Supplementary Material

A. Dataset

A.1. Further imaging details

In the Cell Painting assay cell phenotypes are captured with six generic fluorescent dyes and imaged across five channels. The assay is designed to visualise eight cellular components: nucleus (DNA channel), endoplasmic reticulum (ER channel), nucleoli, cytoplasmic RNA (RNA channel), actin, Golgi, plasma membrane (AGP channel) and mitochondria (Mito channel) [5].

For the data in this study U2-OS cells were incubated in 5µM compounds for 48h, then fixed and stained according to the updated Cell Painting protocol [13]. Plates were imaged on a CellVoyager CV8000 (Yokogawa) with a water-immersion 20× objective (NA 1.0). Excitation and emission wavelengths were as follows for fluorescent channels: DNA (ex: 405nm, em: 445/45nm), ER (ex: 488nm, em: 525/50nm), RNA (ex: 488nm, em: 600/37nm), AGP (ex: 561nm, em: 600/37nm) and Mito (ex: 640nm, em: 676/29nm). The three brightfield images were acquired from different focal z-planes; within, 4µm above and 4µm below the focal plane. Images were saved as 16-bit .tiff files with 2 × 2 binning (998 × 998 pixels).

A.2. Active subset selection

In Fig. 4 we provide an overview of how the active subset was selected. We identified three groups of treated cells based on their CellProfiler features across all 10 plates: the active, partially active and inactive groups. The inactive subset overlapped with the negative control (DMSO) subset. Fig. 4E provides an intuitive visualisation of our choice to train the model with the active subset only. Our results using the entire plate show that using labels which do not correspond to structural and biological differences will reduce the generated image quality. We provide examples of lower quality images in Section B.

Although we do not use it for selecting our subsets, we present a visualisation of Grit scores for our dataset in Fig. 4. Grit is a calculation used in image-based-prifling to define how different a perturbation or compound is from the DMSO controls (`https://github.com/broadinstitute/grit-benchmark`).

B. Further results and figures

B.1. Cell Painting feature breakdown

We present CellProfiler feature correlation matrices between the features extracted from model predicted images and the features from the ground truth Cell Painting in Fig. 5, which are presented as heatmaps. We compare three models: the unlabelled model, and models with perturbation and target as the label for the active subset. The labels were used in both training through AdaGN and sampling with classifier guidance.

CellProfiler features are categorised as different feature groups (texture, radial distribution, intensity, granularity. colocalization, neighbours and area/shape) across the cells, cytoplasm and nuclei [9]. The correlation heatmaps present features after standard feature selection (total of 635 features), which includes dropping highly correlated features and zero-value features. Some feature groups in certain channels have no remaining features after feature selection (nan).

B.2. Brightfield vs Cell Painting

In this study it is notable how well the brightfield images perform in the transfer learning tasks for target matching. This was a surprising result given the limited studies in the literature which employ brightfield for image-based profiling. Although this behaviour could be unique to our dataset, these result pose a challenge to the utility of (fluorescent) label-free Cell Painting methods, which can be computationally intensive and may not necessarily outperform the brightfield modality in its own right.

There are a number of advantages to imaging and profiling without fluorescent staining. Brightfield imaging is cheaper, requires minimal preparation, and does not damage the cells with photo- or cyto-toxic effects. In fluorescent staining, certain combinations of dyes are restricted due to the particular wavelength the dye can be imaged at (spectral overlap). These technical limitations can hinder the ability of the scientist to capture morphological information from the unstainable subcellular compartments. Because of this, there is interest in using cheaper, quicker, less damaging alternatives such as brightfield to perform high-throughput screening and image-based profiling.

We investigate some of the quantitative and qualitative differences between brightfield, Cell Painting and our predicted Cell Painting (from brightfield) in this section.

Overlap in matching target predictions

Firstly, we compare the overlap of the specific matching targets in each of the feature spaces of the different sets of images produced by the models. We also compare the ground truth Cell Painting and brightfield against the model predictions. For two or three models, this would be visualised with a Venn diagram. As we have more than three models to compare, we present the overlapping matching target predictions between models as two matrices in Figs. 6 and 7. We used the active subset study for this analysis.

