Submitted by: Amit Damri

# Data Science in Cell Imaging – Final Project Report

## Introduction

Drug discovery is a slow and expensive process. It is prohibitively difficult to evaluate all potential chemical structures in physical experiments in order to find a promising cure for specific illness situations [1,4]. However, one can reduce the effort by predicting the biological activity or Mechanism of action (MOA) of a query compound based on that of a reference compound with a suitably similar profile [2]. Basically, MoA is the description of how a compound interacts with a target and affects a biological system [2]. In addition, when we talk about compound profiles it means that each sample is put into a single, shared high-dimensional representation space using a profile, and the similarity of a pair of compounds indicates how closely related they are [4].

My motivation in this project is based on this short introduction. Once we have a new or an unknown compound and we want to test it, we can find its profile and compare it with other well-known compounds. By grouping the compounds and according to the group members, we can infer the MoA of the compound that is tested.

The goal of this project is to assess the Mechanism of Action (MoA) of different compounds and grouping compounds according to their effect on the cells.

## Method

I suggest the following method which has two parts: Reconstruction of cell profiling images using UNET Autoencoder [6] and compound classification using Deep Neural Network (DNN) [5]. Figure 1 shows an illustration of the method:

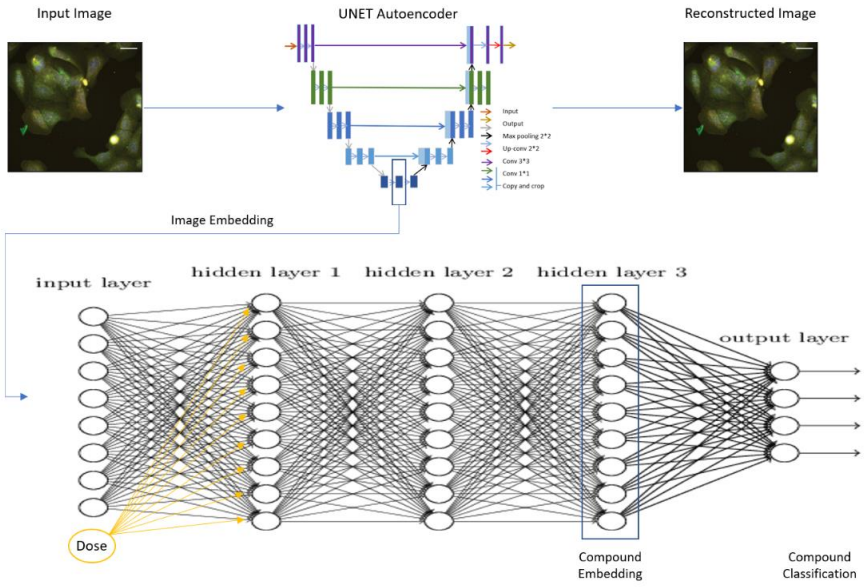


**Figure 1:** Method architecture

The UNET Autoencoder task is to reconstruct the input cell painting image. The final layer of the Encoder architecture consists of an embedding vector of size Mx1 that should give us a compact representation of the input image. I relate to this embedding vector as a "phenotypic signature" of the cell. This embedding extracts M features from the image and learns, based on these features, to decode the embedding vector into its original representation.

I trained the autoencoder only on the control well images to learn the base representation of the cells i.e., to extract the signature of the cells without any compound. After that, I fed into the encoder images of wells with compounds to find their embedded representation. Because the autoencoder were trained only on control images, I thought that the embedding vectors of the other wells will be distant from the control embedding and this distance will contain the influence of the compound. But the distance is not enough, because each compound can affect other features. For this reason, I used a DNN that its goal is to learn an embedding representation of each compound.

The input of the DNN is the embedding vectors (the output of the encoder) of the wells with the compounds. The task was to classify for each vector its correct compound that were used in that well. However, it is reasonable to think that different doses of a compound would act different, therefore I added to the embedding vector another feature that equals to the dose of the compound (thus, I added another neuron to the input layer). Finally, after the training of the network, I used the last hidden layer as a compound embedding vector and analyzed the relations between the different compounds.

Results

The dataset I used is "Multiplex cytological profiling assay" represented by Gustafsdottir et al. The images are of U2OS cells treated with each of 1600 known bioactive compounds and labeled with six labels that characterize seven organelles (the "Cell Painting" assay) [3]. Because of time and memory constraints, I used only one channel, the Hoechst 33342 (nuclei) channel to see the influence of each compound on the nuclei of the cell. Moreover, I chose only one plate to test (plate number 20585 – with ID of 4403). The plate contains 384 wells, each well has 9 fields, 64 of the wells used as controls, wells with no compound (compound name equals to 'DMSO') and the other 320 wells were with compounds.

In the preprocess phase, I split the data into control and compound images. For the compound images I extracted the float value of the dose feature. I resized all images to 256x256 and split each image into 4 patches of 128x128. In addition, I normalized the values of the pixels to be between 0 to 1. Finally, I split the control data into 80% training set and 20% testing set.

For the UNET Autoencoder I used the Mean Squared Error as the loss function and Adam optimizer. The embedding vector is of size 32x1. The convergence graph of the autoencoder and a reconstructed example are presented in figure 2.

