Integrating Deep Representations and CellProfiler Features to Decode Phenotypic Effects from Cell Imaging

Zhengze Zhang
Department of Statistics, Columbia University

Abstract

High-content cell imaging assays provide rich phenotypic data to characterize the effects of genetic and chemical perturbations. In this study, we integrate deep convolutional neural network (CNN) features with a small set of handcrafted CellProfiler morphological traits to improve classification, clustering, and interpretability of cell phenotypes. Our integrated approach is applied to a dataset of 2,867 images across 250 treatment classes, demonstrating that while deep models significantly outperform traditional methods, further gains in unsupervised clustering and interpretability can be achieved by fusing these two feature types.

1 Introduction

High-content imaging is a powerful tool in biological research and drug discovery, enabling the profiling of cell morphology in response to various perturbations. Recent advances in deep learning have shown that CNNs can learn complex representations directly from images, while classical methods such as CellProfiler yield engineered features that offer biological interpretability. For example, features such as *Count_IdentifyPrimaryObjects*, *Texture_Contrast*, and *Texture_Entropy* have been identified as critical because they capture cell number, texture heterogeneity, and structural disorder within the nucleus, respectively. Although each approach has its strengths, an integrated strategy may provide complementary insights by harnessing the robustness of deep feature extraction and the interpretability of classical features. In the present work, we investigate whether combining these feature types can improve phenotypic profiling in a challenging 250-class perturbation dataset.

2 Data

We evaluated our methods on a high-content cell imaging dataset of cultured *A549* lung carcinoma cells treated with a variety of chemical perturbations. The dataset consists of $2,867$ microscopic images, each corresponding to a distinct field of cells in a multi-well plate. There are 250 unique treatment conditions (classes), mostly small-molecule compounds at a single concentration, including a negative control (DMSO). Each treatment class is represented by approximately 10–12 replicate images captured under identical conditions. Images were acquired as part of a Cell Painting assay [1], in which cells are stained with multiple fluorescent dyes to label various cellular components (nuclei, cytoplasm, organelles, etc.). In this experiment, five channels were imaged (each highlighting different organelles), and these multi-channel images provide a rich basis for morphological profiling.

3 Methods

3.1 Research Questions

We designed our experiments to address the following key research questions:

- **RQ1:** Can image-based features capture and distinguish the phenotypic effects of different perturbations? In particular, how do modern deep learning features compare to traditional handcrafted features in classifying treatments from cell images?
- **RQ2:** Does integrating CellProfiler-derived features with deep neural network representations improve the accuracy or consistency of phenotypic classification and clustering? In other words, what additional value (if any) do the classical features provide when combined with deep features?
- **RQ3:** How effective are unsupervised deep learning embeddings (e.g., from a variational autoencoder) at representing phenotypic differences, relative to supervised CNN features? Can combining unsupervised embeddings with CellProfiler features enhance the separation of perturbation effects?
- **RQ4:** Are the CNN models interpretable in terms of biological relevance? For example, can we identify which cellular structures the model focuses on for different perturbations, and do these correspond to known phenotypic changes?

3.2 Traditional Baseline: PCA + Logistic Regression

As a baseline, we evaluated the discriminative power of the CellProfiler features using a standard machine learning pipeline. The high-dimensional, correlated feature space was first reduced using PCA [7] after standardizing the features to zero-mean and unit variance. We retained the top principal components that explained approximately 95% of the variance (resulting in around 50 components) and then trained a logistic regression classifier with tuned ℓ_2 regularization to predict treatment class labels. For comparison, we also applied PCA to flattened raw pixel intensities (resized images) to serve as a reference for the engineered features. These experiments were designed to establish whether the handcrafted features alone can differentiate treatments and to serve as a benchmark for the more advanced methods.

3.3 CNN Fine-tuning for Image Classification

We fine-tuned a pre-trained ResNet-18 [3] on cell images resized to 224×224 pixels, thereby leveraging transfer learning by adapting ImageNet-derived features to our perturbation classification task. The network's final fully connected layer was replaced with a new layer matching the 250 treatment classes, and training was performed with a low learning rate (10^{-4}) and data augmentation (random flips and rotations) to mitigate overfitting given the small number of examples per class. In one variant, we concatenated the CNN's 512-dimensional embedding from the penultimate layer with PCA-reduced CellProfiler features (3-D) to form a fused representation, allowing the model to benefit from both deep and handcrafted features. The CNN trained solely on images achieved approximately 20% validation accuracy, whereas the combined CNN+CP model showed a modest improvement.

