CP-BG-1M: A CONTROLLED MULTI-VIEW BENCHMARK FOR DENSITY AND BACKGROUND SHORTCUTS IN MORPHOLOGY PROFILING

Tim Treis[1], **Nikita Moshkov**[1], **Ghaith Mqawass**[1], **Fabian J. Theis**[1,2]
(1) Helmholtz Munich, Ingolstädter Landstraße 1, 85764 Oberschleißheim, Germany
(2) Technical University of Munich, Arcisstraße 21, 80333 München, Germany
{tim.treis, nikita.moshkov, ghaith.mqawass, fabian.theis}@helmholtz-munich.de

ABSTRACT

We introduce **CP-BG-1M**, a diagnostic framework for detecting cell-density and background-mediated shortcut learning in representation models for high-throughput morphology profiling. The dataset contains $\sim$1 million quality-controlled single-cell tiles from JUMP-CP Target2 dataset imaged across multiple production sites. Each cell is provided in four synchronized views that preserve center-cell morphology while selectively exposing background context or an explicit, source-agnostic density signal. This controlled design enables shortcut testing: morphology-driven embeddings should remain stable across views, whereas shortcut-dependent models show performance drops when background is removed and partial recovery when density is reintroduced without altering morphology. Using a DINOv3 ViT-B baseline with LoRA trained under a chemical-similarity contrastive objective, we reveal strong metric dependence. Segmented representations outperform crops in compound retrieval (recall@10: 0.37–0.38 vs. 0.29–0.30) and phenotypic activity detection (98.67% vs. 67.44–78.41% significant replicate agreement), while also improving batch mixing. These findings show that metric choice can invert model rankings by rewarding shortcut signals, positioning **CP-BG-1M** as a practical tool to diagnose and mitigate cell-density confounder.

1 INTRODUCTION

High-throughput microscopy has become a cornerstone of phenotypic drug discovery, enabling researchers to screen thousands of compounds by observing their effects on cellular morphology. Cell Painting (Bray et al., 2016; Cimini et al., 2023), a standardized multiplexed fluorescence imaging assay, has emerged as the dominant approach, capturing morphological changes across eight major cellular compartments using six stains, but is commonly acquired as five channels because several components are multiplexed into a single channel (e.g., the AGP channel). A persistent challenge is that the strongest sources of variation in these assays are often not mechanistic but experimental: batch effects such as image resolution or cell density differ across sites and batches, and directly affect many downstream measurements. The recent rise of deep learning has promised to transcend these limitations. Vision transformers and convolutional neural networks, when trained on Cell Painting images, now routinely outperform CellProfiler features on benchmark tasks like MOA classification and compound clustering (Kraus et al., 2024; Kim et al., 2025; Moshkov et al., 2024). A limitation is that deep learning models are prone to shortcut learning, a phenomenon when the model abuses spurious correlations with high signal in training data to achieve high performance (Geirhos et al., 2020). For example, in computer vision, instead of learning the properties of the object in classification task, model learns surroundings of this object (Beery et al., 2018). That phenomenon also creates a failure mode for discovery workflows: models can achieve strong benchmark performance by exploiting density- and background-mediated shortcuts (implicitly "counting cells" (Seal et al., 2026)) rather than capturing perturbation-specific single-cell morphology. If unrecognized, this shortcut can misrank compounds, obscure mechanism-specific signals, and limit transfer across experimental settings.

Previous work has provided circumstantial evidence for this shortcut. (Seal et al., 2026) demonstrated that cell count alone serves as a strong baseline for treatment prediction. (Moshkov et al., 2024) showed that models trained on "cells-in-context"-crops retaining background information outperform those trained on tightly cropped single cells, a result they attributed to beneficial contextual cues, but which equally supports the density-shortcut hypothesis. Critically, the field has lacked a controlled experimental framework to diagnose when models are exploiting density versus learning morphology. Existing datasets confound these signals: crops always include background, and segmentation always removes it, making it impossible to isolate density's contribution.

Recent advances in contrastive learning have shown promise for learning morphological representations from Cell Painting data. Methods like CLOOME (Contrastive Learning of Optical Morphology Embeddings) demonstrate that pairing cell images with molecular structure in a CLIP-style framework can produce embeddings that capture both morphological and chemical similarity. However, these approaches inherit the same vulnerability to density shortcuts: when images include background context, models can learn to associate molecular structure with cell count rather than true morphological features. The field currently lacks a standardized method for diagnosing whether such models exploit this shortcut or genuinely learn morphology-driven representations.

We introduce **CP-BG-1M**, a diagnostic dataset and evaluation protocol designed to turn this anecdotal concern into a quantifiable, reproducible measurement. The dataset comprises $\sim$1.1 million quality-controlled single-cell tiles derived from the JUMP Cell Painting Consortium, covering 302 unique compounds imaged across 10 production sites. Crucially, each cell is provided in four synchronized views: (A) a standard 150×150 pixel crop with background and neighboring cells; (B) the same cell with segmentation masks applied, removing background context; (C) the segmented cell with an explicit density encoding: corner patches whose intensity directly reflects field-of-view cell count; and (D) the crop with the same density encoding added. These four views hold the center cell's morphology constant while systematically toggling the model's access to density information and background context in orthogonal ways.

This intervention-based manipulation enables a simple diagnostic test while addressing a key issue: cell density can reflect real biology (e.g., cytotoxicity), so the goal is not to eliminate density information from representations. Instead, segmentation in **CP-BG-1M** does not eliminate density information; it removes *contextual shortcut channels*, such as background intensity statistics, neighbor proximity, and confluence patterns, through which models can infer density in a source-entangled way, while preserving morphology-encoded latent features of toxicity that remain visible within the segmented center cell. We additionally provide an explicit, source-agnostic density signal via corner-patch intensity to isolate the contribution of density alone. Under this setup, morphology-driven representations should be comparatively stable across views because the center cell is identical, whereas shortcut-rewarding benchmarks exhibit a characteristic signature: performance drops when background/context is removed (`crops` $\rightarrow$ `seg`) and partially recovers when density is reintroduced without adding morphology (`seg` $\rightarrow$ `seg_density`). To validate this diagnostic capability, we fine-tune a DINOv3 vision transformer (Siméoni et al., 2025) using LoRA (Low-Rank Adaptation) (Hu et al., 2021) paired with ECFP4 molecular fingerprints, training with a multi-positive contrastive loss that incorporates chemical similarity weighting. This baseline is not proposed as a state-of-the-art model, but rather as a representative example of modern self-supervised vision models adapted to Cell Painting through parameter-efficient fine-tuning. Across the evaluations we consider, the baseline shows clear sensitivity to view-controlled access to background and density, demonstrating that **CP-BG-1M** can diagnose context-mediated density reliance in contemporary morphology models.

Beyond exposing the shortcut, our analysis reveals a secondary benefit of segmentation: improved generalization across batch effects. By removing source-entangled contextual pathways (background intensity statistics, confluence textures, neighbor layout, and other field-of-view correlates that often track plating density and acquisition settings), segmented representations achieve roughly better mixing across production sites while maintaining equivalent biological signal preservation. This suggests that the most harmful shortcut is not density information per se, but density accessed through context and background cues that are entangled with imaging source and therefore limit transferability across experimental settings.

We release **CP-BG-1M** as an evaluation framework for diagnosing shortcut learning in perturbation biology representations. **CP-BG-1M** provides four synchronized views per cell that selectively ex-

pose (i) background context and (ii) an explicit density signal, enabling controlled tests of whether models rely on density-mediated shortcuts. Using identical DINOv3-ViT+LoRA CLIP-style fine-tuning across views, we show that metric choice can invert conclusions about representation quality. Benchmarks tied to background-inclusive aggregate statistics favor crops-based models, whereas perturbation-level biological validation and cross-source robustness favor segmentation-based representations. Concretely, our contributions are:

(1) **CP-BG-1M**: a controlled, multi-view perturbation-biology resource (1.1M cells, 302 compounds, 10 sources) designed for diagnostic evaluation of background- and density-mediated shortcuts.

(2) A shortcut-diagnosis protocol that (i) tests for the characteristic performance signature under controlled access to background and density ($\texttt{crops} \rightarrow \texttt{seg}$, $\texttt{seg} \rightarrow \texttt{seg\_density}$) and (ii) summarizes shortcut susceptibility via a scalar density reliance score $\mathrm{DRS} = \big(m(\texttt{crops}) - m(\texttt{seg})\big)/\big(m(\texttt{seg\_density}) - m(\texttt{seg})\big)$.

(3) Evidence of metric inversion: metric choice can invert conclusions about representation quality. CellProfiler feature prediction favors crops-based models, while compound retrieval, phenotypic activity detection, and cross-source mixing favor segmented representations (Figs. 3-4).

(4) Baseline implementations, checkpoints and standardized splits/evaluations (DINOv3-ViT+LoRA with chemical-similarity-guided contrastive training) to support reproducible benchmarking and future method development, to be released with the bioRχiv preprint.

By making shortcut learning measurable rather than anecdotal, we aim to redirect optimization pressure from proxy signals toward representations that are biologically meaningful and transferable across experimental settings. Importantly, **CP-BG-1M** has the goal to allow for quantifying shortcut reliance under controlled conditions, not to mitigate them.

