Multimodal Perturbation Modeling Using Transcriptomic and Imaging Data

Harshi Saha, Lee Chen, Vivian Hir
6.8700 Fall 2025

Abstract:

Advances in high throughput technologies enable large scale characterization of cellular responses to chemical perturbations, providing insight into molecular and phenotypic states relevant to drug discovery. This includes the Cell Painting assay, which captures cellular morphology through imaging, and the L1000 transcriptomic assay, representing expression of landmark genes [1, 6]. These modalities have been used independently to investigate the biochemical interactions through which drugs produce effects in the body, also known as the mechanism of action (MoA), though with limited success [2, 3, 11]. Integration of these complementary modalities to obtain biologically meaningful models of perturbation effects is underexplored, attributed to the historical lack of large and publicly available multimodal datasets [2, 3]. Despite emerging evidence that combining these modalities improves MoA modeling and biological interpretability, there are no widely accepted state of the art methods in this field, and existing baselines employ simple modality integration techniques with high variation in class specific performance [2, 3, 5].

In this project, we use one of the largest publicly available datasets with paired L1000 transcriptomic and Cell Profiler morphological profiles, released by the Broad Institute, to advance multimodal perturbation modeling with a specific focus on MoA [3]. We integrate both modalities and develop representation learning frameworks, using the resulting unified embeddings as richer representations of drug mechanisms than MoA alone. Such representations can support more comprehensive models of cell responses to perturbations, enabling MoA prediction and biologically interpretable analyses across transcriptomic and phenotypic feature spaces [2, 3, 5]. Examining the learned embeddings can identify modality driven patterns, highlight shared versus unique biological signals, and characterize perturbation specific signatures at the gene, pathway, and organelle levels [2, 3, 5, 12]. Therefore, a representation learning approach to embedding generation and subsequent MoA prediction has the potential to address gaps where robust and consensus state of the art approaches are yet to be established.

In this paper, we implement two models to create embeddings for each perturbagen: a late-fusion supervised contrastive learning model and a representative learning model. Both models used L1000 and Cell Profiler features to create embeddings, after which k-nearest neighbors, Gaussian mixture models, or a multilayer perceptron were used to classify the data points into likely mechanisms of action. Our best model combined the contrastive learning model with k-nearest neighbors, with high specificity for the PLK, HDAC, and MEK inhibitors. However, for other mechanisms of action, the combination proved to be a poor predictor. In all, our study serves as a framework for trialing different models for predicting mechanisms for a noisy biological dataset.

Background:

In drug discovery, there has been an explosion of high throughput methods to characterize cell lines in response to perturbations. Methods include imaging and transcriptomic assays, and increasingly used for these efforts are Cell Painting and L1000 respectively. Cell Painting is a microscopy technique that enables researchers to understand the structural details of a cell at an organelle level, including the nucleus and mitochondria, allowing for morphological and positional analysis [1]. The resulting images are usually processed using the Cell Profiler tool, which uses segmentation techniques to extract continuous numerical features that capture the aforementioned structural details [3, 4, 10, 12]. L1000 on the other hand, is a bulk-transcriptomic assay that measures the expression of ~1000 landmark genes, which capture >80% of the variation in the transcriptome such that the majority of the expression levels of all other genes can be computationally estimated [3, 6, 9]. These assays allow inference of cellular states at two distinct levels, and can help determine how perturbations affect molecular pathways in cells [3, 5, 12].

Prior work in drug discovery has focused on each of these modalities independently to infer the mechanism of action of a compound, largely due to the unavailability of public and large scale datasets that pair these assays for the investigated chemical perturbations [3]. However, in 2022, the Broad

released such a dataset, using which researchers have started exploring the area of multimodal characterization of compound effects using associated the L1000 and Cell Profiler features [3, 5, 12]. Because of the emerging nature of this area, there are yet to be widely accepted state of the art methods for integration of these modalities to characterize drug perturbation and predict MoA class. Early efforts that serve as baselines attempts to incorporate L1000 and CellProfiler modalities for MoA prediction use clustering in the unsupervised case, logistic regression and MLP classifiers in the supervised case, and early and late stage integration of modalities showed comparable performance and high variation in class specific prediction performance, with poor predictive power for the majority of MoA classes [3].

