image would receive a score from the model. The score is a quantification of how the trained model thinks the Col II level of the image. For the well-level prediction, we decided to take an average of all images belong to a specific well, and use the mean as a statistic for classification of the well.

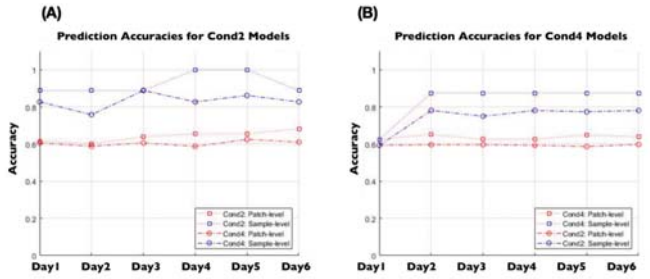

Fig. 3. Comparison of Patch-level Predictions and Sample-level Predictions. Sample-level prediction accuracy is noticeably higher than patch-level prediction

G. Consistency of Models

Because we trained and tested 12 different models, consistency among all these models is imperative. Such consistency may indicate that models are learning similar features regardless of different images under different conditions. To check consistency, we randomly sample 600 patch-level images across all days. Those chosen images were fed to the models so that each model (total of 12) outputs 600 scores for the images. Pairwise correlations for the scores are calculated for every possible pair of models to evaluate the consistency of the models.

III. Results

A. Deep Learning Models can predict Collagen II level

Collagen II prediction accuracies across the models are summarized in Table 1. The above panel contains the accuracy of models trained with condition 2, and the bottom, condition 4. Well-level prediction accuracies are consistently above 80percent for most of the models when tested with the same-condition images. For a couple of models trained with condition 2— day 5 and day 4 models— the accuracies were 100percent. Even when the models were tested with images from different conditions, although the overall accuracy decreases, the accuracies still remain above 70percent for most of the models. Since images from the different conditions were never used for the training and validation, there is a higher number of wells to predict (27 - 32) for this setting.

B. Average-based prediction improves overall accuracy

One interesting pattern found is that well-level prediction is remarkably better than the patch-level prediction. While most of the well-level accuracies hover above 70percent, the patch-level prediction accuracies are around 60percent. From the difference, we may assume that the average-based voting significantly improves accuracy at well-level. (Fig. 3)

C. Models Consistent Starting Day 3

As mentioned in the above method section, pairwise correlations are calculated for all models. A high correlation would mean that output scores for given images are similar for the two selected models. It turns out that models trained at earlier

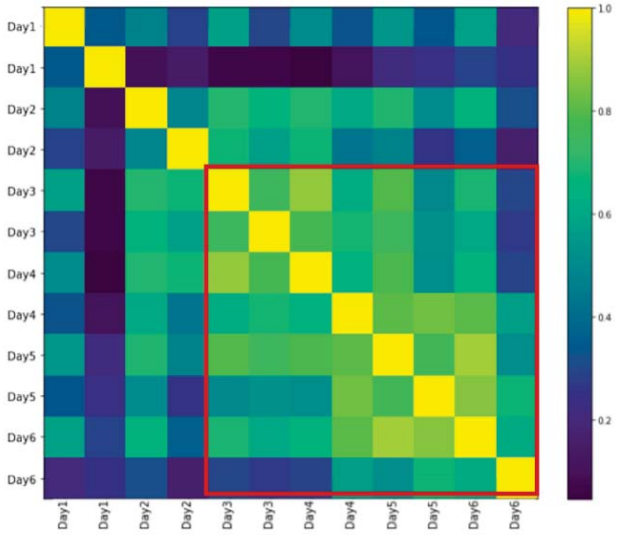

Fig. 4. Correlations Among Outputs of All Models. Pairwise correlation was evaluated for all models. Starting day 3, predictions of models become more correlated suggesting the consistency of the models.

days have relatively low correlation suggesting the models are not that consistent. However, starting day 3, we can clearly see those correlations among the models are remarkably high, as seen in the Fig. 4. Such finding suggests that models are learning similar features from images regardless of conditions or possible batch effects. Also, clearly elevated correlation starting day 3 suggests that three days of cell expansion should be used as minimum days needed so that our computational models could faithfully predict. We could also hypothesize some critical image features, generated by functional features, might arise past day 2.

