Capturing cell heterogeneity in representations of cell populations for image-based profiling using contrastive learning

Robert van Dijk, John Arevalo, Mehrtash Babadi,
Anne E. Carpenter, Shantanu Singh

Abstract

Image-based cell profiling is a powerful tool that compares perturbed cell populations by measuring thousands of single-cell features and summarizing them into profiles, typically by averaging across cells. Although average profiling is commonly used, it fails to capture the heterogeneity within cell populations. We introduce CytoSummaryNet: a machine learning approach for summarizing cell populations that outperforms average profiling in predicting a compound's mechanism of action. CytoSummaryNet uses weakly supervised contrastive learning in a multiple-instance learning framework and provides an easier-to-apply method for aggregating single-cell feature data than previously published strategies. Interpretability analysis suggests that CytoSummaryNet achieves this by downweighting noisy cells (small mitotic cells or those with debris) and prioritizing less noisy cells (large uncrowded cells). Remarkably, CytoSummaryNet may also mitigate batch effects, even though this was not part of the training objective. Finally, the framework is designed to facilitate retraining, employing weak labels derived from perturbation replicates that are readily available in all cell profiling datasets. We show on a public dataset that CytoSummaryNet aggregated profiles result in a 68% increase in the mean average precision of mechanism of action retrieval compared to the commonly used average-aggregated profiles.

Introduction

High-throughput assays enable quantifying cellular responses to perturbations at a large scale. Image-based assays are among the most accessible and inexpensive such technologies that offer single-cell resolution. Cell populations are perturbed with compounds or genetic perturbations, stained, and then imaged. Large amounts of quantitative morphological data are extracted from these microscopy images, generating tabular data comprising single cell profiles. These single cell profiles are then aggregated by a variety of methods to a per-perturbation profile that describes that population's phenotype. The profiles of different cell populations can be compared to identify previously unrecognized cell states induced by different experimental perturbations of interest. This method, called image-based cell profiling, is a powerful tool that can be used for drug discovery, functional genomics, and disease phenotyping [1]. Among other applications, image-based profiling has already been used to predict assay outcomes for compounds [2]–[4], detect leukemia label-free [5], and predict the impact of particular gene mutations [6].

Image-based cell profiling shows great potential, but many steps in its pipeline can still be improved [1]. One of the profiling challenges is to capture both population trends and single-cell variability. Cell populations are known to be heterogeneous [7], [8], and recent studies have yielded many insights into the mechanisms and importance of this characteristic [9]–[12]. Capturing that heterogeneity could improve a profile's information content and, thus, utility in many image-based profiling applications.

Despite its current limitations, so-called population-averaged profiling, where all single-cell features are averaged per feature using either the mean or the median, remains the most commonly-used approach in the field of image-based profiling. This is true regardless of the types of features or the kinds of transformations (e.g., normalization, feature selection, etc.) applied to the profiles [13], [14]. Average profiling is a simple way of summarizing a cell population (hereafter referred to as a sample) into a vector (a sample's profile) with only one value per measured feature. It dramatically decreases the data size (as there are typically thousands of cells per well, hundreds of wells per plate, and multiple plates per experiment) and simplifies downstream analysis.

The biggest drawback of average profiling is that information on cell subpopulations is lost. This can result in identical average-aggregated profiles despite cell populations having distinct internal structures. Additionally, ignoring subpopulations can lead to a quantitatively incorrect interpretation. For example, two cell populations can show correlations among certain features when averaged but show completely different relations when compared after grouping the cells, i.e., Simpson's paradox [15]. Lastly, averaging a sample essentially assumes each measured feature corresponds to a simple unimodal distribution. If this is not the case, e.g., as in the case of heterogeneous cell populations, the average will be a poor summary statistic for the data.

Several methods have been proposed to capture the heterogeneity of cell populations into their corresponding profiles. The most straightforward solution is to incorporate the cell population's dispersion (e.g., standard deviation) for each extracted feature and concatenate these values with the average-aggregated profile. This approach is widely adopted but offers only minor improvements over average profiling [16], although a later study suggested that incorporating the sample histogram – instead of dispersion – may offer improvements [17]. A different approach involves first clustering cells and then calculating the profiles based on their subpopulations [18], [19]. These methods capture more information about subpopulations rather than only incorporating their dispersions but did not significantly improve upon average profiling [16]. Furthermore, they can lead to incomparable profiles across experiments unless the subpopulations are defined beforehand. As well, many cell phenotypes are better described with a continuous rather than a discrete scale [13].

Recently, we improved the performance of average profiling by fusing features' averages, dispersion, and covariances [13]. This method provided ~20% better performance predicting a compound's mechanism of action and a gene's pathway, showing that capturing statistics related to cell population heterogeneity can improve performance on downstream tasks. However, this method has two major limitations. First, it only captures the first- and second-order moments of the data. Second, because it produces a similarity matrix rather than an embedding, it requires recomputing the pairwise similarities among all profiles each time a new profile is included in the dataset. Here, we introduce a novel method that addresses both of these limitations and automatically finds an effective way to aggregate single-cell data to improve the information content of sample profiles.

Results

We propose CytoSummaryNet: a weakly-supervised contrastive learning approach that leverages the naturally available information in profiling experiments. Contrastive learning is a method where data points corresponding to the same entity – for example the same sample – are brought closer together while others are pushed apart in a feature space [20]. In doing so, it aims to capture generalizable features beneficial for a broad range of tasks. Here, we use perturbation identifications (IDs) as labels to train a latent feature space that distinguishes samples with different IDs. In this feature space, profiles of replicates with the same perturbation should be close to each other, while those of different perturbations should be further apart. This labeling approach frames the problem as a multiple-instance learning problem [21], assuming that the replicate wells consist of cells with similar feature distributions, and that different compounds generate populations with distinct feature distributions. Although not every perturbation yields a profile distinct from all others, these assumptions collectively contribute to the development of a feature embedding that captures biologically significant morphological variations. Importantly, we here apply contrastive learning to already-extracted single-cell features, as opposed to raw pixels, making it immediately applicable to most image-based profiling datasets, where extracted features are readily available.

We treat the data as a collection of sets of cells, where each sample (all cells from a single well) corresponds to one set (based on the mathematical definition of "set"). To aggregate cells from such a sample into a profile, the function should possess a few properties. First, it should be capable of handling input samples of arbitrary sizes. Second, because cells within a sample are, by definition, unordered, the function should be permutation invariant. Several methods have been developed for analyzing this type of data [22], [23], and a general formulation for addressing this problem is known as Deep Sets [24]. Zaheer et al. [24] revealed that a function acting as a universal approximator for sets has a specific structure, consisting of a permutation-invariant function and a learnable representation function. This structure provides a foundation for designing neural networks capable of processing unordered data represented as sets.

We therefore chose to employ the Deep Sets formulation to learn a model (CytoSummaryNet) that aggregates single-cell feature data into a profile (CytoSummaryNet profiles) that outperforms population-averaged profiles in predicting a compound's mechanism of action. This is achieved through weakly-supervised contrastive learning within a multiple-instance learning framework, which allows the model to process groups of data – labeled only at the group level rather than the individual data point level – and to make classifications based on the collective information of these groups. We evaluate the training task (replicate retrieval, i.e., retrieving profiles of other replicate wells of the same compound) and the downstream task (retrieving profiles of other compounds annotated with the same mechanism of action) using the mean average precision (mAP) metric. This information retrieval metric indicates whether all the positive examples can correctly be identified without accidentally marking too many negative examples as positive. We evaluated CytoSummaryNet on two fronts: (i) testing its generalizability across unseen compounds and experimental protocols, and (ii) a practical application scenario where we trained the model on a large dataset and subsequently measured its performance on mechanism of action retrieval for that dataset.

We first investigated CytoSummaryNet’s capacity to generalize to unseen compounds and experimental protocols. We performed this analysis on the cpg0001 [25] dataset from the public Cell Painting Gallery, which consists of 384-well plates with identical sample layouts of 90 unique compounds, i.e., there is a single “plate layout” for the entire dataset (See *Experimental Setup: Data*); each plate contains four replicates of each compound in different well positions, plus 24 negative controls. The dataset consists of many experimental plates, most with different experimental conditions to optimize the image-based cell profiling protocol. The dataset can be categorized into four subsets of plates: Stain2, Stain3, Stain4, and Stain5. Each of these subsets corresponds to a specific set of assay conditions that were designed to optimize Cell Painting – the most widely used image-based profiling assay – for detecting morphological phenotypes and grouping similar perturbations together [26].

Each StainX subset was processed as a distinct batch, which means batch effects need to be taken into account. Stain2, Stain3, and Stain4 were split into train, validation, and test plates. Stain5 consists of only test plates and is considered completely out-of-distribution because no data from this subset is used during the training phase. Training plates are employed to update the model weights, validation plates serve to select the optimal model, while test plates are reserved for complete hold-out evaluation. Further, a subset of 18 compounds were designated as *validation compounds*; the wells in the training plates corresponding to these compounds were excluded from training. Compounds that were seen during training are termed *training compounds*; note that (1) wells corresponding to training compounds from the training plates are seen during training, but (2) wells corresponding to training compounds from the test/validation plates are not seen during training. Given the limited number of compounds, this intricate strategy allowed us to evaluate the performance of CytoSummaryNet on 18 compounds that were not seen during training. See *Experimental Setup: Stratification* for further details.

Profiles from CytoSummaryNet trained on Stain2, Stain3, and Stain4 consistently showed significantly higher mAP scores for the replicate retrieval task than the baseline average-aggregated profiles, on the training and validation plates of Stain3 (Figure 1). CytoSummaryNet achieved comparable results on Stain2 and Stain4 (*Supplementary Material A*). CytoSummaryNet aggregation also improves the mAP score compared to average profiling on the test plates, however, the improvement is not statistically significant, indicating some degree of overfitting. The improvement CytoSummaryNet provides is diminished overall for held-out validation compounds compared to compounds included in the training set, indicating some degree of overfitting on the compounds seen during training (Figure 2). The poor performance on Stain5, the subset that was not used for training CytoSummaryNet, suggests that CytoSummaryNet does not generalize to out-of-distribution batches of data.

