PLOS Computational Biology

RESEARCH ARTICLE

Integrating chemical structures as treatments improves representations of microscopy images for morphological profiling

Yemin Yu[1‡], Emre Hayir[2], Neil Tenenholtz[2], Lester Mackey[2], Ying Wei[3], David Alvarez-Melis[2], Ava P. Amini[2], Alex X. Lu[2*]

1 Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong, **2** Microsoft Research, Cambridge, Boston, Massachusetts, United States of America, **3** Department of Computer Science, Zhejiang University, Hangzhou, China

‡ Work primarily done during an internship at Microsoft Research.
* lualex@microsoft.com

Citation: Yu Y, Hayir E, Tenenholtz N, Mackey L, Wei Y, Alvarez-Melis D, et al. (2026) Integrating chemical structures as treatments improves representations of microscopy images for morphological profiling. PLoS Comput Biol 22(7): e1014483. https://doi.org/10.1371/journal.pcbi.1014483

Editor: Virginie Uhlmann, University of Zurich Faculty of Mathematics and Science: Universitat Zurich Mathematisch-Naturwissenschaftliche Fakultat, SWITZERLAND

Received: May 8, 2025

Accepted: June 22, 2026

Published: July 29, 2026

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1014483

Abstract

Recent advances in self-supervised deep learning have improved our ability to quantify cellular morphological changes in high-throughput microscopy screens, a process known as morphological profiling. However, most current methods only learn from images, despite many screens being inherently multimodal, as they involve both a chemical or genetic perturbation as well as an image-based readout. We hypothesized that incorporating chemical compound structures during self-supervised pre-training could improve learned representations of images from high-throughput microscopy screens. We introduce a representation learning framework, MICON (Molecular-Image Contrastive Learning), that models chemical compounds as treatments that induce transformations of cell phenotypes. MICON significantly outperforms classical hand-crafted features such as CellProfiler and existing deep-learning-based representation learning methods in challenging evaluation settings where models must identify reproducible effects of drugs across independent replicates and data-generating centers. We demonstrate that incorporating chemical compound information into the learning process provides small, but consistent improvements in performance and that modeling compounds specifically as treatments outperforms approaches that directly align images and compounds in a single representation space. Our findings point to a new direction for representation learning in morphological profiling, suggesting that methods should explicitly account for the multimodal nature of microscopy screening data.

Author summary

Large-scale microscopy experiments can be useful for many applications, such as discovering new drugs. To analyze these images, scientists will extract

Data availability statement: Code to reproduce our analyses and model is available at github.com/microsoft/MICON. Model weights and representations from our models are available at https://huggingface.co/datasets/Yemin/MICON/tree/main. As the image datasets are readily available from the JUMP Cell Painting Consortium, and too large to re-upload, we instead provide scripts to reproduce our splits and preprocessing. The source data is available at the JUMP Cell Painting AWS instance at https://registry.opendata.aws/cellpainting-gallery/.

Funding: The author(s) received no specific funding for this work.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: E.H, N.T, L.M, D.A.M, A.P.M, and A.X.L are employees of and hold equity in Microsoft.

representations, series of measurements, that quantify how the shape of cells in these images may change when they're treated with drugs. Recently, deep learning models have been used to learn representations of microscopy images. However, most models only learn from image data alone, despite the fact that many microscopy experiments involve treating cells with drugs, and information about the drug could improve learning. In this work, we propose a new way of training deep learning models that uses both images and the chemical structure of drugs to learn representations for microscopy. Unlike prior proposals, our method models drugs as treatments that alter cells in images. We show that our method learns representations more effective for drug screening applications, and that small, but significant improvements are caused by incorporating information about drugs over just using images alone.

Introduction

High-throughput microscopy experiments have become instrumental to drug discovery, guided by the principle that changes in cellular morphology can give insight into drug efficacy and mechanism of action [1,2]. By coupling scalable fluorescent staining methods like Cell Painting [3] with large libraries of small molecules and high-throughput instrumentation, researchers can screen millions of compounds in cell lines. Recent advances in computer vision promise to automate the quantification of increasingly subtle morphological changes in response to these compound treatments [4,5]. The goal of these computational methods is to convert an image into a vector of features, or a representation, that captures cell phenotype, a process called morphological profiling. Representations facilitate the quantitative comparison of phenotypes in microscopy images. For example, to nominate candidate drugs, a researcher might seek to use morphological profiling to identify compounds that yield similar representations — and therefore similar phenotypes — to known effective drugs.

Hand-crafted features, made accessible by the CellProfiler [6] software, have traditionally been used to construct these representations. However, these coarse features often fail to pick up on subtle morphological changes and are confounded by microscopy batch effects that cause irreproducible technical variation. The need for high-quality, robust representations has led researchers to investigate deep-learning based representation learning methods [7]. Since labels for morphological profiling are scarce and frequently unreliable, many representation learning methods train using self-supervised learning, which aims to learn high-quality representations without labeled training data [8]. For morphological profiling, previous studies have explored various self-supervised frameworks [9], including contrastive learning [10–12], masked autoencoding [13,14], and self-distillation [15,16].

While representation learning approaches have shown improvements in sensitivity and robustness relative to hand-crafted features, most methods learn from images exclusively. However, microscopy-based drug screens are inherently multimodal:

perturbations, which may be chemical, genetic, or environmental, are used to induce phenotypic changes in cells. Few works have investigated how integrating multimodal information about perturbations, such as the identities of chemical compounds, can improve representation learning. Some weakly supervised methods predict perturbation identity as a target during training [17], but these methods only use the identity of the chemical compound and not any information about its structure. CLOOME [18] and MolPhenix [19] integrate chemical structure information through a CLIP (Contrastive Language Image Pre-training) [20] objective to align images and chemical compounds in a unified representation space that enables querying of data from one modality given the other. However, morphological phenotypes and chemical compounds are not directly comparable; for example, depending on cell type, cells may produce different phenotypic responses to the same compound. This may create tensions where a single representation of a chemical compound must be aligned to multiple representations across a dynamic range of phenotypic responses.

Rather than assuming cell phenotypes in images and chemical compounds to be directly comparable, we instead model chemical compounds as *treatments* that induce transformations of microscopy images to develop MICON (Molecular-Image Contrastive Learning), a self-supervised representation learning framework for morphological profiling (Fig 1). Given an image of unperturbed cells and a chemical compound, MICON learns how to generate a representation corresponding to a phenotypic transformation of the cells in the image, as if the cells had been treated with the compound. Using the MICON framework, we controllably test the hypothesis that integrating chemical compound information can improve image representation learning in morphological profiling. First, we show that MICON achieves effective representation learning, substantially improving over both CellProfiler features and previously proposed deep learning methods. Second, through ablation of MICON's multimodal component, we demonstrate that integrating chemical compound information provides a small but statistically significant improvement in performance relative to learning from images alone. Finally, supporting our hypothesis that *how* multimodal information is integrated is crucial, we show that modeling chemical compounds as treatments significantly outperforms direct alignment of images and compounds with a CLIP objective. Our results advocate for approaching morphological profiling as a multimodal learning problem and provide insight into how to best integrate images and perturbations.

Materials and methods

MICON: Molecular-image contrastive learning

MICON is a novel representation learning method for morphological profiling experiments that integrates chemical compounds and images into a multimodal pre-training task (Fig 1). MICON models chemical compounds as treatments that induce a transformation in images. Given an image of untreated control cells and a chemical compound, our method transforms the cells in the image as if they had been treated with the chemical compound. In other words, we train MICON to simulate the phenotypic effects of chemical compounds, given a chemical compound and an image of untreated cells as input. We perform this transformation directly in the image representation space, using contrastive losses to align predicted and real representations (Fig 1A). We next describe the two components of our representation learning approach: a perturbation-aware contrastive learning objective that learns faithful representations of phenotypes in real microscopy images and a predicted perturbation-aware contrastive objective that aligns predicted representations and perturbation representations.

Perturbation-aware contrastive learning of representations. As our method relies on alignment of real and predicted perturbation representations, we must first ensure that representations of real phenotypes are robust and capture the phenotypic effects of perturbations. However, learning such a representation is challenging because of microscopy batch effects, which induce differences in images that stem from irreproducible technical variation [21].

Batch effects are known to be a major confounder even in replicates at the same data-generating institution [4]. Our work leverages a dataset (cpg0016, see Datasets and Evaluation) where images are supplied by multiple sources, which introduces further differences between experiments (e.g., images may be collected under different microscopes) and thus

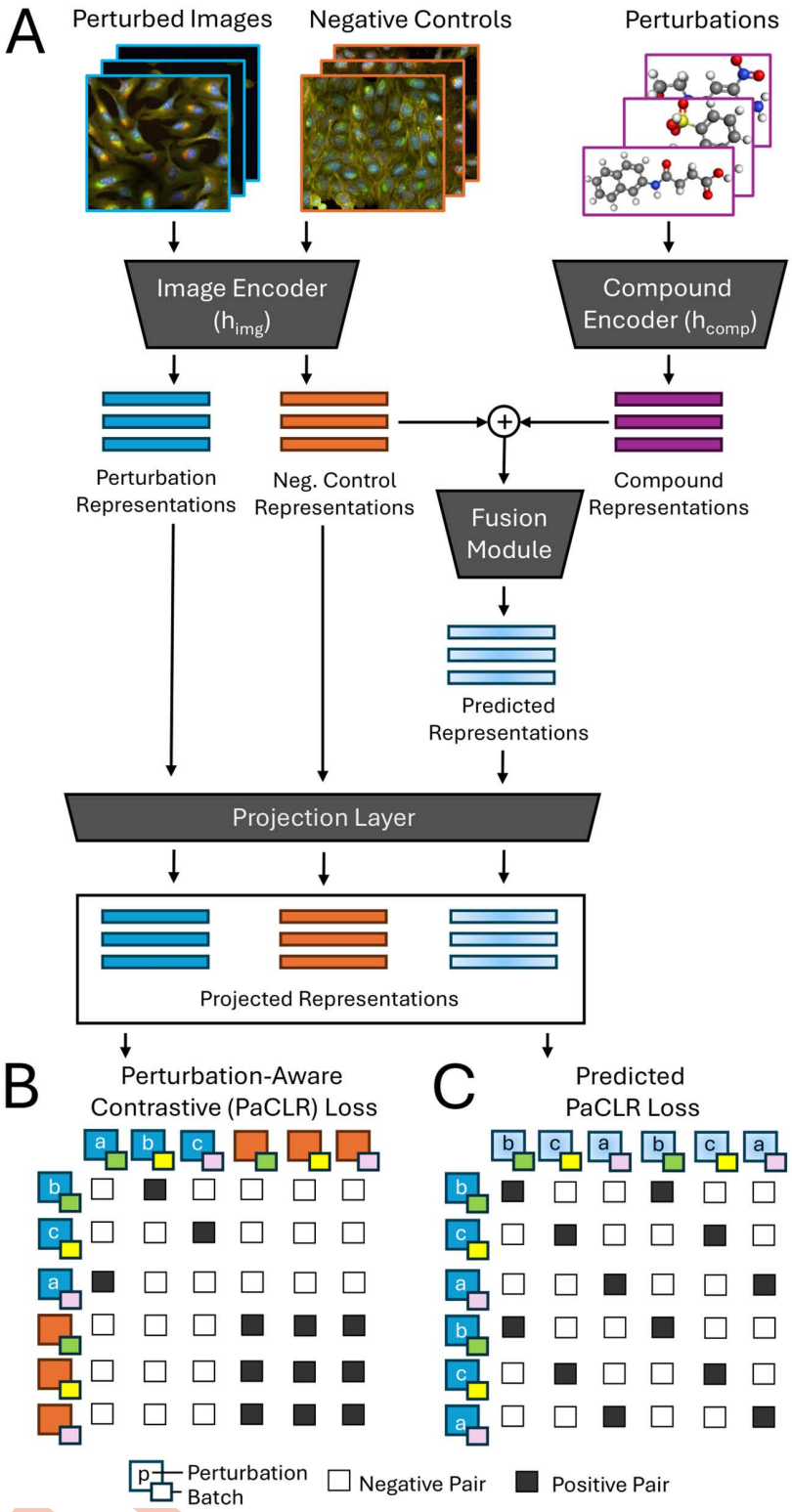

Fig 1. Summary of the MICON framework. (A) An overview of our model. Compound perturbed (blue) images and negative control (orange) images are transformed into representations by an image encoder, while chemical compounds (purple) are transformed into representations by a compound

