Metadata-guided Consistency Learning for High Content Images

Johan Fredin Haslum [1,2,4]   JHASLUM@KTH.SE

Christos Matsoukas [1,2,5]   MATSOU@KTH.SE

Karl-Johan Leuchowius [3]   KARL-JOHAN.LEUCHOWIUS@ASTRAZENECA.COM

Erik Müllers [4]   ERIK.MUELLERS@ASTRAZENECA.COM

Kevin Smith [1,2]   KSMITH@KTH.SE

[1] *KTH Royal Institute of Technology, Stockholm, Sweden*

[2] *Science for Life Laboratory, Stockholm, Sweden*

[3] *Mechanistic and Structural Biology, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden*

[4] *Cell & Molecular Pharmacology, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D, AstraZeneca, Gothenburg, Sweden*

[5] *Cardiovascular, Renal and Metabolism Pathology, Clinical Pharmacology & Safety Sciences, R&D AstraZeneca, Gothenburg, Sweden*

Abstract

High content imaging assays can capture rich phenotypic response data for large sets of compound treatments, aiding in the characterization and discovery of novel drugs. However, extracting representative features from high content images that can capture subtle nuances in phenotypes remains challenging. The lack of high-quality labels makes it difficult to achieve satisfactory results with supervised deep learning. Self-Supervised learning methods, which learn from automatically generated labels has shown great success on natural images, offer an attractive alternative also to microscopy images. However, we find that self-supervised learning techniques underperform on high content imaging assays. One challenge is the undesirable domain shifts present in the data known as *batch effects*, which may be caused by biological noise or uncontrolled experimental conditions. To this end, we introduce Cross-Domain Consistency Learning (CDCL), a novel approach that is able to learn in the presence of batch effects. CDCL enforces the learning of biological similarities while disregarding undesirable batch-specific signals, which leads to more useful and versatile representations. These features are organised according to their morphological changes and are more useful for downstream tasks – such as distinguishing treatments and mode of action.

Keywords: Representational Learning, Domain Shifts, Cell Painting, Fluorescence Microscopy

1. Introduction

High content screening (HCS), or image-based screening of cells treated with potential drug-molecules have become an important tool for pharmaceutical drug discovery and development. Image-based screens enable us to probe, at scale, large panels of compounds and other perturbents. These screens capture rich phenotypes of complex sub-cellular processes and tie them to a multitude of endpoints, such as mode of action, bioactivity and toxicity. Large scale screening efforts of standardized HCS assays such as Cell Painting (Bray et al.,

2016) in the JUMP-CP initiative, have generated phenotypic response data for hundreds of thousands of treatments. However the extent to which, and how, this valuable information can be extracted from cell culture images using machine learning is still an open question – *what is the best way to learn useful representations for fluorescence microscopy data?*

Feature-based extraction methods, like Cell Profiler (Carpenter et al., 2006), have been utilised throughout the years, but recent studies indicate that more information can be uncovered using deep learning (DL) methods (Hofmarcher et al., 2019; Moshkov et al., 2022). However, applying supervised learning methods on high content imaging data is challenging for many reasons.

Labeled data is rarely available in large quantities, since characterizing each treatment for different endpoints can be both challenging as well as prohibitively expensive at scale. And when it is available, it is often noisy and sparse and tailored to specific experimental settings, limiting the applicability for other purposes and downstream tasks.

Self-Supervised Learning Algorithms (SSL) seem to be appealing alternatives for high content imaging assays. SSL approaches use various techniques to circumvent the need for labeled samples, and can in the process potentially mitigate the effects of noisy or erroneous annotations by simply not using any labels. Recent advances in SSL in the natural imaging domain have been shown to generate more useful features for downstream tasks than their supervised counterparts (Caron et al., 2021).

However, self-supervised learning methods come at a cost – they are more data-hungry than their supervised counterparts. Fortunately, although high content screens may suffer from noisy or sparse labels, they do contain plentiful amounts of raw image data. Therefore, HCS would appear to be an excellent candidate application for self-supervised learning.