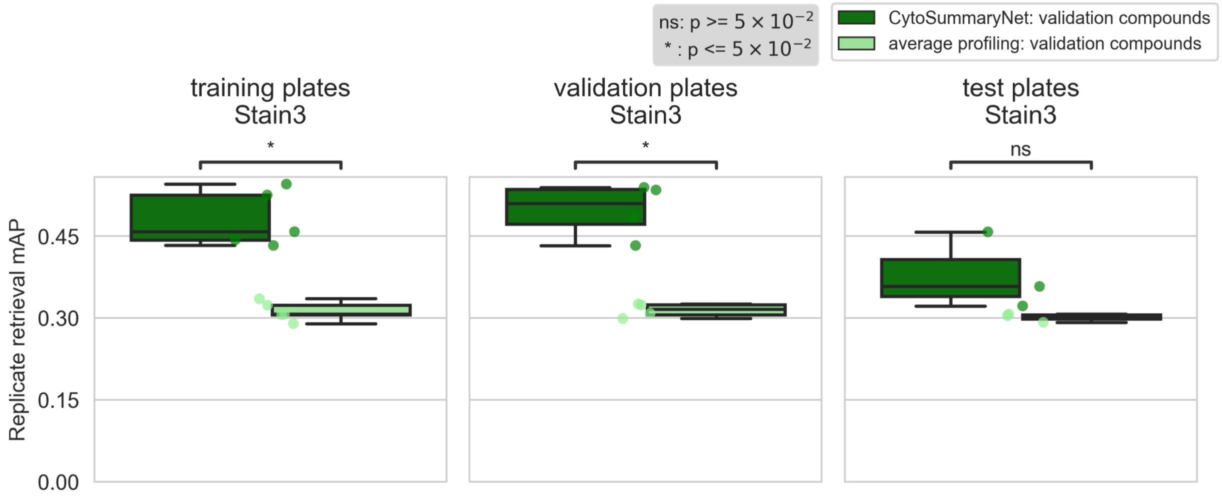

Figure 1: CytoSummaryNet profiles generally outperform average-aggregated profiles for sensitively identifying replicates of a given sample, and partially generalize to unseen experimental protocols (test plates). The box plots illustrate the mAP of replicate retrieval for all validation compounds of Stain3 (averaged per plate) by CytoSummaryNet (dark green) and average (light green) profiles. Welch's t-tests were used to compare the means between CytoSummaryNet and average mAP scores on corresponding data; their p-values are indicated as stars at the top of each plot (ns = not significant).

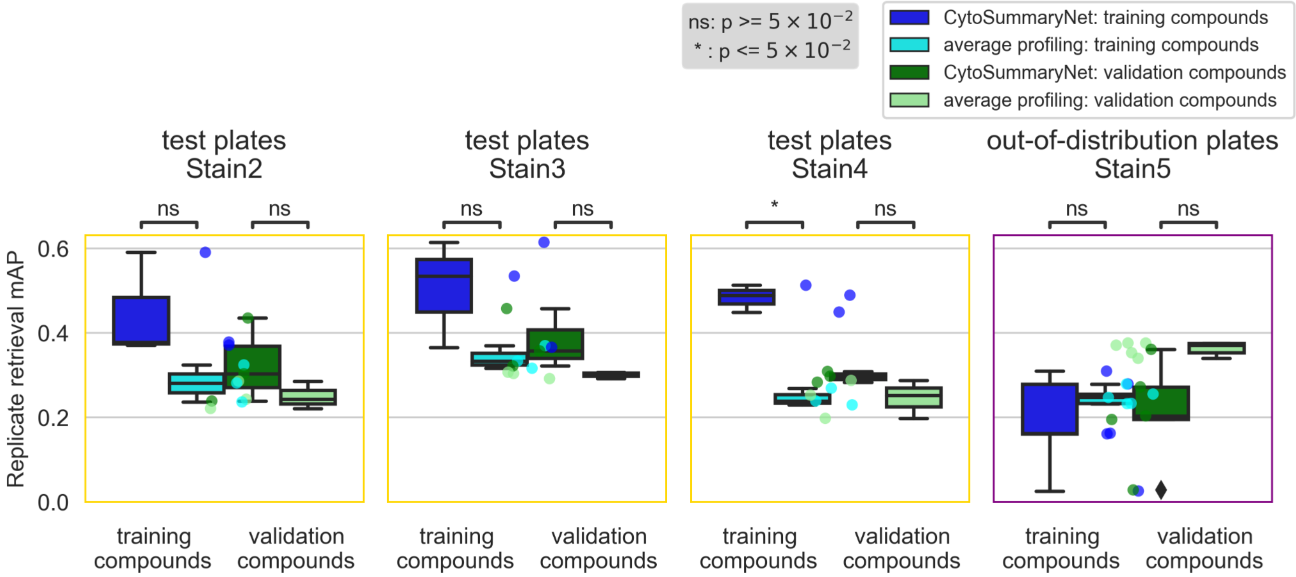

Figure 2: CytoSummaryNet profiles partially generalize to unseen compounds and do not generalize to out-of-distribution batch data (Stain 5). The box plots illustrate the mAP of replicate retrieval for training and validation compounds (averaged per plate) from the test plates of Stain2, Stain3, Stain4, and Stain5; CytoSummaryNet profiles performance in dark blue and dark green respectively, and average profiles performance in cyan and light green respectively. Although the boxes are labeled "training compounds" and "validation compounds", note that all data shown comes from test plates and therefore none of it has been seen during training (see description of stratification in Results *for further details). Welch's t-tests were used to compare the means between CytoSummaryNet and average mAP scores on corresponding data; their p-values are indicated as stars at the top of each plot (ns = not significant).*

Next, we turned to a more challenging task: predicting the mechanism of action class for each compound, rather than simply matching replicates of the exact same compound (Figure 3). Note that this is a downstream task that CytoSummaryNet is not trained on, and so improvements observed on the training and validation plates are more meaningful (unlike in the previous task, where only improvements on the test plate were meaningful). As expected, the baseline mAP for this task is much lower, given in part the well-known imperfect annotation of compounds with respect to their mechanism. Still, we find that CytoSummaryNet profiles significantly improve the mAP scores for mechanism of action retrieval compared to average-aggregated profiles. There is one exception: Stain5 test plates, which require generalizing to out-of-distribution batch data; here, CytoSummaryNet's performance was poor (Figure 3), which is consistent with the replicate retrieval results (Figure 2).

Overall, on the mechanism-retrieval task, the CytoSummaryNet profiles achieve a mAP that is 68%, 61%, and 49% higher than average profiling for all of the training, validation, and test set plates (excluding the out-of-distribution data from Stain5), respectively (Table 1, top panel). Interestingly, average profiling demonstrates superior performance on out-of-distribution Stain5, outperforming CytoSummaryNet with a 36% higher mAP on average.

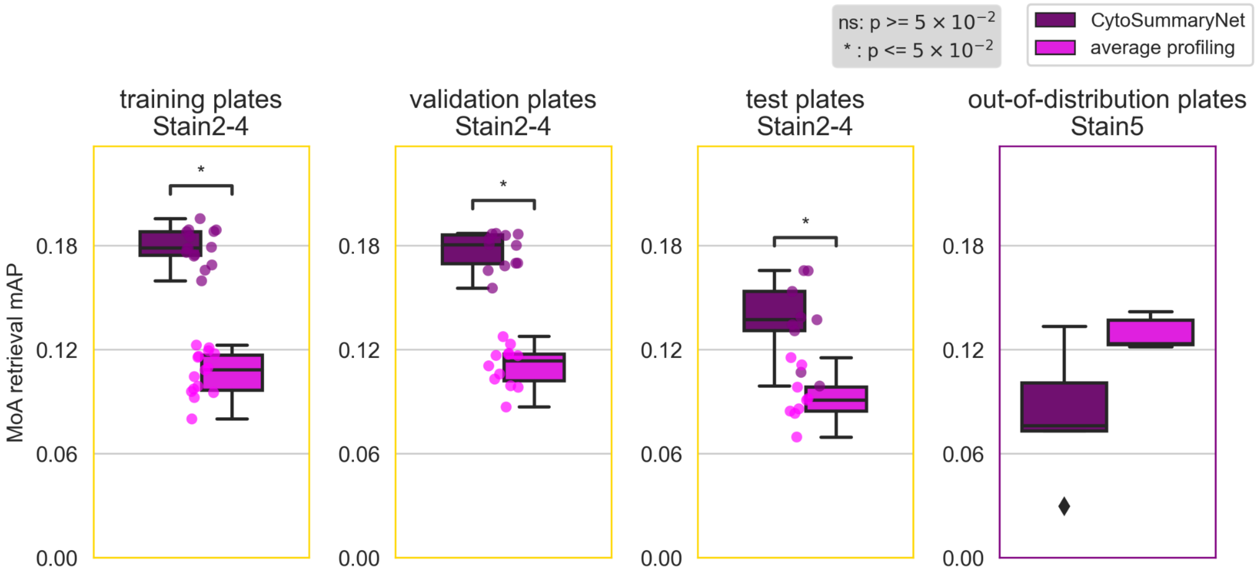

Figure 3: CytoSummaryNet profiles generally outperform average-aggregated profiles in mechanism of action (MoA) retrieval, although not for out-of-distribution batch data (Stain5). The box plots illustrate the average plate-level mAP of mechanism of action retrieval for CytoSummaryNet- (dark purple) and average- (light pink) aggregated profiles. Welch's t-tests were used to compare the means between CytoSummaryNet and average mAP scores; their p-values are indicated as stars inside each plot (ns = not significant).