In this study, we propose to use representation learning frameworks to integrate transcription profiles from the L1000 assay with morphological features extracted from Cell Painting images using Cell Profiler. By applying such methods to these complementary modalities, we aim to derive joint embeddings that capture a more comprehensive model of cellular responses to perturbations. This integration will potentially allow us to improve predictive performance of perturbation effects and help reveal patterns and information in one modality that are uniquely specific to that modality, or that are correlated and even causally associated to the other [3, 12]. In addition to supporting MoA classification, this would allow for downstream analyses to increase biological interpretability of drug mechanisms, and if these efforts are successful, it would not only represent progress for drug discovery in terms of the time and costs saved for selection and prioritization of candidate compounds, but it would also accelerate the identification of mechanistic workings to help predict off-target effects, toxicity, and even guide compound design.

Representation learning, through both early and late fusion methods, has been used to predict one modality from the other, but our approach of using such methods to integrate the modalities specifically for MoA prediction and characterization is still novel and understudied [3, 5]. We chose to build the peturbagen embeddings from scratch using various representation learning frameworks, testing both early and late fusion strategies and their effect on classification performance. Frameworks such as contrastive learning are especially useful for downstream tasks of classifying mechanisms because the resulting embedding space encodes information about the relationships between each mechanism of action. This project therefore explores a novel method and application of the integration of the specific assays, and these investigations can serve as a starting point for establishing state of the art model architectures in the multimodal modeling of chemical perturbations, MoA prediction, and reverse reasoning to understand patterns across the transcriptome and phenome, that are yet to be established.

Results and Discussion:

Data Preparation. The paired L1000 assay gene expression and Cell Profiler data were downloaded from the Cell Painting AWS bucket maintained by the Broad Institute [3]. This dataset includes data from 4 different sources, and the one focused on in this project is the LINCS-Pilot1 dataset, which contains data for chemical perturbations [3, 8, 9]. The LINCS-Pilot1 dataset was split into a training (p=0.8) and testing set (p=0.2), with no datapoints with the same mechanism of action (MoA) being shared across the two groups. The training set was used to develop the representation learning method and for analysis of transcriptomic and morphological patterns associated with generated embeddings clusters, while the testing set was used to evaluate the generalizability of the modeling strategy proposed in this project to characterize chemical perturbations.

The dataset additionally contains SMILES strings, multiple replicates for each of the chemical perturbations in each of the modalities, and MoA annotations. The metadata for both modalities across the datasets was made consistent using processing scripts provided alongside the source data in the reference paper [3]. The L1000 data (samples x genes) had affymetrix microarray probe set ids as column labels, which were converted to gene symbols to ensure downstream interpretability, and literature review was conducted to enable interpretation of morphological characteristics captured by the provided Cell Profiler segmentation features [10].

Based on the results of exploratory data analysis on the LINCS dataset across modalities, the inputs to the models in this project will be L1000 gene expression data and the Cell Profiler derived Cell Painting features. Note that although the raw Cell Painting assay images associated with the chemical

perturbations were made available publicly, due to issues arising from the size of the datasets and the space and computational power available for use in this course project, we will not be incorporating this modality. Furthermore, these modalities will be used to generate embeddings for each of the perturbations via representation learning, and clusterings of these embeddings will be investigated to uncover shared patterns in the transcriptomic and morphological modalities. The MoA annotations provided in the LINCS dataset will be used only for downstream analyses.

Data Preprocessing. Prior to joining, the L1000 and Cell Profiler datasets were preprocessed individually. The L1000 was downloaded as a cell-normalized and scaled dataset and was kept as such prior to joining. The Cell Profiler contained comprehensive measurements of a cell's features (e.g., texture, area, and shape). Measurements independent of the cell, such as its location within the Cell Painting assay well, was removed prior to processing.

Within Cell Profiler measurements, some features did not have any variance between different perturbagens and thus we removed features with a variance of less than 0.01. Still, many features from Cell Profiler are highly correlated. So, we iteratively calculate pairwise correlations between each feature and remove one element of the pair if their correlation is greater than 0.9. Due to the noisiness of the Cell Profiler dataset, we also used RobustScaler (scaling based on median and interquartile range instead of mean and standard deviation) to scale each Cell Profiler Feature. To calculate the median and interquartile range, we subset the data from entries labeled with a perturbagen of DMSO and a MoA of negative control. Finally, although we removed correlated features previously, we ran ZCA whitening to remove remaining covariance between pairs of features.

