Classifying cells after chemical perturbations using feature extraction and machine learning

Project 2 - Applied Data Science GR5243
Val Rogel

November 9, 2024

1 Abstract

Numerous studies have examined the effects of chemical and genetic perturbations on cellular function, yet there remains significant potential to enhance predictive accuracy through via machine and deep learning algorithms still remains. In this study, I used classification techniques to explore whether features extracted from fluorescence microscopy could predict the chemical perturbations of over 150,000 cells. By leveraging the Cellpose segmentation algorithm and both random forest and support vector machine (SVM) algorithms, a classification accuracy of just over 20% may be reached. Although the accuracy is modest, fine-tuning of the classification algorithms and addressing imbalances of the training data are promising means through which precision and recall may be improved.

2 Introduction

Understanding the impact of perturbations on cellular function is a critical avenue of biological research, especially in the field of pharmaceuticals. The increased knowledge of the impacts of chemical perturbations can lead to progress in predicting patient responses to drugs and tailoring compounds to target certain cells more accurately (Lotfollahi et al., 2019).

Cell imaging procedures, which are critical to data collection, have been enhanced by the development of Cell-Painting, an assay which uses fluorescent dyes to capture an image with specific organelles dyed one of six colors (Bray et al., 2016). An example of Cell Painting in practice lies in the 2024 Nature Methods publication of Chandrasekaran et al., in which eight components (nucleus, nucleoli and cytoplasmic RNA, endoplasmic reticulum, Golgi and plasma membrane, mitochondria, and the actin cytoskeleton) of over 75 million cells were stained across five channels. This procedure was used in order to perform a screening of over 400 perturbations on the U2OS and A549 cell lines.

Segmentation such as that offered by Cellpose is especially important in automatically distinguishing between cells in images where they may overlap or be clumped together. On a more general level, segmentation methods aim to separate a given image into small regions which contain individual nuclei, and consequentially, individual cells (Li et al., 2008). Cellpose specifically converts the masks of cells in the training set to vector flows which can be understood by a neural network (Stringer et al., (2021).

Supervised machine learning algorithms may be used to predict labels for data after a training and testing period, wherein the performance is determined through the testing period. Common examples of supervised algorithms include random forest and support vector machine (SVM) (Hallou et al., 2021). Scikit-learn, a Python package, offers machine learning algorithms that can be implemented to complete scaling, splitting into training and testing, and model training and evaluation.

3 Methods

I imported processed data from the Chandrasekaran et al. CPJUMP1 resource dataset. This cell data was originally collected by the researchers with chemical and genetic perturbations available in a library from the Broad Institutes Drug Repurposing Hub. These perturbations consisted of a total of 160 genes and 303 chemical compounds.

The images of the cells were created using fluorescence microscopy. The dataset produced by the researchers originally contained 3 million images with over 75 million cells. I used a smaller sample of 2867 images. This decision was informed by the limitations of RAM. Similarly, some methods were also informed by speed and GPU limitations. These cell images only contained a subset of the original perturbations (250) in order to limit the scope of the work.

3.1 Segmentation of Images

With regards to segmentation parameters, only the membrane staining channel was used for segmentation and feature extraction. The diameters of the cells were estimated on a per-image basis. To control the maximum allowed error in the flows for each mask, "flow_threshold parameter" was set to .5 (neither very strict nor lenient). We can assume a relatively high quality of images from the original dataset, but want to allow cells to be accepted even if the boundaries in the images are somewhat uncertain.

The "cellprob_threshold" was set to the relatively low value of -5, which allow us to find as many cells as possi-

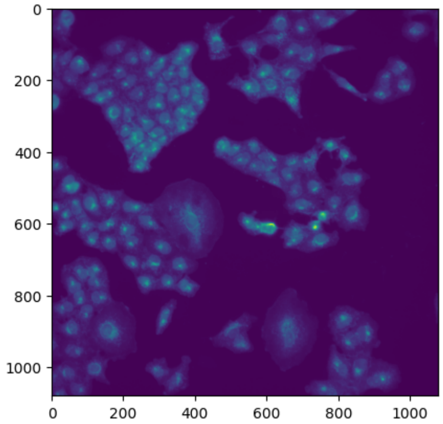

Figure 1: *Example of an image of cells from resource dataset (Chandrasekaran et al., 2024).*

ble within the image, especially if these cells are clumped together or appear to overlap.