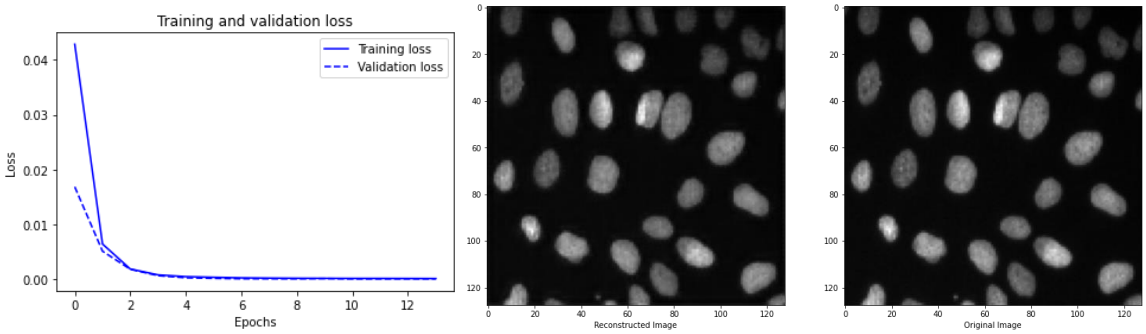


**Figure 2:** From left to right. The UNET Autoencoder convergence graph, An example of a reconstructed image and its original image

The graph shows that the autoencoder succeeded to learn the reconstruction of the images, as well as the reconstructed image that shows only slightly differences. Next, I predicted the embedding vectors for all the compound images using the encoder part of the trained autoencoder. For each vector I added another feature – the dose value.

I split the vectors into train and test sets on the same way as previously (80%/20%) and used these vectors to train the DNN. Because this is a classification problem, I set the target values as the compound names and encoded these values into integers. Total of 320 compounds were used (hard problem with many classes). For the training I used Sparse Categorical Cross Entropy loss and Adam optimizer. I calculated the classification accuracy and in order to compare to performance I used a baseline accuracy of the most frequent compound - I found out that the baseline accuracy is 0.7%.

The convergence graph of the DNN training presented in figure 3. It can be seen that the loss is very high, thus I got that the test accuracy is only 14%. However, compared to the baseline accuracy is much higher.

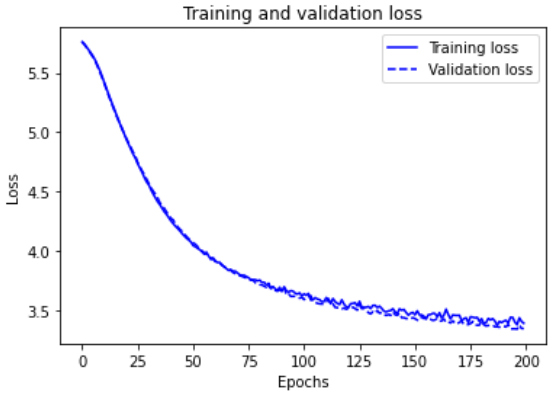


**Figure 3:** DNN training convergence graph

To analyze the results, I used the last hidden layer vector of size 32x1 and used a PCA to embed each vector into a 2D space. I chose 2 frequent compounds (in this data) Anoctinine and Vitexin from the test set and compared their visualization. Figure number 4 shows the results – on the left, each color represents a compound and on the right each symbol represents a dose value. The model succeeded to discriminate between the two compounds (Figure 4 - Left) but it also divided between the same compound but with different dose value (Figure 4 – Right). It makes sense because different amounts of the same compound can affect different sometimes.


**Figure 4:** PCA of 2 compounds – each has a different color (left) and a different dose value (right)

Finally, I reduced the dimensions of all compound vectors using PCA to 2D. The result presented in Figure 5. Each color represents a compound, and each symbol represents a different dose. The visualization shows that different compounds may be close to each other (different colors close to each other), hence I may assume that their MoA is very similar. Moreover, it shows that the same compound can have different affect with different doses.


**Figure 5:** PCA of all compound vectors

## Discussion

As mentioned before, I used only one plate, I did not validate the results with other plates, and used only one channel, hence the results are very limited. In addition, I missed many compounds from different plates, and I could not validate the compound clusters because I miss the domain knowledge. Moreover, according to [4] they suggest that for proper evaluation, one complete compound set, including all concentrations, should be left out of training. Therefore, for future work I will suggest remembering these tips.

However, from the last figure it seems that the suggested architecture succeeded to find a profile vector for each compound, because we can see many dots with the same color close to each other. We can see also that different compounds mapped to the same area and maybe it will be interesting to validate if these compounds have a similar mechanism of action.

To summary, I saw multiple Deep Learning approaches from the last two years, and it seems that this area of research is becoming more popular. Additionally, I learned from [4] about the different techniques that proved to be useful in each stage of the data analysis process of the image-based cell profiling workflow, and I tried to base my project on their ideas.

## References


1. Moshkov, Nikita, et al. "*Predicting compound activity from phenotypic profiles and chemical structures.*" bioRxiv (2022): 2020-12.
2. Chandrasekaran, Srinivas Niranj, et al. "*Image-based profiling for drug discovery: due for a machine-learning upgrade?.*" Nature Reviews Drug Discovery 20.2 (2021): 145-159.
3. Gustafsdottir, Sigrun M., et al. *"Multiplex cytological profiling assay to measure diverse cellular states."* PloS one 8.12 (2013): e80999. http://idr.openmicroscopy.org/webclient/? show=screen 1952
4. Caicedo, Juan C., et al. *"Data-analysis strategies for image-based cell profiling."* Nature methods 14.9 (2017): 849-863.
5. Li, Xiangjie, et al. "Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis." Nature communications 11.1 (2020): 1-14.
6. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "*U-net: Convolutional networks for biomedical image segmentation.*" International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.
7. Google Colab notebook Link: Click here