Although the L1000 and Cell Painting assays were described to be paired assays run on the same well, we were not able to find an identification tag to unify our two datasets. So, we generated our final dataset based on the Cartesian product of the L1000 and Cell Painting datasets (i.e., for groups of data points with the same perturbagen and dosage in the L1000 and Cell Painting dataset, we generated a new result for each L1000 and Cell Painting result combination). Although this approach ensures complete pairing coverage, it also likely introduced many false positives, increased the effective sample size, and reduced the statistical power. More dataset mining is needed to determine the true dataset pairing.

To ensure our model is robust to determine MoAs for new molecules, we separated the experiment replicates such that each perturbagen is only in the training or testing set. The full set of MoAs was available in both the training and testing set.

Model Architecture: To develop richer representations of each mechanism of action (MoA), we built a late-fusion contrastive learning model to learn embeddings of MoAs for each perturbagen. First, we build individual encoder heads for the Cell Profiler and L1000 assays. Each encoder head included a 512 dimension fully-connected layer, batch normalization, a ReLU activation function, dropout, followed by a 256 dimension fully-connected layer. After both datasets were passed through their respective encoder heads, they were concatenated and passed through the fusion head, which included a 256 dimension fully-connected layer, a ReLU activation function, dropout, and a 128 dimension fully-connected layer to produce the latent embedding. Prior to being passed for contrastive learning evaluation, the embeddings are L2-normalized.

To implement contrastive learning, we used SupCon loss, which extended from InfoNCE loss. We labeled perturbagens that had the same MoA as samples that are within the same class and all other perturbagens as samples with a different class. The fully-connected layers are trained based on this loss.

$$L_i^{\text{SupCon}} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(z_i, z_p)/\tau)}{\sum_{a \in A(i)} \exp(\text{sim}(z_i, z_a)/\tau)}$$

Contrastive Learning Results: The late-fusion contrastive learning model was trained on the training set and validated on the validation set.