2 Background and Related Work

Image-based morphological profiling summarizes perturbation effects from multiplexed microscopy into vector representations used for compound retrieval and mechanism-of-action discovery. Cell Painting is a dominant assay for this setting, with widely adopted protocols enabling large-scale, multi-channel profiling (Gustafsdottir et al., 2013; Bray et al., 2016; Cimini et al., 2023). Profiles are classically built by segmenting cells and extracting hand-crafted features (e.g., CellProfiler), followed by normalization and aggregation (Carpenter et al., 2006; Caicedo et al., 2017). Recent work increasingly replaces these descriptors with learned embeddings from deep representation learning, including self-supervised and foundation-model-style encoders trained on Cell Painting data (Pratapa et al., 2021; Kraus et al., 2024; Kim et al., 2025; Moshkov et al., 2024).

A major recent direction is to couple morphological representation learning with external supervision, in particular chemical structure, to bias embeddings toward biologically meaningful axes. CLIP-style image-chemistry objectives enable joint embedding spaces that support cross-modal retrieval and can improve perturbation representation quality (Sanchez-Fernandez et al., 2023; Moshkov et al., 2024). CLOOME is a representative example, pairing cell images with molecular structure in a contrastive framework to learn optical morphology embeddings aligned with chemistry (Sanchez-Fernandez et al., 2023). While these objectives can improve transfer and alignment, they also inherit vulnerability to nuisance shortcuts if the model can exploit confounded signals that correlate with the supervisory modality.

This issue is amplified in large multi-source datasets such as JUMP-CP (Chandrasekaran et al., 2023), where batch effects across sources and acquisition settings are substantial and can induce trade-offs between batch mixing and biological structure preservation (Arevalo et al., 2024). Accordingly, evaluation often combines perturbation-centric retrieval/activity metrics with robustness measures; retrieval-based frameworks such as mAP/`copairs` provide unified, non-parametric quantification of profile strength and similarity (Kalinin et al., 2025). Directly related to our study, cell density is a strong confound in perturbation screens and can act as a shortcut signal: cell count alone can perform remarkably well on bioactivity benchmarks (Seal et al., 2026). Our manuscript targets this gap by using controlled views that selectively expose background context and an explicit density signal, enabling diagnosis of density reliance under otherwise identical training and evaluation.

3 METHODS

3.1 CP-BG-1M DATASET CONSTRUCTION

We derived **CP-BG-1M** from the JUMP Cell Painting Consortium's Target2 compound subset, which comprises 302 unique chemical perturbations imaged across 10 production sites. While we instantiate the framework on JUMP-CP here, the underlying multi-view shortcut diagnostic is broadly applicable: any high-throughput microscopy dataset with per-cell crops (or centroids), segmentation masks (or an equivalent background-removal procedure), and a field-of-view density proxy (e.g., nuclei/cell count) can be converted into the same synchronized views. We chose JUMP-CP because its multi-source design and well-documented technical heterogeneity (plating density, acquisition settings, and nested site/batch structure) make it a stringent stress test where background- and density-mediated shortcuts are most likely to arise and to matter in practice. To ensure balanced compound representation while accounting for toxicity-driven cell yield variation, we applied submodular optimization on scPoli-integrated CellProfiler features to select maximally informative wells, scaling selection inversely by cell count to maintain coverage of cytotoxic compounds. Following segmentation with SPARCSpy and morphology-based quality control filtering (nucleus/cell area ratios, border distance thresholds), we obtained 1,146,049 single-cell crops.

From this filtered set, we generated four synchronized dataset variants: `crops`: 150x150 pixel bounding boxes centered on each cell, retaining background and neighboring cells; `seg`: the same cells with segmentation masks applied (nucleus mask on DAPI channel, cell mask on remaining channels), removing background context; `seg_density`: segmented cells with corner patches encoding field-of-view cell count as pixel intensity ($n_{\text{cells}}/2$, max 506 cells); and `crops_density`: crops with the same density encoding added. These variants implement explicit interventions on the image, masking background and injecting density signal, thereby creating counterfactual views that toggle density/context access while holding the center cell's morphology constant, enabling controlled diagnosis of shortcut learning. Example images can be seen in Figure 1A and Figure S3.

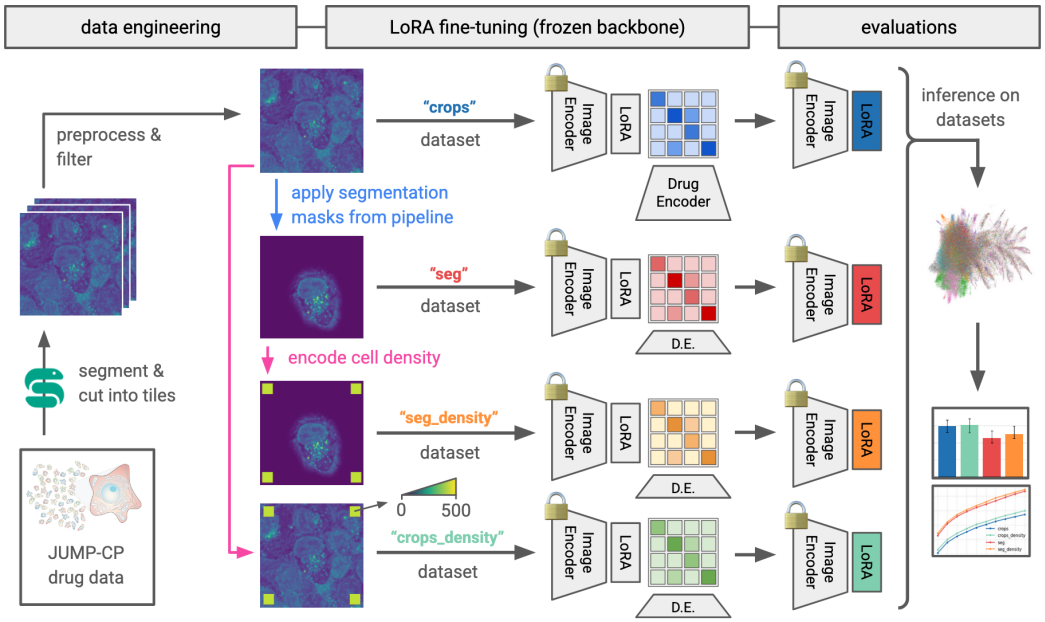


Figure 1: Overview of **CP-BG-1M** dataset creation, LoRA adaptation, and evaluations. Starting from JUMP-CP drug images, we preprocess, segment, and extract single-cell tiles to create four synchronized views: `crops` (background retained), `seg` (background removed), `seg_density` (`seg` + cell-density encoded as corner-patch intensity), and `crops_density` (`crops` + the same density encoding). For each view, a frozen DINOv3 ViT backbone is adapted with a view-specific LoRA module in a CLIP-style setup with a drug encoder, and the resulting embeddings are compared on typical downstream evaluations (retrieval, batch mixing, and CellProfiler feature prediction).

Full details on well selection, segmentation parameters, quality control thresholds, and metadata schema are provided in Appendix A.1.

3.2 Baseline Model Training

To validate **CP-BG-1M**'s diagnostic capability, we trained four separate image encoders, one per dataset variant, using identical architectures and training protocols (Figure 1). Each encoder is a DINOv3 vision transformer (ViT-Base) adapted to 5-channel Cell Painting images via a trainable 1×1 convolution (5→3 channels) and fine-tuned using LoRA (rank 8, α=16). Images are paired with precomputed ECFP4 molecular fingerprints (2048-bit, radius 2) projected through a linear layer into a shared 128-dimensional embedding space.

Training employs a multi-positive contrastive loss where all images of the same compound serve as hard positives, augmented with soft positives from the top-5 chemically similar compounds (Tanimoto similarity >0.25). Chemical similarity weights are ramped via curriculum learning ($\alpha(t)$ from 0 to 2 over steps 1k–3k). We trained for 20,000 steps using AdamW with mixed-precision (bfloat16), batch size 768 (96 unique compounds per batch), and compound-stratified sampling to balance morphologically active and inactive compounds. Architecture details, hyperparameters, and loss formulation are provided in the Appendix (sections A.2, A.3, and A.5).

3.3 Embedding Generation and Postprocessing

Following training, we generated embeddings by running inference with each of the four fine-tuned image encoders on a random subset of 250,000 images per dataset variant. This yields four distinct embedding spaces, sharing cell identities and metadata but differ in their access to density cues during training as described in Appendix A.1.

The embeddings of these single-cell images were then mean-pooled on a per-well basis resulting in a total of 13899 wells for all 4 datasets. These embeddings were then spherized based on the DMSO control wells as described in (Serrano et al., 2025). Next, the datasets were integrated using Harmony (Korsunsky et al., 2019) using a PCA embedding with 50 principal components, based on the quantitative benchmark of (Arevalo et al., 2024). This resulted in embeddings with 13899 observations and 50 features for each of the fine-tuned image encoders. These embeddings were used for all subsequent analysis steps.

