Take-Home Interview Question - Bio+ML

Data

Link to data: https://drive.google.com/file/d/1jW_YaLzstEWaHZ7LrWIWJpt0gZCRW-Z7

You are provided with a parquet file (phh_prod_image_data_oasis_with_dmso.parquet) containing image embedding data from cell painting (Bray et al, 2016) images of pooled primary human hepatocytes treated with molecular perturbations (compound, dose) for 48 hours collected from ~1k unique molecules (OASIS consortium data). The data has been aggregated at a well level by taking the image mean embeddings across 15 fields of view (FOVs, often referred to as "sites") into a single embedding vector for each row in the parquet file (**plate**, **well_id**), leaving us with about 17,408 well-level embedding rows.

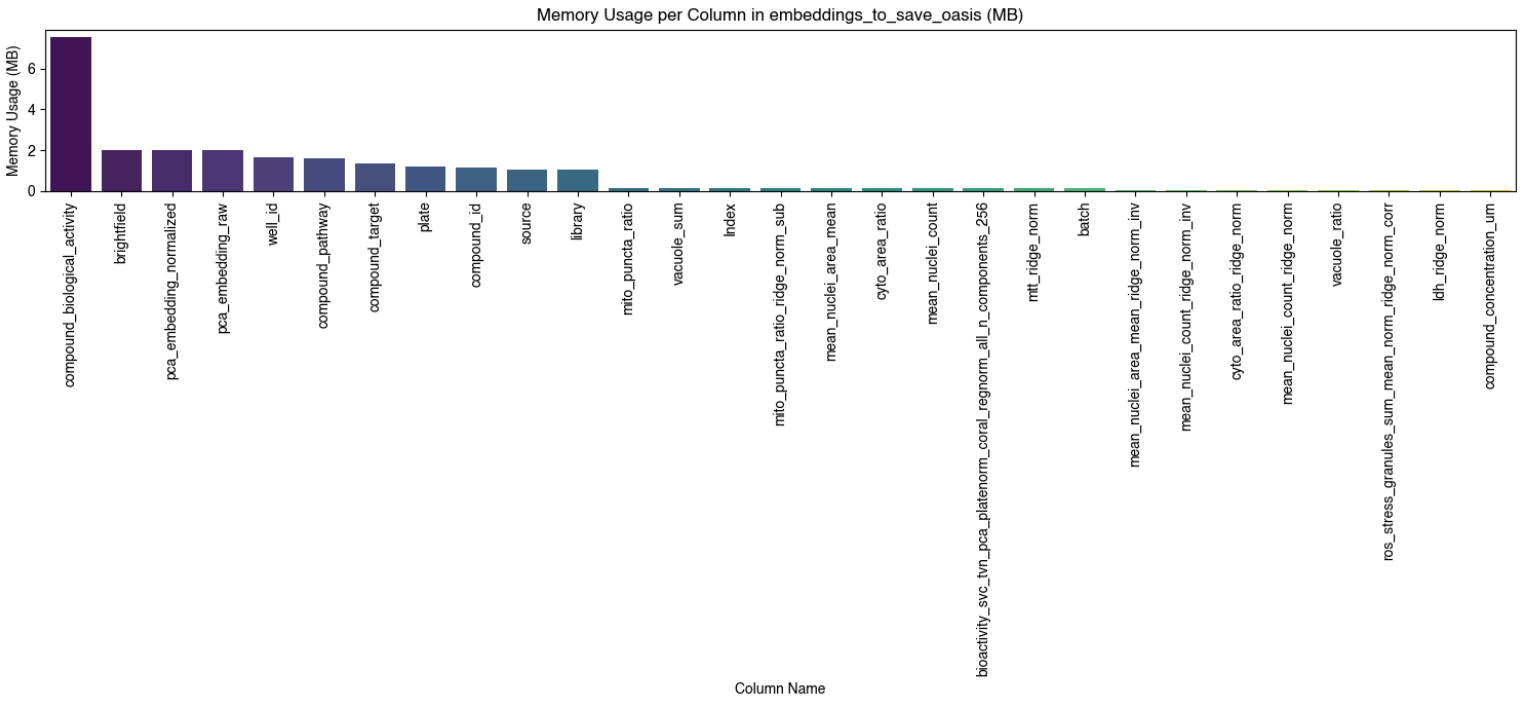
The columns provided are:

- Metadata
  - **library** is the library the molecular perturbation was sourced from - library == oasis was used to filter for these data
  - **source** is the screen batch (technical batch)
  - **plate** is the unique identifier for a 384 well plate
  - **well_id** is the unique identifier for the well position within a 384 well plate
- Biological activity annotations: the vendor's annotations are provided, where available
  - **compound_biological_activity**: a textual blurb describing the compound's activity, ex. *"Vinblastine (sulfate) (Standard) is the analytical standard of Vinblastine (sulfate). This product is intended for research and analytical applications. Vinblastine sulfate is a cytotoxic alkaloid used against various cancer types. Vinblastine sulfate inhibits the formation of microtubule and suppresses nAChR with an IC50 of 8.9 μM."*
  - **compound_taget**: Med Chem Express target annotations. NOTE multiple labels allowed, separated by semicolon and leading space. Example for Vinblastine (sulfate): *Autophagy; Microtubule/Tubulin*
  - **compound_pathway**: Med Chem Express pathway annotations. NOTE multiple labels allowed, separated by semicolon and leading space. Example for Vinblastine (sulfate): *Autophagy; Cell Cycle/DNA Damage; Cytoskeleton*
- Perturbation ID
  - **compound_id** contains IUPAC or common compound name for the molecule. DMSO = control, i.e. no bioactive compound applied
  - **compound_concentration_um** contains the dose which a molecule was dosed, in micromolar units
- Image embeddings
  - **brightfield** contains raw, un-normalized DINOv2 image embeddings from the brightfield channel only (1x768)
  - **pca_embedding_raw** contains DINOv2 embeddings for the five fluorescent channels which were concatenated end-to end (5x768), and then reduced to 128 dimensions with PCA. PCA was fit using only the first batch (first 10 sources).
  - **pca_embedding_normalized** is computed for each source separately as (**pca_embedding_raw_source** - **mean_source_dmso**) / **std_source_dmso**, where **mean_source_dmso** and **std_source_dmso** are the elementwise mean and std over DMSO vehicle control **pca_embedding_raw** embeddings corresponding to a given source.

Additional assay feature columns are also provided:

- Biochemical assay features: contain ground truth viability and toxicity data
  - **mtt_ridge_norm**: MT-Glo viability assay which has been normalized to regress out **well_id** positional effects (1=viable, no change from DMSO. 0 = no viability signal).

    - **ldh_ridge_norm**: LDH cytotoxicity assay which has been normalized to regress out **well_id** positional effect (0=viable, no change from DMSO. 1 = maximum cytotoxicity).
- Image features: the remaining columns contain UNET segmentation features
    - **mean_nuclei_count**: the mean nuclei count in a FOV. Should be positively correlated with **mtt_ridge_norm,** and negatively correlated with **ldh_ridge_norm**
    - **mito_puncta_ratio:** the ratio of fissioned mitochondria (objects < area threshold) to all mitochondria, by area
    - **mito_puncta_ratio_ridge_norm_sub: mito_puncta_ratio** that has been normalized to regress out well_id effect
    - .. other image features corresponding to other endpoints



Task

Your task is to interrogate the data, identify how we should process or normalize embeddings, and how we can extract information from them about phenotype and mechanism of action (MoA).

1. Exploratory data analysis: Please take some time to get acquainted with the data! Ex. How many replicates are there? How many compounds have biological activity annotations? How many unique target and pathway annotations are there? What about source and batch? How many sources have image feature data? Which features are correlated, and which are anti-correlated?

2. Explore the embeddings: pca_embedding_normalized is a good place to start. How can you use the existing embeddings to derive information about phenotype and MoA for a new embedding which doesn't have these annotations? You might consider ideas like clustering, predicting target/pathway, or associating with the biological activity text field. Use the assay feature columns as a baseline.

3. How could you evaluate how good a particular set of embeddings is? You might consider how well they perform for your task above, how strongly they correlate with attributes of the drug perturbation used to treat the cells such as target, pathway, and biological activity (good), and how strongly they associate with

non-perturbation confounding variables such as plate, source, and batch. Use the assay feature columns as a baseline.

