# CellCLIP - Learning Perturbation Effects in Cell Painting via Text-Guided Contrastive Learning

**Mingyu Lu**[*], **Ethan Weinberger**[*], **Su-In Lee**
Paul G. Allen School of Computer Science & Engineering
University of Washington
{mingyulu,ewein,suinlee}@cs.washington.edu

## Abstract

High-content screening (HCS) assays based on high-throughput microscopy techniques such as Cell Painting have enabled the interrogation of cells' morphological responses to perturbations at an unprecedented scale. The collection of such data promises to facilitate a better understanding of the relationships between different perturbations and their effects on cellular state. Towards achieving this goal, recent advances in multimodal contrastive learning could, in theory, be leveraged to learn a unified latent space that aligns perturbations with their corresponding morphological effects. However, the application of such methods to HCS data is not straightforward due to substantial technical artifacts and the difficulty of representing different classes of perturbations (e.g. small molecule vs CRISPR gene knockout) in a single latent space. In response to these challenges, here we introduce CellCLIP, a multi-modal contrastive learning framework for HCS data. CellCLIP leverages pre-trained image encoders coupled with a novel channel encoding scheme to better capture relationships between different microscopy channels in image embeddings, along with natural language encoders for representing perturbations. Our framework outperforms current open-source models, demonstrating the best performance in both profile-to-perturbation and perturbation-to-profile retrieval tasks while also achieving significant reductions in computation time. Code for our reproducing our experiments is available at `www.placeholder.com`.

## 1 Introduction

A grand challenge in cellular biology is understanding the impacts of different perturbations, such as exposure to chemical compounds or gene knockouts, on cellular function. In pursuit of this goal, a number of high-content screening (HCS) assays have been developed that combine image-based deep phenotyping with high-throughput perturbations (Gu et al., 2024; Kudo et al., 2024). For example, the Cell Painting assay (Bray et al., 2016) has been leveraged to profile cells' response to chemical and genetic perturbations (Sivanandan et al., 2023; Ramezani et al., 2025).

Despite the promise of this data, extracting meaningful quantitative representations of cellular states from morphological profiles presents a formidable challenge. Traditional analyses of image-based cellular profiles extracted sets of domain-expert-crafted morphological features implemented in tools like CellProfiler (Caicedo et al., 2017; Carpenter et al., 2006). More recent works have found that self-supervised deep learning methods based on DINO models (Caron et al., 2021; Oquab et al., 2023) or masked autoencoders (MAEs; He et al. (2022)) can capture more subtle changes in cellular morphology, resulting in representations that better agree with known biology.

In order to more explicitly capture the relationships between perturbations' effects on cellular state, a recent line of work has applied multi-modal contrastive learning techniques (Radford et al., 2021) to learn unified representations of both perturbation labels and imaging readouts for Cell Painting data. In particular, recent work has leveraged contrastive learning techniques by treating images and their corresponding perturbation labels as paired samples from different modalities (Fradkin et al., 2024;

[*] Equal contribution.

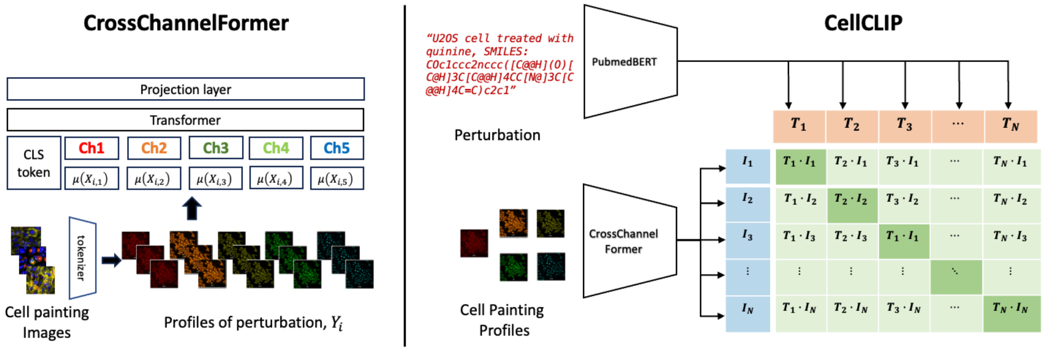

Figure 1: Overview of our proposed framework: CrossChannelFormer (left) processes channel embeddings; CellCLIP (right) employs contrastive learning with *(profile, perturbation)* pairs.

Sanchez-Fernandez et al., 2023). By leveraging the learned representations of perturbations, such methods enable the systematic identification of perturbations with similar morphological effects. Moreover, post-training, these models' perturbation encoders can be queried for novel perturbations not present in the training data to predict the effects of new perturbations.

Despite their promise, previous contrastive learning methods for Cell Painting come with significant drawbacks. For example, microscopy image data exhibits substantial differences from natural images, such as a variable number of channels and less information shared between different channels compared to standard RGB images; however, existing methods either fail to account for these differences (Sanchez-Fernandez et al., 2023) or are not openly available to the community (Fradkin et al., 2024). Moreover, previous works have exclusively focused on chemical perturbations, relying on graph representation learning techniques applied to chemical structures to learn perturbation representations. Thus, these methods cannot be applied to other classes of perturbations (e.g. CRISPR-mediated gene knockouts), and it is unclear how to represent perturbations from different classes in the same input space for contrastive learning.