We find that there is a good overlap between the matching targets found by Cell Painting (both CellProfiler and DINO) and brightfield. When using the perturbation as the

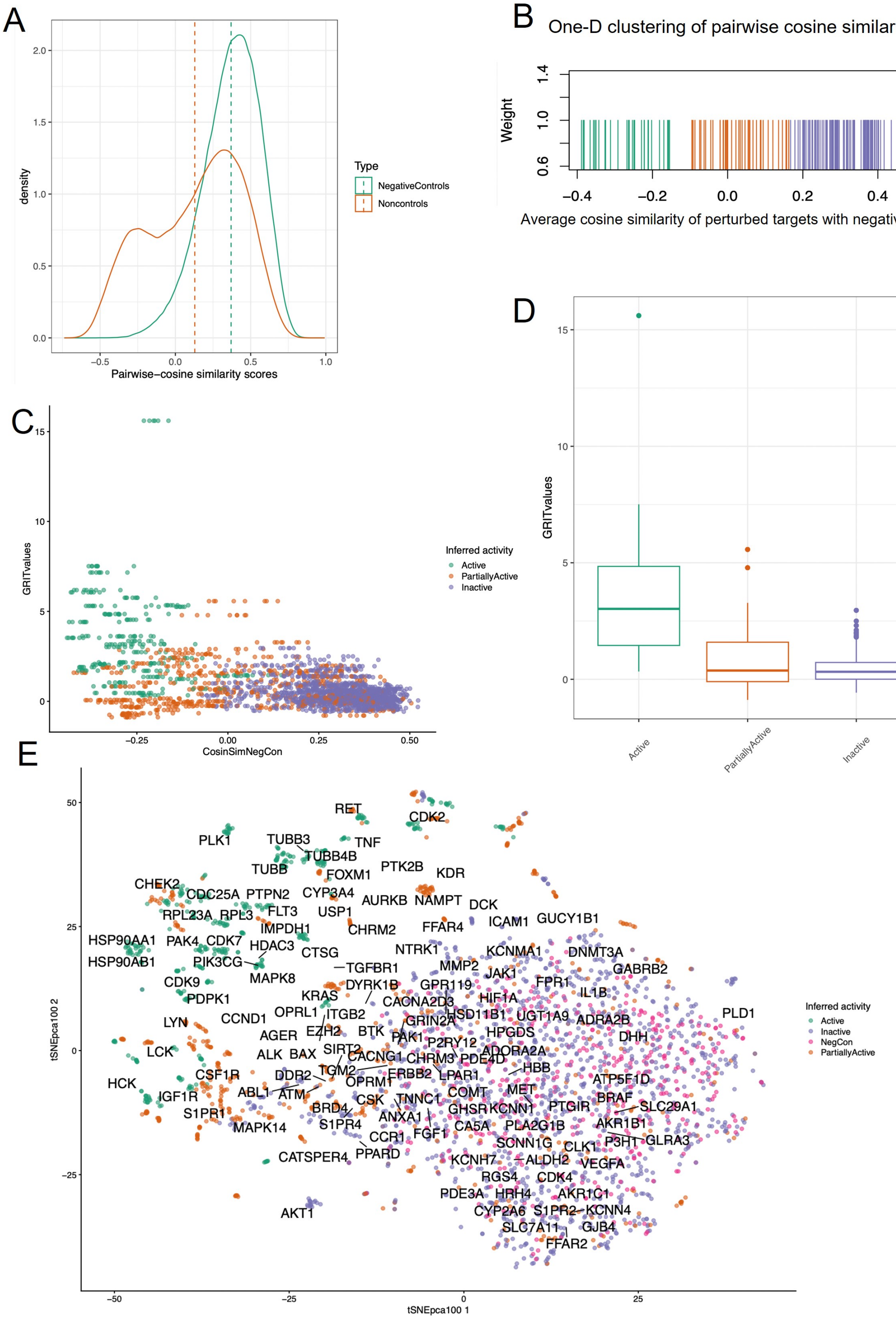


Figure 4: Inferring the target activity with the ground truth Cell Painting CellProfiler features. **A.** The distribution of all the pairwise cosine similarity scores derived from the top 100 PCA dimensions across the negative controls and the drug treatments. **B.** One-dimensional K-means clustering of the average cosine similarity metric computed between the targets and negative controls. **C.** Scatter plot of the Grit values computed for each target and the corresponding cosine similarity metric calculated from the negative controls. **D.** Box plot depicting Grit values across inferred target activity. **E.** Two-dimensional t-SNE plot of all the 10 TARGET-2 plates colored based on the inferred target activity.

