Project Report

# Exploring Perturbation-Induced Morphological Shifts for the Discovery of Novel Drug Candidates

*Student:*
Rahul Ohlan
Amitesh Badkul
{rohlan,
abadkul}@gradcenter.cuny.edu

*Artificial Intelligence:*
Prof. Jonathan Gryak

**Department of Computer Science**
December 16, 2024

# Contents

# 1 Project Abstract

Phenotypic data-driven drug discovery uses cellular morphology to explore the effects of chemical perturbations. In this work, we utilize chemically perturbed microscopy images obtained through the cell painting methodology on MCF7 cell lines to predict the morphological response of cells to novel chemical compounds. Drawing inspiration from Opean AI's text-to-image generation model DALL-E2 [1], we designed a pipeline for cellular image generation based on chemical input. Our approach begins with contrastive pretraining using an cell microscopy InfoLOOB [2]-based loss function to align chemical and images in a shared representation space. We further employ a 2D U-Net-diffusion-based decoder model to generate microscopy images from the latent representations learned by a CLIP [3] model. Using this framework, we generated cellular morphology images for unseen drugs based solely on their chemical structures.

Our findings demonstrate strong alignment between chemical and image modalities, as well as the ability of the model to produce realistic and biologically meaningful perturbation images for novel compounds. The generated images reflect key morphological changes associated with chemical perturbations, underscoring the potential of this approach for virtual phenotypic screening and drug discovery.

However, the current implementation has certain limitations:

- The study is restricted to a single cell line (MCF7) and does not consider variability across multiple cell lines, which may limit the generalizability of the findings.
- The dataset is relatively scarce, with only about 1400 unique chemical compounds available after filtering for specific time and dosage conditions.
- Time and dosage information are not explicitly encoded in the model, which may reduce its ability to capture temporal and dosage-dependent effects on cellular phenotypes.

Addressing these limitations in future work will enhance the robustness and applicability of this framework for broader drug discovery efforts. The code is available at https://github.com/rahul-ohlan/ai$_p$*roject*

# 2 Accomplishments

## 2.1 Major goals of the project

The primary goal of this project is to facilitate interpretable and biologically meaningful embeddings that bridge molecular structures and their corresponding cellular perturbation images. By creating a unified framework, this project aims to enable two key downstream applications:

- **Retrieval:** Efficiently retrieve perturbation images or chemical structures based on a query embedding, enabling comparative analysis and hypothesis generation.
- **Generation:** Generate accurate and biologically relevant perturbation images for novel compounds, aiding virtual phenotypic screening and drug discovery efforts.

This integrative approach seeks to advance our understanding of chemical-biological interactions and streamline the drug discovery process.

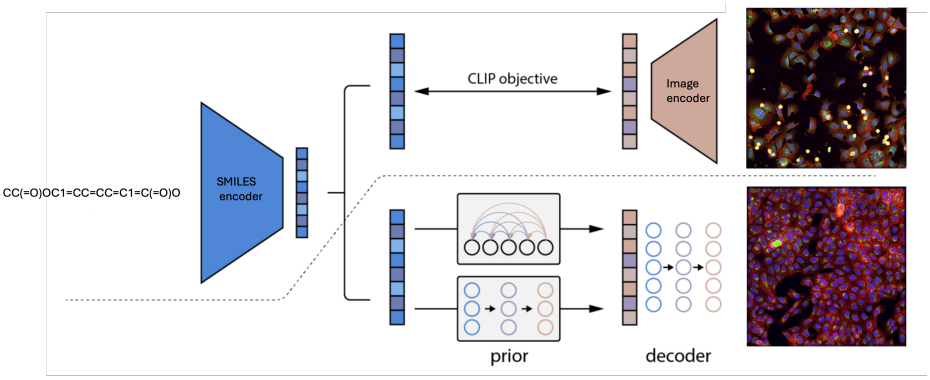



Figure 1: **A.**(Above the dotted line): Alignment of SMILES string representations and microscopy image embeddings in a shared latent space using contrastive learning. **B.** Generation of microscopy images through the image decoder, using SMILES embeddings as input and mapping them to image-CLIP embeddings via a prior model.