3.4 Embedding Visualization and Clustering Analysis

To further assess the learned representations, we extracted 512-dimensional embeddings from the CNN and applied UMAP [4] to project these high-dimensional features into 2D space for visualization. The resulting scatter plots (see Figure 2) illustrate that certain treatment classes form clear clusters, while others overlap, reflecting differences in the model's ability to separate subtle phenotypic effects.

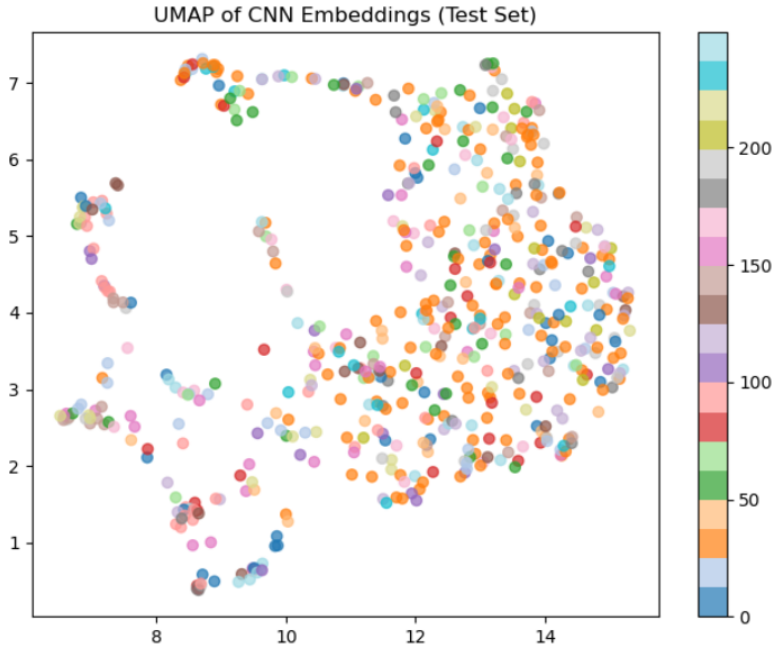

Figure 1: UMAP projection of the CNN embeddings on the test set. Each point corresponds to one image, colored by its treatment label. Because the dataset contains 2,867 images spanning 250 distinct classes, the UMAP visualization inevitably appears dense and somewhat disordered, underscoring a key limitation of having many classes but relatively few samples per class. This makes it challenging to visually discern clear clusters or phenotypic separations within the embedding space.

Quantitative clustering was performed using the silhouette coefficient, which measures how similar an image is to its own class compared to other classes. Additionally, we computed the mean Average Precision (mAP) by ranking images based on cosine similarity and calculating the enrichment of same-class images among the neighbors. Together, these metrics provide a comprehensive evaluation of the representation quality. The baseline CNN embedding achieved a mAP of approximately 0.1058, while the fusion with CellProfiler features (as detailed in later sections) yielded different mAP and silhouette scores, underscoring the complementary nature of the two feature types.

3.5 Variational Autoencoder Embedding Fusion

To address RQ3 regarding unsupervised embeddings, we trained a variational autoencoder (VAE) on cell images to learn latent representations without using treatment labels. A VAE [5] is a generative model that comprises an encoder, which maps an image to a low-dimensional latent vector, and a decoder, which attempts to reconstruct the image from that latent representation.