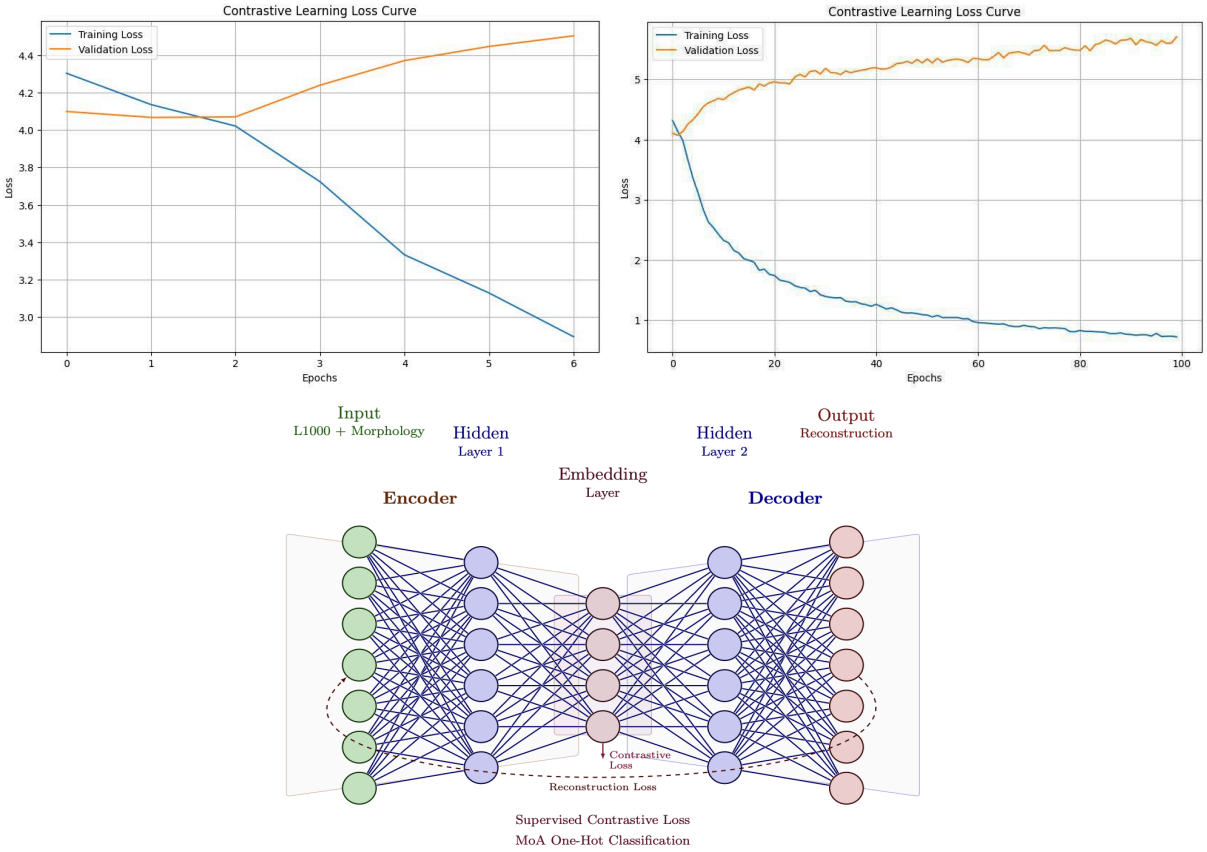

Figure 1. Training and validation loss of a supervised late-fusion contrastive model, with (a) and without (b) early stopping. (c) Architecture of the late-fusion autoencoder trained with supervised contrastive loss.

MoA Class Prediction Setup: To predict the MoA classes on the test embeddings, we used three classifiers in the following order: KNN, MLP, and GMM. We initially used KNN because KNN is the most simple and straightforward one out of the three, and then did MLP because the paper used MLP for classification. As for GMM, we wanted to compare the GMM's performance with the other two.

The MoA class-specific F1 scores were used to evaluate the model's performance. We did classification and model performance evaluation on four types of embeddings: uncollapsed, collapsed, raw+uncollapsed, and raw+collapsed. Uncollapsed embeddings came from the original preprocessed dataset which acted as the control because they contained multiple trials for each perturbation ID dose. On the other hand, the collapsed embeddings came from collapsed data, which was the median of all the features for each perturbation ID dose.

The purpose of having the raw+uncollapsed and the raw+collapsed groups is that the paper tested the model performance on the raw features themselves, with fairly good results. As a result, we wanted to investigate whether adding raw features in addition to the embeddings would lead to improved or worse model performance. Before using a classifier, we performed L2 normalization on the embeddings.

For the KNN, we graphed the cross validation score for a range of k-neighbors from 1 to 100 to find the optimal k for the type of embedding. For contrastive learning, the number of neighbors was 50 for uncollapsed, collapsed, and collapsed+raw. Meanwhile, uncollapsed+raw had 100 neighbors. For the representation learning model, we used 40 neighbors for collapsed, 75 for uncollapsed and raw+collapsed, and 100 for raw+uncollapsed. A cosine similarity metric was used for each KNN.

For the MLP in the contrastive learning model, we used a hidden layer dimension of 32 by 32 with an alpha of 0.0001, an initial learning rate of 0.001, and a batch size of 64. A batch size of 64 was on the higher end, but was used for faster speed compared to 32 or 16. ReLU activation was used. Meanwhile, the MLP in the representation learning model had a hidden layer dimension of 256 by 128, and the other parameters were the same as the ones used in the contrastive learning model.The larger dimensions for the hidden layers came from the representation learning model generating 165-dimensional embeddings, which are much larger than the 6-dimensional embeddings from the contrastive learning model.

For the GMM in the contrastive learning model, we used a full covariance type. Class priors were computed from the training set, and then we trained 1 GMM per class. For the representation learning model, we used PCA to reduce the dimension of the 165-dimensional embeddings to 10 components, and then used the diagonal covariance type to prevent overfitting. Like the previous model, we trained 1 GMM per class.

Contrastive Learning Model Performance: Out of the three models, KNN performed the best, while MLP performed the worst. GMM's performance was close behind that of KNN. 16 out of the 72 MoAs for the KNN had non-zero F1 scores, whereas the paper had 34. Meanwhile, GMM had 14. For the KNN, the classes with the highest F1 scores were PLK, HDAC, and MEK inhibitors. PLK and MEK are kinases, while HDAC is a histone deacetylase. Out of the four groups, the collapsed embeddings performed the best, while the raw+collapsed performed the worst. Compared to the paper's F1 scores, our model performed better for VEGFR, but worse for EGFR. The GMM's three highest-scoring F1 MoA classes were the same as the ones in the KNN, but the MEK scored higher on average while the HDAC inhibitor scored lower on average.