## 2.2 Specific Objectives

To achieve the major goals, this project focuses on the following specific objectives:

1. **Develop and Optimize a Shared Embedding Space:** Establish a shared embedding space that aligns microscopy images of cellular phenotypes with molecular structures. This alignment captures essential biological and chemical features, enabling interpretable embeddings without requiring activity labels. The hypothesis driving this aim is that such alignment will enhance retrieval and predictive capabilities for a wide range of biological responses.

2. **Evaluate Predictive Capabilities for Biological Activity and Image Generation:** Assess the utility of the learned representations in two key downstream tasks:

   - **Retrieval:** Predict and retrieve embeddings for biological outcomes, such as drug efficacy or toxicity, and find related perturbation images or molecular structures.
   - **Generation:** Generate microscopy images based on molecular data, enabling visualization of cellular morphological responses for novel compounds.

   This aim tests the hypothesis that multimodal embeddings can reveal biologically meaningful relationships and generate accurate, interpretable perturbation images in a label-free manner.

Our framework primarily involves training separate encoders for ecnoding chemical strings and microscopy image data, followed by a contrastive learning approach to align these embeddings in a shared space. Fig 1 illustrates the model architecture and pipeline.

## 2.3 Significant findings

The CLIP model was evaluated for its ability to align molecular and microscopy image embeddings. This section presents key findings supported by visualizations and retrieval metrics.

## 2.4 Cross-modal Alignment

The UMAP visualization in Figure 2b shows the clustering of molecular and microscopy image embeddings after training. The distinct but correlated

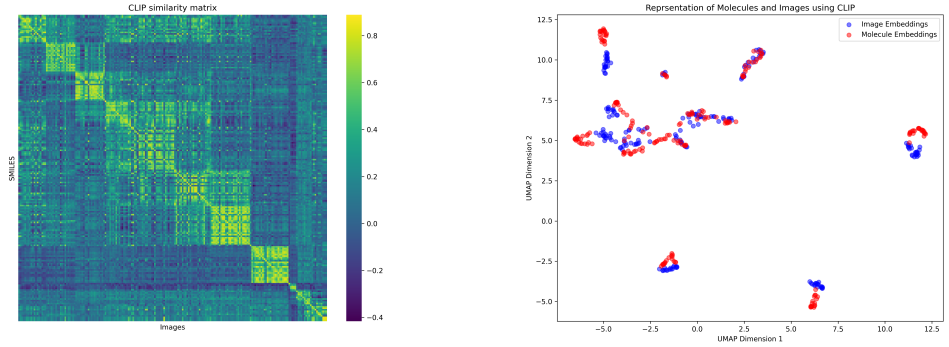
groupings indicate that the model effectively captures shared representations across modalities. Molecule embeddings (red) and image embeddings (blue) align well, demonstrating the model's ability to harmonize the two data modalities.

### 2.4.1 Retrieval Performance

Table 1 presents the text-to-image retrieval performance of the CLIP model.

- The retrieval accuracy at rank 1 (R@1) is 42.5%, indicating the model's precision in retrieving the correct image for a given molecular embedding.
- High R@5 (84.5%) and R@10 (96.5%) values reflect the model's strong performance over broader retrieval sets.

### 2.4.2 Similarity Distribution




(a) CLIP similarity matrix between molecular embeddings (SMILES) and microscopy image embeddings. A strong diagonal trend highlights high instance-level similarity.

(b) UMAP visualization of molecular and microscopy image embeddings after CLIP training. Molecule embeddings (red) and image embeddings (blue) cluster closely, indicating strong cross-modal alignment.

Figure 2: Comparison of CLIP similarity matrix and UMAP visualization, highlighting the cross-modal alignment between molecular and microscopy image embeddings.

