CELL PAINTING INTERACTIVE VISUALIZATION DASHBOARD PROJECT

Team 93 CSE 6242 - Fall 2023

Jennifer Tian, Francis Lin, Hunter Lonon, Stanislav Sheludko

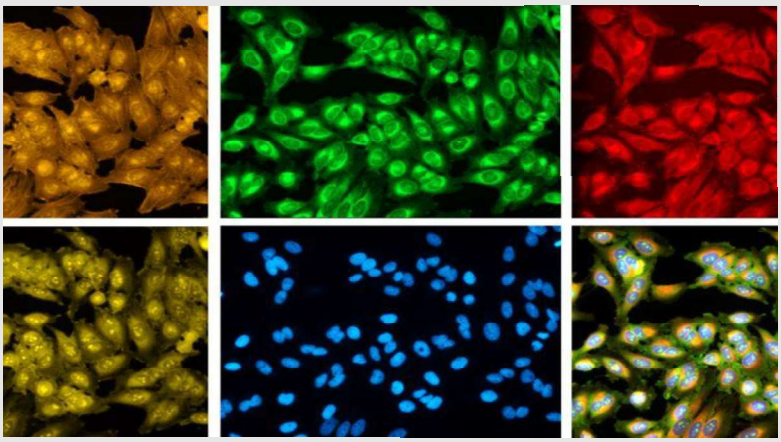
INTRODUCTION

What is the problem?

- Cell painting is a drug discovery strategy, imaging of dyed cells to capture information about cell response to various compounds.

Why is it important?

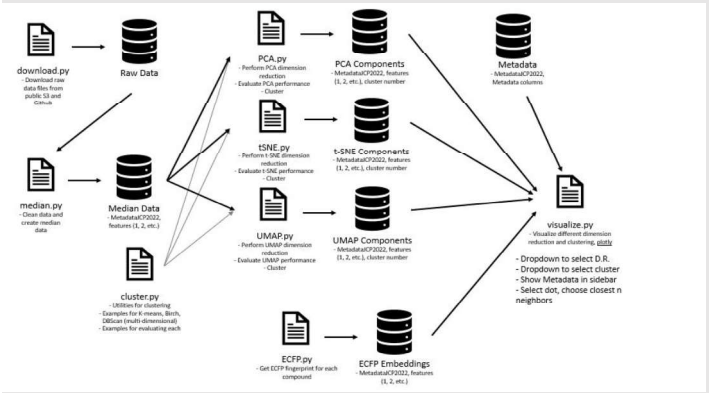
- This dataset is novel. It is also large and computationally expensive to process. There is not a standard for processing and analyzing this type of unlabeled cell image data.
- Our goal is to create an interactive visualization dashboard for researchers to facilitate drug knowledge and discovery.



Credit: Broad Institute of MIT and Harvard

APPROACH

- The JUMP dataset – 115 terabytes of data – consists of images of 1.6 billion cells subjected to 136,000 compounds and genetic changes. Over 4,700 features derived from images.
- We provide 12 different combinations of dimensionality reduction / clustering methods: UMAP, PCA, t-SNE, ECFP/UMAP for feature reduction and k-Means, BIRCH, and DBSCAN for clustering to determine the best approach to handling the dataset.
- We assess the success of each approach through qualitative evaluation of clusters and closest neighbors from the visualization. We also utilize a ChemPub API and rdkit.Chem to assess chemical compound similarities within clusters.



DATA

- The data was downloaded from JUMP public repositories on AWS and Github using boto3 package in python.
- The data set consists of 2,378 parquet files amounting to over 1 million rows (perturbations) and 4,700 columns (features), 50 GB in size.
- Then the data was read into a dataframe using dask package and median values were calculated for each feature based on the compound used, further preprocessing steps were taken to deal with missing values.
- Resulting dataframe shape is 138,906 x 3,648.

EXPERIMENTS & RESULTS

- PCA, UMAP, and t-SNE were used to reduce dimensions for phenotype similarity comparisons.
  - Optimization of feature reduction methods to retain 95% of variance.
- ECFP + UMAP feature reduction applied to compound data for structural comparisons.
- Kmeans, BIRCH and DBSCAN were used on each reduced dataset for clustering.
  - Optimization of clusters based on elbow plot/adjusted rand score.
- Dash/plotly was used to create visualization.
- Results were evaluated by examining pairwise chemical compound similarity using the Tanimoto coefficient for different feature reduction and clustering methods.
- Based on our evaluation, UMAP and PCA were the feature reduction methods that produced the highest degree of chemical compound similarity within one method of clustering. Birch clustering produced the highest chemical compound similarity within clusters overall compared to Kmeans and DBSCAN.