OUT-OF-DISTRIBUTION EVALUATIONS OF CHANNEL AGNOSTIC MASKED AUTOENCODERS IN FLUORESCENCE MICROSCOPY

Christian John Hurry[*], **Jinjie Zhang, Olubukola Ishola, Emma Slade, Cuong Q. Nguyen**
GSK AIML
79 New Oxford Street, London, United Kingdom, WC1A 1DG

ABSTRACT

Developing computer vision for high-content screening is challenging due to various sources of distribution-shift caused by changes in experimental conditions, perturbagens, and fluorescent markers. The impact of different sources of distribution-shift are confounded in typical evaluations of models based on transfer learning, which limits interpretations of how changes to model design and training affect generalisation. We propose an evaluation scheme that isolates sources of distribution-shift using the JUMP-CP dataset, allowing researchers to evaluate generalisation with respect to specific sources of distribution-shift. We then present a channel-agnostic masked autoencoder **Campfire** which, via a shared decoder for all channels, scales effectively to datasets containing many different fluorescent markers, and show that it generalises to out-of-distribution experimental batches, perturbagens, and fluorescent markers, and also demonstrates successful transfer learning from one cell type to another.

1 INTRODUCTION

Phenotypic drug discovery, in which cells or animal models are subject to a perturbation and monitored for a desired change in phenotype, has seen a resurgence due to its success in finding compounds that meet regulatory approval (Zheng et al., 2013; Boutros et al., 2015; Zanella et al., 2010). To quantify the effect of perturbations, it is common to use high content screening (HCS), a method in which batches of cells are stimulated with thousands of compounds in parallel, and multiple markers of changes in phenotype are measured simultaneously. In comparison with modalities based on sequencing technologies, imaging is more time- and cost-effective at scale and has been the main modality of HCS data. This necessitated the development of automated pipelines that extract biologically relevant features from cellular imaging data. Typically, this has involved traditional methods based on cell-segmentation and feature extraction and has been applied in various applications including protein sub-cellular localisation (Pärnamaa & Parts, 2017), quantitative structure-activity relationship modelling (Nguyen et al., 2023) and identifying mechanism of action (Dürr & Sick, 2016; Wong et al., 2023) and markers of drug resistance (Kelley et al., 2023).

In recent years, the volume of high content imaging (HCI) data has increased in scale due to the introduction of the Cell Painting assay (Bray et al., 2016; Seal et al., 2024). This assay provides standardisation for HCI by using a specific set of dyes, optimised to highlight and contrast several cellular compartments of interest (Cimini et al., 2023). The public JUMP-CP dataset was generated using the Cell Painting assay, and contains images of millions of cells, subject to 116,000 compound and 22,000 genetic perturbations, derived from multiple laboratories and institutions (Chandrasekaran et al., 2023). Examples of the application of the JUMP-CP dataset to-date include predicting the mechanism of action of kinase inhibitors (Dee et al., 2024), quantifying differences in data derived from different experimental batches (Arevalo et al., 2024), and predicting the molecular structure of perturbagens from cell imaging (Watkinson et al., 2024).

Recently, large-scale public datasets have been leveraged to produce SOTA *foundation* models in various biological domains, including single cell sequencing data (Yang et al., 2022; Cui et al.,

[*]Corresponding author, christian.j.hurry@gsk.com

2024; Heimberg et al., 2024), protein and complex structure prediction (Abramson et al., 2024), and pathology (Vorontsov et al., 2023; Juyal et al., 2024; Dippel et al., 2024). Foundation models differ from typical deep neural networks as they learn representations which are useful across a range of biologically relevant tasks. To achieve this, they are typically trained in self-supervised fashion, such that representations are learned in a manner which is less biased to a particular task in comparison with fully supervised methods. Recent work has applied self-supervised learning to HCI (Borowa et al., 2024; Bourriez et al., 2024; Bao & Karaletsos, 2023; Kraus et al., 2024) and has shown that models scale with dataset and model size suggesting a route towards a foundation model for HCI (Kraus et al., 2024).

It remains an open challenge to develop a foundation model for HCI, in part due to the variety of sources of distribution-shift that are common to fluorescence microscopy. This includes changes in experimental conditions, instruments of measurement, and the cell types and perturbations under consideration. Another source of distribution-shift that is a particular challenge with HCI is changes to the set of fluorescent markers used to generate images. These differ between experimental screens to highlight the most relevant cellular compartments. The evaluation of models for HCI with respect to different sources of distribution-shift, varies between previous works. In some cases, models are shown to generalise to different microscopy screens, which can include changes in cell types, fluorescent markers and perturbations in combination and is, therefore, the most challenging form of generalisation. This, however, confounds all sources of distribution-shift, making it difficult to assess how changes in model architecture or training protocol affect generalisation to distribution-shift of a specific type. Therefore, to assess and benchmark different models for HCI, it is important to show *explicit* generalisation to *i*) new experimental plates, *ii*) changes in perturbagens, *iii*) changes in fluorescent markers, and *iv*) new microscopy screens. By isolating different sources of distribution-shift, we can provide a more comprehensive analysis of the universality of any given model for HCI.

To address these generalisation requirements we introduce **Campfire**[1] (**C**hannel **a**gnostic **m**orphological **p**rofiling **f**rom **i**maging under a **r**ange of **e**xperimental settings). This is a channel-agnostic masked autoencoder that uses a shared decoder for all channels, such that it scales more effectively to datasets containing a wide variety of fluorescent markers, and that can handle out-of-distribution fluorescent markers during inference. In this work, we propose an evaluation scheme based on the JUMP-CP dataset, that isolates sources of distribution-shift common to fluorescence microscopy, and demonstrate that **Campfire** learns representations of data that generalise across plates, perturbagens, fluorescent markers, and cell type.