The similarity matrix shown in Figure 2a provides further insights:

- A strong diagonal trend indicates high similarity between molecules and images from the same instances.
- Sub-block structures suggest clusters of molecular and image pairs sharing higher inter-modal similarities, potentially reflecting shared functional or structural properties.

These findings demonstrate the CLIP model's ability to learn meaningful relationships between molecular data and microscopy images, paving the way for applications in retrieval and integrated analysis of chemical and biological datasets.

Table 1: Text-to-Image Retrieval Performance of the CLIP Model

| Metric | R@1 | R@5 | R@10 |
|---|---|---|---|
| Value | 42.5% | 84.5% | 96.5% |

## 2.5 Generative Performance

The generative performance of the UNET-based latent diffusion model is evaluated by decoding the latent representations into the original microscopy images with 5 channels. The model was trained using a reconstruction objective, where the generated images aim to replicate the ground truth.

Although the latent diffusion framework is theoretically powerful, it is computationally intensive and very slow to train due to the complexity of noise schedules and the high-dimensional latent space. Given the time constraints and limited opportunities for optimization, the generated images display significant deviations from the ground truth, particularly in the clarity and fidelity of the features, ref Figure 3.

Despite these challenges, the results indicate the model's ability to approximate the structure of the ground truth images, albeit imperfectly. The artifacts in the generated images suggest that further tuning of the noise schedule, model architecture, and training strategies could significantly enhance the quality of the output in future work.

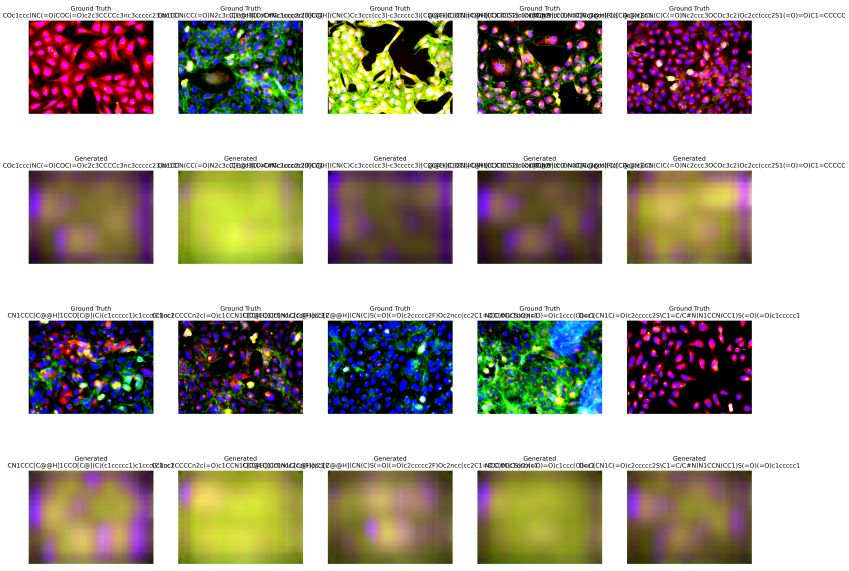


Figure 3: Comparison of ground truth images (top row of each pair) with the corresponding generated images (bottom row of each pair) produced by the UNET-based latent diffusion model. Each image has 5 channels, with color overlays representing distinct features.

# 3 Team Member Contributions

1. **Rahul Ohlan:**

   (a) Managed data acquisition by selecting and adapting datasets such as BBBC021 and LINCS L1000 to address project limitations.

   (b) Played a primary role in implementing the molecular encoder using Morgan fingerprints processed through an MLP architecture.

   (c) Conducted interpretability analyses with UMAP and t-SNE visualizations to uncover relationships between molecular and microscopy embeddings.

(d) Focused on enhancing the contrastive learning framework, optimizing it with InfoNCE and CLOOB loss functions to align molecular and microscopy embeddings in a shared latent space.