To address these challenges, we propose CellCLIP (Figure 1), a contrastive learning framework designed to account for the unique challenges in Cell Painting perturbation data. Our framework employs off-the-shelf pretrained vision models combined with a novel channel-encoding scheme to account for different information in Cell Painting channels when embedding cellular images, and leverages natural language encoders for representing cells' corresponding perturbations. By using natural language to encode perturbations, CellCLIP can be readily applied to arbitrary classes of perturbations. Moreover, by using publicly available pretrained vision and natural language encoders, CellCLIP can be easily applied to new datasets without requiring costly training of image or perturbation encoders from scratch.

We benchmark CellCLIP on profile-perturbation retrieval for unseen compounds (Bray et al., 2016) and its ability to recover biological relationships via perturbation matching (Chandrasekaran et al., 2024). Our results show that CellCLIP achieves strong performance while using readily available off-the-shelf components. Our method also shows promising results in cross-perturbation class matching, providing a scalable and effective solution for biological discovery.

## 2 Background

**Representation Learning for Cell Painting** Self-supervised learning (SSL) deep learning techniques have been successfully applied to experimental microscopy data, with recent studies demonstrating their ability to capture intricate details of cellular morphology (Sivanandan et al., 2023; Doron et al., 2023) better than hand-crafted features as implemented in CellProfiler (Carpenter et al., 2006). Initial applications of SSL methods to Cell Painting reiled on architectures designed for natural RGB images, where information across channels exhibits strong correlations. On the other hand, Cell Painting channels each capture distinct biological structures (e.g. actin via phalloidin, mitochondria via MitoProbe etc.). To account for this, recent works (Bao et al., 2023; Kraus et al.,

2024) have proposed so-called channel-agnostic vision transformers (CA-ViTs) that use separate tokens for each channel in a spatial patch rather than aggregating information across all channels. However, CA-ViTs significantly increase computational costs due to increased number of tokens[1].

**Multi-modal Contrastive Learning** Multi-modal contrastive learning methods such as CLIP (Radford et al., 2021) learn joint latent representations across modalities (e.g. text and image). Such methods optimize a symmetric contrastive loss to maximize the similarity between correct pairs while minimizing it for incorrect ones. Specifically, given a batch of $N$ image-text pairs $\{(I_i, T_i)\}_{i=1}^{N}$, let $f(I)$ and $g(T)$ be the normalized embeddings of an image and text, respectively. Their similarity is defined as, $s_{ij} = \frac{f(I_i)\cdot g(T_j)}{\tau}$ where $\tau$ is a learnable temperature parameter. The CLIP loss consists of two cross-entropy objectives for image-to-text and text-to-image alignment:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{N}\sum_{i=1}^{N}\left[-\log\frac{\exp(s_{ii})}{\sum_{j=1}^{N}\exp(s_{ij})} - \log\frac{\exp(s_{ii})}{\sum_{j=1}^{N}\exp(s_{ji})}\right] \tag{1}$$

While contrastive learning excels in cross-modality alignment, several challenges remain in its application to Cell Painting perturbation data. First, while a thorough body of work exists studying encoding (tokenization) strategies for natural images and natural language (Sennrich, 2015; Wu, 2016; Alexey, 2020), encoding strategies specifically designed for cellular perturbations and Cell Painting images remain largely underexplored. Second, perturbations in Cell Painting experiments can span a variety of classes, including chemical perturbations, gene knockouts, and open reading frames (ORFs). Yet, existing contrastive learning approaches for Cell Painting, such as CLOOME (Sanchez-Fernandez et al., 2023) and MolPhenix (Fradkin et al., 2024) have exclusively focused on chemical peturbations, limiting their applicability. Lastly, model weights for some recent state of the art methods (e.g. (Fradkin et al., 2024)) are not openly available, and training these models from scratch is prohibitively expensive, requiring large datasets and significant GPU resources.

## 3 Method

In this section we introduce CellCLIP, our multi-modal contrastive learning framework for learning a unified latent space of perturbations and corresponding Cell Painting images. We begin by describing our strategy for encoding Cell Painting images in a manner suitable for applying the contrastive loss with perturbation labels (Section 3.1); due to substantial technical effects in optical perturbation screens, this encoding must be done with care, and we cannot naively reuse strategies from natural images. We then proceed to describe our strategy for encoding perturbations (Section 3.2), and for training the model (Section 3.3).