encoder. A fusion module accepts pairs of negative control representations and compound representations and outputs predicted representations (light blue). All representations are passed through a projection layer to form the final projected representations. **(B)** Summary of the perturbation-aware contrastive (PAC) loss. Perturbed images (blue) and negative controls (orange) from the same batches as the perturbed images (batches denoted by small colored boxes) are sampled, with at least two images for each perturbation. Images treated with the same perturbation (a, b, or c) serve as positive pairs (black) for the PaCLR loss. **(C)** Summary of the PaCLR loss on predicted images (abbreviated as Predicted PaCLR loss). Perturbation representations (blue) and predicted representations (light blue) using the same compound (a, b, or c) serve as positive pairs (black) for the predicted PaCLR loss.

https://doi.org/10.1371/journal.pcbi.1014483.g001

even more severe microscopy batch effects [21]. If representations of real phenotypes are not robust, learning to predict the phenotypic impact of perturbations may result in spurious learning of microscopy batch effects, not the perturbation. To learn representations robust to microscopy batch effects, we design a Perturbation-Aware Contrastive Learning of Representations (PaCLR) objective, which leverages metadata about perturbation and microscopy batches to sample positive and negative pairs for contrastive learning.

We first consider a more standard self-supervised contrastive learning objective, i.e., SimCLR [10]. SimCLR trains using self-augmented pairs of images. In this setting, an image $\mathbf{x}$ is encoded by an image encoder $\mathbf{h}_{\text{img}}$, then passed through a projection head g to obtain a projected representation $\mathbf{t} = g(\mathbf{h}_{\text{img}}(\mathbf{x}))$. Given N images with projected representations $\mathbf{t}_1, \cdots, \mathbf{t}_N$, a contrastive loss can be used to pull together the representation of an image and its augmentations (positive pairs), while repelling representations of dissimilar images (negative pairs). For a given image index i, we let j index a differently augmented version of the same image and k iterate over all images other than i in the set (where $k \neq j$ are assumed to be dissimilar to i). The contrastive loss is then given by

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\tau^{-1}\mathtt{sim}(\mathbf{t}_i, \mathbf{t}_j))}{\sum_{k=1}^{N} \mathbb{1}_{[i \neq k]} \exp(\tau^{-1}\mathtt{sim}(\mathbf{t}_i, \mathbf{t}_k))},$$

where τ is the temperature scale hyperparameter for the softmax function and $\mathtt{sim}$ is a similarity measure (in our work, the cosine distance, consistent with standards established by SimCLR [10]).

This naive contrastive strategy poses two problems in the context of morphological profiling. The first is class collision. In the above setting, all other images in a training batch are considered negative samples. In high-throughput imaging however, some images in the same training batch may share the same perturbation and thus carry related morphological changes that we wish to capture. Second, augmented pairs of images are still likely to share microscopy batch effects, which makes a standard contrastive strategy prone to learning confounding technical variation, instead of focusing on the effect of the perturbation alone.

By leveraging metadata about the perturbation and microscopy batch to sample pairs for contrastive learning, the PaCLR loss addresses both of these issues (Fig 1B). First, we use metadata about perturbations to define positive pairs. Instead of using different augmented views of a single instance, we use images of the same perturbation as positive pairs. The underlying assumption is that, if the network learns to align two images of the same perturbation, the shared signal that is learned should be the biological signal, rather than confounding microscopy batch effects (since these images are drawn uniformly from the dataset, positive pairs can come from different microscopy batches, although this is not strictly enforced). Second, we use metadata about microscopy batch to sample negative control images (which are not treated with any compound, but rather a control treatment of only the solvent in which compounds are dissolved in). This means there are samples that share microscopy batch effects, but not perturbation, serving as negative examples, further encouraging the model to be robust to microscopy batch effects. We sample negative control images such that they come from the same microscopy plate as perturbed images or from the same microscopy batch as a fall-back for the rare case that there are no viable control images on the same plate (see Datasets and Evaluation for a description of the structure of microscopy experiments). In addition to image instances perturbed by different compounds, these negative control

images act as negative pairs in training, which further discourages the MICON model from learning about microscopy batch effects, as negative pairs are enforced to come from the same microscopy batches as perturbed images.

The above is implemented through sampling of batches during training. Given an image $\mathbf{x}$, let $\mathbf{p}$ be its perturbation, and $\mathbf{b}$ be its microscopy batch identifier. We first sample T perturbed images from the entire dataset to form the subset $\mathcal{T}_1 = \{\mathbf{x}_1^{\mathcal{T}_1}, ..., \mathbf{x}_T^{\mathcal{T}_1}\}$. Next, for each image in $\mathcal{T}_1$, we sample a second distinct image from the dataset with the same perturbation to form the subset $\mathcal{T}_2 = \{\mathbf{x}_1^{\mathcal{T}_2}, ..., \mathbf{x}_T^{\mathcal{T}_2}\}$, so $\mathbf{p}_j^{\mathcal{T}_1} = \mathbf{p}_j^{\mathcal{T}_2}$ but $\mathbf{x}_j^{\mathcal{T}_1} \neq \mathbf{x}_j^{\mathcal{T}_2}$. Finally, we sample C negative control images from the dataset to form the subset $\mathcal{C} = \{\mathbf{x}_1^{\mathcal{C}}, ..., \mathbf{x}_C^{\mathcal{C}}\}$, such that their microscopy batch identifiers match as many of the perturbed images ($\mathcal{T}_1 \cup \mathcal{T}_2$) as possible given the size of the subset: $\mathbf{b}_j^{\mathcal{C}} \in \{\mathbf{b}_1^{\mathcal{T}_1}, ..., \mathbf{b}_T^{\mathcal{T}_1}\} \cup \{\mathbf{b}_1^{\mathcal{T}_2}, ..., \mathbf{b}_T^{\mathcal{T}_2}\}$, and $|\{\mathbf{b}_1^{\mathcal{C}}, ..., \mathbf{b}_C^{\mathcal{C}}\}| = \min(|\{\mathbf{b}_1^{\mathcal{T}_1}, ..., \mathbf{b}_T^{\mathcal{T}_1}\} \cup \{\mathbf{b}_1^{\mathcal{T}_2}, ..., \mathbf{b}_T^{\mathcal{T}_2}\}|, C)$. Putting these together, the training batch is the union of these three subsets $\mathcal{T}_1 \cup \mathcal{T}_2 \cup \mathcal{C}$.

Finally, we extract a representation $\mathbf{t}_i = g(\mathbf{h}_{\text{img}}(\mathbf{x}_i))$ from each image $\mathbf{x}_i$ using our image encoder $\mathbf{h}_{\text{img}}$. With this notation, we define the PaCLR loss,

$$\mathcal{L}_{\text{PaCLR}}^{\text{real}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sum_{k=1}^{N}\mathbb{1}_{[\mathbf{p}_i=\mathbf{p}_k, i\neq k]}}\sum_{j=1}^{N}\mathbb{1}_{[\mathbf{p}_i=\mathbf{p}_j, i\neq j]}\log\frac{\exp(\tau^{-1}\texttt{sim}(\mathbf{t}_i, \mathbf{t}_j))}{\sum_{k=1}^{N}\mathbb{1}_{[i\neq k]}\exp(\tau^{-1}\texttt{sim}(\mathbf{t}_i, \mathbf{t}_k))}. \quad (1)$$

PaCLR's integration of metadata about perturbation into representation learning is an idea also explored by several previous works, but our implementation of this idea has several key differences. Compared to Set-DINO [22] and WS-DINO [23], we use images treated with the same perturbation as positive pairs for contrastive learning, instead of enforcing similarity in a student-teacher set-up. Compared to weakly supervised proposals [17,24], we use metadata about perturbations for sampling pairs, rather than as predictive targets. Finally, we note that unlike many other prior works which only use perturbed images [18,23], we explicitly include negative control images. We incorporate negative controls because, due to their lack of perturbation treatment, they primarily represent phenotypic changes caused by microscopy batch effects and thus enable disentanglement of microscopy batch effects from perturbation effects.

Predicted perturbation-aware contrastive objective. To incorporate chemical compound information into MICON's learning objective, we first encode chemical compounds into representations using a compound encoder $\mathbf{h}_{\text{comp}}$ (Fig 1A). A fusion module F combines the resulting representations of chemical compounds with the representation of a negative control image. The fusion module predicts representations reflecting if cells were to be treated with a given perturbation (Fig 1A). To ensure these predicted representations are accurate, we align predicted representations with real representations of images of cells actually treated with the compounds using a variant of the perturbation-aware contrastive loss. Unlike the original PaCLR loss (Equation equation 1), the PaCLR loss on the predicted images (abbreviated as Predicted PaCLR loss in Fig 1) is calculated using only perturbation representations, i.e., excluding the negative controls (Fig 1C). However, the negative control images used to generate predicted representations in the fusion module are sampled such that they come from the same microscopy batches as the real perturbation representations.

