Mini Project 2: How do cells respond to different drug treatments

(Due October 30th)

Project link: here

Background

In this study [1], data science techniques were essential to analyze millions of cell images. The paper presents a massive dataset containing three million images of cells that were treated with drugs and had certain genes modified. The data were obtained using fluorescent dyes to label different cellular components, such as the nucleus, cytoskeleton, and mitochondria. Scientists used this dataset to see how different genes and chemicals affect cell morphology (how cells look like). They used machine learning and image processing tools to extract and analyze features from microscopy images to find patterns and relationships between genetic changes and drug treatments.

Prompt: You are provided with a subset of the images present in the dataset. These images have been obtained after treatment with a small number of perturbations. Information about the type of perturbations can be extracted from the title of the file and the metadata provided. How well can you predict the drug treatment from any given image?

Reference Materials:

- **Main Paper** [1]: Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations | Nature Methods
- **Data:** Available to download from the journal website, but a curated subset will be added here too (subset should have cleaner images that can be run on a laptop RAM). Documentation about the curated dataset can be found here and metadata is available in the .csv file.
- **Starter kit**: https://colab.research.google.com/github/pr4deepr/cellpose-colab/blob/main/Cellpose_cell_segmentation_2D_prediction_only.ipynb
- Bonus points awarded if you use CellProfiler to extract cell features.

Expectations:

Writeup Requirements:

- **Abstract:** A concise summary of your findings.
- **Introduction:** Brief background information on the problem

- **Methods:** Detailed description of the data processing and analysis steps. Methods should have a section named *Code Availability*, containing a link to the public github account where the code is available.
- **Results:** Presentation of the analysis. What did you find? It should include references to at least two figures
- **Conclusion:** Summary of your findings and open questions, directions

Output:

- **Project Write-up:** less than 4 pages
- **Colab notebook or a project GitHub repo:** It should be able to run in a Google Colab environment (available through Columbia University). A link should be provided in the Write-up.

Suggested steps:

Unlike in project 1, there are no particular steps for this project. Read the paper and try to answer the prompt as well as you can, using the tools you learnt in class or other tools relevant to the class. A helpful read is the Veridical Data Science textbook by Bin Yu and Rebecca Barter. You can use chapter https://vdsbook.com/11-ls_binary **as a template for exploring your project. Please note that the data they provide is different from the data you are meant to use for the project.**

Resources from last time, which are still relevant:

- **Help in choosing colors for clustering:** https://color.adobe.com/create/color-wheel
- **Good figures**: Ten Simple Rules for Better Figures | PLOS Computational Biology
- **More information on format:** https://www.nature.com/ncomms/submit/article
- **Putting together your report:** https://www.overleaf.com/
- **Github help:** Getting started with your GitHub account - GitHub Docs