### 3.1 Cell Painting Image Encoding

**Image Profiles.** Recent works have developed image foundation models trained on natural images, such as DINOv2 (Oquab et al., 2023), which have demonstrated strong capabilities in capturing global structural image features. Rather than training new models from scratch, we choose to adopt these pretrained models as image encoders in the CellCLIP framework. However, unlike natural images with the standard set of RGB channels, Cell Painting images contain a variable number of channels corresponding to the specific stains used in an experiment. To work around this difference and enable models trained on natural images to be applied to Cell Painting data, we treat each Cell Painting channel as an independent grayscale image and extract embeddings separately. Formally, for each perturbation $i$ we denote the collection of Cell Painting images corresponding to that perturbation as $X_i = \{x_k\}_{k=1}^{N_i}$, where $x_k \in \mathbb{R}^{C\times H\times W}$ denotes an individual image. For each image $x_k$, we may apply a feature extractor $\phi_\theta$ that maps individual channels $x_k^c$ to an embedding, $z_k^c = \phi_\theta(x_k^c) \in \mathbb{R}^m$. This produces a channel-wise embedding matrix,

$$z_k = [z_k^1, z_k^2, \ldots, z_k^C]^T \in \mathbb{R}^{C\times m}. \tag{2}$$

As in previous work (Caicedo et al., 2018), we refer to this matrix as the *profile* of a Cell Painting image.

[1] For an image of size $h \times w$ with $c$ channels, CA-ViTs produce $\frac{h\times w}{p^2} \times c$ tokens assuming a patch size $p$.

**Mean Perturbation Profiles.** The standard CLIP model aligns pairs of natural images and corresponding text annotations. However, attempting to align individual Cell Painting image profiles and corresponding perturbation annotations as done by CLOOME (Sanchez-Fernandez et al., 2023) may produce subpar results. Beyond standard pathologies with the contrastive loss, such as false negative pairs, Cell Painting perturbation screens have a number of technical issues that may impede successful model training. For example, due to variable guide efficiency, cells with guide RNA barcodes corresponding to a specific gene knockdown may not have truly undergone the corresponding perturbation (Papalexi et al., 2021; Weinberger et al., 2024). Similarly, cells exposed to chemical perturbations may have highly varied responses (or no response at all) due to their position in a microscope well rather than properties of the chemical (Wang et al., 2023).

To work around these issues, similar to Fradkin et al. (2024), we instead choose to align each perturbation with an aggregated summary of all images labeled with the perturbation. In particular, we compute the mean profile of a perturbation $i$:

$$\mu(X_i) = \frac{1}{N_i} \sum_{k=1}^{N_i} z_k, \tag{3}$$

where $z_k$ was defined in Equation (2). By aggregating information across cells receiving the same perturbation, we may mitigate noise in individual cells' responses and facilitate more stable training.

**CrossChannelFormer.** Beyond a different number of channels, the relationships between Cell Painting image channels exhibit substantial differences compared to natural image channels. In particular, while natural image channels share a significant amount of information, Cell Painting image channels correspond to stains that each highlight distinct, semantically independent aspects of cellular morphology. Thus, to effectively learn meaningful embeddings of Cell Painting images, it is necessary to explicitly reason between information in different channels (Bao et al., 2023).

To accomplish this task while minimizing computational costs, we introduce CrossChannelFormer, a specialized encoder for CellCLIP (Figure 1). Unlike the standard Vision Transformer (ViT; (Alexey, 2020)), where each input token represents a multi-channel image patch, CrossChannelFormer takes as input mean profiles (Equation (3)) that encode the *global* cellular features associated with a specific stain. Following Bao et al. (2023), we then introduce a set of learnable channel embeddings, $[\texttt{chn}^1, \ldots, \texttt{chn}^C]$, where each $\texttt{chn}^c \in \mathbb{R}^d$ encodes information unique to its respective channel. We then prepend a learnable classifier token $\texttt{cls} \in \mathbb{R}^d$ to the sequence, which aggregates global image features across all channels. The resulting input sequence to the transformer is:

$$[\texttt{cls}, \mu(X_i)^1 + \texttt{chn}^1, \mu(X_i)^2 + \texttt{chn}^2, \ldots, \mu(X_i)^C + \texttt{chn}^C], \tag{4}$$

where $\mu(X_i)^c$ corresponds to the $c$th channel of $\mu(X_i)$. Following the original ViT, we feed the above sequence into a Transformer encoder. The Transformer encoder consists of alternating layers of multi-head self-attention and MLP blocks, with layer normalization applied before each block and residual connections established after each block. The final layer representation of the CLS token serves as the projection in the latent space.

Altogether, our proposed framework provides two major advantages over previous encoding schemes for Cell Painting images. First, our approach allows us to reuse off-the-shelf vision encoders pretrained on natural image data, which are far more plentiful than specialized models pretrained on Cell Painting images. Second, our CrossChannelFormer method allows us to capture the relationships between information in different CellPainting channels with only $C + 1$ tokens, which is far more computationally efficient compared to previously proposed CA-ViT methods.

### 3.2 Perturbation Encoding

Previous contrastive learning methods for Cell Painting rely on perturbation-class-specific encoders to represent perturbation treatments. For instance, CLOOME (Sanchez-Fernandez et al., 2023) encodes chemical compounds by passing Morgan fingerprints (Morgan, 1965) through a simple multilayer perceptron (MLP). This setup is not ideal, as different perturbation types (e.g. chemical compound vs gene knockouts) require distinct encoder networks, making it challenging to incorporate data from multiple perturbation types into the contrastive learning process.