1.1 Related work

To address the challenges of HCI, several works have recently adapted vision transformers (ViT) which can accept input sequences of varying length during training and inference (Dosovitskiy, 2020). Each channel in HCI provides information from a distinct cellular compartment. To utilise this ChannelViT (Bao et al., 2023) and ChAda-ViT (Bourriez et al., 2024) proposed encoding each channel of an image into a set of distinct patch embeddings by applying a shared projection layer to each channel. These patch embeddings were encoded with the channel of origin via a set of learnable embeddings, akin to position encoding. Notably, self-supervised ChAda-ViT demonstrated successful transfer learning from one HCI dataset to another. CA-MAE (Kraus et al., 2024) was also trained in self-supervised fashion using HCI but was trained to reconstruct images of wells from masked input with arbitrary combinations of channels. This work demonstrated that masked autoencoders (MAEs) (He et al., 2022) scale, such that increasing the amount of available HCI and compute lead to better image representations. While not channel-agnostic, ContextViT (Bao & Karaletsos, 2023) showed improved performance on out-of-distribution experimental plates, by appending an additional *context* embedding to the sequence of patch embeddings. Context embeddings were formed by projecting the average patch embedding through a linear layer. To the best of our knowledge, a model has not been explicitly designed to generalise to OOD fluorescent markers.

[1] Code and model checkpoint can be found at https://github.com/GSK-AI/campfire.

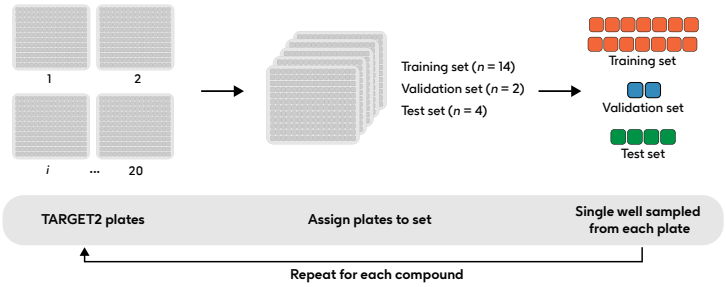

Figure 2.1: A sketch of how wells are assigned to training/validation/test sets for a given compound in TARGET2 plates. For each compound, we select 14 plates from which to sample a single well treated with that compound for the training set, 2 plates to sample a well for the validation set, and 4 plates for the test set. The exception are 60 randomly selected compounds which are all held-out of training. We also hold 5 TARGET2 plates out of training, with all wells added to the test set.

2 METHODOLOGY

2.1 JUMP-CP DATASET

In this work we used Source 3 of the JUMP-CP dataset Chandrasekaran et al. (2023), consisting of images of cells perturbed by small molecules. We considered all 25 TARGET2 and 237 COMPOUND plates, comprised of 384 wells each. To avoid data leakage, we assigned images to training, validation or test sets according to the well from which they were derived. This assignment differed between the TARGET2 and COMPOUND plates.

Each TARGET2 plate contained wells stimulated by 302 compounds, including 8 positive controls and 1 negative control. We randomly selected 60 compounds to hold out of training, excluding the controls. We also held 5 TARGET2 plates out of training, such that we had 20 in-distribution (ID) and 5 out-of-distribution (OOD) plates. As sketched in Fig. 2.1 for each compound we selected 14 plates for training, 2 plates for validation, and 4 plates for testing, sampling a single well from each of these plates for that compound. Consequently, our training/validation/test set was represented by 14/2/4 wells for each compound, where the plates from which these wells were derived may have differed for each compound. Wells within the test set were either *i*) from an ID plate, with ID compound *ii*) from an OOD plate, with ID compound *iii*) from an ID plate, with OOD compound *iv*) from an OOD plate, with OOD compound. Hence, at the cost of less images being included in our training set, this data splitting procedure allowed allowed us to evaluate our model against different sources of distribution-shift, by evaluating performance on different subsets of our test set, isolating the effect of distribution-shift caused by plate of origin or whether or not the compound was seen during training.

We included additional wells in the training set from COMPOUND plates, which in total represented cells stimulated by 58456 different compounds. For these plates, we assigned each well to the training set with probability $p_t = 0.5$ to limit the number of single cell images in our dataset. To produce a consistent evaluation scheme, we did not include COMPOUND plates in our test or validation sets.

2.2 CAMPFIRE: A CHANNEL AGNOSTIC MASKED AUTOENCODER WITH CHANNEL EMBEDDINGS AND CHANNEL SUBSAMPLING

Campfire, sketched in Fig. 2.2, is a channel-agnostic MAE. Well images were preprocessed into cell-centred tiles of size (112,112) for model input. These were further processed by a *3D convolutional layer*, with kernel and stride of shape $(1, P, P)$, where P was the size of a patch, to form $N \times C$ embeddings of dimension D, each representing N non-overlapping (P, P) patches for each channel of a cell-centred tile with C channels. Applying the same convolution to each channel,

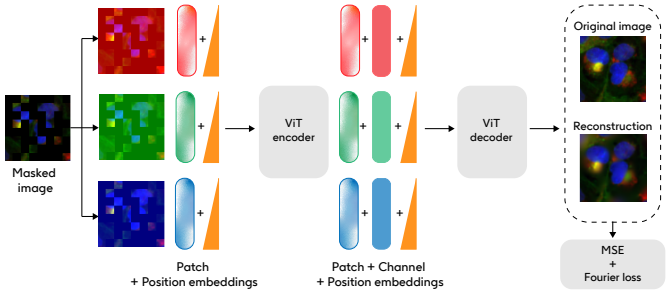

Figure 2.2: A sketch of the **Campfire** architecture.

mapped image patches to embeddings in a channel agnostic manner, such that the model could be trained with images of inconsistent channel ordering, number, or type. Patch embeddings were augmented by adding sinusoidal and RoPE position embeddings (Su et al., 2024; Heo et al., 2025).