Table 1: Top panel: Absolute and relative average improvements in mAP of mechanism of action retrieval between CytoSummaryNet- and average-aggregated profiles for the cpg0001 dataset. The improvements are calculated as $mAP(CytoSummaryNet) - mAP(average)$*. Bottom panel: The same mAP improvements but for the cpg0004 dataset.*

dataset	stratification	mAP improvement Stain2	mAP improvement Stain3	mAP improvement Stain4	average mAP improvement Stain2, Stain3, and Stain4	mAP improvement Stain5
cpg0001	training	0.081 (81%)	0.065 (55%)	0.073 (69%)	0.073 (68%)	
	validation	0.068 (70%)	0.060 (50%)	0.071 (63%)	0.066 (61%)	
	test	0.043 (52%)	0.039 (36%)	0.052 (60%)	0.045 (49%)	-0.047 (-36%)

		mAP improvement
cpg0004	10 μM (training/ validation)	0.021 (68%)
	3.33 μM (test)	0.009 (30%)

We then tested the strategy in a more practical use case: CytoSummaryNet is trained using the compound IDs and then used to infer the improved profiles for the mechanism of action retrieval task directly afterward on the same dataset. For this purpose, we trained the model using the 10 μM dose point data of Batch 1 of the cpg0004 dataset [27], which includes 1,258 unique compounds. The multiple dose points in this dataset allow us to create a separate hold-out test set: we used the 3.33 μM dose point data to test for generalization.

In this context, we found that mechanism of action retrieval is again more successful using CytoSummaryNet profiles versus the average baseline (*Figure 4*; most data points are found in the lower right, using the identity line as the divider). The mAP averaged over all mechanisms of action is 68% and 30% higher when using CytoSummaryNet profiling compared to average profiling for the 10 μM (training/validation) and 3.33 μM (test) dose points, respectively (Table 1, bottom panel). CytoSummaryNet generally amplifies the strength of profiles that already achieve a mAP higher than 0.1 with average profiling, i.e., some signal is present already. The points near the origin show that CytoSummaryNet also amplifies the strength of some profiles that were not visible before using average profiling by increasing the mAP from less than 0.05 to more than 0.1. The use of CytoSummaryNet can thus make readily identifiable mechanisms of action even easier to find and allow for the discovery of previously unfindable mechanisms of action. Few CytoSummaryNet profiles achieve a lower mAP than average profiling, suggesting a negligible drawback to the use of CytoSummaryNet.

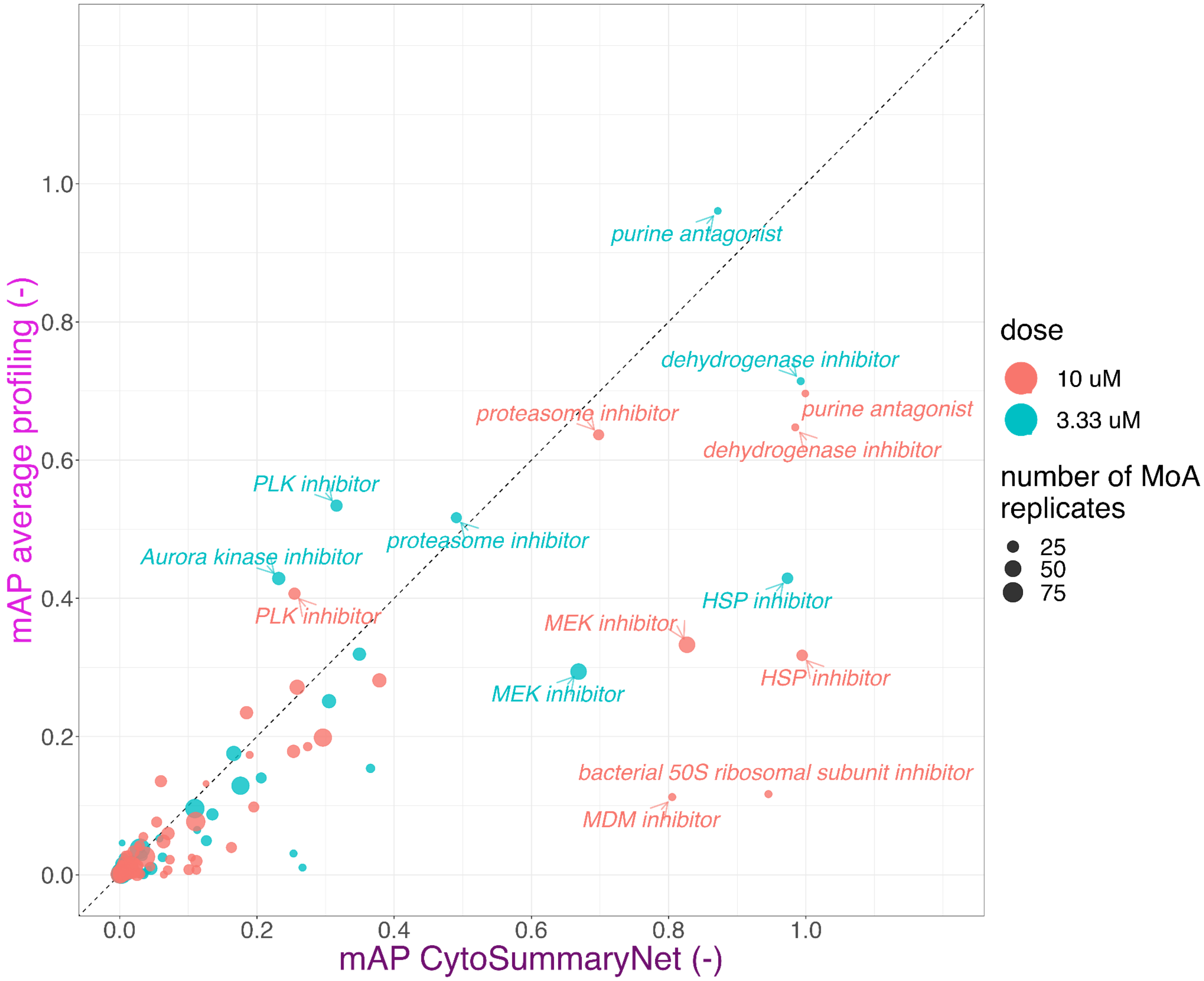

Figure 4: CytoSummaryNet profiles make readily identifiable mechanisms of action even easier to find and allow for the discovery of previously unfindable mechanisms of action when using average-aggregated profiles (cpg0004 dataset). Mean average precision (mAP) of average and CytoSummaryNet-based profiling for mechanism of action retrieval of 3.33 μM (test) and 10 μM (training/validation) dose point compound perturbations. We highlight certain high-performing mechanisms of action. Data points are scaled in size based on the number of mechanism of action (MoA) replicates used to compute them.

