Self-Supervised Learning of Phenotypic Representations from Cell Images with Weak Labels

Jan Oscar Cross-Zamirski
DAMTP & AstraZeneca
University of Cambridge
Cambridge CB3 0WA
jc856@cam.ac.uk

Guy Williams
Discovery Sciences, R&D
AstraZeneca
Cheshire, SK10 4TG
guy.williams2*

Elizabeth Mouchet
Discovery Sciences, R&D
AstraZeneca
Cheshire, SK10 4TG
elizabeth.mouchet*

Carola-Bibiane Schönlieb
DAMTP
University of Cambridge
Cambridge CB3 0WA
cbs31@cam.ac.uk

Riku Turkki †
Discovery Sciences, R&D
AstraZeneca
Gothenburg, 431 50 Mölndal
riku.turkki*

Yinhai Wang †
Discovery Sciences, R&D
AstraZeneca
Cambridge, CB4 0WG
yinhai.wang*

*@astrazeneca.com †Equal contribution

Abstract

We propose WS-DINO as a novel framework to use weak label information in learning phenotypic representations from high-content fluorescent images of cells. Our model is based on a knowledge distillation approach with a vision transformer backbone (DINO), and we use this as a benchmark model for our study. Using WS-DINO, we fine-tuned with weak label information available in high-content microscopy screens (treatment and compound) and achieve state-of-the-art performance in not-same-compound mechanism of action prediction on the BBBC021 dataset (98%), and not-same-compound-and-batch performance (96%) using the compound as the weak label. Our method bypasses single cell cropping as a pre-processing step, and using self-attention maps we show that the model learns structurally meaningful phenotypic profiles.

1 Introduction

Deep learning methods for phenotyping cells from microscopy images have advanced rapidly alongside cellular imaging technology [1]. Profiles captured from images can be used to quantify activity and perform downstream tasks such as mechanism of action (MOA) prediction of cells treated with compounds of interest. These methods hold promise for accelerating drug discovery pipelines [2].

Typically, convolutional neural networks (CNNs) have been the method of choice, however these models can struggle to generalise to new treatments, and can be limited by label information [3]. Rarely are they interpretable, and this may contribute to how machine learning methods have not reached their potential in clinical use [4]. Across medical imaging fields, vision transformers (ViTs) [5] and self-supervised [6] methods have risen as a viable alternative to CNNs with labels [7] [8].

In this study we make the following contributions:

- We introduce Weakly Supervised DINO (WS-DINO), a framework to incorporate weak label information into a self-supervised knowledge distillation construction based on the DINO algorithm [9].

Preprint. Under review.

- We implement WS-DINO, and with the BBBC021 dataset [10] [11] we learn representations with two weak labels - treatment and compound. We achieve state-of-the-art results for not-same-compound (98%), and not-same-compound-and-batch (96%) mechanism of action prediction using the compound as the weak label. We show the learnt representations are biologically meaningful and based on cellular structure with self-attention maps.

2 Background

2.1 Image-based profiling

Imaged-based profiling [2] is the process of extracting features from high-content images (HCI) [12] of cells. The features extracted from these kinds of images are used to build profiles which can assess bioactivity and MOA of the cells in response to treatments with chemical compounds and/or genetic perturbations. In drug discovery these phenotypic profiles are used to compare new treatments with known ones, and downstream applications include routine lead compound identification, drug target screening through CRISPR technology and toxicity assessment [13].

Most phenotypic representation learning methods are based on single cell segmentation [14] [15]. When applied to large datasets this can lead to high computational demands in pre-processing images with segmentation algorithms [16]. Working with single-cells requires aggregation to the population level and may necessitate further computational requirements for corrections due to cellular heterogeneity [17]. In high-content imaging there is metadata which can be considered a *weak* label, examples being treatment compound and concentration. In previous work a classification network was trained predict the weak label as an auxiliary task, with a frozen layer of the network forming the phenotypic representation used in downstream tasks such as MOA prediction [18].

Unsupervised methods such as self-supervised learning constructions [6] using a contrastive loss [19] have been very successful. Self-supervised algorithms such as SwAV [7] and DINO [9] give consistently richer embeddings for downstream tasks compared to a pretrained supervised baseline model in medical imaging [20]. Self-supervised models have performed very well learning feature represenations from single cell microscopy images [3] [14], and recently on the image-level with a multi-scale approach [21].