The sequence of position encoded patch embeddings were then fed through an asymmetric MAE. A random subset of patch positions indexed $1, \ldots, N$ were selected to be masked, with p_m denoting the fraction of masked patches. For all channels, patch embeddings corresponding to the masked positions were removed from input to the encoder. The latent state of the sequence of embeddings was then projected through a linear layer to the smaller embedding dimension of the decoder. This sequence was padded with mask tokens, each representing a patch embedding removed during masking. Each patch embedding was augmented with sinusoidal and RoPE position embeddings, and an additional channel embedding, to encode the channel of origin in the patch. To compute the channel embedding, we averaged all patch embeddings in the batch from the same channel, and projected this through a linear layer shared by all channels. This linear layer, therefore, was trained to produce channel embeddings, given the average patch embedding from that channel.

The latent state of the input to the decoder was passed to an objective function, detailed in App. A, to optimise the reconstruction of the original input images from the masked input. Lastly, at inference time, the decoder of our model was discarded, the full unmasked set of patch embeddings were fed-forward through the encoder, and the average of the latent state of the patch embeddings was used as our final representation of a cell-centred tile.

3 Results

3.1 Model training and reconstruction of cell-centred tiles

Campfire was trained on 16 A100 (80GB) GPUs for 1920 GPU hours, with a per-GPU batch size of 400, for 50 epochs. The patch size was set to $(14, 14)$ and the fraction of patches that were masked from training was set to $p_m = 0.8$. During training we used random flipping, rotation and channel normalisation as additional augmentations. We used the AdamW optimiser with learning rate $lr = 5 \times 10^{-4}$ with weight decay $wd = 0.005$, and applied cosine annealing with 20 epochs of linear warm up with a starting learning rate $lr = 1 \times 10^{-5}$ and minimum learning rate of $\eta = 1 \times 10^{-6}$. These hyperparameters were chosen as the result of an ablation study, detailed in App. B.

The images of the JUMP-CP dataset consisted of 5 fluorescent channels which highlighted the Nucleus (Nu), Actin+Golgi Apparatus+Plasma Membrane (Ac), Mitochondria (M), Endoplasmic Reticulum (ER), and the Nucleolus+cytoplasmic RNA (cyRNA). We trained our model with the Nu channel and two additional channels, Ac and M, selected at random. The ER and cyRNA channels were held-out to evaluate OOD performance. For each batch in training, we sampled a subset of channels, S, from the set of all available channels, with uniform probability, $S \sim \mathcal{U}(\{X \mid X \subseteq \{\text{Nu}, \text{Ac}, \text{M}\}\})$, and only these channels of the images in the batch were passed to the model. Consequently, our model was trained with *multiple views* of the images to be robust to different combinations of ID channels. After training, we found that our model was capable of reconstructing cell-centred tiles from wells not seen during training, as exemplified in Fig. 3.1.

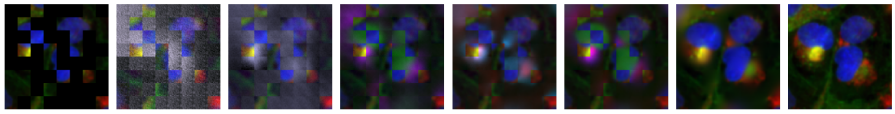

Figure 3.1: An example of the reconstruction of a cell-centred tile. From left to right, we show the masked input, followed by six reconstructed images at different epochs, with epoch increasing from left to right, and lastly the original cell-centred tile to be reconstructed.

3.2 Self-supervised objective function and channel subsampling leads to biologically informed feature extraction

To assess whether **Campfire** distilled useful knowledge from the JUMP-CP dataset, we evaluated our model on several biologically meaningful tasks. In order of increasing complexity, the tasks were to predict the 1-of-9 control compounds of stimulation, using single cell embeddings derived from *i*) plates included in training, or *ii*) plates held-out of training, or to predict the 1-of-60 held-out compounds of stimulation using single cell embeddings derived from *iii*) plates included in training, or *iv*) plates held-out of training.

After training **Campfire** we sampled, from all 25 TARGET2 plates, 100 single cell embeddings from each well stimulated by a control compound, and 30 from each well stimulated by a held-out compound. This kept our sample balanced in the representation of wells (due to the low number of cells in some held-out wells).

For the prediction of control compounds, we collected all embeddings derived from training wells stimulated by control compounds, and divided these into 10 equal subsets. For each subset of embeddings, we trained a linear classifier to predict the 1-of-9 control compounds, using a cross-entropy loss function. Each classifier was trained for 100 epochs, using embeddings from validation wells to perform early stopping with respect to the accuracy. After training, we computed the accuracy of the classifier using embeddings from test wells for the ID and OOD plates, separately.

For prediction of the held-out compounds, we assigned embeddings derived from held-out compound wells in the ID plates to 5 train/test splits via 5-fold cross-validation. Each fold was used to train a linear classifier for 100 epochs, using early stopping with respect to the accuracy. For each fold, we computed the accuracy with test embeddings from the ID and OOD plates, separately.

We summarise the performance of our model in Tab. 3.1. For comparison, we extracted single cell embeddings from the 25 TARGET2 plates using two ImageNet1k baselines and repeated the steps above. We found that across all four tasks, our model outperformed both baseline models, suggesting that our self-supervised objective led to the distillation of useful biological knowledge that can transfer to new tasks.

3.3 Integration of information from multiple fluorescent channels

An ideal model for HCI should integrate information from multiple fluorescent channels, such that model embeddings improve with the number of fluorescent channels used to form them. To assess channel integration with **Campfire**, we first extracted single cell embeddings from the 25 TARGET2 plates multiple times, using the following combinations of fluorescent channels, *i*) Nu channel only

Table 3.1: Comparison of three models on four different tasks: predicting compound of stimulation, either ID or OOD, from single cell image embeddings from either ID or OOD plates. **DinoViT-S8** and **DinoViT-L14** are pretrained on ImageNet1k only, **Campfire** was pretrained on JUMP-CP.