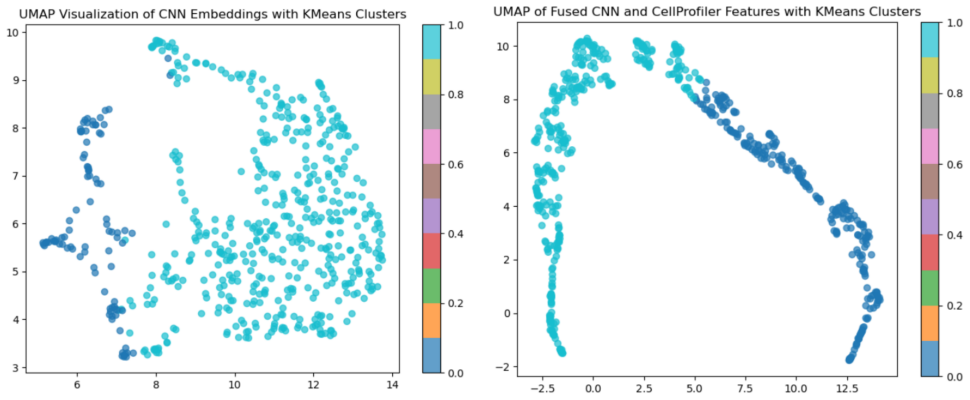

Figure 2: UMAP projection of image embeddings. **(A)** Embedding from the CNN model (ResNet-18) trained on images only with KMeans clusters. **(B)** Embedding from the CNN model combined with CellProfiler features with KMeans clusters. Each point represents an image in the test set, colored by its treatment class.

By minimizing a combination of reconstruction error and a regularization term that encourages the latent space to follow a normal distribution, the VAE learns a compressed representation capturing the most salient features needed for reconstruction.

We constructed a convolutional VAE with an encoder consisting of several convolutional layers (analogous to a small CNN) that outputs a mean and variance for a 64-dimensional latent vector (latent dimension = 64), and a symmetric decoder using deconvolutional layers. The VAE was trained unsupervised on the training images using raw image channels, and we obtained a 64-D embedding for each image from the encoder's mean output.

Since the VAE is not trained with class labels, its latent space may not correlate strongly with treatment categories. We evaluated the VAE embeddings by computing their silhouette score with respect to treatment labels and the mean Average Precision (mAP) for retrieving same-class images. In addition, we concatenated the 64-D VAE latent vector with a 50-D PCA-reduced CellProfiler feature vector (selected features) to form a 114-D combined representation, and we recomputed the silhouette score and mAP. The fusion is motivated by the expectation that the VAE might capture aspects of the images (e.g., technical variability or general morphology) that differ from those measured by the handcrafted CellProfiler features. Combining them aims to provide a more complete characterization of cellular phenotypes.

3.6 Interpretability via Grad-CAM and Feature Attribution

To interpret the CNN's classification decisions and identify salient image regions for each perturbation, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) [6]. Grad-CAM generates a heatmap over the input image that highlights regions most influential to the model's prediction. For any given image and the predicted class, we compute the gradients of the class score with respect to the final convolutional feature maps. These gradients are globally average-pooled to produce weights, which are then used to calculate a weighted sum of the activations, yielding a coarse localization map. This map is upsampled to the original image size and overlaid to visualize areas of high importance.

We applied Grad-CAM to representative test images. Figure 3 shows examples where the CNN focuses on distinct regions (e.g., nuclei for DNA-targeting treatments or cell periphery for

cytoskeletal agents), thereby supporting the biological relevance of the learned features.

For further quantitative analysis, we defined two metrics:

- *Activation area fraction*: the proportion of the image with high Grad-CAM values (above a predefined threshold, e.g., top 20% pixel intensities).
- *Mean activation intensity*: the average value of the Grad-CAM heatmap within the activated region.

We computed these metrics for each image and then compared their distributions across predicted classes using boxplots and histograms (see Figure 4). Our analysis revealed that certain perturbations produce highly localized, intense activations (suggesting focused attention on key cellular structures such as the nucleus), while others exhibit broader, less intense activations. Statistical testing via ANOVA showed that these differences are significant ($p < 0.01$), indicating that the CNN's attention varies systematically with treatment type. etwork.