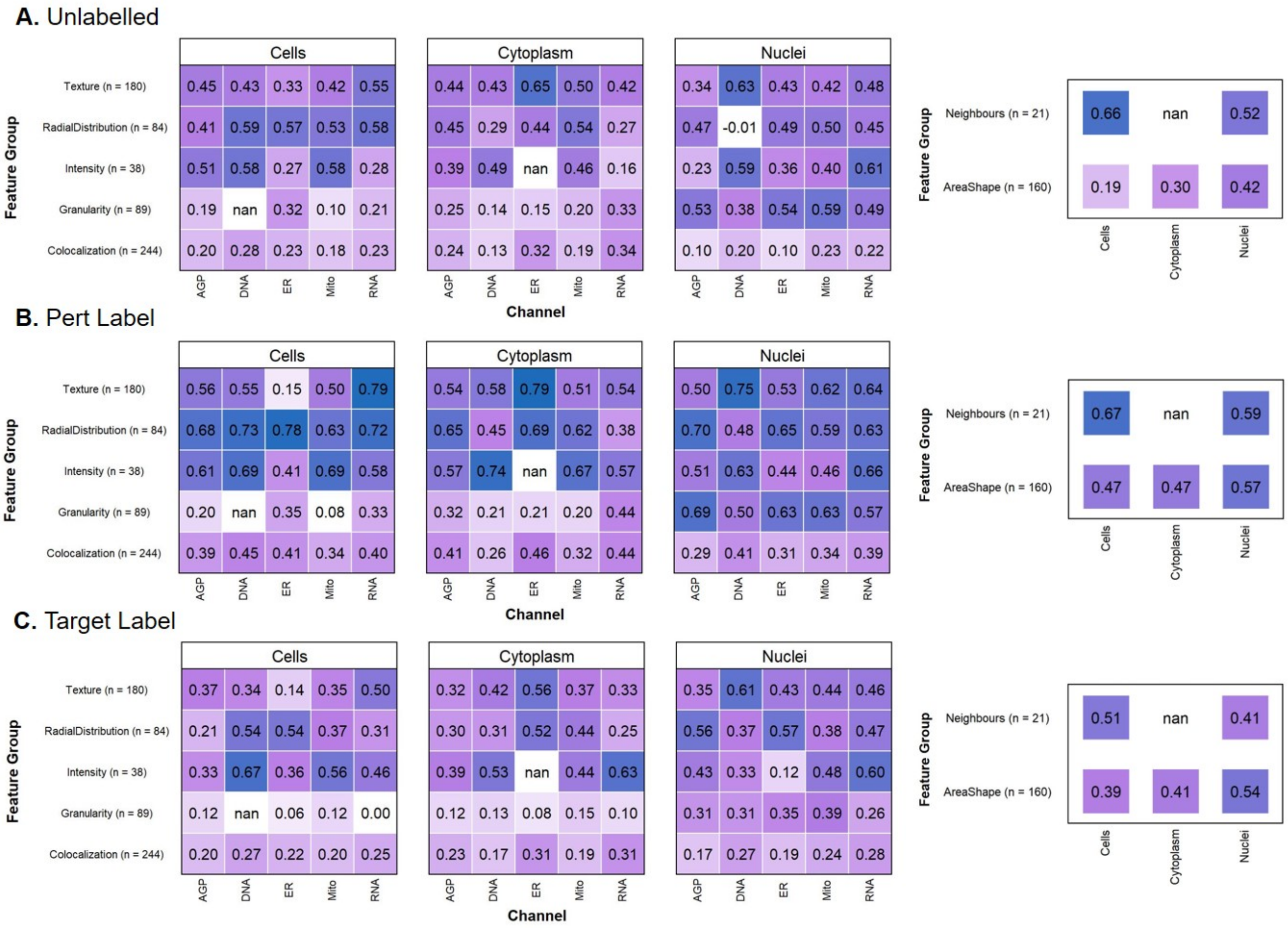


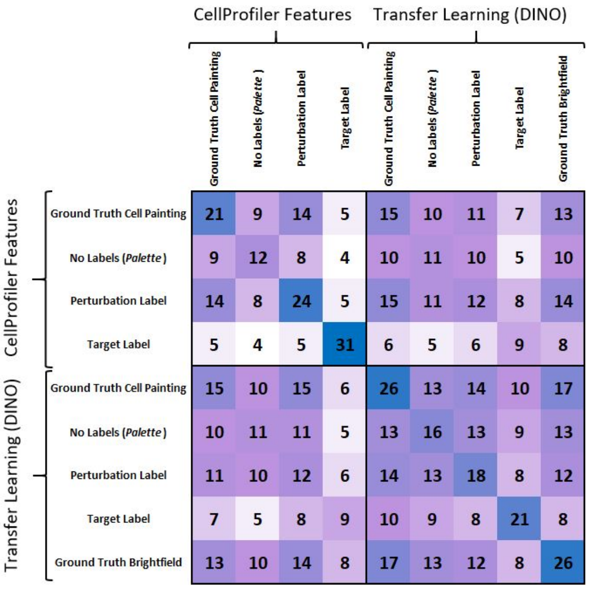
Figure 5: Heatmaps of the mean correlations to the ground truth of features by group/channel for **A.** the unlabelled model (*Palette*), **B.** perturbation as a label (AdaGN and CG) and **C.** target as a label (AdaGN and CG). These features are extracted from the active subset. The number of features for each feature group is also presented (total 635 features). The mean correlations for all the selected features to the ground truth features are **A.** 0.386, **B.** 0.504, **C.** 0.355.

guiding label, comparable performance and reproducibility of matching targets is seen, however most of the other models produce worse results. Transfer learning with DINO generally produced over 50% overlap in matching targets to the CellProfiler features, while also being capable of finding matching targets not in CellProfiler space. Perhaps a combined feature space (CellProfiler and transfer learning) could outperform the best individual models.

Although using the target as the label produced the largest number of matching targets (31), it shared very few matching targets to the other models including the ground truth. Even though it may be an advantage that this model can find different matching targets to the ground truth features, it should not come at the cost of failing to predict the simple-to-predict matching target pairs. Hence, we propose that using this label has not produced reproducible or correct features, rather the model has "brute-forced" similarities between images with the same labels, most likely by adding noise. We discuss this further in the Section B.3.

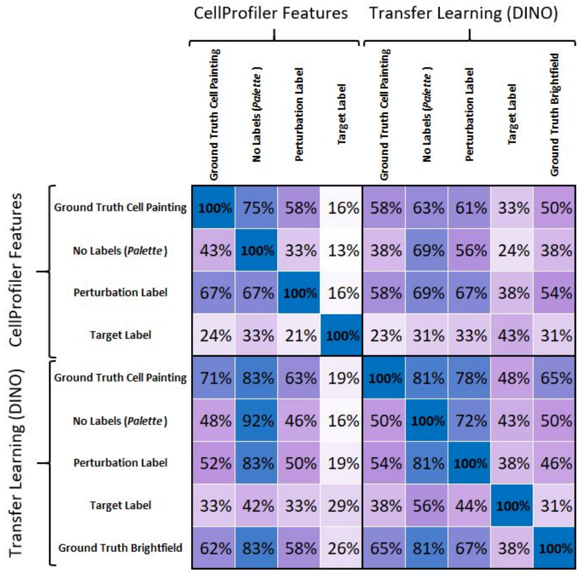