To address this, we adopt a simple approach—representing each perturbation using *text*. Since most perturbations and their associated metadata can be effectively captured through textual descriptions, text serves as an efficient intermediate modality for generalizing across multiple perturbation types. We construct a corresponding text prompt $t \in \mathbb{R}^{d_t}$ that encodes information on cell types and perturbation-specific details. For example, to encode the chemical compound ethotoin, an anticonvulsant used in the treatment of epilepsy, we use the prompt:

*"A cell painting image of U2OS cells treated with ethotoin,*

*with SMILES string: CCN1C(=O)NC(C1=O)C1=CC=CC=C1."*

Similarily, for a CRISPR perturbation, the prompt is structured as:

*"A cell painting image of U2OS cells treated with CRISPR, targeting genes: AP2S1."*

By representing perturbations as text prompts, our approach facilitates encoding arbitrary perturbations from different classes, simplifying training across diverse perturbation types. It can also potentially integrate relevant textual metadata, enhancing perturbation retrieval across experiments.

For the text encoder, we utilize a pretrained BiomedicalCLIP (Zhang et al., 2023) text encoder which is adapted from a domain-specific language model PubMedBERT (Gu et al., 2021).

### 3.3 CellCLIP Training

For CellCLIP training, we adopt the contrastive loss to align profile and text embeddings of the same perturbation while separating those of different perturbations. Given a batch of $N$ paired mean Cell Painting profile embeddings and their corresponding perturbation (text) inputs, $(z_i, t_i)$, we compute their embeddings using a profile encoder $f(\cdot)$ and a perturbation encoder $g(\cdot)$, respectively. The similarity score between a profile $z_i$ and a perturbation text, $t_j$ is then defined as, $s_{i,j} = \frac{f(z_i) \cdot g(t_j)}{\tau}$. We then apply the loss calculation as described in Equation (1).

## 4 Experiment Setup

In this section, we describe the datasets and tasks used to evaluate our framework.

### 4.1 Perturbation-Profile Retrieval

We first benchmarked CellCLIP's performance by assessing its ability to retrieve test set perturbations given corresponding Cell Painting image profiles treated with each perturbation. That is, for a given model we compute perturbation projections along with corresponding mean Cell Painting profile projections in the shared latent space. Given the mean Cell Painting profile embeddings, we then compute cosine similarities with all perturbations' embeddings in the test set and retrieve the top-$k$ most similar perturbations; ideally, the image profile's true perturbation should be contained in this nearest neighbors set. Our evaluation metric, Recall@$k$ (R@$k$), measures whether the correct perturbation appears in the top-$k$ retrieved results, with $k = 1, 5, 10$. We denote this task as *profile-to-perturbation* retrieval

Swapping the roles of perturbations and Cell Painting profiles, we may similarly evaluate *perturbation-to-profile* retrieval, where, given a perturbation embedding, we compute similarities with mean Cell Painting profile embeddings. For this task we again use Recall@$k$ for evaluation.

### 4.2 Perturbation Detection and Matching

To further assess CellCLIP's ability to identify meaningful biological relationships, we also evaluated its image embeddings via the following tasks defined by Chandrasekaran et al. (2024):

- **Perturbation detection** assesses **replicability**, measuring how well replicates across batches of a given perturbation can be distinguished from negative controls.
- **Perturbation matching** evaluates **biological relevance** by identifying perturbations that target the same genes that would induce similar cellular morphological changes. This includes comparisons within the same perturbation class (e.g., compound-compound) and across different classes (e.g., CRISPR-compound).

Following Chandrasekaran et al. (2024) and Kalinin et al. (2024), we use *average precision (AP)* as our primary evaluation metric, defined as:

$$AP = \sum_{k=1}^{n} (R_k - R_{k-1}) P_k \tag{5}$$

where $P_k$ and $R_k$ denote the precision and recall at rank $k$, respectively, based on cosine similarity between query profiles. The ranking is determined by sorting profiles in descending order of cosine similarity to the query. To assess statistical significance, we perform permutation testing by shuffling rankings 100,000 times to construct a null distribution. We then apply multiple comparison corrections (Benjamini & Hochberg, 1995) and filter out non-significant AP values. For each perturbation replicate in different batches, we compute AP scores, which are then averaged to obtain a *mean average precision (mAP)* score representing the perturbation's phenotypic activity. Finally, we use mAP across classes, defined by specific perturbations or gene associations, to evaluate the performance in both tasks.

### 4.3 Datasets

For **retrieval** tasks, following Sanchez-Fernandez et al. (2023), we utilize the dataset from Bray et al. (2016). This dataset consists of 919,874 five-channel Cell Painting images corresponding to 30,616 small-molecule perturbations. We partitioned the dataset into train, validation, and test sets with a 70/10/20 split, resulting in 2115 unseen small molecules in the test set.

For **perturbation detection & matching**, we employ CPJUMP1 (Chandrasekaran et al., 2024), which features 186,925 nine-channel microscopy images.These images include three bright-field channels in addition to six Cell Painting dye channels. They are perturbed across 650 distinct perturbations, including compounds and genetic modifications such as CRISPR and ORF interventions. For each perturbation class, we applied a 70/10/20 split for training, validation, and testing.

Further details about datasets and preprocessing can be found in Appendix A.