D. Feature Extraction through Visualization

One way to understand how and what the models learn is by checking images that scored high and those with low scores from the models. For models shown high correlation with other models, we checked images that receive exceptionally high scores and images with exceptionally low scores, shown in Fig. 5. Not only the images that receive similar extreme scores are consistent but also there exists a feature that was readily noticeable via visual inspection. It seems that the crowdedness of cells is negatively correlated with the scores given. Most of the images that received high scores tend to have less crowded cells in the images, while the low score ones are noticeably more crowded than the high score ones.

E. Generalizability of the Models

We prepared another set of data to assess generalizability of our trained models. All of the newly attained samples, however, were measured to have all strong Collagen II expression levels, the scenario where our models are not trained to classify. Yet, we hypothesize that our models could 1) correctly

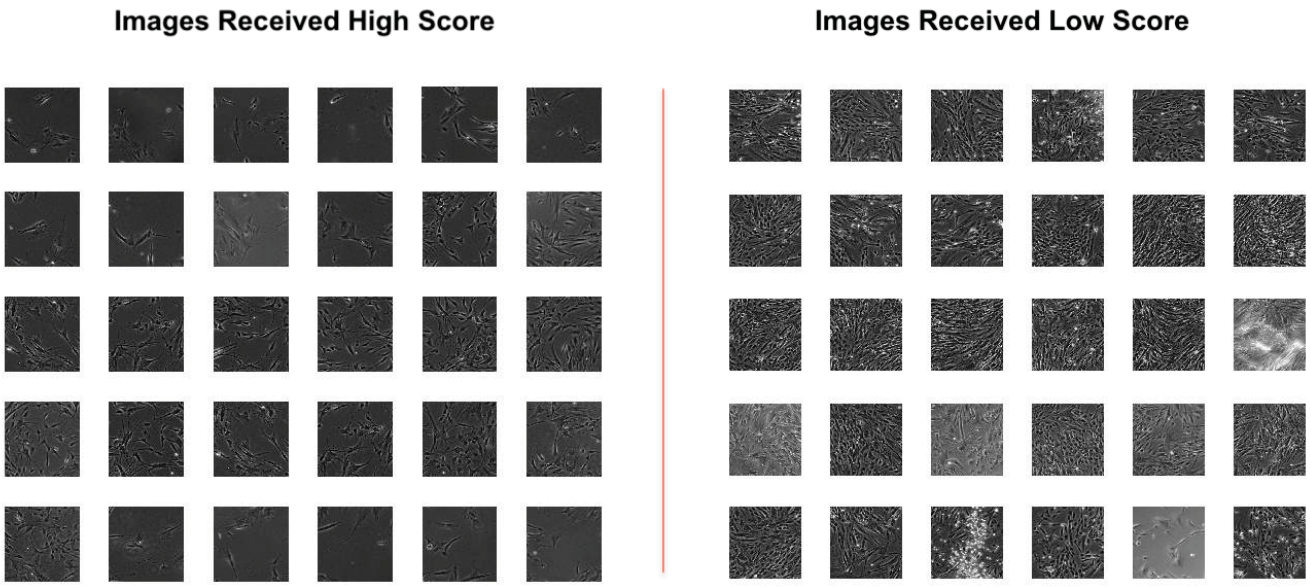

Fig. 5. Images Received Exceptionally High / Low Scores from the Models. Crowdedness of cells in an image seems to be negatively correlated with scores.

predict them as positive and 2) differentiate between the extremely high Collagen II samples from ordinary high samples. The newly acquired dataset contains four different conditions, each with a unique donor. Images from the new dataset were fed to models trained with condition 2 from the original data. The results are summarized in the Fig. 6. All of the 288 wells were measured to have strong Collagen II expression level, and our models predicted 269 out of 288 (94.4 per cent) to be positive, when 0.5 was used as a threshold. For the samples from donor A, positive regression lines for all days suggest that our models were able to differentiate a wide spectrum of the collagen II expression levels. The samples from donor B shows positive trends starting a model from the fourth day. The samples from donor C show positive trends starting day 2 to day 4, yet the trend is lost starting day 5. The samples from donor D show positive trends for all days except for the second day. Such findings above demonstrate that our models possess generalizability to some extent, yet they are not robust enough to be universal classifiers shown by the samples from donor C.