The cells were segmented using Cellpose, with the default model "cyto" which is trained strictly on the training set from Cellpose. All 2867 images were successfully segmented.

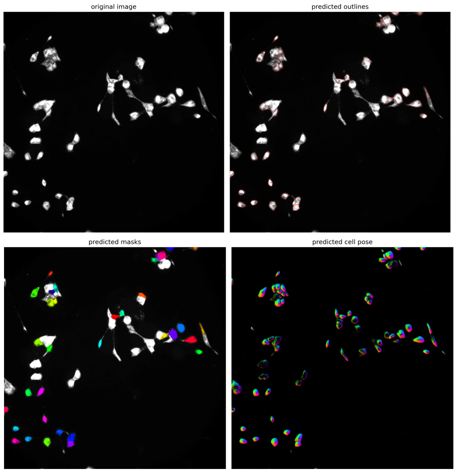

Figure 2: *Original image (top left), predicted outlines (top right), predicted masks (bottom left), and predicted cell pose (bottom right)*

3.2 Feature Extraction

Scikit-image was used to carry out feature extraction from the image. The following features were extracted from the masks: area, perimeter, eccentricity, solidity, and the intensity (mean, median, and standard deviation). Metadata from the original study was used to supplement the feature extraction data for a total of 150759 cells over the 2867 images in the dataset (and a corresponding mean number of cells per image of 52.58).

3.3 Classification

Random forest classification and support vector machine (SVM) were used with the goal to determine if one of these common classification algorithms perform better in classification of cells by chemical perturbation treatments. Both underwent an 80-20 train test split and were trained on all seven extracted features: area, perimeter, eccentricity, solidity, mean intensity, median intensity, and standard deviation of intensity. SVM was undertaken with the default kernel of the radial basis function (rbf) and was run both with and without the use of scaling.

Accuracy, precision, and recall were compared between the two.The precision score is defined as the ratio of true positives and all positives (true and false positives) while the recall is defined as the ratio between the true positives and the sum of true positives and false negatives (Kramer, 2016).

3.4 Code availability

Code link: https://github.com/valcsuremm/GU-5423

4 Results

4.1 Features

Of the 250 chemical compounds in the dataset, the five most common chemical perturbations among the images are the following compounds: DMSO (n=33848), BQ-788 (n=1191), zamifenacin (n=1180), sacubitril (n=1146), and M-25 (n=1127). The least common compounds as described by the metadata are GSK1070916 (n=108), NVP-HSP990 (n=86), AZ191 (n=74), PF-431396 (n=71), and UNC2025 (n=64).

The following histograms visualize the spread of eccentricity and mean intensity for the two most and least common compounds relative to the features for the full dataset of all images as informed by the metadata. See appendix for histograms of area and median intensity.

As demonstrated by the histograms, the distribution of the features for cells treated by particular compounds do not necessarily match the overall distribution of the features.

4.2 Random forest classification

The random forest classification, produced with an 80-20 train-test split yields a 21.96% accuracy. The classification report yielded precision and recall for each compound.

While the precision and recall for this classification were largely between 0.00 and 0.02, there were a few compounds with notably better classification. It should be noted that precision was set to 0 in instances where the compound was not predicted.

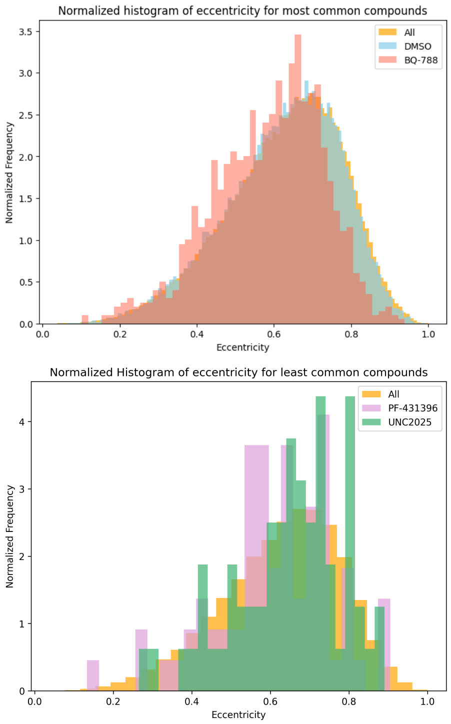

Figure 3: *Eccentricity, normalized by sample size, for DMSO and BQ-788 treated samples relative to full dataset (top) and for PF-431396 and UNC2025 (bottom)*

Compound	Precision	Recall	Support
ponatinib	.67	0.61	36
TG-101348	.57	.64	33
DMSO	.24	0.94	6803
cediranib	.30	.31	68
ingenol-mebutate	.25	.27	175
briciclib	.24	.28	58

The above compounds have a corresponding f1-score $\geq .25$.