2.2 DINO - self-supervised learning with knowledge distillation

DINO from Caron *et al.* [9] is a self-supervised learning approach which incorporates aspects of knowledge distillation [22]. Uniquely combining self-supervised learning with a Vision Transformer (ViT) [5] backbone, DINO can learn features shown to perform well at *k*-NN clustering tasks.

DINO consists of student network g_{θ_s}, parameterized by θ_s, trained with set of image crops to match the teacher network g_{θ_t}, parameterized by θ_t, which sees a different set of crops. For each image x a set of views V is generated which contains the two global crops $x^{g,1}$ and $x^{g,2}$ as well as eight local crops. The student is passed the set of global and local crops V, while the teacher sees only the global crops. The student network is trained to maximise the agreement between the outputs of g_{θ_s} and g_{θ_t}. To achieve this, the parameters of θ_s are found by minimizing the cross-entropy loss:

$$\min_{\theta_s} \sum_{x \in \{x^{g,1}, x^{g,2}\}} \sum_{\substack{x' \in V \\ x' \neq x}} H(P_t(x), P_s(x')) \tag{1}$$

Where $H(a, b) = -a \log b$ and P_s and P_t are the probability distributions of the student and teacher respectively (defined in the Appendix). Teacher weights θ_t are frozen during each epoch of student training and updated iteratively with an exponential moving average based on the previous weights of the student network with the formula: $\theta_t \leftarrow \lambda\theta_t + (1 - \lambda)\theta_s$, where λ is the momentum parameter.

Both the student and the teacher network have a ViT [5] backbone. The inputs are small patches of fixed size ($N \times N$) in a non-overlapping grid, which are embedded (alongside a positional embedding) with a linear layer. An additional learnable [CLS] token with a projection head output h is added to the embeddings and sent to the transformer encoder. The attention mechanism allows the ViT to

synthesise information across the whole image as self-attention layers in the Transformer globally update the attention token embeddings.

3 Methods

3.1 WS-DINO

We propose a **W**eakly **S**upervised form of self-**di**stillation with **no** labels (WS-DINO) as an adaptation to the DINO algorithm. The aim of the approach is to retain the benefits of self-supervised learning with knowledge distillation while also incorporating the weak label information as part of the training.

We introduce the notation x_{i,y_i} to represent the ith field of view of a fluorescent channel in the dataset which has been treated with treatment or compound y_i - the weak label. The superscript contains the crop information: g for global and l for local crops. When generating the sets of different views V_t (seen by teacher) and V_s (seen by student) for training, we enforce the constraint that the global and local crops are sampled from different images with the same weak label. We define the sets V_t and V_s for the randomly sampled ordered pair (i, j) where $i \neq j$ and $y_i = y_j$:

$$V_t = \{x^{g,1}_{i,y_i}, x^{g,2}_{i,y_i}\} \tag{2}$$

$$V_s = V_t \cup \left\{ x^{l,k}_{j,y_j} : \quad k \in \{1, ..., n\} \right\} \tag{3}$$

Where the superscript details the kth local crop of n total crops (default: $n = 8$). Sampling the local crops from a different image is in contrast to sampling random crops with different augmentations from the same image, which is the method of DINO. For WS-DINO we minimize the loss:

$$\min_{\theta_s} \sum_{x \in V_t} \sum_{\substack{x' \in V_s \\ x' \neq x}} H(P_t(x), P_s(x')) \tag{4}$$

In this section we have presented the key details of our adaptation to the DINO algorithm from the original paper [9]. We provide implementation details in section 3.4 and in our GitHub repository. [1]

3.2 Dataset and augmentation

We evaluated our experiments with the publicly available BBBC021 dataset [10], which consists of MCF-7 breast cancer cells exposed to several chemical compounds for 24h. The cells are stained with three fluorescent labels: DNA, F-actin, and β-tubulin. For training and evaluation, we used the subset which has been labelled for MOA evaluation [11] which consists of 12 unique MOA, 38 unique compounds and 103 unique treatments (compound/concentration pairs) across 10 experimental batches. There are 1320 3-channel images in the dimethyl sulfoxide (DMSO) treated control group, and 2528 3-channel images with MOA annotations. Several studies have used the annotated dataset to evaluate classical and machine learning strategies [11] [14] [15] [18] [21] [23] [24] [25] [26] [27].