2. **Amitesh Badkul:**

(a) Led the development of the microscopy image encoder using the ResNet architecture to extract crucial morphological features for downstream tasks.

(b) Contributed significantly to troubleshooting and optimizing the CLIP-based framework to align chemical and microscopy embeddings, despite dataset challenges.

(c) Investigated challenges in contrastive learning and suggested alternative strategies like one-hot encoded dosage levels and dataset augmentation to improve alignment.

(d) Conducted preliminary testing of generative outputs and explored integrating a lightweight MLP-based prior model to enhance microscopy image generation.

3. **Collaborative Efforts:**

(a) Both team members collaborated on designing and refining the pipeline, drawing inspiration from DALL-E2 and CLOOME models for multimodal data alignment and image generation.

(b) Jointly addressed dataset limitations and model performance issues, iterating solutions to improve predictive and generative tasks.

(c) Shared responsibility in preparing and delivering the final presentation, emphasizing key findings, challenges, and future directions.

(d) Maintained comprehensive documentation and ensured the code base was clean, modular, and well-organized to facilitate reproducibility and future scalability.

# 4 Changes/Problems

## 4.1 Changes in approach

Our initial efforts began with the BBBC021 cell painting dataset, a widely used resource containing a large number of microscopy images corresponding

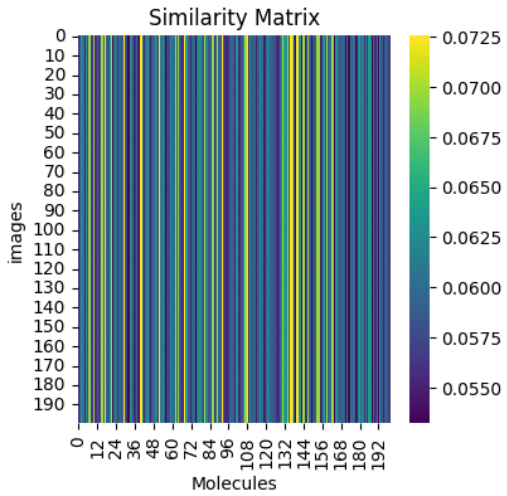



Figure 4: Similarity matrix between a randomly sampled set of molecules and corresponding microscopy images from the set after training CLIP

to 133 chemical compounds. This dataset has been extensively employed for tasks such as mechanism of action prediction. Given its diversity and relevance, it was an appealing choice for training a model using contrastive learning. However, during the early stages of experimentation, we encountered several challenges that necessitated substantial changes in our approach.

### 4.1.1 Challenges with Contrastive Learning

We first attempted to train a CLIP-like model using molecular embeddings derived from Morgan fingerprints of SMILES strings and microscopy image embeddings from the BBBC021 dataset. Despite rigorous training, the model failed to learn meaningful cross-modal representations (as shown in Figure 4). One major limitation of the dataset was the lack of representational diversity for molecules. The embeddings of molecular structures were highly similar across the dataset, which made it difficult for contrastive learning to effectively align chemical and image embeddings.

To address this issue, we incorporated additional contextual information by conditionally encoding one-hot representations of dosage levels along with the molecular embeddings. Our hypothesis was that incorporating dosage as a factor would enhance the embedding separation for different conditions. However, this adjustment did not improve the performance, as the

embeddings still overlapped heavily across different dosages, failing to capture meaningful differences in the data.

### 4.1.2 Exploring Self-Supervised Pretraining

In response to these challenges, we reviewed the literature for potential solutions and identified self-supervised learning as a promising strategy to generate more meaningful representations [4] of microscopy images. Specifically, we employed the SimCLR [5] framework for self-supervised pretraining. As shown in Figure 5, two random augmentations are applied to each image, and contrastive learning is used to pull embeddings of augmented versions of the same image closer, while pushing embeddings of distinct images apart.