Model	ID compound + ID plate	ID compound + OOD plate	OOD compound + ID plate	OOD compound + OOD plate
DinoViT-S8 (Caron et al., 2021)	0.418 ± 0.006	0.322 ± 0.007	0.195 ± 0.006	0.187 ± 0.003
DinoViT-L14 (Oquab et al., 2023)	0.356 ± 0.012	0.281 ± 0.009	0.158 ± 0.009	0.152 ± 0.002
Campfire	**0.460 ± 0.016**	**0.375 ± 0.007**	**0.229 ± 0.008**	**0.220 ± 0.002**

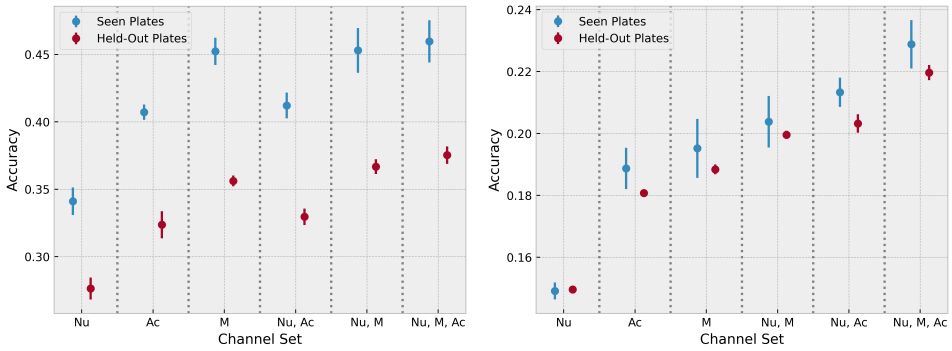

Figure 3.2: Accuracy of a linear classifier predicting 1-of-9 control compounds (left panel) or 1-of-60 held-out compounds (right) from single cell embeddings. Each column represents a linear classifier trained on single cell embeddings derived from images comprised of different sets of fluorescent channels. Channels shown: Nucleus (N), Actin+Golgi apparatus+Plasma Membrane (Ac), Mitochondria (M) and their combinations.

ii) M channel only *iii*) Ac channel only *iv*) Nu and Ac channels *v*) Nu and M channels, and *vi*) Nu, Ac and M channels. For each combination of channels, we extracted single cell embeddings for all cell-centred tiles, hence, extracting embeddings corresponding to *identical* sets of images for each channel combination. We then followed the steps outlined in Sec. 3.2 to sample single cell embeddings and train linear classifiers to predict the compound of stimulation.

In the left panel of Fig. 3.2, we show the accuracy of classifiers that predict control compounds. From the accuracy of classifiers trained with a single channel, we found significant difference in the quality of information that each channel provided. The M channel was found to be the most informative channel, alone. It was always found beneficial to include additional fluorescent channels; the Nu channel and Ac channel in combination provided higher accuracy than the Ac channel alone, and despite the Nu channel being the least informative channel, the Nu and M channels in combination provided higher accuracy than the M channel alone.

We compared this with the accuracy of classifiers trained to predict the held-out compounds in the right panel of Fig. 3.2. Comparing the rank of each combination of channels, in terms of their predictive accuracy, we found that the rank was not consistent between the tasks of predicting control and held-out compounds. Although the rank of single channel accuracy was consistent, we found that the combination of the Nu and Ac channels outperformed the combination of the Nu and M channels, in the prediction of the held-out compounds, despite M being the single most informative channel. This suggests that the interaction between fluorescent channels impacted the quality of image embeddings, and that embedding quality is not maximised by merely selecting the single most informative channels. From this we concluded that our model successfully integrates information across fluorescent channels.

3.4 Generalisation to held-out fluorescent channels

To evaluate generalisation to OOD fluorescent channels, we extracted single cell embeddings from the 25 TARGET2 plates using the Nu, ER and cyRNA channels (although the Nu channel was ID, we included it in the OOD channel set as it is unlikely for a nucleus channel to be excluded in a screen). We sampled single cell embeddings and trained linear classifiers according to the steps in Sec. 3.2.

We compared the performance of **Campfire** to our strongest baseline **DinoViT-S8** (for which all channels are OOD, as it is pretrained on ImageNet1k only). In Tab. 3.2, the accuracy of **DinoViT-S8** was found to increase with OOD channels across all 4 tasks. Having seen that single channels provide different amounts of information in Sec. 3.3, this implied that an informative channel was held-out during training. Conversely, when comparing ID to OOD channels, **Campfire** showed lower accuracy on all 4 tasks, as is expected due to distribution-shift. Despite this, **Campfire** out-

Table 3.2: Comparison of 2 models on 4 different biological tasks, predicting compound of stimulation, either ID or OOD, from single cell image embeddings from either ID or OOD plates. Results are shown when inference is performed with either ID or OOD sets of channels. **DinoViT-S8** was pretrained on ImageNet1k, **Campfire** was pretrained on JUMP-CP data, with the ID channel set. Metric is the accuracy shown over 10 training splits (ID compound) or 5 folds (OOD compound), with ± indicating standard deviation.

Model / Channel Set	ID compound + ID plate	ID compound + OOD plate	OOD compound + ID plate	OOD compound + OOD plate
Campfire / ID	**0.460 ± 0.016**	**0.375 ± 0.007**	**0.229 ± 0.008**	**0.220 ± 0.002**
DinoViT-S8 / ID	0.418 ± 0.006	0.322 ± 0.007	0.195 ± 0.006	0.187 ± 0.003
Campfire / OOD	0.44 ± 0.01	0.353 ± 0.07	**0.206 ± 0.009**	**0.199 ± 0.003**
DinoViT-S8 / OOD	**0.457 ± 0.009**	**0.356 ± 0.009**	0.204 ± 0.003	0.191 ± 0.001

performed **DinoViT-S8** on the task of highest difficulty, predicting the held-out compound from OOD plates, suggesting that **Campfire** distilled useful biological knowledge that transfers to OOD fluorescent channels.