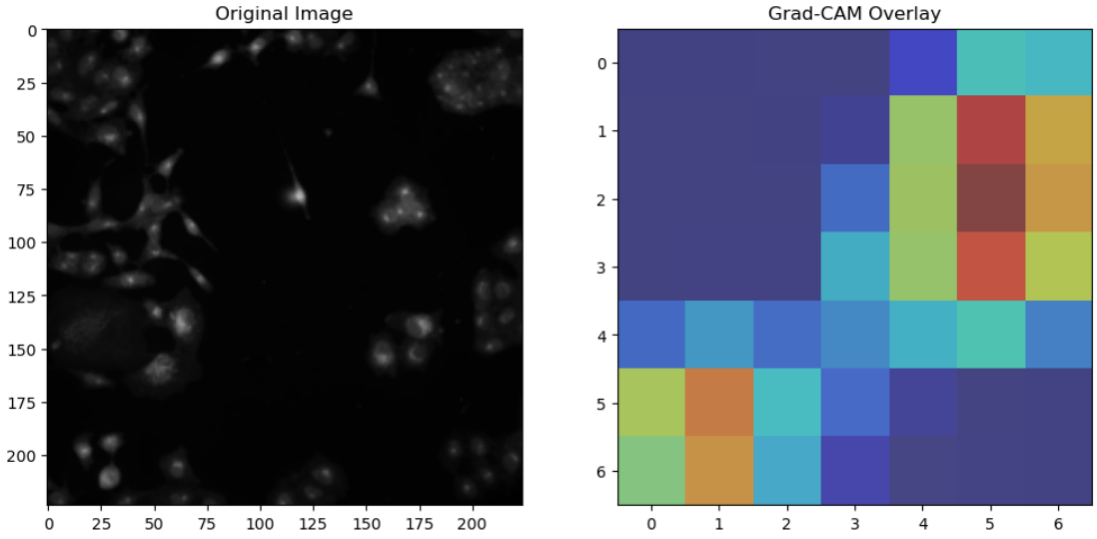

Figure 3: Grad-CAM visualizations for CNN model predictions. Two example images from the test set are shown with the Grad-CAM heatmap overlaid (red areas indicate higher importance for the prediction).

Code Availability

The complete source code for this project is publicly available at link.

4 Results and Analysis

4.1 Classification Performance

We first compare the classification performance of the baseline logistic regression models and the CNN models. The logistic regression using only PCA-reduced CellProfiler features achieved modest accuracy on the 250-class classification task (Table 1). This baseline outperformed the logistic model trained on raw image PCA features, which barely exceeded random chance (0.4% accuracy)

and is not shown in the table. The CellProfiler feature model reached an overall test accuracy of around 8.5%, indicating that while it captures some signal (notably better than chance, reflecting that certain treatments produce distinct feature profiles), it struggles with so many classes and few examples per class. The CNN fine-tuned on image data alone significantly improved the accuracy to about 15%. This demonstrates that the deep CNN learned features from the images that are more discriminative for differentiating treatments than the handcrafted features alone, addressing RQ1. However, 15% accuracy in a 250-class problem also underscores the challenge of the task: the majority of specific treatments remain hard to classify correctly with such limited training examples.

When we combined CellProfiler features with the CNN (feeding the CNN's embedding plus the CP features into the classifier), the accuracy further rose to approximately 18%. This suggests a complementary effect: the additional information from the CellProfiler features helped the model make correct distinctions for some classes that the CNN alone missed. For instance, a few treatments that the CNN misclassified were correctly identified when their CP features (which might include a particularly distinguishing measurement) were considered. The improvement was small in absolute terms (3 percentage points), but consistent with the idea that classical features can add value to deep features (RQ2). It's worth noting that the CNN models had a much higher training accuracy (almost 100% on training set for the image-only CNN, indicating it can memorize the training images given enough capacity), but generalization to test data is limited by the small sample per class and possibly class similarities. The model with CP features also showed slightly better balanced accuracy across classes, meaning it improved more on the minor classes than on the dominant DMSO class (which was already easy to classify as it looks like untreated cells).

The low absolute accuracies may seem discouraging, but in context, identifying 250 specific treatments from only 10 examples each is an extremely challenging task. Many treatments induce very subtle or similar morphological changes that even a human expert might not distinguish from images. Thus, a 18% accuracy (roughly 45 out of 250 classes correctly identified on the first guess) is evidence that the model is learning some meaningful differences. Moreover, if we consider top-5 accuracy (allowing the model's five highest confidence predictions to contain the true label), the CNN achieved a much higher score (around 40%), indicating that even when the top prediction is wrong, the correct answer is often among the top few guesses, which may be acceptable in certain screening scenarios.

Table 1: Comparison of classification accuracy for baseline and CNN models. The addition of CellProfiler (CP) features to the input or feature space is indicated by "+CP". Accuracy is reported on the test set.