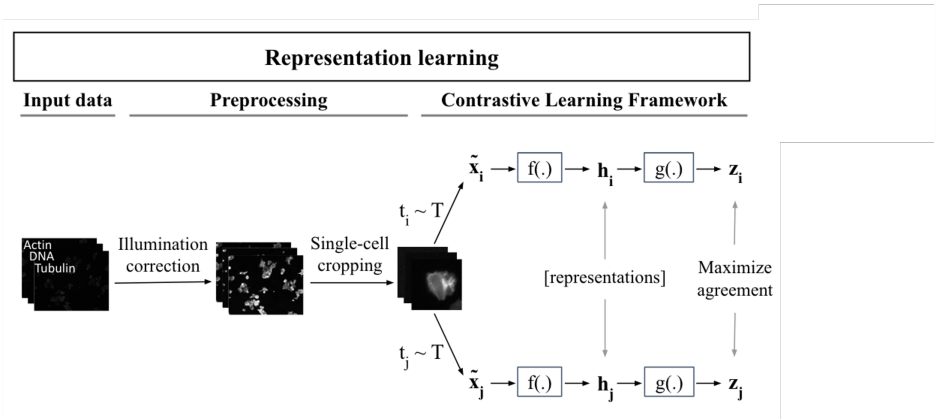



Figure 5: For representation learning [4], we feed single-cell images to the contrastive learning framework, wehre the single-cell representations are extracted.

Our hypothesis was that microscopy images treated with the same chemical at different concentrations should form separate clusters in the learned representation space, reflecting distinct perturbation effects. However, as seen in Figure 6, this approach also fell short of expectations. The embeddings of perturbation images corresponding to the same chemical overlapped significantly, indicating that the model could not distinguish between different concentrations or effects effectively. This overlap suggests that the SimCLR framework, despite its strengths, may not be sufficient to handle the complexity of chemical perturbation data in this context.

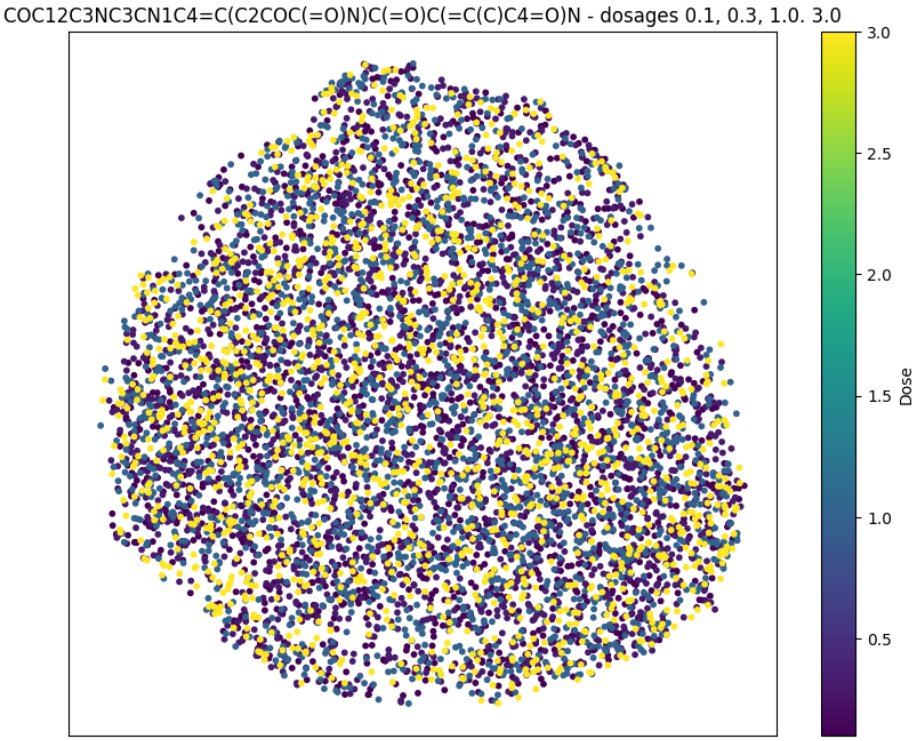



Figure 6: Visualization of perturbation embeddings after self-supervised pretraining using the SimCLR framework. Embeddings corresponding to perturbations of the same chemical overlap significantly, indicating poor separability of conditions.