### 4.4 Implementation details

**Profile & Perturbation Encoding** We utilized a series of pre-trained models (Appendix B) to generate profiles following the encoding strategy described in Section 3.1. For generating text prompts, we adopted the template: *"A {cell type} treated with {perturbation}, with {detailed perturbation information, such as SMILE or target genes}".*

**CellCLIP Backbone** Our CrossChannelFormer backbone consists of a transformer model with 12 layers, 8 attention heads, and a 512-dimensional embedding space. For the text encoder, we employ the pre-trained PubMedBERT from BiomedCLIP (Zhang et al., 2023) from Microsoft Research.

**Model Training** For retrieval evaluation on Bray et al. (2016), CellCLIP was trained with 50 epochs using a batch size of 768 and an AdamW optimizer. The learning rate was set at $2e^{-4}$ with cosine annealing and restart. The temperature parameter, $\tau$, is initialized from the pretrained BiomedicalCLIP. For perturbation detection and matching in CPJUMP1 (Chandrasekaran et al., 2024), given the limited number of unique perturbations, we reused our model trained on the Bray et al. (2016) dataset and fine-tuned it for another 50 epochs, using the same parameter settings as during their initial training. More details about training CellCLIP and other baselines can be found in Appendix B.

## 5 Results

### 5.1 Each component of CellCLIP improves alignment

We began by assessing the effectiveness of CellCLIP's approaches for encoding perturbations and mean Cell Painting profiles by evaluating their impact on retrieval performance for unseen molecules. In particular, starting with CLOOME's proposed encoding scheme, where individual images encoded using ResNet50 are aligned with chemical perturbations encoded using Morgan

| Vision Encoder | Perturb. Encoder | Train time (hr) | Chem-to-Profile (%) | | | Profile-to-Chem (%) | | |
|---|---|---|---|---|---|---|---|---|
| | | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| ResNet50 | Morgan Fingerprint + MLP | 48.3 | 0.56 | 1.37 | 2.31 | 0.23 | 0.91 | 1.41 |
| ResNet50 | PubMedBERT | 48.3 | 0.56 | 2.11 | 3.10 | 0.24 | 1.96 | 3.84 |
| CrossChannelFormer | PubMedBERT | 10.2 | 0.56 | 2.26 | 3.44 | 0.56 | 2.26 | 4.34 |
| CrossChannelFormer (M) | PubMedBERT | **1.4** | **1.08** | **3.68** | **6.01** | **1.05** | **3.78** | **5.76** |

M: Mean perturbation profiles used for alignment rather than individual images.

Table 1: Retrieval performance of different vision and perturbation encoder combinations for Chem-to-Profile and Profile-to-Chem tasks on 2,115 unseen small molecules. We evaluate retrieval using mean-pooled Cell Painting image embeddings, reporting recall at rank 1, 5, and 10. The channel profiles for CellCLIP training are generated from DINOv2 as described in Equation (2).

fingerprints combined with an MLP, we gradually replaced each of CLOOME's components with those of CellCLIP and assessed each change's impact on model performance (Table 1).

We first found that replacing the Morgan fingerprint encoder with our proposed natural language approach significantly improved retrieval performance, suggesting that perturbation information can indeed be effectively captured through textual descriptions. We next replaced CLOOME's ResNet50 encoder from with our CrossChannelFormer method for capturing relationships across different Cell Painting channels, and we found that this change again leads to a substantial improvement in retrieval performance. This results illustrates the importance of explicitly accounting for the relationships between different channels when learning embeddings of Cell Painting images. Notably, beyond providing an increase in performance, this change also resulted in substantially reduced training time, with 4.8 times faster training compared to the original CLOOME model.

Finally, we considered the impact of using mean Cell Painting profiles for each perturbation during training rather than attempting to align individual Cell Painting images. We found that this change led to yet another jump in performance while further reducing training time.

Altogether, these results demonstrate that each component of our CellCLIP framework contribute to significantly stronger retrieval performance.

### 5.2 Effects of Different Image Profile Encoding

CellCLIP provides a flexible framework for integrating off-the-shelf pretrained image foundation models into Cell Painting analyses (Section 3.2). To understand the impact of different vision encoding backbones on CellCLIP's performance, we conducted an ablation study where we varied CellCLIP's image encoder while holding all other aspects of our framework constant. Specifically, for this experiment we considered DINOv1 along with DINOv2 models of varying sizes. Aligning with results for natural images, we found that increases in model size broadly led to increased performance on our retrieval tasks (Table 2).

To understand the impact of using image encoder backbones originally trained on natural images versus those trained directly on Cell Painting data, we also applied OpenPhenom-S/16, an openly available masked autoencoder model pretrained on Cell Painting data (Kraus et al., 2024). Interestingly, we found that using OpenPhenom-S/16 did not result in superior performance compared to DINO models trained on natural images. This suggests that, despite not being originally trained on microscopy data, foundation models trained on diverse natural image distributions combined with small tweaks to account for differences in channels as in CellCLIP can achieve competitive performance on Cell Painting data.