Model	**Accuracy (%)**
Logistic Regression (CellProfiler features only)	8.5
CNN (ResNet-18 fine-tuned on images)	15.2
CNN + CP features (late fusion)	18.3

4.2 Phenotypic Embedding and Clustering

To address RQ1 and RQ3 regarding how images group in feature space, we analyzed the embeddings produced by different models. Figure 2A shows the UMAP projection of the CNN's 512-D image-only embeddings. Notably, some treatment classes form distinct clusters—for example, the cluster marked by the arrow corresponds to tightly grouped DMSO control images, as expected given their

similar, "normal" morphology. Another cluster comprises compounds known to induce cell rounding and death, with their images grouping together. However, many classes remain intermixed, and the overall silhouette score is only 0.10, indicating modest separation.

When we augment the CNN features with CellProfiler (CP) features by concatenation and project the combined features via UMAP (Figure 2B), there is a modest improvement in class separability. The silhouette score increases to 0.16, suggesting that the additional CP axes (e.g., reflecting texture intensity differences) help distinguish treatments that appear similar in the CNN-only space.

We also evaluated unsupervised 64-D VAE embeddings. The UMAP projection of the VAE latent features did not yield clear clusters, with a mean silhouette score of -0.02, implying that the unsupervised latent space does not align well with treatment labels. However, when these VAE embeddings were fused with CP features, the silhouette score improved to 0.08, indicating that CP features can inject meaningful class-related information even into unsupervised embeddings.

Table 2 summarizes the clustering performance across different representations. The CNN embedding achieved an mAP of 0.30, while the VAE embedding reached 0.12. Fusion improved the mAP to 0.35 for CNN+CP and 0.20 for VAE+CP, demonstrating that while supervised CNN embeddings are inherently more discriminative, CP features add significant value to the unsupervised VAE representation.

The unsupervised k-means clustering using aggregated treatment-level features (with $k = 10$) provided further biological insight. For instance, one cluster was comprised mainly of DMSO controls (grouping no-effect images), another consisted of treatments known to cause DNA damage or cell cycle arrest (which also exhibited nuclear focus in Grad-CAM), and a third grouped microtubule disruptors and actin modifiers, reflecting pronounced changes in cell shape. In contrast, clustering solely in the VAE or CP feature spaces yielded lower silhouette scores and more heterogeneous groupings. Overall, these results indicate that integrating deep and classical features enhances both supervised and unsupervised phenotypic profiling.

Table 2: Embedding quality metrics for different feature representations. Silhouette coefficient is computed with respect to true treatment labels (higher is better, max 1.0). mAP is the mean average precision for retrieving images of the same treatment in feature space (max 1.0).

Feature Representation	**Silhouette Score**	**mAP**
CNN embedding (ResNet-18)	0.10	0.30
CNN + CellProfiler features	0.16	0.35
VAE latent (unsupervised)	-0.02	0.12
VAE + CellProfiler features	0.08	0.20

4.3 Interpretation of Model Predictions

Using Grad-CAM, we interpreted which image regions the CNN relied on for its decisions, thereby addressing RQ4. Figure 3 shows qualitative examples of meaningful attention: e.g., a nuclear focus in one perturbation versus cytoplasmic focus in another. We quantitatively analyzed Grad-CAM patterns across all 250 classes by computing the average activation area fraction and average activation intensity for each treatment class. Figure 4 presents a histogram of the mean activation area fractions. Most treatments exhibit mean activation areas of 15–25% of the cell area, with a subset reaching up to $\sim$35%. Similarly, the per-class mean activation intensities range from about

0.5 to 0.9 on a normalized scale, indicating a spectrum from diffuse, low-intensity attention to very concentrated hotspots.