### 4.1.3 Switching to a New Dataset

To overcome the limitations of the BBBC021 dataset and to stay within the timeline constraints, we switched to the LINCS L1000 Dataset, which contains paired chemical perturbations for over 30,000 chemicals. This dataset is well-suited for cross-modal learning as it provides both molecular and gene expression profiles for each chemical perturbation, offering greater diversity and richer representations compared to BBBC021.

Using a subset of this dataset, we trained our model on paired chemical perturbations and corresponding embeddings. This pivot allowed us to take advantage of the scale and diversity of the dataset, leading to more meaningful and distinguishable representations. Although the timeline lim-

ited further optimization, the use of this dataset significantly improved the alignment between molecular and gene expression data, marking a key step forward in our approach.

#### 4.1.4 Summary of Approach Evolution

The iterative changes in our approach underscore the complexity of the task and the limitations of existing methods when applied to chemical perturbation datasets. While initial efforts with BBBC021 faced challenges, the shift to the Giga science dataset [6] enabled us to train a more robust model within the given constraints. This microscopy dataset includes 919 265 five-channel fields of view, representing 30 616 tested compounds Moving forward, more advanced methods, such as hybrid frameworks combining supervised and self-supervised techniques or incorporating additional omics data, could be explored to address these challenges.

We successfully implemented the two major parts of our work: the CLIP-based framework and the prior-based generative model. The CLIP-based framework consisted of options for multiple loss contrastive loss functions such as infoNCE loss and CLOOB loss. The first part is meant for the initial alignment of chemical compounds and cell microscopy images, similar to all CLIP-based models. However, we didn't utilize transformer-based models (like ChemBERTa, ChemBERTa-2, and SMILES-BERT) to obtain chemical representations. Instead, we used morgan fingerprints as input to our chemical encoder. We realized this would add another computational cost to our project, and the existing state-of-the-art works in similar tasks utilize the morgan fingerprint as the chemical representation [7]. These morgan fingerprints are processed through a simple multi-layer perceptron (MLP). Our cell microscopy image encoder included a ResNet model; we didn't explore vision transformer-based models as they have provided similar performance compared to ResNet but at a higher computational cost [8]. Moreover, we obtained a plethora of negative results, which caused us to change our methodology.

We initially used the BBBC021 dataset, which contained 113 compounds; however, only 38 of those had experimentally validated phenotypic effects, and we had around 12,700 corresponding cell microscopy images for varying dosages for each chemical. Since CLIP-based and contrastive learning models typically require large amounts of training data to correctly align between the two modalities, we perform cropping of the cellular microscopy

images using the cell profiler software [9] to augment our training data to around 93,000 images. However, our CLIP model still wasn't able to learn the proper relations between the two modalities, possibly due to the one-to-many relationship between compounds and cell microscopy images. CLIP is designed for one-to-one image-text pairings, and this mismatch can hinder its ability to learn accurate associations. Providing multiple images for a single chemical can confuse the model, as it expects each image to correspond to a distinct chemical. This misalignment can impede the model's capacity to establish clear and accurate associations between modalities [10]. Another reason could be cropping eliminates the surrounding cellular environment, which is often crucial for understanding biological relationships. Microscopy images often rely on spatial context, such as the arrangement of cells or their interactions, to depict meaningful biological effects. Individual cropped cells may not capture the diversity of phenotypic changes observed in the original, full microscopy images. The lack of variability in cropped images might cause the model to overfit redundant features, hindering generalization. In our first step itself, we notice this fact, and to rectify this, we try out two methods: a) we perform one-hot encoding of the dosage information along with the chemical representations before passing through the chemical encoder, and b) we test out our model with the most biologically significant dosage 10 nM. CLIP frameworks are particularly sensitive to the quality and diversity of image-text pairs [11].