### 5.3 Perturbation Detection & Matching through CellCLIP

Finally, we assessed the replicability and biological relevance (Section 4.3) of learned Cell Painting image embeddings on unseen perturbations in JUMPCP. For this evaluation, only image embeddings (and not perturbation embeddings) are required. Thus, here we not only compared between multimodal contrastive learning methods, but also assessed CellCLIP's performance compared to unimodal self-supervised learning (SSL) and weakly supervised learning (WSL) methods trained to predict perturbation labels. As an additional baseline, we also considered CellProfiler features.

| Image Encoding Backbone | Chem-to-Profile Retrieval (%) | | | Profile-to-Chem Retrieval (%) | | |
|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| DINOv1 | 0.75 | 2.22 | 3.78 | 0.66 | 1.90 | 3.73 |
| DINOv2 (small) | 1.18 | 3.35 | 5.05 | 0.99 | 3.16 | 4.82 |
| DINOv2 (base) | 1.11 | 3.63 | 5.48 | 1.04 | 3.35 | 5.24 |
| DINOv2 (large) | 1.08 | 3.61 | 5.67 | 0.99 | 3.45 | 5.50 |
| **DINOv2 (giant)** | **1.08** | **3.68** | **6.01** | **1.05** | **3.78** | **5.76** |
| CA-MAE (OpenPhenom-S/16) | 0.99 | 3.64 | 5.29 | 0.75 | 3.35 | 5.39 |

Table 2: Retrieval performance of CellCLIP trained with profiles generated from various pretrained imaging models on Chem-to-Profile and Profile-to-Chem tasks. Results are reported as Recall@$k$ (%) for $k = 1, 5, 10$.

| Method | Detection | Matching | |
|---|---|---|---|
| | | Within Perturb. | Across Perturb. |
| CellProfiler | 0.463 | 0.293 | **0.072** |
| *Multi-modal contrastive learning* | | | |
| CellCLIP (CLIP-B/16) | 0.593 | 0.434 | 0.024 |
| CellCLIP (DINOv2-Large) | 0.612 | **0.467** | 0.033 |
| CellCLIP (DINOv2-Giant) | **0.663** | 0.385 | 0.043 |
| CellCLIP (OpenPhenom-S/16) | 0.596 | 0.211 | 0.036 |
| CLOOME | 0.538 | 0.199 | 0.028 |
| *Weakly supervised models* | | | |
| ViT-L/16 | 0.513 | 0.217 | 0.028 |
| *Channel-agnostic MAE ViTs* | | | |
| OpenPhenom-S/16 | 0.357 | 0.219 | 0.031 |

Table 3: Comparison of different models for perturbation detection and matching within and across perturbation classes. Results are reported as mean average precision (mAP) across perturbations.

Overall, our findings demonstrate that contrastive learning approaches significantly outperform both WSL and SSL approaches. Within the considered contrastive learning approaches, we found that CellCLIP with a DINOv2-Giant backbone outperformed all other approaches for detecting replicates of the same perturbation. This result suggests that our approach can distinguish distinct perturbation effects more accurately compared to previous work.

For perturbation matching, where we compare profiles targeting the same genes within the same perturbation class (e.g., compounds vs compounds), CellCLIP (DINOv2-Large) again performs the best among all approaches. Notably, for cross-class perturbation matching, (e.g. CRISPR vs compounds), the overall mean Average Precision (mAP) for all machine-learning-based methods remains low, even falling below the baseline CellProfiler. This suggests that, despite targeting the same gene, morphological changes across different perturbation classes remain highly distinct, aligning with previous findings reported in Chandrasekaran et al. (2024).

## 6 Conclusion

In this work, we addressed the challenge of learning meaningful representations of Cell Painting images by introducing CellCLIP, a multi-modal contrastive learning framework that unifies perturbations across classes through textual descriptions. As part of our framework, we also developed CrossChannelFormer, a transformer-based architecture that efficiently captures channel dependencies and processes profile data while reducing computational costs. Our results demonstrate that CellCLIP improves retrieval performance, and replicates detection, and generalization across perturbation types. Overall, CellCLIP offers a promising solution for analyzing high-content morphological screening data, and future work will explore the impact of various contrastive losses and the contributions of each Cell Painting channel to downstream performance.

## 7 Meaningfulness Statement

We define meaningful representations of life as embeddings produced by methods that capture established biological features while enabling the discovery of novel ones, ultimately enhancing downstream task performance. Our approach leverages contrastive learning to integrate textual perturbation information with cell painting profiles. This integration captures underlying perturbation effects across diverse classes and facilitates the identification of perturbations with similar morphological signatures.

## References

Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth c x 16 x 16 words. *arXiv preprint arXiv:2309.16108*, 2023.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature protocols*, 11(9):1757–1774, 2016.

Juan C Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Joseph D Barry, Harmanjit Singh Bansal, Oren Kraus, et al. Data-analysis strategies for image-based cell profiling. *Nature methods*, 14(9):849–863, 2017.

Juan C Caicedo, Claire McQuin, Allen Goodman, Shantanu Singh, and Anne E Carpenter. Weakly supervised learning of single-cell feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9309–9318, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, et al. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7:1–11, 2006.

Safiye Celik, Jan-Christian Huetter, Sandra Melo, Nathan Lazar, Rahul Mohan, Conor Tillinghast, Tommaso Biancalani, Marta Fay, Berton Earnshaw, and Imran S Haque. Biological cartography: Building and benchmarking representations of life. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022.

Srinivas Niranj Chandrasekaran, Beth A Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods*, pp. 1–8, 2024.