Pre-processing: To remove systematic variations in pixel intensities across each field of view, we used a CellProfiler [16] pipeline to perform an illumination correction. A smoothing function with a filter size of 320 pixels was used to create an illumination correction function. Each image was corrected by dividing all pixel intensities by the illumination correction function. This methodology is consistent with best practice established by the JUMP Cell Painting Consortium [28]. After correction, images were resized using bicubic interpolation to a size of 640×512 pixels. A maximum pixel value cut-off of $10,000$ was enforced before normalizing each image to have a mean of 0 and standard deviation of 1. The scripts to replicate the pre-processing are available in our GitHub repository.

Post-processing: Next we take the median feature embedding from four 224×224 crops around the centre of each image (equivalent to a 448×448 pixel centre crop split into four non-overlapping

[1] `https://github.com/crosszamirski/WS-DINO`

crops). To aid in correcting for batch effects (which can be significant in HCI screens) and to capture the range of variation of unperturbed cells, we applied typical variation normalization (TVN) [15] as a post-processing correction. TVN is a technique where a principal component analysis (PCA) transformation without dimensionality reduction (whitening) is learned using the DMSO embeddings (aggregated to one embedding per image from the median of four crops around the centre). The mean of each of the field-level embeddings is taken across a plate, followed by median aggregation to treatment level resulting in 103 embeddings - in line with the studies following Ando *et al.* [15].

3.3 Performance validation and evaluation

We evaluated the corrected and aggregated feature embeddings with two metrics ubiquitous in representation learning studies using the BBBC021 dataset: not-same-compound (NSC) matching [11] and not-same-compound-and-batch (NSCB) matching [15] with the MOA labels. NSC matching is a 1-Nearest-Neighbour (1-NN) match for each given treatment to the nearest neighbour in representation space which is **not** of the same compound. Cosine distance was used as the distance measure. NSCB matching is the same method as NSC matching with the additional constraint to exclude treatments from the same compound **and** batch as the given treatment. A small number of treatments have a MOA label only present in a single batch and these were excluded in NSCB evaluation.

3.4 Training

For all the models the following variables are kept constant: network architecture (ViT-S/8 backbone with 3-layer multi-layer perceptron head), size and number of crops: 2 global 224×224 random resize crops (range = $0.1 - 0.2$ of full size image) and 8 local 96×96 random resize crops (range = $0.04 - 0.08$), teacher momentum = 0.99, and batch size = 16. There was no batch normalization, weight decay or gradient clipping. The optimizer was adamW [29]. We applied horizontal and vertical flips (with a probability of 0.5 per sample) as augmentations in training. Learning rate and temperature parameters are controlled with adaptive cosine schedules with linear warm-up. After 10 epochs of linear warmup, the learning rate = 4×10^{-6} and the teacher temperature = 0.04. The learning rate was decayed to 3×10^{-6} over 400 epochs, and the teacher temperature kept constant after warmup. All other parameters unless otherwise stated are default DINO parameters. A separate model was trained for each of of the three channels. Every model was initialized with weights from DINO trained on RGB ImageNet [9] [30] (grayscale images were converted to RGB). Embeddings of size 384 were extracted for each channel then concatenated to a size of 1152 followed by L2 normalization. We provide a full PyTorch implementation in our GitHub repository.

4 Results and Discussion

As a baseline we trained the unsupervised DINO network [9]. First, we evaluated with ImageNet weights only (transfer learning). Next we fine-tuned the model on the BBBC021 annotated dataset without weak labels. We then trained WS-DINO with the same network and parameters. Two weak labels y were evaluated: treatment and compound. Additionally we trained with MOA as the label y, although this cannot be considered a weak label. The images in training were sampled using a weighted random sampler enforcing even sampling of images from each weak label group.

Table 1: A summary of the results of our experiments. Further details of mean NSC and NSCB scores in training with a discussion of best epoch selection is provided in the Appendix.

