COMPARING REPRESENTATIONS OF BIOLOGICAL DATA LEARNED WITH DIFFERENT AI PARADIGMS, AUGMENTING AND CROPPING STRATEGIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in computer vision and robotics enabled automated large-scale biological image analysis. Various machine learning approaches have been successfully applied to phenotypic profiling. However, it remains unclear how they compare in terms of biological feature extraction. In this study, we propose a simple CNN architecture and implement weakly-supervised, self-supervised, unsupervised and regularized learning of image representations. We train 16 deep learning setups on the 770k image dataset under identical conditions, using different augmenting and cropping strategies. We compare the learned representations by evaluating multiple metrics for each of three downstream tasks: i) distance-based similarity analysis of known drugs, ii) classification of drugs versus controls, iii) clustering within cell lines. We also compare training times and memory usage. Among all tested setups, multi-crops and random augmentations generally improved performance across tasks. We show that self-supervised models have competitive performance and can be trained up to 11 times faster than others. We demonstrate pros and cons of using regularized learning. We observe that no single combination of augmenting and cropping strategies consistently delivered top performance across tasks and recommend prospective research directions.

1 INTRODUCTION

With recent advances in robotics and deep learning methods, automated large-scale biological image analysis has become possible. Different microscopy technologies allow to collect imaging data of samples under various treatment conditions. Then, images are processed to extract meaningful biological features and compare samples across cohorts. As opposed to carefully engineered features used in the past, deep learning approaches are widespread and automatically distil relevant information directly from the data (Moen et al., 2019).

A lot of approaches, following different paradigms of machine learning, have been successfully applied to image-based phenotypic profiling: from fully supervised approaches (Godinez et al., 2017; Kraus et al., 2017) to generative adversarial learning (Hu et al., 2019; Goldsborough et al., 2017; Radford et al., 2016) and self-supervision (Robitaille et al., 2021; Zhang et al., 2020). However, it remains unclear how these approaches align with each other in terms of biological feature extraction. The direct comparison is close to impossible, as many aspects differ between the studies: imaging technologies, datasets, learning approaches and model architectures, implementations and hardware.

In the emergent field of self-supervised learning, a key role of random data augmentations and multiple image views has recently been shown (Caron et al., 2021). Their synergetic impact on learning image representations has not yet been rigorously studied. In this paper, we compare different deep learning setups in their ability to learn representations of drug-treated cancer cells. We propose a simple CNN architecture and implement four approaches to learn representations: the weakly-supervised, the unsupervised (with and without regularization) and the self-supervised. We train the models on the same dataset of 770k cell images with and without random image augmentations, with single and multi-crops. The other training conditions are kept identical. We compare the learned representations in the three downstream analysis tasks, discuss their performance and provide the comparison summary table.

Our main contributions are:

- implementations of 16 deep learning setups, including state-of-the-art methods trainable within limited resources,
- a systematic comparison of learned representations.

2 RELATED WORK

Weak supervision has been a popular choice to learn medical image representations and has proven its efficiency (Caicedo et al., 2018; Lu et al., 2020). When analyzing samples corresponding to different treatments, patients, or any experimental conditions, those are often used as weak labels. In our case, there are 693 conditions with different combinations of drugs and cell lines. However, the effects of those combinations are largely unknown, so we restrict ourselves into using two labels only: drug vs control (supposedly, effect vs no effect).

A recent approach to understand morphological features of cancer cells by Longden et al. (2021) follows an unsupervised perspective. The authors apply a deep autoencoder to learn 27 continuous morphological features. However, their model does not work with raw images. It uses 624 extracted numerical features as input, and applies a series of linear layers to reconstruct them. Here, we use a convolutional autoencoder instead, to learn more features directly from the data.

Several approaches for learning representations of cell images are based on generative adversarial networks (Arjovsky et al., 2017; Gulrajani et al., 2017). Such models often have two components: the generator and the discriminator network, trained simultaneously in a competitive manner. In this work, we implement a similar idea in the form of regularization: we use a deep convolutional autoencoder as generator, and a weakly-supervised classifier as discriminator. Both networks share the same stack of layers, responsible for learning representations, while optimizing different loss functions. In this setting, the computational time and memory usage remain comparable to the aforementioned approaches.

Finally, self-supervision has recently emerged in bioinformatics to address problems like cell segmentation, annotation and clustering (Lu et al., 2019; Santos-Pata et al., 2021). Most recently, a self-supervised contrastive learning framework has been proposed by Ciortan & Defrance (2021) to learn representations of scRNA-seq data. The authors follow SimCLR (Chen et al., 2020b) in the implementation of contrastive loss and show that their approach compares favorably with state-of-the-art (SOTA) methods in a downstream clustering task. Here, we train a self-supervised CNN backbone, following BYOL (Grill et al., 2020). Unlike SimCLR, this approach does not need negative pairs, yet it was shown to have a superior performance.

In spite of the great interest in deep-learning-based approaches to learning representations of biological data, there have been very few attempts to fairly compare those. A comparison of AI-based methods to predict cell function has come out lately (Padi et al., 2020). However, it was primarily focused on collating traditional machine learning versus deep learning. Brief general comparisons of recent AI approaches can be found in reviews and surveys (Moen et al., 2019; Chandrasekaran et al., 2021; Nguyen et al., 2019), but they lack details and cannot inform decision making. Recently, a thorough comparison of data-efficient image classification models has been published by Brigato et al. (2021). The authors evaluated 10 models on 6 different datasets. Eventually, this analysis focuses on classification tasks only. Although we see papers illustrating how a single study can benefit from multiple AI paradigms (Chen et al., 2020a), it remains unclear which approach is preferable in a particular representation learning task. In this study, we attempt to address this question for analysis of images of cancer cell lines growing in tissue cultures.