Self-attention maps

In Fig. 8 we present self-attention maps for each of the ground truth channels using pretrained DINO weights [8]. Self attention maps provide a visualisation of which 8×8 patches the vision transformer network places most emphasis on when calculating a feature representation of the image. While the Cell Painting channels' self-attention maps are slightly sharper in their segmentation properties, the brightfield channels show fairly reliable segmentation of cellular structure. It may be the case that modern computer vision architectures such as self-supervised, attention-based transformer networks have unlocked the brightfield as a valid modality for image-based profiling. This has not been



		CellProfiler Features				Transfer Learning (DINO)				
		Ground Truth Cell Painting	No Labels (*Palette*)	Perturbation Label	Target Label	Ground Truth Cell Painting	No Labels (*Palette*)	Perturbation Label	Target Label	Ground Truth Brightfield
CellProfiler Features	Ground Truth Cell Painting	21	9	14	5	15	10	11	7	13
	No Labels (*Palette*)	9	12	8	4	10	11	10	5	10
	Perturbation Label	14	8	24	5	15	11	12	8	14
	Target Label	5	4	5	31	6	5	6	9	8
Transfer Learning (DINO)	Ground Truth Cell Painting	15	10	15	6	26	13	14	10	17
	No Labels (*Palette*)	10	11	11	5	13	16	13	9	13
	Perturbation Label	11	10	12	6	14	13	18	8	12
	Target Label	7	5	8	9	10	9	8	21	8
	Ground Truth Brightfield	13	10	14	8	17	13	12	8	26

Figure 6: Matrix of the total number of shared matching targets (NN top 5) predicted between each of the models/ground truth modalities, in both CellProfiler Feature space and transfer learning (DINO) feature space. From the active subset, with labels included through AdaGN and CG.