Model	Weak Label	NSC	NSCB
DINO with ImageNet weights only	None	91%	82%
DINO finetuned on BBBC021	None	92%	95%
WS-DINO finetuned on BBBC021	Treatment	92%	90%
	Compound	**98%**	**96%**
	MOA	100%	100%

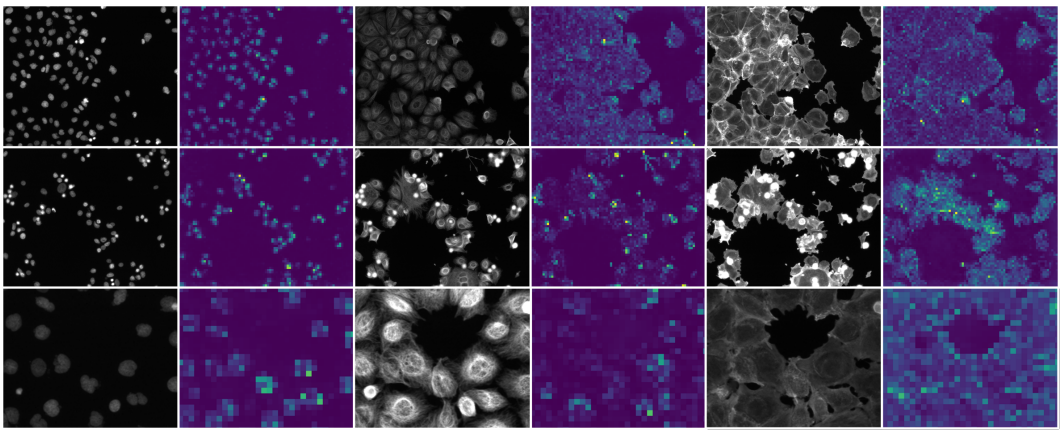

Figure 1: Examples of BBBC021 images with their coupled self-attention maps. Left two columns: DNA channel. Middle two columns: β-tubulin channel. Right two columns: F-actin channel. The first two rows show the images at full size, and the bottom row displays a section of an image zoomed in by a scale factor of four. Produced from WS-DINO weights with compound as the weak label.

WS-DINO achieves state-of-the-art results on BBBC021 using the compound as the weak label, outperforming all known previous approaches using this dataset in NSC and NSCB MOA prediction (Table 1 and Appendix). It is notable that using treatment as the weak label does not improve upon the unsupervised approach (DINO). In this dataset all images of the same treatment are from the same batch, however this is not true for the different compounds and MOAs which are expressed with different images across multiple batches. Hence we propose that WS-DINO may provide effective batch correction for datasets with weak label classes with images representing multiple batches.

We present self-attention maps (Figure 1) which reveal the segmentation properties of the algorithm. Such visualisations increase confidence in the model by demonstrating the network is learning biologically and structurally meaningful features.

We use MOA as a *psuedo-weak* label as a proof-of-concept for our method. However, we propose future work to evaluate MOA prediction using datasets with partial MOA labels. One advantage of WS-DINO is that it is adaptable to datasets with some known MOA information, a property of many real datasets used in drug discovery.

WS-DINO synthesises a powerful self-supervised framework with a way to implicitly incorporate the informative weak labels in learning phenotypic representations. Our method can contribute to accelerating drug discovery pipelines by clustering phenotypes in a biologically meaningful hierarchy. The proposed framework is general and the method is not specific to DINO. The sampling of pairs of images with the same weak label is a concept adaptable to other networks. Self-supervised methods are an active field of research and new algorithms have outperformed DINO on ImageNet classification, for example Masked Siamese Networks [31]. Future work could incorporate these studies which would be adaptable in a very similar way to WS-DINO. Finally, we will expand this research to other datasets with more classes and labels, as well as different downstream tasks.

Acknowledgments and Disclosure of Funding

Jan Oscar Cross-Zamirski is funded by the UKRI-BBSRC DTP (UK Research and Innovation and the Biotechnology and Biological Science Research Council Doctoral Training Partnership) studentship grant 21-3688 and AstraZeneca.

Carola-Bibiane Schönlieb acknowledges support from the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC grants EP/S026045/1 and EP/T003553/1, EP/N014588/1, EP/T017961/1, the Wellcome Trust 215733/Z/19/Z and 221633/Z/20/Z, the Leverhulme Trust project Unveiling the invisible, the European Union Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 777826 NoMADS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute.

Authors Guy Williams, Elizabeth Mouchet, Riku Turkki and Yinhai Wang are employees of AstraZeneca. AstraZeneca provided the funding for this research and provided support in the form of salaries for the authors, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. We would also like to acknowledge the use of the Scientific Computing Platform (SCP) within AstraZeneca.