Density reliance score (DRS) To quantify the extent to which an evaluation rewards background- and density-mediated shortcuts, we compute a derived diagnostic from the controlled views. For any metric $m(\cdot)$ (higher is better), we define $\Delta_{\mathrm{bg}} = m(\texttt{crops}) - m(\texttt{seg})$ and $\Delta_{\mathrm{dens}} = m(\texttt{seg\_density}) - m(\texttt{seg})$, and report $\mathrm{DRS} = \Delta_{\mathrm{bg}}/\Delta_{\mathrm{dens}}$. We compute DRS using the same summary statistic reported for each metric (e.g., median across features or folds). DRS is intended as an interpretable scalar diagnostic (not a new benchmark): values $\gg 1$ indicate background/context provides gains beyond explicit density, values near 1 indicate explicit density largely explains the gain, and negative values indicate that segmentation improves the metric, consistent with background/context acting as a confound for that evaluation.

4 RESULTS

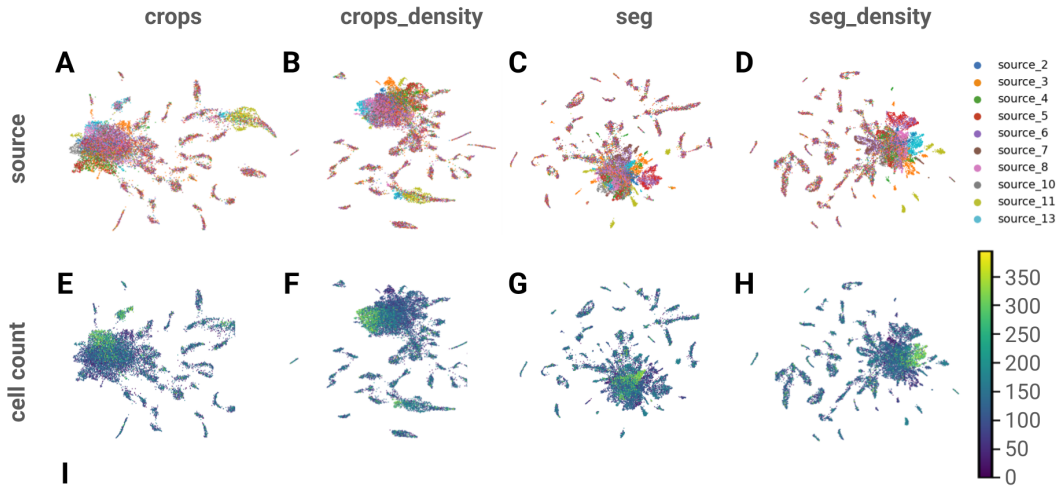
4.1 SEGMENTATION HELPS REMOVE TECHNICAL BATCH EFFECTS

A key challenge in learning representations from multi-source Cell Painting data, and in particular data from the JUMP-CP consortium, is disentangling biologically meaningful variation from technical batch effects (Arevalo et al., 2024). It is likely that many potential density shortcuts would manifest as source-specific artifacts, due to differences in cell plating density, confluence at the time of imaging, and drug dilution processes. To test whether removing background context mitigates these confounds, we computed integration metrics using the scIB-metrics package (Luecken et al., 2022) on embeddings from each of the four fine-tuned encoders.

We assessed representation quality along two complementary axes:

- batch correction: quantifying the mixing of cells across the 10 production sources (higher scores indicate reduced batch effects), and
- quantifying preservation of compound-level structure (higher values indicate that cells from the same compound remain neighbors in embedding space).

The total score represents a weighted average of these components. Figure 2 shows UMAP projections colored by source and well-level cell count alongside quantitative metrics, a qualitative biological sanity check with the Target2 control compounds can be found in Fig. S7.



I

Dataset	Segmented	Cell density	Batch correction	Bio conservation	Total
crops			0.25	0.46	0.37
crops_density		✅	0.31	0.46	0.40
seg	✅		0.44	0.50	0.48
seg_density	✅	✅	0.43	0.50	0.47

Figure 2: Segmentation improves batch correction without sacrificing biological signal. Panels A–H: UMAP projections of the embeddings colored by source and well-level cell count for each dataset variant: (A+E) `crops`, (B+F) `crops_density`, (C+G) `seg`, (D+H) `seg_density`. Panel I: Quantitative integration metrics computed using scib-metrics.

Segmented datasets perform substantially better than their non-segmented counterparts (+0.11 score for `crops` vs `seg` and +0.07 for `crops_density` vs `seg_density`). The UMAP projections visually confirm this pattern: embeddings from segmented variants show increased mixing across sources (Figure 2C–D) compared to crops-based variants (Figure 2A–B), where, for example, a larger cluster of wells from source 11 remains visible.

This improved batch correction with preserved biological structure suggests that background context introduces source-specific confounders (differences in cell density, imaging artifacts, plate effects) that the model encodes in its features, rather than biologically meaningful morphological information. Segmentation effectively strips away a layer of technical variation that would otherwise contaminate learned embeddings. For downstream applications requiring cross-study generalization, such as mechanism-of-action prediction or compound screening across institutions, this batch robustness is essential for reliable biological inference.

Notably, adding explicit density encoding to segmented images (`seg_density`) does not degrade the total integration score relative to pure segmentation (`seg`), achieving very similar scores of 0.47 and 0.48. This indicates that the density shortcut is not merely cell count per se, but rather the spatial distribution patterns and background context that correlate with source-specific imaging protocols. The corner patch encoding provides density information in a source-agnostic format, allowing the model to access cell count without learning source-specific visual artifacts.

4.2 Segmentation Helps Models Learn Biologically Coherent Representations

To assess whether learned representations capture genuine perturbation biology versus technical artifacts, we evaluated embeddings across complementary tasks probing compound-level organization and phenotypic activity detection. If models learn robust morphological features, embeddings should organize cells primarily by compound identity (biological signal) rather than other technical confounders, such as the generating laboratory ("source") or cell density.

4.2.1 Segmentation Improved Compound Retrieval

We quantified embedding space quality using compound-level recall@k: for each well, we count whether any of its k nearest neighbors shares the same compound (Figure 3A). This metric directly measures whether embeddings cluster by the morphological phenotype induced by a given perturbation as opposed to confounding factors.

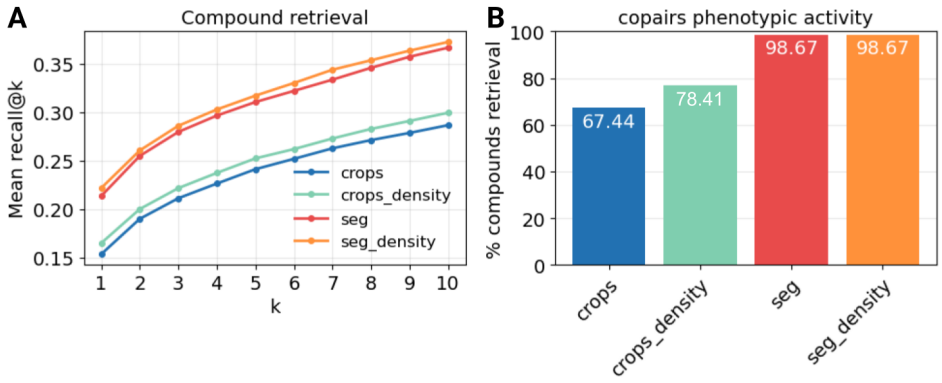


Figure 3: Segmentation improves the retrieval of identical perturbations. Panel A: Compound-level retrieval: mean recall@k measuring whether wells' k-nearest neighbors share the same compound (out of 302 unique labels). Embeddings from models trained on segmented data (red, orange) achieve a roughly 7% higher recall than those trained on simple crops (blue, green), demonstrating embeddings organize by compound identity rather than density. Panel B: Each bar shows the percentage of the total 302 unique compounds labeled as phenotypically active based on copairs replicate retrieval.

The embeddings from models trained on segmented data (seg, seg_density) substantially outperform crops-trained models across all k values, achieving recall@10 of 0.37-0.38 compared to 0.29-0.30 for crops-based variants, a 25% relative improvement. This gap is systematic across the entire recall curve, indicating a fundamental difference in embedding structure: when background context is removed, representations organize primarily by compound identity. In contrast, crops-trained models exhibit density-driven clustering, where wells from different compounds at similar densities (e.g., two toxic compounds with low cell counts) are embedded closer than cells from the same compound at different densities.

Critically, while both density-encoded variants consistently outperform their counterparts by a small margin across all k values, this gap is substantially smaller than the difference between segmented and crops-trained models (0.37-0.38 vs 0.29-0.30). This pattern suggests that explicit density encoding provides modest additional signal for compound discrimination. While cell count does carry some biological information about perturbation effects, the majority of the information retained in the background seems to bias model towards technical confounders.

4.2.2 Phenotypic Activity Detection: Segmentation Improves Robustness of Representations

The superior embedding organization of segmented models directly translates to improved detection of biologically active compounds. Using Copairs (Kalinin et al., 2025), a non-parametric replicability metric that quantifies whether compound replicates produce more similar profiles than expected by chance, we assessed what percentage of the 302 compounds achieve significant replicate agreement (average precision p-value <0.05) in each embedding space (Figure 3B).