		CellProfiler Features				Transfer Learning (DINO)				
		Ground Truth Cell Painting	No Labels (*Palette*)	Perturbation Label	Target Label	Ground Truth Cell Painting	No Labels (*Palette*)	Perturbation Label	Target Label	Ground Truth Brightfield
CellProfiler Features	Ground Truth Cell Painting	100%	75%	58%	16%	58%	63%	61%	33%	50%
	No Labels (*Palette*)	43%	100%	33%	13%	38%	69%	56%	24%	38%
	Perturbation Label	67%	67%	100%	16%	58%	69%	67%	38%	54%
	Target Label	24%	33%	21%	100%	23%	31%	33%	43%	31%
Transfer Learning (DINO)	Ground Truth Cell Painting	71%	83%	63%	19%	100%	81%	78%	48%	65%
	No Labels (*Palette*)	48%	92%	46%	16%	50%	100%	72%	43%	50%
	Perturbation Label	52%	83%	50%	19%	54%	81%	100%	38%	46%
	Target Label	33%	42%	33%	29%	38%	56%	44%	100%	31%
	Ground Truth Brightfield	62%	83%	58%	26%	65%	81%	67%	38%	100%

Figure 7: Matrix of the values from Fig. 6 expressed as a percentage of the total number of the value of the diagonal in the same column. i.e. the values in the first column are the % of targets predicted by each model as a percentage of the targets predicted in the CellProfiler feature space extracted from the ground truth Cell Painting images.

possible previously due to the lower resolution and higher noise of brightfield images. Our findings, alongside other recent studies [22] provide motivation for further work investigating image-based profiling with brightfield images.

Additionally, in Figs. 9 and 10 we compare the self-attention maps of the predicted Cell Painting channels to the ground truth channels. While self-attention maps with 8×8 patches do not reveal fine-scale structure, we can see that the larger scale structural properties of the channels are replicated relatively well in the predicted Cell Painting channels.

B.3. Background noise

We noticed that some of the predicted images were noisy across the whole image, which was particularly visible in the background (Fig. 11). The model would not add noise to all the predicted images in the test set, just a small number with certain class labels. This could be seen as a form of overfitting, where the model has learned to output irrelevant noise patterns which make images of the same class more similar. This is reflected in the metrics in Tables 1 and 2, where target matching improves despite reconstruction quality dropping in all metrics.

This was most common with target as the class label. One way to think about this is that if we were to construct a classifier to predict the target from the input images, this would be a very difficult task (in fact, this simple problem motivates much of the field of image-based profiling in drug discovery) compared to using the perturbation. Hence we emphasise the importance of a sensible choice of class labels. We should not introduce labels which are too ambitious for the network, and which may prevent learning a faithful reconstruction. Instead, we propose that the utility of class-guided image-to-image diffusion is through using simple labels to guide learning important structural fundamentals.

B.4. Experimental batch effects

The experimental batch effect is always a consideration in image-based profiling, and many studies have focused on tackling it [31, 45, 2]. We provide a few remarks relating to our study and the batch effect in this section.

If there is a batch effect in the ground truth, a faithful reconstruction would preserve it. Since we are producing potentially entire plates of images, all batch correction/normalisation methods which can be applied to real Cell Painting can also be applied to the predicted images (for example TVN [2] which uses variation in DMSO controls to correct for experimental batch variation). For this reason, we intentionally tested each replicate of our models on a single test batch. However, we present a model trained on 8 plates and simultaneously tested on 2 unseen plates to study the batch effect. We present a 2-D t-SNE plot of this