In this work, we find that self-supervised learning techniques do not exhibit the same successes for HCS as they do for the natural image domain. The reason for this failure is an issue at the heart of most high content screening datasets – *batch effects*[1]. Batch effects are a well-known problem in HCS referring to undesirable domain shifts present in the data caused by biological noise or difficult-to-control experimental conditions. They are confounding factors that can dominate the signal, *e.g.* minute variations in the fluorescent stains used or the length of incubation times can lead to differences in the acquired images., even when great efforts are undertaken to keep experimental conditions consistent. These confounding factors are related to the batch the data was generated in – not the biological process of interest. As such, they are problematic for SSL, which tends to be misled into learning features dominated by batch-related factors, rendering them of little use for predictions and downstream tasks. Whereas strong phenotypic effects could still be captured across batches (such as strong cellular toxicity), more subtle effects are typically lost.

To counteract this, we propose a modification to current self-supervised learning methods that we call Cross-Domain Consistency Learning (CDCL).

The main idea behind CDCL is to utilize readily-available metadata to enforce consistent representations of treatments across batches during self-supervised learning. This forces the model to focus on the biological signal of relevance while disregarding confounding factors related to batch effects. We demonstrate the effectiveness of the proposed approach when

---

1. Note that the term "batch" in the context of HCS refers to a group of experiments generated under similar conditions, as opposed to the samples presented to the model for learning at each iteration in the ML context. To avoid confusion we use the term "mini-batch" when we refer to the latter.

combined with SSL methods in two fluorescence microscopy datasets. Further, comparisons show that our method results in versatile and biologically meaningful features – better suited for down-stream analysis. Our contributions are summarised as follows:

- We showcase the limitations of current SSL methods in high content imaging attributed to batch effects.

- We propose CDCL, a method that counteracts the shortcomings for this type of data – allowing the application of SSL methods for high content screening.

- We demonstrate that the representations learned using our method are more relevant for downstream tasks.

These findings, along with additional ablation studies, suggest that SSL-inspired methods can indeed be utilised for sparse and noisy high content screening, opening the door to benefit from deep learning where it was previously difficult. The code to reproduce our work will be released soon.

2. Related Work

Feature extraction from high content imaging microscopy data has historically been done using software like Cell Profiler (Carpenter et al., 2006), relying on single cell segmentation followed by human-crafted feature extraction. This approach has proven useful for down stream analysis task such as mode of action identification, bioactivity- and toxicity-prediction (Simm et al., 2018) (Warchal et al., 2019) (Trapotsi et al., 2022). More recent works have shown the benefits of deep learning-based feature extraction (Moshkov et al., 2022).

Deep learning based features have been used both directly from ImageNet trained weights (Moshkov et al., 2022) as well as models trained for the particular task at hand (Hofmarcher et al., 2019). Task-specific models have shown superior performance. However, training task-specific models has proven challenging due to the lack of reliable task specific labels (Caicedo et al., 2018). Therefore, many have turned to using readily available metadata as training labels, using information such as treatment as "weak"-labels. This approach has shown slight performance benefits for training feature extractors (Caicedo et al., 2018; Moshkov et al., 2022).

Self-supervised learning methods use various techniques to circumvent the need for labeled samples. SSL approaches have made significant improvements in recent years and are closing the performance gap to supervised learning on commonly used datasets such as ImageNet (Grill et al., 2020; Caron et al., 2021). At the core of the most recent SSL approaches is the use of consistency learning between positive pairs, enforcing similar output distributions (Chen and He, 2021). Contrary to previous similarity based approaches, these methods do not require negative examples, thus separating them from pure contrastive methods. This also means that they do not require large batches of negative samples or hard negative mining in order to balance the loss (Chen et al., 2020). They can be applied more easily and have been shown empirically to produce state-of-the-art results (Chen and He, 2021; Caron et al., 2021; Grill et al., 2020).

Contrastive learning in high content imaging – previous works have explored the use of contrastive learning based approaches in HCI, with metric learning approaches using contrastive losses being used both in supervised (Ando et al., 2017) as well as weakly supervised manners (Caicedo et al., 2018). This was followed by more recent work exploring contrastive self-supervision methods (Janssens et al., 2021; Perakis et al., 2021). Beyond this other methods such as Archetypal Analysis (Siegismund et al., 2021) have been explored for representational learning. However, all of these works primarily explore small datasets with few samples, treatment and task specific classes in a translucent setting, thus avoiding working on problems associated with batch effects. Further, none of the recent successful SSL methods that enforce consistency instead of contrastive losses have yet been explored in this context. Therefore we explore the use of newer SSL methods in HCS datasets, focusing on their ability to learn good representation in larger and more challenging datasets.