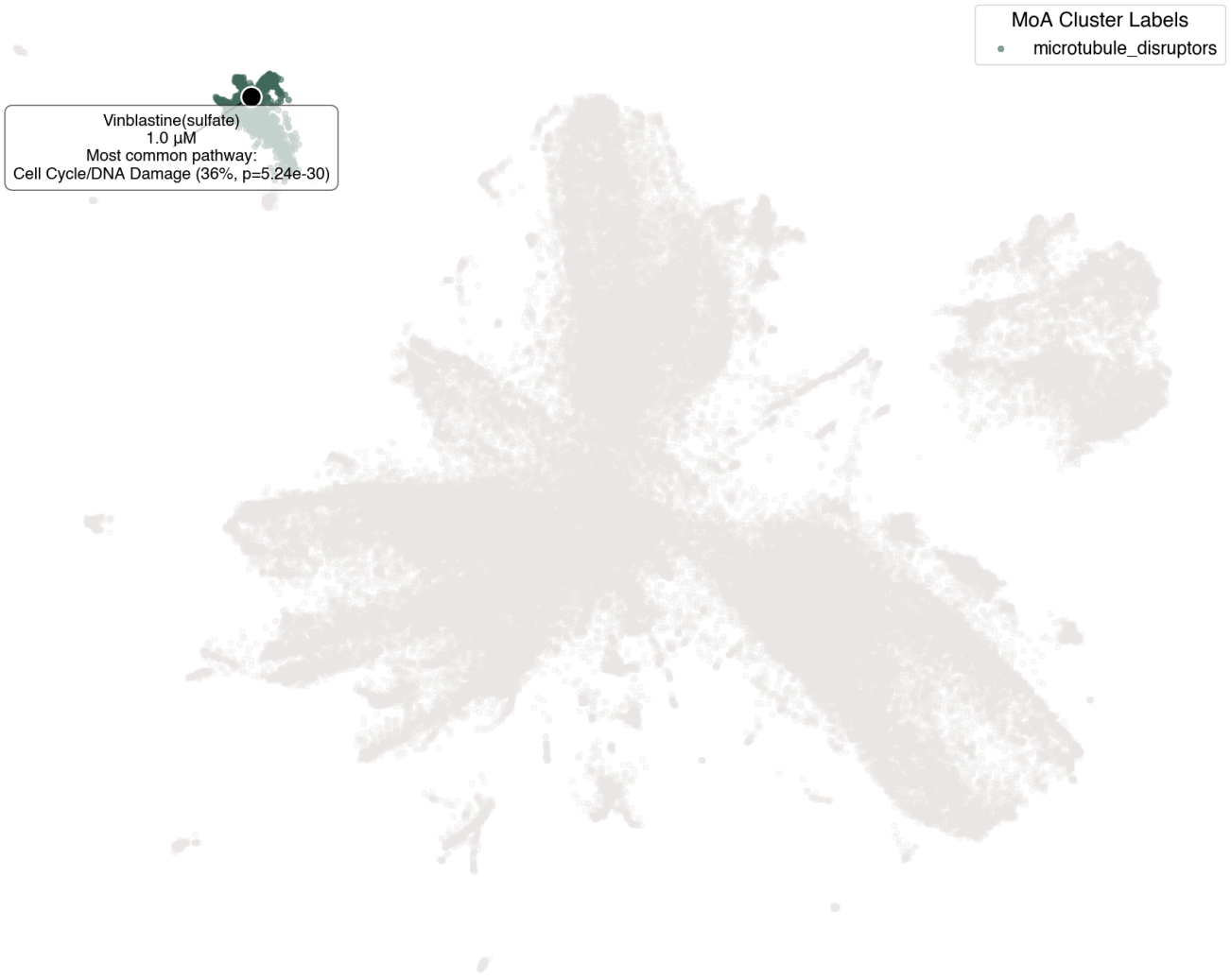
4. Given your evaluation criteria, can you come up with any methods to normalize or process the either the raw pca embeddings or brightfield embeddings to obtain useful and phenotypically meaningful representations? Can you derive better cell painting embeddings than pca_embedding_normalized?

5. What are some examples of drug attributes the embeddings (pca_embedding_normalized, or your brightfield or cell painting embeddings) can capture well, and that are captured poorly?

BONUS

1. Can you pull in any publicly available biological annotation data sources to aid in the identification of useful information in these embeddings?

Example 2D projection of all bioactive well-level image data (>75k wells), with microtubule inhibitor cluster containing Vinbalstine:



Sourced from HESI 2025 poster https://colab.research.google.com/drive/1BPps4GzQTp0GjfsUZaOPXA1VUhwTnuHt

Expectations

1. Please provide Python code and analysis/report. A Jupyter notebook report with supporting .py files is a good option; other options like .py code and a report in Google Docs also work.

2. Please feel free to use any tools you'd like to help complete this task. For example, Cursor, claude code, chatgpt, using any software packages or data online, or Googling are all OK.
3. Please don't hesitate to ask us questions: no question is too small.