Formally, we use a fusion model F to combine the representation of a chemical compound $\tilde{\mathbf{p}}$ and the representation of a negative control image $\mathbf{x}^c$ into a predicted representation $\tilde{\mathbf{t}} = F(g(\mathbf{h}_{\text{img}}(\mathbf{x}^c)), \mathbf{h}_{\text{comp}}(\tilde{\mathbf{p}}))$. Then, given $N/2$ perturbation representations and their perturbation treatments, denoted as the tuple $(\mathbf{t}_i, \mathbf{p}_i)$, and $N/2$ predicted representations and their perturbation treatments, denoted as the tuple $(\tilde{\mathbf{t}}_j, \tilde{\mathbf{p}}_j)$, we define the following PaCLR loss on the predicted images over the predicted representations and the real perturbation representations:

$$\mathcal{L}_{\text{PaCLR}}^{\text{pred}} = -\frac{2}{N}\sum_{i=1}^{N/2}\frac{1}{\sum_{k=1}^{N/2}\mathbb{1}_{[\mathbf{p}_i=\tilde{\mathbf{p}}_k]}}\sum_{j=1}^{N/2}\mathbb{1}_{[\mathbf{p}_i=\tilde{\mathbf{p}}_j]}\log\frac{\exp(\tau^{-1}\text{sim}(\mathbf{t}_i, \tilde{\mathbf{t}}_j))}{\sum_{k=1}^{N/2}\exp(\tau^{-1}\text{sim}(\mathbf{t}_i, \tilde{\mathbf{t}}_k))}.$$

The final loss for MICON combines the PaCLR objectives on the real versus predicted images. We weight these loss components equally:

$$\mathcal{L} = \mathcal{L}^{\text{real}}_{\text{PaCLR}} + \mathcal{L}^{\text{pred}}_{\text{PaCLR}}$$

Datasets and evaluation

Training dataset. We leveraged a curated subset of the JUMP Cell Painting Consortium's (JUMP-CP) cpg0016 dataset, a large-scale dataset cell painting microscopy images of drug-treated U2-OS cells [25]. We chose cpg0016 as it consists of data from 12 data-generating sources, enabling us to test robustness to microscopy batch effects more extensively.

Each source collects images of cells cultured on *plates*, which are sub-divided into 384 or 1536 *wells*, where each well is capable of containing an independent experiment treated with a different perturbation. A robot-controlled microscope takes images of the experiment in each well, typically multiple images of different parts of the well, where each distinct image of the same well is a *field of view (FOV)*. Plates are further organized into *batches*, which we term a *microscopy batch* to distinguish from a batch of data for model training, where plates in the same microscopy batch are imaged on the same day. While performing the same perturbation experiment across different wells, plates, microscopy batches, or sources can all generate microscopy batch effects, microscopy batch effects are generally considered more pronounced across different microscopy batches or sources, as opposed to different wells on the same plate, or different plates in the same microscopy batch [21] (supporting this, we show the average distance between negative control CellProfiler features across batches and sources in JUMP-CP in S1 Fig). Further details on the dataset and data generation setup can be found in the original JUMP-CP paper [25].

We curated several kinds of plates and wells from the cpg0016 dataset for training and evaluating our models. First, we curated positive control wells (abbreviated as POS-CTL in this work) from the entire dataset; these reflect 8 compounds selected to have the most distinct morphological changes in a pilot study. Most plates in cpg0016 have 4 well replicates of each of these 8 positive controls. Second, we curated JUMP-Target-2-Compound plates (abbreviated as Target-2 in this work), which image 301 compounds selected as likely to induce phenotypic effects in cells, and are measured in replicate by every source. Third, we curated a subset of 184 compounds from the Compound plates of cpg0016 (see Data Stratification and Compound-Replicate Matching for further details). Finally, every plate includes negative control wells treated with Dimethyl Sulfoxide (DMSO), the solvent used for all compound perturbations, which acts as a control where no compound is dissolved in DMSO. We curate negative controls from all plates from which we sampled POS-CTL, Target-2, and Compound wells. Together, our dataset comprises 87,282 of the 1,096,069 wells in cpg0016. Note that for our loss, we consider a well to be an image $\mathbf{x}_i$, and sample a random FOV for the well. This has the impact that no FOVs from the same well are ever within the same batch of data during training of our method.

Data stratification and compound-replicate matching. We purposed cpg0016 for both training, and for compound-replicate matching evaluations, a standard strategy for evaluating if representations robustly identify phenotypes in the presence of microscopy batch effects [4,26]. The principle underlying compound-replicate matching is that cells treated with the same perturbations should have similar phenotypes, and therefore have similar representations, even if experiments originate from different microscopy batches or sources. Compound-replicate matching is typically implemented through nearest-neighbor retrieval [26–29], which means for data stratification, in addition to defining *training* and *validation* datasets for representation learning, we also need to define compound-replicate matching *retrieval* and *query* datasets for evaluation. During evaluation, we retrieve the nearest neighbor in the retrieval dataset for each well from the query dataset, and measure whether these share the same perturbation treatment. There should be no overlap between the query dataset and each of the model training and validation datasets, to avoid overfitting to images seen during representation learning.

Our data stratification strategy is motivated by two guiding use cases for morphological profiling: first, the identification of reproducible biological changes in unseen data for sources and compounds seen during training, which we refer to as "in-distribution (ID) generalization"; and second, the identification of these changes in unseen data for sources and compounds **not** seen during training, which we refer to as "out-of-distribution (OOD) generalization". The first use case captures scenarios such as those when a source trains a model on its own data and then seeks to evaluate it on newly-produced data. The second use case captures settings such as those when a distinct source seeks to use a pre-existing model, trained on data from other sources, to analyze their data, or when a source seeks to use a pre-existing model on new compounds not seen in training. To capture ID and OOD performance for *both* unseen sources and unseen compounds, we produce four distinct data splits on different subsets of data within the cpg0016 dataset (Figs 2 and 3).

We used the POS-CTL dataset to evaluate model generalization to unseen sources (Fig 2A). We stratified the dataset by source in two ways to test ID and OOD performance (Fig 2B and 2C), with metadata about splits available in Table E in S2 Text). For the ID setting (Fig 2B), we used the 6 cpg0016 sources that have at least 10 POS-CTL plates.

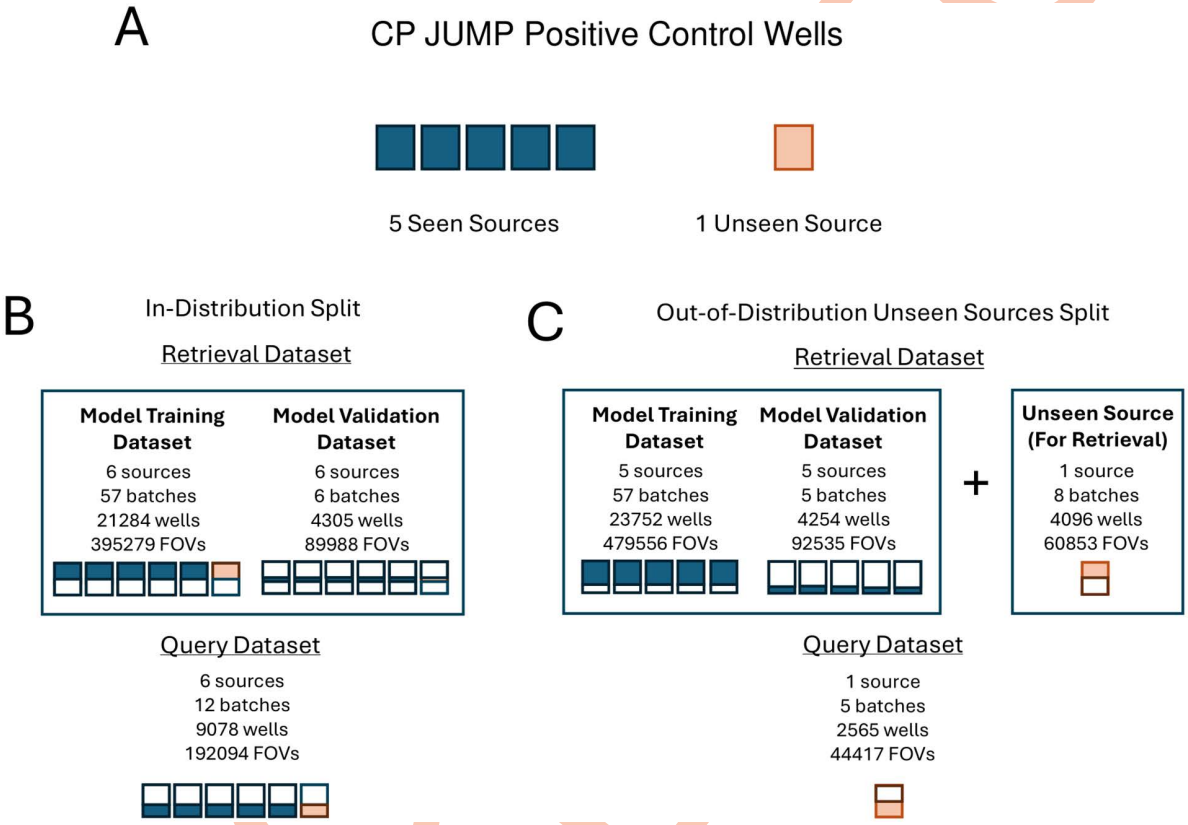

Fig 2. Stratification of positive control wells into in-distribution and out-of-distribution datasets. (A) POS-CTL wells from a total of six sources are used. For our out-of-distribution evaluation, five sources are designated as seen (blue boxes) during model training, and 1 source is designated as unseen (orange boxes). In panels B and C, filled parts of the boxes indicate data from each source in that split, while unfilled (white) parts of the boxes indicate held-out data (note that filled/unfilled ratios are not proportionately scaled to split ratios, and are just intended to illustrate where splits are disjoint). **(B)** The in-distribution evaluation trains the representation learning model on a subset of microscopy batches from all six sources, holding out one batch from each source for validation. The training and validation dataset are used as the retrieval dataset for compound-replicate matching, with unseen microscopy batches from all six sources used as the query dataset. **(C)** The out-of-distribution evaluation trains and validates the representation learning model on the five seen sources. For evaluation, the retrieval dataset is the training and validation data plus a subset of microscopy batches from the unseen source, and the query dataset is a disjoint subset of microscopy batches from the unseen source.