3. Methods & Experimental Setup

We start by investigating the extent to which recent SSL-based methods can be utilized for representational learning in high content imaging datasets. We introduce, as baselines, methods to learn representations using weakly-supervised learning of treatments. We then introduce two recent state-of-the-art consistency based SSL methods: BYOL and DINO. In Section 4, we show that current SSL methods fail to learn useful biological features and we pinpoint the cause of it – batch effects. Finally, we propose a method to counteract the domain shifts caused by batch effects, which we call Cross-Domain Consistency Learning (CDCL).

3.1. Baselines

High content screenings are rarely accompanied with large quantities of task-specific labels, which limits the applicability of supervised learning. This lack of labels can be attributed to the challenging nature of defining ground truth labels for each treatment and doing so at scale. However, metadata such as compound and concentration information is commonly available for each setting. Such data is therefore often used as "weak" labels (Caicedo et al., 2018; Moshkov et al., 2022).

Effectively considering each treatment a unique class and assuming that each treatment generates a different phenotypic response. We use the "weakly"-supervised method as the baselines for this work, where the prediction of treatments is the primary task, and predicting mode of action is a down-stream task (when such information is available). Both tasks are trained using a cross-entropy loss in a supervised learning setup.

3.2. Self-Supervised Learning

We consider two of the most successful consistency-based SSL methods, DINO (Caron et al., 2021) and BYOL (Grill et al., 2020), which we briefly describe below.

Both methods use a teacher-student setup to learn a useful feature extractor. A feature extractor is a network intended for other down-stream tasks, typically one of the mainstream architectures *e.g.* DEIT (Touvron et al., 2021) or RESNET (He et al., 2016). The teacher and student network are both presented views of the same image – the student network
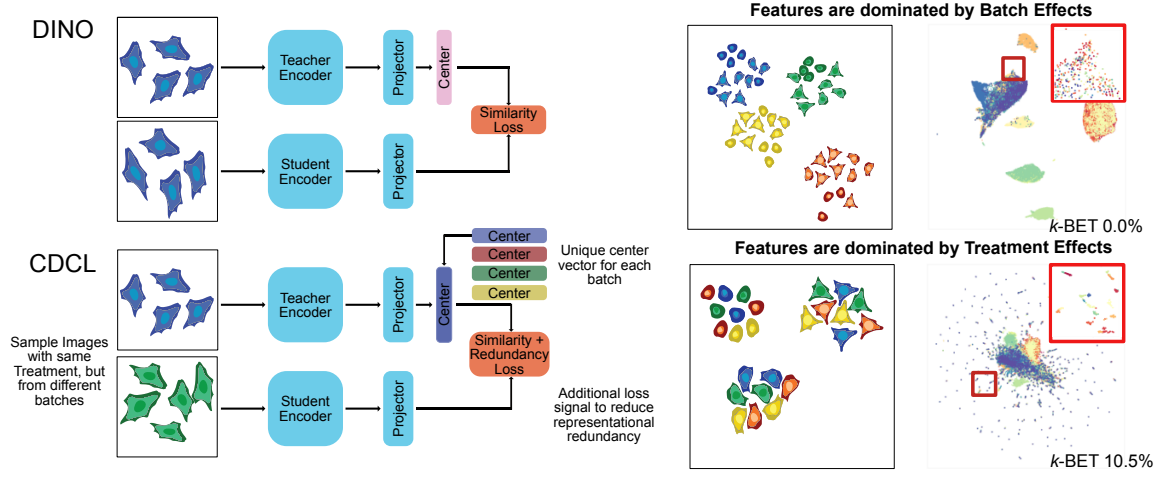
Figure 1: *Our approach.* The upper-left diagram depicts an SSL method, in this case DINO, applied to HCS data. The illustration in the upper-right shows feature embeddings that typically result from DINO – dominated by batch effects. The lower diagrams depict our approach, CDCL, and the resulting feature embeddings which correctly capture treatment effects and ignore batch effects.

receives different augmented views of the image presented to the teacher. Both consist of a feature extractor followed by projection heads implemented through MLP layers. The teacher network f_θ^T does not back-propagate any learning signal. It gets updated using the exponential moving average of the student's weights after each iteration. The student network f_θ^S is trained to mimic the teacher network, when feed different augmented views of the same image.