These results suggest that the cell morphology features from CellProfiler and the transcriptomic information from L1000 assay may be predictive of MoA class to some extent. A possible reason why collapsed performed the best out of the four groups is that by using a dataset that took the median of the features, this prevented outliers from influencing the embeddings. As for the raw+collapsed group's low performance, not normalizing the raw features before training the model may be a factor.

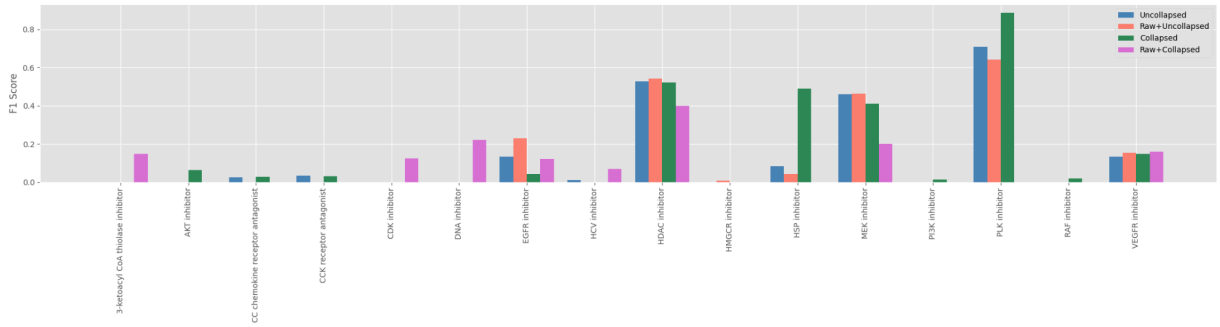

Figure 2. The F1-scores of the nonzero MoA classes using the KNN for the contrastive learning model. Blue is uncollapsed, orange is raw+uncollapsed, green is collapsed, and pink is raw+collapsed.

Representation Learning Model Performance: Compared to the contrastive learning model, the F1 scores for all three classifiers were worse, as there were much fewer classes with nonzero F1 scores, and these values were lower. The KNN returned 3 nonzero F1 MoA classes, the highest one being the VEGFR inhibitor, which was also found in the KNN in the contrastive learning model. The MLP performed the worst, as it only returned the VEGFR inhibitor, with F1 scores ranging from 0.12 to 0.28 across the 4 groups. On the other hand, the GMM performed the best out of the three, as it returned 9 nonzero F1 MoA classes. However, only 2 out of 9 classes had more than one group, as the rest only had one; these values were below 0.15, with the exception of the uncollapsed group having a F1 score of 0.6 for HDAC inhibitor. It is worth noting that the GMM did not return VEGFR inhibitor as a nonzero F1 score MoA class.

Relative to our supervised contrastive loss model, a possible reason for the poorer performance is the limited number of samples in both our uncollapsed and collapsed schemes for the training dataset. The collapsed scheme resulted in n=1,560 points for our dataset. For both embedding and unsupervised

methods, we want significantly more samples to account for the diverse gene expression and morphology landscape of our dataset. The supervised contrastive loss model incorporated additional information about the MoA, leading to its increased performance relative to the representation learning model.

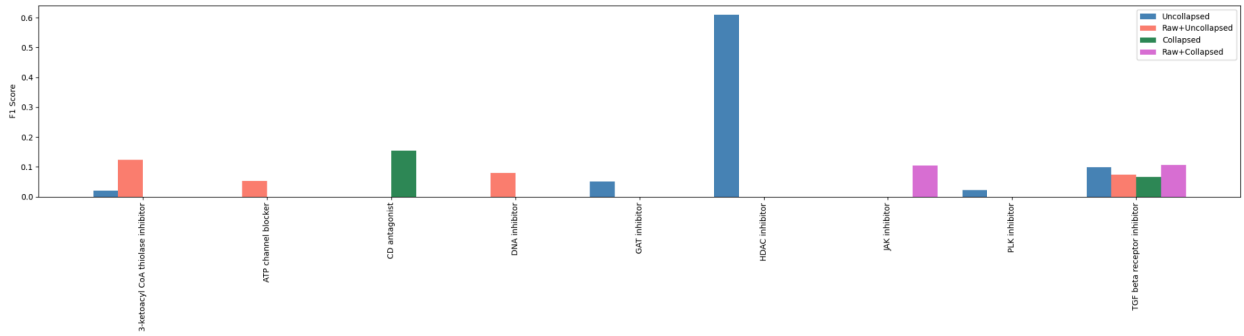

Figure 3. The F-1 scores of the nonzero MoA classes using the GMM for the representation learning model. Blue is uncollapsed, orange is raw+uncollapsed, green is collapsed, and pink is raw+collapsed.