Models trained on segmented data achieve 98.67% retrieval of active compounds, while crops-trained models retrieve only 67.44%-78.41%. This 30% gap implies that in a real phenotypic screen, embeddings from crops-trained models could miss nearly one-third of hits. The near-perfect performance of segmented variants demonstrates that when embeddings are organized by compound identity (Figure 3A), they naturally excel at distinguishing perturbation-driven phenotypes from control conditions.

The phenotypic activity failures of crops-trained models arise from their density-driven organization: when a compound's replicates span different density regimes (due to well-to-well variation in confluence or toxicity), crops-trained models embed them far apart because density variation dominates the embedding distance. This causes the copairs metric to score these as "non-replicating" despite genuine morphological activity. In contrast, segmented models, forced to ignore background density cues, learn representations where replicates cluster by their shared compound-driven morphology regardless of density differences.

Together, these results demonstrate that **CP-BG-1M**'s controlled manipulation of density cues reveals a fundamental trade-off: crops-trained models optimize for density-correlated patterns at the expense of biological signal, while segmented models, forced to ignore background context, learn representations that capture perturbation-driven morphology.

4.3 Traditional Morphology Benchmarks Favor Density-Exploiting Models

To understand why crops-trained models excel in traditional benchmarks despite performing worse at biological tasks (Section 4.1 and Section 4.2), we evaluated all four variants on a traditional benchmark: predicting CellProfiler features. We trained ridge regression models to predict 599 compound-level CellProfiler features (selection described in Appendix A.2), assessing per-feature R^2 via 10-fold cross-validation (Figure 4A). To diagnose whether this benchmark rewards density learning, we also trained models to predict field-of-view-level cell count from embeddings, stratifying folds by compound to prevent data leakage (Figure 4B).

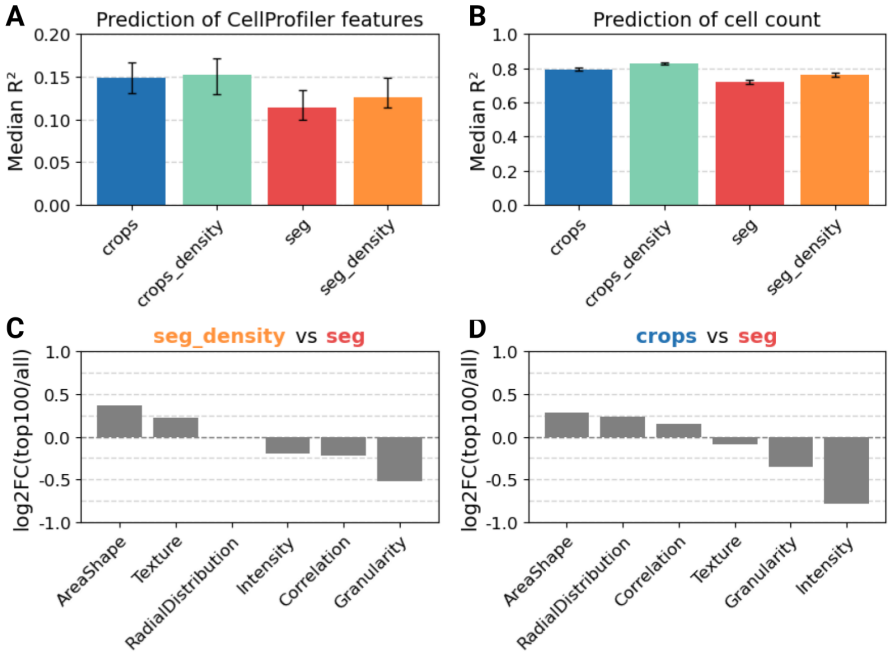


Figure 4: CellProfiler feature and cell count prediction highlights density-dependent gains. Panel A: Median R^2 per model with 95% bootstrap CIs over features. Panel B: Median R^2 per model with 95% bootstrap CIs over folds. Panel C-D: Feature-family enrichment among the top-100 features, ranked by ΔR^2 for `seg_density` - `seg` (C) and `crops` - `seg` (D), computed as $\log_2[(n_{fam}/n_{\text{sel}})/(N_{fam}/N_{\text{all}})]$.

Crops-trained models achieve higher CellProfiler feature prediction performance (median R^2: 0.149 for `crops`, 0.153 for `crops_density`) than segmented models (0.114 for `seg`, 0.126 for `seg_density`). However, all four models, including those trained exclusively on segmented cells, predict cell count with remarkably high accuracy: (0.794 for `crops`, 0.828 for `crops_density`) compared to segmented models (0.723 for `seg`, 0.761 for `seg_density`).

This reveals that cell density information is recoverable even from segmented single-cell images, likely through toxicity-induced morphological changes, such as apoptosis-related shape changes, stress granules, and organelle disruption that the model learns to associate with low-density conditions at the compound level. Critically, crops-trained models achieve only marginally better cell count prediction than segmented models, yet show substantially larger gaps in CellProfiler feature prediction. This suggests that the CellProfiler benchmark advantage of crops-trained models derives not merely from better density estimation, but from learning background spatial patterns, neighbor proximity, and field-of-view statistics that correlate with aggregate feature values without requiring single-cell morphological understanding.

To identify which CellProfiler features drive these performance gaps, we ranked and sorted the ΔR^2 of the prediction task and computed enrichment for feature families among the top 100. Features most improved by explicit density encoding (`seg_density` vs `seg`, panel 4C) are enriched for AreaShape and Texture families, indicating that controlled density information helps predict gross

morphological measurements and spatial texture patterns, both of which plausibly correlate with compound-level toxicity and thus cell count. In contrast, features most improved by background context (`crops` vs `seg`, panel 4D) show enrichment for AreaShape, Correlation, and especially Intensity families. The strong Intensity enrichment is particularly revealing: background pixels contribute directly to field-of-view intensity measurements, and models trained on crops can exploit empty background regions and neighboring cell signals to predict these aggregate statistics without learning individual cell morphology.

We can make this trade-off more concrete using the density reliance score (DRS) derived from our controlled views. On CellProfiler feature prediction (Fig. 4A), the median performance gap between `crops` and `seg` is $\Delta_{\mathrm{bg}} = 0.149 - 0.114 = 0.035$, while the gain from adding explicit density to segmented inputs is $\Delta_{\mathrm{dens}} = 0.126 - 0.114 = 0.012$, yielding $\mathrm{DRS} \approx 2.9$. This indicates that most of the benchmark advantage of crops-based models is attributable to background/neighborhood context beyond what is explained by a scalar density signal. In contrast, for predicting cell count itself (Fig. 4B), $\Delta_{\mathrm{bg}} = 0.794 - 0.723 = 0.071$ and $\Delta_{\mathrm{dens}} = 0.761 - 0.723 = 0.038$, giving $\mathrm{DRS} \approx 1.9$, consistent with density being a strong but not exclusive driver. Together, these DRS values support a two-pathway view of confounding: (1) morphology-encoded correlates of toxicity that remain accessible under segmentation (explaining high cell-count predictability even for `seg`), and (2) additional background-mediated shortcuts (intensity statistics, neighbor proximity, confluence patterns) that inflate aggregate-feature benchmarks and can dominate model ranking without corresponding gains in perturbation-specific morphology.

Models trained only on segmented cells are forced to learn morphology-encoded density which retains biological signal and generalizes across sites (Section 4.1). Crops-trained models can bypass this by learning background-encoded density which yields higher benchmark scores but performs worse at biological tasks (Section 4.2) and embeds source-specific artifacts (Section 4.1).

This analysis suggests a flaw in current evaluation practices when optimizing for aggregate feature prediction. Models achieving state-of-the-art CellProfiler feature prediction may be learning compound-level density signatures and background statistics rather than the single-cell morphological features necessary for mechanism-of-action discovery and cross-institutional generalization.

4.4 Evaluation Metric Choice Determines Model Rankings

Our controlled views reveal that the apparent "best" representation depends strongly on the evaluation metric, and that common benchmarks can systematically reward density- and background-mediated shortcuts. On the traditional benchmark of CellProfiler feature prediction (Fig. 4A), crops-based models rank highest, consistent with the fact that many aggregate features encode field-of-view context (background intensity statistics, neighbor proximity, and confluence patterns) that is directly available in `crops` and `crops_density`. In contrast, when we evaluate biological validity using perturbation-centric criteria such as compound retrieval and phenotypic activity detection (Fig. 3), segmented data (`seg`, `seg_density`) rank highest, indicating embeddings are organized by reproducible perturbation phenotypes rather than density-correlated artifacts. Finally, on cross-site robustness (Fig. 2), segmentation yields substantially improved source mixing while preserving compound-level structure, highlighting that the shortcut signal is also source-specific and limits transfer across experimental settings. Together, these results show that metric choice does not merely change the scale of reported improvements but can invert model rankings: evaluations tied to background-inclusive aggregate statistics favor shortcut-exploiting representations, whereas perturbation-level replicability and cross-source generalization favor representations constrained to single-cell morphology. **CP-BG-1M** is designed to make this dependence explicit and to support experiment-aware model selection for phenotypic screening.