3 DATA

The initial dataset comprises 1.1M high-resolution grey-scale images of drug-treated cancer cell populations growing adherently *in vitro*. It captures 693 unique combinations of 21 cell lines and 33 drugs at 5 different drug concentrations, multiple time points and biological replicates. Analyzing such datasets often presents a challenge of identifying previously unknown drug-induced morphological patterns and is useful to inform future clinical applications (such as combination therapy).

We carefully subset the initial data to obtain a balanced dataset of two labels: samples with supposedly the strongest drug effect (i.e., the highest drug concentration, the latest time point) and controls (no drugs, any time point). We end up with about 770k image crops of size 64x64. It is important to note that some drugs did not provoke any effect on resistant cell lines, so the corresponding images of drugs and controls look similar. Some other drugs showed growth arrest, which resulted in drug-treated images being similar to early time point controls, where the cells have not grown yet. By balancing the dataset to contain such cases (**Fig. 1**), we expected the models to learn specific morphological differences, instead of superficial features like cell location in the crop, cell population density, amount of grey, etc.

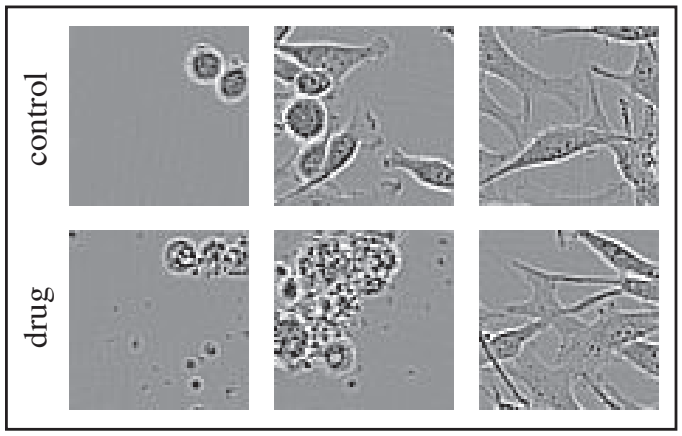

Figure 1: Typical examples of control and drug images (M14 cell line). On the left, an early time point of the control (cells have not grown yet) is shown against a strong drug effect (fragmented or dead cells). On the right, the end time points for the control and an ineffective drug are depicted. In the middle, an example of intermediate growth of a control sample versus another cytotoxic drug is given. Note the similarity between drugs and controls.

4 METHODS

4.1 MODEL ARCHITECTURES

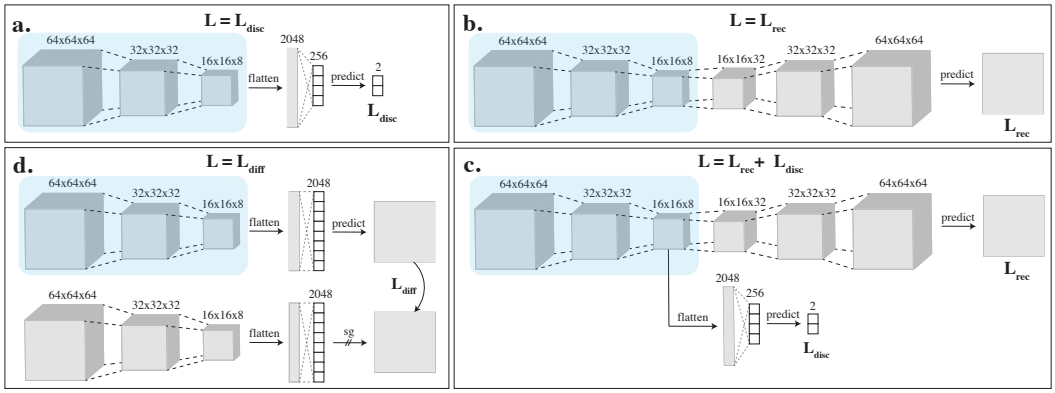

Figure 2: Graphical overview of models. **a.** A weakly-supervised deep classifier with a categorical cross-entropy discrimination loss (drug vs control). **b.** A convolutional autoencoder with a binary cross-entropy reconstruction loss. **c.** A regularized convolutional autoencoder: models a and b, sharing the CNN backbone, trained simultaneously. **d.** A self-supervised CNN backbone with a mean-squared error difference loss (BYOL).

To learn image representations, we applied four machine learning paradigms:

a. weakly-supervised learning: a classifier (with two labels only),

b. unsupervised learning: an autoencoder (with classic encoder-decoder architecture),

c. regularized learning: a combination of models **a** and **b** (with encoder-classifier-decoder architecture),

d. self-supervised learning: the CNN backbone, trained following BYOL (Grill et al., 2020).

As seen on **Figure 2**, the four architectures contain the same CNN backbone, which is used to produce image representations for the downstream analysis. It was important to use the same stack of layers to ensure fair comparison of methods in retrieving relevant biological information. However, image representations are learned solving substantially different tasks: discrimination between drug and control images (model **a**), reconstruction of images (models **b-c**), minimization of image representation difference between the online and the target networks (model **d**).

