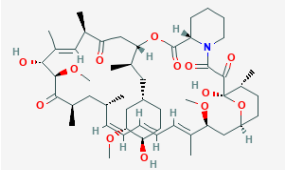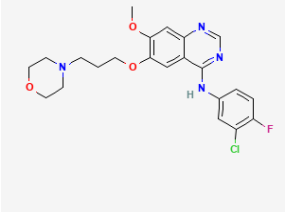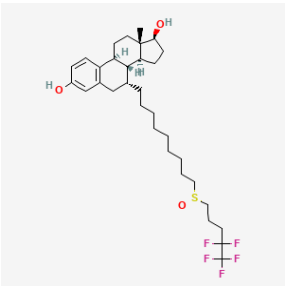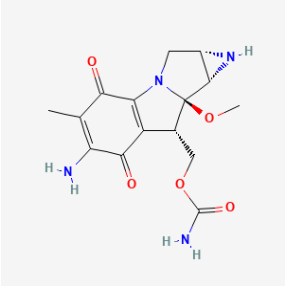# Hackathon Challenge #2: Drug Combination Response Prediction

Major goal: development of interpretable machine learning models to predict response to drugs and drug combinations. Model development can lead to clinically relevant information on drug mechanisms and which patients should receive drugs or drug combinations at an individual level.

Many cancer drugs have shown clinical benefit to various patients, but it is well-known that individual tumor response to effective drugs can vary. While some drug response variation may simply be due to cancer type (e.g., lung vs skin cancer), most reasons for individual variation to specific drugs remain unclear, and predicting individual variation to drug response is a complex challenge. In some cases drug response is due to resistance mechanisms caused by genetic mutations, or changes to the expression of genes in tumor cells.
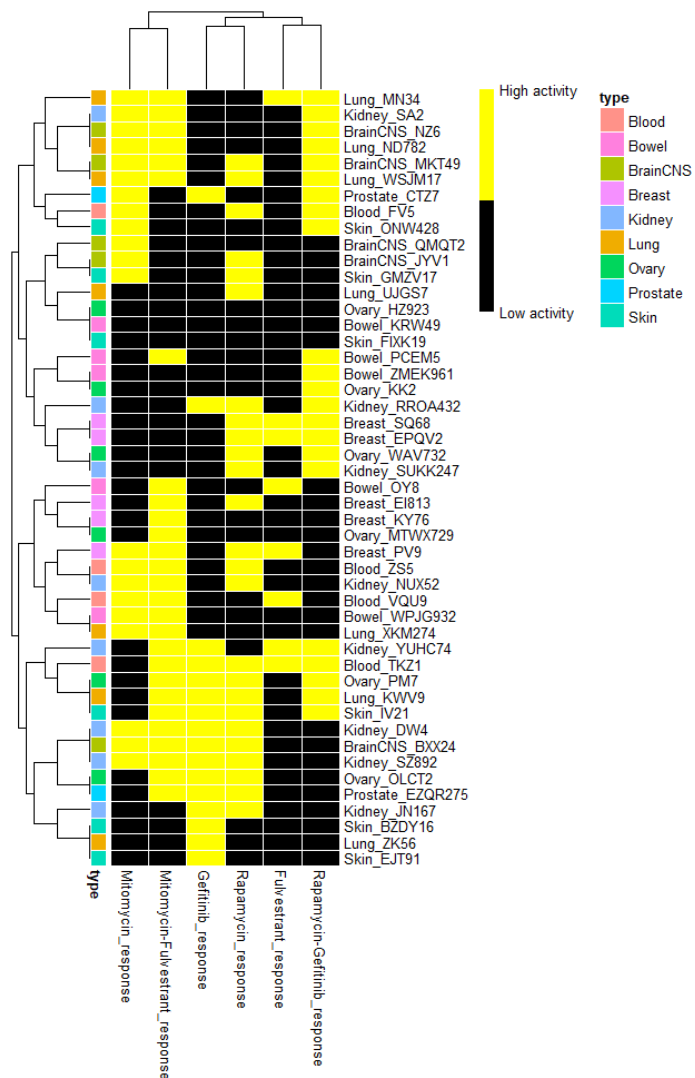
A research group has examined 4 FDA-approved drugs with variable efficacy across a panel of 9 different cancer types and wants to determine molecular features that predict drug response. Popular assays of molecular characteristics were performed across 48 cellular models of these 9 cancer types. The cellular models are known as "cell lines", which originate from tumor samples that grow in lab environments indefinitely. Based on information of the cancer type and given molecular features, development of predictive models that predict response to individual response or combinations will be the goal of this challenge, which will be scored by evaluation of prediction accuracy on test data that will be provided after submission. To develop models, participants will be provided with a drug response. Drug response was measured by determining the inhibition of cancer growth or cell death after transient drug exposure. In the data provided, values of 1 represent samples that were considered responsive (i.e., effective in killing cancer cells) to a drug or drug combination, and 0 represents a lack of response (representing cancers that are resistant to the given therapy). Matching molecular features are provided for the cell line samples where drugs were tested.