5 DISCUSSION

Our study shows that metric choice can invert conclusions about representation quality: representation models can achieve strong quantitative performance by exploiting experimental correlates of perturbation rather than learning perturbation-specific single-cell morphology. Using **CP-BG-1M**, we turn this concern into a controlled, counterfactual test. By providing four synchronized views of the same cell that orthogonally toggle access to background context and an explicit density signal, we transform "implicit cell counting" from an anecdotal explanation into a measurable diagnostic. Across a representative modern training setup (foundation ViT encoder adapted via LoRA and guided by chemical similarity), we observe a consistent pattern: representations trained with background context exhibit stronger sensitivity to density- and source-linked variation, while segmentation-based views reduce these shortcuts and yield embeddings that better reflect reproducible perturbation phenotypes.

A central implication is that evaluation choice can invert model rankings. On aggregate-statistics benchmarks such as CellProfiler feature prediction, crops-based models score highest, despite performing worse on perturbation-centric evaluations. This is expected because many classical features summarize field-of-view properties (background intensity statistics, neighbor proximity, confluence textures) that are directly available in `crops` and can be inferred without improving single-cell morphology understanding. Our feature-family enrichment analysis supports this interpretation: the largest performance gains attributable to background context concentrate in intensity- and correlation-related features, which are intrinsically sensitive to background pixels and neighborhood structure. In contrast, when we evaluate representations using perturbation-centric criteria, compound retrieval and replicate agreement, segmented views perform substantially better and recover a much larger fraction of significant, reproducible perturbations. These metrics more directly reflect the downstream goal of phenotypic screening: identifying perturbations that induce consistent morphological effects despite variation in experimental conditions.

Because cell density can reflect real biology (e.g., cytotoxicity), the goal is not to remove density information from representations. Our results instead suggest that segmentation does not eliminate density information but it instead removes contextual shortcut channels while preserving morphology-encoded correlates of toxicity. Cell count is not purely technical since toxicity can induce morphological states that co-occur with reduced cell yield, making density partially recoverable even from segmented single-cell images. This is consistent with our finding that all models, including those trained only on segmented inputs, predict cell count with high accuracy. The problematic shortcut is therefore not the existence of density signal per se, but how the signal is accessed. Background-inclusive training provides models with additional pathways to density and confluence via spatial statistics and imaging artifacts that are entangled with source-specific acquisition conditions. Segmentation removes these contextual pathways, forcing the model to rely on morphology-encoded correlates that are more likely to generalize. Notably, reintroducing density in a controlled, source-agnostic form (`seg_density`) preserves batch-mixing performance, consistent with the hypothesis that the most harmful shortcut arises from source-entangled visual correlates rather than the scalar count itself.

These observations have practical consequences for both users and method developers. For practitioners using learned representations for hit finding, mechanism-of-action discovery, or transfer across experimental settings, our results suggest that constraining access to background context, or explicitly controlling for density, can improve robustness and replicate coherence. Conversely, if the task is explicitly to predict field-of-view or aggregate imaging statistics, background-inclusive models may be appropriate and should not be interpreted as single-cell morphology specialists.

More broadly, we recommend reporting shortcut susceptibility alongside task performance: (i) how well embeddings predict cell count, (ii) deltas under controlled access to background and density (`crops`→`seg` and `seg`→`seg_density`), (iii) robustness to batch/source variation, and (iv) a scalar summary such as DRS, interpreted as metric-specific evidence that an evaluation benefits from shortcut access rather than improved morphology.

CP-BG-1M also enables algorithmic research aimed at shortcut-resistant learning. Because the four views hold the center cell constant while manipulating shortcut access, they provide a natural testbed for training objectives that penalize reliance on context (e.g., invariance constraints between `crops` and `seg`), explicit nuisance-factor removal (e.g., density-adversarial objectives), or multi-instance

formulations that separate single-cell morphology from field-of-view statistics. Importantly, our baseline is not intended as a state-of-the-art method; rather, it demonstrates that shortcut learning persists under contemporary ingredients (foundation vision transformers, parameter-efficient fine-tuning, and chemical-similarity-guided contrastive training). The broader lesson is that supervision from chemistry, while valuable, does not by itself prevent exploitation of experimental correlates when they align with training signal.

Our work has limitations. **CP-BG-1M** is derived from the JUMP-CP Target2 compound set, imaged in U2OS cells, and thus focuses on a specific assay setting and a finite set of perturbations. Whether the observed shortcut patterns generalize to other cell lines, imaging platforms, or assay configurations remains an open question we intend to address in future work; the multi-view diagnostic design is in principle applicable to any dataset with per-cell crops, segmentation masks, and a field-of-view density proxy, but empirical validation on additional settings is needed before broad generalization can be claimed. Extending the framework to additional Cell Painting datasets and cell lines is a direct next step we are actively pursuing.

Segmentation quality and crop sizing can also influence conclusions; although we apply morphology-based QC, residual segmentation artifacts may remain. Finally, density is sometimes biologically meaningful, particularly for cytotoxic mechanisms, and "removing density" is not universally desirable. The goal is not to eliminate density signal, but to prevent models from exploiting source-entangled contextual correlates as a proxy for phenotype.

In summary, **CP-BG-1M** provides a controlled evaluation framework that makes density- and background-mediated shortcut learning measurable. By demonstrating that common benchmarks can reward shortcut access and invert conclusions, we aim to encourage evaluation practices that better reflect phenotypic screening objectives: representations that capture reproducible perturbation morphology and generalize beyond the specific experimental settings in which they were trained.

ACKNOWLEDGEMENTS

We thank Shantanu Singh, Alan Muñoz, and Alessandro Palma for helpful discussions and feedback on this work.

T.T., N.M., and G.M. are supported by Helmholtz Munich. T.T. is a doctoral researcher in the Munich School for Data Science (MUDS). G.M. is a doctoral researcher funded by the European Union's Horizon research and innovation programme under the Marie Skłodowska-Curie Actions Doctoral Networks grant agreement No. 101120466 (AiChemist). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Code and data will be made publicly available at `https://github.com/theislab/CP-BG-1M`.

LARGE LANGUAGE MODEL USAGE

Large language models were used as general-purpose writing assistance tools during the preparation of this manuscript, including for grammar checking and phrasing suggestions. All scientific content, experimental design, analysis, and conclusions are solely the work of the authors.

REFERENCES

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

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. pp. 472–489, 2018.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M Gustafsdottir, Christopher C Gibson, and Anne E Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.*, 11(9):1757–1774, September 2016.

Juan C Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Joseph D Barry, Harmanjit Singh Bansal, Oren Kraus, Mathias Wawer, Lassi Paavolainen, Markus D Herrmann, Mohammad Rohban, Jane Hung, Holger Hennig, John Concannon, Ian Smith, Paul A Clemons, Shantanu Singh, Paul Rees, Peter Horvath, Roger G Linington, and Anne E Carpenter. Data-analysis strategies for image-based cell profiling. *Nat. Methods*, 14(9):849–863, September 2017. doi: 10.1038/nmeth.4397.

Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, Polina Golland, and David M Sabatini. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.*, 7(10):R100, October 2006.

Srinivas Niranj Chandrasekaran, Jeanelle Ackerman, Eric Alix, D Michael Ando, John Arevalo, Melissa Bennion, Nicolas Boisseau, Adriana Borowa, Justin D Boyd, Laurent Brino, Patrick J Byrne, Hugo Ceulemans, Carolyn Ch'ng, Beth A Cimini, Djork-Arne Clevert, Nicole Deflaux, John G Doench, Thierry Dorval, Regis Doyonnas, Vincenza Dragone, Ola Engkvist, Patrick W Faloon, Briana Fritchman, Florian Fuchs, Sakshi Garg, Tamara J Gilbert, David Glazer, David Gnutt, Amy Goodale, Jeremy Grignard, Judith Guenther, Yu Han, Zahra Hanifehlou, Santosh Hariharan, Desiree Hernandez, Shane R Horman, Gisela Hormel, Michael Huntley, Ilknur Icke, Makiyo Iida, Christina B Jacob, Steffen Jaensch, Jawahar Khetan, Maria Kost-Alimova, Tomasz Krawiec, Daniel Kuhn, Charles-Hugues Lardeau, Amanda Lembke, Francis Lin, Kevin D Little, Kenneth R Lofstrom, Sofia Lotfi, David J Logan, Yi Luo, Franck Madoux, Paula A Marin Zapata, Brittany A Marion, Glynn Martin, Nicola Jane McCarthy, Lewis Mervin, Lisa Miller, Haseeb Mohamed, Tiziana Monteverde, Elizabeth Mouchet, Barbara Nicke, Arnaud Ogier, Anne-Laure Ong, Marc Osterland, Magdalena Otrocka, Pieter J Peeters, James Pilling, Stefan Prechtl, Chen Qian, Krzysztof Rataj, David E Root, Sylvie K Sakata, Simon Scrace, Hajime Shimizu, David Simon, Peter Sommer, Craig Spruiell, Iffat Sumia, Susanne E Swalley, Hiroki Terauchi, Amandine Thibaudeau, Amy Unruh, Jelle Van de Waeter, Michiel Van Dyck, Carlo van Staden, Michał Warchoł, Erin Weisbart, Amélie Weiss, Nicolas Wiest-Daessle, Guy Williams, Shan Yu, Bolek Zapiec, Marek Żyła, Shantanu Singh, and Anne E Carpenter. JUMP cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, pp. 2023.03.23.534023, March 2023.