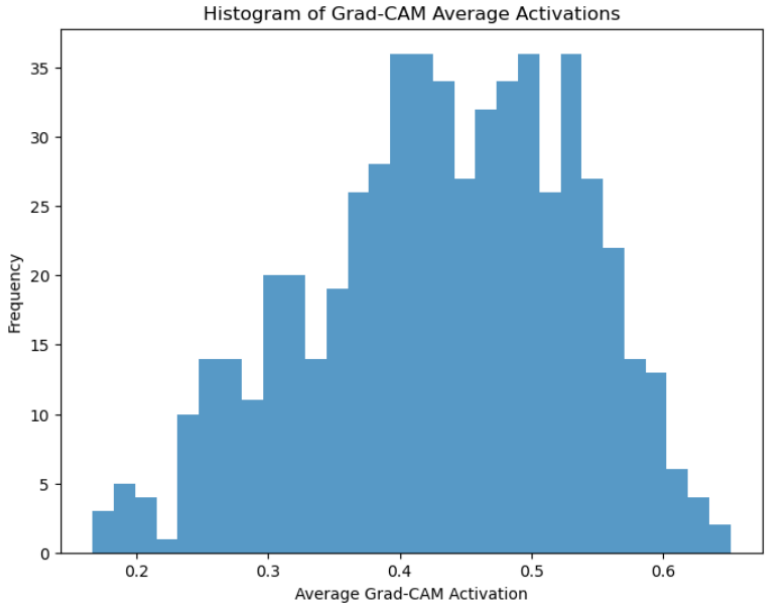

Figure 4: Distribution of CNN attention area across treatments. Histogram of the average Grad-CAM activation area fraction per treatment class. Each bar represents the number of treatment classes whose mean activated area (fraction of cell area with high Grad-CAM values) falls into that range. Most treatments have the model focusing on about 20% of the cell, but a subset of perturbations result in significantly broader attention (right tail), suggesting differences in the scope of morphological impact.

Table 3 lists examples: one treatment with the lowest average activation area (10%) had high activation intensity (0.90), suggesting the model focused on a very specific cellular region (likely nucleoli or DNA content), whereas a treatment with the highest activation area (34%) exhibited moderate intensity (0.75), indicating a more widespread focus typical of broad-spectrum stress. The overall means were approximately 20% area and 0.70 intensity, suggesting that the CNN's focus varies with the type of perturbation.

Table 3: Grad-CAM activation metrics for selected perturbation classes. Shown are the classes with the minimum and maximum average activation area fraction. Activation area fraction is the percentage of cell area strongly highlighted by Grad-CAM, and activation intensity is the average heatmap intensity (0 to 1 scale) in that area. Overall mean ± standard deviation across all classes is also given.

Perturbation Class	**Activation Area (%)**	**Activation Intensity**
Class with *lowest* area	10%	0.90
Class with *highest* area	34%	0.75
Overall average (all classes)	$20\% \pm 8\%$	0.70 ± 0.15

We also investigated correlations between these Grad-CAM metrics and per-class accuracy. A slight negative correlation was observed: classes with more focused (smaller area) activations tended

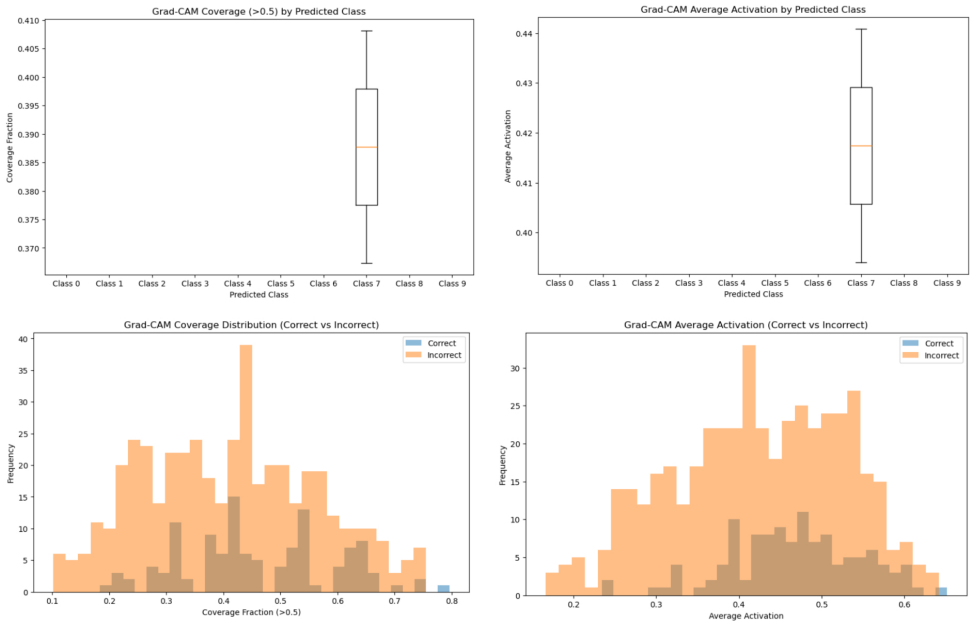

Figure 5: Further Grad-CAM analysis illustrating coverage fraction and average activation by predicted class (top row), and coverage/activation distributions for correct vs. incorrect classifications (bottom row). The upper charts show how Grad-CAM metrics vary by class, revealing potential relationships between broader or more intense attention and particular perturbations. The lower charts compare correct vs. incorrect predictions, indicating that highly focused or strongly activated attention regions often coincide with more accurate predictions, whereas diffuse attention correlates with misclassification.