In DINO, the student f_θ^S and the teacher f_θ^T share the same architecture – a feature extractor followed by a 3-layer MLP layer θ_{Proj}. Given two views of the same image, x_1 and x_2, the model f_θ^S is trained to minimize cross-entropy loss of the output probability of the student $P_S = softmax(f_\theta^S(x_1))$ and the teacher $P_T = softmax(f_\theta^T(x_2) - C)$. C is a centering vector, defined as the exponential moving average of the teacher's mean pre-softmax activations over each mini-batch. BYOL has an asymmetrical architecture. Similarly to DINO, both the teacher and the student consist of a feature extractor, followed by a projection MLP-based head. However, the student has an additional MLP network θ_{Pred} on top of the projection head. The f_θ^S model is trained to minimize the normalized mean square error loss between the outputs of the student $\theta_{Pred}^S(\theta_{Proj}^S(f_\theta^S(x_1)))$ and the teacher $\theta_{Proj}^T(f_\theta^T(x_2))$. The main difference between the two methods is the type of augmentations they use. BYOL utilizes two augmented views $V = 2$ of the same image whereas, DINO uses eight views $V = 8$ in total. In detail, for each image, DINO generates one large augmented view, called a global crop, and three smaller views, called local crops. All crops are used by the student but only the global views are passed through the teacher. The loss is then defined as

$$\mathcal{L}_{DINO} = \sum_{x \in \{x_1^g, x_2^g\}} \sum_{x' \in \{V\} x' \neq x} -P_T(x) log(P_S(x'))$$

3.3. Cross-Domain Consistency Training

Our method, Cross-Domain Consistency Training (CDCL) is a modification that can be readily applied to common SS frameworks such as DINO and BYOL described above. CDCL is designed to utilize readily-available metadata to enforce the learning of consistent representations of batches during self-supervision. As such, it is more suitable for the challenges associated with HCS data.

Consistency-based SSL methods learn by enforcing consistent representations given different augmentations of the same image. In a similar spirit, CDCL enforces consistent representations given a pair of images from different batches, but with the same treatment. When applied to DINO, CDCL creates 1 global crop and 3 local crops from two images sampled from the same treatment but *from different batches*, see Figure 1. When applied to BYOL, CDCL augments two images sampled from the same treatment but from different batches. Then, these views are used as inputs to the teacher and the student network, as described in Section 3.2. Specifically for DINO, we introduce two additional changes that improve performance even further:

1. Batch-wise centering. We use a centering vector for each batch to mitigate feature collapse.

2. An additional consistency-based loss. We add an extra loss function that increases representational diversity and improves convergence, even for small mini-batches.

These improvements are described below.

Batch-wise centering In DINO, the center vector C is continuously updated using an exponential moving averaging when training. This mechanism is meant to stop the network from representational collapse *i.e.* collapsing to trivial solutions. In CDCL, we use a similar approach but maintain a center vector C_d for each of the domains D. This is done such that the sample probability output of the teacher P_T is dependent on the domain d the input image x_1 comes from $P_T = softmax(f_\theta^S(x_1) - C_d)$.

A domain can be considered a batch or a plate, depending on the need. The motivation behind this choice is to ensure that each batch/plate/domain is evenly distributed across the latent space.

Additional Consistency Loss Self-supervised methods like DINO use large mini-batches, but this is computationally demanding. In (Zbontar et al., 2021), the Barlow loss was introduced to reduce the feature representational redundancy and the need for large mini-batches. This loss minimizes the mean square error between the cross-correlation matrix R of P_T and P_S and the unit matrix I. Effectively penalizing any covariance between output dimensions, while still enforcing a similar representation between samples. We add the Barlow loss as additional loss to reduce the GPU requirements and increase the representational diversity, which is defined as

$$\mathcal{L}_{Barlow} = \sum_i (1 - R_{ii})^2 + \alpha \sum_i \sum_{j \neq i} R_{ii}^2$$

The combined loss is then defined as $\mathcal{L} = \lambda \mathcal{L}_{DINO} + (1 - \lambda)\mathcal{L}_{Barlow}$ regulated by a constant $\lambda \in [0, 1]$. Throughout this work we use $\lambda = 0.5$.