3.5 Transfer learning from one microscopy screen to another

To assess the extent that **Campfire** could transfer knowledge from the JUMP-CP dataset to a different microscopy screen, we finetuned **Campfire** to detect changes in macrophage morphology. We considered a dataset comprised of 4, 384-well, plates containing macrophages subject to different gene knock-out (Mehrizi et al., 2023). Two plates contained macrophages with M1 polarisation, while the other 2 contained macrophages with M2 polarisation. We froze the parameters of **Campfire** and attached a 2-layer MLP (with hidden dimensions 1024 and 128) to its last layer. The resultant model was then trained using a triplet loss objective function (Schroff et al., 2015) using one M1 plate and one M2 plate, leaving the other two plates held-out of training for later evaluation. For each mini-batch we sampled several cell-centred tiles from the same set of wells, and computed the mean model embedding for each well in the batch. These *well-level* embeddings were then passed to the triplet loss objective function, where we treated well-level embeddings derived from wells subject to the same/different gene knock-out as positive/negative samples. After 500 epochs of training, we ran inference over all four plates, and computed the well-level embedding for each well. We repeated this process for the **DinoViT-S8** model.

Our dataset contained wells subject to positive and negative control stimulations that had, respectively, a known impact or lack of impact on macrophage morphology. We computed the Z'-score (Zhang et al., 1999) for each pairwise combination of positive and negative controls (calculation detailed in App. C) . This score indicated the size of the statistical difference between the groups of embeddings subject to two different stimulations.

As shown in Fig. 3.3, we found that **Campfire** was demonstrably better than **DinoViT-S8** at distinguishing between M1 macrophages that had been subject to positive or negative controls. To a greater extent, this was also true for M2 macrophages, for which **DinoViT-S8** failed to discern between stimulations. Despite this difference, the models were comparable when discerning between macrophages of different polarisation, regardless of stimulation. That **DinoViT-S8** could discern between macrophages of different polarisation, which were derived from different plates, but not between positive and negative controls, suggested it was prone to plate effects. From this we concluded that pretraining on HCI allowed **Campfire** to mitigate batch effects without additional training constraints during finetuning.

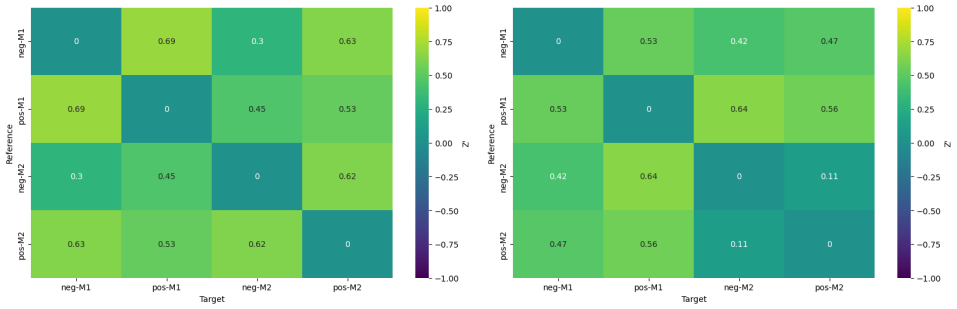

Figure 3.3: Z'-score measuring statistical difference between model embeddings from a reference and target compound of stimulation. Model embeddings are derived from **Campfire** (left) pretrained on JUMP-CP, and **DinoViT-S8** (right) pretrained on ImageNet1k. Both models have been finetuned on a macrophage dataset. Z'-score is shown for negative (neg) and positive (pos) controls for plates with macrophages in either M1 or M2 polarisation.

4 Conclusion

Developing a channel-agnostic model that can generalise to new experimental plates, compounds of perturbation, OOD fluorescent channels and cell types is essential if a foundation model for HCI is to be realised. The development of a foundation model for HCI not only requires specialised model design and dataset curation, but also proper evaluation, to ascertain how choices in architecture, dataset, and training protocol affect the generalisation of models for HCI. In this work we proposed a method of splitting the JUMP-CP dataset that allowed us to scrutinise how model design and training protocol impacted the ability of a model to generalise under a specific source of distribution-shift.

We developed a channel-agnostic model for HCI that, when trained in self-supervised fashion, could predict the compound of stimulation from images at single cell resolution. Our model, **Campfire**, generalised to OOD experimental batches, compounds, and sets of fluorescent channels. Of these sources of distribution-shift, generalising to a new set of fluorescent channels was found to be the most challenging. All images in the JUMP-CP dataset consisted of 5 fluorescent channels which limited the number of channels that could be included in training, in order to form an appropriate OOD evaluation set. Training over a broader set of fluorescent channels would likely have improved the generalisation of our model under this source of distribution-shift. While we trained **Campfire** with the JUMP-CP dataset exclusively, training with screens of different cell types will be crucial for a HCI foundation model. This may also improve generalisation to OOD fluorescent channels, as ID fluorescent markers will capture broader variation due to differences in the morphology of different cell types. Despite the challenges of generalising to OOD channels, **Campfire** successfully demonstrated transfer learning to a microscopy screen containing new cell types. Our evaluation scheme allows us to compare model performance with respect to different distribution shifts, and compare this with their performance when transfer learning to a new microscopy screen, where these models would be used in practice, hopefully increasing the rate at which a foundation model for HCI can be realised.

While our work focussed on predicting the compound of perturbation, our proposed method to split the wells of the JUMP-CP dataset into training, validation and test sets may be used to evaluate model performance on different biologically relevant tasks, while still isolating the different sources of distribution-shift we considered. Future work towards the development of a universal model for HCI should focus on leveraging data from multiple sources and increasing the number of fluorescent channels and cell types included in training, in order to produce high quality embeddings under realistic sources of variation in images produced by HCS.

Meaningfulness Statement

Life as we know it is a multi-scale phenomena, from the structure of DNA, to cell state, tissue function, and organ interactions. A meaningful representation of life should capture how changes at one scale affect another. Our work focusses on how perturbations at the sub-cellular level affect

cell phenotype. By evaluating channel agnostic models for high content screening with respect to different distribution shifts, we hope to aid the development of a foundation model for high content imaging which provides robust representations of cells subject to perturbation that capture fine-grained information of cell phenotype.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

John Arevalo, Ellen Su, Jessica D Ewald, Robert van Dijk, Anne E Carpenter, and Shantanu Singh. Evaluating batch correction methods for image-based cell profiling. *Nature communications*, 15 (1):6516, 2024.