https://doi.org/10.1371/journal.pcbi.1014483.g002

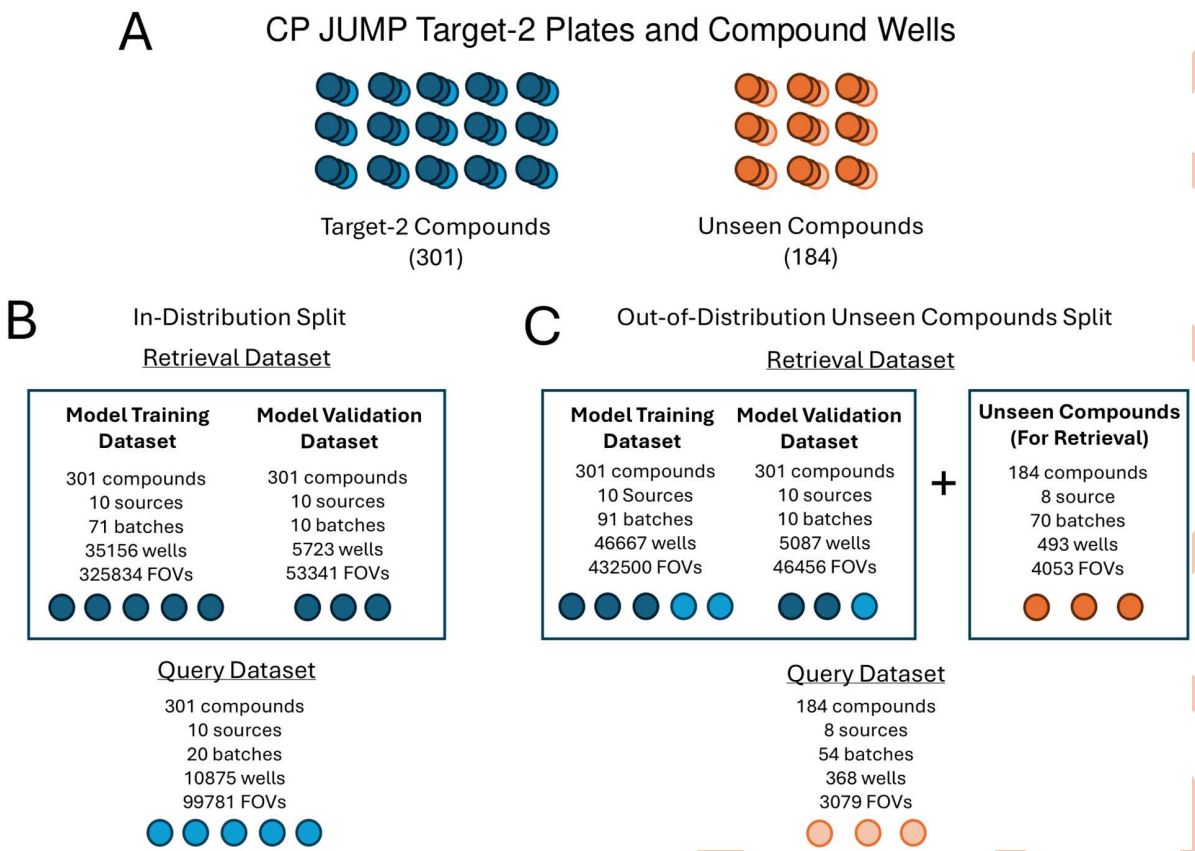

Fig 3. Stratification of Target-2 and Compound plates into in-distribution and out-of-distribution datasets. (A) The 301 Target-2 compounds (blue) are designated as seen compounds for training representation learning models, while 184 unseen compounds (orange) are sampled from the Compound plates and held out. **(B)** The in-distribution evaluation trains the representation learning model on 80% of microscopy batches from all Target-2 plates, reserving 20% of microscopy batches for the query dataset. **(C)** The out-of-distribution evaluation trains the representation learning model on all microscopy batches from the Target-2 plates. The training dataset is combined with 493 wells for 184 unseen compounds to form the retrieval dataset and evaluated on a query dataset of a disjoint set of 368 wells of held-out, unseen compounds from the Compound plates.

https://doi.org/10.1371/journal.pcbi.1014483.g003

We trained representation learning models on a subset of data from all 6 sources stratified by microscopy batch, and we held out one batch from each source as validation data for model selection. This training dataset is recycled as the retrieval dataset, but we reserved unseen batches from all 6 sources as the query dataset (Fig 2B and Table E in S2 Text). We split microscopy batches into the retrieval dataset versus the query dataset at a 85–15 ratio. For the OOD setting (Fig 2C), we used the same 6 cpg0016 sources as before, but created a training dataset containing data from just 5 of the 6 sources. The retrieval dataset includes data from the training dataset, but also a subset of microscopy batches from the unseen source (50% of batches). We added the unseen source to the retrieval dataset because this allows us to quantify two settings for a source not included in training data: first, if the source uses the pre-trained model to compare between images generated internally, and second, if the source uses the pre-trained model to compare their images to another different source (NSB and NSS respectively, see Compound-Replicate Matching Implementation). The query dataset is a disjoint 50% of microscopy batches from the unseen source. Dataset stratification is randomized over 3 random seeds, and this is a source of uncertainty in our error bars (in addition to randomizing weights, where applicable for the method).

We used the Target-2 dataset to evaluate generalization to unseen compounds. We stratified the Target-2 dataset to test ID performance, and used a mixture of Target-2 and Compound plates to test OOD performance (Fig 3, with metadata about splits available in Table F in S2 Text). For the ID setting (Fig 3B), we reserved 20% of the microscopy batches as the query dataset. The remaining 80% of batches are split into training and validation datasets (holding out one batch from each source for validation), and additionally used as the retrieval dataset. For the OOD setting (Fig 3C), we trained representation learning models using all batches from the Target-2 plates. To evaluate, we extended the dataset by identifying compounds from the Compound plates that elicited strong phenotypic effects across sources. For each source we ranked compounds in the Compound plates based on the replicate-averaged distance between the CellProfiler [30] image features for the compound versus features for the negative control spatially closest to the well in the plate map. We then nominated compounds in the top 10% of greatest distances for at least four sources, yielding a total of 184 compounds. All of these compounds have disjoint mechanisms of action (MoAs) with the Target-2 compounds. Our retrieval dataset combines the training dataset with a subset of wells (493 wells) for the unseen compounds, while the query dataset contains 368 unseen wells (2 wells per compound). This stratification scheme means that the 184 unseen compounds can be matched to 485 compounds (both the 301 compounds seen and the 184 compounds unseen), making the compound-replicate matching task more difficult and penalizing models if they overfit to compounds seen in training.

Compound-replicate matching implementation. To implement compound-replicate matching, we identified the nearest neighbor in the retrieval dataset, using the cosine distance of the representation, for each well in the query dataset. We reported the accuracy of this retrieval, defined as the fraction of wells where the nearest well in the retrieval dataset is treated with the same perturbation. As each well is typically imaged with multiple FOVs, we averaged the representations of all FOVs to form a well-level representation. To test if representations are robust to microscopy batch effects caused by differences in microscopy batch and source, we further restricted matching to the nearest neighbor in a different microscopy batch (Not in Same Batch - NSB) or source (Not in Same Source - NSS).

Mechanism of action enrichment. In addition to compound replicate matching, we also evaluated models on mechanism of action (MoA) enrichment. We use the BBBC036 dataset, previously proposed for this evaluation by [24]. We further implement this evaluation using images from the JUMP-CP compound dataset. MoA labels were obtained from the Drug Repurposing Hub [31], and for this task only compounds that are labeled in the Drug Repurposing Hub with a corresponding MoA was used. A query compound was excluded from the task if there was no other compound with a shared MoA. This evaluation aims to assess if compounds with biologically similar effects are clustered in representation space, and measures retrieval of distinct compounds with similar MoAs. In this evaluation, all other compounds are ranked using a query compound. The top 1% of matching compounds is evaluated for enrichment in compounds with the same MoA as the query compound, relative to the other 99% of compounds in the ranked list, reported using the odds-ratio (or fold-of-enrichment). Finally, the odds-ratio is averaged across all query compounds (see S1 Text for a full description of how metrics are calculated). As the BBBC036 is distinct from JUMP CP, MoA enrichment evaluation measures performance on a fully held-out dataset (BBBC036), as well as the more in-distribution JUMP-CP dataset.

In addition to reporting the average fold-of-enrichment following [24], we reported the proportion of compounds with a significant enrichment (below a p-value of 0.05 for the permutation test used to calculate enrichment), as a more robust metric. Many MoAs have few compounds associated with them, so for individual MoAs, all compounds with the same MoA may be ranked in the top 1% of retrieved hits. This leads to an infinite odds ratio, which the authors deal with by imputing the size of the background set when this happens. However, this imputation strategy can significantly inflate the average fold-of-enrichment metric as the size of the background set (1534 for BBBC036 and 941 for JUMP-CP) is typically much higher than the typical finite fold-of-enrichment (on average 7–11 fold). This can lead to exaggerated differences between methods with minimal differences in behavior (e.g., if one method ranks all matching compounds within the top 1%, and another compound ranks all but one matching compound in the top 1%).

Implementation

Image preprocessing. To remove systematic variation in pixel intensity from uneven illumination in the field of view (FOV), we applied illumination correction [32]. Each image was corrected by dividing all pixel intensities by the illumination correction function, calculated by taking a median filter on the average image across all FOVs in a plate. Consistent with previous representation learning studies on cell painting images [24], to reduce the disk size of the dataset, images are compressed from 16-bit TIFF files to 8-bit PNG files, rescaling each channel independently between the 0.05th and 99.95th percentile of pixels intensity-wise. To further save disk space and improve I/O speed, we take a crop of 50% of the image height and width from the center of each image. For data augmentation and to allow large batch sizes, the processed images are further re-sized to 224x224 resolution during training. Following previous works [21,28,33], we also evaluated the effect of spherizing representations: we first rescaled features with mean absolute deviation (MAD) normalization, and then we applied spherizing, which transforms the feature space such that negative control representations have an identity covariate matrix.

Architecture and training details. We use ResNet101 [34], pre-trained on ImageNet, as our image encoder $\mathbf{h}_{\text{img}}$, following previous works showing that ImageNet features are effective at morphological profiling [23,35–37]. We replaced the final fully-connected layer with a new randomly initialized layer of hidden dimension size 256 and fine-tune only this new layer, with the pre-trained encoder frozen during training. Since models pre-trained on ImageNet typically expect 3-channel inputs, we preprocessed 5-channel microscopy images by converting each channel separately into a 3-channel image by duplicating the image 3 times along the channel axis. Each microscopy channel is separately preprocessed in this manner and inputted into the encoder to output an encoding of dimension 256. We concatenate the outputted encodings from the 5 original microscopy channels to obtain a final image representation with dimension $256 \times 5 = 1280$.

To encode chemical compounds, we represented them as ECFP4 fingerprints with a radius of 2 [38], which represents molecular structures as circular neighborhoods of atoms. The ECFP4 fingerprint transforms a molecular structure to a bit-vector embedding which captures patterns within a two-bond radius around each atom, enabling efficient comparison of molecular structures. The compound encoder, $\mathbf{h}_{\text{comp}}$, is trained from scratch. $\mathbf{h}_{\text{comp}}$ contains 4 fully-connected hidden layers of dimensionality 2048 each and interleaved with batch normalization and ReLU activations, receives as input the ECFP4 fingerprint, and outputs a molecular encoding of dimension 2048. We also tested starting from a pre-trained 3D molecular encoder, UniMol [39], but observed in initial experiments that this did not improve performance.

The fusion module accepts as input a concatenation of a negative control image embedding and a molecule embedding. The model is implemented as a two-layer multi-layer perceptron with leaky-ReLU activations. Finally, all image representations (i.e., both representations of real images and predicted representations) are passed through a projection head g, also implemented as a two-layer multi-layer perceptron with leaky-ReLu activations. Together, these components (the image encoder $\mathbf{h}_{\text{img}}$, the chemical compound encoder $\mathbf{h}_{\text{comp}}$, the fusion module F, and the projection head g) are combined into the overall MICON framework (Fig 1). The entire model is optimized end-to-end with the overall MICON loss, with the exception of the frozen components of the image encoder.

We trained using the AdamW optimizer [40] using a ReduceLROnPlateau learning rate scheduler using batch size 64 for 30 epochs on two NVIDIA A100-40G graphic cards. We oversampled negative controls to form half of the training batch for our POS-CTL models and one eighth of the training batch for our Target-2 models. We perform a linear warm-up for 2000 steps. All models are trained 3 times from random seeds. We saved checkpoints at every 2000 steps, and use the best-performing checkpoint on the validation dataset for testing. Detailed hyper-parameters and architectural details can be found in Table G in S2 Text.