3.4. Datasets

We use two different HCI datasets for our experiments, both using setups similar to or the same as the standardized Cell Painting Assay (Bray et al., 2016), which relies on multiplexed fluorescence microscopy data to capture phenotypic response data of cell lines treated with either small molecules or siRNAs. We use RXRX1-HUVEC (Taylor et al., 2019) as our primary dataset and we run additional experiments on CPG0004 (Way et al., 2022), both described below. We use DeiT-S as the backbone for all settings as we found experimentally that it outperforms similar capacity CNN-based architectures when applied to high content screening data.

RXRX1-HUVEC is a subset of the fluorescent microscopy dataset RXRX1 (Taylor et al., 2019), originally designed to study representational learning under batch effects. The subset contains 59,000 6-channel images, consisting of 24 distinct batches *i.e.* replicates of the same experiment. Every experiment contains the same 1139 unique siRNA treatments, which are considered as distinct classes. The task for this dataset is to identify the treatment across batches. To this end we split the data at the batch-level and use 16 batches for training, 4 for validation and 4 for testing, in a 6-fold cross validation setup.

CPG0004 is a dataset from the Cell Painting Gallery (Way et al., 2022). The full dataset contains over 9000 unique compound-concentration pairs of 5-channel images. We use a well-annotated subset of 540 unique treatments with known Mode of Action (MoA) data, where each treatment belongs to one of 50 MoA. The primary task for this dataset is to predict treatment, the downstream task is to predict MoA. CPG0004 consists of 136 plates, with most plate layouts being replicated 5 times. We use a 5-fold cross-validation setup, splitting the data by replicate group and using three for training, one for validation and test respectively.

4. Experiments

Our experiments explore the use of representational learning, in particular self-supervised learning, for high content screening data. We begin in Section 4.1 by establishing supervised baselines and then showing the unexpected failure of SSL methods on this type of problem. We pinpoint the source of the problem to batch effects. In Sections 4.2 through 4.4 we apply our proposed method, CDCL, and demonstrate that it can mitigate batch effects and improves performance, making self-supervision feasible on HCS data. Further ablation studies suggest that our method learns more robust and distinguishing features that leads to more useful features for downstream analysis.

Implementation details For all the experiments we use DeiT-S models pretrained on ImageNet. For establishing the supervised baselines the the DeiT model is followed by a linear layer. For the SSL methods the same model is used as a feature extractor $f_{\theta_{Ext}}$ and we assess performance using the k-NN evaluation protocol on the extracted features, similarly to (Caron et al., 2021). All models are trained using AdamW (Loshchilov and Hutter, 2017) as the optimizer. Hyper-parameters are selected through grid search based on results from the validation set. We use learning rate warm-up for 1000 iterations followed by cosine annealing.

Dataset	Method	Lin. Accuracy ↑	k-NN Accuracy ↑	k-NN Accuracy - norm ↑
RXRX1-HUVEC	Supervised	52.56 ± 5.08	51.44 ± 5.47	53.52 ± 5.53
	SSL-DINO	14.46 ± 2.65	14.28 ± 2.04	25.04 ± 3.29
	SSL-BYOL	4.04 ± 0.86	9.38 ± 1.45	15.1 ± 2.74
	SSL-DINO-CB	30.08 ± 8.21	41.52 ± 8.5	55.3 ± 7.29
	SSL-BYOL-CB	46.24 ± 4.41	42.06 ± 4.19	47.44 ± 4.02
	CDCL	49.64 ± 4.93	53.56 ± 6.61	$\mathbf{63.1 \pm 6.33}$

Table 1: *Main results* We report the average cross-validation accuracy on the RXRX1-HUVEC test sets, using three evaluation protocols described in Section 4. Unlike the natural domain, off-the-shelf SSL learning performs significantly worse than supervised methods. Our proposed methods make self-supervision feasible.
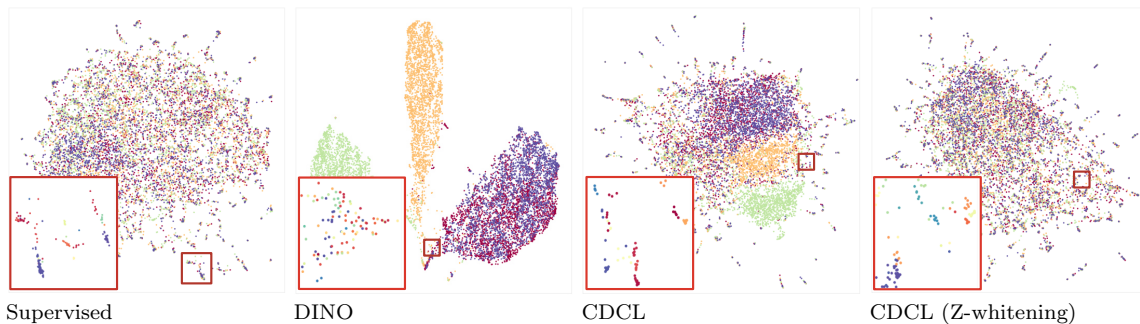