Visualizations of Embeddings: Using the dimensionality reduction techniques t-SNE and UMAP for visualization of the embeddings did not show any significant clustering patterns. Most, if not all, the MoA classes were scattered across both plots for the four types of embeddings, resulting in no clear pattern. However, the VEGFR inhibitor (label 70, brown) and the PLK inhibitor (label , pink) classes formed a small cluster in the t-SNE and UMAP for the collapsed embeddings using KNN in contrastive learning. The t-SNE and UMAP plots for the uncollapsed embeddings were not used for further analysis because the data had many trials containing the same perturbagen ID dose, which resulted in correlated data points and an overconfident prediction.

Overall, the visual results of the t-SNE and UMAP plots generally aligned with the F1 score results for model performance, because few MoA classes had nonzero F1 scores, suggesting that the classification performed poorly.

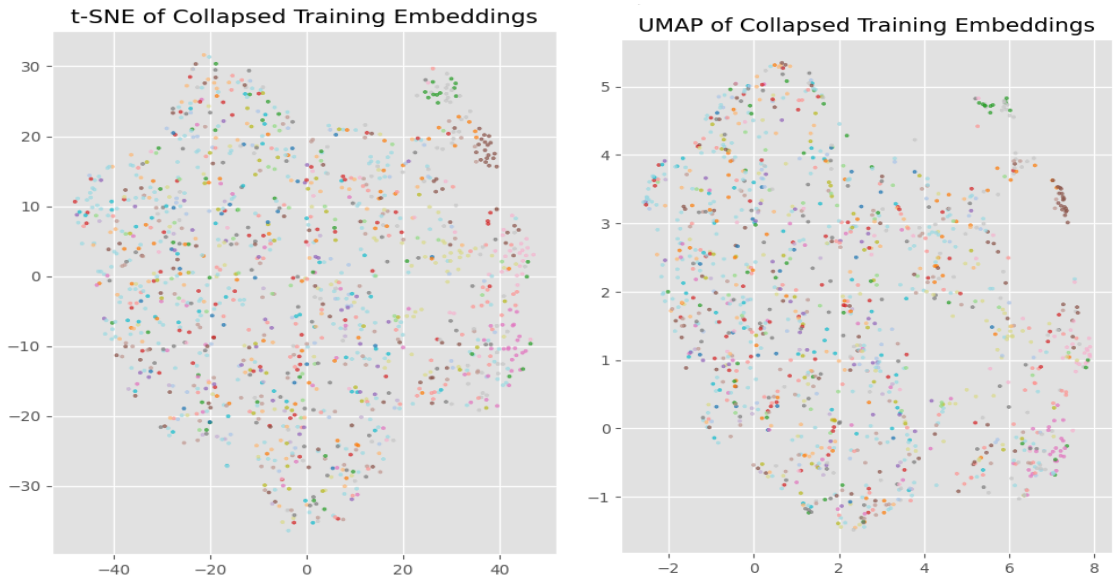

Figure 4. The a) t-SNE and b) UMAP plots for the collapsed training embeddings obtained from using the KNN in the contrastive learning model. There are 72 MoA classes.

Future Goals:

To investigate how well our model classifies mechanisms of action, we will classify new data from the Bray dataset. Although both the LINCS and Bray datasets contain L1000 and Cell Profiler assays, we expect that there will be slight experimental methodological changes that will allow us to test the robustness of our models. To do this, we will calculate the embedding for new perturbagen using the associated gene expression and Cell Profiler data and determine its class using k-nearest neighbors. In the case that our model is effective in classifying samples from the Bray dataset, then we have learned non-dataset-specific embeddings of the mechanisms of action.

Currently, our task is to learn embeddings that are more granular to the MoA. In our original dataset, the MoA is annotated using a one-hot encoding, which loses information on which MoAs are more similar to each other. Intuitively, two protease inhibitor pathways such as a serine proteases inhibitor pathway and matrix metalloproteinase pathway should be similar, though the one-hot encoding does not express this similarity. In our embedding space, we'll be able to learn these similarities and provide insight on which genes and cell morphologies correspond with each mechanism of action.