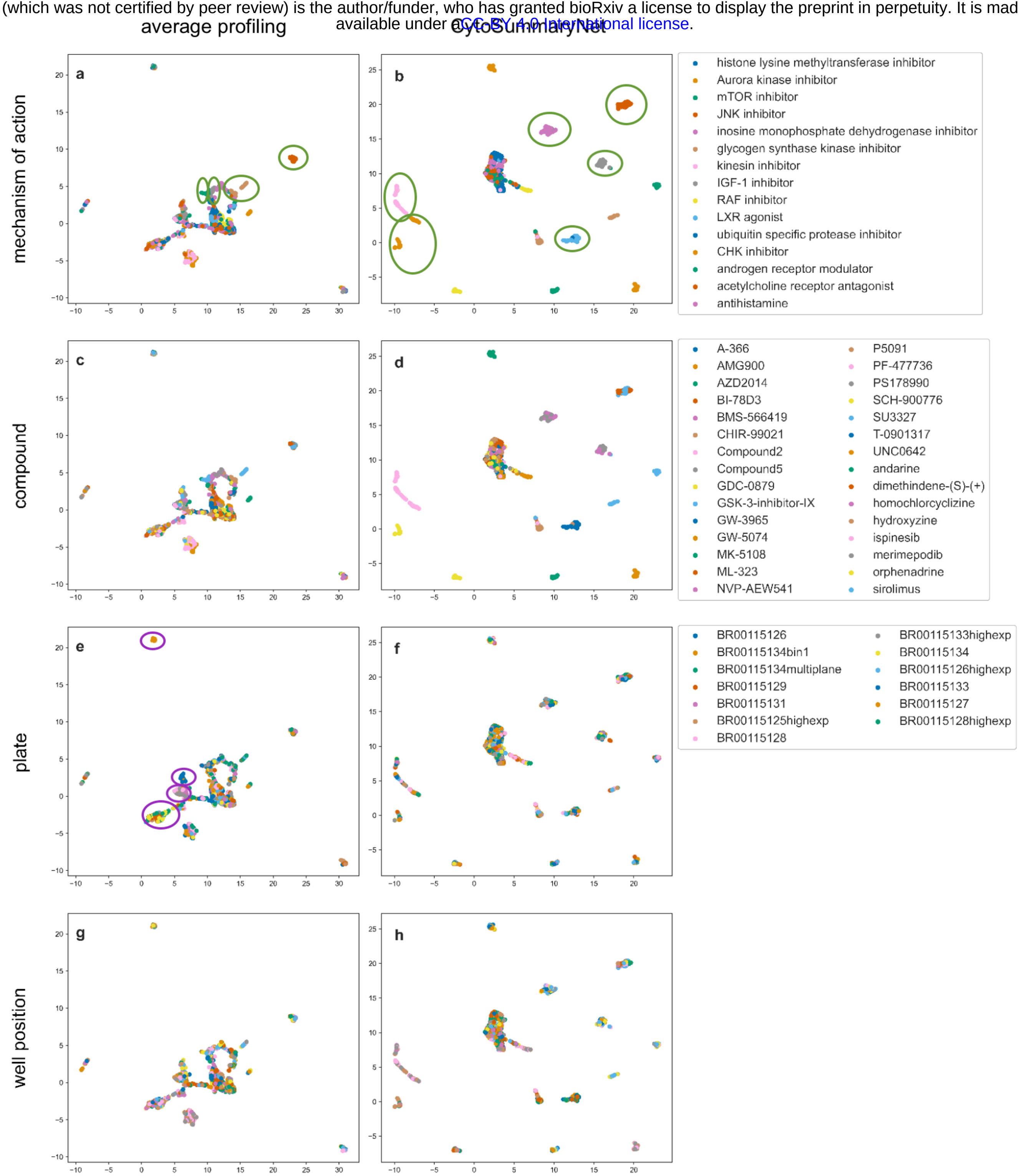

Figure 5: CytoSummaryNet profiles show a better ability to distinguish similar samples and overcome batch effects than average-aggregated profiles (cpg0001 dataset). UMAP of the average- (left column) and CytoSummaryNet- (right column) aggregated profiles of the top 15 mechanisms of action, based on CytoSummaryNet's mAP scores for mechanism of action retrieval, from all used cpg0001 Stain3 plates. The UMAP was created using $n\_neighbors = 15$ *and cosine similarity as a distance measure. The profiles are colored based on their corresponding annotated mechanism of action, compound, plate, and well position from top to bottom, respectively. Same mechanism of action profile clusters and isolated plate clusters that were visible for multiple* $n\_neighbors$ *values are annotated in green and purple ellipses, respectively.*

Interpretability analysis

We next interrogated CytoSummaryNet to better understand how it was achieving improved results for replicate and mechanism of action retrieval. As discussed in the introduction, some improved performance was previously achieved by using second-order moments in addition to the average [13]. Here, we performed supporting experiments (*Supplementary Material B*) using a toy dataset which indicates that CytoSummaryNet is able to learn second-order moments (covariance) as well as third order moments (skewness). We then use uniform manifold approximation and projection (UMAP [28]) and saliency analyses to investigate what CytoSummaryNet has learned.

A UMAP shows that the CytoSummaryNet profiles are better able to distinguish similar samples (Figures 5b and 5d) and overcome batch effects (Figure 5f) than average-aggregated profiles (Figures 5a, 5c, and 5e). For visualization purposes, only plates from Stain3 are shown; Stain2 and Stain4 showed similar results (data not shown). The UMAP of the CytoSummaryNet profiles shows six clusters of compound profiles that are easily distinguishable from the rest of the profiles; within each cluster, the compounds share the same mechanism of action annotation. The remaining clusters located beyond the central main cluster are still organized based on their compound replicates despite not being grouped by their mechanism of action; no clusters with a mixture of mechanisms of action are seen, and no clusters are formed according to plates.

By contrast, the average-aggregated profiles' UMAP shows only four clusters of compound profiles that are easily distinguishable from the rest of the profiles and within each, share a common mechanism of action (Figures 5a and 5c). The other compound profiles are either part of the main cluster in the center of the plot or one of the three smaller clusters in the top-left, bottom-right, or left of the plot. The top-left cluster contains profiles from only a single plate with a very different experimental protocol than the others which involved microscopy image pixel binning before extracting the single-cell features (Figure 5e; Supplementary Material F also shows how this plate (BR0015134bin1) differs strongly from other plates in the Stain3 subset using a similarity analysis). At the edges of the main cluster lie three more clusters of single plates.

Because the improved results could stem from prioritizing certain features over others during aggregation, we investigated each cell's importance during CytoSummaryNet aggregation by calculating a relevance score for each. The features considered were extracted using CellProfiler [29], with a pipeline that measures thousands of single-cell features. The relevance score, which assesses the importance of each cell in the aggregation process by CytoSummaryNet, is computed by combining sensitivity analysis (SA) and critical point analysis (CPA). SA evaluates the model's predictions by analyzing the partial derivatives in a localized context, while CPA identifies the input cells with the most significant contribution to the model's output (see *Methods* for details).

The combination of CellProfiler features that most highly correlate with the SA and CPA combined relevance score (Table 2) point to [features associated with] crowded cells (which are smaller and often have elevated DNA signal from adjacent cells measured in the cytoplasm or at the cell edge) being down-weighted in the final aggregated profile, while bigger and more isolated cells receive a larger weight. *Supplementary Material E* lists an additional 15 most correlated features.

Table 2: Top 5 CellProfiler features based on their positive and negative Pearson correlation coefficient with the SA and CPA combined relevance score. The scores were calculated for a single test plate of cpg0001 Stain3 (200922_015124-V).

Feature category	Feature name	Correlation coefficient
Cytoplasm	Cytoplasm_Correlation_K_DNA_Brightfield	0.72
Cells	Cells_AreaShape_MeanRadius	0.71
Cells	Cells_AreaShape_MaximumRadius	0.70
Cells	Cells_AreaShape_MedianRadius	0.70
Cells	Cells_AreaShape_Area	0.68

Feature category	Feature name	Correlation coefficient
Cells	Cells_Intensity_MeanIntensityEdge_DNA	-0.74
Cytoplasm	Cytoplasm_Intensity_MeanIntensityEdge_DNA	-0.72
Cytoplasm	Cytoplasm_Intensity_UpperQuartileIntensity_DNA	-0.71
Cytoplasm	Cytoplasm_Intensity_MeanIntensity_DNA	-0.69
Cytoplasm	Cytoplasm_Correlation_K_Brightfield_DNA	-0.67

We examined an image of a well, chosen based on the corresponding profile's relative increase in mechanism of action retrieval mAP score using CytoSummaryNet profiling compared to average profiling (*Figure 6)*. The most relevant cells are larger, generally isolated from other cells (not touching them), and do not contain spots of high-intensity pixels (green arrow). The least relevant cells exhibit the opposite behavior; they are more often clumped together (bottom purple arrow) or contain spots of high-intensity pixels (top left and top right purple arrows). These findings agree with the conclusions drawn from Table 3 despite being extracted from a separate experimental batch (Stain2 instead of Stain3), indicating that prioritizing certain types of cells is a generalizable way for CytoSummaryNet to improve upon average profiling.

Computation time and storage requirements comparison

To assess the practicalities of CytoSummaryNet, we quantified the computation time and storage capacity requirements of CytoSummaryNet profiling compared to average profiling. CytoSummaryNet profiling adds computation time and storage requirements during aggregation but decreases the final storage size of aggregated profiles compared to average profiling (Table 3). CytoSummaryNet training requires an intermediate form of single-cell data storage for quick access and takes one to two days depending on the type of CPU (or GPU) used, here an Intel(R) Xeon(R) Platinum 8375C CPU @ 2.90GHz and 2.4 GHz 8-Core Intel Core i9, respectively. The intermediate data storage size is a fraction (0.44) of the original raw SQLite file due to the conversion of double precision to single precision numbers. The latter is also why the final aggregated profiles take up less storage space when using CytoSummaryNet aggregation instead of the average, despite CytoSummaryNet profiles containing more features.

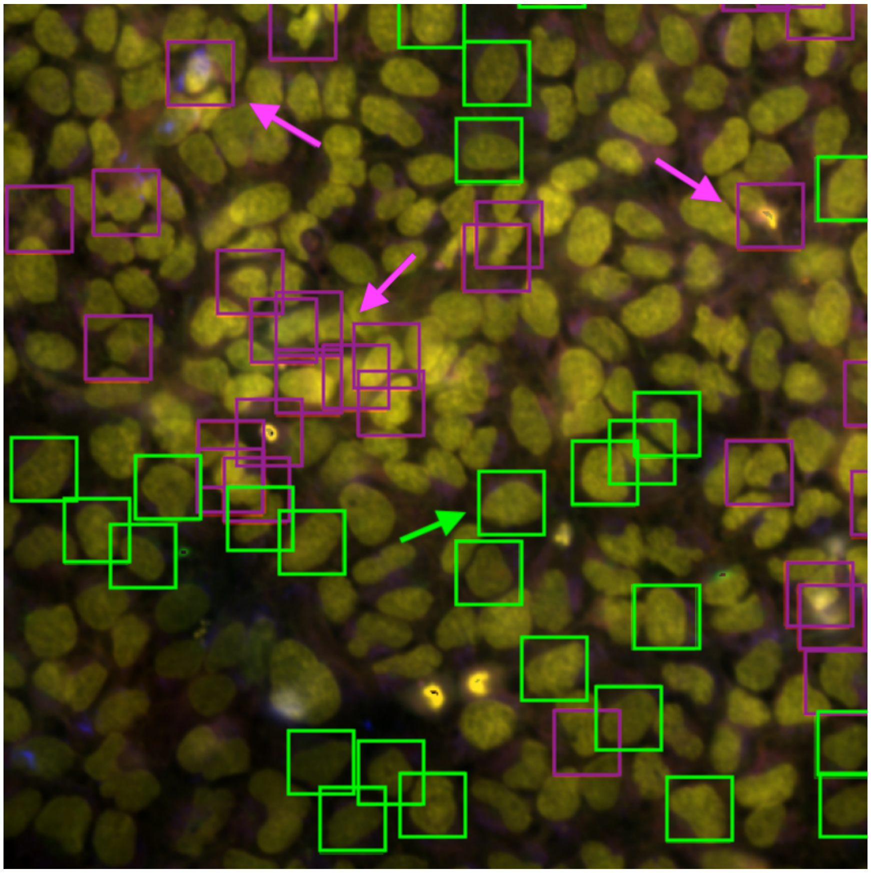

Figure 6: Five-channel combined microscope image of one of the four fields of view for a well in plate cpg0001 Stain2 BR00112197binned. The most relevant cells are annotated with green boxes and the least relevant cells are annotated with purple boxes. Three cells characteristic for low relevance scores are explicitly labeled with purple arrows. One cell characteristic for high relevance scores is explicitly labeled with a green arrow. The other three images from this well are shown in Supplementary Material G.