Supervised   DINO   CDCL   CDCL (Z-whitening)

Figure 2: *Feature embeddings.* UMap representation of the test set feature representation of each method, colored by batch belonging.

Evaluation protocol We evaluate model performance using three different strategies depending on the type of training. For supervised models, we append a linear layer on top of the feature extractor and we fine-tune the full model. For self-supervised learning, we use the model's output features to make predictions using the following methods: *i)* by adding a aux layer that is learned independently of the feature extractor, which does not back-propegate any learning signal to the feature extractor). *ii)* using the k-NN evaluation protocol. We visualize the feature space based on the models output features using UMAP (McInnes et al., 2018) in Figures 1 and 2. In both the k-NN evaluation and UMAP feature visualisation we also perform feature normalization as a post processing step as commonly done in HCS datasets. We use Z-norm whitening for batch wise feature normalization, such that each batch has zero mean a unit variance, based on the controls. We further include the k-BET-score (Büttner et al., 2019), to measure sample mixing between batches. This metric ranges from 0 to 1, with a higher value indicating more overlap between samples from different batches.

4.1. Self-Supervised Learning

In the Introduction we asked *can self-supervised methods like DINO and BYOL learn meaningful representations for high content screening image data?* Given the relatively large size of high content screening datasets – and their sparse and noisy labels – one would expect these methods would work particularly well in this setting. Through a series of experi-

Dataset	Method	Train Set k-BET ↑	Test Set k-BET ↑	Test Set - norm k-BET ↑
RXRX1-HUVEC	Supervised	15.78 ± 4.35	24.05 ± 17.76	38.20 ± 25.20
	SSL-DINO	00.00 ± 00.00	6.87 ± 9.45	27.71 ± 23.76
	SSL-BYOL	00.02 ± 00.03	17.00 ± 13.72	57.90 ± 20.49
	SSL-DINO-CB	2.31 ± 1.29	12.54 ± 9.19	45.06 ± 26.60
	SSL-BYOL-CB	35.39 ± 10.32	35.73 ± 17.52	52.12 ± 28.04
	CDCL	10.45 ± 3.59	22.26 ± 14.50	52.89 ± 31.20

Table 2: *k-BET* We report the average cross-validation k-BET score on the RXRX1-HUVEC, train and test sets.

ments, we explore how off-the-shelf self-supervised methods DINO and BYOL perform on high content data.

The results for RXRX1-HUVEC appear in Table 1, where we compare a standard supervised learning approach (top row) with self-supervision using DINO (SSL-DINO) and BYOL (SSL-BYOL). The performance using self-supervision drops precipitously across all three metrics when compared to the standard supervised approach. This stands in stark contrast to the natural domain, where SSL methods perform on par with their supervised counterparts (Caron et al., 2021). Some property of high content data causes a drop in performance disproportionate to what is expected.

To understand the root of the problem, we visualise how the extracted features are organised in Figure 1. We use UMAP (McInnes et al., 2018) to project the extracted features to a two-dimensional plane and visualize the 2D projections of the feature space after training. Each color represents the batch the data was created in. Evidently, the features are clustered based on the batch they belong to, instead of biologically relevant properties, revealing the dominance of batch effects. This explains the unexpected failure of SSL methods we observed above. In Table 2, we quantify this effect by comparing the k-BET score of the supervised features versus the self-supervised features. Using kbet, we measure how well the data clusters according to batch and how well it clusters according to treatment. Clearly, self-supervised learning features are dominated by batch effects.

One way to counteract batch effect is z-norm whitening, as explained in Section 3.4. Results from Table 1 show that both DINO and BYOL's k-NN performance is improved significantly, however, this is not enough to combat the problem. A solution that deals with the root of the problem is needed.