Yujia Bao and Theofanis Karaletsos. Contextual vision transformers for robust representation learning. *arXiv preprint arXiv:2305.19402*, 2023.

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth c x 16 x 16 words. *arXiv preprint arXiv:2309.16108*, 2023.

Adriana Borowa, Dawid Rymarczyk, Marek Żyła, Maciej Kańduła, Ana Sánchez-Fernández, Krzysztof Rataj, Łukasz Struski, Jacek Tabor, and Bartosz Zieliński. Decoding phenotypic screening: A comparative analysis of image representations. *Computational and Structural Biotechnology Journal*, 23:1181–1188, 2024.

Nicolas Bourriez, Ihab Bendidi, Ethan Cohen, Gabriel Watkinson, Maxime Sanchez, Guillaume Bollot, and Auguste Genovesio. Chada-vit: Channel adaptive attention for joint representation learning of heterogeneous microscopy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11556–11565, 2024.

Michael Boutros, Florian Heigwer, and Christina Laufer. Microscopy-based high-content screening. *Cell*, 163(6):1314–1325, 2015.

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.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D Boyd, Laurent Brino, et al. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *BioRxiv*, pp. 2023–03, 2023.

Beth A Cimini, Srinivas Niranj Chandrasekaran, Maria Kost-Alimova, Lisa Miller, Amy Goodale, Briana Fritchman, Patrick Byrne, Sakshi Garg, Nasim Jamali, David J Logan, et al. Optimizing the cell painting assay for image-based profiling. *Nature protocols*, 18(7):1981–2013, 2023.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pp. 1–11, 2024.

William Dee, Ines Sequeira, Anna Lobley, and Gregory Slabaugh. Cell-vision fusion: A swin transformer-based approach for predicting kinase inhibitor mechanism of action from cell painting data. *Iscience*, 27(8), 2024.

Jonas Dippel, Barbara Feulner, Tobias Winterhoff, Timo Milbich, Stephan Tietz, Simon Schallenberg, Gabriel Dernbach, Andreas Kunft, Simon Heinke, Marie-Lisa Eich, et al. Rudolfv: a foundation model by pathologists for pathologists. *arXiv preprint arXiv:2401.04079*, 2024.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Oliver Dürr and Beate Sick. Single-cell phenotype classification using deep convolutional neural networks. *Journal of biomolecular screening*, 21(9):998–1003, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Graham Heimberg, Tony Kuo, Daryle J DePianto, Omar Salem, Tobias Heigl, Nathaniel Diamant, Gabriele Scalia, Tommaso Biancalani, Shannon J Turley, Jason R Rock, et al. A cell atlas foundation model for scalable search of similar human cells. *Nature*, pp. 1–3, 2024.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2025.

Dinkar Juyal, Harshith Padigela, Chintan Shah, Daniel Shenker, Natalia Harguindeguy, Yi Liu, Blake Martin, Yibo Zhang, Michael Nercessian, Miles Markey, Isaac Finberg, Kelsey Luu, Daniel Borders, Syed Ashar Javed, Emma L Krause, Raymond Biju, Aashish Sood, Allen Ma, Jackson Nyman, John Shamshoian, Guillaume Chhor, Darpan Sanghavi, Marc Thibault, Limin Yu, Fedaa Najdawi, Jennifer A. Hipp, Darren Fahy, Benjamin Glass, Eric Walk, John Abel, Harsha Vardhan pokkalla, Andrew H. Beck, and Sean Grullon. PLUTO: Pathology-universal transformer. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*, 2024. URL https://openreview.net/forum?id=7yn50e6tVX.

Megan E Kelley, Adi Y Berman, David R Stirling, Beth A Cimini, Yu Han, Shantanu Singh, Anne E Carpenter, Tarun M Kapoor, and Gregory P Way. High-content microscopy reveals a morphological signature of bortezomib resistance. *Elife*, 12:e91362, 2023.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11757–11768, 2024.

Rahil Mehrizi, Arash Mehrjou, Maryana Alegro, Yi Zhao, Benedetta Carbone, Carl Fishwick, Johanna Vappiani, Jing Bi, Siobhan Sanford, Hakan Keles, et al. Multi-omics prediction from high-content cellular imaging with deep learning. *arXiv preprint arXiv:2306.09391*, 2023.

Cuong Q Nguyen, Dante Pertusi, and Kim M Branson. Molecule-morphology contrastive pretraining for transferable molecular representation. *arXiv preprint arXiv:2305.09790*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Tanel Pärnamaa and Leopold Parts. Accurate classification of protein subcellular localization from high-throughput microscopy images using deep learning. *G3: Genes, Genomes, Genetics*, 7(5): 1385–1392, 2017.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Srijit Seal, Maria-Anna Trapotsi, Ola Spjuth, Shantanu Singh, Jordi Carreras-Puigvert, Nigel Greene, Andreas Bender, and Anne E Carpenter. Cell painting: a decade of discovery and innovation in cellular imaging. *Nature methods*, pp. 1–15, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Siqi Liu, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023.

Gabriel Watkinson, Ethan Cohen, Nicolas Bourriez, Ihab Bendidi, Guillaume Bollot, and Auguste Genovesio. Weakly supervised cross-modal learning in high-content screening. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Daniel R Wong, David J Logan, Santosh Hariharan, Robert Stanton, Djork-Arné Clevert, and Andrew Kiruluta. Deep representation learning determines drug mechanism of action from cell painting images. *Digital Discovery*, 2(5):1354–1367, 2023.

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4(10):852–866, 2022.

Fabian Zanella, James B Lorens, and Wolfgang Link. High content screening: seeing is believing. *Trends in biotechnology*, 28(5):237–245, 2010.