Baselines and ablations. We compared MICON to the following baselines and ablations: hand-crafted, image-based features ("CellProfiler"); an image-only, self-supervised contrastive learning model ("SimCLR"); and a CLIP-based multi-modal model that aligns image and chemical compound representations ("CLIP").

CellProfiler [30] features – which are hand-designed, image-based texture, shape, and intensity features – were provided by and downloaded from the JUMP-CP dataset [25]. Models trained using SimCLR [10] provided an image-only, self-supervised contrastive learning baseline to validate our strategy of integrating chemical compound information in a multi-modal framework. To specifically evaluate if MICON treating perturbations as treatments was effective, we trained multi-modal baseline models using CLIP [20], which directly aligns image and chemical compound representations in a shared embedding space. A CLIP-based approach for multimodal learning between microscopy images and chemical compounds was previously proposed by CLOOME [18] and extended in MolPhenix [19]. As these works use different architectures, datasets, and pre-trained models, they do not provide appropriate baseline controls to MICON. We trained a CLIP baseline model using the same dataset and encoder architectures as MICON to enable direct comparison of the method itself, instead of transferring these models, which could result in disparities in performance from these other factors. Thus, we refer to the resulting baseline model as "CLIP."

Both the SimCLR and CLIP models use the same architectures, dataset, and stopping criterion as MICON listed in the Archiecture and Training Details, when applicable to the model (i.e., SimCLR and CLIP also use ResNet101 pre-trained on ImageNet, while CLIP also uses our ECFP4-based compound encoder trained from scratch). All models are trained 3 times with distinct random seeds. We also test features extracted from the ResNet101 pre-trained encoder that we initialized SimCLR and CLIP with, with no further training, to confirm that SimCLR and CLIP improve over this initialization.

In addition to SimCLR and CLIP, we also fine-tuned DINOv3 models on our datasets, a state-of-the-art self-supervised vision method. We initialize the VIT-Base model trained by [41], and fine-tuned on our dataset following their method, for 30 epochs with ReduceLROnPlateau learning rate scheduler starting with a learning rate of 10^{-3}. The rest of the optimizer settings are identical to the setup in Architecture and Training Details. We also test a DINOv3 model without fine-tuning to confirm that fine-tuning induces improvements in performance.

Finally, we extracted features from OpenPhenom-S/16 [14], a foundation model for cell painting trained using masked autoencoding,

Results

To validate MICON, we ran 8 experiments that reflect different challenges for morphological profiling, including integration across batches and sources as well as generalization to microscopy batch effects and perturbations unseen in training (Fig 4. We benchmarked MICON against baselines and ablated the PaCLR loss on the predicted images – and therefore information about chemical structures – to evaluate how integration of chemical compounds improves representation learning in morphological profiling. Our results demonstrate that our full MICON method using the PaCLR loss on the predicted images results in a sizable improvement over the CellProfiler, SIMCLR, and CLIP baselines. While performance improvements over using just the PaCLR loss alone were smaller, and thus not always statistically significant for individual experiments, pooling experiments indicated that the PaCLR loss on the predicted images significantly improves performance (F-ratio value of 9.249, p-value 0.0228 with a one-way repeated measures ANOVA). This observation suggests that modeling chemical compounds as treatments during training is effective at improving representations of microscopy images. We also observe that across all settings, all self-supervised learning methods improve over a pre-trained ResNet encoder on ImageNet, suggesting that some form of self-supervised fine-tuning on microscopy images is beneficial to performance regardless of method. In subsequent sections, we expand on the results of each experiment. We report 1-nearest neighbor matching results in the main manuscript, with top-3 and top-5 results are available in Tables H-J in S2 Text.

MICON improves representations of images

We first trained MICON and the baseline models on the ID (in-distribution) training dataset of the POS-CTL wells, on a total of 21,284 wells from 6 sources (Fig 2B). The positive control wells represent 8 compounds nominated to be morphologically distinct from DMSO and from each other [25]. We used this positive control dataset to measure if representations

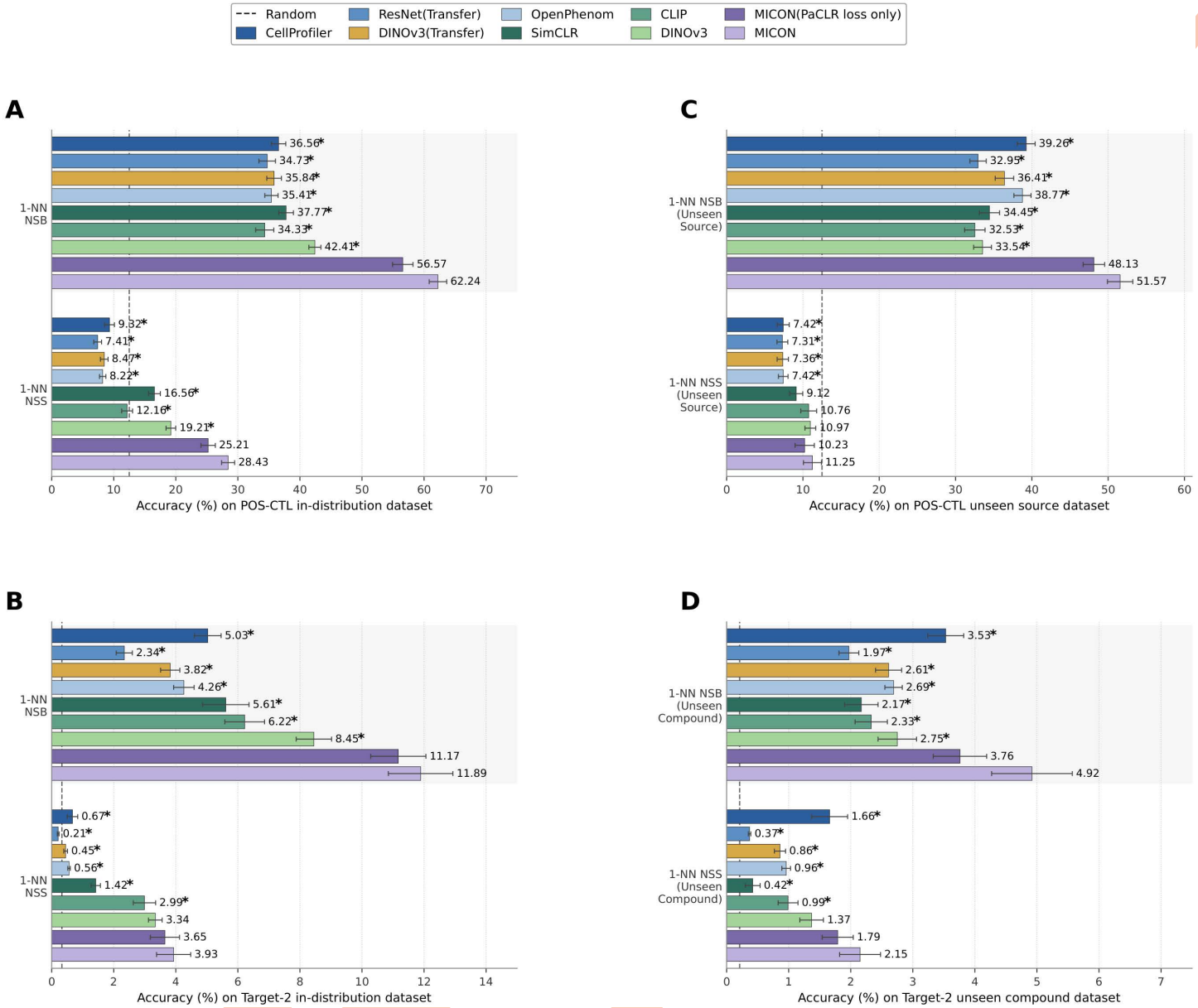

Fig 4. Compound-replicate matching accuracy for representation learning methods across evaluation settings. Retrieval accuracies on the **(A)** POS-CTL ID query dataset, **(B)** Target-2 ID query dataset, **(C)** POS-CTL OOD query dataset (unseen source), and **(D)** Target-2 OOD query dataset (unseen compounds). NSB designates retrieval of the nearest neighbor not in the same batch; NSS designates retrieval of the nearest neighbor not in the same source. $n = 3$ models trained and evaluated with different dataset stratifications and different random seeds; the mean accuracy is labeled, and the error bars represent standard deviation. Asterisk indicates that MICON significantly outperforms the baseline, defined as $p < 0.05$ on an unpaired one-tailed t-test with performance across the $n = 3$ random seeds as trials.

https://doi.org/10.1371/journal.pcbi.1014483.g004

can identify reproducible effects of chemical compounds that induce relatively strong phenotypic effects across increasing batch effects. Using the trained models, as well as CellProfiler features, we extracted representations for 9,078 unseen wells (Fig 2B; a subset of these representations are shown as a UMAP in S2 Fig. We then measured compound-replicate matching accuracy, specifically how frequently the nearest-retrieved NSB (not in same batch) and NSS (not in same

source) neighbor was treated with the same chemical compound (Fig 4A and Table A in S2 Text). CellProfiler features, the common-practice standard in the field, yielded a 1-NN NSB accuracy of 36.56% and a 1-NN NSS accuracy of 9.32% on the positive control test dataset, where expected accuracy of random guessing across 8 compounds is 12.5%. In contrast, MICON achieved a 1-NN NSB accuracy of 62.24% and a NSS accuracy of 28.43%, significantly outperforming both CellProfiler features as well SimCLR, (37.77% 1-NN NSB and 16.56% NSS) and CLIP, (34.33% 1-NN NSB and 12.16% NSS) (Fig 4A and Table A in S2 Text). DINOv3, as a more state-of-the-art self-supervised learning method, performs better than SimCLR and CLIP (42.41% 1-NN NSB and 19.21% NSS), but is still outperformed by MICON. Finally, we observe that OpenPhenom-S/16, a microscopy foundation model, generally performs comparable to or worse than CellProfiler (35.41% 1-NN NSB and 8.22% NSS).

We then asked whether these performance gains were due to MICON's multimodal integration of chemical structures or if they could be fully attributed to the PaCLR component, which operates exclusively on images. To test this, we ablated the part of the MICON loss on the predicted images, which is the only component that integrates chemical compounds, and trained a model using just the PaCLR loss. While this ablated model underperformed the full MICON model, it still outperformed the baselines, achieving a 1-NN NSB accuracy of 56.57% and a 1-NN NSS accuracy of 25.21% (Fig 4A and Table A in S2 Text). Together, these results suggest that, while the majority of MICON's strong performance is driven by the PaCLR loss, the integration of chemical compound information further improved the quality of its representations.

MICON outperforms baseline methods on unseen sources

We next sought to understand if MICON generalized to sources unseen in training. Many representation learning methods are designed to learn representations invariant to microscopy batch effects. This is true not just of MICON, whose contrastive method minimizes distances between images treated with the same perturbation across different sources, but also CLIP-based approaches [18,19], which map images treated with the same perturbation across sources to the same chemical compound representation. Hence, a major question with these representation learning methods is if these learned invariances generalize, or if they are overfit to sources seen during training.