IV. Discussion

Our study highlighted the potential of deep neural network models used for early predictions of the functional properties of MSCs in cell manufacturing. The trained neural network models showed above 70-80percent well-level accuracy boosted by average-based prediction. Dividing original images into patch level small images not only boosts prediction accuracy but also is more suitable because otherwise, the deep neural network would require too much memory for the input images. The trained models show a high level of consistency, especially past day 2. Such consistency is important because it might mean, although separately trained, models are picking up similar features in predicting chondrogenesis capability. The fact that correlations are more clearly elevated

	Cond2 Images		**Cond4 Images**	
Cond2-trained model	**Patch-level**	**Sample-level**	**Patch-level**	**Sample-level**
Day 1	1994/3240 (62%)	8/9 (89%)	6323/10440 (61%)	24/29 (83%)
Day 2	1952/3240 (60%)	8/9 (89%)	6144/10440 (59%)	22/29 (76%)
Day 3	2071/3240 (64%)	8/9 (89%)	5895/9720 (61%)	24/27 (89%)
Day 4	2124/3240 (66%)	9/9 (100%)	6138/10440 (59%)	24/29 (83%)
Day 5	2125/3240 (66%)	9/9 (100%)	6529/10440 (63%)	25/29 (86%)
Day 6	2210/3240 (68%)	8/9 (89%)	6379/10440 (61%)	24/29 (83%)
	Cond4 Images		**Cond2 Images**	
Cond4-trained model	**Patch-level**	**Sample-level**	**Patch-level**	**Sample-level**
Day 1	1800/2880 (62%)	5/8 (62%)	6840/11520 (59%)	19/32 (59%)
Day 2	1877/2880 (65%)	7/8 (88%)	6873/11520 (60%)	25/32 (78%)
Day 3	1805/2880 (63%)	7/8 (88%)	6871/11520 (60%)	24/32 (75%)
Day 4	1808/2880 (63%)	7/8 (88%)	6841/11520 (59%)	25/32 (78%)
Day 5	1868/2880 (65%)	7/8 (88%)	6551/11160 (59%)	24/31 (77%)
Day 6	1841/2880 (64%)	7/8 (88%)	6890/11520 (60%)	25/32 (78%)

TABLE I
Prediction Accuracies of Models across All 6 Days

starting day 3 could be a piece of useful information for the cell culturing perspective because it is only then the image classifier could robustly predict. The consistency among the models was confirmed when we manually check images that models predicted to be chondrogenesis positive vs. images that models predicted to be low chondrogenesis negative. It should be noted, however, that the features that seem relevant and apparent based on our visual inspection are inadvertently biased and subjective. One feature seemed highly associated with the chondrogenesis capability of the samples: crowdedness of cells. Some possible hypothesis on why crowdedness might matter are 1) that extensive cell proliferation at an early stage of cell culturing indicates those cells are of a cell type that is not related to chondrogenesis or 2) that such early proliferation leads to early depletion of limited resources. Future studies could focus on answering how and why such crowdedness of cells might be related to chondrogenesis. With all the aforementioned findings, some limitations should also be noted. First, our study does not provide a universal machine learning classifier that can be used for any MSCs images. Our models were trained and tested on images taken under a very consistent procedure. Images generated from noticeably different conditions, such as different donors, other functional property measures, different data acquisition procedures, and manufacturing protocols, could not produce results as robust we presented. However, our study still highlighted the potential of the machine learning approach being used for predicting functional properties in cell manufacturing. Future studies should aim to improve the generalizability of this approach, possibly by trying to remove batch effects and training models with more data collected through various procedures. Once the generalizability is improved, a universal machine learning classifier can be achieved to improve the efficiency and quality control of cell manufacturing.

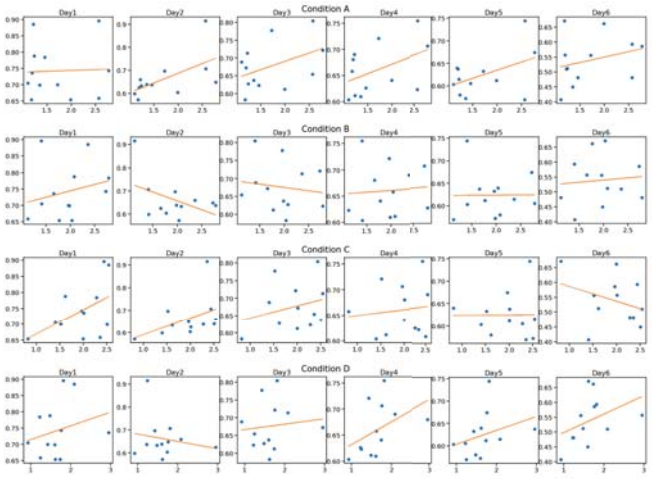

Fig. 6. Prediction of New Samples to the Trained Models. Although all samples were positive, for many models, samples with higher Collagen II level received higher score as seen by positive correlation of the scatter plots.