Michael Doron, Théo Moutakanni, Zitong S Chen, Nikita Moshkov, Mathilde Caron, Hugo Touvron, Piotr Bojanowski, Wolfgang M Pernice, and Juan C Caicedo. Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*, 2023.

Philip Fradkin, Puria Azadi, Karush Suri, Frederik Wenkel, Ali Bashashati, Maciej Sypetkowski, and Dominique Beaini. How molecules impact cells: Unlocking contrastive phenomolecular retrieval. *arXiv preprint arXiv:2409.08302*, 2024.

Jiacheng Gu, Abhishek Iyer, Ben Wesley, Angelo Taglialatela, Giuseppe Leuzzi, Sho Hangai, Aubrianna Decker, Ruoyu Gu, Naomi Klickstein, Yuanlong Shuai, et al. Mapping multimodal phenotypes to perturbations in cells and tissue with crisprmap. *Nature Biotechnology*, pp. 1–15, 2024.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Markus Hofmarcher, Elisabeth Rumetshofer, Djork-Arne Clevert, Sepp Hochreiter, and Gunter Klambauer. Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *Journal of chemical information and modeling*, 59(3):1163–1171, 2019.

Alexandr A Kalinin, John Arevalo, Loan Vulliard, Erik Serrano, Hillary Tsang, Michael Bornholdt, Bartek Rajwa, Anne E Carpenter, Gregory P Way, and Shantanu Singh. A versatile information retrieval framework for evaluating profile strength and similarity. *bioRxiv*, 2024.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11757–11768, 2024.

Takamasa Kudo, Ana M Meireles, Reuben Moncada, Yushu Chen, Ping Wu, Joshua Gould, Xiaoyu Hu, Opher Kornfeld, Rajiv Jesudason, Conrad Foo, et al. Multiplexed, image-based pooled screens in primary cells and tissues with perturbview. *Nature Biotechnology*, pp. 1–10, 2024.

Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Efthymia Papalexi, Eleni P Mimitou, Andrew W Butler, Samantha Foster, Bernadette Bracken, William M Mauck III, Hans-Hermann Wessels, Yuhan Hao, Bertrand Z Yeung, Peter Smibert, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature genetics*, 53(3):322–331, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Meraj Ramezani, Erin Weisbart, Julia Bauman, Avtar Singh, John Yong, Maria Lozada, Gregory P Way, Sanam L Kavari, Celeste Diaz, Eddy Leardini, et al. A genome-wide atlas of human cell morphology. *Nature Methods*, pp. 1–13, 2025.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Cloome: contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nature Communications*, 14(1):7339, 2023.

Rico Sennrich. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Srinivasan Sivanandan, Bobby Leitmann, Eric Lubeck, Mohammad Muneeb Sultan, Panagiotis Stanitsas, Navpreet Ranu, Alexis Ewer, Jordan E Mancuso, Zachary F Phillips, Albert Kim, et al. A pooled cell painting crispr screening platform enables de novo inference of gene function by self-supervised deep learning. *bioRxiv*, pp. 2023–08, 2023.

Zitong Jerry Wang, Romain Lopez, Jan-Christian Hütter, Takamasa Kudo, Heming Yao, Philipp Hanslovsky, Burkhard Hoeckendorf, Rahul Moran, David Richmond, and Aviv Regev. Multi-contrastivevae disentangles perturbation effects in single cell images from optical pooled screens. *bioRxiv*, pp. 2023–11, 2023.

Ethan Weinberger, Ryan Conrad, and Tal Ashuach. Modeling variable guide efficiency in pooled crispr screens with contrastivevi+. *arXiv preprint arXiv:2411.08072*, 2024.

Yonghui Wu. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

# APPENDIX

## A DATASETS & PREPROCESSING

**Bray et al. (2016)** The dataset[2] consists of 919,265 five-channel microscopy images with resolutions 520 × 696 corresponding to 30,616 different molecules. These images were captured using 406 multi-well plates, with each image representing a view from a sample within one well. Six adjacent views collectively form one sample.

Sanchez-Fernandez et al. (2023) refined the dataset by removing images that were out of focus, exhibited high fluorescence, or contained untreated control cells. The final dataset comprises 759,782 images linked to 30,404 unique molecules, split into training (674,357), validation (28,632), and test (56,793) sets. The processed dataset is publicly available at[3]. Next, for retrieval evaluation on unseen compounds Section 4.3, following Sanchez-Fernandez et al. (2023), we removed samples from the test set that corresponded to the same molecule and plate to mitigate plate effects. The remaining samples are referred to as the final ”test set”, which consists of 2,115 compounds.