For the regularized model, we adopted a particular implementation where a classifier and an autoencoder are trained in turns, optimizing different loss functions (**Fig. 2c**). Our idea was to encourage the autoencoder to learn representations that would bear differences between drug and control images, while still delivering high quality image reconstructions. In this formulation, the classifier acts as a regularizer. Although similar models have been utilized in chemo- and bioinformatics tasks (Gómez-Bombarelli et al., 2018; Rong et al., 2020), to our knowledge this implementation has not been tested previously in the analysis of biological images.

4.2 Training setups

Each model was trained under 4 conditions of presence and absence of random image augmentations and multi-crops, giving rise to 16 training setups in total. Since the dataset is naturally grayscale, we only applied random resized crops, horizontal flips and Gaussian blurs to augment. Note that data augmentations are intrinsic to the self-supervised approach. Therefore, we tested single and double augmenting (while preprocessing and/or while training) for model **d**. In the one-crop setting, we used single 64x64 images. For multi-crop, we added 4 random resized crops applied to 64x64 images: 2 of about half-size, and 2 more of about quarter-size (5 crops in total).

We implemented the 16 described setups and trained them using Nvidia GeForce RTX 2060 with 6 GB of memory only. We chose the CNN backbone architecture, batch size and other common hyperparameters by running grid search and finding the best average performance across models, achievable within reasonable training time and hardware memory constraints.

For the self-supervised model, we additionally optimized three BYOL parameters: *projection_size*, *projections_hidden_size* and *moving_average_decay*. We trained the model 100 times, sampling parameters from predefined ranges. We found that the model with an equal number of neurons for hidden and projection layers worked consistently and achieved the lowest MSE loss for our data among testable parameter sets. That parameter set was used for final training and evaluation.

The classifiers in **a** and **c** were trained by optimizing categorical cross-entropy loss. The autoencoders in **b** and **c** were trained with binary cross-entropy loss and softmax activation. We trained all models for 50 epochs, using Adam optimizer with a constant learning rate of 0.0001. A batch size of 256 was used. We defined the same early stopping criterion, which checks a simple divergence condition on the loss function. We used the same data splits with 10% for the validation set to test classification accuracy and reconstruction quality, while training.

4.3 Validation and evaluation

We validated the models by monitoring the corresponding loss functions, classification accuracy and image reconstruction quality for training and validation sets (see **Supplementary** materials). For each of three downstream tasks, we evaluated several metrics as described below.

4.3.1 Distance metrics

First, we compared the learned representations in their ability to capture similarity of known drugs. Let S_1 and S_2 be the sets of images of two drugs, known to have similar effects, and C be the set of control images. We calculate the following two metrics to quantify similarity between S_1 and S_2:

- $D(S_1, S_2) = \underset{u \in S_1, v \in S_2}{\text{median}} (||u - v||)$,
  i.e., the median Euclidean distance between any two images (u, v) of two sets.

- $d(S_1, S_2) = \frac{\hat{D} - D(S_1, S_2)}{\hat{D}}$, where $\hat{D} = \frac{1}{2}[D(S_1, C) + D(S_2, C)]$,
  i.e., the normalized difference between drug-to-control and drug-to-drug distances.

4.3.2 CLASSIFICATION METRICS

Next, we compared the learned features by their performance in the binary classification task, defined initially for the weakly-supervised model. We used a pretrained stack of layers of each model to generate codes and then trained a classifier with two linear layers to differentiate between drugs and controls. We used the same data splits for all models and trained them for 25 epochs with SGD optimizer and batch size of 1024. We ran grid search over learning rate, momentum and weight decay parameters to achieve the best training and validation accuracy. To comprehensively evaluate the performance of classifiers, we calculated the following metrics for each cell line individually: accuracy, precision, recall and area under ROC.

4.3.3 CLUSTERING METRICS

Finally, we compared the learned features in the clustering task. For each cell line, we pulled the corresponding images from the validation set, obtained their representations and further reduced dimensionality with UMAP (McInnes et al., 2020). We clustered the resulting embeddings with HDBSCAN (McInnes et al., 2017) and evaluated several metrics: number of identified clusters and percent of noise points, Silhouette score and Davies-Bouldin similarity measure.

For each cell line, we ran the analysis multiple times, varying two parameters: i) *n_neighbors*, responsible for constraining the size of the local neighborhood in UMAP, and ii) *min_cluster_size*, representing the smallest grouping size in HDBSCAN. We adopted the following procedure to find the best partitions:

i. select Silhouette scores above median,
ii. for those, select Davies-Bouldin scores below median,
iii. within the rest, select the lowest percent of noise,
iv. if multiple parameter sets are left, pick the one of max number of clusters.

This logic was motivated by zero correlation between the Silhouette and the Davies-Bouldin measures, and by the objective to find as many "clean" clusters as possible.

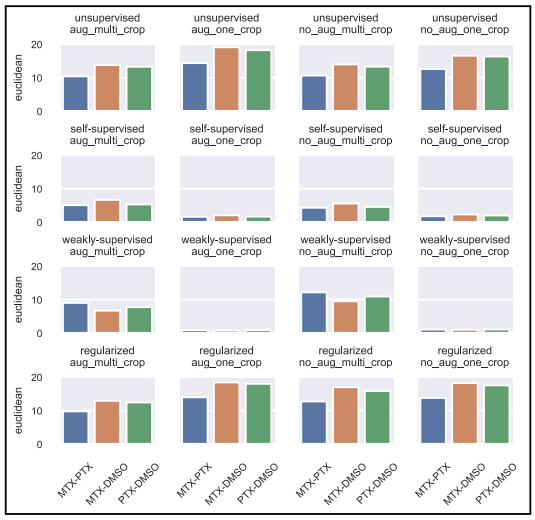

Figure 3: D(MTX, PTX), D(MTX, DMSO), D(PTX, DMSO) for M14 cell line. The four model architectures are organized in rows (from top to down): the unsupervised, the self-supervised , the weakly-supervised, the regularized. In columns, the four different setups are given (from left to right): augmentations + multi-crops, augmentations + single crops, no augmentations + multi-crops, no augmentations + single crops.