To assess generalization to unseen sources, we designed an OOD (out-of-distribution) experiment using the POS-CTL wells, in which the training dataset consists of 5 seen sources and the query dataset consists wholly of a sixth source unseen during training (Fig 2C). For each representation learning method, we evaluated compound-replicate matching performance on the unseen source test dataset. We found that all representation learning methods dropped in performance when sources are unseen in training (Fig 4C and Table A in S2 Text). In the NSB setting, SimCLR (34.45%), CLIP (32.53%), and DINOv3 (33.54%) no longer outperformed CellProfiler (39.26%), suggesting that performance above CellProfiler may be due to overfitting to seen sources. However, MICON (51.57%) still significantly outperformed CellProfiler, SimCLR, and CLIP, indicating MICON's capacity to generalize to unseen sources. In the NSS setting, MICON's performance (11.25%) was not significantly better than that of SimCLR (9.14%), CLIP (10.76%), or DINOv3 (10.97%), although MICON was significantly better than CellProfiler (7.32%). We note that the unseen source NSS experiment is the only setting where all methods underperformed random guessing, suggesting that generalization to a source unseen during training remains an unsolved problem.

MICON demonstrates robust performance over larger compound datasets with unseen compounds

While our previous experiments focused on data representing just 8 chemical compound treatments, most morphological profiling experiments assay a greater number of perturbations. Therefore, we next sought to evaluate the performance of MICON when trained on datasets with a larger number of chemical compounds. We trained a MICON model on a total of 35,156 wells subsetted from the Target-2 plates, which image 301 perturbations replicated across all sources (Fig 3B; a subset of these representations are shown as a UMAP in S3 Fig).

As with our POS-CTL well experiments, we evaluated NSB and NSS compound-replicate matching accuracy on an unseen query dataset. Due to the greater number of chemical compounds, this task becomes much more challenging – the accuracy expected from random guessing is 0.33% (1 in 301). The CellProfiler, SimCLR, CLIP, DINOv3, and OpenPhenom-S/16 baselines exhibited performance ranging from 4.26 – 8.45% for NSB accuracy and from 0.21 – 3.34% for NSS accuracy, while MICON significantly improved on these baselines, providing 11.89% NSB accuracy and 3.93% NSS accuracy (Fig 4B). We show examples of the images retrieved by MICON versus CellProfiler in S4 Fig.

One limitation of our compound-replicate evaluation is that representation learning strategies that incorporate information about chemical compounds (i.e., both MICON and CLIP) may potentially overfit to compounds seen during training, which may result in a poorer representation of phenotypes induced by new, unseen compounds. To test this, we constructed an independent, OOD test dataset of wells treated with 184 compounds unseen during training (Fig 3C). We sampled these wells from the Compound plates of the JUMP-CP cpg0016 dataset and evaluated whether models trained on the Target-2 plates could still identify reproducible effects in images across these unseen compounds. To make this retrieval task more challenging, we allowed our nearest neighbor classifiers to retrieve for unseen compounds, compounds seen during training. We did this by curating a query set that only consists of the 184 seen compounds, but including both seen and unseen compounds in the retrieval dataset (for a total of 485 compounds, 301 seen and 184 unseen).

We observed that the baseline representation learning methods largely were not robust on compounds unseen in the training dataset, resulting in accuracies mostly well below that of CellProfiler features, despite being originally superior on seen compounds (Fig 4D). While CellProfiler features had an NSB accuracy of 3.53% and an NSS of 1.66%, SimCLR resulted in an NSB of 2.17% and an NSS of 0.72%, CLIP had an NSB of 2.33% and an NSS of 0.99%, OpenPhenom had a NSB of 2.69% and a NSS of 0.97%, and DINOv3 had an NSB of 2.75% and a NSS of 1.79%. In contrast, MICON performed significantly better than all baselines on unseen compounds, achieving an NSB of 4.92% and an NSS of 2.15% (Fig 4D and Table B in S2 Text). These results demonstrate that incorporating information about perturbations into representation learning yields improvements generalization to new, unseen perturbations. We note unlike the out-of-distribution experiments with POS-CTL wells (which held out sources from training), we perform above random guessing with unseen compounds, suggesting that generalizing to unseen compounds is less challenging than generalizing to unseen sources.

MICON benefits from microscopy batch effect correction

While our evaluations thus far used the raw representations from MICON and baselines, in practice, batch effect correction is often applied *post hoc* to representations from morphological profiling. One common transformation is to spherize representations for each plate such that the negative control representations have an identity covariate matrix [28,33]. MICON is designed to correct for batch effects through its training set-up, which enforces similarity between images with the same perturbation across different batches. We thus next sought to confirm that performance improvements from MICON were not because of trivial correction of batch effects, such that common adopted standards for batch effect correction could equalize the performance of baselines that do not inherently build in batch effect correction.

We assessed the impact of microscopy batch effect correction on MICON representations relative to the baseline methods by conducting compound-replicate matching, but with spherized representations, across the 8 evaluation settings (Fig 5 and Tables C-D in S2 Text). Spherizing substantially improved the performance of all representations evaluated. Interestingly, SimCLR, CLIP, OpenPhenom-S/16, and DINOv3 underperformed or performed similarly to CellProfiler in almost all experiments (except for the in-distribution Target-2 setting for DINOv3). These observations suggest that the previously observed improvements from the SimCLR, CLIP, OpenPhenom-S/16, and DINOv3 baselines without batch effect correction (Fig 4) may be driven by an inherent robustness to microscopy batch effects in the feature representations. Thus, this performance difference is eliminated when CellProfiler features are post-processed with batch effect correction (Fig 5). Critically, MICON representations still significantly outperformed all self-supervised representations in all experiments, and

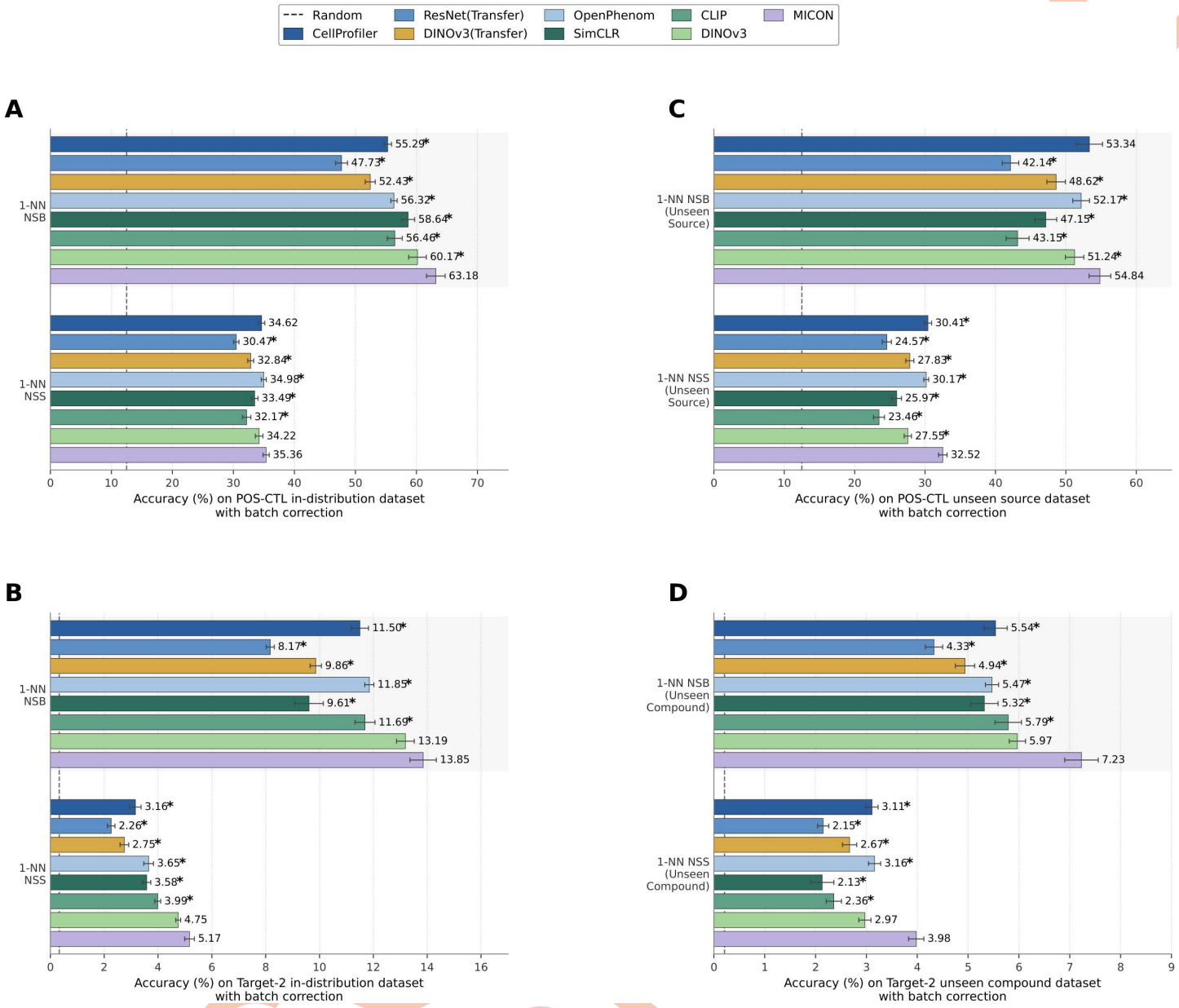

Fig 5. Compound-replicate matching accuracy for representation learning methods across evaluation settings, when microscopy batch effect correction is applied via spherizing. Retrieval accuracies on the **(A)** POS-CTL ID query dataset, **(B)** Target-2 ID query dataset, **(C)** POS-CTL OOD query dataset (unseen source), and **(D)** Target-2 OOD query dataset (unseen compounds). NSB designates retrieval of the nearest neighbor not in the same batch; NSS designates retrieval of the nearest neighbor not in the same source. $n = 3$ models trained and evaluated with different dataset stratifications and different random seeds; the mean accuracy is labeled, and the error bars represent standard deviation.. Asterisk indicates that MICON significantly outperforms the baseline, defined as $p < 0.05$ on an unpaired one-tailed t-test with performance across the $n = 3$ random seeds as trials.

https://doi.org/10.1371/journal.pcbi.1014483.g005

were significantly better or similar to CellProfiler features in all experiments (significantly better in 6 of 8 settings). These results suggest that MICON performance is driven by superior recognition of phenotypic effects, not just robustness to microscopy batch effects (Fig 5).

MICON achieves mechanism of action enrichment

Although compound-replicate matching tests if representations are able to identify reproducible biological signal, it does not assess whether compounds are grouped sensibly in representation space (i.e., compounds that induce biologically similar phenotypes clustering together). We thus evaluated our models on mechanism of action (MoA) enrichment [24], which tests if compounds with the same MoA are enriched in the top-ranked retrievals for a given query compound.