References

[1] Pratapa, A., Doron, M., & Caicedo, J. C. (2021). Image-based cell phenotyping with deep learning. *Current opinion in chemical biology*, 65, 9–17. https://doi.org/10.1016/j.cbpa.2021.04.001

[2] Chandrasekaran, S.N., Ceulemans, H., Boyd, J.D. & Carpenter, A. E. (2021). Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nat Rev Drug Discov* 20, 145–159. https://doi.org/10.1038/s41573-020-00117-w

[3] Lu, A.X., Kraus, O.Z., Cooper, S. & Moses, A.M. (2019). Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. *PLoS Comput Biol* 15(9): e1007348. https://doi.org/10.1371/journal.pcbi.1007348

[4] Weissler, E., Naumann, T., Andersson, T., Ranganath, R.,Elemento, O., Luo, Y. *et al.* (2021). The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*. 22. http://dx.doi.org/10.1186/s13063-021-05489-x

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. *et al.* (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint:* https://doi.org/10.48550/arXiv.2010.11929

[6] Doersch, C., & Zisserman, A. (2017). Multi-task Self-Supervised Visual Learning *arXiv preprint:* https://doi.org/10.48550/arxiv.1708.07860

[7] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. & Joulin, A. (2020) Unsupervised learning of visual features by contrasting cluster assignments. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* https://doi.org/10.48550/arXiv.2006.09882

[8] Matsoukas, C., Haslum, J.F., Söderberg, M., & Smith, K. (2021). Is it Time to Replace CNNs with Transformers for Medical Images? *arXiv preprint:* https://doi.org/10.48550/arxiv.2108.09038

[9] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. *Proceedings of the International Conference on Computer Vision (ICCV).* https://doi.org/10.48550/arXiv.2104.14294

[10] Caie, P.D., Walls, R.E., Ingleston-Orme, A., Daya, S., Houslay, T., Eagle, R., Roberts, M.E. & Carragher, N.O. (2010). High-content phenotypic profiling of drug response signatures across distinct cancer cells. *Molecular cancer therapeutics 9(6)*, 1913–1926. https://doi.org/10.1158/1535-7163.mct-09-1148

[11] Ljosa, V., Caie, P.D., Ter Horst, R., Sokolnicki, K.L., Jenkins, E.L., Daya, S., Roberts, M.E., Jones, T.R., Singh, S., Genovesio, A. *et al.* (2013). Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *Journal of biomolecular screening 18(10)*, 1321–1329. https://doi.org/10.1177/1087057113503553

[12] Bickle, M. (2010). The beautiful cell: high-content screening in drug discovery. *Anal. Bioanal. Chem.* 398, 219–226. https://doi.org/10.1007/s00216-010-3788-3

[13] Trapotsi, M.-A., Mouchet, E., Williams, G., Monteverde, T., Juhani, K., Turkki, R. *et al.* (2022). Cell Morphological Profiling Enables High-Throughput Screening for PROteolysis TArgeting Chimera (PROTAC) Phenotypic Signature. *ACS Chemical Biology* 2022 17 (7), 1733-1744. https://doi.org/10.1021/acschembio.2c00076

[14] Perakis, A., Gorji, A., Jain S., Chaitanya, K., Rizza, S., & Konukoglu, E. (2021). Contrastive Learning of Single-Cell Phenotypic Representations for Treatment Classification. *arXiv preprint:* https://doi.org/10.1007%2F978-3-030-87589-3_58

[15] Ando, D.M., McLean, C.Y. & Berndl, M. (2017). Improving phenotypic measurements in high-content imaging screens. *BioRxiv preprint* p. 161422. https://doi.org/10.1101/161422

[16] McQuin, C., Goodman, A., Chernyshev, V., Kamentsky, L., Cimini, B.A. *et al.* (2018). CellProfiler 3.0: Next-generation image processing for biology. *PLOS Biology* 16(7): e2005970. https://doi.org/10.1371/journal.pbio.2005970

[17] Rohban, M. H., Abbasi, H. S., Singh, S. & Carpenter, A. E. (2019). Capturing single-cell heterogeneity via data fusion improves image-based profiling. *Nat. Commun.* 10, 2082. https://doi.org/10.1038/s41467-019-10154-8

[18] Caicedo, J. C., McQuin, C., Goodman, A., Singh, S. & Carpenter, A. E. (2018). Weakly Supervised Learning of Single-Cell Feature Embeddings. *Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, 9309–9318. https://doi.org/10.1109/cvpr.2018.00970

[19] Hadsell, R., Chopra S., & LeCun Y. (2006). Dimensionality Reduction by Learning an Invariant Mapping, *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 1735-1742. https://doi.org/10.1109/CVPR.2006.100.