One area to explore in the future is to incorporate representations of chemical structures such as SMILES strings or Morgan fingerprints in the dataset. The dataset we used in this project has cell morphology features from Cell Profiler and gene expression data from L1000, but we did not use the molecular or chemical information regarding the perturbagens.

A second area of exploration is to engineer the L1000 assay into gene program expressions. Each gene program describes a group of genes and their collective function (e.g., the group GO:0005975 describes a group of genes responsible for carbohydrate metabolism). By adding this information, we hope that the model will incorporate information about functional networks of genes and how they result in changes in the cell environment, leading to a MoA.

Another future consideration is to use other model architectures for generating embeddings, which then can be used for MoA prediction. Potential model architectures include unsupervised contrastive learning or pretrained transformers. Also, other classifiers can be used for predicting MoA classes, such as multi-class logistic regression. Besides using different classifiers for MoA class prediction, another method that can be used to evaluate the model's performance is to train the model on another dataset for generalizability, such as the Bray dataset. However, this would require the model to address the effect of different cell types on MoA class prediction; LINCS uses the A549 cell line (lung cancer), while Bray uses the U2OS cell line (bone cancer).

Comparison with Original Proposal:

The final version of this project consists of a few alterations from the original proposal, in the modalities used and the extent of downstream analysis. Initially, the project aimed to incorporate L1000, raw Cell Painting images, and Cell Profiler derived features for each chemical perturbation to generate embeddings and predict MoA. However, the Cell Painting images were not used due to issues with size and accessibility, discovered after beginning data processing. Despite this, the integration of the kept modalities, generation of embeddings through representation learning architectures, and usage of embeddings for downstream MoA classification aligned with the specific aims stated in the proposal. The other difference was in the specific aims of interpreting modality specific performance and patterns at the GSEA and organelle morphology level by MoA class, which was not undertaken due to lack of embedding clustering by MoA and inconsistent class wise predictive power of the investigated models.

Though we were not able to accomplish all of the goals we set out to pursue in our original proposal, we were able to implement the pipeline of multimodal feature integration to generate embeddings via representation learning for downstream MoA prediction, which was our main goal for the project. This also means that we were able to address the concerns of our proposal reviews, which revolved around the integration of modalities and the use of representation learning to generate embeddings. We were able to address the former by pairing modalities by compound and dosage, and we addressed the latter using

two different strategies. However, we were not able to accomplish all of the specific aims for downstream analysis that we outlined in our initial proposal, a concern expressed by one reviewer, but we were only able to determine that we would be unable to do this after actually implementing and evaluating models.

Commentary on Experience:

The LINCS dataset from the Broad Institute was well-suited for the scope of this project, providing MoA and dose annotations for over a thousand unique chemical perturbations, along with multiple replicates of complete L1000 and Cell Profiler derived features for each compound dose combination. Although we initially aimed to also incorporate raw Cell Painting images as an additional modality, this proved not to be feasible as the data was on the scale of terabytes. Even with the available features being those necessary in terms of both content and scale for this project, the provided metadata and column names were not intuitive to navigate, requiring literature review. Integrating modalities was also challenging because sample level pairing was not consistent across the assays, and we attempted to address this through Cartesian products on compound dose pairs, and aggregating replicates using the median.

The most significant challenges arose during the model design and implementation stage, since the integration of L1000 and Cell Profiler modalities is a new area, meaning that there are no standard approaches for generating joint embeddings. Within the scope of this project, we were able to explore two embeddings models and three MoA classification strategies, but the performances fell short of expectations, failing to consistently outperform baselines from much simpler models published by the Broad. We were unable to adhere to the schedule we laid out in the proposal due to this, as well as because we previously had limited knowledge and understanding of the concepts with which we worked, both in the computational and biological feature aspects. We met and discussed the project regularly outside of class, but progress did not happen at the rate that we planned, and a potential approach to combat this would be to split up the model implementation so that different methods could each have been tuned and tested simultaneously by members, to more efficiently explore a larger solution space.