4.3 Support vector machine classification

The SVM classification with the 'rbf' kernel, also produced with an 80-20 train-test split, yielded a 22.46% accuracy without scaling. Unlike the random forest classification, precision and recall were 0.0 for all compounds except for DMSO, which had a precision of .22, recall of 1.00, f1-score of .37, and a support of 6770. As above, it should be noted that precision was set to 0 in instances where the compound was not predicted.

With the addition of scaling through scikit-learn's StandardScaler, results appeared more promising. Accuracy over all compounds was 23.36%. Like in the random forest classification, precision and recall often fell between 0 and .02. Below are compounds which contributed to the nonzero accuracy with an f1-score above .25:

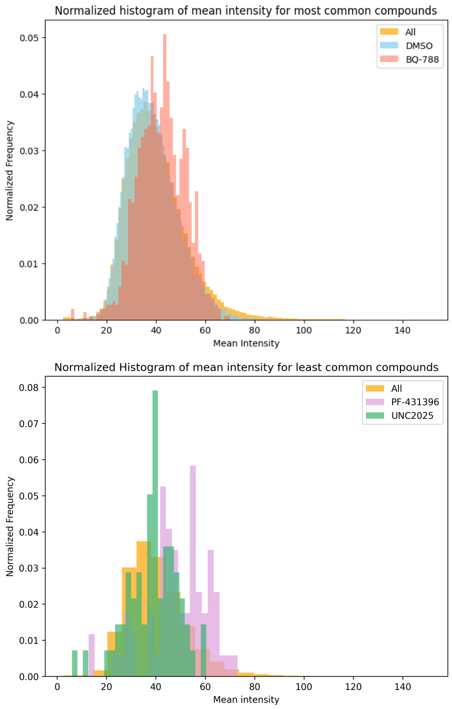

Figure 4: *Mean intensity, normalized by sample size, for DMSO and BQ-788 treated samples relative to full dataset (top) and for PF-431396 and UNC2025 (bottom)*

Compound	Precision	Recall	Support
TG-101348	.56	.58	33
ponatinib	.49	.67	46
LDN-212854	.4	0.33	30
DMSO	.23	1	6853
briciclib	.24	0.38	58
cediranib	.32	.26	72
ingenol-mebutate	0.22	0.31	142

5 Conclusion

Random forest classification and support vector machine are considered to have comparable performance (Schroff, 2008). Though the accuracy appear comparable, based upon the precision and recall, random forest outperformed the latter. Prior to scaling support vector machine with the default kernel parameter did not correctly classify any non-DMSO treated cells. It appears as though the SVM with the 'rbf' kernel only predicted DMSO as the perturbation. Scaling led to an improvement on prediction metrics of precision and recall among non-DMSO cells, and it is possible that the unbalanced nature of the dataset has contributed to the low accuracy.

Nonetheless, SVM with scaling yielded the highest accuracy by 1.3 percentage points. This fits with the existing notions that SVM tends to have the highest accuracy regardless of the scaling/normalization method (Shahriyari, 2019). However, since other classification algorithms were not utilized, it is possible that a different algorithm, such as gradient boosting machine (GBM) could outperform SVM in this instance.

The segmentation and classification had decidedly poor outcomes relative to machine learning classification tasks in nonbiological research such as transportation mode detection via smartphone, where accuracy may fall between 50 and 90%. Binary classification or "easier" tasks may experience much higher accuracy rates (Vakili et al. 2020).

Finally, it is important to remark upon the bias introduced through the set of compounds in this study, which was created with the compounds available from the Broad Institute's Drug Repurposing Hub (Chandrasekaran et al., 2024). The compounds are known to bind with greater specificity than randomly selected compounds might, creating an avenue for bias not only in my work, but in the work of Chandrasekaran et al. as remarked in the 2024 publication.