ACKNOWLEDGMENT

This research was supported by funds from The Marcus Foundation, The Georgia Research Alliance, and the Georgia Tech Foundation through their support of the Marcus Center for Therapeutic Cell Characterization and Manufacturing (MC3M) at Georgia Tech.

REFERENCES

[1] T. R. Heathman, A. W. Nienow, M. J. McCall, K. Coopman, B. Kara, and C. J. Hewitt, "The translation of cell-based therapies: clinical landscape and manufacturing challenges," Regen Med, vol. 10, no. 1, pp. 49-64, 2015, doi: 10.2217/rme.14.73.

[2] A. Terzic, M. A. Pfenning, G. J. Gores, and C. M. Harper, Jr., "Regenerative Medicine Build-Out," Stem Cells Transl Med, vol. 4, no. 12, pp. 1373-9, Dec 2015, doi: 10.5966/sctm.2015-0275.

[3] A. Kaiser et al., "Towards a commercial process for the manufacture of genetically modified T cells for therapy," Cancer gene therapy, vol. 22, no. 2, p. 72, 2015.

[4] R. Haddock et al., "Manufacturing cell therapies: The paradigm shift in health care of this century," NAM Perspectives, 2017.

[5] V. Mundra, I. C. Gerling, and R. I. Mahato, "Mesenchymal stem cell-based therapy," Mol Pharm, vol. 10, no. 1, pp. 77-89, Jan 7 2013, doi: 10.1021/mp3005148.

[6] C. R. Harrell, B. S. Markovic, C. Fellabaum, A. Arsenijevic, and V. Volarevic, "Mesenchymal stem cell-based therapy of osteoarthritis: Current knowledge and future perspectives", Biomedicine and Pharmacotherapy, vol. 109. Biomedicine and Pharmacotherapy, pp. 2318–2326, 2019.

[7] M. Mendicino, A. M. Bailey, K. Wonnacott, R. K. Puri, and S. R. Bauer, "MSC-based product characterization for clinical trials: an FDA perspective," Cell Stem Cell, vol. 14, no. 2, pp. 141-5, Feb 6 2014, doi: 10.1016/j.stem.2014.01.013.

[8] J. C. Caicedo et al., "Data-analysis strategies for image-based cell profiling," Nat Methods, vol. 14, no. 9, pp. 849-863, Aug 31 2017, doi: 10.1038/nmeth.4397.

[9] M. W. Klinker, R. A. Marklein, J. L. Lo Surdo, C. H. Wei, and S. R. Bauer, "Morphological features of IFN-gamma-stimulated mesenchymal stromal cells predict overall immunosuppressive capacity," Proc Natl Acad Sci U S A, vol. 114, no. 13, pp. E2598-E2607, Mar 28 2017, doi: 10.1073/pnas.1617933114.

[10] M. Fujitani et al., "Morphology-based non-invasive quantitative prediction of the differentiation status of neural stem cells," J Biosci Bioeng, vol. 124, no. 3, pp. 351-358, Sep 2017, doi: 10.1016/j.jbiosc.2017.04.006.

[11] S. M. Richardson et al., "Mesenchymal stem cells in regenerative medicine: focus on articular cartilage and intervertebral disc regeneration," Methods, vol. 99, pp. 69-80, 2016.

[12] Y. Jiang et al., "Human cartilage-derived progenitor cells from committed chondrocytes for efficient cartilage repair and regeneration," Stem Cells Translational Medicine, vol. 5, no. 6, pp. 733-744, 2016.

[13] D. Bosnakovski, M. Mizuno, G. Kim, S. Takagi, M. Okumura, and T. Fujinaga, "Chondrogenic differentiation of bovine bone marrow mesenchymal stem cells (MSCs) in different hydrogels: influence of collagen type II extracellular matrix on MSC chondrogenesis," Biotechnology and bioengineering, vol. 93, no. 6, pp. 1152-1163, 2006.

[14] Schon, B. S., Schrobback, K., van der Ven, M., Stroebel, S., Hooper, G. J., and Woodfield, T. B. (2012). Validation of a high-throughput microtissue fabrication process for 3D assembly of tissue engineered cartilage constructs. Cell Tissue Res 347, 629-642.

[15] Simonyan, Karen, Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.