[20] Truong, T., Mohammadi, S. & Lenga, M. (2021). How Transferable are Self-supervised Features in Medical Image Classification Tasks? *Proceedings of Machine Learning for Health*, in *Proceedings of Machine Learning Research* 158:54-74. https://doi.org/10.48550/arXiv.2108.10048

[21] Janssens, R., Zhang, X., Kauffmann, A., de Weck, A., & Durand, E. Y. (2021). Fully unsupervised deep mode of action learning for phenotyping high-content cellular images. *Bioinformatics (Oxford, England)*, btab497. Advance online publication. https://doi.org/10.1093/bioinformatics/btab497

[22] Hinton, G., Vinyals, O. & Dean, J. (2015). Distilling the Knowledge in a Neural Network. *arXiv preprint:* https://doi.org/10.48550/arxiv.1503.02531

[23] Singh, S., Bray, M. A., Jones, T. R., & Carpenter, A. E. (2014). Pipeline for illumination correction of images for high-throughput microscopy. *Journal of microscopy*, 256(3), 231–236. https://doi.org/10.1111/jmi.12178

[24] Godinez, W. J., Hossain, I., Lazic, S. E., Davies, J. W., & Zhang, X. (2017). A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics (Oxford, England)*, 33(13), 2010–2019. https://doi.org/10.1093/bioinformatics/btx069

[25] Pawlowski N., Caicedo J.C., Singh S., Carpenter A.E. & Storkey A. (2016). Automating morphological profiling with generic deep convolutional networks. *BioRxiv preprint:* https://doi.org/10.1101/085118

[26] Lafarge, M.W., Caicedo, J.C., Carpenter, A.E., Pluim, J.P., Singh, S. & Veta, M. (2019). Capturing Single-Cell Phenotypic Variation via Unsupervised Representation Learning. *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning, PMLR*. https://pubmed.ncbi.nlm.nih.gov/35874600

[27] Goldsborough, P., Pawlowski, N., Caicedo, J.C., Singh, S. & Carpenter, A.E. (2017). CytoGAN: Generative Modeling of Cell Images. *BioRxiv preprint:* https://doi.org/10.1101/227645

[28] Cimini, B.A., Chandrasekaran, N., Kost-Alimova, M., Miller, L., Goodale, A., Fritchman, B. *et al.* (2022). Optimizing the Cell Painting assay for image-based profiling. *BioRxiv preprint:* https://doi.org/10.1101/2022.07.13.499171

[29] Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. *ICLR 2019*. https://doi.org/10.48550/arXiv.1711.05101

[30] Deng, J., Dong, W., Socher, R., Li, L.-J., Kai, L. & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition.* https://doi.org/10.1109/CVPR.2009.5206848

[31] Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F. Vincent, P., Joulin, A., Rabbat, M. & Ballas, N. (2022). Masked Siamese Networks for Label-Efficient Learning. *arXiv preprint:* https://doi.org/10.48550/arXiv.2204.07141

A Probability distributions of student and teacher

The probability distributions for the student P_s and teacher P_t are defined by:

$$P_s(x)^{(i)} = \frac{\exp\left(g_s(x)^{(i)}/\tau_s\right)}{\sum_{k=1}^{K} \exp\left(g_s(x)^{(k)}/\tau_s\right)}$$

for K dimensions and temperature parameters $\tau_s > 0$ and $\tau_t > 0$ [9].

B Further training details

The hyperparameters are fixed across all of the models we train. Many of the parameters are the default values from the original DINO implementation. We optimised the remaining parameters such as learning rate with a small subset of BBBC021 trained without weak labels. In our final models the batch size is limited by GPU memory - a batch size of 16 is small and most likely a larger batch size would improve performance. The models were trained with two GPUs on the AstraZeneca Scientific Computing Platform with a maximum allocation of 32G memory for each GPU.

Table 2: The results of training with NSC and NSCB scores of the best performing epoch as well as the mean NSC and NSCB scores between 50 and 250 epochs of training each model.

Model	Weak Label	Best NSC/NSCB	Mean NSC/NSCB
DINO finetuned on BBBC021	None	92% / 95%	90% / 90%
WS-DINO finetuned on BBBC021	Treatment	92% / 90%	89% / 83%
	Compound	**98% / 96%**	96% / 93%
	MOA	100% / 100%	99% / 99%

We determined the best epoch by following the method of Janssens *et al.* [21]. We calculated NSC for each model and selected the epoch with the highest score. This method was chosen to allow a comparison of our model to other studies, however we note there are drawbacks to selecting best epoch with MOA information. In Table 2 the mean values of NSC and NSCB over 200 epochs of training are displayed for each model, and the results show improved performance for NSC and NSCB MOA matching was observed over multiple epochs of training using both compound and MOA as the weak label. The values are lower than the best values as this range contains both under- and over-fit models. We provide epoch data and training logs in our GitHub repository.