*Table 3: CytoSummaryNet profiling adds computation time and storage requirements during aggregation but decreases the storage size of aggregated profiles compared to average profiling. The average of various computation times and storage sizes are listed for both CytoSummaryNet and average profiling. *CytoSummaryNet training time is highly dependent on the CPU (or GPU) used.*

method	**preprocessing time per plate (seconds)**	**intermediate data storage size per plate (GB per raw SQLite GB)**	**model training time (hours)**	**profile aggregation computation time per plate (seconds)**	**csv file size of aggregated profiles per plate (MB)**
CytoSummaryNet profiling	~50	0.44	~26 to ~51*	20.1	~8
average profiling	~40	N/A	N/A	0.4	~30

Discussion

This study proposes a Deep Sets-based model "CytoSummaryNet" that automatically finds the best way to aggregate single-cell population profiles by using weakly supervised contrastive learning. This method learns an aggregation strategy by receiving single-cell morphological feature input data and providing the aggregated profile as an output. CytoSummaryNet generally outperforms average profiling in both replicate retrieval and mechanism of action retrieval, perhaps in part due to its correction of batch effects. The relative increase in performance depends on how CytoSummaryNet is applied. For mechanism of action retrieval on our primary dataset, cpg0001, CytoSummaryNet achieves 36% to 60% better performance than average profiling on hold-out plates created with different experimental protocols than the training plates, provided that these hold-out plates are drawn from the same batch distribution. Similarly, a 30% improvement is achieved compared to average profiling on a second dataset, cpg0004 3.33 µM hold-out data from the same batch. For mechanism of action retrieval on cpg0001 training and validation plates, CytoSummaryNet achieves 50% to 81% better performance than average profiling, which is again affirmed by the 68% performance increase found for the cpg0004 training data. Note that mechanism of action retrieval is a downstream task unrelated to the task for which CytoSummaryNet is trained; therefore it is meaningful to report results on not just the test plates but also the training and validation plates. Overall, these results improve upon the ~20% gains we previously observed using covariance features [13] and importantly, CytoSummaryNet also effectively overcomes the challenge of recomputation after training, making it easier to use. CytoSummaryNet profiling adds computation time and storage requirements during the training process but also decreases the storage size of aggregated profiles compared to average profiling.

We noted that CytoSummaryNet has more trouble with generalizing to test plates, not seen during training, both those within distribution (Stain2-4) and out-of-distribution (Stain5). It's worth mentioning that our decision to average the mAP scores per plate was made to enhance clarity and reduce noise. However, this decision also means we have a limited number of data points available for significance testing. This likely mattered for the within-distribution test plates in the replicate retrieval task on the cpg0001 dataset, where CytoSummaryNet's improvement over average profiling is visible but not statistically significant. We also investigated why CytoSummaryNet encounters difficulties in generalizing to two Stain2 test plates in particular. The two outliers had more dramatically different experimental protocols than other plates: plate BR00112199 was imaged with multiple z-planes, which were then z-projected prior to further processing. All other plates had only a single z-plane imaged. The other plate (BR00113821) had far fewer cells seeded on the plates; 1250 instead of the usual 2500, which would clearly impact a method that weights features relating to crowded cells differently.

We also could explore generalizability in the cpg0004 dataset, which had the same compounds tested at two doses. The mAP scores for mechanism of action retrieval using CytoSummaryNet profiling are lower for the unseen 3.33 µM dose point data than for the seen 10 µM dose point data. However, CytoSummaryNet profiling still improves upon average profiling, showing that CytoSummaryNet is able to extract some general biological features; it is just not able to completely generalize to unseen data. A possible explanation is that CytoSummaryNet learns certain feature patterns that are not (as strongly) present in lower dose points. Additionally, as noted before, CytoSummaryNet tends to amplify already existing strong phenotypes. The 10 µM dose point data may therefore show larger increases in mAP for mechanism of action retrieval than the 3.33 µM dose point data because higher dose points often cause stronger phenotypic profiles than lower dose points.

When trained on and applied to data within a given dataset (e.g., the StainX subsets), CytoSummaryNet creates a better-organized feature space for discerning different compounds than average profiling, and remarkably mitigates plate-related batch effects almost entirely, even though this was not part of the training objective. Still, not all compounds with the same mechanism of action cluster well. As discussed elsewhere [27], there are many reasons that compounds with a particular annotated mechanism of action would not have an identical impact on a cell system; for example, many of the compounds have more than one annotated mechanism or have unannotated off-target effects. As well, the cell type or dose of a given compound may not be optimal for its impact to be visible in the cells chosen. Finally, technical artifacts, such as the influence of well position effects might introduce a bias.

We identified the likely mechanism by which the learned CytoSummaryNet aggregates cells: the most salient cells are generally larger and more isolated from other cells, while the least salient cells appear to be smaller and more crowded, and tend to contain spots of high-intensity pixels (whether dying, debris or in some stage of cell division). It makes sense that a compound’s effects would be more visible in large, flat, uncrowded cells given that overall, phenotypes are more visible in these conditions. This is consistent with the recent finding that "cells outside the locally crowded areas generally display more phenotypic variance and abundant state-specific genetic interactions" [30], and the success of identifying genotype-phenotype associations by separating cells touching others from isolated cells [31].

We found that the proposed CytoSummaryNet’s performance is limited to data that was created under similar experimental conditions as the training data. Although inconvenient, this is not a major limitation in practice, because many experiments are performed using the same experimental protocol and analysis pipeline. Even in cases when a new model must be trained (as tested here on the cpg0004 10 µM dataset), the labels required for training the weakly-supervised model are almost always available in profiling experiments. The only requirement is having a sufficient number of compounds with multiple replicates available; though not exhaustively analyzed, our data suggests that around 90 compounds with four replicates each may be sufficient.

CytoSummaryNet could be improved further by increasing its generalization capabilities so that it does not need to be retrained for each dataset or experiment. This might be achieved by adding more variation to the training data, e.g., by combining multiple training sets that represent technical variation, as was recently successful for training feature extraction models more generally [32].

Our analysis indicates that CytoSummaryNet has learned a form of cell quality control, selectively strengthening profile accuracy. Although our results suggest that CytoSummaryNet is able to learn second-order and possibly third-order moments (as discussed in *Supplementary Material B*), substantiating this capability is complex and warrants dedicated research efforts. Looking forward, identifying and abstracting general aggregation principles from CytoSummaryNet - such as the exclusion of smaller cells prior to aggregation - could not only enhance average profiling techniques but also enrich our understanding of cellular heterogeneity. We encourage further exploration into CytoSummaryNet's architecture, anticipating its broad applicability in processing diverse single-cell tabular datasets, which span morphological as well as other omics data types, including single-cell RNA sequencing.

Conclusion

Our proposed CytoSummaryNet provides an easier-to-apply and better-performing method for aggregating single-cell image feature data than previously published strategies and the average profiling baseline. Based on an interpretability analysis, it is likely that CytoSummaryNet achieves this by performing some form of quality control by filtering out noisy cells (small mitotic cells or those with debris) and prioritizing less noisy cells (large uncrowded cells). Remarkably, CytoSummaryNet could also mitigate batch effects, even though this was not part of the training objective. This shows that the learned latent representation of CytoSummaryNet prioritizes biological signal over technical variance, both on the cell level and the plate level. CytoSummaryNet cannot effectively be directly transferred to unseen datasets; however, it can readily be re-trained on new data and infer the improved profiles directly after because the labels required for training are naturally available in cell profiling experiments. This method could help improve results in future cell profiling studies.

Acknowledgments

Funding for the project was provided by the National Institutes of Health NIGMS (R35 GM122547 to AEC) and an internship funded by the Massachusetts Life Sciences Center, Stichting dr. Hendrik Muller's Vaderlandsch Fonds, Stichting de Fundatie van de Vrijvrouwe van Renswoude te ’s-Gravenhage, and Fund International Experience/Holland Scholarship (to RvD). The authors appreciate helpful comments from Juan Caicedo, Alex Kalinin, Alán Muñoz González, Niranj Chandrasekaran, Srijit Seal, Ellen Su, Beth Cimini, and Erin Weisbart.

Declaration of interests

The Authors declare the following competing interests: S.S. and A.E.C. serve as scientific advisors for companies that use image-based profiling and Cell Painting (A.E.C: Recursion, SyzOnc, Quiver Bioscience; S.S.: Waypoint Bio, Dewpoint Therapeutics, Deepcell) and receive honoraria for occasional talks at pharmaceutical and biotechnology companies. R.v.D is an employee of CellVoyant. All other authors declare no competing interests.

Methods

Architecture

Our proposed model follows the general Deep Sets architecture [24]. The Deep Sets architecture can process permutation invariant data and learn to estimate first and second-order moments from the input data. Permutation invariance is necessary for aggregating sets of single-cell data into a sample profile, because the order in which the cells are processed should not influence the output. Estimating the first and second-order moments is especially important as the previous study by Rohban et al. has shown that this can improve sample profile strength [13]. An experiment was performed to show that this is still the case in a weakly supervised contrastive learning setting, as discussed in *Supplementary Material B*.