**CPJUMP1** CPJUMP1 (Chandrasekaran et al., 2024) comprises 186,925 nine-channel (3 bright field channels) microscopy images of resolution 1080 × 1080. This dataset features a comprehensive collection of perturbations conducted on U2OS and A549 cell lines, including 52 replicates. For perturbations, it includes 301 small-molecule compounds ( 46 controls), 335 sgRNAs (CRISPR) targeting 175 genes ( 88 control sgRNAs), and 175 ORFs (45 controls) for the corresponding genes. (Chandrasekaran et al., 2024) also provides annotations of associations between genes and compounds, enhancing its utility for exploring gene function and compound effects. We split the dataset into 70/10/20 for training, validation, and testing. Raw data, relevant metadata, and gene annotation can be found in[4]

**Image Preprocessing** For both datasets, our preprocessing followed the protocols established by Sanchez-Fernandez et al. (2023) and Hofmarcher et al. (2019), which consist of converting the original TIF images from 16-bit to 8-bit and removing the 0.0028 % of pixels with the highest values[5]. To maintain channel consistency between datasets, we processed the CPJUMP1 images and reordered channels to match the five-channel format of Bray et al. (2016).

## B ADDITIONAL DETAILS ABOUT TRAINING & BASELINES

### B.1 BASELINES FOR RETRIEVAL EVALUATION

We compared CellCLIP against random baselines and CLOOME to assess the impact of different training modalities on retrieval performance. Since molPhenix (Fradkin et al., 2024) is trained on proprietary datasets and its implementation is not publicly available, we did not include it in this work.

**Random** For the random baseline, given a perturbation projection embedding, we randomly select a text embedding.

**CLOOME** We follow the official implementation available at [6]. For retrieval evaluation with Bray et al. (2016), we use the best-performing hyperparameters reported in (Sanchez-Fernandez et al., 2023), employing ResNet50 as the vision encoder and a four-layer MLP as the molecule encoder, excluding the Hopfield layer. The training loss follows the original contrastive loss, using raw images paired with a max-pooled combination of Morgan and RDKit count-based fingerprints[7], resulting in an 8192-bit input representation. The training setup includes a batch size of 256, the

[2] http://gigadb.org/dataset/100351

[3] https://ml.jku.at/software/cellpainting/dataset/

[4] https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods

[5] TIF files preprocessing

[6] https://github.com/ml-jku/cloome

[7] The official sources for the RDKit library

AdamW optimizer, and a learning rate of $1e^{-3}$ with cosine annealing with restart. The learnable temperature parameter $\tau$ was set to 14.3. The model is trained for 70 epochs. We use the same training parameter above for CLOOME's variants in Table 1.

**Pretrained Models for CellCLIP Profiles** For generating channel embeddings (profiles), Section 3, for CellCLIP training, we experimented with a range of image foundation models, including DINO, DINOv2 (small, base, large, giant), CLIP (L16, B14), SigCLIP (so400m, B-16), and CA-MAE (OpenPhenom-S/16). Checkpoints for all models are available on Hugging Face[8]. Cell painting images were treated as grayscale images with a multi-crop strategy and preprocessed using their respective preprocessors.

### B.2 Perturbation detection & matching

For perturbation detection and matching, we compared CellCLIP with the best-performing pretrained models, including DINOv2-giant and DINOv2-large, against CellProfiler, CLOOME, and pretrained models, including CA-MAE (Openphenom)[9] and ViT-L/16[10]. The evaluation pipelines were implemented, following in[11].

**CellProfiler** For CellProfiler features of CPJUMP1, we utilize embeddings provided by Chandrasekaran et al. (2024)[12].

**CLOOME** Since CLOOME is designed specifically for small molecules, we fine-tuned the model using only small molecules within CPJUMP1 training set, employing the same parameters used in training with the (Bray et al., 2016) dataset.

**CellCLIP** Similarly, for CellCLIP, we used the model pretrained from Bray et al. (2016) and fine-tuned with CPJUMP1 using the same parameters as pretraining.

**Weakly Supervised Learning (WSL)** For the weakly supervised baseline, following Kraus et al. (2024), we constructed a Vision Transformer (ViT) Large with a patch size of 16, modified to accommodate five channels, serving as the backbone (Alexey, 2020). We attached a classifier head to this backbone and trained the model for 10 epochs using a learning rate of $1e^{-3}$, incorporating weight decay. The batch size was set at 256. For perturbation detection and evaluation experiments in CPJUMP, we utilized the output from just before the classifier head as the learned embeddings.

### B.3 Batch Effect Correction

For the detection and matching evaluation in CPJUMP1, we follow the approach of Celik et al. (2022). First, we fit a PCA kernel[13] on all control images (profiles) across experimental batches (e.g., the assay's plate wells). Then, we transform all embeddings using this PCA kernel. Next, for each experimental batch, we fit a separate `StandardScaler` on the transformed embeddings of the controls from step 2 and use it to normalize the remaining embeddings from that batch. For kernel selection, we experimented with RBF, polynomial, and linear kernels, selecting the best-performing kernel for each method.

### B.4 Implementation Details

This study employs the PyTorch package tutorial (version 2.2.1). All experiments are conducted on systems equipped with 64 CPU cores and the specified NVIDIA GPUs. Models trained with the largest possible batch size on 8 RTX-6k GPUs.

[8] https://huggingface.co/

[9] https://huggingface.co/recursionpharma/OpenPhenom

[10] https://github.com/pprp/timm

[11] https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods/blob/main/benchmark/1.0.calculate-map-cp.ipynb

[12] https://github.com/jump-cellpainting/2024_Chandrasekaran_NatureMethods

[13] https://scikit-learn.org/1.5/modules/generated/sklearn.decomposition.KernelPCA.html