C Further results

Table 3: A comparison of our method to other selected studies using the BBBC021 annotated dataset.

Type	Method	Reference	Single Cell	NSC	NSCB
Weakly Supervised	**WS-DINO**	**This work**	**No**	**98%**	**96%**
	CNN with Mixup	Caicedo *et al.* 2018 [18]	Yes	95%	89%
Classical Features	CellProfiler	Singh *et al.* 2014 [23]	Yes	90%	85%
	Factor Analysis	Ljosa *et al.* 2013 [11]	Yes	94%	77%
Supervised	Multiscale-CNN	Godinez *et al.* 2017 [24]	No	93%	N/A
Unsupervised	Contrastive Learning	Perakis *et al.* 2021 [14]	Yes	96%	95%
	UMM Discovery	Janssens *et al.* 2020 [21]	No	97%	85%
	VAE+	Lafarge *et al.* 2019 [26]	Yes	93%	82%
	CytoGAN: LSGAN	Goldsborough *et al.* 2017 [27]	No	68%	N/A
Transfer Learning	Deep Metric Network	Ando *et al.* 2017 [15]	Yes	96%	95%
	Inception V3 Pretrained	Pawlowski *et al.* 2016 [25]	No	91%	N/A

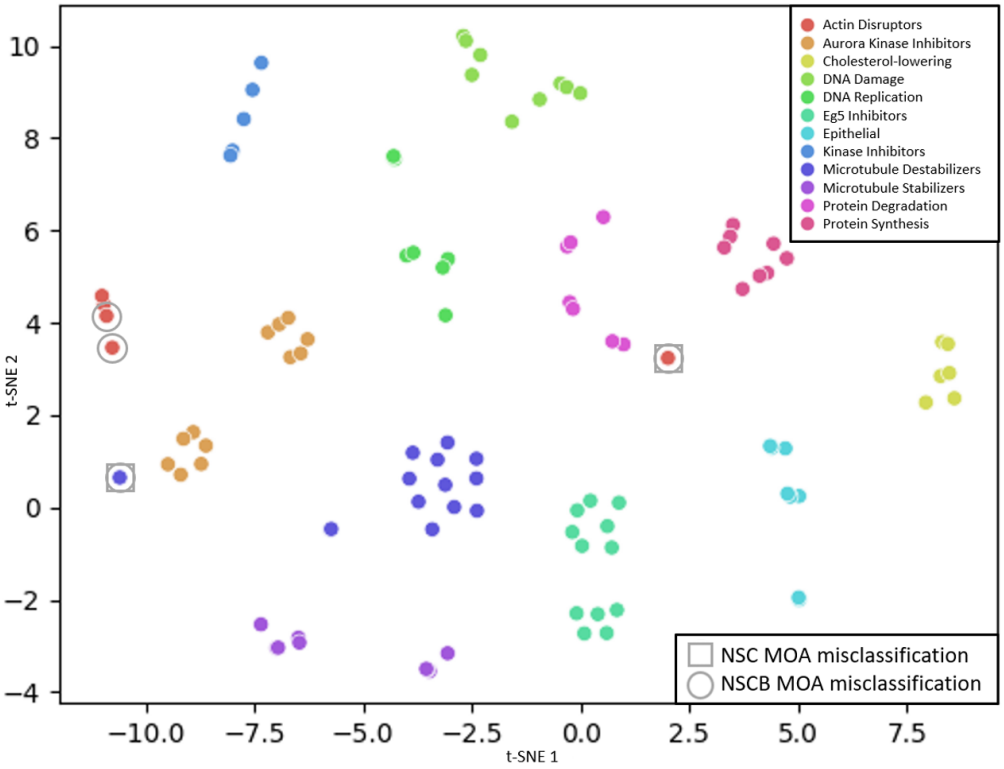

Figure 2: Two-dimensional t-SNE plot of each aggregated treatment feature from 200 epochs of training WS-DINO with compound as the weak label: 98% NSC and 96% NSCB MOA classification.