5 RESULTS

5.1 DISTANCE-BASED DRUG SIMILARITY ANALYSIS

Pemetrexed (PTX) and Methotrexate (MTX) are two drugs that have similar chemical structures and both inhibit folate-related enzymes. Over the years, they have been successfully applied to cure many types of cancer (non-small cell lung cancer, pleural mesothelioma, lymphoma, etc. (Ruszkowski et al., 2019)) We applied distance-based analysis to evaluate how close PTX and MTX are to each other in terms of learned features, and how distant they both are from controls (images of cells under no treatment).

In order to do that, we picked all images related to PTX and MTX drugs from the validation set. Then, we randomly picked the same number of control images (DMSO). We calculated $D(\text{MTX, PTX})$, $D(\text{MTX, DMSO})$, $D(\text{PTX, DMSO})$ on image representations, which resulted in around 3600 distances for each cell line and pair on average. Based on *a-priori* knowledge of efficiency and similarity of the drugs, we expected MTX-PTX distances to be consistently lower than of MTX-DMSO and PTX-DMSO. That expectation was not met by all models, as shows **Figure 3**.

Analyzing distances for M14 cell line, we observed that in the latent space the two drugs (PTX and MTX) were closer to each other than either of them to controls (DMSO) for two models only: the unsupervised (**Fig. 3**, row #1) and the regularized (row #4) ones. The distances for the self-supervised and the weakly-supervised models (rows #2-3) were rather on the same level. Strikingly, the one-crop setup for both of them (columns #2 and #4) resulted in distances close to zero, which implies that information in the learned representations was insufficient to characterize drug effects. Multi-crop setting, in turns, caused large increase in distances, which suggests information gain. Nonetheless, it was not enough to capture dissimilarity between drugs and controls in this case.

We repeated the same analysis for each of 21 cell lines. We found that with the exception of weakly-supervised models, all produced lower average MTX-PTX distances, compared to MTX-DMSO and PTX-DMSO. Also, the median normalized differences d turned out to be the largest for the self-supervised model (**Tab. 2**). This suggests that the space of learned features of weakly-supervised models is likely to contain more trivial information about the drug effects, rather than features of altered morphology.

5.2 Classification of drugs versus controls

Figure 4 shows classification results for a few picked cell lines. All models show comparable performance, crossing 0.6 accuracy bottom line and reaching 0.7 in many cases. However, it is only the weakly-supervised model (row #3) that achieved 0.8 accuracy for the HT29 cell line and delivered consistently higher performance in all setups. This was expected due to identical problem formulation in representation learning. Interestingly, the other three models have also shown rather high, rival performance on this task. That implies that all models have a potential in detecting drug effects in time-series imaging data (e.g., to predict drug onset times for different concentrations).

Table 2 contains four classification metrics for each training setup, evaluated on the entire dataset. Median performance for 21 cell lines is reported. The regularized model with single crops and augmentations showed the highest overall accuracy (0.76 ± 0.07) and ROAUC (0.76 ± 0.06), though the weakly-supervised model was the most robust across settings. Both, the weakly- and self-supervised models delivered their best performance under multi-crop setting.

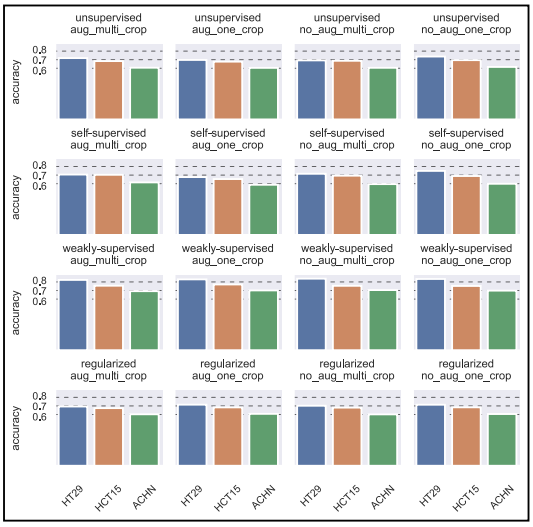

Figure 4: Binary classification accuracy (drug vs control) for three picked cell lines: HT29, HCT15, ACHN. The four model architectures are organized in rows (from top to down): the unsupervised, the self-supervised , the weakly-supervised, the regularized. In columns, the four different setups are given (from left to right): augmentations + multi-crops, augmentations + single crops, no augmentations + multi-crops, no augmentations + single crops.

5.3 Clustering analysis within cell lines

Figure 5A presents mean numbers of identified clusters across models and settings. Varying the clustering parameters to encourage smaller or bigger partitions resulted in relatively large confidence intervals. However, even the lower bounds exceeded n=2 clusters, which would correspond to the trivial case of differentiating between drugs and controls (effect vs no effect), in the majority of

cases. That indicates that the learned representations allow studying the data in more depth (e.g., finding similarities in concentration-dependent morphological drug effects).