After segmentation, thousands of features are measured for each cell, *Figure 7a*. These feature vectors are aggregated using a function $f(\boldsymbol{X})$, where $\boldsymbol{X} \in \mathbb{R}^{\mathrm{MxD}}$ is the input set of single cell profiles, to get a single profile representing the cell population. There are multiple ways of defining the aggregation function $f(x)$. Our proposed architecture, *Figure 7b*, consists of two functions φ and ρ, which are simple fully connected neural networks, capable of approximating arbitrary polynomials. First, the model transforms the input set $\boldsymbol{X}$ by transforming each single-cell profile (row) $x_m \in \mathbb{R}^{\mathrm{D}}$ of the set using neural network $\varphi\colon \mathbb{R}^{\mathrm{D}} \rightarrow \mathbb{R}^{\mathrm{N}}$. φ consists of a single fully connected layer with 2048 nodes followed by a leaky ReLU activation layer. All of these nonlinear representations $\varphi(x_m)$ are summed, collapsing the cell dimension M: $\sum_{m=1}^{M} \varphi(x_m)$. The output $z \in \mathbb{R}^{\mathrm{N}}$ is then processed by the projection network $\rho\colon \mathbb{R}^{\mathrm{N}} \rightarrow \mathbb{R}^{\mathrm{L}}$, which applies more nonlinear transformations to create a final representation v in the loss space ($v = \rho(z)$). The function ρ consists of two subsequent fully connected layers of 512 and 2048 nodes respectively, both followed by leaky ReLU activations.

Loss function

The model is trained using contrastive learning by computing the Supervised Contrastive (SupCon) loss [33]. In order to train contrastive learning models, one has to define positive and negative sample pairs. We define a positive pair as two samples perturbed with the same compound and a negative pair as two samples perturbed with different compounds. The SupCon loss is different from the commonly used SimCLR [34] in only one aspect: it takes into account all positive pairs of a certain sample instead of only one at a time, *equation 1*. SupCon loss pulls positive pairs together in the embedding space, while simultaneously pushing apart negative pairs, *Figure 7c*. In this study, cosine similarity is used as the distance metric. The loss shows benefits for robustness to natural corruptions, hyperparameter settings, and inherently performs hard positive and hard negative mining when used in combination with cosine similarity [33].

$$L^{sup} = \sum_{i \leqq I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} log \frac{exp(\, v_i \bullet v_p \,/\, \tau\,)}{\sum_{a \in A(i)} exp(\, v_i \bullet v_a / \,\tau\,)} \tag{1}$$

I: total number of samples

τ: temperature constant (hyperparameter)

$P(i)$: all positive samples for the current sample i

$A(i)$: all negative samples for the current sample i

$v_i = \rho(\sum_{m=1}^{M} \varphi(x_m))$, where $x_m \in \boldsymbol{X}_i$

Model evaluation

After training the model, the projection network ρ is often discarded, and the summed representation z is used for downstream analysis instead [24]. However, ρ is not discarded here because the projection it has learned is tied to the evaluation task. One of the main applications of image-based cell profiling is discovering the unknown mechanism of action of a certain compound. To that end, in addition to replicate retrieval (the training task), the proposed model is evaluated using mechanism of action retrieval. Mechanism of action retrieval is evaluated by quantifying a profile's ability to retrieve the profile of other compounds with the same annotated mechanism of action. If the model has learned to amplify the phenotypic signature of a sample's profile, finding other compounds with the same mechanism of action should also become easier.

The performance in mechanism of action retrieval and replicate retrieval are compared between model-based profiling, which uses the learned aggregation, and average profiling. The performance is quantified by calculating the mean average precision (mAP). Average precision (AP) is an information retrieval metric that indicates whether the model can correctly identify all the positive examples without accidentally marking too many negative examples as positive. Specifically, the metric is equal to the area under the precision-recall curve. The calculation for the AP of a single query is shown in *equation 2*. To compute the AP, a rank order of sample profiles is required. The top of the rank order corresponds to profiles most similar to the queried profile and vice versa. This rank order is created by calculating the cosine similarity between all K sample profiles.

$$AP = \sum_{k=1}^{K} (r(k) - r(k - 1))p(k) \qquad (2)$$

K: total number of sample profiles
$p(k)$: precision at the k-th position of the rank order
$r(k)$: recall at the k-th position of the rank order

A compound’s AP is calculated by taking the average AP over all its replicate compound profiles, while a mechanism of action’s AP additionally averages over the same mechanism of action compound profiles. The mAP is the mean AP over all of the compounds. If a compound does not have other compounds with the same mechanism of action, the compound and its replicates are left out of the mAP computation for mechanism of action retrieval.

Interpretability analysis

Deep learning models are notoriously difficult to interpret. However, interrogating them has led to useful new insights in previous studies [35] and allows users to better understand the reasoning behind a model’s output. In this study, a combination of sensitivity analysis (SA) [36] and a method similar to critical point analysis (CPA) used in the PointNet study [37] was used to investigate possible biological foundations for the model’s output.

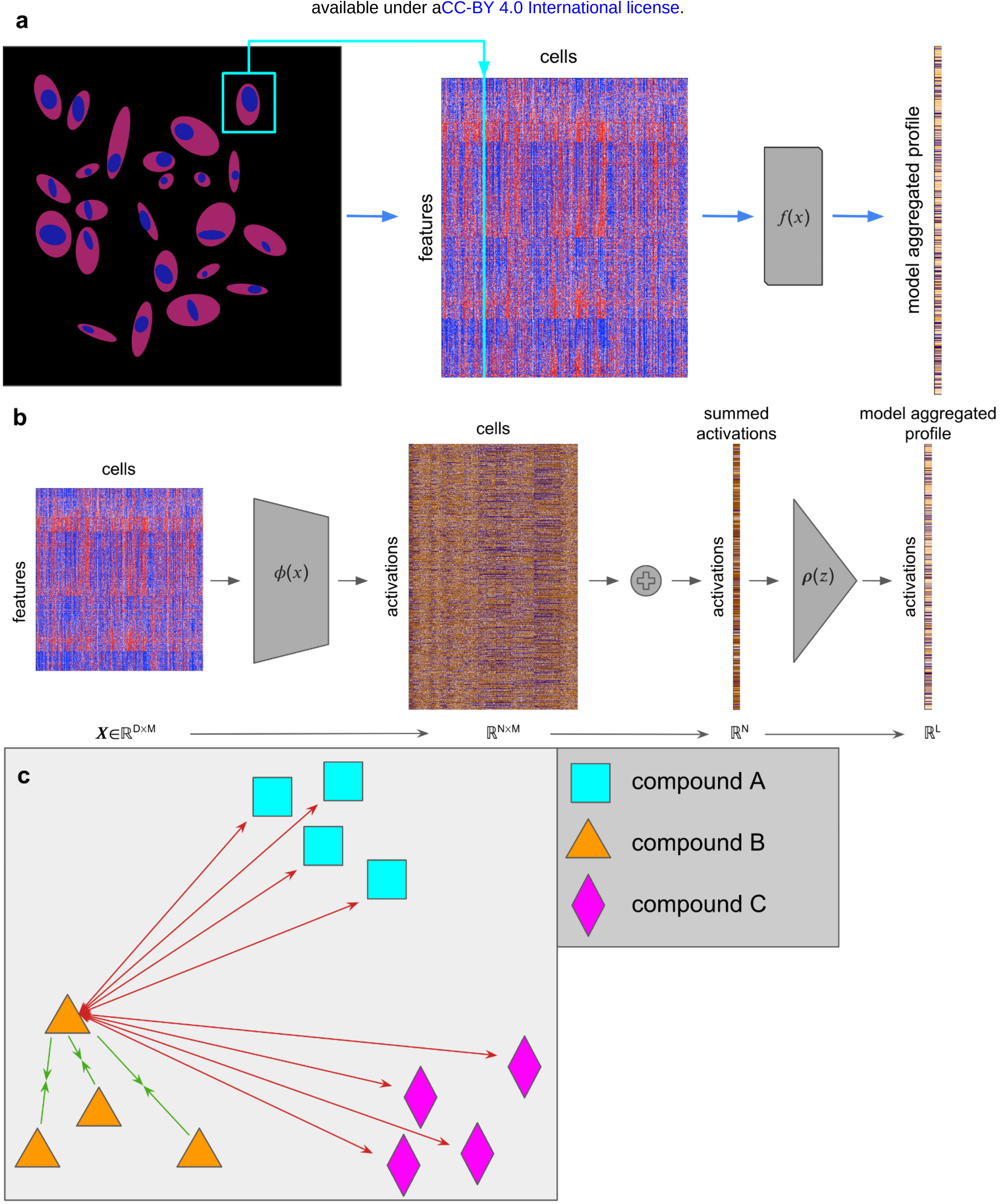

Figure 7: (a) Thousands of features are extracted from each segmented cell in microscopy images of wells. A learned function $f(x)$ aggregates this data into a single feature vector: the sample's profile. (b) An in-depth look at the model architecture used in this study. The model consists of three elements: a function $\varphi(x)$, which maps the input data from $\mathbb{R}^D$ to $\mathbb{R}^N$ space, a summation, which collapses the cell dimension, and $\rho(z)$, which maps the collapsed representation from $\mathbb{R}^N$ to $\mathbb{R}^L$ space. (c) During training, replicate compound profiles are forced to attract each other (green arrows) and simultaneously repel every other compound (red arrows) in the learned feature space. Here, all forces are drawn for a single profile of compound B.

SA explains a model's prediction based on locally evaluated partial derivatives. The partial derivatives of a model's output $f(\boldsymbol{X})$ are calculated with respect to each entry in the input matrix $x_{m,d}$ by backpropagating the loss function L^{sup}. Afterward, the absolute value is taken of these partial derivatives, *equation 3*. These values can then be summed over either the cell or feature dimensions to respectively get the feature or cell relevance scores. In this case, the relevance score per cell is computed. The analysis assumes that the most relevant input values are those to which the model's output is the most sensitive. Thus, inputs that receive a high relevance score will, when changed, make it more or less likely for the model to make a certain prediction. As a result, high relevance values can also characterize input patterns that the model would like to see removed to improve its performance for the predicted class. These patterns, e.g., noise, may not be linked to the class of interest.

$$R_{m,d} = ||\frac{\delta}{\delta x_{m,d}} f(\boldsymbol{X})|| \tag{3}$$

To counteract some of the potential noisy predictions of SA, CPA is used additionally for calculating the relevance scores. Since the model architecture is permutation invariant, each input cell vector is processed independently. In the PointNet study [37], CPA consisted of finding the input points with the maximum value for each feature. These points were found to form the skeleton of the input, meaning that they are the most relevant points for defining it. Their model's permutation invariant operation was a max pooling operation that inherently selected these points. In this study, a summation is used instead, although the reasoning is the same: cells with high activation values before the permutation invariant operation contribute more to the output of the model than those with low activation values. The CPA relevance score was calculated per cell by taking the L1-norm of their respective activations of the first fully connected layer.