Beth A Cimini, Srinivas Niranj Chandrasekaran, Maria Kost-Alimova, Lisa Miller, Amy Goodale, Briana Fritchman, Patrick Byrne, Sakshi Garg, Nasim Jamali, David J Logan, John B Concannon, Charles-Hugues Lardeau, Elizabeth Mouchet, Shantanu Singh, Hamdah Shafqat Abbasi, Peter Aspesi, Jr, Justin D Boyd, Tamara Gilbert, David Gnutt, Santosh Hariharan, Desiree Hernandez, Gisela Hormel, Karolina Juhani, Michelle Melanson, Lewis H Mervin, Tiziana Monteverde, James E Pilling, Adam Skepner, Susanne E Swalley, Anita Vrcic, Erin Weisbart, Guy Williams, Shan Yu, Bolek Zapiec, and Anne E Carpenter. Optimizing the cell painting assay for image-based profiling. *Nat. Protoc.*, 18(7):1981–2013, July 2023.

Carlo De Donno, Soroor Hediyeh-Zadeh, Amir Ali Moinfar, Marco Wagenstetter, Luke Zappia, Mohammad Lotfollahi, and Fabian J Theis. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat. Methods*, 20(11):1683–1692, November 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, November 2020.

Sigrun M Gustafsdottir, Vebjorn Ljosa, Katherine L Sokolnicki, J Anthony Wilson, Deepika Walpita, Melissa M Kemp, Kathleen Petri Seiler, Hyman A Carrel, Todd R Golub, Stuart L Schreiber, Paul A Clemons, Anne E Carpenter, and Alykhan F Shamji. Multiplex cytological profiling assay to measure diverse cellular states. *PLoS One*, 8(12):e80999, December 2013. doi: 10.1371/journal.pone.0080999.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Alexandr A Kalinin, John Arevalo, Erik Serrano, Loan Vulliard, Hillary Tsang, Michael Bornholdt, Alán F Muñoz, Suganya Sivagurunathan, Bartek Rajwa, Anne E Carpenter, Gregory P Way, and Shantanu Singh. A versatile information retrieval framework for evaluating profile strength and similarity. *Nat. Commun.*, 16(1):5181, June 2025.

Vladislav Kim, Nikolaos Adaloglou, Marc Osterland, Flavio M Morelli, Marah Halawa, Tim König, David Gnutt, and Paula A Marin Zapata. Self-supervision advances morphological profiling by unlocking powerful image representations. *Sci. Rep.*, 15(1):4876, February 2025.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, Dominique Beaini, Maciej Sypetkowski, Chi Vicky Cheng, Kristen Morse, Maureen Makes, Ben Mabey, and Berton A Earnshaw. Masked autoencoders for microscopy are scalable learners of cellular biology. *ArXiv*, abs/2404.10242, April 2024.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

Nikita Moshkov, Michael Bornholdt, Santiago Benoit, Matthew Smith, Claire McQuin, Allen Goodman, Rebecca A Senft, Yu Han, Mehrtash Babadi, Peter Horvath, Beth A Cimini, Anne E Carpenter, Shantanu Singh, and Juan C Caicedo. Learning representations for image-based profiling of perturbations. *Nat. Commun.*, 15(1):1594, February 2024.

Aditya Pratapa, Michael Doron, and Juan C Caicedo. Image-based cell phenotyping with deep learning. *Curr. Opin. Chem. Biol.*, 65:9–17, December 2021. doi: 10.1016/j.cbpa.2021.04.001.

Ana Sanchez-Fernandez, Elisabeth Rumetshofer, Sepp Hochreiter, and Günter Klambauer. Contrastive learning unlocks bioimaging databases for queries with chemical structures. *Nat. Commun.*, 2023. doi: 10.1038/s41467-023-42328-w.

Niklas A Schmacke, Sophia C Mädler, Georg Wallmann, Andreas Metousis, Marleen Bérouti, Hartmann Harz, Heinrich Leonhardt, Matthias Mann, and Veit Hornung. SPARCS, a platform for genome-scale CRISPR screening for spatial cellular phenotypes. *Systems Biology*, June 2023.

Jacob Schreiber, Jeffrey Bilmes, and William Stafford Noble. apricot: Submodular selection for data summarization in python. *arXiv [cs.LG]*, pp. 1–6, June 2019.

Srijit Seal, William Dee, Adit Shah, Natacha Cerisier, Andrew Zhang, Esteban Miglietta, Katherine Titterton, Ángel Alexander Cabrera, Daniil Boiko, Alex Beatson, Gregory Slabaugh, Olivier Taboureau, Jordi Carreras Puigvert, Shantanu Singh, Ola Spjuth, Andreas Bender, and Anne E. Carpenter. Counting cells can accurately predict small-molecule bioactivity benchmarks. *Nat. Commun.*, February 2026. doi: 10.1038/s41467-026-68725-5. URL https://www.nature.com/articles/s41467-026-68725-5.

Erik Serrano, Srinivas Niranj Chandrasekaran, Dave Bunten, Kenneth I Brewer, Jenna Tomkinson, Roshan Kern, Michael Bornholdt, Stephen J Fleming, Ruifan Pei, John Arevalo, Hillary Tsang, Vincent Rubinetti, Callum Tromans-Coia, Tim Becker, Erin Weisbart, Charlotte Bunne, Alexandr A Kalinin, Rebecca Senft, Stephen J Taylor, Nasim Jamali, Adeniyi Adeboye, Hamdah Shafqat Abbasi, Allen Goodman, Juan C Caicedo, Anne E Carpenter, Beth A Cimini, Shantanu Singh, and Gregory P Way. Reproducible image-based profiling with pycytominer. *Nat. Methods*, 22(4):677–680, April 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

A Appendix

A.1 Creation of the **CP-BG-1M** dataset

The JUMP-CP consortium produced a large-scale, publicly available Cell Painting dataset comprising over 116,000 unique chemical and over 15,000 genetic perturbations imaged across multiple contributing partners (”sources”), both from academia and industry (Chandrasekaran et al., 2023). It contains both the raw images and morphological profiles from close to 2 billion of treated cells, imaged across the standardized 5-channel Cell Painting assay (DNA, RNA, ER, AGP, Mito). This multi-source structure provides both the scale necessary for large-scale model training and bio discovery, but also introduces significant batch effects as reported in (Arevalo et al., 2024). We derived our CP-BG-1M dataset by subsampling and filtering this resource to obtain balanced compound coverage with quality-controlled single-cell crops.

With the goal of providing a maximally diverse yet balanced image dataset, we considered only the Target2 subset of JUMP-CP. A set of 302 unique chemical perturbations was included as control across all experimental batches in separate plates, ensuring that every compound appears in every source, a property essential for disentangling compound effects from batch effects. Selecting an informative subset from this resource required addressing two challenges: the hierarchical batch structure inherent to JUMP-CP (where each source generated multiple batches, each containing multiple plates, each with multiple wells) and the uneven cell yields across compounds due to varying toxicity. We first processed CellProfiler features following the pipeline described in (Chandrasekaran et al., 2023) and (Serrano et al., 2025), going from 4763 features to 599, excluding source 9 due to substantial batch effects arising from its use of 384-well plates rather than the 96-well format standardized across other sources. Then we integrated the processed features using scPoli (De Donno et al., 2023), which we specifically selected for its ability to model hierarchical batch effects, allowing us to account for source, batch, and plate-level variation simultaneously. From this integrated feature space, we applied sub-modular optimization via apricot (Schreiber et al., 2019) to select the n most informative wells per compound per source. The number of wells n was inversely scaled by cell count in each well, guaranteeing a minimum of $n \geq 3$ wells per compound while allocating additional wells to more toxic compounds, which yield fewer cells per well. This scaling ensures that compounds with strong cytotoxic effects, which are often biologically interesting but underrepresented in cell-level sampling, maintain sufficient coverage in the final dataset. The UMAPs of the integrated latent space and the selected wells are shown in Figure S1.

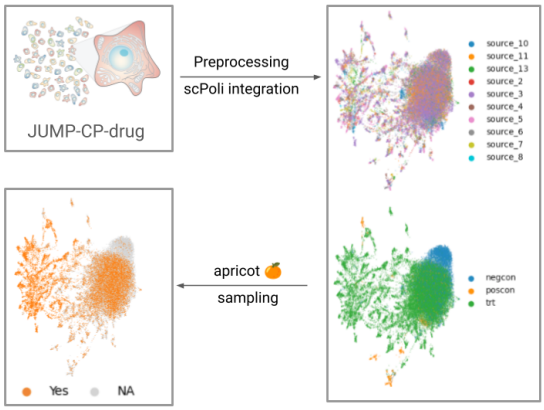


Figure S1: Construction of the CP-BG-1M dataset from JUMP-CP. Starting from the JUMP-CP Target2 compound set (top left), we processed CellProfiler morphological features and integrated them using scPoli to account for hierarchical batch effects. The integrated feature space (top right) shows mixing across the 10 contributing sources while preserving separation between negative controls (negcon), positive controls (poscon), and treatment wells (trt). We then applied sub-modular optimization via apricot to select maximally informative wells (bottom left), with selection scaled inversely by cell count to ensure adequate representation of cytotoxic compounds. Orange points indicate selected wells; gray points (NA) were excluded from the final dataset.