Although mean numbers of clusters look similar, the quality of partitions differed substantially across cell lines, as follows from the Silhouette score barplots (**Fig. 5B**). The weakly-supervised model produced the poorest scores for the three picked cell lines. Close-to-zero and even negative values suggest that the clusters were mainly overlapping. In such cases, obtained partitions are far less trustworthy and any follow-up analysis on them is controversial. The self-supervised model delivered better scores, though the top competitive performance was shown by the other two models: the unsupervised and the regularized ones. Interestingly, the highest scores for the COLO205 cell line appear with low numbers of identified clusters, suggesting there is only a few morphological patterns to be found in this cell line. The mean statistics across all cell lines are given in **Table 2**.

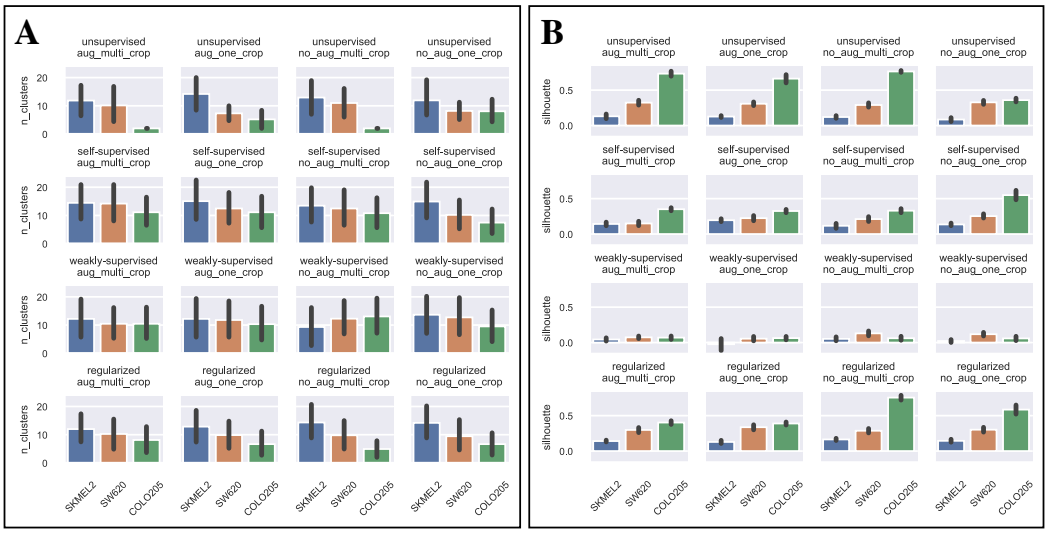

Figure 5: Clustering analysis for three picked cell lines: SKMEL2, SW620, COLO205. Mean numbers of identified clusters (**A**) and mean Silhouette scores (**B**) are shown with confidence intervals. The four model architectures are organized in rows (from top to down): the unsupervised, the self-supervised , the weakly-supervised, the regularized. In columns, the four different setups are given (from left to right): augmentations + multi-crops, augmentations + single crops, no augmentations + multi-crops, no augmentations + single crops.

5.4 Training times and memory usage

All models were trained using Nvidia GeForce RTX 2060 with 6 GB memory. With batch size of 256, steady memory consumption was around 4.3 and 4.7 GB for single and multi-crops, respectively. Batch size of 512 resulted in cuda-out-of-memory error in all setups.

Unlike memory usage, training times differed largely for 4 model architectures and cropping strategies (**Tab. 1**). The self-supervised model was the only one to meet the early stopping criterion, which resulted in *remarkably* small training times. The `one_crop` training stopped after 16/50 epochs, whereas `multi_crop` made 7/50 epochs only. The other models were trained for all 50 epochs.

Table 1: Training time (hours)

	Unsupervised	**Self-supervised**	**Weakly-supervised**	**Regularized**
`one_crop`	7	**1.5**	2.5	9
`multi_crop`	35	**4**	11	45

5.5 SUMMARY OF COMPARISON

Table 2: Summary of comparison. Median distance and classification metrics are given with median absolute deviations. Mean clustering metrics are given with standard deviations. All metrics satisfy *the-higher-the-better*. Top performance for each model and task is highlighted in bold.