Ji-Hu Zhang, Thomas DY Chung, and Kevin R Oldenburg. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *Journal of biomolecular screening*, 4(2):67–73, 1999.

Wei Zheng, Natasha Thorne, and John C McKew. Phenotypic screens as a renewed approach for drug discovery. *Drug discovery today*, 18(21-22):1067–1073, 2013.

A MODEL OBJECTIVE FUNCTION

During pretraining **Campfire** was trained to optimise the following objective function,

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\left[\lambda_s MSE(\hat{y}_i, y_i) + \lambda_\ell MSE(\mathcal{F}[\hat{y}_i|\ell], \mathcal{F}[y_i|\ell]) + \lambda_h MSE(\mathcal{F}[\hat{y}_i|h], \mathcal{F}[y_i|h])\right], \quad (1)$$

where y_i and $\hat{y}_i$ are the unmasked image, and image reconstruction, respectively, and $MSE(x_1, x_2) = \frac{1}{M}\sum_{j=1}^{M}(x_{1,j} - x_{2,j})^2$ for $x_1, x_2 \in \mathcal{R}^M$ is the L2 reconstruction loss and $|\mathcal{D}|$ is the number of samples in the dataset. The hyperparameters $\lambda_{s,l,h} \in [0, 1]$ control the weight given to each term in the loss function. $F[x|h/\ell]$ represents a high/low pass filter function. The high pass filter was performed by applying a 2D fast Fourier transform (FFT), setting the outer-most Fourier coefficients to zero, and applying the inverse 2D FFT. Similarly, the low-pass filter involved applying the 2D FFT, removing the inner-most Fourier coefficients and performing the inverse 2D FFT. Here $h, \ell \in [0, 1]$ represent the fraction of Fourier coefficients removed in the high and low pass filters, respectively. Consequently, our model was trained to reconstruct uncorrupted images, and also optimise the reconstruction of images subject to high and low pass filters, as in recent works (Juyal et al., 2024; Kraus et al., 2024).

B ABLATION STUDY

In order to determine the optimum hyperparameters of our model we performed a three-stage ablation study which sought to optimise hyperparameters for: *i*) the objective function, *ii*) neural network architecture, and *iii*) optimiser.

In the first stage of the ablation study, we optimised the hyperparameters of the loss function

$$\begin{aligned}
\mathcal{L}_i &= -\mathcal{L}_{spatial,i} - \mathcal{L}_{filter,i} - \mathcal{L}_{frequency,i} && (2)\\
\mathcal{L}_{spatial,i} &= \lambda_s MSE(\hat{y}_i, y_i) && (3)\\
\mathcal{L}_{filter,i} &= \lambda_h MSE(\mathcal{F}[\hat{y}_i|h], \mathcal{F}[y_i|h]) + \lambda_\ell MSE(\mathcal{F}[\hat{y}_i|\ell], \mathcal{F}[y_i|\ell]) && (4)\\
\mathcal{L}_{frequency,i} &= \lambda_f L_1(FFT(\hat{y}_i), FFT(y_i)). && (5)
\end{aligned}$$

Here y_i and $\hat{y}_i$ are the input image and the reconstructed image, respectively. The 2D fast Fourier transform (FFT) of an image is represented here by $FFT(y_i)$. The per-sample loss function $\mathcal{L}_i$ was a weighted combination of $\mathcal{L}_{spatial,i}$ which optimised the model for reconstruction of the input

images, $\mathcal{L}_{filter,i}$ which optimised the reconstruction of the images filtered for high and low frequencies, and $\mathcal{L}_{frequency,i}$ which optimised the model for reconstruction of images in the frequency domain. The weights of each loss term were controlled by λ_s, λ_f and λ_h and λ_l.

We performed hyperparameter optimisation via grid search, training a model for different configurations of the hyperparameters in the above loss function, with the values stated in Tab. B.1. For each trial, we trained a model for 50 epochs on 2 A100 GPUs with a global batch size of 512, AdamW optimiser with L2 regularisation 0.002, a linearly increasing learning rate in the first 10 epochs from 1×10^{-5} to 1×10^{-4}, followed by cosine annealing for the next 40 epochs from 1×10^{-4} to 1×10^{-6}. The patch size of the model was set to $(16, 16)$, cell-centred tiles were sized at (112,112), patches were masked during training with fraction $p_m = 0.75$, and we trained the model using the 20 TARGET2 plates only, excluding the COMPOUND plates from the hyperparameter optimisation. We selected the best configuration based on the validation loss and validation reconstruction error, and highlight this configuration in Tab. B.1.

After finding the optimal configuration for the hyperparameters for the loss function, we performed the second stage our hyperparameter optimisation, focussing on hyperparameters concerning the neural network architecture. In this stage of hyperparameter optimisation, we varied the size of the patch size of the cell-centred tiles, whether or not we initialised our encoder with ImageNet1k weights, the size of our encoder (from the standard sizes Large or Huge provided by HuggingFace), the rate of stochastic depth in the attention blocks (Drop Path Rate), and whether patches were masked at random (Sync Mask is False) or patches are masked such that patches from the same position but different channel are masked together (Sync Mask is True). We trained a model for each of the configurations of these hyperparameters in Tab. B.2 for 50 epochs using 2 A100 GPUs with a global batch size of 128, and AdamW optimiser with L2 regularisation 0.002, a linearly increasing learning rate in the first 10 epochs from 1×10^{-5} to 1×10^{-4}, followed by cosine annealing for the next 40 epochs from 1×10^{-4} to 1×10^{-6}. We used the optimal loss hyperparameters from the first stage of hyperparameter optimisation. We highlight the optimal configuration in Tab. B.2. We then perturbed the p_m alone, and found that increasing the mask ratio $p_m = 0.8$ led to higher performance.