Despite these challenges, working on this project was highly rewarding, as the area of multimodal MoA characterization, and even the use of L1000 and Cell Profiler derived Cell Painting features, was new to all of us. Through literature review, data exploration, and model implementation, we appreciated the value that these high throughput assays provide, both in theory and in practice, when it comes to understanding the effects of chemical perturbations on cellular biology. It was also interesting to learn about the different computational methods researchers are developing to model MoA and characterize associated changes in and across the transcriptomic and phenomic levels, and we recognized first hand the difficulties of doing so. This helped us contextualize multimodal perturbation modeling, especially through conversation with Dr. Shantanu Singh of the Broad who is a recognized figure in this area.

Division of Labor:

For the methodological aspects of the project, Harshi defined the problem statement, significance, and approach, conducted the literature review, found and prepared the dataset, conducted exploratory data analysis, and investigated the biological meaning of the contained features. In addition, Lee curated and integrated the features for modeling, defined data splits, designed and implemented the two model architectures to generate embeddings for each perturbagen, and obtained preliminary results. Also, Vivian conducted dimensionality reduction on the generated embeddings, implemented the three classification strategies going from embeddings to MoA, and visualized class specific model performance. These splits represent the general division, but all of us met regularly to discuss and provide guidance to each other for all mentioned project areas, especially for modeling approaches and data interpretation.

For the writing and presentation aspects of the project, the work was split for the proposal, the mid-course and final report, and the final presentation, to reflect the areas each member worked on as described above. Overall, we collaborated effectively as a team and communicated regularly through weekly updates and discussions after lecture and recitation sessions. However, an area of improvement would be to divide tasks such that work could be done concurrently instead of sequentially, since the way we set about this project made it so that the work of each person could not begin until another finished theirs.

Though this could have made our process more efficient, the task allocations were equitable, and overall, we were satisfied with the way that our work in this project was divided and executed.

Code: https://github.com/harshi-saha/mlcb-project

References:

[1] Bray, MA., Singh, S., Han, H. et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc 11, 1757–1774 (2016). https://doi.org/10.1038/nprot.2016.105
[2] Ha, S.V., Jaensch, S., Kańduła, M.M. et al. Cross modality learning of cell painting and transcriptomics data improves mechanism of action clustering and bioactivity modelling. Sci Rep 15, 23010 (2025). https://doi.org/10.1038/s41598-025-05914-0
[3] Haghighi, M., Caicedo, J.C., Cimini, B.A. et al. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. Nat Methods 19, 1550–1557 (2022). https://doi.org/10.1038/s41592-022-01667-0
[4] McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, et al. (2018) CellProfiler 3.0: Next-generation image processing for biology. PLoS Biol 16(7): e2005970. https://doi.org/10.1371/journal.pbio.2005970
[5] Moshkov, N., Becker, T., Yang, K. et al. Predicting compound activity from phenotypic profiles and chemical structures. Nat Commun 14, 1967 (2023). https://doi.org/10.1038/s41467-023-37570-1
[6] Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., et al. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell, 171(6), 1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049
[7] Zhang, X., Yoon, J., Bansal, M. et al. (2023). Multimodal Representation Learning by Alternating Unimodal Adaptation. CVPR 2024. https://doi.org/10.48550/arXiv.2311.10707
[8] Bray MA, Gustafsdottir SM, Rohban MH, Singh S, et al. A dataset of images and morphological profiles of 30,000 small-molecule treatments using the Cell Painting assay. Gigascience. 2017 Dec 1;6(12):1-5. doi: 10.1093/gigascience/giw014
[9] Natoli, Ted; Way, Gregory; Lu, Xiaodong; Cimini, Beth; Logan, David; Karhohs, Kyle; et al. (2020). L1000 data for LINCS profiling complementarity analysis. figshare. Dataset. https://doi.org/10.6084/m9.figshare.13181966.v2
[10] Cimini, B.A., Chandrasekaran, S.N., Kost-Alimova, M. et al. Optimizing the Cell Painting assay for image-based profiling. Nat Protoc 18, 1981–2013 (2023). https://doi.org/10.1038/s41596-023-00840-9
[11] Cirinciani, M. et al. (2024). Drug mechanism: A bioinformatic update. Biochemistry and Biophysics Reports. Advance online publication.https://doi.org/10.1016/j.bcp.2024.116078
[12] Seal, S., Carreras-Puigvert, J., Trapotsi, MA. et al. Integrating cell morphology with gene expression and chemical structure to aid mitochondrial toxicity detection. Commun Biol 5, 858 (2022). https://doi.org/10.1038/s42003-022-03763-5