	aug		no_aug	
	multi_crop	one_crop	multi_crop	one_crop
Unsupervised				
d(MTX, PTX)	**0.17 ± 0.00**	0.17 ± 0.00	0.16 ± 0.00	0.20 ± 0.00
D^{-1}(MTX, PTX)	**0.11 ± 0.02**	0.08 ± 0.01	0.11 ± 0.02	0.09 ± 0.01
Accuracy	0.72 ± 0.06	0.70 ± 0.06	0.72 ± 0.05	**0.75 ± 0.07**
Precision	0.80 ± 0.06	0.75 ± 0.05	0.75 ± 0.04	**0.86 ± 0.06**
Recall	0.66 ± 0.11	0.69 ± 0.10	0.72 ± 0.11	**0.65 ± 0.12**
ROAUC	0.72 ± 0.06	0.69 ± 0.06	0.70 ± 0.05	**0.75 ± 0.06**
# clusters	4 ± 2	**4 ± 2**	4 ± 2	3 ± 1
Not noise, %	93 ± 6	**93 ± 5**	94 ± 5	94 ± 5
Silhouette	0.32 ± 0.14	**0.34 ± 0.17**	0.35 ± 0.16	0.32 ± 0.08
(Davies-Bouldin)$^{-1}$	0.92 ± 0.79	**0.99 ± 0.83**	0.94 ± 0.89	0.80 ± 0.27
Self-supervised				
d(MTX, PTX)	0.27 ± 0.00	**0.24 ± 0.00**	0.25 ± 0.00	0.2 ± 0.00
D^{-1}(MTX, PTX)	0.22 ± 0.02	**0.69 ± 0.06**	0.26 ± 0.02	0.64 ± 0.09
Accuracy	**0.62 ± 0.05**	0.60 ± 0.04	0.61 ± 0.04	0.61 ± 0.05
Precision	**0.69 ± 0.05**	0.63 ± 0.04	0.69 ± 0.05	0.69 ± 0.05
Recall	**0.54 ± 0.14**	0.63 ± 0.10	0.56 ± 0.12	0.55 ± 0.13
ROAUC	**0.62 ± 0.04**	0.59 ± 0.03	0.61 ± 0.04	0.61 ± 0.05
# clusters	5 ± 3	4 ± 3	4 ± 2	**3 ± 1**
Not noise, %	93 ± 4	94 ± 4	95 ± 5	**95 ± 4**
Silhouette	0.29 ± 0.09	0.32 ± 0.06	0.34 ± 0.09	**0.34 ± 0.12**
(Davies-Bouldin)$^{-1}$	0.74 ± 0.15	0.75 ± 0.14	0.86 ± 0.47	**0.92 ± 0.63**
Weakly-supervised				
d(MTX, PTX)	-0.15 ± 0.00	**0.03 ± 0.00**	-0.18 ± 0.00	0.01 ± 0.00
D^{-1}(MTX, PTX)	0.14 ± 0.03	**1.47 ± 0.26**	0.1 ± 0.02	1.20 ± 0.19
Accuracy	0.73 ± 0.05	0.73 ± 0.05	**0.75 ± 0.05**	0.73 ± 0.05
Precision	0.73 ± 0.05	0.75 ± 0.04	**0.75 ± 0.04**	0.75 ± 0.04
Recall	0.77 ± 0.12	0.75 ± 0.11	**0.77 ± 0.11**	0.77 ± 0.10
ROAUC	0.72 ± 0.05	0.73 ± 0.05	**0.74 ± 0.05**	0.73 ± 0.05
# clusters	5 ± 4	3 ± 1	**4 ± 1**	6 ± 5
Not noise, %	90 ± 5	91 ± 7	**89 ± 8**	87 ± 7
Silhouette	0.13 ± 0.08	0.14 ± 0.09	**0.13 ± 0.08**	0.12 ± 0.07
(Davies-Bouldin)$^{-1}$	0.55 ± 0.17	0.54 ± 0.14	**0.56 ± 0.17**	0.53 ± 0.10
Regularized				
d(MTX, PTX)	0.17 ± 0.00	**0.19 ± 0.00**	0.15 ± 0.00	0.18 ± 0.00
D^{-1}(MTX, PTX)	0.12 ± 0.02	**0.08 ± 0.01**	0.09 ± 0.02	0.08 ± 0.01
Accuracy	0.73 ± 0.07	**0.76 ± 0.07**	0.70 ± 0.05	0.72 ± 0.06
Precision	0.79 ± 0.05	**0.83 ± 0.05**	0.75 ± 0.04	0.80 ± 0.06
Recall	0.70 ± 0.11	**0.68 ± 0.11**	0.66 ± 0.11	0.66 ± 0.09
ROAUC	0.73 ± 0.06	**0.76 ± 0.06**	0.69 ± 0.05	0.72 ± 0.06
# clusters	4 ± 1	4 ± 2	**3 ± 1**	4 ± 2
Not noise, %	93 ± 5	93 ± 4	**94 ± 5**	94 ± 5
Silhouette	0.32 ± 0.06	0.30 ± 0.09	**0.35 ± 0.15**	0.33 ± 0.12
(Davies-Bouldin)$^{-1}$	0.76 ± 0.20	0.77 ± 0.26	**0.99 ± 0.92**	0.94 ± 0.68

6 DISCUSSION

In this study, we used the distance-based analysis to validate and compare models. We took images of two drugs (PTX and MTX), known to be structurally and functionally similar, evaluated and compared their distances to control images in the space of learned features. However, this analysis

stays limited to the choice of drugs. Although PTX and MTX made the best example for this dataset to use *a-priori* knowledge in validation and comparison of learned features, the results can not be generalized for any pair of drugs.

A common practice to evaluate learned representations is to apply them to different tasks and datasets. Often, linear evaluation and transfer learning scenarios are tested. However, this is the case when representations are learned from multi-class general purpose datasets (e.g., ImageNet). On the contrary, biological imaging datasets are specific. It has been reported that even SOTA models trained on ImageNet drop their performance significantly on such datasets (Grill et al., 2020). In this study, we had a large imbalanced unlabelled dataset of 1.1M cell images under 693 different conditions over time. We sampled from it in the way to formulate a balanced binary classification problem, which in turn drastically limited further transfer learning applications.

To the date, no consensual measure to evaluate clustering results has been proposed (Palacio-Niño & Berzal, 2019). A number of metrics, such as Adjusted Rand Index, Silhouette score, Normalized Mutual Information, etc., are typically used together to compare results. Most metrics, however, require the ground truth labelling, which were not available in this study. Besides, the clustering itself can be approached in many different ways, using the classical or the newly developed deep-learning based algorithms (Ciortan & Defrance, 2021). In this study, we only intended to fairly compare clustering results, obtained under identical conditions (same algorithm, grid search parameters, evaluation metrics, etc.)