In the last stage of our ablation analysis, we optimised the learning rate and weight decay of our AdamW optimiser. We trained a model for each of the configurations of these hyperparameters in Tab. B.3 for 50 epochs using 2 A100 GPUs with a global batch size of 400, and AdamW optimiser with L2 regularisation 0.002, a linearly increasing learning rate in the first 10 epochs from 1×10^{-5} to 1×10^{-4}, followed by cosine annealing for the next 40 epochs from the learning rate start value for the given trial to 1×10^{-6}. The optimal configuration is shown in Tab. B.3.

Table B.1: Hyperparameter combinations tested during ablation study for loss function. Each row corresponds to a model trained with the stated configuration of hyperparameters. Best configuration is highlighted in bold.

Trial	λ_s	λ_h	λ_l	λ_f	h	ℓ
1	1.0	0.00	0.00	0.00	NA	NA
2	0.75	0.00	0.00	0.25	NA	0.1
3	0.75	0.00	0.25	0.00	0.1	NA
4	0.34	0.00	0.33	0.33	0.1	0.1
5	0.25	0.00	0.50	0.25	0.1	0.1
6	0.25	0.00	0.25	0.50	0.1	0.1
7	0.99	0.01	0.00	0.00	NA	NA
8	0.74	0.01	0.00	0.25	NA	0.1
9	0.74	0.01	0.25	0.00	0.1	NA
10	0.33	0.01	0.33	0.33	0.1	0.1
11	0.24	0.01	0.50	0.25	0.1	0.1
12	0.24	0.01	0.25	0.50	0.1	0.1
13	0.50	0.00	0.50	0.00	0.1	NA
14	0.25	0.00	0.75	0.00	0.1	NA
15	0.00	0.00	1.00	0.00	0.1	NA
16	0.75	0.00	0.25	0.00	0.05	NA
17	0.50	0.00	0.50	0.00	0.05	NA
18	0.25	0.00	0.75	0.00	0.05	NA
19	0.00	0.00	1.00	0.00	0.05	NA
20	0.90	0.00	0.10	0.00	0.1	NA
21	0.90	0.00	0.10	0.00	0.3	NA
22	0.90	0.00	0.10	0.00	0.5	NA
23	0.90	0.00	0.10	0.00	0.7	NA
24	0.75	0.00	0.25	0.00	0.2	NA
25	0.75	0.00	0.25	0.00	0.25	NA
26	**0.75**	**0.00**	**0.25**	**0.00**	**0.3**	**NA**
27	0.75	0.00	0.25	0.00	0.35	NA

Table B.2: Hyperparameter combinations tested during ablation study for neural network architecture. Each row corresponds to a model trained with the stated configuration of hyperparameters. Best configuration is highlighted in bold.

Trial	Patch Size	p_m	Drop Path Rate	Sync Mask	ImageNet1k weights	Encoder Size
1	16	0.75	0	True	True	Large
2	**14**	**0.75**	**0**	**True**	**True**	**Large**
3	8	0.75	0	True	True	Large
4	16	0.75	0.1	True	True	Large
5	14	0.75	0.1	True	True	Large
6	8	0.75	0.1	True	True	Large
7	16	0.75	0.2	True	True	Large
8	14	0.75	0.2	True	True	Large
9	8	0.75	0.2	True	True	Large
10	16	0.75	0.3	True	True	Large
11	14	0.75	0.3	True	True	Large
12	8	0.75	0.3	True	True	Large
13	16	0.75	0	True	False	Large
14	14	0.75	0	True	False	Large
15	8	0.75	0	True	False	Large
16	16	0.75	0	False	True	Large
17	14	0.75	0	False	True	Large
18	8	0.75	0	False	True	Large
19	14	0.75	0	True	True	Huge
20	16	0.75	0	True	True	Huge
21	14	0.75	0	True	False	Huge
22	16	0.75	0	True	False	Huge

Table B.3: Hyperparameter combinations tested during ablation study for the optimiser. Each row corresponds to a model trained with the stated configuration of hyperparameters. Best configuration is highlighted in bold.

Trial	Learning Rate	Weight Decay
1	1.0e-4	2.0e-3
2	5.0e-5	2.0e-3
3	1.5e-4	2.0e-3
4	2.0e-4	2.0e-3
5	1.0e-4	5.0e-3
6	1.0e-4	1.0e-2
7	5.0e-4	2.0e-3
8	1.0e-4	1.0e-3
9	3.0e-4	5.0e-3
10	4.0e-4	5.0e-3
11	**5.0e-4**	**5.0e-3**
12	6.0e-4	5.0e-3

C Z'-SCORE FROM MODEL EMBEDDINGS

The Z'-score is typically computed for scalar values. Here we describe the steps taken to compute the Z'-score to measure the statistical size effect between different groups of model embeddings, corresponding to different compound stimulations.

The dataset we finetuned **Campfire** with in Sec. 3.5 was comprised of plates with wells that were treated with either negative or positive controls. Control compounds, with a priori negative or positive effect on cell morphology, can be used to measure assay quality, but were here used to evaluate model embeddings. From the M1 and M2 plates we computed the well-level embeddings for each well stimulated by positive or negative controls. These wells were labelled {neg-M1, neg-M2, pos-M1, pos-M2}.

Our aim was to compute the Z' score for each pairwise combination of these labels. To do so we converted our model embeddings to a single value corresponding to the difference in the clusters of embeddings for each label. Given two groups of embeddings, X and Y with labels ℓ_x and ℓ_y, we treated X as the *reference* group, and Y as the *target* group. We computed the mean embedding from the group X, μ_x, and subtracted this from all embeddings in both groups i.e $X = X - \mu_x$ and $Y = Y - \mu_x$. We then computed the mean embedding of Y, denoted μ_y, and for each embedding in X and Y we computed the normalised scalar projection on to μ_y. Hence, for the target and reference group, each embedding was represented by a single value quantifying its alignment with the target group. We computed the mean, μ_t and μ_r, and standard deviation, σ_t and σ_r, of these values for the target (t) and reference (r) group and then computed Z' using the formula,

$$Z' = 1 - 3\frac{\sigma_r + \sigma_t}{|\mu_r - \mu_t|}. \tag{6}$$