Next, we run a Snakemake pipeline that orchestrates SPARCSpy (Schmacke et al., 2023) to segment the raw images from all sources for all chemical perturbations available in the data repository of the JUMP-CP consortium into individual cells, providing masks for both cell and nuclei. From this dataset, we then extracted the cells corresponding to the wells identified by apricot in the previous step. From these we then removed all cells that did not at least have 128 pixel distance from the images' border to enable correct downstream cropping. Furthermore, we removed all cells where the size of nucleus and cell segmentation masks as well as the ratio of these was outside of the 2.5 % - 97.5 % quantiles to remove likely segmentation artifacts. The distributions of these values and the respective thresholds are shown in Figure S2.

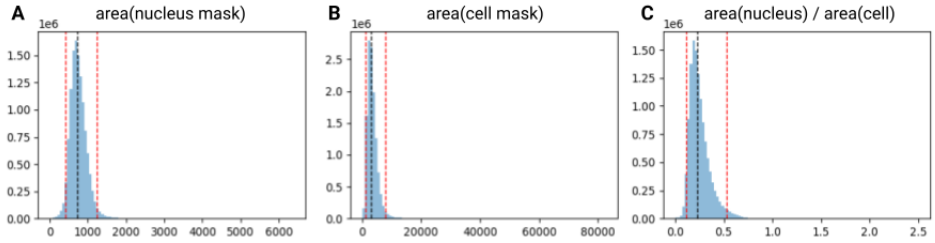


Figure S2: Morphology-based quality control filtering. Distributions of (A) nucleus area, (B) cell area, and (C) nucleus-to-cell ratio for segmented crops. Red dashed lines mark 2.5 % - 97.5 % quantile thresholds; cells outside these bounds were excluded as likely segmentation artifacts.

This workflow yielded a dataset of 1,146,049 unique cells. Based on these, we generated the first dataset `crops` by cropping a bounding box of 150x150 pixels around the cell masks' centroid. Next, we used the respective segmentation masks for these cells to generate dataset `seg` by subsetting the DAPI channel to the nucleus mask and the other 4 channels to the cell mask. Then we generated two more datasets by encoding the number of cells in the field-of-view the crop originates from as square patches in the corner of the tile. Since the maximal number of cells in a given FOV was 506, we chose to encode the intensity directly as $n_{cells}/2$, conveniently matching the uint8 value range in which the images were stored. By embedding this cell density in the `seg` dataset, we generated dataset `seg_density` and by embedding it in the `crops` dataset, we generated dataset `crops_density`. An example image of this can be seen in Figure S3.

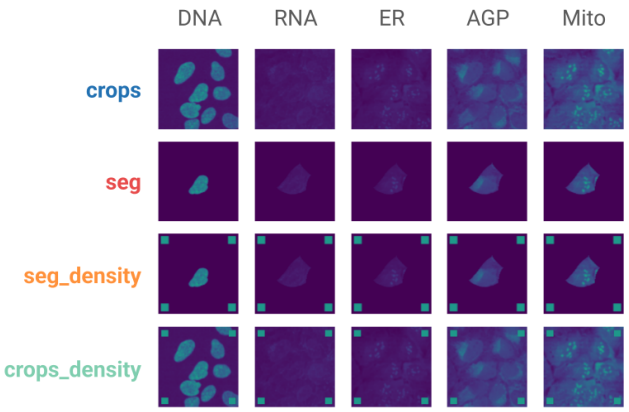


Figure S3: Dataset variants for evaluating background and density effects. The same cell shown across the five Cell Painting channels for each variant: `crops`: full crops with background, `seg`: segmentation-masked single cells, `seg_density`: segmented with cell-density patch, `crops_density`: crops with cell-density patch. Corner patches encode field-of-view cell count as pixel intensity.

A.2 COMPOUND ACTIVITY CLASSIFICATION

To ensure a balanced representation of morphologically active and inactive compounds in every batch during model training, we computed a per-compound activity score based on the Mahalanobis distance from DMSO controls. For this, we processed the CellProfiler features of only source 4 as described in (Chandrasekaran et al., 2023) and (Serrano et al., 2025), arriving at 599 features. Then, for each plate separately, we calculated the Mahalanobis distance of each well from the local control distribution of DMSO-treated wells in PCA space. Distances were aggregated per compound using the median across replicates. Statistical significance was assessed via an empirical null distribution constructed by resampling DMSO wells (10,000 permutations), followed by the Benjamini-Hochberg correction for multiple testing. Compounds with adjusted $p < 0.05$ were classified as morphologically active. This stratification was used during training batch construction to ensure the model encountered a balanced ratio of active and inactive compounds, preventing the loss from being dominated by uninformative negative controls. The distribution of active and inactive compounds in the integrated feature space is shown in Figure S4.

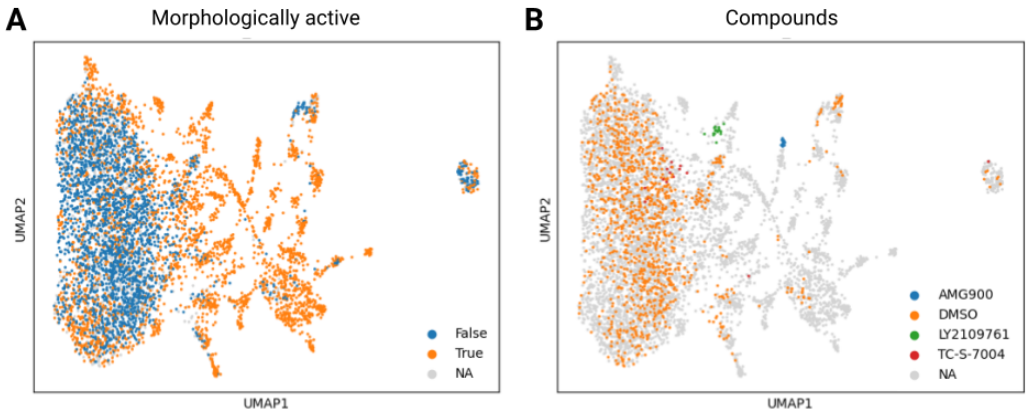


Figure S4: Morphological activity classification for training stratification. (A) UMAP of the scPoli-integrated feature space colored by morphological activity status. Compounds were classified as active (orange) if their median Mahalanobis distance from DMSO controls was significantly elevated (BH-adjusted $p < 0.05$); inactive compounds are shown in blue, and controls (NA) in gray. (B) Same embedding highlighting selected compounds: DMSO (negative control), and three examples of active compounds that form distinct clusters in morphological space.

A.3 MOLECULAR FINGERPRINT COMPUTATION

To represent compounds for contrastive training, we computed molecular fingerprints from InChI identifiers provided in the JUMP-CP metadata. We first standardized each molecule using RDKit's MolStandardize module: selecting the largest fragment, canonicalizing tautomers, and applying strict sanitization where possible. For molecules with hypervalent phosphorus atoms, a common source of sanitization failures, we implemented a repair step that converts redundant P=O double bonds to P-O-$^-$ single bonds. Successfully standardized molecules were then converted to canonical, isomeric SMILES for fingerprint computation.

We chose ECFP4 fingerprints (radius 2, 2048 bits) as the chemical representation for model training rather than learned embeddings such as ChemBERTa. This decision was motivated by three considerations: (1) using a fixed, deterministic chemical representation isolates the image encoder as the sole variable under study across our four dataset conditions; (2) ECFP substructures map directly to interpretable chemical fragments, facilitating analysis of learned chemical–morphology correspondences; and (3) static fingerprints ensure consistent gradients across experimental conditions, avoiding confounded training dynamics that could arise from co-adaptation between jointly trained encoders.

All fingerprints were computed as hashed count vectors with $\log(1 + x)$ transformation to compress count magnitudes. In addition to ECFP4, we computed several alternative representations for release

alongside the dataset: ECFP6 (radius 3, 4096 bits), FCFP4 (radius 2, 2048 bits, pharmacophoric features), multi-radius ECFP (radii 1–3 with 1024 bits each concatenated, 3072 bits), physicochemical and VSA descriptors, and 3D WHIM descriptors from energy-minimized conformers (MMFF94s force field, lowest-energy conformer from 3 candidates). These were not used in this study but are provided with the datasets to facilitate future benchmarking and alternative modeling approaches.

A.4 Phenotypic activity assessment with copairs

copairs (Kalinin et al., 2025) is Python toolkit designed to evaluate image-based profiles, based on average precision (AP) for retrieval task. We specifically use *phenotypic activity* metric from this toolkit, that identifies perturbations that are phenotypically different from negative controls and that is reproducible across replicates of a given perturbation (one at a time). The output of the metric is AP for the replicate retrieval task against a null distribution of negative controls and p-value. Perturbations that pass adjusted p-value 0.05 threshold, are considered active and we report the share of such perturbations for each model in Figure 3 and AP with p-values in Figure S5.