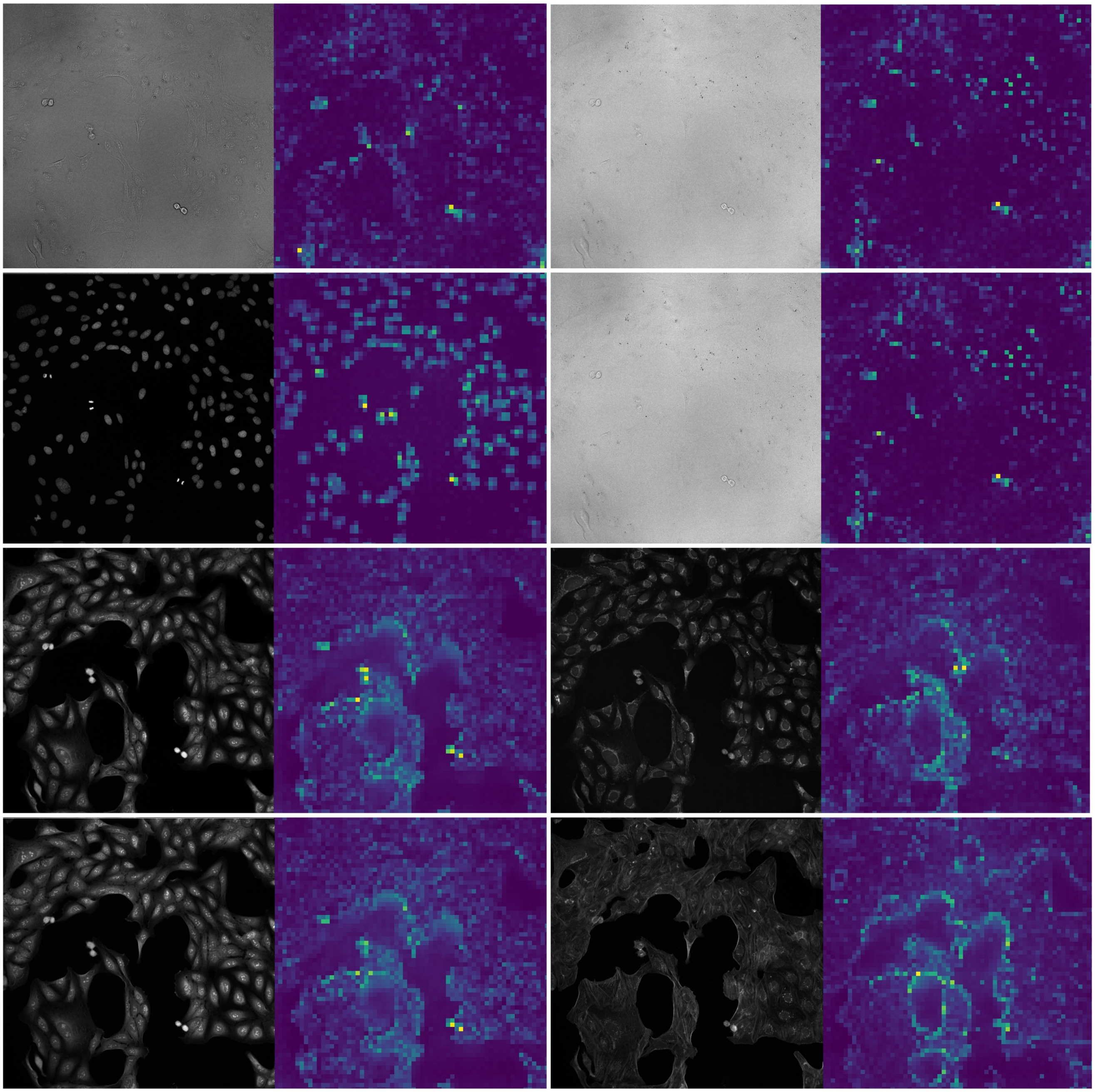


Figure 8: An example of paired self-attention maps for ground truth images with transfer learning (DINO) weights. Left (top to bottom): Brightfield 1, DNA, RNA, ER. Right (top to bottom): Brightfield 2, Brightfield 3, Mito, AGP.

investigation in Fig. 12. There was no notable batch effect between our ground truth plates, and we saw this replicated in our models (with and without labels). Promisingly, the CellProfiler feature spaces extracted from model predicted images overlapped well with the ground truth feature space. This is an improvement upon prior studies, which induced a phantom "batch effect" between the predicted and real feature spaces [16]. This is important to assess, as correlation (Fig. 5) does not account for feature space overlap.