Introduction

For this project, you will download one of Recursion's recently released datasets that is drawn from a series of experiments focused on repurposing compounds for the treatment of COVID19 and answer a series of exploratory questions. Questions are intentionally under-specified, which gives you the opportunity to:

1. demonstrate your familiarity with generalizable "best-practices" for investigating a new dataset,
2. have the freedom to apply whatever ideas/approaches you've got in your back-pocket to address the questions, and
3. demonstrate your ability to distill high-level experimental objectives down to well-justified, quantifiable metrics.

Data summary

Our robots perform experiments and collect images of cells in 1536-well plates.. Cells isolated in each well are exposed to a specific set of conditions. In this dataset, we have exposed human renal cortical epithelial cells (HRCE) and African green monkey kidney epithelial cells (Vero) to Covid19 virus or two separate control conditions. Every image consists of five channels, each focusing a different cellular compartment or structure that is illuminated by fluorescent dyes, and for each well, we image four tiled locations, referred to as sites. We have 3 disease conditions,

1. Mock, - media taken from a live cell culture so it's got all the "cell juice" but lacks any virus
2. UV Inactivated SARS-CoV-2 - Virus in media, but all the viruses are inactivated with UV
3. Active SARS-CoV-2 - The actual CoronaVirus Condition

Once the images are collected, we transform them into embeddings with our own internal deep learning model. This creates 1024 features for you to do the analysis with, instead of downloading 450GB worth of images. There are two pieces of data you will need to download, the metadata and the embeddings. You can ignore the images for this work sample.

Data Description

Metadata - The metadata can be found in metadata.csv and downloaded from here. The schema of the metadata is as follows:

Attribute:	Description
site_id:	Unique identifier of a given site
well_id:	Unique identifier of a given well
cell_type:	Cell type tested
experiment:	Experiment identifier
plate:	Plate number within the experiment
well:	Location on the plate
site:	Indication of the location in the well where image was taken (1, 2, 3 or 4)
disease_condition:	The disease condition tested in the well (mock, irradiated or viral)
treatment:	Compound tested in the well
treatment_conc:	Compound concentration tested (in uM)
SMILES:	Formula of tested compound (as CXSMILES/ChemAxon Extended SMILES)

Features - The deep learning embeddings can be found in embeddings.csv and downloaded from here (n.b. this is 1.4GB). Each row in the csv has a site_id as described in the metadata schema. The remaining 1024 columns are the embedding for that respective site.

This dataset consists of 4 experiments in 2 different cell types, denoted in the experiment and cell_type columns, respectively. Each experiment is setup as a compound dose-response screen, in which "healthy" is represented as the UV Inactivated SARS-CoV-2 population (disease_condition) and the "diseased" is represented as the Active SARS-CoV-2 population (disease_condition; where treatment is NaN). Compound treatments (treatment) are present with a range of applied concentrations (treatment_conc) *and are applied to "diseased" wells*. Experiments consist of multiple plates and up to 1380 distinct wells (isolated mini-experiments) in each plate. The different sites for each well represent different views of that well— in these experiments there are 4 sites/records for each well. Experiments can be considered to be fully independent replicates of the entire screen, generally executed on different days/weeks. site_id and well_id are unique identifiers for all experiment, plate, well/site combinations.

Prompt

Your objective is to develop a method/pipeline to discriminate between the healthy and disease states of this set of experiments and communicate how well it works. With your approach, you will *also* predict/score all the compound-treated wells and communicate the sufficiency/applicability of your method for the purpose of identifying "good rescues", where "rescue" is defined as a compound-treatment which makes "disease" wells look like "healthy" wells.

We *strongly* recommend you "show your work", including any EDA that contributes to your choice of models or other answers. You may use any approach that you can **feasibly develop within a ~2 hour time-frame** (e.g. whatever preprocessing you want, linear or non-linear models, decision trees, regression/classification, etc.)— this is an exercise in EDA and MVP analysis pipeline development, not a quest to invent a brand-new method for evaluating biological data (though, if you find a way to do that too, well, show your work!). In other words, you should carefully budget your time and bias your efforts towards understanding the implications of your approach on the generalizability of your pipeline and interpretability of treatment scores— potentially at the expense of "better" solutions that take too long to develop & vet. Mentions of alternative solutions and approaches are encouraged if they can be justified by your observations. Your work may be used as a foundation for your subsequent interviews.

Given your selected approach, please provide brief answers and any supporting graphs/plots to the following questions:

Model selection

1.a) What is your chosen approach/model(s), and why did you choose it?

1.b) What assumptions does your method make about the features?

1.c) How well does your approach discriminate between healthy and disease states of the provided data?

Model generalization

2.a) How well does your model(s) generalize to different plates, experiments, or cell types?

2.b) If you were provided new data from new experiments (i.e. HRCE-3 and VERO-3), how would you expect your model(s) to perform?

Treatment scoring

3.a) What are the top 20 compounds, as determined by your measure of "rescue"?

3.b) What are the strengths and weaknesses to your selected approach to scoring treatment conditions?

3.c) Would you judge the treatment predictions/scores to be necessary and/or sufficient to conclude the treatment effectively rescues the disease effect (makes disease look like healthy)? Why or why not?

Responses will be reviewed for the following:

- Clear articulation of answers and presentation of supporting data (tables, plots, etc).
- Clean, concise, and executable code that effectively leverages whatever tools you choose to employ
- Effective use of your selected approach and the available data to draw appropriate conclusions

---