In both these approaches, there are several issues:

1. For CLIP to learn meaningful associations, the chemical representations need to distinctly correspond to the visual content, in this case, the cell microscopy images. Additional dosage information doesn't significantly alter the overall representation or is overshadowed by more prominent features, and that's why the model struggles to establish a clear link between the image and its corresponding chemical representations [12].

2. Filtering by dosage might inadvertently remove contextual information present in the original dataset. This loss can hinder the model's capacity to understand the nuanced relationships between chemical compounds and their effects, as depicted in microscopy images.

We can see these results through the uniform similarity matrix obtained for these variations. Inside the CLIP framework, we use two different types of loss infoNCE [13] and CLOOB [2].

The InfoNCE loss is formulated as:

$$L_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\text{sim}(x_i, y_i)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\text{sim}(x_i, y_j)/\tau\right)} \tag{1}$$

Where:

- $x_i$ and $y_i$: Represent the embeddings of positive pairs (e.g., image and its corresponding text/chemical embedding).
- $\text{sim}(x, y)$: Similarity function (e.g., cosine similarity) between embeddings $x$ and $y$.
- $\tau$: Temperature parameter controlling the sharpness of the distribution.
- $N$: Number of samples in the batch.

The CLOOB loss employs a leave-one-out mechanism and is defined as:

$$L_{\text{CLOOB}} = \frac{\tau}{N}\sum_{i=1}^{N}\left[-\text{sim}(x_i, y_i) + \log\sum_{j=1}^{N}\exp\left(\text{sim}(x_i, y_j)/\tau\right)\right] \tag{2}$$

Where:

- $\tau = \frac{1}{\text{inv\_tau}}$: Temperature parameter inverse.
- $\text{sim}(x_i, y_i)$: Similarity between positive pairs.
- $\log\sum_{j=1}^{N}$: Includes all terms except the diagonal (leave-one-out mechanism).

One of the key related works, CLOOME, focuses on training a model for chemical retrieval based on perturbed microscopy images. Their approach leverages the Cell Painting dataset, which contains 30k compounds and 284,035 six-channel images, amounting to approximately 300GB of data. However, training on such a large-scale dataset was infeasible for us due to significant memory limitations, time constraints, and the challenges posed by our dataset.

Our dataset had its own inherent limitations:

1. Limited size: It contained far fewer compounds and microscopy images compared to CLOOME.
2. One-to-many mapping: Multiple images corresponded to a single chemical due to varying dosages, complicating the contrastive learning process.
3. Insufficient training data: This posed a critical bottleneck for achieving meaningful contrastive alignment.

By the time we realized the impact of these constraints, time was a limiting factor. Consequently, we shifted our focus toward completing the second part of the pipeline—the generative framework. Unlike CLOOME, which does not explore generative modeling, we aimed to demonstrate the feasibility of enhancing microscopy image generation from chemical perturbations. Specifically, we utilized CLOOME's pre-trained CLIP model encoders and decoders within our generative framework to highlight the potential of a simple MLP-based prior model in improving the generated microscopy images.

Due to time constraints, we were unable to implement the prior model, an additional innovative step in the DALL-E 2 architecture. The authors of DALL-E 2 claim that incorporating a prior enhances the diversity of generated images, which would be a valuable extension of our work. As part of future efforts, implementing a prior model could further improve our model's performance and sophistication in generating diverse molecular and microscopy images.

# 5 Impact

## 5.1 Impact on the domain

Our research introduces a novel adaptation of the DALL-E 2 [1] framework to the domain of molecular and microscopy-based image generation, bridging the gap between gene expression data and visual representations of biological phenomena. By aligning gene expression profiles with molecular embeddings and leveraging advanced multimodal techniques, this work enables the exploration of complex relationships between genotypic and phenotypic data. The proposed architecture offers transformative potential in drug discovery and precision medicine by facilitating data-driven insights into molecular interactions and biological systems. Additionally, our approach lays a foundation

for integrating multimodal deep learning into computational biology, advancing the capabilities of generative modeling for high-dimensional biomedical data. This work not only enriches the toolkit available to researchers but also addresses the increasing demand for interpretable, scalable models in biosciences.