In this study, we have demonstrated a number of ways to analyze large biological datasets with different representation learning paradigms. Similar approaches can be applied to address actual problems in healthcare and biotech industry (e.g., deriving drug onset times, characterizing concentration-dependent pharmacodynamics, exploring opportunities for combination therapy, etc.) In this context, it is important for the scientific community to see that SOTA methods (such as BYOL) can be successfully trained on large datasets within reasonable time using limited resources.

7 Conclusion

We applied different AI paradigms to analyze a large unlabelled dataset of drug treated cancer cell lines. We implemented four different models and trained them under four different settings, combining augmentations and multi-crops. We kept the training parameters identical to ensure fair comparison of learned representations. We used Nvidia GeForce RTX 2060 with 6 GB only to train all models. The learned representations of 16 setups (model + setting) were evaluated in 3 downstream tasks: i) distance-based similarity analysis of known drugs, ii) classification of drugs versus controls, iii) clustering within cell lines. Multiple metrics were used to quantify performance on each task. We make the following observations summarizing our analysis:

- Multi-crops and augmentations generally improve performance in downstream tasks, as expected. Of 16 setups tested on 3 tasks each, only once the model with no augmentations and single crops produced the best performance.
- The self-supervised model showed very competitive performance and was the fastest to train. Strikingly, we managed to train it on the 770k dataset using a moderate GPU within 1.5 and 4 hours only (for single and multi-crops, respectively). Additionally, double augmenting resulted in improved performance on 2 of 3 downstream tasks.
- Overall, the regularized autoencoder produced the most informative features. It delivered the best scores for classification and clustering tasks and the second best for distance-based drug similarity analysis. However, it also required significantly more time to train.
- No single combination of model (architecture) and setting (augmenting and cropping strategy) consistently outperformed the others. Within each model, the top performance on 3 tasks was often shown by different settings.

Our results suggest a combination of regularized and self-supervised learning as the most promising mechanism to efficiently learn biologically meaningful representations. To achieve top performance in a particular application, we recommend to extensively evaluate the strength of domain-specific regularization, as well as augmenting and cropping strategies.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

Lorenzo Brigato, Björn Barz, Luca Iocchi, and Joachim Denzler. Tune it or don't use it: Benchmarking data-efficient image classification, 2021.

JC Caicedo, C McQuin, A Goodman, S Singh, and AE Carpenter. Weakly supervised learning of single-cell feature embeddings. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 9309–9318, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021.

SN Chandrasekaran, H Ceulemans, and JD et al. Boyd. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat. Rev. Drug Discov.*, 20:145–159, 2021.

L Chen, Y Zhai, Q He, W Wang, and M Deng. Integrating deep supervised, self-supervised and unsupervised learning for single-cell rna-seq clustering and annotation, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020b.

M Ciortan and M Defrance. Contrastive self-supervised clustering of scrna-seq data. *BMC Bioinform.*, 22:280, 2021.

WJ Godinez, I Hossain, SE Lazic, JW Davies, and X Zhang. A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics*, 33(13):2010–2019, 2017.

Peter Goldsborough, Nick Pawlowski, Juan C Caicedo, Shantanu Singh, and Anne E Carpenter. Cytogan: Generative modeling of cell images. *bioRxiv*, 2017.

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, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.*, 4(2):268–276, 2018.

Bo Hu, Ye Tang, Eric I-Chao Chang, Yubo Fan, Maode Lai, and Yan Xu. Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks. *IEEE J. Biomed. Health Inform.*, 23(3):1316–1328, 2019. ISSN 2168-2208.

OZ Kraus, BT Grys, and J et al. Ba. Automated analysis of high-content microscopy data with deep learning. *Mol. Syst. Biol.*, 13(4):924, 2017.

J Longden, X Robin, M Engel, J Ferkinghoff-Borg, I Kjær, ID Horak, MW Pedersen, and R Linding. Deep neural networks identify signaling mechanisms of erbb-family drug resistance from a continuous cell morphology space. *Cell Rep.*, 34(3):108657, 2021.

Alex X Lu, Oren Z Kraus, Sam Cooper, and Alan M Moses. Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. *PLOS Comp. Biol.*, 15 (9):1–27, 2019. doi: 10.1371/journal.pcbi.1007348.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images, 2020.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

E Moen, D Bannon, and T et al. Kudo. Deep learning for cellular image analysis. *Nat. Methods*, 16: 1233–1246, 2019.

G Nguyen, S Dlugolinsky, and M et al. Bobák. Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. *Artif. Intell. Rev.*, 52:77–124, 2019.

Sarala Padi, Petru Manescu, Nicholas Schaub, Nathan Hotaling, Carl Simon, Kapil Bharti, and Peter Bajcsy. Comparison of artificial intelligence based approaches to cell function prediction. *Inform. Med. Unlocked*, 18:100270, 2020. ISSN 2352-9148.

Julio-Omar Palacio-Niño and Fernando Berzal. Evaluation metrics for unsupervised learning algorithms, 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

Michael C Robitaille, Jeff M Byers, Joseph A Christodoulides, and Marc P Raphael. A self-supervised machine learning approach for objective live cell segmentation and analysis. *bioRxiv*, 2021.

Zhiwei Rong, Qilong Tan, Lei Cao, Liuchao Zhang, Kui Deng, Yue Huang, Zheng-Jiang Zhu, Zhenzi Li, and Kang Li. Normae: Deep adversarial learning model to remove batch effects in liquid chromatography mass spectrometry-based metabolomics data. *Anal. Chem.*, 92(7):5082–5090, 2020.