4.2. Cross-Domain Consistency Learning

In Section 3, we outline simple changes to self-supervised learning that should mitigate the issues caused by batch effects. Instead of presenting augmented pairs of the same image to the network during learning, we propose to use readily available metadata to select pairs of samples that have the same treatments but belong to different batches. The intuition for this approach comes from the guiding principle of self-supervised learning – that the network should learn common features and ignore sources of noise. In HCS data, each image should in principle contain two dominating signals. The first is related to the treatment and its biological effect. The other is comprised of confounding factors, *i.e.* batch effects.

Pre-trained method	Finetuned	1-NN NSC MOA Acc ↑
IMAGENET	✓	21.8 ± 4.0
Weakly Supervised	✗	11.3 ± 2.1
CDCL	✗	14.3 ± 1.8
Weakly Supervised	✓	21.0 ± 5.0
CDCL	✓	$\mathbf{22.6} \pm 4.5$

Table 3: *CPG0004* results, reporting the 1-NN Not-Same-Compound MoA accuracy in the test set.

By providing two images of the same treatment but different batches to the network, the common signal should only be the one of interest, the biological. The sampling strategy is not tied to the treatment information. It can be done using any metadata that hint similar biological processes but come from a different experiment. In this work, we use treatment-based sampling for consistency and fair comparison. We explore the effect of this strategy in DINO and BYOL, following the implementation described in Section 3.3.

When cross-batch learning is applied, we observe a significant increase in performance of self-supervised methods (see Table 1). When z-norm whitening is applied, the accuracy further improves by a large margin. The feature maps visualized in Figure 2 indicate an obviously decreased feature dependence of the batch effects. Evidently, cross-batch learning helps to learn features that cluster based on their biological effects and reduce dependence on batch effects. However, the performance is still not on par with supervised learning. For this, we need to apply the full CDCL method, including batch-wise centering and the Barlow loss.

When we apply batch-wise centering along with the additional consistency loss, as described in Section 3.3, we find that DINO with CDCL not only outperforms BYOL, but it significantly outperforms the supervised version – even without fine-tuning. CDCL-DINO results in well-behaved features, alleviated from batch effects, and pushes the score by +9.5% over its supervised counterpart.

4.3. CDCL improves down-stream classification

We demonstrated that CDCL produces well behaved features that improve accuracy and mitigate batch effects. But *are these features useful for downsteam tasks? – is there an advantage of CDCL pre-training?* We answer these questions using CPG0004, where predicting the mode-of-action (MOA) is the down-stream task. We report the results in Table 3. Without fine-tuning using MOA labels, CDCL outperformes its weakly supervised counterpart by 3%, demonstrating the versatility of the features learned with CDCL. When fine-tuned, models trained with CDCL outperform all others, including the fully supervised model.

4.4. CDCL increases robustness in exploratory data settings

One of the main purposes of high content screens is to explore the compound space for promising new candidate drugs. Given its exploratory nature, a larger number of unique

Dataset	Setting	Method	Test Acc. k-NN ↑	Test Acc. k-NN - norm ↑
	I	Supervised	17.47 ± 1.94	18.5 ± 1.89
	I	CDCL	17.68 ± 1.48	**20.8 ± 2.15**
	II	Supervised	49.73 ± 4.47	51.68 ± 4.41
RXRX1-HUVEC	II	CDCL	51.53 ± 5.73	**61.43 ± 5.03**
	III	Supervised	40.85 ± 3.69	42.2 ± 3.57
	III	CDCL	45.53 ± 5.18	**54.47 ± 5.39**

Table 4: *Ablation* Experiments comparing the supervised baseline and CDCL method when trained on data subset scenarios commonly seen in HCS datasets.

compounds is often prioritized over a higher number of replicates. This has a direct performance impact on weakly-supervised learning since it creates settings with many classes with few samples, as well as a few classes with many samples (the controls). The result is a severe class imbalance. To assess performance in different exploratory settings, we subsample RXRX1-HUVEC and we simulate three different scenarios: *I)* many samples per class – using only the controls, *II)* few samples per class – using only the treatments and, *III)* using 50% of the treatments. We compare CDCL with the weakly supervised baselines and we report our findings in Table 4. For all cases, we observe a large performance drop, compared to the full dataset. However, CDCL consistently outperforms its weakly supervised counterpart – especially when Z-norm whitening is used. Particularly when only a half of the treatments are available, suggesting that the method is capable of generating features that are more useful for distinguishing new unseen treatments.