## 5.2 Impact on the individual

**Amitesh Badkul:** The research has been a learning process, helping me better understand how chemical perturbations influence cellular states, one of the most critical concepts in drug discovery. The cell exposed to small molecules or drugs causes chemical perturbation and changes in morphology, signaling pathways, and cellular function. Such changes provide insights into efficacy, toxicity, and mechanisms of action of compounds. While my previous focus has been on chemical-protein interactions, this project thus allowed me to explore how a molecular structure influences cellular systems and fills an important gap in my knowledge. Implementing the CLIP framework to align molecular data with cellular microscopy features. By working on this, I have learned more about techniques using multimodal learning. I also gained practical experience in aligning such complex data types by optimizing contrastive loss functions such as InfoNCE and CLOOB and surmounting challenges that may be intrinsically different from other biological systems. In summary, this taught me how state-of-the-art AI methods can be modified for specific problems at the frontier of computational biology. The most important lesson from this project was how difficult it is to translate machine learning techniques from other fields into biology-based drug discovery. Unlike fields such as computer vision or natural language processing, biology presents complex data and necessitate models that have interpretable results. This project reinforced the value of integrating molecular and cellular-level data to build interpretable models for drug discovery. It aligned well with the goals of Prof. Xie's lab on accelerating drug discovery using computational methods. Overall, this experience expanded my skill set, from and has prepared me to contribute meaningfully to future work in this field.

# References

[1] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[2] A. Fürst, E. Rumetshofer, J. Lehner, V. T. Tran, F. Tang, H. Ramsauer, D. Kreil, M. Kopp, G. Klambauer, A. Bitto *et al.*, "Cloob: Modern hopfield networks with infoloob outperform clip," *Advances in neural information processing systems*, vol. 35, pp. 20 450–20 468, 2022.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[4] A. Perakis, A. Gorji, S. Jain, K. Chaitanya, S. Rizza, and E. Konukoglu, "Contrastive learning of single-cell phenotypic representations for treatment classification," in *Machine Learning in Medical Imaging: 12th International Workshop, MLMI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 12*. Springer, 2021, pp. 565–575.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[6] M.-A. Bray, S. M. Gustafsdottir, M. H. Rohban, S. Singh, V. Ljosa, K. L. Sokolnicki, J. A. Bittker, N. E. Bodycombe, V. Dančík, T. P. Hasaka *et al.*, "A dataset of images and morphological profiles of 30 000 small-molecule treatments using the cell painting assay," *Gigascience*, vol. 6, no. 12, p. giw014, 2017.

[7] A. Sanchez-Fernandez, E. Rumetshofer, S. Hochreiter, and G. Klambauer, "Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures," *Nature Communications*, vol. 14, no. 1, p. 7339, 2023.

[8] A. Palma, F. J. Theis, and M. Lotfollahi, "Predicting cell morphological responses to perturbations using generative modeling," *bioRxiv*, pp. 2023–07, 2023.

[9] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat *et al.*, "Cellprofiler: image analysis software for identifying and quantifying cell phenotypes," *Genome biology*, vol. 7, pp. 1–11, 2006.

[10] A. Bulat, Y. Ouali, and G. Tzimiropoulos, "Fff: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 172–14 182.

[11] C. Gong, D. Wang, M. Li, V. Chandra, and Q. Liu, "Keepaugment: A simple information-preserving data augmentation approach," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1055–1064.

[12] S. Kornblith, L. Li, Z. Wang, and T. Nguyen, "Guiding image captioning models toward more specific captions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 259–15 269.

[13] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.