M Ruszkowski, B Sekula, and A et al. Ruszkowska. Structural basis of methotrexate and pemetrexed action on serine hydroxymethyltransferases revealed using plant models. *Sci. Rep.*, 9:19614, 2019.

Diogo Santos-Pata, Adrián F Amil, Ivan Georgiev Raikov, César Rennó-Costa, Anna Mura, Ivan Soltesz, and Paul FMJ Verschure. Entorhinal mismatch: A model of self-supervised learning in the hippocampus. *iScience*, 24(4):102364, 2021. ISSN 2589-0042.

Ruiyi Zhang, Yunan Luo, Jianzhu Ma, Ming Zhang, and Sheng Wang. scpretrain: Multi-task self-supervised learning for cell type classification. *bioRxiv*, 2020.

A QUALITY OF IMAGE RECONSTRUCTIONS

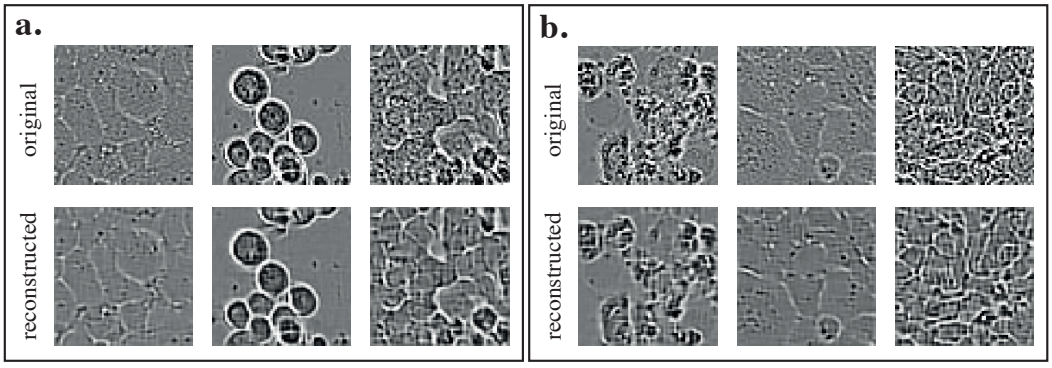

Figure 6: Random examples of reconstructed and original images for the unsupervised (**a**) and the regularized (**b**) models. Regularization did not harm the quality of reconstructions. The learning capacity of the CNN backbone was sufficient to capture normal and altered morphology of the cells.

B CLUSTERING OF HCT CELL LINE REPRESENTATIONS

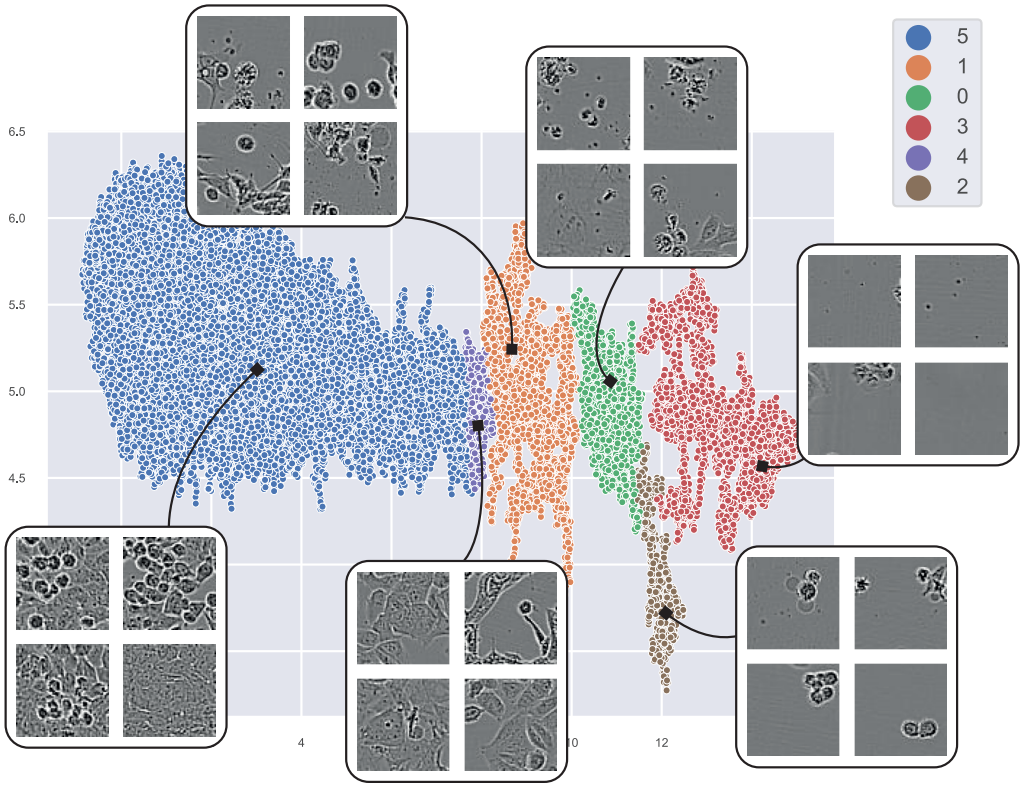

Figure 7: Clustering example of 20480 images (HCT15 cell line) with random cluster representatives. Each point is a 2D UMAP embedding of the learned image representations (self-supervised model). Clusters found by HDBSCAN are highlighted in colors. The left cluster (blue) contains drugs of no effect on HCT15. The right cluster (red) contains the drugs of the strongest effect.