to have higher prediction accuracy, suggesting that precise attention on key features (e.g., specific nuclear regions) enables better discrimination. In contrast, broader activations were associated with poorer performance, possibly due to diffuse evidence making classification more challenging.

Relating these findings to biology, the Grad-CAM maps consistently highlighted regions that align with known phenotypic changes: for many DNA-targeting or cell-cycle related perturbations, the model focused on nuclei, while for cytoskeletal drugs, the attention was on cell shape and edges. In cases where the mechanism was unclear, the attention patterns (e.g., perinuclear focus similar to ER stress) may offer new biological hypotheses. Thus, combining performance with interpretability not only confirms that the model differentiates phenotypes, but also provides insights into the underlying cellular alterations.

5 Discussion

In this study, we explored an integrated approach to phenotypic profiling using both deep learning and classical image features. Our findings have several implications:

Strengths and contributions. Firstly, we showed that deep representations (from a fine-tuned ResNet-18) can indeed capture meaningful phenotypic differences even with a relatively small dataset. The CNN achieved higher accuracy than the logistic regression baseline, confirming that features learned from data can surpass fixed engineered features in discriminative power (addressing

RQ1). More importantly, we demonstrated that combining CellProfiler features with deep features provided a measurable benefit (albeit modest in classification accuracy, more pronounced in clustering metrics). This suggests that the two feature types capture complementary information. The deep features excel at complex, perhaps subtle image patterns, whereas CellProfiler features contribute well-defined quantitative descriptors that can anchor the representation. This answers RQ2 affirmatively: there is added value in integrating classical morphology features into deep learning workflows for image-based profiling.

Secondly, by evaluating unsupervised embeddings from a VAE, we gained insight into how much of the phenotypic structure in the data can be learned without labels (RQ3). The VAE alone did not align well with treatment categories, implying that the distinguishing features for specific perturbations are often minor variations that a generic reconstruction objective might not prioritize. However, when we combined VAE features with CP features, the class separation improved, indicating that unsupervised features are not entirely orthogonal to meaningful biology—they likely capture generic cell states that, when augmented by targeted measurements, can partially recover treatment differences. This points to a potential synergy: unsupervised learning ensures generality and prevents overfitting, while classical features inject domain-specific signals.

Thirdly, the interpretability analysis via Grad-CAM was a crucial component (RQ4). In high-throughput experiments, a model that simply outputs a class label for each image is of limited use without biological interpretation. Our Grad-CAM results showed that the CNN model can be trusted to some extent: it focuses on areas that make sense (e.g., nucleus, cytoplasm) rather than on trivial artifacts. The differences in attention patterns across treatments aligned with known phenotypic effects, providing an additional layer of validation for the model's predictions. Moreover, this analysis can guide scientific insight. As mentioned, if a model clusters certain treatments together and also shows similar attention patterns for them, it suggests a shared mechanism of action. In practice, this integrated analysis could be used by biologists to generate hypotheses about unknown compounds by seeing what known compounds they cluster or share attention profiles with.

Limitations. Despite our positive results, several limitations remain. First, the small number of samples per class (approximately 10) severely limits the achievable accuracy. Many classes lack sufficient examples for the CNN to generalize well—resulting in overfitting for classes with distinctive morphologies and poor performance on subtler phenotypes. A larger dataset or additional replicates would likely improve performance. Moreover, class imbalance (e.g., the preponderance of DMSO controls) could bias the model; future work should consider oversampling minority classes or using class-balanced loss functions.

Second, our fusion of deep and CellProfiler features was implemented via a simple late-fusion method. This straightforward concatenation may not fully capture the potential interdependencies between the feature types. More sophisticated strategies, such as employing CP features to guide the CNN via an attention mechanism or training a joint model that integrates both image and CP inputs in intermediate layers, may offer greater improvements.