We evaluated MoA enrichment on a curated subset of the cpg0016 compound plates, using spherized features extracted by each of our trained models plus CellProfiler features (Table 1). For the trained models we used the top performing replicate of each model from the Target-2 OOD setting. In addition to average fold-of-enrichment, we also report the percent of compounds with significant enrichment. The latter metric is more robust to when there is an infinite odds-ratio for some compounds, which occurs sometimes as there are few MoAs matched for some query compounds. While previous benchmarks imputed the size of the background set when this occurs [24], this can lead to inflated differences in values when there are small differences in behavior of models.

We observe that MICON achieves a higher performance, both on average fold-of-enrichment and percent significant compounds than other methods, suggesting that the improvements in performance MICON achieved in compound-replicate retrieval also translate into better clustering of compounds with similar mechanisms of action. All self-supervised models (SimCLR, CLIP, DINOv3, and MICON) outperformed CellProfiler features by a significant margin by both metrics, unlike our compound-replicate matching experiments. This may be because unlike compound-replicate retrieval, MoA enrichment metrics allow for retrieval of compounds from the same batch, making it less challenging of a batch effect setting.

Next, we sought to similarly investigate MoA enrichment on the out-of-distribution BBBC036 dataset (Table 2). Conversely, besides outperforming CellProfiler, all self-supervised methods performed comparably. Although SimCLR exhibited slightly higher performance on average fold-of-enrichment than the other methods, it had the same percentage of significant compounds to the other methods, suggesting this performance may be due to inflation from imputation. Supporting the observation that all self-supervised learning methods perform similarly, we examined the overlap in significant compounds between methods (S5 Fig). All self-supervised methods retrieved compounds with the same MoAs for almost the same compounds. CellProfiler retrieved same-MoA compounds for a set of query compounds distinct from any self-supervised method. Overall, this suggests that MICON's improvements in retrieving biological signal is more restricted to in-distribution datasets.

MICON can synthesize phenotypic effects of unseen compounds

The PaCLR loss on the predicted images means that MICON models should additionally be able to predict the phenotypic effect of compound perturbations. As a proof-of-concept test of this capability, we assessed how frequently the predicted representations from MICON models could retrieve a representation from a real image of cells perturbed with the same

Table 1. Mechanism of action enrichment for the cpg0016 compound dataset. We report the average fold-of-enrichment with imputation (imputing the size of the background set as in [24]) and the percent of significant compounds.

	Average Fold-of-Enrichment	Percent Significant Compounds
CellProfiler	20.69	11.45%
SimCLR	21.07	14.39%
CLIP	27.47	14.60%
DINOv3	35.39	15.97%
MICON	37.52	16.18%

https://doi.org/10.1371/journal.pcbi.1014483.t001

Table 2. Mechanism of action enrichment for the BBBC036 dataset. We report the average fold-of-enrichment with imputation (imputing the size of the background set as in [24]) and the percent of significant compounds.

	Average Fold-of-Enrichment	Percent Significant Compounds
CellProfiler	25.62	9.20%
SimCLR	32.83	11.23%
CLIP	31.14	11.59%
DINOv3	31.23	11.23%
MICON	31.26	11.23%

https://doi.org/10.1371/journal.pcbi.1014483.t002

Table 3. Nearest neighbor retrieval accuracies (%) using real perturbation representations vs. generated counterfactual representations, across evaluation settings. NSB designates retrieval of the nearest neighbor not in the same batch; NSS designates retrieval of the nearest neighbor not in the same source. Error represents the standard deviation across 3 independent replicates.

POS-CTL	ID NSB	ID NSS	Unseen Sources NSB	Unseen Sources NSS
Random	12.50			
MICON (Real Images)	63.18 ± 3.53	35.36 ± 3.21	54.84 ± 1.94	32.52 ± 2.32
MICON (Generated)	54.75 ± 3.21	26.14 ± 3.37	47.26 ± 3.46	25.93 ± 2.02
TGT-2	ID NSB	ID NSS	Unseen Compounds NSB	Unseen Compounds NSS
Random	0.33			
MICON (Real Images)	13.85 ± 2.79	5.17 ± 1.28	7.23 ± 1.73	3.98 ± 0.67
MICON (Generated)	13.64 ± 2.03	4.76 ± 0.93	2.54 ± 0.58	1.07 ± 0.21

https://doi.org/10.1371/journal.pcbi.1014483.t003

compound. We replicated each of the previous retrieval experiments, except using predicted representations instead of real ones. For each real perturbation representation originally used in compound-replicate matching, we replaced that representation with a predicted representation, using the spatially-nearest negative control as the basis for generation. This set-up enabled direct comparison of retrieval performance with predicted representations versus real representations.

We observed that, for seen compounds, retrieval with the predicted representations only slightly declined in performance compared to with the real perturbation representations (Table 3), even when retrieving across sources and with unseen sources. For compounds unseen in training, the decline in performance was more substantial, but, notably, performance with predicted representations was still several folds higher than the accuracy from random guessing. These results suggest that MICON's predicted representations capture phenotypic effects of compounds seen in training, with some ability to capture the phenotypic impact of completely unseen compounds. However, we caveat that the overall performance is low especially for unseen compounds, suggesting that MICON is likely not useful for prediction yet in its current state.

Discussion

In this work, we introduced MICON, a molecular-image representation learning method for morphological profiling. MICON integrates two losses during training using contrastive learning: a perturbation-aware contrastive (PaCLR) loss that encourages representations of real images to be invariant to microscopy batch effects, and a predicted PaCLR loss that aligns representations of synthetic images synthesized using chemical compound information with those of real images of cells treated with the compound. Both components substantially advance representation learning. Even stand-alone, MICON's PaCLR loss enables identification of reproducible phenotypic effects of compounds under batch effects, achieving improvements over both CellProfiler features and state-of-the-art representation learning methods in computer vision.

We attribute this to the use of perturbation and microscopy batch metadata to mine both positive and negative samples for representation learning.

We showed that MICON'sPaCLR loss on the predicted images further improves performance in identifying phenotypic effects in images relative to models trained with just the PaCLR loss. This result demonstrates that integrating multimodal information about perturbation treatments can improve uni-modal representations of images, establishing directions for future morphological profiling methods. We also demonstrated that *how* images and chemical compounds should be integrated is crucial. MICON models chemical compounds as a transformation of images, and this approach strongly outperforms the CLIP-like strategy of directly aligning images and chemical compounds in a shared embedding space.

Our evaluation also sheds novel insights into representation learning for cell painting. Taking advantage of the large-scale and multi-source nature of the JUMP-CP dataset, we designed evaluation splits that allow us to assess if morphological profiling methods can maintain performance in realistic application scenarios. Specifically, we evaluated performance when comparing images from different data-generating sources, and when sources or compounds are unseen during representation learning (Figs 2 and 3). While evaluations that stratify across sources have been designed to benchmark batch effect correction methods [21], our study systematically benchmarked different representation learning methods in a controlled way, using the same architectures and training data. Previous evaluations of representation learning typically only use datasets collected by a single source and do not stratify compounds/sources such that some are unseen during training [7,16,18]. Here, our benchmarks exposed that some representation learning methods previously reported to be effective for morphological profiling may be overfitting and thus overestimating their performance. For example, while SimCLR outperforms CellProfiler features when representing seen compounds, it underperforms CellProfiler on compounds unseen during training. In contrast, MICON is robust and demonstrates consistent improvements over both CellProfiler and baseline representation learning methods across all scenarios evaluated. MICON's improvements hold even when compounds are unseen in training and even when considering more challenging batch effects like differing imaging sources.

Future work should explore if MICON stays robust when deployed on larger and less curated datasets. In this work, we focused on a small number of perturbations that were more likely to have strong phenotypic effects. However, drug screening experiments are usually systematic in nature and will thus include many compounds that elicit minimal phenotypic impact on cells relative to negative controls. This impacts both training and evaluation. In training, as MICON's PaCLR loss forces the model to distinguish compound-treated images from negative controls, it is not clear if scaling the number of compounds in the training dataset will adversely affect learning. Future work should analyze the trade-off between indiscriminately scaling the number of images and/or compounds versus downsampling datasets to just the stronger phenotypic effects. For evaluation, curating compounds that have minimal but reproducible effects without risking selecting compounds that have no reproducible phenotypic impacts on cells is challenging, but we think evaluating the sensitivity of models to more subtle phenotypes is crucial in future work. In this work, we focused on the POS-CTL wells and the Target-2 plates of cpg0016, which are reproduced multiple times by multiple sources. In contrast, some compounds in cpg0016 are only assayed by a subset of sources and once by each source. This reflects another trade-off in training data that should be explored in future work. Namely, it is unclear if focusing training on a small subset of well-reproduced compounds can explain some of the effectiveness of MICON or if expanding the number of compounds at the cost of less coverage across sources would further improve our method. Finally, there are opportunities in interpretation that we have not explored in this work: future work should investigate if the compound encoder attends to parts of chemical compounds that are responsible for biological mechanism of action.

While we demonstrated that MICON is capable of synthesizing the phenotypic effects of some unseen compounds as a consequence of its pretraining task, our evaluation was an initial proof-of-concept to determine if any predictive signal was present. It is unclear the effectiveness of this strategy compared to other proposals like CLIP-based strategies [18,19], and future work should do a comparative analysis. A more thorough exploration of hyperparameters could further advance

our model's capabilities. In our experiments, we equally weighted the PaCLR losses on predicted versus real images. We demonstrated that although we have some signal in synthesizing the phenotypic effects of unseen compounds, our performance is weak, and one possibility is that in study, we dedicated too little of our loss to generating predictions. Increasing the weight on this loss, either throughout training or during later epochs, may enhance performance in producing synthetic image representations.

Our current proposed model is restricted to chemical compounds as input perturbations. Extending MICON to additional perturbation modalities, such as genetic or environmental perturbations, could similarly model them as treatments that induce a transformation of images. MICON's framework enables prediction of representations and, hence, of the phenotypic effects of unseen compounds. While optimizing this capability is beyond the scope of our work, the ability to predict the phenotypic effects of perturbations prior to experimentation could make high-throughput drug screens more resource efficient by triaging compounds that are more likely to be effective. Towards this capability, our proof-of-concept demonstrates that generation is possible directly in a learned latent space, paralleling similar work with conditional generative models [42] that produce synthetic images that reflect perturbation phenotypes. In sum, through multimodal modeling of perturbations as treatments on images, MICON provides a new direction for representation learning in morphological profiling, paving the way for new methods that explicitly leverage the multimodal nature of microscopy screens.