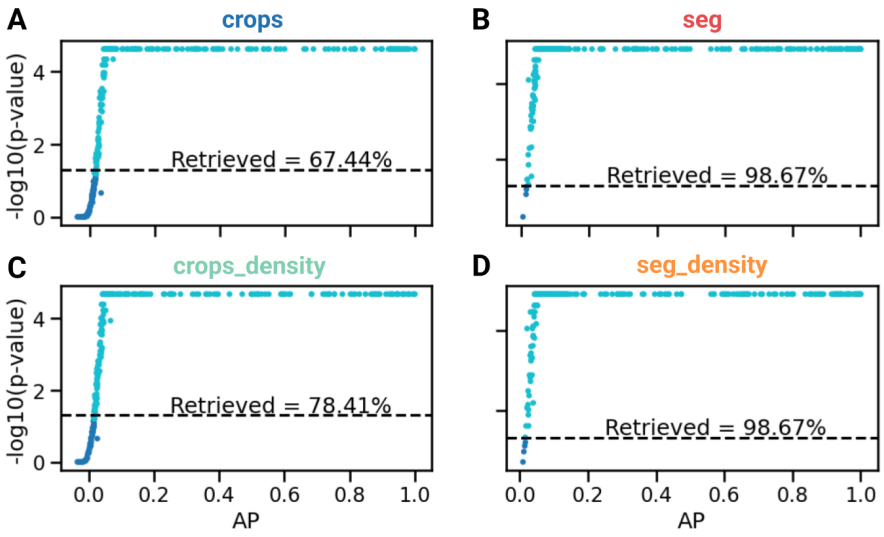


Figure S5: Phenotypic activity detection via copairs replicate retrieval. Each point represents one compound's average precision (AP) and significance (-log10 p-value). Horizontal dashed line marks p=0.05 threshold. Percentage indicates fraction of compounds with significant replicate agreement. Segmented models detect 98.67% of active compounds vs 67.44%-78.41% for crops-based models, demonstrating that density-driven clustering causes crops models to miss compounds whose replicates span different density regimes. Panel A shows results for `crops`, panel B for `seg`, panel C for `crops_density` and panel D for `seg_density`.

A.5 CONTRASTIVE-LOSS AND CHEMICAL-SIMILARITY-GUIDED FINE-TUNING

We frame the training objective not as learning a joint image-molecule embedding space, but as using chemical structure to guide the fine-tuning of the image encoder. The molecule encoder, a single linear projection of precomputed ECFP4 fingerprints, serves as a fixed anchor providing training signal. This design isolates the image encoder as the sole learned component and ensures that chemical similarity provides a deterministic supervision signal.

Since the generated **CP-BG-1M** dataset contains only 302 unique chemical perturbations, we employ a multi-positive contrastive loss. While standard CLIP defines exactly one positive per anchor, this study extends this to a multi-positive formulation where all images of the same compound serve as hard positives. Additionally, we augmented by soft positives from chemically similar compounds. We calculate the Tanimoto coefficient between all compounds as a measure of chemical similarity as follows:

$$T(\mathbf{a}, \mathbf{b}) = \frac{\sum_i a_i \wedge b_i}{\sum_i a_i + \sum_i b_i - \sum_i a_i \wedge b_i} \tag{1}$$

yielding a precomputed similarity matrix $\mathbf{S} \in [0,1]^{C \times C}$ for all C training compounds.

Since chemical similarity doesn't guarantee similar function, we only let the soft-positive weight to a small degree. For a batch of N images spanning K compounds with label encoding $\mathbf{M} \in \{0,1\}^{N \times K}$, hard positives are $\mathbf{W}_{\text{hard}} = \mathbf{M}\mathbf{M}^\top$. Soft positives from each compound's top-κ chemical neighbors (κ=5, $\tau_{\min}$=0.25) are weighted as:

$$w_{kk'} = \left(\frac{T(\mathbf{f}_k, \mathbf{f}_{k'}) - \tau_{\min}}{1 - \tau_{\min}} \right)^{\gamma} \cdot \lambda \cdot \alpha(t) \tag{2}$$

with γ=1, λ=0.5, and curriculum $\alpha(t)$ ramping linearly from 0 to 2 over steps 1k - 3k.

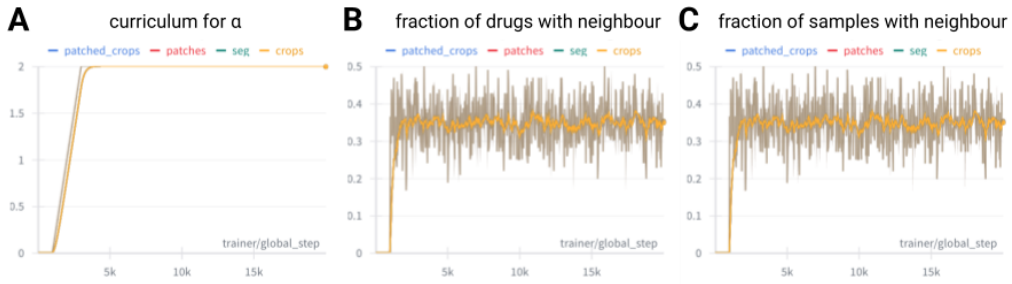


Figure S6: Chemical-similarity neighbor curriculum and neighbor statistics. (A) Curriculum factor $\alpha(t)$ controlling the strength of chemically similar soft-positive neighbors, ramped from 0 to 2 between training steps 1k - 3k and held constant thereafter. (B) Fraction of unique compounds per batch that have at least one eligible chemical neighbor present in the same batch. (C) Fraction of samples per batch that have at least one such neighbor sample available for soft-positive weighting. EMAs for curves are shown for all four datasets, but are the same due to identical ordering.

The combined weights $\mathbf{W} = \mathbf{W}_{\text{hard}} + \mathbf{M}\mathbf{W}_{\text{chem}}\mathbf{M}^\top$ yield the bidirectional loss:

$$\mathcal{L} = -\frac{1}{2N} \sum_i \left[\log \sum_j w_{ij} \cdot \text{softmax}(\mathbf{L})_{ij} + \log \sum_j w_{ji} \cdot \text{softmax}(\mathbf{L}^\top)_{ji} \right] \tag{3}$$

where $\mathbf{L} = \exp(\theta) \cdot \mathbf{Z}_{\text{img}} \mathbf{Z}_{\text{mol}}^\top$ with learned temperature θ.

A.6 EXAMPLE OF COMPOUND CLUSTERING IN PROCESSED UMAPS

The Target2 plates contain a set of control compounds with known phenotypes, these are shown in Figure S7.

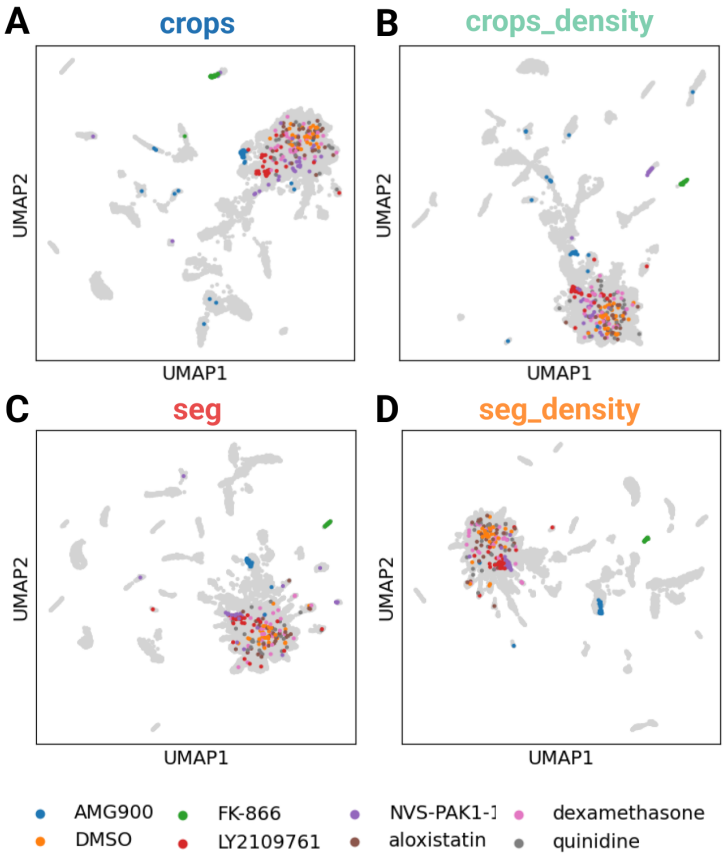


Figure S7: UMAPs of well-level embeddings highlighting Target2 control compounds across the four **CP-BG-1M** input views. Each panel shows the same embedding pipeline used throughout the manuscript (mean-pooled per well, spherized on DMSO controls, Harmony-integrated in PCA space), visualized with UMAP for (A) `crops`, (B) `crops_density`, (C) `seg`, and (D) `seg_density`. Colored points denote wells treated with the same control compounds listed in the legend (including DMSO); all other wells are shown in light gray.