The relevance scores of SA and CPA are min-max normalized per well and then combined by addition. The combination of the two is again min-max normalized, resulting in the combined relevance score. This score was used instead of the separate relevance scores because averaging was expected to cancel out some of the potentially noisy predictions. We show that the combined relevance score indeed achieves higher Pearson correlations with the CellProfiler features than the separate relevance scores, *Supplementary Material E*. Cells with the highest (>0.8) or lowest (<0.2) combined relevance scores are denoted as the 'most relevant' or 'least relevant' cells, respectively.

Multiple methods are used to investigate what the model has learned. First, the model and average-aggregated profiles of the top 15 same mechanism of action compound pairs, based on their mAP for mechanism of action retrieval, are visualized using a UMAP [28]. Four different values were tried for the UMAP hyperparameter $n\_neighbors$, which balances local versus global structure in the data. From these UMAPs, one was chosen based on the best presentation of the clusters that were consistently visible throughout all UMAPs. The second method calculates the Pearson correlation between all of the input CellProfiler features and the combined relevance scores to get a better understanding of what features the model prioritizes.

Finally, the most and least relevant cells are visualized and analyzed in the raw microscopy images to link the model's output to the underlying cell biology of the compound perturbations. Plates from multiple StainX subsets are used for these analyses to find commonalities between them.

Experimental setup

Data

We tested this method separately on the cpg0001 dataset [25], from the JUMP consortium [38] and batch 1 of the cpg0004 dataset [27]. Both datasets are available from the Cell Painting Gallery on the Registry of Open Data on AWS https://registry.opendata.aws/cellpainting-gallery/.

The 384-well plates that make up the cpg0001 dataset each contain 4 replicates of each of the 90 different compound perturbations. This dataset was created with the aim of optimizing the analysis pipeline for image-based cell profiling, resulting in a lot of technical variation. The analysis pipeline varied in dye concentration, cell permeabilization, cell seeding, exposure, pixel binning, compound dose, or microscopy method (confocal versus widefield). Each well was seeded with U2OS cells (a bone cancer cell line) in the solvent dimethyl sulfoxide (DMSO). Each well was either perturbed with one of 90 compounds or used as a negative control using only the solvent. The well position of each compound perturbation was fixed, i.e., the same plate layout was used across all plates. In total, 42 plates from four different optimization experiments called Stain2, Stain3, Stain4, and Stain5 were used. Unlike plates within an experiment, these were carried out at different times and therefore introduce additional technical variance, e.g., due to changes in laboratory conditions, sample manipulation, or instrument calibration.

The cpg0004 dataset (Batch 1) was created with a single analysis pipeline, however, it contains much more biological variation due to the 1.258 different compound perturbations used. These datasets used 384-well plates. The cells were stained using the Cell Painting protocol [39], after which images were taken with a microscope. The dataset used A549 cells (an adenocarcinomic cell line) in a DMSO solvent. The wells on each plate were perturbed with 56 different compounds in six different doses. Every compound was replicated 4 times per dose, with each replicate on a different plate. This requires the model to find replicates across plates during training instead of within plates like in the cpg0001 dataset. In this study, only the highest and second-highest dose points are used: 10 µM and 3.33 µM. 28 different plate layouts were used across all 136 plates.

Preprocessing

After the cells were segmented in each image, features were extracted using CellProfiler [29]. A subset of 1324 features was taken which were available in all plates of cpg0001, because CytoSummaryNet requires a fixed number of input features and to reduce the model's computational burden. Similarly, a subset of 1745 features was taken from the plates in cpg0004. The complete list of these feature names can be found in the GitHub repository of this project https://github.com/carpenter-singh-lab/2023_vanDijk_CytoSummaryNet. The features were standardized on the plate level to reduce batch effects, resulting in zero mean and unit variance features across all cells in the plate. The negative control wells were then removed from all plates because, by definition, these do not have a strong profile.

A commonly used pipeline was employed to calculate the average-aggregated profiles. Only the selected subset of features was used to allow for a fair comparison with the model-aggregated profiles. After calculating the average-aggregated profile for each well in a certain plate, the features were RobustMAD normalized by subtracting their median and dividing by their mean absolute deviation. The final average profiles were acquired by applying feature selection using a variance and correlation threshold. The full data processing workflows are available at https://github.com/jump-cellpainting/pilot-data-public (for cpg0001) and https://github.com/broadinstitute/lincs-cell-painting (for cpg0004). For the model, the outputs were used directly as the aggregated profiles.

Stratification

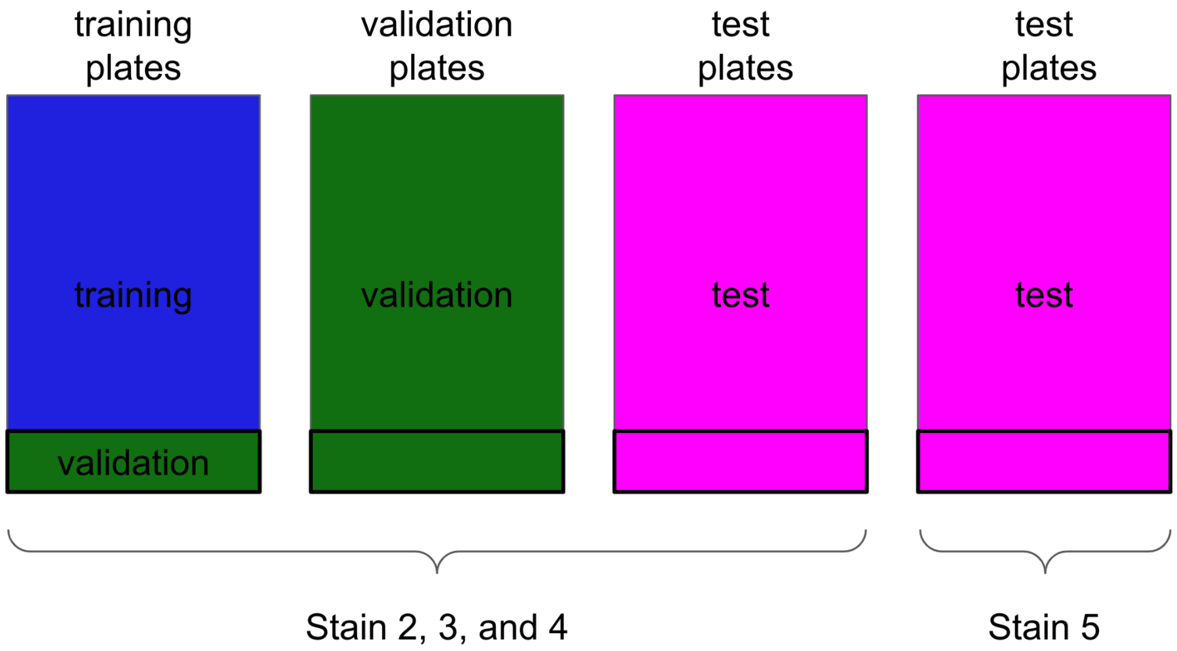

Figure 8: Stratification of the cpg0001 dataset per plate. All training plates are split into training and validation compounds. This compound stratification is perpetuated throughout all plates as indicated by the boxes. Stain2, Stain3, and Stain4 all contain training, validation, and test plates. Stain5 only contains test plates.

The cpg0001 dataset was stratified based on two variables: plate and compound (see *Figure 8*). First, the 90 compounds are split into 72 (80%) training compounds and 18 (20%) validation compounds. Due to the limited number of available compounds in this dataset, we did not define a held-out test set of compounds. Second, except for the Stain5 dataset, all available plates were stratified per dataset, resulting in 5 training, 4 validation, and 3 test plates for Stain2, Stain3, and Stain4.

Six plates from Stain5 were used as an out-of-distribution test set with different experimental conditions than the training data. *Supplementary Material C* shows the plate names per dataset for this stratification. All wells corresponding to training compounds in the training plates are used to train the model weights. Validation compounds from both the training and validation plates are used during the validation epochs to select the best model. None of the test plate data is ever seen during training or validation epochs. These stratifications allowed for isolated evaluation of the model's ability to generalize to unseen compound distributions and unseen plate distributions, i.e., technical variation.

The cpg0004 dataset was stratified in a different way. The 10 μM dose point data was used as the training and validation set and the 3.33 μM data as the test set. In the end, the proposed model is supposed to be trained using the compound replicates and then used to infer the improved profiles for mechanism of action retrieval directly afterward. This stratification simulates that use case and additionally evaluates the model's performance on unseen but similar data, i.e., the 3.33 μM data.

In cpg0001, each StainX subset corresponds to a specific set of assay conditions that were designed to optimize Cell Painting. Further, even within two of the StainX subsets – Stain2 and Stain3 – each plate had slightly different variations of the assay conditions. These differences lead to strong batch effects. While our primary objective in this study is not to address the batch or experiment effect problem directly, we intentionally selected test set plates to ensure they represent the most divergent conditions within the Stain2, Stain3, and Stain4 subsets, maximizing their out-of-distribution nature. Further, the Stain5 plates are inherently out-of-distribution relative to the other StainX subsets. This deliberate choice allows us to comprehensively assess the model's performance, including its limitations when faced with out-of-distribution data.The test plates for Stain2, Stain3, and Stain4 were chosen based on how dissimilar they are to the training and validation data. *Supplementary Material F* describes how this similarity is measured.