6 References

Bray, MA., Singh, S., Han, H. et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc 11, 1757–1774 (2016). https://doi.org/10.1038/nprot.2016.105

Chandrasekaran, S.N., Cimini, B.A., Goodale, A. et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. Nat Methods 21, 1114–1121 (2024). https://doi.org/10.1038/s41592-024-02241-6

Hallou, A., Yevick, H., Dumitrascu, B., & Uhlmann, V. (2021). Deep learning for bioimage analysis. arXiv preprint arXiv:2107.02584.

Kramer, O. (2016). Scikit-Learn. In: Machine Learning for Evolution Strategies. Studies in Big Data, vol 20. Springer, Cham. https://doi.org/10.1007/978-3-319-33383-0_5

Li G, Liu T, Nie J, Guo L, Chen J, Zhu J, Xia W, Mara A, Holley S, Wong ST. Segmentation of touching cell nuclei using gradient flow tracking. J Microsc. 2008 Jul;231(Pt 1):47-58. doi: 10.1111/j.1365-2818.2008.02016.x. PMID: 18638189.

Lotfollahi, M., Wolf, F.A. & Theis, F.J. scGen predicts single-cell perturbation responses. Nat Methods 16, 715–721 (2019). https://doi.org/10.1038/s41592-019-0494-8

Schroff, F., Criminisi, A., & Zisserman, A. (2008, September). Object Class Segmentation using Random Forests. In BMVC (pp. 1-10).

Shahriyari, L. (2019). Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. Briefings in Bioinformatics, 20(3), 985-994.

Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. Nature methods, 18(1), 100-106.

Vakili, M., Ghamsari, M., & Rezaei, M. (2020). Performance analysis and comparison of machine and deep learning algorithms for IoT data classification. arXiv preprint arXiv:2001.09636.

7 Appendix

Additional graphs are displayed to the right as mentioned in the Results section above.

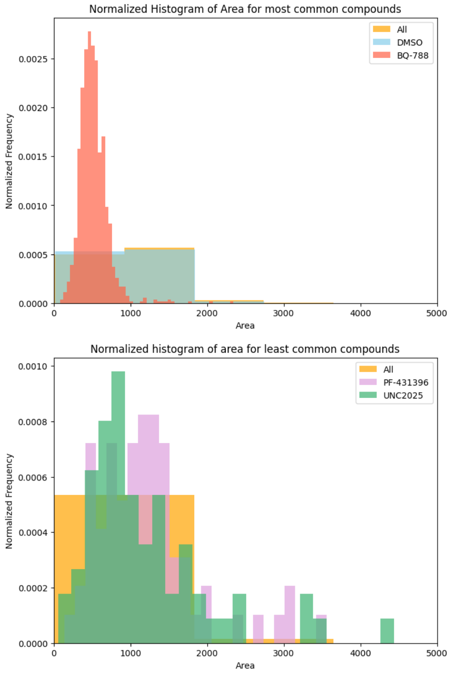

Figure 5: *Area, normalized by sample size, for DMSO and BQ-788 treated samples relative to full dataset (top) and for PF-431396 and UNC2025 (bottom)*

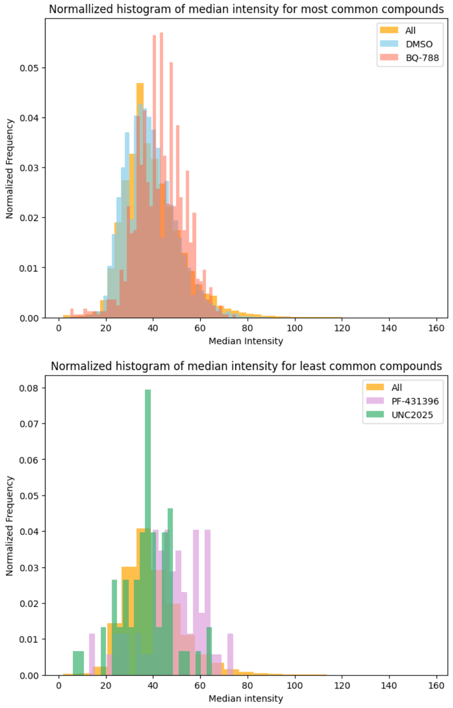

Figure 6: *Median intensity, normalized by sample size, for DMSO and BQ-788 treated samples relative to full dataset (top) and for PF-431396 and UNC2025 (bottom)*