Drug background:

Drug	Structure	Pubchem link	Mechanism of Action
Rapamycin		https://pubchem.ncbi.nlm.nih.gov/compound/Sirolimus	mTOR inhibitor
Gefitinib		https://pubchem.ncbi.nlm.nih.gov/compound/Gefitinib	EGFR inhibitor
Fulvestrant		https://pubchem.ncbi.nlm.nih.gov/compound/fulvestrant	Estrogen receptor antagonist
Mitomycin		https://pubchem.ncbi.nlm.nih.gov/compound/mitomycin	DNA damaging agent

Participants are provided with drug response data for each drug. Rapamycin-Gefitinib combination and Fulvestrant-Mitomycin combination have also been provided. There are 6 different drug response models needed (4 for each monotherapy, and 2 combinations).

An overview of the drug activity in the samples that can be used for training is shown in the following heatmap:

Sample feature background:

Each sample has many features provided. Due to their high number, it is highly recommended to give attention to feature reduction/selection prior to building models. Engineering features based on the base features provided is allowed and recommended. Consistent with the goal of attempting to use machine learning to predict drug response based on tumor information prior to

any treatment being determined, molecular features provided in this challenge were measured in samples *before* treatment.

Cancer type labels:
Information of the broad tissue origin is labeled according to the following 9 classifications:
`"Breast"    "BrainCNS" "Bowel"    "Lung"      "Blood"      "Skin"`
`"Ovary"      "Prostate" "Kidney"`

Mutation data:
Mutations are changes to the genetic sequence in the genome. In this challenge, mutations to the protein coding gene sequences have been taken for each gene of every sample using whole-genome sequencing (WGS). Mutations are considered to play a defining causal role in cancer development and progression. In fact, anti-cancer drugs are commonly designed to be given to patients with specific mutations (known as "targeted therapies"). Despite capability to target key mutations with drugs, presence of specific mutations often fails to predict cancer response to targeted therapies.

After "high-throughput sequencing" of the entire genome (containing >20,000 protein coding genes), mutations which were considered likely to impact protein function were identified in each sample. In the data provided, these mutations are represented in a binary format, with 1 representing a mutation, and 0 representing a non-mutated (normal) gene. These features are labeled as "mut_" followed by the gene's ID.

RNA data:
Ribonucleic acid (RNA) is the genetic product of genes that are read and transcribed from DNA in the nucleus. RNA is later converted into protein, which is considered the functional output of our genetic code. Proteins will perform numerous proteins and their abundance in the cell is mostly determined by the RNA levels of corresponding gene sequences. RNA levels (sometimes referred to as mRNA or transcript abundance) have innumerable associations with cellular functions, including tumor response to various drugs. RNA is most frequently measured for every gene by RNA sequencing (RNA-seq) assays.

In data provided, RNA-seq data has been processed in a standardized format. After performing a standard length normalization that converts counts of genes read in sequencing to Fragments Per Kilobase Million (FPKM), it has been provided in the format of log2(FPKM+1). These features are labeled as "rna_" followed by the gene's ID.

Model Requirements

Any model type that can be executed in the coding environment provided can be used. Popular machine learning tools and libraries are likely helpful. There are three general requirements:
1) Be able to make predictions on new data when provided in the same format as the training data
2) Predict drug/combination response in the same format as the labeled respective drug response columns (1 for samples that respond to the drug, 0 for samples which do not)
3) Have some form of interpretability. Assessment of the most important feature of a model's predictions is required. Popular methods like calculation of Shapley values can

be used, and there is no limitation on methods used for determination of important features.

Submission and Scoring Requirements

1. Coding scripts for model generation and saved model objects.

2. Model descriptions and predictions on test data (Level 1 of submission scoring). Scoring will be based on accuracy.
   a. An overview of descriptions and metrics for models used
   b. A form of predictions on test data, as well as identification of the key feature of the model used in a prediction

3. Short answer questions related to aspects of model training and interpretation (Level 2 of submission scoring). For the top scoring teams in predictive accuracy, short answer questions will be evaluated and have 20% weight.
   a. Questions in a text file will be provided, which can be directly edited for answer submission

Details of form submission of prediction results and questions will be provided in Code Ocean data assets. Scores on part 2 will be calculated automatically from the accuracy of each prediction required in the form.