Data augmentation

Data augmentation was used to increase the generalization of the model to unseen samples. During training, for each batch, a number of cells was randomly sampled with replacement from every well. The number of cells that was sampled was itself a number sampled from a Gaussian distribution. Although they remained mostly similar, this created unique sets of cells for each compound in every batch. The number of sets of cells that was sampled for each compound in each epoch is a tunable hyperparameter. Each sampled set could consist of cells from a single well or a combination of two at random. This decision was based on a coin flip, which was performed for each sample. This last form of augmentation should help decrease plate-layout effects, which introduce a technical bias into the single-cell feature data based on which well position the population is in.

The model training process consists of many tunable hyperparameters, which were chosen using a random search. The AdamW optimizer [40] is used with a learning rate of 5×10^{-4} and weight decay 10^{-2}. The model is trained for 100 epochs or until the best mAP is achieved on the validation compounds of the training and validation plates. An overview of all the tunable parameters, including the data augmentation parameters, is given in *Supplementary Material D*.

References

[1] S. N. Chandrasekaran, H. Ceulemans, J. D. Boyd, and A. E. Carpenter, “Image-based profiling for drug discovery: due for a machine-learning upgrade?,” *Nat. Rev. Drug Discov.*, vol. 20, no. 2, pp. 145–159, Feb. 2021.

[2] J. Simm *et al.*, “Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery,” *Cell Chem Biol*, vol. 25, no. 5, pp. 611–618.e3, May 2018.

[3] N. Moshkov *et al.*, “Predicting compound activity from phenotypic profiles and chemical structures,” *Nat. Commun.*, vol. 14, no. 1, p. 1967, Apr. 2023.

[4] S. Seal, J. Carreras-Puigvert, M.-A. Trapotsi, H. Yang, O. Spjuth, and A. Bender, “Integrating cell morphology with gene expression and chemical structure to aid mitochondrial toxicity detection,” *Commun Biol*, vol. 5, no. 1, p. 858, Aug. 2022.

[5] M. Doan *et al.*, “Label-Free Leukemia Monitoring by Computer Vision,” *Cytometry A*, vol. 97, no. 4, pp. 407–414, Apr. 2020.

[6] J. C. Caicedo *et al.*, “Cell Painting predicts impact of lung cancer variants,” *Mol. Biol. Cell*, vol. 33, no. 6, p. ar49, May 2022.

[7] S. J. Altschuler and L. F. Wu, “Cellular heterogeneity: do differences make a difference?,” *Cell*, vol. 141, no. 4, pp. 559–563, May 2010.

[8] K. A. Janes, “Single-cell states versus single-cell atlases - two classes of heterogeneity that differ in meaning and method,” *Curr. Opin. Biotechnol.*, vol. 39, pp. 120–125, Jun. 2016.

[9] A. Marusyk and K. Polyak, “Tumor heterogeneity: causes and consequences,” *Biochim. Biophys. Acta*, vol. 1805, no. 1, pp. 105–117, Jan. 2010.

[10] D. Deb *et al.*, “Combination Therapy Targeting BCL6 and Phospho-STAT3 Defeats Intratumor Heterogeneity in a Subset of Non-Small Cell Lung Cancers,” *Cancer Res.*, vol. 77, no. 11, pp. 3070–3081, Jun. 2017.

[11] L. Keller and K. Pantel, “Unravelling tumour heterogeneity by single-cell profiling of circulating tumour cells,” *Nat. Rev. Cancer*, vol. 19, no. 10, pp. 553–567, Oct. 2019.

[12] J. Goveia *et al.*, “An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates,” *Cancer Cell*, vol. 37, no. 1, pp. 21–36.e13, Jan. 2020.

[13] M. H. Rohban, H. S. Abbasi, S. Singh, and A. E. Carpenter, “Capturing single-cell heterogeneity via data fusion improves image-based profiling,” *Nat. Commun.*, vol. 10, no. 1, p. 2082, May 2019.

[14] J. C. Caicedo *et al.*, “Data-analysis strategies for image-based cell profiling,” *Nat. Methods*, vol. 14, no. 9, pp. 849–863, Aug. 2017.

[15] C. Trapnell, “Defining cell types and states with single-cell genomics,” *Genome Res.*, vol. 25, no. 10, pp. 1491–1498, Oct. 2015.

[16] V. Ljosa *et al.*, “Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment,” *J. Biomol. Screen.*, vol. 18, no. 10, pp. 1321–1329, Dec. 2013.

[17] A. Janosch, C. Kaffka, and M. Bickle, “Unbiased Phenotype Detection Using Negative Controls,” *SLAS Discov*, vol. 24, no. 3, pp. 234–241, Mar. 2019.

[18] L.-H. Loo, H.-J. Lin, R. J. Steininger 3rd, Y. Wang, L. F. Wu, and S. J. Altschuler, “An approach for extensibly profiling the molecular states of cellular subpopulations,” *Nat. Methods*, vol. 6, no. 10, pp. 759–765, Oct. 2009.

[19] F. Fuchs *et al.*, “Clustering phenotype populations by genome-wide RNAi and multiparametric imaging,” *Mol. Syst. Biol.*, vol. 6, p. 370, Jun. 2010.

[20] P. H. Le-Khac, G. Healy, and A. F. Smeaton, “Contrastive Representation Learning: A Framework and Review,” *IEEE Access*, vol. 8, pp. 193907–193934, 2020.

[21] Maron and Lozano-Pérez, “A Framework for Multiple-Instance Learning,” *Adv. Neural Inf. Process. Syst.*, Jun. 1997, [Online]. Available: https://proceedings.neurips.cc/paper/1997/file/82965d4ed8150294d4330ace00821d77-Paper.pdf

[22] H. Edwards and A. Storkey, “Towards a Neural Statistician,” *arXiv [stat.ML]*, Jun. 07, 2016. [Online]. Available: http://arxiv.org/abs/1606.02185

[23] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space,” *arXiv [cs.CV]*, Jun. 07, 2017. [Online]. Available: http://arxiv.org/abs/1706.02413

[24] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola, “Deep Sets,” *arXiv [cs.LG]*, Mar. 10, 2017. [Online]. Available: http://arxiv.org/abs/1703.06114

[25] S. N. Chandrasekaran *et al.*, “Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations,” *bioRxiv*, p. 2022.01.05.475090, Jan. 05, 2022. doi: 10.1101/2022.01.05.475090.

[26] B. A. Cimini *et al.*, “Optimizing the Cell Painting assay for image-based profiling,” *bioRxiv*, p. 2022.07.13.499171, Jul. 13, 2022. doi: 10.1101/2022.07.13.499171.

[27] G. P. Way *et al.*, “Morphology and gene expression profiling provide complementary information for mapping cell state,” *Cell Syst*, vol. 13, no. 11, pp. 911–923.e9, Nov. 2022.

[28] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” *arXiv [stat.ML]*, Feb. 09, 2018. [Online]. Available: http://arxiv.org/abs/1802.03426

[29] D. R. Stirling, M. J. Swain-Bowden, A. M. Lucas, A. E. Carpenter, B. A. Cimini, and A. Goodman, “CellProfiler 4: improvements in speed, utility and usability,” *BMC Bioinformatics*, vol. 22, no. 1, p. 433, Sep. 2021.

[30] F. Heigwer *et al.*, “A global genetic interaction network by single-cell imaging and machine learning,” *Cell Syst*, vol. 14, no. 5, pp. 346–362.e6, May 2023.

[31] M. Tegtmeyer *et al.*, “High-dimensional phenotyping to define the genetic basis of cellular morphology,” *bioRxiv*, p. 2023.01.09.522731, Jan. 09, 2023. doi: 10.1101/2023.01.09.522731.

[32] N. Moshkov *et al.*, “Learning representations for image-based profiling of perturbations,” *bioRxiv*, p. 2022.08.12.503783, Aug. 15, 2022. doi: 10.1101/2022.08.12.503783.

[33] P. Khosla *et al.*, “Supervised Contrastive Learning,” *arXiv [cs.LG]*, Apr. 23, 2020. [Online]. Available: http://arxiv.org/abs/2004.11362

[34] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” in *Proceedings of the 37th International Conference on Machine Learning*, H. D. Iii and A. Singh, Eds., in Proceedings of Machine Learning Research, vol. 119. PMLR, 13--18 Jul 2020, pp. 1597–1607.

[35] S. Chakraborty *et al.*, “Interpretability of deep learning models: A survey of results,” in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2017, pp. 1–6.

[36] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models,” *arXiv [cs.AI]*, Aug. 28, 2017. [Online]. Available: http://arxiv.org/abs/1708.08296

[37] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation,” *arXiv [cs.CV]*, Dec. 02, 2016. [Online]. Available: http://arxiv.org/abs/1612.00593

[38] A. Mullard, “Machine learning brings cell imaging promises into focus,” *Nat. Rev. Drug Discov.*, vol. 18, no. 9, pp. 653–655, Sep. 2019.

[39] M.-A. Bray *et al.*, “Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes,” *Nat. Protoc.*, vol. 11, no. 9, pp. 1757–1774, Sep. 2016.

[40] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” Sep. 27, 2018. Accessed: Jun. 20, 2022. [Online]. Available: https://openreview.net/pdf?id=Bkg6RiCqY7

[41] R. Vemulapalli and D. W. Jacobs, “Riemannian Metric Learning for Symmetric Positive Definite Matrices,” *arXiv [cs.CV]*, Jan. 10, 2015. [Online]. Available: http://arxiv.org/abs/1501.02393