5. Conclusion

In this work we show that self-supervised learning for high content screens is extremely susceptible to batch effects, and fails to show the same promising results observed in the natural domain. We identify the root causes of the issue and propose methods to counteract them. We demonstrate that using CDCL, self-supervised learning can not only match but exceed supervised learning performance in the HCS drug discovery setting. We also show that it is more robust when used on non-idea exploratory datasets. This is a step towards more domain agnostic features in biomedical image analysis. The extent to which this is applicable to medical and natural data is an open question which we leave as feature work.

Acknowledgments

References

D Michael Ando, Cory Y McLean, and Marc Berndl. Improving phenotypic measurements in high-content imaging screens. *BioRxiv*, page 161422, 2017.

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.

Maren Büttner, Zhichao Miao, F Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. A test metric for assessing single-cell rna-seq batch correction. *Nature methods*, 16(1): 43–49, 2019.

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*, pages 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. *arXiv preprint arXiv:2104.14294*, 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(10):1–11, 2006.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

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.

Rens Janssens, Xian Zhang, Audrey Kauffmann, Antoine de Weck, and Eric Y Durand. Fully unsupervised deep mode of action learning for phenotyping high-content cellular images. *Bioinformatics*, 37(23):4548–4555, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Claire McQuin, Matthew Smith, Allen Goodman, Rebecca Senft, Yu Han, Mehrtash Babadi, Peter Horvath, et al. Learning representations for image-based profiling of perturbations. *bioRxiv*, 2022.

Alexis Perakis, Ali Gorji, Samriddhi Jain, Krishna Chaitanya, Simone Rizza, and Ender Konukoglu. Contrastive learning of single-cell phenotypic representations for treatment classification. In *International Workshop on Machine Learning in Medical Imaging*, pages 565–575. Springer, 2021.

Daniel Siegismund, Mario Wieser, Stephan Heyse, and Stephan Steigele. Self-supervised representation learning for high-content screening. In *Medical Imaging with Deep Learning*, 2021.

Jaak Simm, Günter Klambauer, Adam Arany, Marvin Steijaert, Jörg Kurt Wegner, Emmanuel Gustin, Vladimir Chupakhin, Yolanda T. Chong, Jorge Vialard, Peter Buijnsters, Ingrid Velter, Alexander Vapirev, Shantanu Singh, Anne E. Carpenter, Roel Wuyts, Sepp Hochreiter, Yves Moreau, and Hugo Ceulemans. Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell Chemical Biology*, 25(5): 611–618.e3, 2018. ISSN 2451-9456. doi: https://doi.org/10.1016/j.chembiol.2018.01.015. URL https://www.sciencedirect.com/science/article/pii/S2451945618300370.

J Taylor, B Earnshaw, B Mabey, M Victors, and J Yosinski. Rxrx1: An image set for cellular morphological variation across many experimental batches. In *International Conference on Learning Representations (ICLR)*, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

Maria-Anna Trapotsi, Elizabeth Mouchet, Guy Williams, Tiziana Monteverde, Karolina Juhani, Riku Turkki, Filip Miljkovic, Anton Martinsson, Lewis Mervin, Kenneth R Pryde, et al. Cell morphological profiling enables high-throughput screening for proteolysis targeting chimera (protac) phenotypic signature. *ACS Chemical Biology*, 17(7):1733–1744, 2022.

Scott J Warchal, John C Dawson, and Neil O Carragher. Evaluation of machine learning classifiers to predict compound mechanism of action when transferred across distinct cell lines. *SLAS DISCOVERY: Advancing Life Sciences R&D*, 24(3):224–233, 2019.

Gregory P. Way, Ted Natoli, Adeniyi Adeboye, Lev Litichevskiy, Andrew Yang, Xiaodong Lu, Juan C. Caicedo, Beth A. Cimini, Kyle Karhohs, David J. Logan, Mohammad H. Rohban, Maria Kost-Alimova, Kate Hartland, Michael Bornholdt, Srinivas Niranj Chandrasekaran, Marzieh Haghighi, Erin Weisbart, Shantanu Singh, Aravind Subramanian, and Anne E. Carpenter. Morphology and gene expression profiling provide complementary information for mapping cell state. *bioRxiv*, 2022. doi: 10.1101/2021.10.21.465335. URL https://www.biorxiv.org/content/early/2022/10/12/2021.10.21.465335.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.