Supporting information

S1 Text. Supplementary methods. Supplementary methods describing how permutation tests for mechanism of action enrichment were conducted.
(PDF)

S2 Text. Supplementary tables. Supplementary tables A-G.
(PDF)

S1 Fig. Heat map showing cosine similarity between aggregated well-level profiles of CellProfiler features across negative control wells for each batch. Batches (x and y-axes) are labeled by their source in CP JUMP.
(PNG)

S2 Fig. Image representations from CellProfiler, SimCLR, CLIP, and MICON for the POS-CTL dataset, reduced using UMAP and visualized as a 2D scatterplot. For each representation method, the top plot is colored by source, and the bottom plot is colored by perturbation, to visualize robustness to batch effects and clustering of biological phenotype, respectively. $n = 2{,}000$ randomly sampled images are shown.
(PNG)

S3 Fig. Image representations from CellProfiler, SimCLR, CLIP, and MICON for a subset of the Target-2 dataset, reduced using UMAP and visualized as a 2D scatterplot. For each representation method, the top plot is colored by source, and the bottom plot is colored by perturbation, to visualize robustness to batch effects and clustering of biological phenotype, respectively. 10 compounds with accuracy higher than 0% across all methods are randomly sampled to avoid cluttering the visualization with too many labels. $n = 2{,}000$ randomly sampled images are shown.
(PNG)

S4 Fig. NSB/NSS retrieval results using images sampled from the Target-2 dataset. Representations of the query image (first column) are used to retrieve other images in the dataset. Negative control shows a negative control from the same batch. CellProfiler-NSB and CellProfiler-NSS designate the nearest neighbor of the query image in CellProfiler features in a different batch and source respectively. MICON-NSB and MICON-NSS are these retrievals but using MICON features, while MICON-Gen-NSB and MICON-Gen-NSS are retrievals using predicted representations. Blue frame images

are treated with the same perturbation and orange frames indicate the treatments are from different perturbations/negative control. Channels are colored so that DNA is shown in blue, ER is shown in green, AGP (Actin, Golgi, Plasma Membrane) in red, mitochondria in magneta, and RNA in cyan. Channels are rescaled to the full intensity range of the image. A representative crop is taken for each image.
(PNG)

S5 Fig. Venn diagram showing overlap in compounds significantly enriched in retrieving compounds with the same mechanism of action for the BBBC036 dataset across methods.
(PNG)

Acknowledgments

We thank Sean Whitzell for assistance in dataset storage and management.

Author contributions

Conceptualization: Neil Tenenholtz, Lester Mackey, David Alvarez-Melis, Ava P. Amini, Alex Lu.

Data curation: Emre Hayir, Alex Lu.

Formal analysis: Yemin Yu, Emre Hayir.

Investigation: Yemin Yu, Emre Hayir.

Methodology: Yemin Yu, Emre Hayir, David Alvarez-Melis, Ava P. Amini, Alex Lu.

Project administration: Ying Wei, Ava P. Amini, Alex Lu.

Software: Yemin Yu, Emre Hayir.

Supervision: Neil Tenenholtz, Lester Mackey, Ying Wei, David Alvarez-Melis, Ava P. Amini, Alex Lu.

Validation: Yemin Yu.

Writing – original draft: Yemin Yu, Alex Lu.

Writing – review & editing: Yemin Yu, Emre Hayir, Neil Tenenholtz, Lester Mackey, Ying Wei, David Alvarez-Melis, Ava P. Amini, Alex Lu.

References

1. Lin S, Schorpp K, Rothenaigner I, Hadian K. Image-based high-content screening in drug discovery. Drug Discov Today. 2020;25(8):1348–61. https://doi.org/10.1016/j.drudis.2020.06.001 PMID: 32561299
2. Ziegler S, Sievers S, Waldmann H. Morphological profiling of small molecules. Cell Chem Biol. 2021;28(3):300–19. https://doi.org/10.1016/j.chembiol.2021.02.012 PMID: 33740434
3. Bray M-A, Singh S, Han H, Davis CT, Borgeson B, Hartland C, et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc. 2016;11(9):1757–74. https://doi.org/10.1038/nprot.2016.105 PMID: 27560178
4. Caicedo JC, Cooper S, Heigwer F, Warchal S, Qiu P, Molnar C, et al. Data-analysis strategies for image-based cell profiling. Nat Methods. 2017;14(9):849–63. https://doi.org/10.1038/nmeth.4397 PMID: 28858338
5. Pratapa A, Doron M, Caicedo JC. Image-based cell phenotyping with deep learning. Curr Opin Chem Biol. 2021;65:9–17. https://doi.org/10.1016/j.cbpa.2021.04.001 PMID: 34023800
6. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 2006;7(10):R100. https://doi.org/10.1186/gb-2006-7-10-r100 PMID: 17076895
7. Tang Q, Ratnayake R, Seabra G, Jiang Z, Fang R, Cui L, et al. Morphological profiling for drug discovery in the era of deep learning. Brief Bioinform. 2024;25(4):bbae284. https://doi.org/10.1093/bib/bbae284 PMID: 38886164
8. Rani V, Nabi ST, Kumar M, Mittal A, Kumar K. Self-supervised Learning: A Succinct Review. Arch Comput Methods Eng. 2023;30(4):2761–75. https://doi.org/10.1007/s11831-023-09884-2 PMID: 36713767

9. Kim V, Adaloglou N, Osterland M, Morelli FM, Halawa M, König T, et al. Self-supervision advances morphological profiling by unlocking powerful image representations. BioRxiv. 2023;:2023–04.
10. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International conference on machine learning, 2020. 1597–607.
11. Perakis A, Gorji A, Jain S, Chaitanya K, Rizza S, Konukoglu E. 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, 2021. 565–75.
12. Lin A, Lu A. Incorporating knowledge of plates in batch normalization improves generalization of deep learning for microscopy images. Machine Learning in Computational Biology. PMLR. 2022. 74–93.
13. He K, Chen X, Xie S, Li Y, Dollar P, Girshick R. Masked Autoencoders Are Scalable Vision Learners. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 15979–88. https://doi.org/10.1109/cvpr52688.2022.01553
14. Kraus O, Kenyon-Dean K, Saberian S, Fallah M, McLean P, Leung J, et al. Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. 11757–68. https://doi.org/10.1109/cvpr52733.2024.01117
15. Caron M, Touvron H, Misra I, Jegou H, Mairal J, Bojanowski P, et al. Emerging Properties in Self-Supervised Vision Transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 9630–40. https://doi.org/10.1109/iccv48922.2021.00951
16. Doron M, Moutakanni T, Chen ZS, Moshkov N, Caron M, Touvron H, et al. Unbiased single-cell morphology with self-supervised vision transformers. bioRxiv. 2023.
17. Caicedo JC, McQuin C, Goodman A, Singh S, Carpenter AE. Weakly Supervised Learning of Single-Cell Feature Embeddings. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2018;2018:9309–18. https://doi.org/10.1109/CVPR.2018.00970 PMID: 30918435
18. Sanchez-Fernandez A, Rumetshofer E, Hochreiter S, Klambauer G. CLOOME: contrastive learning unlocks bioimaging databases for queries with chemical structures. Nat Commun. 2023;14(1):7339. https://doi.org/10.1038/s41467-023-42328-w PMID: 37957207
19. Fradkin P, Azadi P, Suri K, Wenkel F, Bashashati A, Sypetkowski M. How Molecules Impact Cells: Unlocking Contrastive PhenoMolecular Retrieval. In: 2024. https://doi.org/arXiv:240908302
20. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S. Learning transferable visual models from natural language supervision. In: International conference on machine learning, 2021. 8748–63.
21. Arevalo J, Su E, Ewald JD, van Dijk R, Carpenter AE, Singh S. Evaluating batch correction methods for image-based cell profiling. Nat Commun. 2024;15(1):6516. https://doi.org/10.1038/s41467-024-50613-5 PMID: 39095341
22. Yao H, Hanslovsky P, Huetter J-C, Hoeckendorf B, Richmond D. Weakly Supervised Set-Consistency Learning Improves Morphological Profiling of Single-Cell Images. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024. 6978–87. https://doi.org/10.1109/cvprw63382.2024.00691
23. Cross-Zamirski JO, Williams G, Mouchet E, Schönlieb CB, Turkki R, Wang Y. Self-supervised learning of phenotypic representations from cell images with weak labels. 2022. https://arxiv.org/abs/2209.07819
24. Moshkov N, Bornholdt M, Benoit S, Smith M, McQuin C, Goodman A, et al. Learning representations for image-based profiling of perturbations. Nat Commun. 2024;15(1):1594. https://doi.org/10.1038/s41467-024-45999-1 PMID: 38383513
25. Chandrasekaran SN, Ackerman J, Alix E, Ando DM, Arevalo J, Bennion M. JUMP Cell Painting dataset: morphological impact of 136,000 chemical and genetic perturbations. bioRxiv. 2023;:2023–03.
26. Chen ZS, Pham C, Wang S, Doron M, Moshkov N, Plummer B, et al. CHAMMI: A benchmark for channel-adaptive models in microscopy imaging. In: Advances in Neural Information Processing Systems 36, 2023. 19700–13. https://doi.org/10.52202/075280-0865
27. Ljosa V, Caie PD, Ter Horst R, Sokolnicki KL, Jenkins EL, Daya S, et al. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. J Biomol Screen. 2013;18(10):1321–9. https://doi.org/10.1177/1087057113503553 PMID: 24045582
28. Ando DM, McLean CY, Berndl M. Improving phenotypic measurements in high-content imaging screens. BioRxiv. 2017;:161422.
29. Janssens R, Zhang X, Kauffmann A, de Weck A, Durand EY. Fully unsupervised deep mode of action learning for phenotyping high-content cellular images. Bioinformatics. 2021;37(23):4548–55. https://doi.org/10.1093/bioinformatics/btab497 PMID: 34240099
30. McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW. CellProfiler 3.0: Next-generation image processing for biology. PLoS Biology. 2018;16(7):e2005970.
31. Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, et al. The Drug Repurposing Hub: a next-generation drug library and information resource. Nat Med. 2017;23(4):405–8. https://doi.org/10.1038/nm.4306 PMID: 28388612
32. Singh S, Bray M-A, Jones TR, Carpenter AE. Pipeline for illumination correction of images for high-throughput microscopy. J Microsc. 2014;256(3):231–6. https://doi.org/10.1111/jmi.12178 PMID: 25228240
33. Way GP, Natoli T, Adeboye A, Litichevskiy L, Yang A, Lu X, et al. Morphology and gene expression profiling provide complementary information for mapping cell state. Cell Syst. 2022;13(11):911-923.e9. https://doi.org/10.1016/j.cels.2022.10.001 PMID: 36395727
34. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770–8. https://doi.org/10.1109/cvpr.2016.90

35. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics. 2016;32(12):i52–9. https://doi.org/10.1093/bioinformatics/btw252 PMID: 27307644
36. Pawlowski N, Caicedo JC, Singh S, Carpenter AE, Storkey A. Automating morphological profiling with generic deep convolutional networks. BioRxiv. 2016;:085118.
37. Godinez WJ, Hossain I, Lazic SE, Davies JW, Zhang X. A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics. 2017;33(13):2010–9. https://doi.org/10.1093/bioinformatics/btx069 PMID: 28203779
38. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–54. https://doi.org/10.1021/ci100050t PMID: 20426451
39. Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z. Uni-mol: A universal 3d molecular representation learning framework. ChemRxiv. 2023.
40. Kingma DP. Adam: A method for stochastic optimization. 2014. https://arxiv.org/abs/1412.6980
41. Siméoni O, Vo HV, Seitzer M, Baldassarre F, Oquab M, Jose C. Dinov3. In: 2025. https://doi.org/arXiv:250810104
42. Palma A, Theis FJ, Lotfollahi M. Predicting cell morphological responses to perturbations using generative modeling. bioRxiv. 2023;:2023–07.