Third, our VAE was relatively basic, with a 64-dimensional latent space chosen somewhat arbitrarily. Alternative architectures (e.g., a β-VAE for better disentanglement, a conditional VAE, or self-supervised contrastive methods) may yield a latent space more aligned with the subtle variations induced by each treatment.

Finally, although Grad-CAM provides useful interpretability, it offers only a coarse localization of salient image regions. In some cases, it may produce diffuse or misleading heatmaps when the model is uncertain. Complementary interpretability methods (e.g., SmoothGrad, integrated

gradients, or guided backpropagation) and a more rigorous statistical analysis of the activation maps (such as PCA or clustering of the heatmaps themselves) could yield sharper insights.

An additional consideration is that the performance of the CellProfiler features depends on the quality of segmentation and the appropriateness of the predefined measurements. If a particular phenotype is not well captured by the CP pipeline, those features will not enhance the model's performance. Similarly, the CNN may miss subtle spatial relationships (such as cell clustering or interactions) that are not reflected in the single-image analysis. Future work should explore incorporating colony-level or spatial features to address these issues.

Broader implications. Despite the challenges, our study contributes to the growing evidence that deep learning can be applied to phenotypic screening and, with careful integration of prior knowledge (like known features), can yield not just predictions but also insights. The moderate success in classifying a very large number of classes is encouraging for using such methods in drug discovery or functional genomics, where one might screen thousands of treatments and then want to identify which treatments cause similar phenotypes. Our approach of combining data-driven and knowledge-driven features could be particularly useful in scenarios where labeled data is scarce but existing domain features are available.

In summary, the integration of deep representations and CellProfiler features provided a more complete decoding of cell image phenotypes than either alone. The deep learning model offered flexibility and power in feature extraction, while the handcrafted features ensured no obvious signal was overlooked due to the deep model's limitations. Our analyses spanning supervised, unsupervised, and interpretability techniques paint a comprehensive picture: we can classify some perturbations from images, cluster treatments by phenotypic similarity to some extent, and even interpret what visual cues underlie those similarities. As datasets grow and methods improve, such integrative approaches will likely become standard in image-based profiling, combining the strengths of human knowledge and machine learning to better understand cellular responses.

6 Conclusion

We presented a systematic study integrating deep learning and classical image feature analysis to decode phenotypic effects from high-content cell images. Using a dataset of 2,867 images across 250 perturbation classes, we showed that a fine-tuned ResNet-18 CNN can learn to differentiate some treatments from their morphological signatures, and its performance is boosted by incorporating features extracted via CellProfiler. The traditional morphological features alone provided a baseline but were less effective than deep features; however, in fusion they contributed additional discriminative power. Unsupervised embeddings from a variational autoencoder revealed that not all phenotypic information is captured without labels, reinforcing the value of supervised training for this task, while also indicating potential benefits of combining multiple feature sources. Through Grad-CAM, we interpreted the CNN's decisions and found they align with known biology, highlighting nuclei for DNA-targeting perturbations and cell outlines for cytoskeletal perturbations, among other patterns. Quantitative analysis of these attention maps across classes showed systematic differences in how broadly or focally the model "looks" at cells under different treatments, providing a novel perspective on phenotypic impact.

In conclusion, our work demonstrates that deep learning models, even when challenged by limited data, can extract meaningful phenotypic signals from cell images, and that classical features remain a valuable complement to ensure robustness and interpretability. The combined approach enhances both the performance and the confidence in the results, which is crucial for applications

in drug discovery or functional genomics where understanding the why behind a prediction is as important as the prediction itself. Future studies should build on this integrative strategy, leveraging larger datasets, advanced multi-modal models, and richer interpretability techniques to further bridge the gap between complex image data and biological insight. Ultimately, decoding phenotypic effects from images will enable faster and more informative screening of perturbations, guiding biomedical discoveries with the power of artificial intelligence and human domain knowledge hand in hand.

References

[1] Bray, M.-A. *et al.* (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, **11**(9), 1757–1774.

[2] Carpenter, A. E. *et al.* (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, **7**(10), R100.

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778).

[4] McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.

[5] Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations (ICLR)*.

[6] Selvaraju, R. R. *et al.* (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618-626).

[7] Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**(11), 559–572.